A COST-SENSITIVE APPROACH TO TERNARY CLASSIFICATION

A Thesis

Submitted to the Faculty of Graduate Studies and Research

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Computer Science

University of Regina

By

Bing Zhou

Regina, Saskatchewan

July, 2012

UNIVERSITY OF REGINA

FACULTY OF GRADUATE STUDIES AND RESEARCH

SUPERVISORY AND EXAMINING COMMITTEE

Bing Zhou, candidate for the degree of Doctor of Philosophy in Computer Science**,** has presented a thesis titled, ***A Cost-Sensitive Approach to Ternary Classification,*** in an oral examination held on July 16, 2012. The following committee members have found the thesis acceptable in form and content, and that the candidate demonstrated satisfactory knowledge of the subject material.

External Examiner:	*Dr. An Aijun, York University
Supervisor:	Dr. Yiyu Yao, Department of Computer Science
Committee Member:	Dr. Malek Mouhoub, Department of Computer Science
Committee Member:	Dr. Howard Hamilton, Department of Computer Science
Committee Member:	Dr. Donald Stanley, Department of Mathematics & Statistics
Chair of Defense:	Dr. Ronald Camp, Kenneth Levene Graduate School

*Participated via SKYPE

Abstract

Bayesian inference and rough set theory provide two approaches to data analysis. There are close connections between the two theories as they both use probabilities to express uncertainties and knowledge about data. Several proposals have been made to apply Bayesian approaches to rough sets. This thesis draws results from two probabilistic rough set models, namely, decision-theoretic rough set models (DTRSM) and confirmation-theoretic rough set models (CTRSM) to propose a new Bayesian rough set model (BRSM) for cost-sensitive ternary classification. I argue that although the two classes of models share many similarities in terms of making use of Bayes' theorem and a pair of thresholds to produce three regions, their semantic interpretations and hence intended applications are different. By integrating the two, I propose a unified model of Bayesian rough sets and apply the model to develop ternary classification. In developing the Bayesian rough set model, I focus on three fundamental issues, namely, the interpretation and calculation of a pair of thresholds, the estimation of probabilities, and the interpretation of the three regions used by rough set theory.

Email spam filtering is used as a real world application to show the usefulness of the proposed model. Instead of treating email spam as a binary classification problem, I argue that a three-way decision approach will provide a way that is more meaningful to users for precautionary handling of their incoming emails. Three email folders instead of two are produced in a three-way spam filtering system. A suspected folder is added

to allow users to further examine suspicious emails, thereby reducing the misclassification rate. In contrast to other ternary email spam filtering methods, my approach focuses on issues that are less studied in previous work, that is, the computation of required thresholds to define the three email categories and the interpretation of the cost-sensitive characteristics of spam filtering. Instead of having the user supply the thresholds based on their intuitive understanding of the intolerance for errors, I systematically calculate the thresholds based on the decision-theoretic rough set model. The cost of making the decision is interpreted as the loss function for Bayesian decision theory. The final decision is made by choosing the possible decision for which the overall cost is minimum. Experimental results on several benchmark datasets show that the new approach reduces the error rate of misclassifying a legitimate email to spam and demonstrates a better performance from a cost perspective.

Finally, I propose and investigate two extensions of the basic model. One concerns multi-class classification and the other concerns multi-stage ternary classification. These two extensions make the model more applicable to solving real world problems.

Acknowledgements

There are lots of people I would like to thank. First, and foremost, I must thank my supervisor, Dr. Yiyu Yao, for convincing me to become a Ph.D. student and for always making sure that I received the kind of support that I needed along the way. Without his knowledge, perceptiveness and encouragement I would not have finished this thesis.

I would like to thank NSERC, the Faculty of Graduate Studies and Research and Dr. Yiyu Yao for supplying my funding. Without their support of the sciences, many valuable research efforts might not have grown beyond a passing thought.

Thank you to all of my friends and colleagues whose collaborative and personal support was essential for the completion of this research.

Finally, I would like to say a big 'thank you' to my family: my husband, my parents, parents-in-law, and my two lovely sons for their constant love, support, and encouragement throughout my graduate career, which has imbued everything in my life with value.

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

In this chapter, I discuss the scope, goal, methodology and contributions of the thesis. I identify several difficulties with binary classifications, give motivations for introducing ternary classification, and propose a cost-sensitive model of ternary classification by combining rough sets and Bayesian inference.

1.1 Difficulties with Binary Classification

In data mining and machine learning, classification is a typical example of supervised learning [58]. Given a collection of objects (i.e., a training set), where each object is described by a set of attributes and one of the attributes is the class, the goal of classification is to find a model for the class attribute as a function of the values of the other attributes such that previously unseen objects will be assigned to a class as accurately as possible. In supervised learning, we provide the algorithms with a set of inputs and their correct answers (classes); the algorithm learns the associations

between the inputs and the outputs.

Many classification problems have been treated as binary classifications. Even when the class attributes in a classification problem have more than two values (multi-class), it can be translated into a binary classification. Learning methods, such as decision trees [50, 71, 72], support vector machines [1, 12], and neural networks [87], are suitable for learning binary classifiers. In binary classification, the class attribute of the training set has two values (i.e., the actual class). The classifier assigns one of the two classes (i.e., the predicted class) to each object based on its properties. For example, in a medical decision-making process, assume that a patient either has a disease (i.e., in the class) or does not have the disease (i.e., not in the class), a classifier needs to predict whether a patient has the disease based on his/her symptoms. This simple formulation has a few limitations.

First, binary classification requires a definite decision. However, for many real world tasks, it is difficult to make definite decisions. For instance, a doctor may not be able to make a positive diagnosis right away when a patient's symptoms are insufficient to support a particular disease. Instead of making a decision that could produce faulty results, a better choice is to perform diagnostic tests to collect more evidence. Many real world decision-making problems become easier and more efficient by adding a third option.

Second, one goal of a typical classification learning algorithm is to minimize the number of misclassified examples [27, 58, 71, 72]. This goal may not be appropriate if the costs of different types of misclassification vary. It may be preferable to incorporate such costs into the classification process. Cost-sensitive learning is one of

the challenging problems in the current stage of data mining and machine learning research [2, 18, 19, 42]. As an example of different kinds of classification errors having different costs, in medical diagnosis, not treating a cancer patient could cause death or injury. On the other hand, resources will be wasted and harm will be done by unnecessarily treating a patient who does not have cancer. In credit card fraud detection, failure to detect fraud could be very expensive; again resources will be wasted for investigating non-fraud. Therefore, the goal of cost-sensitive learning is to minimize the expected cost of misclassification. This may be difficult to achieve in binary classifications because a forced definite decision may come with a larger cost.

1.2 Motivations of Ternary Classification

In order to address the limitations of binary classification, a cost-sensitive three-way decision approach is proposed in this thesis. The main advantages of adding a third option are explained by the following two examples.

Example 1 *Consider a story given in the book by Savage [26, 81]. "Your wife has just broken five good eggs into a bowl when you come in and volunteer to finish making the omelet. A sixth egg, which for some reason must either be used for the omelet or wasted altogether, lies unbroken beside the bowl. You must decide what to do with this unbroken egg."*

Consider first a binary-decision scenario, namely, *to break it into the bowl containing the other five, or to throw it away without inspection.* Depending on the state of the egg, each of these two actions will have some consequences and they are given by

the following 2×2 pay off matrix:

	Good	Rotten
Break into bowl	six-egg omelet	no omelet, and five good eggs destroyed
Throw away	five-egg omelet, and one good egg destroyed	five-egg omelet

If the sixth egg is good and you made the right decision by breaking it into the bowl, your wife will be happy to see a finished six-egg omelet. If the sixth egg is bad and you made the right decision of throwing it away, your wife will still be happy to see a finished five-egg omelet. Wrong decisions cost much more. In one case, if the sixth egg is good but you throw it away, one good egg will be wasted. In the worst case, the sixth egg is bad but you break into the bowl, five good eggs will be destroyed, and you had better think about how to explain your decision to your wife. In binary decision-making, each action comes with the cost of ruining some good eggs.

Is there a better way of doing this? A better choice would be not to break the sixth egg into the bowl or throw it away, but to *break it into a saucer* for inspection. By adding this third action, the consequence changed to the following 3×2 pay off matrix:

	Good	Rotten
Break into bowl	six-egg omelet	no omelet, and five good eggs destroyed
Throw away	five-egg omelet, and one good egg destroyed	five-egg omelet
Break into saucer	six-egg omelet, and a saucer to wash	five-egg omelet, and a saucer to wash

By further examining the state of the sixth egg, the cost is reduced from ruining good eggs to washing a saucer.

A similar three-way decision making process is used in both the omelet example and the previous medical example, where the doctor needed to decide whether to treat the patient, not treat the patient, or perform diagnostic tests to collect more information [62]. We can extend this three-way process to general decision-making problems for making acceptance, rejection, or deferment decisions.

Example 2 *In spam filtering systems, there are two actual classes of email (i.e., legitimate and spam) and the spam filter needs to decide whether to accept a new email as legitimate or reject it as spam. As shown in Figure 1.1-(a), assume that the filter uses a discriminant function* $f(x)$ *with* $0 \leq f(x) \leq 1$ *to measure the legitimacy of email* x *and then compares the value of* $f(x)$ *with a threshold* γ *to decide whether to accept or reject the email.*

If the value of $f(x)$ is close to 1, the spam filter is more likely to accept x as legitimate. If the value of $f(x)$ is close to 0, the spam filter is more likely to reject x as spam. If the value of $f(x)$ is neither close to 1 nor close to 0 (e.g., $f(x) = 0.5$),

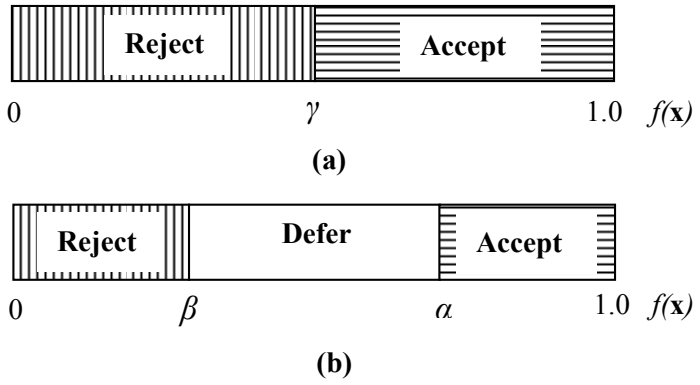

Figure 1.1: (a) Binary spam filtering with a single threshold. (b) Ternary spam filtering with a pair of thresholds.

it is difficult to decide whether to accept or reject x. In other words, if a forced binary-decision is made, neither decision will be ideal.

A better choice would be neither to accept nor to reject email x, but to instead make a further examination by collecting additional information. As shown in Figure 1.1-(b), when the value of $f(x)$ is in between a pair of thresholds α and β (i.e., $\beta < f(x) < \alpha$), x will be classified into the suspected folder. A pair of thresholds is used in three-way classification to produce three email folders instead of two. The first threshold α determines whether it is necessary for a re-examination and the second threshold β determines whether to reject the email. Three-way email filtering is especially useful when users view their emails under a time constraint. They may view the accepted folder immediately, delay the processing of suspected folders, and delete the rejected folder without viewing.

Three-way decision making, also known as ternary classification has been studied and applied in fields such as medical decision-making [54, 62, 83], social judgement

theory [86], hypothesis testing in statistics [105], management sciences [30, 109], and peering review process [106]. For instance, Wald [105] considered a sequential hypothesis testing model, in which a pair of thresholds is used for accepting a hypothesis, rejecting a hypothesis, or further testing based on two thresholds on probability values. Pauker and Kassirer [62] proposed a threshold approach to clinical decision making. A pair of a "testing" threshold and a "test-treatment" threshold on probability is used, with testing threshold determining whether to perform a diagnostic test on a patient, and the test-treatment threshold determining whether to treat the patient immediately. Robinson [77] suggested adding a boundary region marked unsure in email spam filtering problem, in addition to the binary legitimate/spam classification results to reduce the misclassification error. Similar approaches can also be found in many other fields and disciplines, such as data warehouse [92], information retrieval [49], Bayesian significant tests in statistical process control [109], and decision making in environment management [30]. A theory of three-way decision making was recently proposed by Yao [123].

1.3 Overview of the Proposed Approach

The proposed cost-sensitive ternary classifier is constructed by drawing connections between two well-known data analysis tools, namely, Bayesian inference and rough set theory.

Bayesian inference offers a collection of statistical methods for data analysis [6, 8]. In Bayesian data analysis, a priori probability is used to capture our belief about an

event or a hypothesis before observing the evidence or data, and Bayes' theorem is used to update the a priori probability into an a posteriori probability through a likelihood function when evidence becomes available. A classifier based on Bayesian methods usually makes binary decisions by choosing the class with the higher a posteriori probability.

Pawlak rough set theory provides another approach to data analysis [63, 64]. It is a method for discovering knowledge based on approximating a crisp set (i.e., conventional set) in terms of a pair of sets called the lower and upper approximations of the set or equivalently three regions. Probabilistic rough set models [68, 89–91, 116, 124, 125, 147] consider a pair of thresholds on probabilities for defining probabilistic approximations. Based on set approximations, the object space can be divided into three regions, called the positive region, the negative region and the boundary region. Decision rules generated from these three regions can be used to build a ternary classifier.

A number of proposals have been made to apply Bayesian approaches to rough sets. Yao et al. [116, 124, 125] proposed a decision-theoretic rough set (DTRS) model, in which the pair of thresholds can be systematically calculated based on the well established Bayesian decision theory. Ślęzak and Ziarko [89–91] proposed a Bayesian rough set (BRS) model and a rough Bayesian model to explicitly use Bayes' theorem in formulating a probabilistic rough set model. In a series of papers [33, 65, 66], Pawlak used Bayes' theorem to explain the probabilistic relationship between conditions and decisions in decision rules. Each of these studies focuses on a specific perspective on Bayesian approaches to rough sets. The implications of Bayesian decision theory and

Bayesian inference for rough sets have not been fully explored.

The main advantages of combining Bayesian inference and rough sets for data analysis are given below. We can apply the probabilistic and statistical methods provided by Bayesian inference to quantify uncertainty in data and determine the class with the lowest expected cost; in the same framework, we can extend binary decision-making to ternary decision-making based on the three probabilistic regions defined in rough set theory.

1.4 Contributions of the Thesis

In this thesis, a cost-sensitive approach is introduced by fully exploring the implications of Bayesian decision theory and Bayesian inference for rough set theory. More precisely, the results from decision-theoretic rough set models are used to interpret and compute a pair of thresholds based on Bayesian decision theory, with the aid of more practically operable notions such as cost, risk, benefit etc.; the techniques used in existing Bayesian rough set models are adopted and extended to estimate probability based on Bayes' theorem and inference; and a framework of three-way decisions is used to interpret rules from the three regions for building ternary classifiers. The main contributions are summarized below.

A Complete Model of Bayesian Rough Sets

In the existing studies on Bayesian approaches to rough sets, attention has been mainly paid to mathematical constructions and formal properties of various notions,

but the semantic of these models has not been explicitly studied and made clear. Although phrases such as "rough Bayesian model" and "Bayesian rough sets" have been introduced and used in the literature, the implications of Bayesian decision theory and Bayesian inference for rough sets have not been fully explored. In this thesis, a complete model of Bayesian rough sets is introduced.

A Framework for Building Cost-Sensitive Ternary Classifiers

The three regions defined in rough set theory provide a formal way to model the three-way decisions for building ternary classifiers, which are complementary to binary classifiers. More specifically, the positive, negative and boundary regions are viewed as the regions of acceptance, rejection and deferment in a ternary classification. Building on these ideas, this thesis introduces a framework of cost-sensitive ternary classification.

A Cost-Sensitive Approach to Email Spam Filtering

Email spam filtering is used as a real world application to show the usefulness of the proposed model. Two issues that are rarely studied in existing email spam filtering systems are addressed: the computation of required thresholds and the interpretation of the cost-sensitive characteristics of spam filtering. Different sets of thresholds are produced by varying the values of loss functions. The final decision is made by choosing the possible decision for which the overall cost is minimum. Experiment results show that the new approach demonstrates better performances in cost-sensitive evaluations. The new approach can also be applied to the elimination of other web-based

material similar to spam, such as web pages and blogs with slight modifications.

Multi-Class Classification and Adaptive Learning

As extensions of the proposed model, two extensions associated with three-way decision classifications are also proposed in this thesis. One concerns a formulation of multi-class classification and the other concerns an adaptive learning algorithm that automatically classifies the deferred examples.

1.5 Thesis Organization

Chapter 2 reviews the main features of probabilistic rough set models. An understanding of unsolved problems in existing models enables me to show that there is a lack of a complete and well-developed Bayesian rough set model, which motivates the work described in this thesis.

Chapter 3 presents a unified framework of three-way Bayesian data analysis. Section 3.1 discusses two applications of Bayesian inference. Section 3.2 explains the computation of thresholds based on decision-theoretic rough set models. Section 3.3 explains the foundation of sequential three-way decisions based on confirmation-theoretic rough set models. Section 3.4 describes how to estimate probabilities in both models based on Bayesian inference. Finally, Section 3.5 provides a framework of three-way decisions to construct and interpret rules from the three regions for building ternary classifiers.

Chapter 4 demonstrates the usefulness of the proposed framework by applying it

to a typical example of text classification, namely, email spam filtering. In contrast to other ternary email spam filtering methods, my approach focuses on issues that are rarely studied in previous work.

Chapter 5 evaluates the performance of the proposed approach by comparing it with the traditional naive Bayesian spam filter and two existing ternary spam filters. The experimental results on several benchmark datasets show that the new approach reduces the rate of misclassifying a legitimate email to spam and demonstrates a better performance from a cost perspective.

Chapter 6 investigates two extensions of the proposed model. One is multi-class classification and the other is sequential decision-making.

Chapter 7 summarizes the contribution of the thesis and outlines some possible future research directions.

Chapter 2

A CRITICAL REVIEW OF PROBABILISTIC ROUGH SETS

In this chapter, I review main results of probabilistic rough set models. I examine carefully the salient features of each model and its unsolved problems. Such an understanding enables me to show that there is a lack of a complete and well-developed Bayesian rough set model and hence the contributions of this thesis.

2.1 A Brief History

Rough set theory was introduced by Pawlak [63, 64] in early 1980s as a tool for analyzing data represented in a tabular form. A central notion of the theory is the indiscernibility of objects and induced lower and upper approximations of a set due to indiscernibility. The first probabilistic rough set (PRS) model, called the 0.5-probabilistic rough set model [116], was proposed by Wong and Ziarko [107] and

Pawlak et al. [68]. A threshold of 0.5 on probability is used to define probabilistic lower and upper approximations, or equivalently three probabilistic regions, of a set. The threshold 0.5 can be intuitively interpreted based on the notion of the majority rule. Yao et al. [116, 124, 125] proposed a generalized probabilistic model, called a decision-theoretic rough set (DTRS) model, by considering a pair of thresholds on probabilities for defining probabilistic approximations. The pair of thresholds can be systematically calculated based on the well established Bayesian decision theory and interpreted in terms of more practically operable notions such as cost, risk, benefit etc. Herbert and Yao [40] integrated game theory and decision-theoretic rough set model to introduce a game-theoretic rough set (GTRS) model.

Based on the notion of graded set inclusion, Ziarko [147] introduced a variable precision rough set (VPRS) model by using a pair of thresholds on a set-inclusion function. The derived approximations are equivalent to a special case of the decision-theoretic rough set model. The main results of VPRS model were later explicitly re-expressed in terms of thresholds on probability instead of a set-inclusion function [148].

In a series of papers [33, 65, 66], Pawlak advocated a research direction in studying connections between Bayesian methods and rough set approaches. Greco et al. [32] argued that it may be insufficient to consider only probability values when formulating a probabilistic rough set model. They introduced a parameterized model by considering a probability and a confirmation measure. Two pairs of thresholds are used for defining rough set three regions, one pair on the probability and the other pair for Bayesian confirmation measure. Ślęzak and Ziarko [89–91] conducted a series of studies by drawing a natural correspondence between the fundamental notions of

rough sets and statistics. Yao and Zhou [131, 132] introduced a naive Bayesian rough set (NBRS) model by modifying results from the standard naive Bayesian model for ternary classification.

One of the main objectives of this thesis is to weave together results from these existing studies and formulate a more complete model of Bayesian rough sets.

2.2 Pawlak Rough Set Theory

In Pawlak's rough set model [64], informations about a finite set of objects are represented in an information table with a finite set of attributes. Formally, an information table can be expressed as:

$$S = (U, At, \{V_a \mid a \in At\}, \{I_a \mid a \in At\}),$$

where

U is a finite nonempty set of objects called the universe,

At is a finite nonempty set of attributes,

V_a is a nonempty set of values for $a \in At$,

$I_a : U \to V_a$ is an information function.

The information function I_a maps an object in U to a value of V_a for an attribute $a \in At$, that is, $I_a(x) \in V_a$.

An equivalence relation can be defined with respect to a subset of attributes $A \subseteq$

At, denoted as R_A, or simply R,

$$\begin{aligned} xRy &\iff \forall_{a\in A} I_a(x) = I_a(y) \\ &\iff I_A(x) = I_A(y). \end{aligned}$$

The relation R is reflexive, symmetric and transitive. Two objects x and y in U are equivalent or indiscernible by the set of attributes A if and only if they have the same values on all attributes in A.

The equivalence relation R induces a partition of U, denoted by U/R. The subsets in U/R are called equivalence classes, which are the building blocks to construct rough set approximations. The equivalence class containing x is defined as:

$$[x] = \{y \in U \mid xRy\}.$$

One can use logic formulas defined in [137] to describe an equivalence class. In the simplest case, a subset $\{a\} \subseteq At$ contains one attribute a, $[x]_{\{a\}}$ can be described by an atomic formula $(a = v)$ indicating that all the objects in $[x]_{\{a\}}$ having the same attribute value v on attribute a, that is, $I_a(x) = v$. When A contains more than one attribute, $[x]_A$ can be described by the conjunction of attribute value pairs $\bigwedge_{i=1}^{n} I_{a_i}(x) = v_{a_i}$, or simply written as $\bigwedge_{i=1}^{n} v_{a_i}$. In other words, the description of an equivalence class $[x]_A$ can be represented by a tuple with n components, written as $Des(x) = (v_{a_1}, v_{a_2}, ..., v_{a_n})$. The characteristic of each object $x \in U$ can be represented by the description of its equivalence class.

Consider an equivalence relation R on U. The equivalence classes induced by the partition U/R are the basic blocks to construct Pawlak's rough set approximations. For a subset $C \subseteq U$, the lower and upper approximations of C with respect to U/R

are defined by:

$$\begin{aligned}
\underline{apr}(C) &= \{x \in U \mid [x] \subseteq C\} \\
&= \bigcup\{[x] \in U/R \mid [x] \subseteq C\}; \\
\overline{apr}(C) &= \{x \in U \mid [x] \cap C \neq 0\} \\
&= \bigcup\{[x] \in U/R \mid [x] \cap C \neq \emptyset\}.
\end{aligned} \tag{2.1}$$

Based on the rough set approximations of C, one can divide the universe U into three pair-wise disjoint regions: the positive region $\mathrm{POS}(C)$ is the union of all the equivalence classes that is included in C; the negative region $\mathrm{NEG}(C)$ is the union of all equivalence classes that have an empty intersection with C; and the boundary region $\mathrm{BND}(C)$ is the difference between the upper and lower approximations:

$$\begin{aligned}
&\mathrm{POS}(C) = \underline{apr}(C), \\
&\mathrm{NEG}(C) = U - \mathrm{POS}(C) \cup \mathrm{BND}(C), \\
&\mathrm{BND}(C) = \overline{apr}(C) - \underline{apr}(C).
\end{aligned} \tag{2.2}$$

It can be verified that $\mathrm{NEG}(C) = \mathrm{POS}(C^c)$, where C^c is the complement of C.

The relationships between the basic notions of rough sets are illustrated in Figure 2.1. A set is said to be rough if its boundary region is non-empty, otherwise the set is crisp. Rough set theory thus provides a way to classify objects into three regions based on their descriptions. That is, rough set theory leads to a ternary classifier [121], which is complementary to widely used binary classifiers. For this reason, I prefer the formulation by three regions to the formulation by a pair of lower and upper approximations.

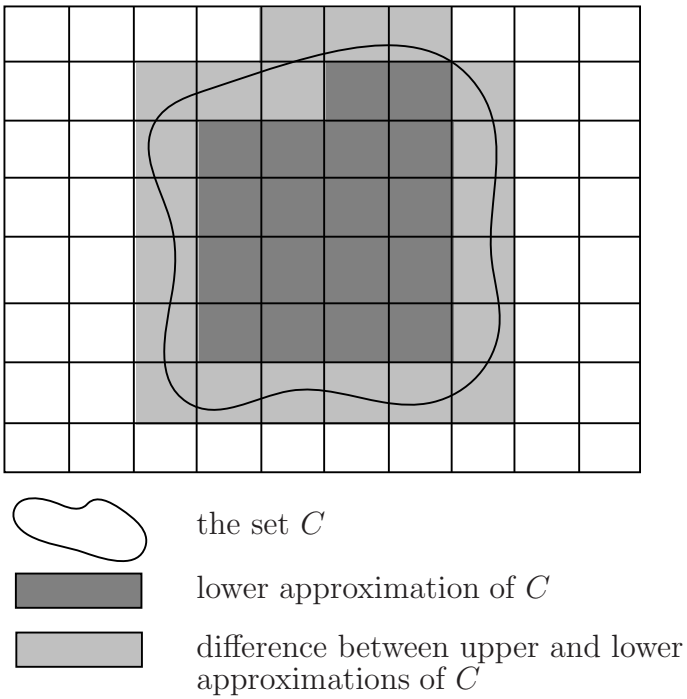

Figure 2.1: Lower and upper approximations in rough sets

Example 3 *Table 2.1 is an information table taken from [46]. There are seven objects, three attributes* $\{Age, Lower\ Extremity\ Motor\ Score\ (LEMS), Walk\}$*, and Walk is the class attribute with two values. In Table 2.1, if attribute* $A = \{Age\}$ *is chosen, we can obtain the following family of equivalence classes, or a partition of* U*:*

$$U/R_{\{Age\}} = \{\{x_1, x_2, x_6\}, \{x_3, x_4\}, \{x_5, x_7\}\}.$$

If we consider attribute $A = \{LEMS\}$*, the family of equivalence classes is:*

$$U/R_{\{LEMS\}} = \{\{x_1\}, \{x_2\}, \{x_3, x_4\}, \{x_5, x_6, x_7\}\}.$$

Table 2.1: An information table

	Age	LEMS	Walk
x_1	16-30	50	yes
x_2	16-30	0	no
x_3	31-45	1-25	no
x_4	31-45	1-25	yes
x_5	46-60	26-49	no
x_6	16-30	26-49	yes
x_7	46-60	26-49	no

If we consider attribute $A = \{Age, LEMS\}$*, the family of equivalence classes is:*

$$U/R_{\{Age,LEMS\}} = \{\{x_1\}, \{x_2\}, \{x_3, x_4\}, \{x_5, x_7\}, \{x_6\}\}.$$

By choosing different sets of attributes from the information table, we can get different partitions. For partition $U/R_{\{Age,LEMS\}}$*, given a set*

$$C = \{x \in U \mid I_{Walk}(x) = yes\} = \{x_1, x_4, x_6\},$$

the lower and upper approximation of C *are:*

$$\underline{apr}(C) = \{x_1, x_6\},$$
$$\overline{apr}(C) = \{x_1, x_3, x_4, x_6\}.$$

The three regions are:

$$\text{POS}(C) = \underline{apr}(C) = \{x_1, x_6\},$$

$$\text{BND}(C) = \overline{apr}(C) - \underline{apr}(C) = \{x_3, x_4\},$$

$$\text{NEG}(C) = U - \text{POS}(C) \cup \text{BND}(C) = \{x_2, x_5, x_7\}.$$

Set C is rough since the boundary region is not empty.

2.3 0.5-Probabilistic Rough Set Model

The positive and negative regions defined based on equation (2.2) in Pawlak rough sets must be completely certain. An equivalence class is in the positive region if and only if it is fully contained in the set. This may be too restrictive to be practically useful in real applications. An attempt to use probabilistic information for approximations was suggested by Pawlak, Wong, Ziarko [68] to allow some tolerance of errors, in which the degrees of overlap between equivalence classes $[x]$ and a set C to be approximated are considered. A conditional probability is used to state the degree of overlapping and is defined as:

$$Pr(C \mid [x]) = \frac{|C \cap [x]|}{|[x]|}, \tag{2.3}$$

where $|\cdot|$ denotes the cardinality of a set, and the conditional probability is written as $Pr(C \mid [x])$ representing the probability that an object belongs to C given that the object is described by $[x]$. Here $[x]$ is a simplified form of $Des(x)$ describing the properties of x. Pawlak and Skowron [67] called the conditional probability a rough membership function.

According to the above definitions, the three regions defined in equation (2.2) can be equivalently defined by:

$$\begin{aligned} \text{POS}(C) &= \{x \in U \mid Pr(C \mid [x]) = 1\}, \\ \text{BND}(C) &= \{x \in U \mid 0 < Pr(C \mid [x]) < 1\}, \\ \text{NEG}(C) &= \{x \in U \mid Pr(C \mid [x]) = 0\}. \end{aligned} \tag{2.4}$$

They are defined by using the two extreme values, 0 and 1, of probabilities. They are of a qualitative nature; the magnitude of the value $Pr(C \mid [x])$ is not taken into account.

The 0.5-probabilistic rough set model [68] is based essentially on the majority rule. An object x is put into the positive region of set C if the majority of its equivalent classes $[x]$ are in C. That is,

$$\begin{aligned} \text{POS}_{0.5}(C) &= \{x \in U \mid Pr(C \mid [x]) > 0.5\}, \\ \text{BND}_{0.5}(C) &= \{x \in U \mid Pr(C \mid [x]) = 0.5\}, \\ \text{NEG}_{0.5}(C) &= \{x \in U \mid Pr(C \mid [x]) < 0.5\}. \end{aligned} \tag{2.5}$$

The boundary region consists of those objects whose conditional probabilities are exactly 0.5, which represents maximal uncertainty.

2.4 Decision-Theoretic Rough Set Model

Yao et al. [116, 124, 125] introduced a more general probabilistic model, called a decision-theoretic rough set (DTRS) model, in which a pair of thresholds α and β

with $\alpha > \beta$ on the probability is used to define three probabilistic regions. The (α, β)-probabilistic positive, boundary and negative regions are defined by [124, 125]:

$$\begin{aligned} \mathrm{POS}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C \mid [x]) \geq \alpha\}, \\ \mathrm{BND}_{(\alpha,\beta)}(C) &= \{x \in U \mid \beta < Pr(C \mid [x]) < \alpha\}, \\ \mathrm{NEG}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C \mid [x]) \leq \beta\}. \end{aligned} \tag{2.6}$$

The 0.5-probabilistic rough set model is a special case of decision-theoretic rough set model, which is formulated based on a particular choice of α and β values, namely, $\alpha = \beta = 0.5$.

Unlike the qualitative Pawlak approximations, probabilistic approximations introduce certain levels of error in both the positive and boundary regions. More precisely, Pawlak regions and (α, β)-probabilistic regions are linked together by:

$$\begin{aligned} &\mathrm{POS}(C) \subseteq \mathrm{POS}_{(\alpha,\beta)}(C), \\ &\mathrm{BND}_{(\alpha,\beta)}(C) \subseteq \mathrm{BND}(C), \\ &\mathrm{NEG}(C) \subseteq \mathrm{NEG}_{(\alpha,\beta)}(C). \end{aligned} \tag{2.7}$$

Probabilistic three regions may be interpreted in terms of costs of different types of classification decisions [119, 121]. One obtains larger positive and negative regions by introducing classification errors in trade of a smaller boundary region so that the total classification cost is minimum [122]. Considering the errors introduced, the three regions are semantically interpreted as the following three-way decisions [119, 121, 122]. We accept an object x to be a member of C if the conditional probability is greater than or equal to α, with an understanding that it comes with an $(1 - \alpha)$-level acceptance error and associated cost. We reject x to be a member of C if the

conditional probability is less than or equal to β, with an understanding that it comes with an β-level of rejection error and associated cost. We neither accept nor reject x to be a member of C if the conditional probability is between of α and β, instead, we make a decision of deferment. The boundary region does not involve acceptance and rejection errors, but it is associated with cost of deferment. The three probabilistic regions are obtained by considering a trade-off between various classification costs.

2.5 Bayesian Rough Set Model

Ślęzak and Ziarko [90, 91] introduced a Bayesian rough set (BRS) model. A priori probability $Pr(C)$ is used to replace 0.5 in the 0.5-probabilistic rough set model as a threshold for defining three regions:

$$\begin{aligned} \mathrm{POS}_B(C) &= \{x \in U \mid Pr(C \mid [x]) > Pr(C)\}, \\ \mathrm{BND}_B(C) &= \{x \in U \mid Pr(C \mid [x]) = Pr(C)\}, \\ \mathrm{NEG}_B(C) &= \{x \in U \mid Pr(C \mid [x]) < Pr(C)\}. \end{aligned} \tag{2.8}$$

They also suggested to compare two likelihood functions $Pr([x] \mid C)$ and $Pr([x] \mid C^c)$ directly when neither a posteriori probability $Pr(C \mid [x])$ nor a priori probability $Pr(C)$ is derivable from data. That is,

$$\begin{aligned} \mathrm{POS}_B(C) &= \{x \in U \mid Pr([x] \mid C) > Pr([x] \mid C^c)\}, \\ \mathrm{BND}_B(C) &= \{x \in U \mid Pr([x] \mid C) = Pr([x] \mid C^c)\}, \\ \mathrm{NEG}_B(C) &= \{x \in U \mid Pr([x] \mid C) < Pr([x] \mid C^c)\}. \end{aligned} \tag{2.9}$$

Ślęzak [89] further drew a natural correspondence between the fundamental notions of rough sets and statistics. The set to be approximated corresponds to a hypothesis and an equivalence class to a piece of evidence; the three probabilistic regions correspond to the cases that the hypothesis is verified positively, negatively, or undecided based on the evidence. Based on such a correspondence, Ślęzak introduced a rough Bayesian model [89], in which probabilistic approximations are defined based on a pair of thresholds on the ratio of the a priori and a posteriori probabilities. More importantly, their studies explicitly used the Bayes' theorem in formulating a probabilistic model.

2.6 Parameterized Model based on Bayesian Confirmation

Greco et al. [32] introduced a parameterized rough set model by considering a pair of thresholds on a Bayesian confirmation measure, in addition to a pair of thresholds on probability. The Bayesian confirmation measure is denoted by $c([x], C)$ which indicates the degree to which an equivalence class $[x]$ confirms the hypothesis C. Given a Bayesian confirmation measure $c([x], C)$ and a pair of thresholds (s, t) with $t < s$, three (α, β, s, t)-parameterized regions are defined by:

$$\begin{aligned}
\mathrm{PPOS}_{(\alpha,\beta,s,t)}(C) &= \{x \in U \mid Pr(C \mid [x]) \geq \alpha \wedge c([x], C) \geq s\}, \\
\mathrm{PBND}_{(\alpha,\beta,s,t)}(C) &= \{x \in U \mid (Pr(C \mid [x]) > \beta \vee c([x], C) > t) \wedge \\
&\quad (Pr(C \mid [x]) < \alpha \vee c([x], C) < s)\},
\end{aligned}$$

$$\text{PNEG}_{(\alpha,\beta,s,t)}(C) \quad = \quad \{x \in U \mid Pr(C \mid [x]) \leq \beta \wedge c([x], C) \leq t\}. \tag{2.10}$$

There is no general agreement on a Bayesian confirmation measure. Choosing an appropriate confirmation measure for a particular application may not be an easy task. The ranges of the values of different confirmation measures are different. This makes it an even more difficult task to interpret and set the thresholds.

2.7 Variable Precision Rough Set Model

Ziarko [147] proposed variable precision rough set (VPRS) model. The basic idea of VPRS is based on a function that measures the relative degree of misclassification. Suppose X and Y are two sets, the degree of misclassification between X and Y is defined as:

$$mc(X,Y) = \begin{cases} 1 - \frac{|X \cap Y|}{|X|}, & |X| > 0, \\ 0, & |X| = 0. \end{cases}$$

Obviously,

$$X \subseteq Y \Longleftrightarrow mc(X,Y) = 0. \tag{2.11}$$

By generalizing this relation, one can get a z-majority relation: $0 \leq z < 0.5$

$$X \subseteq_z Y \Longleftrightarrow mc(X,Y) \leq z. \tag{2.12}$$

The z-majority relation can be explained by the degree of X included in Y, namely, the inclusion degree. Based on the z-majority relation, the z-positive, negative and

boundary regions are defined as:

$$\begin{aligned}
\mathrm{ZPOS}_z(C) &= \{x \in U \mid [x] \subseteq_z C\} \\
&= \{x \in U \mid mc([x], C) \leq z\}, \\
\mathrm{ZNEG}_z(C) &= (\mathrm{ZPOS}_z(C^c))^c \\
&= (\{x \in U \mid mc([x], C^c) \leq z\})^c \\
&= \{x \in U \mid mc([x], C) \geq 1 - z\}, \\
\mathrm{ZBND}_z(C) &= (\mathrm{ZPOS}_z(C) \cup \mathrm{ZNEG}_z(C))^c \\
&= \{x \in U \mid z < mc([x], C) < 1 - z\}. \qquad (2.13)
\end{aligned}$$

The VPRS model based on one threshold z is called the symmetric VPRS. By using a pair of thresholds (l, u), $0 \leq l < u \leq 1$, one can define a nonsymmetric VPRS model as [45]:

$$\begin{aligned}
\mathrm{ZPOS}_{(l,u)}(C) &= \{x \in U \mid mc([x], C) \leq l\}, \\
\mathrm{ZNEG}_{(l,u)}(C) &= \{x \in U \mid mc([x], C) \geq u\}, \\
\mathrm{ZBND}_{(l,u)}(C) &= \{x \in U \mid l < mc([x], C) < u\}. \qquad (2.14)
\end{aligned}$$

Symmetric VPRS model is a special case of nonsymmetric VPRS model, where $l = z$, $u = 1 - z$ and $z < 0.5$.

The relationships between probability and the relative degree of misclassification can be expressed as:

$$Pr(C \mid [x]) = \frac{|C \cap [x]|}{|[x]|} = 1 - mc([x], C). \qquad (2.15)$$

Based on equation (2.15), the corresponding relationship between decision-theoretic

rough set model and VPRS model can be defined as:

$$\begin{aligned}
\mathrm{POS}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C \mid [x]) \geq \alpha\} \\
&= \{x \in U \mid 1 - Pr(C \mid [x]) \leq 1 - \alpha\} \\
&= \{x \in U \mid mc([x], C) \leq l\} \\
&= \mathrm{ZPOS}_{(l,u)}(C), \\
\mathrm{NEG}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C \mid [x]) \leq \beta\} \\
&= \{x \in U \mid 1 - Pr(C \mid [x]) \geq 1 - \beta\} \\
&= \{x \in U \mid mc([x], C) \geq u\} \\
&= \mathrm{ZNEG}_{(l,u)}(C), \\
\mathrm{BND}_{(\alpha,\beta)}(C) &= \{x \in U \mid \beta < Pr(C \mid [x]) < \alpha\} \\
&= \{x \in U \mid 1 - \alpha < 1 - Pr(C \mid [x]) < 1 - \beta\} \\
&= \{x \in U \mid l < mc([x], C) < u\} \\
&= \mathrm{ZBDN}_{(l,u)}(C),
\end{aligned} \tag{2.16}$$

where $l = 1 - \alpha$ and $u = 1 - \beta$. From this point of view, VPRS is a special case of decision-theoretic rough set model.

2.8 Discussion

Each of the probabilistic models introduces three approximation regions. Although three regions are similar in form, they have different semantics interpretations. More specifically, the three regions defined respectively by Ślęzak et al. [90, 91] and Greco et al. [32–35] have a very different interpretation from those of the decision-theoretic

rough set models. It may not be appropriate to interpret the former as probabilistic approximations of C. Rather, they are interpreted as classification of evidences (i.e., equivalence classes). The details of the sematic difference are explained in Section 3.3.

A probabilistic rough set model must address at least the following three issues:

(i) interpretation and computation of thresholds;

(ii) estimation of conditional probability $Pr(C \mid [x])$;

(iii) interpretation and applications of three regions in data analysis.

The existing probabilistic rough set models focus mainly on the first issue, the estimation of the conditional probability has not received much attention. The definition of a rough membership function [67] is a simple way to do it, but is of limited value due to the requirement of a large-sized sample. Dembczyński et al. [15] suggested a statistical model in which probabilities are estimated based on the maximization of a likelihood function. In the naive Bayesian rough set (NBRS) model introduced by Yao and Zhou [131,132], the estimation of the a posteriori probability is translated into the estimation of the likelihood function based on Bayes' theorem and naive conditional independence assumption. Many studies of probabilistic rough sets concentrate more on a theoretical development. The interpretation and application of three regions for real world applications remain to be partially unsolved. The recent proposal of three-way decisions for interpreting three regions is a promising direction [121, 122].

Table 2.2 summarizes the results of a detailed comparison of probability rough set models. The differences of models lie in their choices of thresholds. None of the

Table 2.2: Comparison of probabilistic rough set models

RS models	main features	unsolved/partially solved issues
0.5-probabilistic RS model	one threshold	(ii)&(iii)
Decision-theoretic RS model	a pair of thresholds	(ii)
VPRS model	one threshold or a pair of thresholds	(i),(ii)&(iii)
Parameterized RS model	two pairs of thresholds	(i),(ii)&(iii)
Bayesian RS model	a priori probability as threshold	(ii)&(iii)

models solved the above mentioned three issues satisfactorily. These observations lead to the current study on developing a more complete Bayesian rough set model.

Studies on Bayesian rough sets [90, 91], rough Bayesian sets [89], and parameterized rough sets [32–35] introduce different types of three regions based on Bayesian statistics and Bayesian confirmation theory. In this thesis, I redefine the notion of Bayesian rough sets by only using the Bayes' theorem in the above three studies. My formulation of the model is based on probabilistic regions from decision-theoretic rough set models [116, 124, 125]. The rationale for this choice will be explained in the next chapter by drawing attention to the semantics differences between different types of probabilistic rough set models and difficulties in interpreting some of their notions.

One can easily apply the methodology in this thesis to develop a different version of a Bayesian rough set model with respect to any other definitions of three regions, by focusing on the three fundamental issues. With further investigations, it may be possible to produce a more grand Bayesian rough set model that encompasses these models as its special cases.

Chapter 3

A UNIFIED FRAMEWORK OF THREE-WAY BAYESIAN DATA ANALYSIS

In this chapter, we first discuss two applications of Bayesian inference. One is to draw conclusions based on available evidence and the other is the evaluation of evidence. The former can be formulated by using decision-theoretic rough sets [116, 124, 125] and the latter by using confirmation-theoretic rough sets [35, 90, 91]. By combining the two, we propose a unified framework of Bayesian rough sets. A cost-sensitive ternary classifier is built within the framework.

Our formulation of Bayesian-confirmation model is different from the proposed model by Greco et al [35]. They used both a probability function and a Bayesian confirmation measure, and defined three regions by a pair of thresholds on probability

and another pair of thresholds on the Bayesian confirmation. Their combination of probability and Bayesian confirmation leads to semantics difficulty. In contrast, we only use a Bayesian confirmation measure to define a Bayesian-confirmation model. We use the phrase "Bayesian confirmation model" while Greco uses "parameterized rough set model".

3.1 Overview

The basic ideas of the proposed Bayesian rough set model are explained by using examples.

3.1.1 Two Applications of Bayesian Inference

Bayesian inference uses probability for quantifying uncertainty in inferences. In Bayesian data analysis, a priori probability of a hypothesis is updated into a posteriori probability after observing some evidence [6]. Bayesian inference can be used to address two issues and, hence, leads to two types of applications. First, we use the degree to which evidence supports a hypothesis to classify objects based on their satisfiability of the hypothesis. That is, we classify an object as satisfying or not satisfying the hypothesis if the object positively supports or is against the hypothesis beyond a certain level. Second, we can evaluate the quality of different pieces of evidence (i.e., how much the a posteriori probability increases or reduces after observing the evidence). That is, we can either weight or select pieces of evidence according to their confirmations of the hypothesis.

My main focus in this thesis is on the first classification task. However, an understanding of the semantics differences behind these two applications enables me to demonstrate the flexibility of the proposed Bayesian rough set model. The main ideas of the two applications of Bayesian inference are illustrated by examples.

Suppose we have a data table from a hospital historical database. In this table, there are a set of patients and a set of attributes indicating patients' symptoms (e.g., cough) with regard to a certain disease (e.g., lung cancer).

The first application concerns the diagnosis. Given a patient with certain symptoms (i.e., the evidence), what are the chances that the patient has lung cancer (i.e., the hypothesis)?

In an ideal case, the information in the data table is complete, the probability of the patient has lung cancer can be estimated by the number of patients who have lung cancer and symptom cough divides the number of people who has symptom cough. In many cases, we may only have limited information on hand. Can we still predict the probability? Bayesian inference provides an answer to this question. Suppose that we have some prior knowledge about lung cancer (i.e., the probability of an arbitrary person having lung cancer). When the doctor sees a new patient, he/she receives evidence (i.e., cough) about lung cancer. The evidence is related to lung cancer by a conditional probability, called likelihood. Bayes' theorem, also called Bayes' law or Bayes' rule named after Thomas Bayes, offered a solution for this problem. In Bayes' theorem, the a posteriori probability can be calculated from the a priori probability and the likelihood function,

$$Pr(lung\ cancer \mid cough) \;=\; \frac{Pr(cough \mid lung\ cancer) \cdot Pr(lung\ cancer)}{Pr(cough)},$$

where $Pr(lung\ cancer)$ is a priori probability of an arbitrary people with lung cancer. $Pr(lung\ cancer \mid cough)$ is a posteriori probability that a patient with lung cancer after observing evidence cough, and $Pr(cough \mid lung\ cancer)$ is the likelihood of evidence cough related to lung cancer.

Once we obtain the a posteriori probability, how do we make decisions based on the value of the a posteriori probability? Rough set theory provides us a way for three-way decision making [121,122,142]. When applying Bayesian inference to rough sets, one may view the set C as a hypothesis that an object is in C and an equivalence class as evidence that an object is in the equivalence class. This immediately leads to the definition of three probabilistic regions defined in decision-theoretic rough set models (equation (2.6)),

$$\begin{aligned}
\mathrm{POS}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C|[x]) \geq \alpha\}, \\
\mathrm{BND}_{(\alpha,\beta)}(C) &= \{x \in U \mid \beta < Pr(C|[x]) < \alpha\}, \\
\mathrm{NEG}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C|[x]) \leq \beta\}.
\end{aligned}$$

They can be used to build a ternary classifier for three-way decisions. The doctor can make a decisions of treatment when the probability is greater than or equal to α, namely, $Pr(lung\ cancer \mid cough) \geq \alpha$, and of not treatment when $Pr(lung\ cancer \mid cough) \leq \beta$. In the case when the probability lies in between α and β, the doctor can perform a medical test to further examine the patient.

The second application concerns which symptoms or medical tests provide more information when diagnosing a disease. A doctor may decide to perform a particular test in order to revise a priori probability in the process of diagnosing and treating

lung cancer.

For example, if the probability of a patient has lung cancer given that he/she has cough increased from a priori probability (i.e., the probability without seeing any evidence), then cough is considered as supporting evidence for lung cancer. If there are many possible tests that may be used, how do we decide which one is more informative? Assume there are two tests that can be performed. Consider a Bayesian confirmation measure defined by [21, 24, 25, 98] $Pr(C \mid [x]) - Pr(C)$. The three probabilistic regions are defined in confirmation-theoretic rough set models as follows [35, 90, 91],

$$\begin{aligned}
\mathrm{POSC}_{(s,t)}(C) &= \{x \in U \mid Pr(C|[x]) - Pr(C) \geq s\}, \\
\mathrm{BNDC}_{(s,t)}(C) &= \{x \in U \mid t < Pr(C|[x]) - Pr(C) < s\}, \\
\mathrm{NEGC}_{(s,t)}(C) &= \{x \in U \mid Pr(C|[x]) - Pr(C) \leq t\},
\end{aligned}$$

where (s, t) is a pair of thresholds and $s > t$. The results of a Bayesian confirmation model also offer three-way decisions for evaluating symptoms or tests. More specifically, if $Pr(C|[x]) - Pr(C) \geq s$, the symptoms of x support C or these tests should be performed to confirm that x has lung cancer. If $Pr(C|[x]) - Pr(C) \leq t$, the symptoms of x are against C or these tests should be performed to rule out x has lung cancer. If $t < Pr(C|[x]) - Pr(C) < s$, the symptoms of x are neutral to C or these tests are not informative.

Alternatively, the value of a Bayesian confirmation measure can be used to help the doctor to decide which medical test should be performed. Based on the historical data, the results of *Test1* provide a better indication that a patient has lung cancer

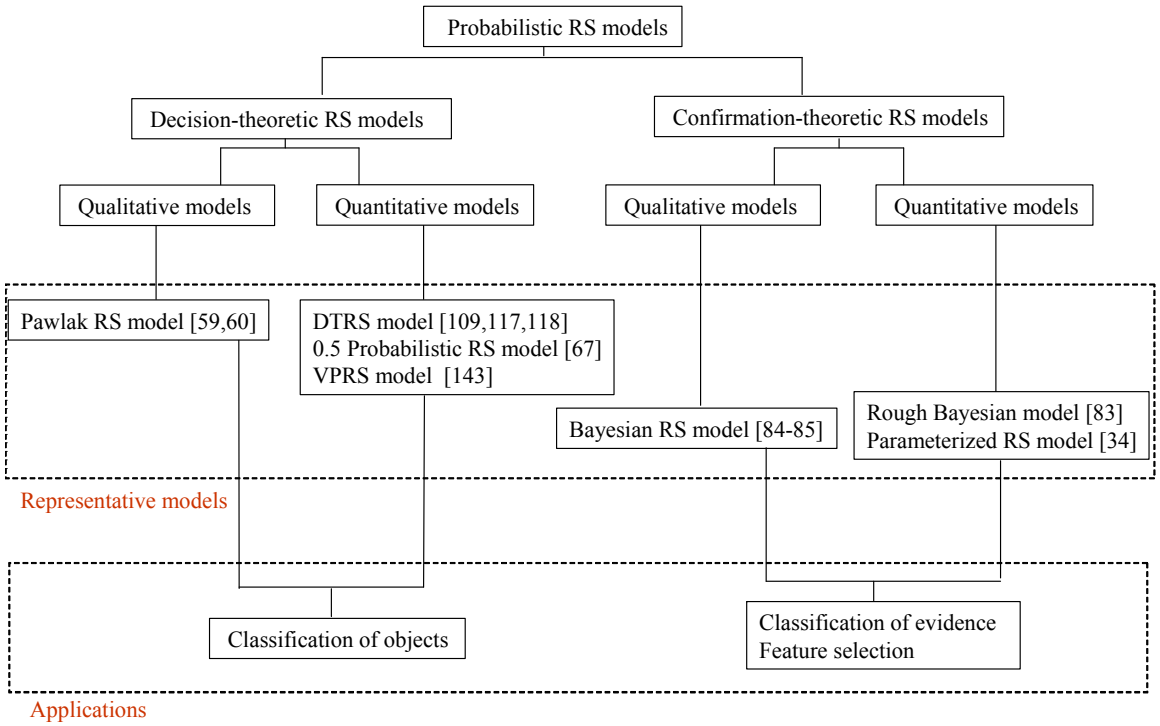

Figure 3.1: Categorization of probabilistic rough set models

than the results of *Test2* if

$$Pr(lung\ cancer \mid Test1) - Pr(lung\ cancer) > Pr(lung\ cancer \mid Test2) - Pr(lung\ cancer).$$

Then, the doctor might want to perform *Test1* instead of *Test2*.

By combining decision-theoretic rough set models and confirmation-theoretic rough set models, I propose a unified framework of Bayesian rough sets. The categorization of these two types of probabilistic rough set models is shown in Figure 3.1.

3.1.2 Cost-Sensitive Three-Way Approach

In many classification and decision making problems, we need to consider the costs caused by different classification errors.

Two examples given in Chapter 1 are both cost-sensitive decision-making problems. In the egg example, two types of wrong decisions come with different costs. If the sixth egg is good but you throw it away, one good egg will be wasted. If the sixth egg is bad but you break into bowl, five good eggs will be destroyed. Email spam filtering is also a cost-sensitive task, misclassifying a legitimate email into spam usually costs more than misclassifying a spam email into legitimate.

Now I want to incorporate costs associated with each decision into the decision making process. Bayesian decision theory provides a solution for combining utility theory and probability theory, in which the decision is made based on minimum risk decision rules [20].

Similar to the egg example, a 3×2 cost matrix can be drawn for the medical example. In this matrix, there are six types of costs, called loss functions in Bayesian decision theory. If the patient has cancer and the doctor makes the right decision to treat the patient, the cost is $\lambda(Treatment \mid cancer)$. If the patient does not have cancer and the doctor makes the right decision to not treat, the cost is $\lambda(\textit{Non-treatment} \mid \textit{non-cancer})$. Wrong decisions normally cost more. If the patient does not have cancer but the doctor decides to treat as cancer, the cost for false treatment is $\lambda(Treatment \mid \textit{non-cancer})$. The worst case is that the patient has cancer but the doctor decides to not treat, the cost for delayed treatment is $\lambda(\textit{Non-treatment} \mid cancer)$. There are also costs for making a deferment decision

(i.e., further testing). The cost of delaying treatment and performing diagnostic tests to patient with cancer is $\lambda(\textit{Further testing} \mid \textit{cancer})$, to patient without cancer is $\lambda(\textit{Further testing} \mid \textit{non-cancer})$. This is a cost-sensitive task because the six loss functions are not equal. For instance, the cost incurred by not treating a cancer patient is the highest since it could cause death or injury.

	cancer	non-cancer
Treatment	$\lambda(Treatment \mid cancer)$	$\lambda(Treatment \mid non\text{-}cancer)$
Further testing	$\lambda(Further\ testing \mid cancer)$	$\lambda(Further\ test \mid non\text{-}cancer)$
Non-treatment	$\lambda(Non-treatment \mid cancer)$	$\lambda(Non-treatment \mid non\text{-}cancer)$

Based on these six types of loss functions, we can calculate the expected cost for making each decision by combining loss functions with its corresponding probabilities as:

$$\begin{aligned}
R(Treatment \mid cough) &= \lambda(Treatment \mid cancer)Pr(cancer \mid cough) \\
&\quad + \lambda(Treatment \mid non\text{-}cancer)Pr(non\text{-}cancer \mid cough), \\
R(Further\ testing \mid cough) &= \lambda(Further\ testing \mid cancer)Pr(cancer \mid cough) \\
&\quad + \lambda(Further\ testing \mid non\text{-}cancer)Pr(non\text{-}cancer \mid cough), \\
R(Non\ treatment \mid cough) &= \lambda(Non-treatment \mid cancer)Pr(cancer \mid cough) \\
&\quad + \lambda(Non-treatment \mid non\text{-}cancer)Pr(non\text{-}cancer \mid cough).
\end{aligned}$$

According to the minimum risk decision rules [20], the doctor will choose the decision with the minimum expected cost.

Since the three-way decisions are defined by a pair of thresholds α and β, it is possible to obtain the pair of thresholds (α, β) by minimizing expected cost, as

demonstrated in decision-theoretic rough set model [116, 124, 125].

Based on the informal discussion so far, we propose a unified model of Bayesian rough sets for three-way decisions. The rest of the chapter is organized as follows. Section 3.2 explains the computation of thresholds based on decision-theoretic rough set models. Section 3.3 explains the foundation of sequential three-way decisions based on confirmation-theoretic rough set models. Section 3.4 presents how to estimate probabilities in both models based on Bayesian inference. Finally, Section 3.5 provides a framework of three-way decisions to construct and interpret rules from the three regions for building ternary classifiers.

3.2 Decision-Theoretic Rough Set Models

Decision-theoretic rough set (DTRS) models provide a systematic method to calculate a pair of thresholds based on the well established Bayesian decision theory, which establishes a basis for three-way data analysis. This section reviews and summarizes main results of DTRS based on discussions in [116, 124, 125].

3.2.1 Overview of Bayesian Decision Theory

Bayesian decision theory is a fundamental statistical approach that makes decisions under uncertainty based on probabilities and costs associated with decisions. Following Duda and Hart [20], the basic ideas of Bayesian decision theory are briefly summarized. Let $\Omega = \{w_1, \ldots, w_s\}$ be a finite set of s states and let $\mathcal{A} = \{a_1, \ldots, a_m\}$ be a finite set of m possible actions. The loss function $\lambda(a_i|w_j)$ for taking action a_i

when the state is w_j is given by a $m \times s$ matrix:

	w_1	w_2	$\cdots$	w_s
a_1	$\lambda(a_1\|w_1)$	$\lambda(a_1\|w_2)$	$\cdots$	$\lambda(a_1\|w_s)$
a_2	$\lambda(a_2\|w_1)$	$\lambda(a_2\|w_2)$	$\cdots$	$\lambda(a_2\|w_s)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_m	$\lambda(a_m\|w_1)$	$\lambda(a_m\|w_2)$	$\cdots$	$\lambda(a_m\|w_s)$

Let $Pr(w_j|\mathbf{x})$ be the conditional probability of an object x being in state w_j given that the object is described by $\mathbf{x}$. For an object with description $\mathbf{x}$, suppose action a_i is taken. Since $Pr(w_j|\mathbf{x})$ is the probability that the true state is w_j given $\mathbf{x}$, the expected loss associated with taking action a_i is given by:

$$R(a_i|\mathbf{x}) = \sum_{j=1}^{j=s} \lambda(a_i|w_j) Pr(w_j|\mathbf{x}). \tag{3.1}$$

The quantity $R(a_i|\mathbf{x})$ is also called the conditional risk.

Given a description $\mathbf{x}$, a decision rule is a function $\tau(\mathbf{x})$ that specifies which action to take. That is, for every $\mathbf{x}$, $\tau(\mathbf{x})$ takes one of the actions, $a_1, \ldots, a_m$. The overall risk $\mathbf{R}$ is the expected loss associated with a given decision rule. Since $R(\tau(\mathbf{x})|\mathbf{x})$ is the conditional risk associated with action $\tau(\mathbf{x})$, the overall risk is defined by:

$$\mathbf{R} = \sum_{\mathbf{x}} R(\tau(\mathbf{x})|\mathbf{x}) Pr(\mathbf{x}), \tag{3.2}$$

where the summation is over the set of all possible descriptions of objects. If $\tau(\mathbf{x})$ is chosen so that $R(\tau(\mathbf{x})|\mathbf{x})$ is as small as possible for every $\mathbf{x}$, the overall risk $\mathbf{R}$ is minimized. Thus, the optimal Bayesian decision procedure can be formally stated as

follows. For every $\mathbf{x}$, compute the conditional risk $R(a_i|\mathbf{x})$ for $i = 1, \ldots, m$ defined by equation (3.1) and select the action for which the conditional risk is minimum. If more than one action minimizes $R(a_i|\mathbf{x})$, a tie-breaking criterion is used.

3.2.2 Decision-Theoretic Rough Sets

A decision-theoretic rough set model is constructed by a straightforward application of the Bayesian decision theory [116, 124, 125]. With respect to a subset $C \subseteq U$, we can form a set of two states $\Omega = \{C, C^c\}$. To derive the three regions in rough set theory, the set of actions is given by $\mathcal{A} = \{a_P, a_B, a_N\}$, where a_P, a_B, and a_N represent the three actions in classifying an object x, namely, deciding $x \in \text{POS}(C)$, deciding $x \in \text{BND}(C)$, and deciding $x \in \text{NEG}(C)$, respectively. The loss function is given by a 3×2 matrix:

	C (P)	C^c (N)
a_P	$\lambda_{PP} = \lambda(a_P\|C)$	$\lambda_{PN} = \lambda(a_P\|C^c)$
a_B	$\lambda_{BP} = \lambda(a_B\|C)$	$\lambda_{BN} = \lambda(a_B\|C^c)$
a_N	$\lambda_{NP} = \lambda(a_N\|C)$	$\lambda_{NN} = \lambda(a_N\|C^c)$

In the matrix, λ_{PP}, λ_{BP} and λ_{NP} denote the losses incurred for taking actions a_P, a_B and a_N, respectively, when an object belongs to C, and λ_{PN}, λ_{BN} and λ_{NN} denote the losses incurred for taking these actions when the object does not belong to C.

The expected losses associated with taking different actions for objects in $[x]$ can be expressed as:

$$\begin{aligned} R(a_P|[x]) &= \lambda_{PP}Pr(C|[x]) + \lambda_{PN}Pr(C^c|[x]), \\ R(a_B|[x]) &= \lambda_{BP}Pr(C|[x]) + \lambda_{BN}Pr(C^c|[x]), \\ R(a_N|[x]) &= \lambda_{NP}Pr(C|[x]) + \lambda_{NN}Pr(C^c|[x]). \end{aligned} \tag{3.3}$$

The Bayesian decision procedure suggests the following minimum-risk decision rules:

(P) If $R(a_P|[x]) \leq R(a_B|[x])$ and $R(a_P|[x]) \leq R(a_N|[x])$, decide $x \in \text{POS}(C)$;

(B) If $R(a_B|[x]) \leq R(a_P|[x])$ and $R(a_B|[x]) \leq R(a_N|[x])$, decide $x \in \text{BND}(C)$;

(N) If $R(a_N|[x]) \leq R(a_P|[x])$ and $R(a_N|[x]) \leq R(a_B|[x])$, decide $x \in \text{NEG}(C)$.

When equalities hold, tie-breaking criteria are used so that each object is put into only one region.

By $Pr(C|[x]) + Pr(C^c|[x]) = 1$, we can simplify the rules by using only the probabilities $Pr(C|[x])$ and the loss function λ. Consider a special kind of loss functions with [116, 124, 125]:

$$\text{(c0).} \qquad \lambda_{PP} \leq \lambda_{BP} < \lambda_{NP}, \quad \lambda_{NN} \leq \lambda_{BN} < \lambda_{PN}. \tag{3.4}$$

That is, the loss of classifying an object x belonging to C into the positive region $\text{POS}(C)$ is less than or equal to the loss of classifying x into the boundary region $\text{BND}(C)$, and both of these losses are strictly less than the loss of classifying x into the negative region $\text{NEG}(C)$. The reverse order of losses is used for classifying an object not in C. Under condition (c0), we can simplify decision rules (P)-(N) as

follows. For the rule (P), the condition $R(a_P|[x]) \leq R(a_B|[x])$ can be expressed as:

$$\begin{aligned}
& R(a_P|[x]) \leq R(a_B|[x]) \\
\iff\ & \lambda_{PP}Pr(C|[x]) + \lambda_{PN}Pr(C^c|[x]) \leq \lambda_{BP}Pr(C|[x]) + \lambda_{BN}Pr(C^c|[x]) \\
\iff\ & \lambda_{PP}Pr(C|[x]) + \lambda_{PN}(1 - Pr(C|[x])) \leq \lambda_{BP}Pr(C|[x]) + \lambda_{BN}(1 - Pr(C|[x])) \\
\iff\ & Pr(C|[x]) \geq \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}. \qquad (3.5)
\end{aligned}$$

Similarly, other conditions of the three rules can be expressed as:

$$\begin{aligned}
R(a_P|[x]) \leq R(a_N|[x]) &\iff Pr(C|[x]) \geq \frac{(\lambda_{PN} - \lambda_{NN})}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}, \\
R(a_B|[x]) \leq R(a_P|[x]) &\iff Pr(C|[x]) \leq \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}, \\
R(a_B|[x]) \leq R(a_N|[x]) &\iff Pr(C|[x]) \geq \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}, \\
R(a_N|[x]) \leq R(a_P|[x]) &\iff Pr(C|[x]) \leq \frac{(\lambda_{PN} - \lambda_{NN})}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}, \\
R(a_N|[x]) \leq R(a_B|[x]) &\iff Pr(C|[x]) \leq \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}.
\end{aligned}$$

By introducing three parameters:

$$\begin{aligned}
\alpha &= \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})}, \\
\beta &= \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})}, \\
\gamma &= \frac{(\lambda_{PN} - \lambda_{NN})}{(\lambda_{PN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{PP})}. \qquad (3.6)
\end{aligned}$$

We can express concisely the decision rules (P)-(N) as:

(P) If $Pr(C|[x]) \geq \alpha$ and $Pr(C|[x]) \geq \gamma$, decide $x \in \mathrm{POS}(C)$;

(B) If $Pr(C|[x]) \leq \alpha$ and $Pr(C|[x]) \geq \beta$, decide $x \in \mathrm{BND}(C)$;

(N) If $Pr(C|[x]) \leq \beta$ and $Pr(C|[x]) \leq \gamma$, decide $x \in \mathrm{NEG}(C)$.

Each rule is defined by two out of the three parameters.

The conditions of rule (B) suggest that it maybe reasonable to impose the constraint $\alpha > \beta$ so that the boundary region may be non-empty. By setting $\alpha > \beta$, we obtain the following condition on the loss function [116]:

$$\text{(c1).} \qquad \frac{\lambda_{NP} - \lambda_{BP}}{\lambda_{BN} - \lambda_{NN}} > \frac{\lambda_{BP} - \lambda_{PP}}{\lambda_{PN} - \lambda_{BN}}. \tag{3.7}$$

The conditions (c0) and (c1) imply that $1 \geq \alpha > \gamma > \beta \geq 0$. In this case, after tie-breaking, the following simplified rules are obtained [116]:

$$\begin{aligned}
&\text{(P)} && \text{If } Pr(C|[x]) \geq \alpha, \text{ decide } x \in \text{POS}(C);\\
&\text{(B)} && \text{If } \beta < Pr(C|[x]) < \alpha, \text{ decide } x \in \text{BND}(C);\\
&\text{(N)} && \text{If } Pr(C|[x]) \leq \beta, \text{ decide } x \in \text{NEG}(C).
\end{aligned}$$

The parameter γ is no longer needed.

From the rules (P), (B), and (N), the (α, β)-probabilistic positive, negative and boundary regions are given, respectively, by:

$$\begin{aligned}
\text{POS}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C|[x]) \geq \alpha\},\\
\text{BND}_{(\alpha,\beta)}(C) &= \{x \in U \mid \beta < Pr(C|[x]) < \alpha\},\\
\text{NEG}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C|[x]) \leq \beta\}.
\end{aligned} \tag{3.8}$$

Thus, the formulation of a decision-theoretic rough set model not only produces three probabilistic regions, but also provides a theoretical basis for and a practical interpretation of the probabilistic rough sets. The thresholds are systematically calculated from a loss function that can be easily interpreted in more operable terms, including profits, risk, cost, etc.

Table 3.1: Loss function of a medical example

	C (P)(cancer)	C^c (N)(non-cancer)
a_P(Treatment)	$\lambda_{PP} = \$200$	$\lambda_{PN} = \$600$
a_B(Further testing)	$\lambda_{BP} = \$300$	$\lambda_{BN} = \$300$
a_N(Non-treatment)	$\lambda_{NP} = \$1000$	$\lambda_{NN} = \$0$

3.2.3 An Example

The process of deriving the required thresholds from loss functions can be demonstrated by using the previous discussed medical example. In this case, there are two states regarding a disease: C (P) denoting cancer and C^c (N) denoting non-cancer. There are three actions: a_P for treating the patient, a_B for making a deferred decision (i.e., perform a diagnostic test such as X-ray), and a_N for not treating the patient.

Table 3.1 gives the loss functions of the medical diagnostic cost, the pair of thresholds α and β is calculated according to equation (3.6) as:

$$\begin{aligned} \alpha &= \frac{(\lambda_{PN} - \lambda_{BN})}{(\lambda_{PN} - \lambda_{BN}) + (\lambda_{BP} - \lambda_{PP})} = \frac{600 - 300}{(600 - 300) + (300 - 200)} = 0.75, \\ \beta &= \frac{(\lambda_{BN} - \lambda_{NN})}{(\lambda_{BN} - \lambda_{NN}) + (\lambda_{NP} - \lambda_{BP})} = \frac{300 - 0}{(300 - 0) + (1000 - 300)} = 0.30. \end{aligned}$$

Suppose the probability of a patient having cancer given his/her symptom (e.g., cough) is $Pr(cancer \mid cough) = 0.10$, the doctor may decide not to treat the patient because $Pr(cancer \mid cough) \leq \beta$.

3.3 Confirmation-Theoretic Rough Set Models

Based on Bayes' theorem and Bayesian confirmation theory [21,24,25,98], alternative models of probabilistic rough sets have been proposed and studied [32, 35, 89–91]. In this section, I present a confirmation-theoretic framework to summarize the main results from these studies and show their differences from the (α, β)-probabilistic approximations. To differentiate these models from decision-theoretic models, I refer to them as (Bayesian) confirmation-theoretic rough set models. To differentiate the derived three regions from probabilistic approximation regions, I refer to them as (Bayesian) confirmation regions. For simplicity, I assume that, whenever a ratio is used, a probability in the denominator is not zero.

3.3.1 Bayesian Inference

Bayes' theorem [6] shows the relation between two conditional probabilities that are the reverse of each other. It expresses the conditional probability (or a posteriori probability) of an event H after E is observed in terms of a priori probability of H, probability of E, and the conditional probability of E given H. The Bayes' theorem is expressed as follows:

$$Pr(H|E) = \frac{Pr(E|H)}{Pr(E)} Pr(H), \tag{3.9}$$

where $Pr(H)$ is a priori probability of H that H happens or is true, $Pr(H|E)$ is the a posteriori probability that H happens after observing E, and $Pr(E|H)$ is the likelihood of H given E. Through Bayes' theorem, a difficulty to estimate probability $Pr(H|E)$ is expressed in terms of an easy to estimate likelihood $Pr(E|H)$. This makes

Bayes' theorem particularly useful in data analysis and pattern classification.

Bayesian inference derived from Bayes' theorem can be expressed in a more general form. Replacing H with a continuous parameter set Θ, E with observations y, and Pr with a probability-density function p, results in the following,

$$p(\Theta|y) = \frac{p(y|\Theta)}{p(y)}p(\Theta), \tag{3.10}$$

where $p(\Theta)$ is the set of a priori distributions of parameter set Θ before y is observed, $p(y \mid \Theta)$ is the likelihood of y under a model, and $p(\Theta \mid y)$ is the joint a posteriori distribution, sometimes called the full a posteriori distribution, of parameter set Θ that expresses uncertainty about parameter set Θ after taking both the a priori and data into account [6, 8, 29]. Since there are usually multiple parameters, Θ represents a set of j parameters ($\Theta = \theta_1, ..., \theta_j$). The denominator

$$p(y) = \int p(y \mid \Theta)p(\Theta)d\Theta, \tag{3.11}$$

defines a priori predictive distribution of y, and can be set to an unknown constant $\boldsymbol{c}$. $p(y)$ indicates what y should look like before y has been observed. The presence of $p(y)$ normalizes the joint a posteriori distribution, $p(\Theta \mid y)$, ensuring it is a proper distribution and integrates to one.

By removing $p(y)$ from equation (3.10), the relationship changes from '*equals*' (=) to '*proportional to*' ($\propto$), the formulation of Bayes' theorem becomes,

$$p(\Theta|y) \propto p(y|\Theta)p(\Theta). \tag{3.12}$$

This form can be stated as the unnormalized joint a posteriori being proportional to the likelihood times a priori [6, 8, 29].

There are a few advantages of Bayesian inference. First, Bayesian inference allows informative priors so that prior knowledge or results of a previous model can be used to inform the current model. Second, Bayesian inference estimates $p(hypothesis \mid data)$, instead of $p(data \mid hypothesis)$. The term 'hypothesis testing' suggests it should be the hypothesis that is tested, given the data, not the other way around. Third, Bayesian inference includes uncertainty in the probability model, yielding more realistic predictions.

3.3.2 Qualitative Bayesian Confirmation

In a qualitative interpretation of Bayesian confirmation theory [21,24,25,98], evidence E confirms hypothesis H, disconfirms H, or is neutral with respect to H whenever the a posteriori probability $Pr(H|E)$ increases from the a priori probability $Pr(H)$, decreases from $Pr(H)$ or is unchanged from $Pr(H)$. Festa [24] referred to this interpretations of confirmation as P-incremental confirmation, to reflect the fact that the hypothesis H is confirmed by the evidence E when the initial probability of H increases as an effect of E. This qualitative notion of P-incremental confirmation can be precisely defined as follows:

$$\begin{cases} E \text{ confirms } H & \text{iff } Pr(H|E) > Pr(H) \\ & \text{iff } \frac{Pr(H|E)}{Pr(H)} > 1, \\ E \text{ is neutral w.r.t } H & \text{iff } Pr(H|E) = Pr(H) \\ & \text{iff } \frac{Pr(H|E)}{Pr(H)} = 1, \\ E \text{ disconfirms } H & \text{iff } Pr(H|E) < Pr(H) \\ & \text{iff } \frac{Pr(H|E)}{Pr(H)} < 1. \end{cases}$$

In this definition, the condition $Pr(H|E) > Pr(H)$ is equivalently expressed as $Pr(H|E)/Pr(H) > 1$, assuming that $Pr(H) \neq 0$.

Bayesian confirmation theory also relies on the Bayes' theorem for the computation of $Pr(H|E)$. More specifically, two additional representations can be further derived. According to Bayes' theorem, we have

$$\frac{Pr(H|E)}{Pr(H)} = \frac{Pr(E|H)}{Pr(E)}. \tag{3.13}$$

Thus, one can re-express conditions for qualitative confirmation by using likelihood and the probability of evidence [21, 24, 25, 98]:

$$\begin{cases} E \text{ confirms } H & \text{iff } \frac{Pr(E|H)}{Pr(E)} > 1, \\ E \text{ is neutral w.r.t } H & \text{iff } \frac{Pr(E|H)}{Pr(E)} = 1, \\ E \text{ disconfirms } H & \text{iff } \frac{Pr(E|H)}{Pr(E)} < 1. \end{cases}$$

Let H^c denote the complement of hypothesis H. It follows that $Pr(H) + Pr(H^c) = 1$. By the computation of the probability of evidence,

$$Pr(E) = Pr(E|H)Pr(H) + Pr(E|H^c)Pr(H^c), \tag{3.14}$$

we have,

$$\begin{aligned} \frac{Pr(E|H)}{Pr(E)} > 1 \iff & \frac{Pr(E|H)}{Pr(E|H)Pr(H) + Pr(E|H^c)Pr(H^c)} > 1 \\ \iff & Pr(E|H) > Pr(E|H)Pr(H) + Pr(E|H^c)Pr(H^c) \\ \iff & (1 - Pr(H))Pr(E|H) > Pr(E|H^c)Pr(H^c) \\ \iff & \frac{Pr(E|H)}{Pr(E|H^c)} > 1. \end{aligned} \tag{3.15}$$

It follows that the conditions for confirmation can also be equivalently expressed

through a likelihood ratio:

$$\begin{cases} E \text{ confirms } H & \text{iff } \frac{Pr(E|H)}{Pr(E|H^c)} > 1, \\ E \text{ is neutral w.r.t } H & \text{iff } \frac{Pr(E|H)}{Pr(E|H^c)} = 1, \\ E \text{ disconfirms } H & \text{iff } \frac{Pr(E|H)}{Pr(E|H^c)} < 1. \end{cases}$$

The three distinct but equivalent expressions of the conditions of confirmation provide an understanding of Bayesian confirmation from different perspectives. The first focuses on the comparison of a priori and a posteriori probability of H; the second focuses on whether E is more probable conditional on H than it is unconditionally; and the third focuses on the likelihood ratio of H given E.

When applying qualitative Bayesian confirmation theory to rough sets, one may view the set C as a hypothesis that an object is in C and an equivalence class as evidence that an object is in the equivalence class. This immediately leads to a definition of three qualitative Bayesian confirmation regions, namely, the positive confirmation region, the non-confirmation (or neutral) region, and the negative confirmation (or disconfirmation) region:

$$\begin{aligned} \text{POSC}(C) &= \{x \in U \mid Pr(C|[x]) > Pr(C)\} \\ &= \{x \in U \mid \frac{Pr(C|[x])}{Pr(C)} > 1\} \\ &= \{x \in U \mid \frac{Pr([x]|C)}{Pr([x])} > 1\} \\ &= \{x \in U \mid \frac{Pr([x]|C)}{Pr([x]|C^c)} > 1\}, \\ \text{BNDC}(C) &= \{x \in U \mid Pr(C|[x]) = Pr(C)\} \\ &= \{x \in U \mid \frac{Pr(C|[x])}{Pr(C)} = 1\} \end{aligned}$$

$$
\begin{aligned}
&= \{x \in U \mid \frac{Pr([x]|C)}{Pr([x])} = 1\} \\
&= \{x \in U \mid \frac{Pr([x]|C)}{Pr([x]|C^c)} = 1\}, \\
\text{NEGC}(C) &= \{x \in U \mid Pr(C|[x]) < Pr(C)\} \\
&= \{x \in U \mid \frac{Pr(C|[x])}{Pr(C)} < 1\} \\
&= \{x \in U \mid \frac{Pr([x]|C)}{Pr([x])} < 1\} \\
&= \{x \in U \mid \frac{Pr([x]|C)}{Pr([x]|C^c)} < 1\}. \qquad (3.16)
\end{aligned}
$$

These three regions were introduced and studied by Ślęzak and Ziarko [90, 91] in a Bayesian rough set model. They interpreted them as probabilistic rough set approximations of C. I argue that they have a very different semantic interpretations and it may not be appropriate to interpret them straightforwardly as probabilistic approximations of C.

The three confirmation regions and the three (α, β)-probabilistic approximation regions are similar in form and both are defined in probabilistic terms. It is very tempting to treat them as the same by viewing $Pr(H)$ as a threshold on the a posteriori probability [90, 91]. However, it must be realized that there are significant semantics differences between them. Similar to Pawlak approximation regions, the three confirmation regions defined by equation (3.16) are of a qualitative nature. They are defined based on whether an equivalence class confirms, is neutral with respect to, or disconfirms C. There is no consideration of the degree of confirmation. In contrast, the (α, β)-probabilistic approximation regions are of a quantitative nature. According to Bayesian confirmation theory, one may interpret the three confirmation regions as follows. For an equivalence class inside the positive confirmation region POSC(C), the

evidence that an object is in the equivalence class positively confirms the hypothesis C; evidence from the negative confirmation region negatively confirms, or disconfirms, the hypothesis; evidence from the boundary region neither confirms nor disconfirms, is neutral with respect to, the hypothesis. In some sense, the three confirmation regions are a classification of various pieces of evidence according to their confirmation of the set C. They may not be viewed directly as approximations of C. In contract, (α, β)-probabilistic regions are a classification of objects as approximation of C. For these reasons, I refer to one as confirmation regions and the other as probabilistic approximation regions.

3.3.3 Quantitative Bayesian Confirmation

A quantitative conception of Bayesian confirmation theory employs a Bayesian confirmation measure that quantifies the degree to which evidence confirms hypothesis. There is no general agreement on a quantitative measure of confirmation [99]. Many Bayesian confirmation measures have been proposed and studied [24, 25]. A number of properties to be satisfied by such measures have been introduced and examined. Festa [24] suggested that a Bayesian confirmation measure $c(E, H)$ must satisfy the condition of probability increment, that is, $c(E, H) = f(Pr(H|E), Pr(E))$, must be an increasing function of $Pr(H|E)$ and a decreasing function of $Pr(E)$. Corresponding to the previously discussed three forms of qualitative confirmation, we have the

following P-incremental confirmation measures:

$$
\begin{aligned}
c_d(E,H) &= Pr(H|E) - Pr(H), \\
c_r(E,H) &= \frac{Pr(H|E)}{Pr(H)} = \frac{Pr(E|H)}{Pr(E)}, \\
c_r^+(E,H) &= \frac{Pr(E|H)}{Pr(E|H^c)},
\end{aligned} \tag{3.17}
$$

where a similar notational system of Festa is used.

Fitelson [25] argued that a confirmation measure must be consistent with qualitative interpretation of confirmation in the sense that

$$
c(E,H) \begin{cases} > 0, & Pr(H|E) > Pr(H) \\ = 0, & Pr(H|E) = Pr(H) \\ < 0, & Pr(H|E) < Pr(H) \end{cases} \tag{3.18}
$$

A measure satisfying this property is called a relevance measure [25]. Corresponding to the three representations of qualitative Bayesian confirmation, we can have the following Bayesian confirmation measures that are relevance measures:

$$
\begin{aligned}
c_d(E,H) &= Pr(H|E) - Pr(H), \\
c_{nr}(E,H) &= \frac{Pr(H|E)}{Pr(H)} - 1 = \frac{Pr(E|H)}{Pr(E)} - 1, \\
c_{nr}^+(E,H) &= \frac{Pr(E|H)}{Pr(E|H^c)} - 1, \\
c_{lr}(E,H) &= \log \frac{Pr(H|E)}{Pr(H)} = \log \frac{Pr(E|H)}{Pr(E)}, \\
c_{lr}^+(E,H) &= \log \frac{Pr(E|H)}{Pr(E|H^c)}.
\end{aligned} \tag{3.19}
$$

Confirmation measures c_{nr} and c_{nr}^+ may be viewed as normalized version of c_r and c_r^+ that satisfy the constraint given by equation (3.18). Additional Bayesian confirmation measures and their interpretations can be found in [24, 25, 35].

When applying a Bayesian confirmation measure to define parameterized confirmation regions, one may expect that evidence in the positive region must confirm the hypothesis beyond a certain level, evidence in the negative region must disconfirm the hypothesis beyond another level, and evidence in the boundary region fails to confirm or disconfirm the hypothesis beyond the required levels. In the context of rough set theory, the value $c([x], C)$ is the degree to which the evidence $y \in [x]$ confirms the hypothesis $y \in C$, where $y \in U$. Given any Bayesian confirmation measure $c([x], C)$ and a pair of thresholds (s, t), with $t < s$, on the confirmation measure, we define three (s, t)-confirmation regions as:

$$\begin{aligned} \mathrm{POSC}_{(s,t)}(C) &= \{x \in U \mid c([x], C) \geq s\}, \\ \mathrm{BNDC}_{(s,t)}(C) &= \{x \in U \mid t < c([x], C) < s\}, \\ \mathrm{NEGC}_{(s,t)}(C) &= \{x \in U \mid c([x], C) \leq t\}. \end{aligned} \tag{3.20}$$

They are pair-wise disjoint.

There are several difficulties for the application of quantitative Bayesian confirmation regions. Recall that there is no general agreement on a Bayesian confirmation measure. Choosing an appropriate confirmation measure for a particular application may not be an easy task. The ranges of the values of different confirmation measures are different. This makes it an even more difficult task to interpret and set the thresholds for the desired levels of confirmation or disconfirmation. For different confirmation measure, we may need to lay out different guidelines for setting thresholds.

Consider now a minimal requirement of (s, t)-confirmation regions. According to intended interpretation of Bayesian confirmation, it is reasonable to require that

(s,t)-confirmation is consistent with qualitative confirmation. That is, the pair of thresholds must be chosen so that it satisfies the following conditions: if evidence quantitatively confirms a hypothesis beyond a level s, the evidence must qualitative confirms the hypothesis; if evidence quantitatively disconfirms a hypothesis below a level t, the evidence must qualitatively disconfirms the hypothesis. If $c([x], C)$ is a relevance measure, the condition $t < 0 < s$ will satisfy this requirement. In general, the reverse implications are not true. This immediately leads to the following linkage between qualitative and quantitative confirmation regions:

$$\begin{aligned} &\mathrm{POSC}_{(s,t)}(C) \subseteq \mathrm{POSC}(C), \\ &\mathrm{BNDC}(C) \subseteq \mathrm{BNDC}_{(s,t)}(C), \\ &\mathrm{NEGC}_{(s,t)}(C) \subseteq \mathrm{POSC}(C). \end{aligned} \tag{3.21}$$

The linkage of qualitative and quantitative confirmation regions is a kind of reverse relation of the linkage of Pawlak and (α, β)-probabilistic approximations. One may conclude that Bayesian confirmation theory provides a new class of probabilistic rough set models. It is important to distinguish the class of decision-theoretic rough set models and the class of confirmation-theoretic rough sets models. That is, we can use decision-theoretic rough set models for the classification of objects based on their satisfiability of the hypothesis; we use confirmation-theoretic rough sets models to evaluate the quality of different pieces of evidences.

For the three representations of qualitative Bayesian confirmation, they produce the same confirmation regions. This is no longer true for their corresponding quantitative confirmation measures. For confirmation measure $c_d([x], C) = Pr(C|[x]) - Pr(C)$,

we have the following three regions:

$$\begin{aligned}
\mathrm{POSC}_{(s,t)}(C) &= \{x \in U \mid Pr(C|[x]) - Pr(C) \geq s\}, \\
\mathrm{BNDC}_{(s,t)}(C) &= \{x \in U \mid t < Pr(C|[x]) - Pr(C) < s\}, \\
\mathrm{NEGC}_{(s,t)}(C) &= \{x \in U \mid Pr(C|[x]) - Pr(C) \leq t\}.
\end{aligned} \tag{3.22}$$

Based on the range of the values of c_d, it is reasonable to require that $t < 0 < s$, with $[s, 1]$ for positive confirmation, (t, s) for neutral confirmation, and $[-1, t]$ for negative confirmation. For confirmation measure, $c_r^+([x], C) = \frac{Pr([x]|C)}{Pr([x]|C^c)}$, one can immediately obtain the three regions introduced by Ślęzak [89] in a rough Bayesian model:

$$\begin{aligned}
\mathrm{POSC}_{(s',t')}(C) &= \{x \in U \mid \frac{Pr([x]|C)}{Pr([x]|C^c)} \geq s'\}, \\
\mathrm{BNDC}_{(s',t')}(C) &= \{x \in U \mid t' < \frac{Pr([x]|C)}{Pr([x]|C^c)} < s'\}, \\
\mathrm{NEGC}_{(s',t')}(C) &= \{x \in U \mid \frac{Pr([x]|C)}{Pr([x]|C^c)} \leq t'\}.
\end{aligned} \tag{3.23}$$

The pair of thresholds (s', t') on c_r^+ is different from the pair (s, t) on c_d. It is reasonable to require that $t' < 1 < s'$, with $[s', +\infty)$ for positive confirmation, (t', s') for neutral confirmation, and $[0, t']$ for negative confirmation. To some degree, it might be easier to interpret thresholds on confirmation measure c_d, as it is the difference between a posteriori and a priori probability. Interpretations of thresholds on confirmation measures such as c_r^+, c_{lr} or c_{lr}^+ are not as intuitive.

3.3.4 Combination of Probabilistic Approximation and Bayesian Confirmation

Greco et al. [35] proposed a parameterized rough set model by combining (α, β)-probabilistic approximation regions and (s, t)-Bayesian confirmation regions. In their paper, they defined a pair of lower and upper approximations. I re-express their results in terms of three regions as follows. Suppose $0 \leq \beta < \alpha \leq 1$, $c([x], C)$ is a Bayesian confirmation measure, and $t < s$. Three parameterized regions are defined as:

$$\begin{aligned}
\mathrm{PPOS}_{(\alpha,\beta,s,t)}(C) &= \{x \in U \mid Pr(C|[x]) \geq \alpha \wedge c([x], C) \geq s\} \\
&= \mathrm{POS}_{(\alpha,\beta)}(C) \cap \mathrm{POSC}_{(s,t)}(C), \\
\mathrm{PBND}_{(\alpha,\beta,s,t)}(C) &= \{x \in U \mid (Pr(C|[x]) > \beta \vee c([x], C) > t) \wedge \\
&\quad (Pr(C|[x]) < \alpha \vee c([x], C) < s)\} \\
&= (\mathrm{POS}_{(\alpha,\beta)}(C) \cap \mathrm{NEGC}_{(s,t)}(C)) \cup \\
&\quad (\mathrm{BND}_{(\alpha,\beta)}(C) \cup \mathrm{BNDC}_{(s,t)}(C)) \cup \\
&\quad (\mathrm{NEG}_{(\alpha,\beta)}(C) \cap \mathrm{POSC}_{(s,t)}(C)), \\
\\
\mathrm{PNEG}_{(\alpha,\beta,s,t)}(C) &= \{x \in U \mid Pr(C|[x]) \leq \beta \wedge c([x], C) \leq t\} \\
&= \mathrm{NEG}_{(\alpha,\beta)}(C) \cap \mathrm{NEGC}_{(s,t)}(C). \qquad (3.24)
\end{aligned}$$

That is, both positive and negative regions are the common parts of the corresponding (α, β)-probabilistic approximations and (s, t)-Bayesian confirmation, and the boundary regions is the union of the two boundary regions and the disagreement parts of (α, β)-probabilistic approximations and (s, t)-Bayesian confirmation.

In addition to the difficulty in choosing an appropriate Bayesian confirmation measure and a pair of thresholds on the measure, the parameterized model faces a new challenge of interpreting the combined results. Since probabilistic approximation regions and Bayesian confirmation regions represent semantically different concepts, it is necessary to move beyond a simple method of combination. One must address the issue regarding the rationale and interpretation of such a combination. However, this should not be considered as a criticism of a combined model, but instead serving as a motivation for further studies.

3.3.5 Discussions

Applications of Bayesian methods to rough set theory result in two classes of probabilistic rough set models. The two classes share many similarities. Both of them make use of Bayes' theorem and both employ the same technique of using a pair of thresholds to produce three regions. However, their semantic interpretations and hence intended applications are very different. Decision-theoretic rough set models, developed based on Bayesian decision theory and Bayesian classification, concern the approximation of a set of objects by three probabilistic regions. In contrast, confirmation-theoretic rough set models, developed based on Bayesian confirmation theory, classify evidences related to equivalence classes into three confirmation regions.

In existing studies on confirmation-theoretic rough set models, attention has been paid mainly to mathematical constructions and formal properties of various notions. The semantics of these models, in the context of rough set theory, have not been explicitly studied and made clear. There exists confusion about different classes of

models. For example, the connections and differences between Bayesian rough sets and rough Bayesian sets are not clearly stated. In this thesis, I use the term "Bayesian rough sets" to indicate a combination of Bayesian methods and rough set theory, or applications of Bayesian methods to rough set theory, in a similar way that the term of "probabilistic rough sets" is used to denote a combination of probabilistic methods and rough set theory or applications of probabilistic methods to rough set theory. In other words, Bayesian rough sets are interpreted as Bayesian approaches to rough sets, involving Bayesian statistics, Bayesian inference, Bayesian decision theory, and Bayesian confirmation theory.

The plurality of Bayesian confirmation measures imposes another practical difficulty in applying confirmation-theoretic rough set models. Unlike decision-theoretic rough set models, there is, in general, a lack of guidelines and systematic procedures for interpreting and computing the required thresholds. For this reason, I choose probabilistic approximation regions of decision-theoretic rough set models for the development of a Bayesian rough set model. As future research, by using the same methodology one might build an alternative Bayesian rough set model based on confirmation regions of confirmation-theoretic models.

3.4 Estimating Probabilities Based on Bayesian Inference

Bayesian inference, based on Bayes' theorem, is used to design procedures for computing a posteriori probability required by Bayesian rough sets.

3.4.1 Inferring a Posteriori Probability by Bayes' Theorem

In the (α, β)-probabilistic approximations, the a posteriori probabilities are not always directly derivable from data. In such cases, we need to consider alternative ways to calculate their values. A commonly used method is to apply the Bayes' theorem:

$$Pr(C|[x]) = \frac{Pr(C)Pr([x]|C)}{Pr([x])}, \tag{3.25}$$

where

$$Pr([x]) = Pr([x]|C)Pr(C) + Pr([x]|C^c)Pr(C^c).$$

It reduces the problem of estimating the a posteriori probability $Pr(C|[x])$ of class C given $[x]$ into estimating the a priori probability $Pr(C)$ of class C, and the likelihood $Pr([x]|C)$ of C given $[x]$. There are many methods to estimate likelihood from data such as naive Bayes and belief networks, which makes probabilistic rough set models practically useful.

In Bayesian classification, one often uses a monotonically increasing function of the conditional probability to construct an equivalent classifier. One of such functions is the odds defined by $O(\cdot) = \frac{Pr(\cdot)}{1-Pr(\cdot)}$. The odds version of Bayes' theorem is given by:

$$\begin{aligned} O(C|[x]) &= \frac{Pr(C|[x])}{Pr(C^c|[x])} = \frac{Pr([x]|C)}{Pr([x]|C^c)} \cdot \frac{Pr(C)}{Pr(C^c)} \\ &= \frac{Pr([x]|C)}{Pr([x]|C^c)} \cdot O(C). \end{aligned} \tag{3.26}$$

It shows how to update a priori odds $O(C)$ into a posteriori odds $O(C|[x])$ through the likelihood ratio $\frac{Pr([x]|C)}{Pr([x]|C^c)}$. Posterior odds is useful in many applications. For example, one may not be interested in the actual value of the posterior probability, but the probability of C given $[x]$ is how many times than the probability of C^c given $[x]$. For

the positive region, we have:

$$\begin{aligned} Pr(C|[x]) \geq \alpha &\iff O(C|[x]) \geq \frac{\alpha}{1-\alpha} \\ &\iff \frac{Pr(C|[x])}{Pr(C^c|[x])} \geq \frac{\alpha}{1-\alpha} \\ &\iff \frac{Pr([x]|C)}{Pr([x]|C^c)} \cdot \frac{Pr(C)}{Pr(C^c)} \geq \frac{\alpha}{1-\alpha} \\ &\iff \frac{Pr([x]|C)}{Pr([x]|C^c)} \geq \frac{Pr(C^c)}{Pr(C)} \cdot \frac{\alpha}{1-\alpha}. \end{aligned} \tag{3.27}$$

Similar expressions can be obtained for the negative and boundary regions. Consequently, we can express the three regions in terms of the likelihood ratio as:

$$\begin{aligned} \text{POS}_{(\alpha,\beta)}(C) &= \{x \in U \mid \frac{Pr([x]|C)}{Pr([x]|C^c)} \geq \frac{Pr(C^c)}{Pr(C)} \cdot \frac{\alpha}{1-\alpha}\}, \\ \text{BND}_{(\alpha,\beta)}(C) &= \{x \in U \mid \frac{Pr(C^c)}{Pr(C)} \cdot \frac{\beta}{1-\beta} < \frac{Pr([x]|C)}{Pr([x]|C^c)} < \frac{Pr(C^c)}{Pr(C)} \cdot \frac{\alpha}{1-\alpha}\}, \\ \text{NEG}_{(\alpha,\beta)}(C) &= \{x \in U \mid \frac{Pr([x]|C)}{Pr([x]|C^c)} \leq \frac{Pr(C^c)}{Pr(C)} \cdot \frac{\beta}{1-\beta}\}. \end{aligned} \tag{3.28}$$

The conditions about the likelihood ratio are obtained by the same transformation of the original thresholds α and β with an adjustment of the a priori odds of C.

Another widely used monotonically increasing transformation of probability function is the logit transformation defined by $\text{logit}(Pr(\cdot)) = \log(O(\cdot)) = \log \frac{Pr(\cdot)}{1-Pr(\cdot)}$. By applying logarithm function to equation (3.27), we can express the three regions as:

$$\begin{aligned} \text{POS}_{(\alpha,\beta)}(C) &= \{x \in U \mid \log \frac{Pr([x]|C)}{Pr([x]|C^c)} \geq \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\alpha}{1-\alpha}\}, \\ \text{BND}_{(\alpha,\beta)}(C) &= \{x \in U \mid \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\beta}{1-\beta} < \log \frac{Pr([x]|C)}{Pr([x]|C^c)} < \\ &\quad \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\alpha}{1-\alpha}\}, \\ \text{NEG}_{(\alpha,\beta)}(C) &= \{x \in U \mid \log \frac{Pr([x]|C)}{Pr([x]|C^c)} \leq \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\beta}{1-\beta}\}. \end{aligned} \tag{3.29}$$

Again, conditions about the value of the logit function are obtained from the same transformation of original thresholds α and β with an adjustment of the a priori odds.

Both the likelihood ratio and the logarithm of likelihood ratio are examples of Bayesian confirmation measures. By looking at the forms of the last two new but equivalent definitions of probabilistic approximations, one may confuse them with Bayesian confirmation regions. Since the a priori odds of C is involved, the thresholds are in fact not on the corresponding Bayesian confirmation measure, but on a Bayesian confirmation modified by the a priori odds.

3.4.2 Naive Bayesian Rough Sets

Naive Bayesian classification provides an effective method to estimate the likelihood by representing an object as a feature vector and assuming that the features are probabilistically independent [134]. Its application to rough set theory leads to a Naive Bayesian rough set model [131, 132].

The description of an equivalence class $[x]$ can be represented by a feature vector $\text{Des}(x) = (v_{a_1}, v_{a_2}, ..., v_{a_n})$. In a Bayesian rough set model, we need to estimate a joint probabilities $Pr([x]|C) = Pr(v_1, v_2, ..., v_n|C)$ and a joint probability of $Pr([x]) = Pr(v_1, v_2, ..., v_n)$. In practice, it is difficult to analyze the interactions between the components of $[x]$, especially when the number n is large. A common solution to this problem is to calculate the likelihood based on the naive conditional independence assumption [29]. That is, we assume each component v_i of $[x]$ to be conditionally independent of every other component v_j for $j \neq i$. Although this assumption may seem overly simplistic, but many empirical studies showed its effectiveness for classification problems [17, 47, 134].

Formally, the probabilistic independence assumptions are given by:

$$\begin{aligned} Pr([x]|C) &= Pr(v_1, v_2, ..., v_n|C) = \prod_{i=1}^{i=n} Pr(v_i|C), \\ Pr([x]|C^c) &= Pr(v_1, v_2, ..., v_n|C^c) = \prod_{i=1}^{i=n} Pr(v_i|C^c). \end{aligned} \tag{3.30}$$

By inserting them into equation (3.29), we can compute the logarithm of the likelihood ratio as:

$$\log \frac{Pr([x]|C)}{Pr([x]|C^c)} = \log \prod_{i=1}^{i=n} \frac{Pr(v_i|C)}{Pr(v_i|C^c)} = \sum_{i=1}^{i=n} \log \frac{Pr(v_i|C)}{Pr(v_i|C^c)}. \tag{3.31}$$

Now, the three probabilistic regions can be computed by:

$$\begin{aligned} \mathrm{POS}_{(\alpha,\beta)}(C) &= \{x \in U \mid \sum_{i=1}^{i=n} \log \frac{Pr(v_i|C)}{Pr(v_i|C^c)} \geq \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\alpha}{1-\alpha}\}, \\ \mathrm{BND}_{(\alpha,\beta)}(C) &= \{x \in U \mid \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\beta}{1-\beta} < \sum_{i=1}^{i=n} \log \frac{Pr(v_i|C)}{Pr(v_i|C^c)} < \\ &\quad \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\alpha}{1-\alpha}\}, \\ \mathrm{NEG}_{(\alpha,\beta)}(C) &= \{x \in U \mid \sum_{i=1}^{i=n} \log \frac{Pr(v_i|C)}{Pr(v_i|C^c)} \leq \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\beta}{1-\beta}\}. \end{aligned} \tag{3.32}$$

The individual probabilities can be estimated from the frequencies of the training data:

$$\begin{aligned} Pr(v_i|C) &= \frac{|m(a_i = v_i) \cap C|}{|C|}, \\ Pr(C) &= \frac{|C|}{|U|}, \end{aligned}$$

where $|\cdot|$ denotes the cardinality of a set, and $m(a_i = v_i) = \{x \in U \mid I_{a_i}(x) = v_i\}$ denotes the set of all objects having the property expressed by the formula $a_i = v_i$. Thus, a simple method for probability estimation is obtained in a naive Bayesian rough set model [131, 132].

3.4.3 A Binary Probabilistic Independence Rough Set Model

In this subsection, I consider a binary probabilistic independent model [20] for a specific classification problem to show the usefulness of Naive Bayesian rough sets. In this model, all the feature vectors are binary valued. That is, an object x is represented by $[x] = (v_1, v_2, ..., v_n)$ where $v_i = 1$ if feature i is present for the object x and $v_i = 0$ if i is not present for x.

Let $p_i = Pr(v_i = 1|C)$ denote the probability of a feature presenting in an object belong to class C, and $q_i = Pr(v_i = 1|C^c)$ denote the probability of a feature presenting in an object belong to class C^c. We can rewrite equation (3.31) as:

$$\begin{aligned}
\sum_{i=1}^{i=n} \log \frac{Pr(v_i|C)}{Pr(v_i|C^c)} &= \sum_{i=1}^{i=n} \log \frac{Pr(v_i = 1|C)^{v_i} Pr(v_i = 0|C)^{1-v_i}}{Pr(v_i = 1|C^c)^{v_i} Pr(v_i = 0|C^c)^{1-v_i}} \\
&= \sum_{i=1}^{i=n} \log \frac{p_i^{v_i}(1-p_i)^{1-v_i}}{q_i^{v_i}(1-q_i)^{1-v_i}} \\
&= \sum_{i=1}^{i=n} v_i \log \frac{p_i(1-q_i)}{q_i(1-p_i)} + \sum_{i=1}^{i=n} \log \frac{1-p_i}{1-q_i} \\
&= \sum_{i=1}^{i=n} w_i v_i + w_0, \qquad (3.33)
\end{aligned}$$

where $w_i = \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$, $i = 1, 2, ..., n$, and $w_0 = \sum_{i=1}^{i=n} \frac{1-p_i}{1-q_i}$. With these notations, the three probabilistic regions can be computed by:

$$\begin{aligned}
\mathrm{POS}^B_{(\alpha,\beta)}(C) &= \{x \in U \mid \sum_{i=1}^{i=n} w_i v_i \geq \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\alpha}{1-\alpha} - w_0\}, \\
\mathrm{BND}^B_{(\alpha,\beta)}(C) &= \{x \in U \mid \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\beta}{1-\beta} - w_0 < \sum_{i=1}^{i=n} w_i v_i < \\
&\quad \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\alpha}{1-\alpha} - w_0\}, \\
\mathrm{NEG}^B_{(\alpha,\beta)}(C) &= \{x \in U \mid \sum_{i=1}^{i=n} w_i v_i \leq \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\beta}{1-\beta} - w_0\}.
\end{aligned}$$

The value w_i may be viewed as a weight of feature i.

3.4.4 An Example

Table 3.2 is an information table that contains eight patients. Each patient is represented by five symptoms whose values are either 1 (i.e., a particular symptom is present) or 0 (i.e., the symptom is absent). Class is the decision attribute that denotes whether a patient has cancer (Class=1) or non-cancer (Class=0).

In practice, the probability or the odds of patients with cancer can be more accurately estimated based on past experience. In this example, I estimate the probability based on the frequencies of the training examples:

$$Pr(C) = Pr(\text{Class} = 1) = \frac{|\{o_1, o_3, o_4, o_6, o_7\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8\}|} = \frac{5}{8},$$

$$Pr(C^c) = Pr(\text{Class} = 0) = \frac{|\{o_2, o_5, o_8\}|}{|\{o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8\}|} = \frac{3}{8}.$$

We can get $\log(Pr(C^c)/Pr(C)) = -0.22$, where base 10 is used. The probabilities p_i's and q_i's can be estimated by:

$$p_i = \frac{\text{the number of cancer patients with the } i\text{th symptom present}}{\text{the number of cancer patients}},$$
$$q_i = \frac{\text{the number of non-cancer patients with the } i\text{th symptom present}}{\text{the number of non-cancer patients}}.$$

For example, for $S_4 = 1$, we have:

$$p_4 = Pr(S_4 = 1|\text{Class} = 1) = \frac{|\{o_1, o_3, o_5, o_6\} \cap \{o_1, o_3, o_4, o_6, o_7\}|}{|\{o_1, o_3, o_4, o_6, o_7\}|} = \frac{3}{5},$$
$$q_4 = Pr(S_4 = 1|\text{Class} = 0) = \frac{|\{o_1, o_3, o_5, o_6\} \cap \{o_2, o_5, o_8\}|}{|\{o_2, o_5, o_8\}|} = \frac{1}{3}.$$

Table 3.2: The training set

Patients	S_1	S_2	S_3	S_4	S_5	Class
o_1	1	1	0	1	1	1
o_2	1	0	1	0	0	0
o_3	0	0	1	1	0	1
o_4	1	0	0	0	1	1
o_5	0	1	0	1	1	0
o_6	1	0	0	1	1	1
o_7	1	0	1	0	1	1
o_8	1	0	1	0	0	0

Table 3.3: The testing set

Patients	S_1	S_2	S_3	S_4	S_5	Class
o_9	0	0	0	1	1	?
o_{10}	0	1	0	1	0	?

Thus, the weight factor $w_4 = \log \frac{p_4(1-q_4)}{q_4(1-p_4)} = 0.48$. Similarly, we can estimate the probabilities for the remaining attributes:

$$
\begin{aligned}
p_1 &= \frac{4}{5}, \quad q_1 = \frac{2}{3}, \\
p_2 &= \frac{1}{5}, \quad q_2 = \frac{1}{3}, \\
p_3 &= \frac{2}{5}, \quad q_3 = \frac{2}{3}, \\
p_5 &= \frac{4}{5}, \quad q_5 = \frac{1}{3}.
\end{aligned}
$$

They produce the following weights:

$$w_1 = 0.30, \quad w_2 = -0.30, \quad w_3 = -0.48, \quad w_5 = 0.90.$$

Thus,

$$\begin{aligned}\sum_{i=1}^{i=5} w_i v_i &= w_1 v_1 + w_2 v_2 + w_3 v_3 + w_4 v_4 + w_5 v_5 \\ &= 0.30 \times v_1 + (-0.30) \times v_2 + (-0.48) \times v_3 + 0.48 \times v_4 + 0.90 \times v_5.\end{aligned}$$

The pair of thresholds on this discriminant function according to the values of loss functions in Table 3.1 (page 45) is calculated as:

$$\begin{aligned}& \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\alpha}{1-\alpha} - w_0 \\ = \;& \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\lambda_{PN} - \lambda_{BN}}{\lambda_{BP} - \lambda_{PP}} - \sum_{i=1}^{n} \log \frac{1-p_i}{1-q_i} \\ = \;& -0.22 + \log \frac{600-300}{300-200} - (-0.62) = 0.88,\end{aligned}$$

$$\begin{aligned}& \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\beta}{1-\beta} - w_0 \\ = \;& \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\lambda_{BN} - \lambda_{NN}}{\lambda_{NP} - \lambda_{BP}} - \sum_{i=1}^{n} \log \frac{1-p_i}{1-q_i} \\ = \;& -0.22 + \log \frac{300-0}{1000-300} - (-0.62) = 0.03.\end{aligned}$$

We can derive three decision rules for processing a patient represented by $(v_1, v_2, v_3, v_4, v_5)$:

1. If $0.30 \times v_1 + (-0.30) \times v_2 + (-0.48) \times v_3 + 0.48 \times v_4 + 0.90 \times v_5 \geq 0.88$, the doctor will treat the patient;

2. If $0.03 < 0.30 \times v_1 + (-0.30) \times v_2 + (-0.48) \times v_3 + 0.48 \times v_4 + 0.90 \times v_5 < 0.88$, the doctor will further diagnose the patient;

3. If $0.30 \times v_1 + (-0.30) \times v_2 + (-0.48) \times v_3 + 0.48 \times v_4 + 0.90 \times v_5 \leq 0.03$, the doctor will not treat the patient.

Consider two new patients in Table 3.3. For the patient o_9, we have:

$$\begin{aligned} & 0.30 \times v_1 + (-0.30) \times v_2 + (-0.48) \times v_3 + 0.48 \times v_4 + 0.90 \times v_5 \\ = \; & 0.30 \times 0 + (-0.30) \times 0 + (-0.48) \times 0 + 0.48 \times 1 + 0.90 \times 1 \\ = \; & 1.38 > 0.88. \end{aligned}$$

The doctor will treat o_9. For the patient o_{10}, we have:

$$\begin{aligned} & 0.30 \times v_1 + (-0.30) \times v_2 + (-0.48) \times v_3 + 0.48 \times v_4 + 0.90 \times v_5 \\ = \; & 0.30 \times 0 + (-0.30) \times 1 + (-0.48) \times 0 + 0.48 \times 1 + 0.90 \times 0 \\ = \; & 0.18. \end{aligned}$$

In this case, since the value is in between the two thresholds $0.03 < 0.18 < 0.88$, the doctor can not make an immediate decision, and further testing is needed.

3.5 Building Ternary Classifiers with Probabilistic Rough Sets

The three probabilistic regions can be used to build a ternary classifier for three-way classification. The introduction of a third choice makes a ternary classifier more advantageous than a binary classifier under certain conditions.

3.5.1 Three-Way Classifications

There are extensive studies on inferring classification rules from rough set approximations [36, 37, 69]. The majority of them focus on rule discovery methods and characterization of rules in terms of statistical measures such as generality, accuracy, coverage, etc. [48, 66, 102, 108, 126]. However, there is an insufficient treatment on semantics and implications of the discovered rules. Recently, the notion of three-way decisions was introduced to provide a new interpretation of rules derived in rough set theory [121, 122, 142]. That is, classification rules in rough set theory produce a ternary classifier [119].

Recall that a main result of decision-theoretic rough set model is the (α, β)-probabilistic positive, boundary and negative regions defined by:

$$\begin{aligned} \mathrm{POS}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C|[x]) \geq \alpha\}, \\ \mathrm{BND}_{(\alpha,\beta)}(C) &= \{x \in U \mid \beta < Pr(C|[x]) < \alpha\}, \\ \mathrm{NEG}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C|[x]) \leq \beta\}. \end{aligned} \tag{3.34}$$

They provide three-way classification of objects with respect to C. A pair of thresholds on probabilities defines three regions for classification in rough set theory.

Let $\mathrm{Des}(x)$ denote the description of objects in the equivalence class $[x]$. Based on the probabilistic three regions, we can derive following three types of positive, boundary and negative rules for classification:

$$\begin{aligned} &\mathrm{Des}(x) \longrightarrow_P C, \quad \text{if } [x] \subseteq \mathrm{POS}(C); \\ &\mathrm{Des}(x) \longrightarrow_B C, \quad \text{if } [x] \subseteq \mathrm{BND}(C); \\ &\mathrm{Des}(x) \longrightarrow_N C, \quad \text{if } [x] \subseteq \mathrm{NEG}(C). \end{aligned} \tag{3.35}$$

Although these three types of rules have the same form, they have different interpretations, and hence lead to different decisions and actions. To be consistent with the interpretation of three regions, we can make one of three decisions when classifying an object. A positive rule is a rule for acceptance: if $[x] \subseteq \text{POS}(C)$, we accept x to be an instance of C, that is, we accept $x \in C$. A negative rule is a rule for rejection: if $[x] \subseteq \text{NEG}(C)$, we reject x to be an instance of C, that is, we reject $x \in C$. A boundary rule is a rule for deferment: if $[x] \subseteq \text{BND}(C)$, we neither accept nor reject $x \in C$; instead, we defer such a definite decision. Accordingly, we also call the three types of rules the acceptance, deferment and rejection rules. By applying the three types of rules, one can easily classify an object through a decision of acceptance, rejection or deferment. The result is a ternary classifier.

Ternary classifiers derived from probabilistic rough sets can be used in many applications. Consider the medical example discussed earlier. A positive rule indicates that the doctor will treat the patient immediately, a negative rule indicates that the doctor decides not to treat the patient, and a boundary rule indicates that a deferred decision is made for a patient due to insufficient information and further diagnose is required.

3.5.2 Comparison with Two-Way Classifications

In Bayesian classification, a monotonically increasing function of $Pr(C|[x])$ is called a discriminant function [20]. There are typically two ways to use a discriminant function for classification. One is to use the discriminant function to rank objects and let a user to classify objects by reading through the ranked list. The other is to set a

threshold. This produces a binary classifier: objects whose values are above or equal to the threshold are accepted as instances of the class and whose values below the threshold are rejected. The Bayesian decision theory can be used to systematically compute the threshold [20].

Formally, Bayesian two-way classification can be described as follows. Let $\gamma \in [0, 1]$ denote a threshold on the a posteriori probability $Pr(C|[x])$. One can divide the set of objects into two regions as approximations of C, namely, the γ-probabilistic positive and negative regions:

$$\begin{aligned} \mathrm{POS}_\gamma(C) &= \{x \in U \mid Pr(C|[x]) \geq \gamma\}, \\ \mathrm{NEG}_\gamma(C) &= \{x \in U \mid Pr(C|[x]) < \gamma\}. \end{aligned} \tag{3.36}$$

They in turn induce two types of rules, that is, the positive rules for acceptance and the negative rules for rejection. Note that the probability $Pr(C|[x])$ can be estimated by the same methods discussed in the last section.

The differences between binary and three-way classifications are demonstrated in Figure 3.2. As stated in [122], a three-way classification is advantageous under the condition $\beta < \gamma < \alpha$. In this case, for the two intervals $[0, \beta]$ and $[\alpha, 1]$, binary and ternary classifier make the same decision of acceptance and rejection, respectively, and for interval (α, β), a ternary classifier chooses a deferment rather than an acceptance or a rejection. It follows that,

$$\begin{aligned} &\mathrm{POS}_{(\alpha,\beta)}(C) \subseteq \mathrm{POS}_\gamma(C), \\ &\mathrm{BND}_\gamma(C) \subseteq \mathrm{BND}_{(\alpha,\beta)}(C), \\ &\mathrm{NEG}_{(\alpha,\beta)}(C) \subseteq \mathrm{NEG}_\gamma(C). \end{aligned} \tag{3.37}$$

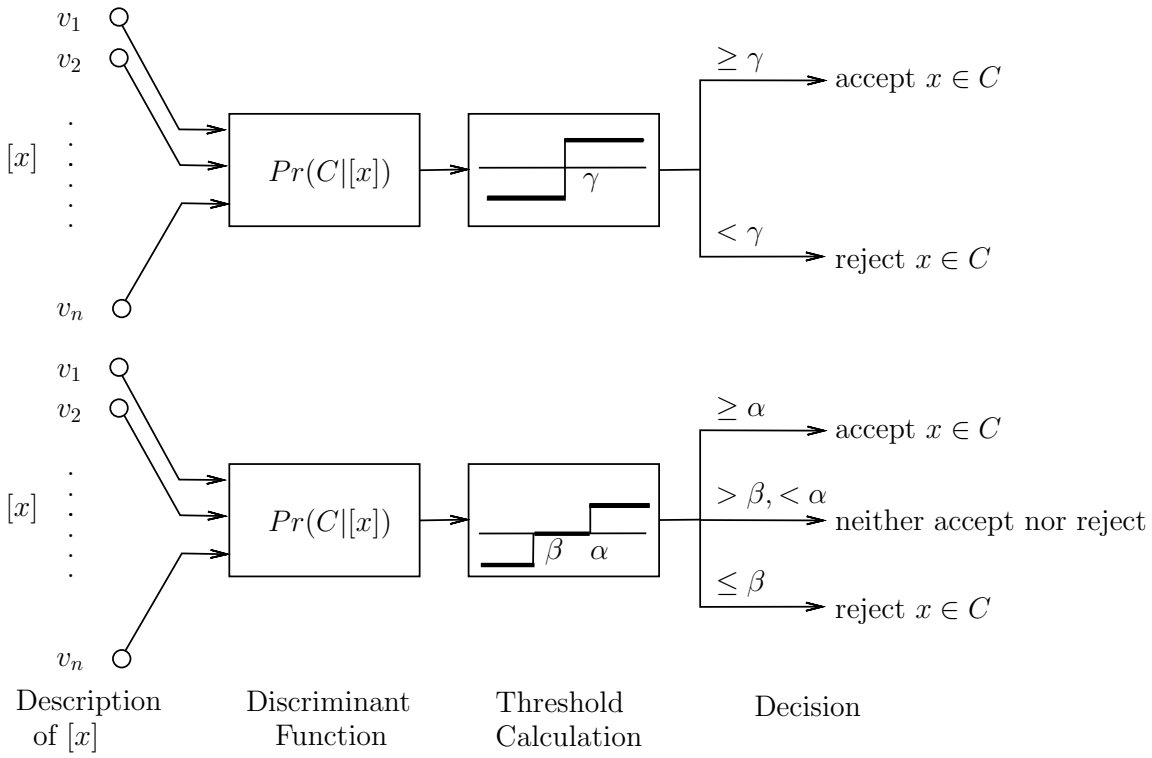

Figure 3.2: Comparison of binary and ternary classifiers

That is, a ternary classifier move some objects from the positive and negative regions of a binary classifier into a deferment region.

Binary and ternary classifiers can be compared in terms of various types of classification errors and costs [122]. A binary classifier may produce two types of errors, namely, incorrect acceptance and incorrect rejection. A ternary classifier may produce two additional types of errors, namely, deferment of positive and deferment of negative. Table 3.4 and Table 3.5 are two confusion matrices resulting from binary and ternary classifiers, respectively. The symbols in the two tables denote the following numbers:

Table 3.4: Confusion matrix resulting from a binary classifier

	C (P)	C^c (N)
accept	$TP = \|C \cap \mathrm{POS}_\gamma(C)\|$	$FP = \|C^c \cap \mathrm{POS}_\gamma(C)\|$
reject	$FN = \|C \cap \mathrm{NEG}_\gamma(C)\|$	$TN = \|C^c \cap \mathrm{NEG}_\gamma(C)\|$

TP (True Positive) = the number of correctly classified positive examples,

FP (False Positive)= the number of incorrectly classified negative examples,

FN (False Negative)= the number of incorrectly classified positive examples,

TN (True Negative)= the number of correctly classified negative examples,

and

AP (Accepted Positive)= the number of correctly accepted positive examples,

AN (Accepted Negative)= the number of incorrectly accepted negative examples,

DP (Deferred Positive)= the number of deferred positive examples,

DN (Deferred Negative) = the number of deferred negative examples,

RP (Rejected Positive) = the number of incorrectly rejected positive examples,

RN (Rejected Negative) = the number of correctly rejected negative examples.

Although the four numbers of a binary classifier have the same interpretation as those in a ternary classifier, new names are used for the latter to emphasize their corresponding decisions [122].

Binary and ternary classifiers can be compared with respect to errors. The acceptance error is the proportion of accepted examples which are actually not in class C. Let $AccE_2$ and $AccE_3$ denote the acceptance errors of the binary and three-way

Table 3.5: Confusion matrix resulting from a three-way classifier

	C (P)	C^c (N)
accept	$AP = \lvert C \cap \mathrm{POS}_{(\alpha,\beta)}(C)\rvert$	$AN = \lvert C^c \cap \mathrm{POS}_{(\alpha,\beta)}(C)\rvert$
defer	$DP = \lvert C \cap \mathrm{BND}_{(\alpha,\beta)}(C)\rvert$	$DN = \lvert C^c \cap \mathrm{BND}_{(\alpha,\beta)}(C)\rvert$
reject	$RP = \lvert C \cap \mathrm{NEG}_{(\alpha,\beta)}(C)\rvert$	$RN = \lvert C^c \cap \mathrm{NEG}_{(\alpha,\beta)}(C)\rvert$

classifiers, respectively. They are defined by:

$$\begin{aligned} AccE_2 &= \frac{FP}{TP+FP} = \frac{|C^c \cap \mathrm{POS}_\gamma(C)|}{|\mathrm{POS}_\gamma(C)|}, \\ AccE_3 &= \frac{AN}{AP+AN} = \frac{|C^c \cap \mathrm{POS}_{(\alpha,\beta)}(C)|}{|\mathrm{POS}_{(\alpha,\beta)}(C)|}. \end{aligned} \tag{3.38}$$

Similarly, the rejection error is the proportion of rejected examples which are actually in class C. The rejection errors of a binary classifier and a three-way classifier are defined respectively by:

$$\begin{aligned} RejE_2 &= \frac{FN}{FN+TN} = \frac{|C \cap \mathrm{NEG}_\gamma(C)|}{|\mathrm{NEG}_\gamma(C)|}, \\ RejE_3 &= \frac{RP}{RP+RN} = \frac{|C \cap \mathrm{NEG}_{(\alpha,\beta)}(C)|}{|\mathrm{NEG}_{(\alpha,\beta)}(C)|}. \end{aligned} \tag{3.39}$$

By the condition $\beta < \gamma < \alpha$, we can express $\mathrm{POS}_\gamma(C)$ as $\mathrm{POS}_{(\alpha,\beta)}(C) \cup M$, where $M = \{x \mid \gamma \leq Pr(C|[x]) < \alpha\}$. Therefore, the difference between acceptance errors

of binary and ternary classifiers is given by:

$$\begin{aligned} AccE_2 - AccE_3 &= \frac{|C^c \cap \mathrm{POS}_\gamma(C)|}{|\mathrm{POS}_\gamma(C)|} - \frac{|C^c \cap \mathrm{POS}_{(\alpha,\beta)}(C)|}{|\mathrm{POS}_{(\alpha,\beta)}(C)|} \\ &= (1 - \frac{|C \cap \mathrm{POS}_\gamma(C)|}{|\mathrm{POS}_\gamma(C)|}) - (1 - \frac{|C \cap \mathrm{POS}_{(\alpha,\beta)}(C)|}{|\mathrm{POS}_{(\alpha,\beta)}(C)|}) \\ &= \frac{|C \cap \mathrm{POS}_{(\alpha,\beta)}(C)|}{|\mathrm{POS}_{(\alpha,\beta)}(C)|} - \frac{|C \cap \mathrm{POS}_\gamma(C)|}{|\mathrm{POS}_\gamma(C)|} \\ &= \frac{|C \cap \mathrm{POS}_{(\alpha,\beta)}(C)|}{|\mathrm{POS}_{(\alpha,\beta)}(C)|} - \frac{|C \cap (\mathrm{POS}_{(\alpha,\beta)}(C) \cup M)|}{|\mathrm{POS}_{(\alpha,\beta)}(C) \cup M|} \\ &= \frac{|C \cap \mathrm{POS}_{(\alpha,\beta)}(C)|}{|\mathrm{POS}_{(\alpha,\beta)}(C)|} - \frac{|C \cap \mathrm{POS}_{(\alpha,\beta)}(C)| + |C \cap M|}{|\mathrm{POS}_{(\alpha,\beta)}(C)| + |M|}. \end{aligned} \tag{3.40}$$

By the definition of probabilistic regions of binary and ternary classifications, and the condition $\beta < \gamma < \alpha$, we can establish the following facts:

$$\begin{aligned} |C \cap \mathrm{POS}_{(\alpha,\beta)}(C)| &= |C \cap (\bigcup\{[x] \mid \frac{|C \cap [x]|}{|[x]|} \geq \alpha\})| \\ &= |\bigcup\{C \cap [x] \mid \frac{|C \cap [x]|}{|[x]|} \geq \alpha\}| \\ &= \sum_{\frac{|C\cap[x]|}{|[x]|} \geq \alpha} \mid C \cap [x] \mid \\ &\geq \sum_{\frac{|C\cap[x]|}{|[x]|} \geq \alpha} \alpha \mid [x] \mid \\ &= \alpha \mid \mathrm{POS}_{\alpha,\beta}(C) \mid, \end{aligned} \tag{3.41}$$

$$\begin{aligned} |C \cap M| &= |C \cap (\bigcup\{[x] \mid \gamma \leq \frac{|C \cap [x]|}{|[x]|} < \alpha\})| \\ &= |\bigcup\{C \cap [x] \mid \gamma \leq \frac{|C \cap [x]|}{|[x]|} < \alpha\}| \\ &= \sum_{\gamma \leq \frac{|C\cap[x]|}{|[x]|} < \alpha} \mid C \cap [x] \mid \\ &\leq \sum_{\gamma \leq \frac{|C\cap[x]|}{|[x]|} < \alpha} \alpha \mid [x] \mid \\ &= \alpha \mid M \mid, \end{aligned} \tag{3.42}$$

where the summation is over all equivalence classes satisfying a certain condition.

Thus, equation (3.40) can be rewritten as:

$$\begin{aligned}
& \frac{|C \cap \mathrm{POS}_{(\alpha,\beta)}(C)| \cdot (|\mathrm{POS}_{(\alpha,\beta)}(C)| + |M|) - |\mathrm{POS}_{(\alpha,\beta)}(C)| \cdot (|C \cap \mathrm{POS}_{(\alpha,\beta)}(C)| + |C \cap M|)}{|\mathrm{POS}_{(\alpha,\beta)}(C)| \cdot (|\mathrm{POS}_{(\alpha,\beta)}(C)| + |M|)} \\
= & \frac{|C \cap \mathrm{POS}_{(\alpha,\beta)}(C)| \cdot |M| - |\mathrm{POS}_{(\alpha,\beta)}(C)| \cdot |C \cap M|}{|\mathrm{POS}_{(\alpha,\beta)}(C)| \cdot (|\mathrm{POS}_{(\alpha,\beta)}(C)| + |M|)} \\
\geq & \frac{\alpha|\mathrm{POS}_{(\alpha,\beta)}(C)| \cdot |M| - \alpha|\mathrm{POS}_{(\alpha,\beta)}(C)| \cdot |M|}{|\mathrm{POS}_{(\alpha,\beta)}(C)| \cdot (|\mathrm{POS}_{(\alpha,\beta)}(C)| + |M|)} \\
\geq & 0. \qquad (3.43)
\end{aligned}$$

Therefore, $AccE_2 \geq AccE_3$. Similarly, it can be shown that $RejE_2 \geq RejE_3$. That is, both acceptance and rejection errors of the three-way classification are lower than the corresponding errors in binary classification. A ternary classifier reduces the acceptance or rejection errors by introducing deferment errors.

According to the optimality of Bayesian decision theory, the trade-off made by a ternary classifier, under the condition $\beta < \gamma < \alpha$, leads to lower overall cost of classification. The conditions on the loss function to ensure $\beta < \gamma < \alpha$ and a comparison of binary and ternary classification with respect to classification costs can be found in [122].

3.5.3 An Example

Suppose we have a set of 700 patients, which are divided into four equivalence classes of 300, 200, 100, and 100 patients, respectively. Let C denote a subset of 364 cancer patients. The relationships between C and the equivalence classes are given in Table 3.6. An integer in each cell of the second and third columns is the number of patients in the corresponding set. The number in each cell of the last two columns is a conditional probability.

Table 3.6: Relationships between C and the equivalence classes

	$C \cap E$	$C^c \cap E$	$Pr(C\|E)$	$Pr(C^c\|E)$
$E_1(300)$	270	30	0.90	0.10
$E_2(200)$	80	120	0.40	0.60
$E_3(100)$	9	91	0.09	0.91
$E_4(100)$	5	95	0.05	0.95

Table 3.7: Loss function of another medical example

	C (P)(cancer)	C^c (N)(non-cancer)
a_P(Treat)	$\lambda_{PP} = 0$	$\lambda_{PN} = 10$
a_B(Further test)	$\lambda_{BP} = 5$	$\lambda_{BN} = 5$
a_N(Not Treat)	$\lambda_{NP} = 90$	$\lambda_{NN} = 0$

Using the loss functions provided in Table 3.7, by inserting them in equation (3.6), we obtain the following thresholds:

$$\alpha = 0.50,\ \beta = 0.06,\ \gamma = 0.10.$$

According to these thresholds, the three probabilistic regions of C of the ternary classifier are given by:

$$\mathrm{POS}_{(0.50,0.06)}(C) = E_1,$$

$$\mathrm{BND}_{(0.50,0.06)}(C) = E_2 \cup E_3,$$

$$\mathrm{NEG}_{(0.50,0.06)}(C) = E_4,$$

and the two probabilistic regions of the binary classifier are given by:

$$\mathrm{POS}_{0.10}(C) = E_1 \cup E_2,$$

$$\mathrm{NEG}_{0.10}(C) = E_3 \cup E_4.$$

Note that the boundary region of a binary classifier is always empty and is added for notational simplicity.

We compute the acceptance errors of the the ternary and binary classifiers based on equation (3.38):

$$\begin{aligned} AccE_2 &= \frac{|C^c \cap \mathrm{POS}_{0.10}(C)|}{|\mathrm{POS}_{0.10}(C)|} \\ &= \frac{|C^c \cap (E_1 \cup E_2)|}{|E_1 \cup E_2|} \\ &= \frac{|(C^c \cap E_1) \cup (C^c \cap E_2)|}{|E_1 \cup E_2|} \\ &= \frac{150}{500} = 0.30, \end{aligned}$$

$$\begin{aligned} AccE_3 &= \frac{|C^c \cap \mathrm{POS}_{(0.50,0.06)}(C)|}{|\mathrm{POS}_{(0.50,0.06)}(C)|} \\ &= \frac{|C^c \cap E_1|}{|E_1|} \\ &= \frac{30}{300} = 0.10. \end{aligned}$$

It can be seen that the acceptance error of the ternary classifier is lower than the acceptance error of the binary classifier. Similarly, we can compute the rejection

errors based on equation (3.39):

$$\begin{aligned} RejE_2 &= \frac{|C \cap \mathrm{NEG}_{0.10}(C)|}{|\mathrm{NEG}_{0.10}(C)|} \\ &= \frac{|C \cap (E_3 \cup E_4)|}{|E_3 \cup E_4|} \\ &= \frac{|(C \cap E_3) \cup (C \cap E_4)|}{|E_3 \cup E_4|} \\ &= \frac{14}{200} = 0.07, \end{aligned}$$

$$\begin{aligned} RejE_3 &= \frac{|C \cap \mathrm{NEG}_{(0.50,0.06)}(C)|}{|\mathrm{NEG}_{(0.50,0.06)}(C)|} \\ &= \frac{|C \cap E_4|}{|E_4|} \\ &= \frac{5}{100} = 0.05. \end{aligned}$$

The rejection error of the ternary classifier is lower than the rejection error of the binary classifier.

For the equivalence class E_1, both the ternary and the binary classifiers make a decision of acceptance. The associated cost is given by:

$$\begin{aligned} R^t(\mathrm{Accept}|E_1) = R^b(\mathrm{Accept}|E_1) &= \lambda_{PP} Pr(C|E_1) + \lambda_{PN} Pr(C^c|E_1) \\ &= 0 \times 0.90 + 10 \times 0.10 = 1. \end{aligned}$$

For the equivalence class E_2, the ternary classifier makes a deferment decision by putting E_2 into the boundary region, while the binary classifier makes a decision of acceptance. The associated cost are:

$$\begin{aligned} R^t(\mathrm{Defer}|E_2) &= \lambda_{BP} Pr(C|E_2) + \lambda_{BN} Pr(C^c|E_2) \\ &= 5 \times 0.40 + 5 \times 0.60 = 5, \end{aligned}$$

and

$$\begin{aligned} R^b(\text{Accept}|E_2) &= \lambda_{PP}Pr(C|E_2) + \lambda_{PN}Pr(C^c|E_2) \\ &= 0 \times 0.40 + 10 \times 0.60 = 6. \end{aligned}$$

For the equivalence class E_3, the ternary classifier makes a deferment decision by putting E_3 into the boundary region, while the binary classifier makes a decision of rejection. The associated cost are:

$$\begin{aligned} R^t(\text{Defer}|E_3) &= \lambda_{BP}Pr(C|E_3) + \lambda_{BN}Pr(C^c|E_3) \\ &= 5 \times 0.09 + 5 \times 0.91 = 5, \end{aligned}$$

and

$$\begin{aligned} R^b(\text{Reject}|E_3) &= \lambda_{NP}Pr(C|E_3) + \lambda_{NN}Pr(C^c|E_3) \\ &= 90 \times 0.09 + 0 \times 0.91 = 8.1. \end{aligned}$$

For the equivalence class E_4, both the ternary and the binary classifiers make a decision of rejection. The associated cost is:

$$\begin{aligned} R^t(\text{Reject}|E_4) = R^b(Reject|E_4) &= \lambda_{NP}Pr(C|E_4) + \lambda_{NN}Pr(C^c|E_4) \\ &= 90 \times 0.05 + 0 \times 0.95 = 4.5. \end{aligned}$$

The overall risk for the ternary and binary classifier can be computed by plugging the above costs into equation (3.2):

$$\begin{aligned} \mathbf{R}^t &= \sum_{i=1}^{i=4} R(\tau(E_i)|E_i)Pr(E_i) \\ &= R^t(\text{Accept}|E_1)Pr(E_1) + R^t(\text{Defer}|E_2)Pr(E_2) + R^t(\text{Defer}|E_3)Pr(E_3) \\ &\quad + R^t(\text{Reject}|E_4)Pr(E_4) \\ &= 1 \times \frac{300}{700} + 5 \times \frac{200}{700} + 5 \times \frac{100}{700} + 4.5 \times \frac{100}{700} = 3.21, \end{aligned}$$

$$
\begin{aligned}
\mathbf{R}^b &= \sum_{i=1}^{i=4} R(\tau(E_i)|E_i)Pr(E_i) \\
&= R^b(\text{Accept}|E_1)Pr(E_1) + R^b(\text{Accept}|E_2)Pr(E_2) + R^b(\text{Reject}|E_3)Pr(E_3) \\
&\quad + R^b(\text{Reject}|E_4)Pr(E_4) \\
&= 1 \times \frac{300}{700} + 6 \times \frac{200}{700} + 8.1 \times \frac{100}{700} + 4.5 \times \frac{100}{700} = 3.94.
\end{aligned}
$$

The overall cost of the ternary classifier is lower than the cost of the binary classifier.

3.6 Summary

In this chapter, a complete model of Bayesian rough sets is proposed by investigating three main components of the model. The required thresholds for defining probabilistic approximations are interpreted and computed based on a decision-theoretic rough set model. The Bayes' theorem and Bayesian inference are used to estimate the required probability. In particular, a naive Bayesian model is studied. Finally, a new semantic interpretation of probabilistic rough set is examined based on a framework of three-way decisions. The introduction of a third choice makes a ternary classifier to be more advantageous than a binary classifier under certain conditions.

Chapter 4

THREE-WAY EMAIL SPAM FILTERING

Email spam filtering is a typical example of automated text categorization. In this chapter, I demonstrate the usefulness of the proposed framework by applying it to email spam filtering.

4.1 Automated Text Categorization

Text categorization, also known as document classification or topic spotting, is the task of automatically classifying an electronic document into a set of categories based on its contents [84,94]. It has attracted a lot of interests from researchers over the last few decades since manually classifying the large volume of documents in an organizations' databases or generated from the Internet can be too expensive, or simply not feasible given the time constraints of the applications. Automatic document classifi-

cation is useful for a wide range of applications such as webpage classification, email spam filtering, news article classification, and many more.

The document classification task can be divided into three different sorts: supervised document classification where documents have predefined categories, unsupervised document classification (i.e., document clustering) where the class labels of documents are unknown, and semi-supervised document classification where only parts of the documents are labeled [58]. Pattern recognition and machine learning have been applied to document classification. A number of classification algorithms, such as naive Bayes classifier, k-nearest neighbor (kNN), decision trees, neural networks and Support Vector Machines (SVM), have been used to classify documents [1, 12, 50, 58, 71, 72, 87]. The technique details of these classifiers are explained as follows.

4.1.1 Rule-Based Techniques

In the earliest approaches to text categorization, examples are classified by handcrafted rules constructed from field experts [43, 76, 104]. Rule-based techniques to text classification problems are appealing due to their understandability. Later, rules are automatically generated from the training data to guide the classification task. Rule-based approaches were widely used in early work with acceptable classification accuracies.

Suppose we are trying to classify a document x into a set of predefined categories $\{C_1, C_2, ..., C_m\}$. The general form of a classification rule can be written as: $Des(x) \longrightarrow C_i$, where the left-hand side is a set of preconditions (e.g., frequencies of

a set of keywords in x), and the right-hand side is a decision which gives the class of documents matching the preconditions. The preconditions may be in a number of different forms. Boolean matching may be used, in which the existence of certain keywords in texts forms the preconditions. Using the frequency of keywords to form preconditions is another common method to interpret classification rules.

Rules can either be directly extracted from data, with for example, RIPPER [11], CN2 [10], and 1R [42] algorithms, or indirectly extracted from other classification models, such as decision trees [72] and neural networks [87] algorithms, etc. Rule-based classifiers attempt to produce mutually exclusive rules, that is, rules that are independent of each other where every document is covered by at most one rule. Some rule learning methods use exhaustive coverage, that is, every possible combination of attribute values are considered and each example is covered by at least one rule.

A rule covers a document x if the keywords in x satisfy the precondition of the rule. To resolve inconsistencies, all rules may be applied to ensure that any rules that match this document agree in classification. Disagreements cause the document to be deemed undecidable, in which case the document is assigned to a special undecided class, or presented to the user for a more educated opinion [13,44,133]. Rules can also be rank ordered according to their priority. They can either be ranked based on their quality (rule-based ordering) or grouped together based on their classes (class-based ordering) [100,128]. The highest ranked rule is used to classify the document. If none of the rules are triggered, the document will be assigned to the default class.

Rule-based approach is highly expressive, but rules are easily readable and understood by humans who could possibly change them for fine-tuning or improve-

ment. Rules can classify documents rapidly with performance comparable to decision trees [100]. The requirements of rule-based techniques are very low in comparison to those of more complex methods. However, rule-based techniques suffer from high computational complexity. Many dimensionality reduction methods have been used to aid rule induction [3, 85].

4.1.2 Statistically Based Techniques

The rule-based approach was appropriate when few machine-readable texts were available and computational power was expensive. Recent approaches apply machine learning algorithms to automatically classify documents into different categories. With these approaches, a set of decision criteria is learned automatically from the training data. If a learning method is statistically based, it is called a statistically based approach. In statistical text classification, the differences between documents are expressed statistically as the likelihood of certain events, rather than some manually generated rules.

A statistically based approach employs probability theory to text classification task. The selected probabilistic classifier computes the conditional probability $Pr(C_i \mid x)$, where $1 \leq i \leq m$. The document is assigned to the category with the highest conditional probability. In other words, probabilistic classifiers are trained to minimize the probability of error whenever a decision is made. In the case of binary classifications, suppose we only have two decision classes (i.e., C_1 and C_2), for a given

document x, the probability of error is calculated as:

$$Pr(\text{error} \mid x) = \begin{cases} Pr(C_1 \mid x) & \text{if we decide } C_2 \\ Pr(C_2 \mid x) & \text{if we decide } C_1. \end{cases}$$

We can minimize the probability of error by deciding C_1 if $Pr(C_1 \mid x) \geq Pr(C_2 \mid x)$, and C_2 if $Pr(C_2 \mid x) \geq Pr(C_1 \mid x)$.

The main advantage of the statistically based approach is efficiency, since calculating probabilities is relatively an easy task for computers. On the other hand, its main drawback is the lack of intuitiveness. It is much easier to understand the semantics of a set of classification rules than large tables of conditional probabilities. Especially since the intuitive, popular understanding of probability involves equiprobable randomness. Nevertheless, statistically based methods are acquiring wide interests as a feasible solution to efficiency problems in text categorization tasks.

4.1.3 Other Techniques

Other classification methods, such as Support Vector Machine (SVM) and vector space model (VSM) are based on linear algebra. In SVM model, we look for a decision surface that is maximally far away from any data point. This distance from the decision surface to the closest data point determines the margin of the classifier.The VSM is another popular text categorization technique [60, 80, 103], in which each document is represented by a term (i.e., keyword) vector. A document is assigned to the most likely category based on vector similarities (e.g., the cosine formula) since documents with similar contents have similar vectors [80, 103]. The dimensionality of the feature space is a main issue for vector-based models, especially for long documents. Many

algorithms are not applicable due to this reason and requires very large amounts of storage space.

Neural network [44] was proposed as an alternative to various conventional classification methods. The basic idea of neural networks is inspired by the structure of biological neural networks, which consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. One can think a neural network as an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Neural networks are non-linear models, which makes them flexible in modeling real world complex relationships.

Decision trees can be seen as indirect methods of rule-based classification. k-nearest neighbor (kNN) uses k closest points (nearest neighbors) for performing classification. k-nearest neighbors of an object x are data points that have the k smallest distance to x. One can use different distance metric to compute distance between objects. Unlike decision tree induction and rule-based systems, kNN classifier does not build models explicitly. Therefore, classifying unknown objects are relatively expensive when using kNN.

4.2 Workflow of a Spam Filtering System

Over the years, various anti-spam technologies and softwares have been developed. One popular approach is to treat spam filtering as a classification problem [2, 79]. Many classification algorithms from machine learning can be applied to automatically

classify incoming emails into different categories based on the contents of emails [14, 29, 55, 58, 61, 79, 82]. Among these algorithms, the naive Bayes classifier has received much attention and served as a base for many open source projects and commercial products, due to its simplicity, computational efficiency and good performance [5, 7, 28, 75, 77, 111]. The naive Bayes classifier, along with many other classification algorithms, typically treat spam filtering as a binary classification problem, that is, the incoming email is either spam or non-spam. In reality, this simple treatment is too restrictive and could result in losing vital information by misclassifying a legitimate email to spam. For example, a user could miss an important job offer just because the email contains "congratul" (i.e., a common word in email spam filter word list) in its header. On the other hand, misclassifying a spam email to non-spam also brings unnecessary costs and waste of resources.

The workflow of an email spam filtering system is based on the bag of words model [110]. Several pre-processing steps are required before information can be used by a filter. The main steps involved in a spam filter are illustrated in Figure 4.1 [38].

At the first step, the whole text of every email need to be scanned to obtain words, phrases or meta-data as tokens. The set of tokens is then represented by a format (e.g., real values) required by the machine learning algorithm used. These values indicate a number of things such as the presence of a token or the frequency with which a given token occurs.

At the second step, words common to both spam and non-spam classes will be removed. For example, words, such as "to," "will" and "be," provide very little information as to the class of the email message. At the same time, a step called

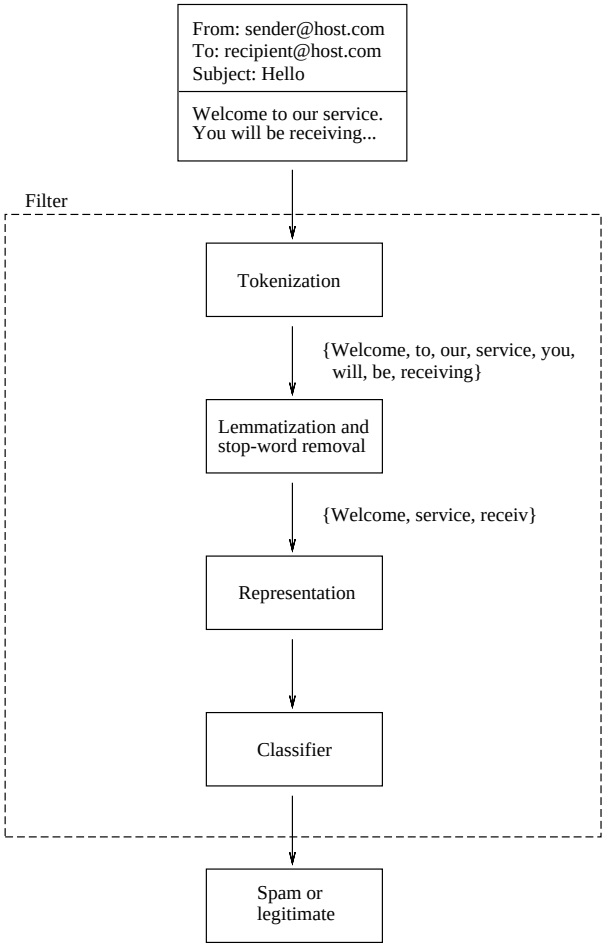

Figure 4.1: An illustration of some of the main steps involved in a spam filter

lemmatization (stemming) reduces words to their root forms. For example, "receiving" and "received" will be treated as "receiv" rather than two distinct words.

At the third step, feature selection is often performed to reduce the size of the set of selected tokens to ensure the performance of algorithms. There are two commonly used feature selection methods in email spam filtering. Document Frequency (DF) is the number of emails in which a token occurs. Information Gain (IG) measures

the number of bits of information obtained for category prediction by knowing the presence or absence of a token in an email [80]. It is assumed that rare tokens are non-informative for category prediction. Tokens whose DF or IG are less than some predetermined thresholds will be removed.

At the last step, after a set of emails have been scanned, they can be represented in a data table (i.e., the training set). We have a set of tokens as attributes, one of these attributes is class (e.g., legitimate or spam). Each email is a record of the collection. The attribute values are frequencies of tokens. A learning algorithm can be performed on the training set, so we can find a model (e.g., a decision tree or a set of rules) to measure the hamminess or spamminess of each email. Finally when a new email comes in, we can apply the model and predict its class.

The last step is a typical classification process in data mining and it is the part that I want to improve on.

4.3 Related Works on Ternary Email Spam Filtering

In a binary spam filtering system, two email folders are generated that contain spam and non-spam email messages, respectively. Ideas of ternary email spam filtering has been discussed in some previous literatures. Robinson [77] suggested to add a boundary region marked unsure to the classification results. Yin et al. [112] suggested to call these messages that could reasonably be considered either spam or good as gray mail, and proposed four prototype methods for detecting them. Zhao et al. [136] proposed

a classification schema based on rough set theory to classify the incoming emails into three categories: spam, non-spam, and suspicious. Siersdorfer et al. [88] introduced a framework of using restrictive methods and ensemble-based meta methods for junk elimination. In their approaches, classifiers for a given topic make a ternary decision on a document: they can accept the document for the topic, reject it for the topic, or abstain if there is insufficient evidence for acceptance or for rejection. Zhou and Yao [142] proposed a three-way decision approach based on Bayesian decision theory. Ternary classifications were also enabled in some anti-spam applications, such as SpamBayes [97], Bogofilter [7], and SpamAssassin [96].

The main advantage of the ternary email spam filtering is that it allows the possibility of non-commitment, i.e., of refusing to make a decision in close cases. This is a useful option if the cost of being indecisive is not too high. The undecided cases must be re-examined by collecting additional information, thereby classify emails with fewer errors.

Suppose each incoming email x is represented by a feature vector $Des(x) = (x_1, x_2, ..., x_n)$, where $x_1, x_2, ..., x_n$ represent the occurrence of tokens in an email message. Let C denote the legitimate class, and C^c denote the spam class. Based on the description of email, we can use a discriminant function $f(x)$ for classification. For example, in naive Bayes classifier, $f(x)$ is the a posteriori probability or its monotonic transformations (e.g., the a posteriori odds). In Support Vector Machine (SVM) method, $f(x)$ is the distance between the given email and the decision hyperplane. There are typically two ways to use a discriminant function for classification. One is to use the discriminant function to rank emails and let a user read through the ranked

list. The other is to classify emails into several categories based on some thresholds on $f(x)$. Three existing approaches on ternary email spam filtering are reviewed as follows.

Robinson's Approach

Given an email x described by its feature vector $Des(x) = (x_1, x_2, ..., x_n)$. Robinson [77] calculates the spamminess of each selected token based on its occurrence x_i and then combines these individual probabilities into an overall probability for x. The per-word probability $Pr(C^c \mid x_i)$ is calculated as follows,

$$Pr(C^c \mid x_i) = \frac{(s \times p) + (m_{x_i} \times \frac{Pr(x_i|C^c)}{Pr(x_i|C^c)+Pr(x_i|C)})}{s + m_{x_i}}, \tag{4.1}$$

where s is the strength we want to give to our background information, p is our assumed probability, based on our general background information, that a word we do not have any other experience of will first appear in a spam, and m_{x_i} is the number of emails we have received that contain x_i. In practice, the values for s and p are found through testing to optimize performance. Reasonable starting points are 1 and 0.5 for s and p, respectively.

The per-word probabilities are combined into an overall probability based on Fisher's inverse chi-square procedure [101]. The combined probability H is calculated as follows,

$$H = Chi^{-1}(-2 \ln \prod_{i=1}^{n} Pr(C^c \mid x_i), 2n), \tag{4.2}$$

where $Chi^{-1}()$ is the inverse chi-square function, used to derive a probability from a chi-square-distributed random variable. H indicates the hamminess of an given email. The combined probability S that represents the spamminess of the email is

also calculated:

$$S = Chi^{-1}(-2\ln\prod_{i=1}^{n}(1 - Pr(C^c \mid x_i)), 2n). \tag{4.3}$$

The final probability is:

$$I = \frac{1 + H - S}{2}. \tag{4.4}$$

A given email is classified as spam if the value of I is near 1, is classified as legitimate if I is near 0, and is classified as uncertain when I is near 0.5.

Robinson's approach does not need naive independence assumption when combining the individual probabilities, but it is based on an assumption that a randomly chosen e-mail containing x_i would be spam in a world where half the e-mails were spam and half were ham. Whereas in reality, legitimate and spam emails are usually not equally distributed. In addition, the selection of the thresholds 1, 0.5 and 0 is based on intuition with little analysis.

Rough Set Approach

Zhao et al. [136] introduced a rough set approach to automatically learning rules to classify emails into three categories: spam, non-spam and suspicious. A set of decision rules is induced by a genetic algorithm intergraded in the rough set tool kit, Rosetta [78]. The strength of each rule is evaluated by its accuracy. For an email x, suppose a rule r that matches x is given in the form $Des(x) \rightarrow C$, where the left-hand side is the conjunction of features that describes x (or equivalently the equivalence class $[x]$), and the right-hand side is the non-spam class label C. Let $RUL(x)$ denote the set of these types of rules with non-spam as consequent that matches x. The

certainty of x being in the non-spam class is measured as follow,

$$Certainty_x = \frac{\sum_{r \in RUL(x)} accuracy(r)}{\mid RUL(x) \mid}, \tag{4.5}$$

where $\mid \cdot \mid$ denote the cardinality of a set, and the accuracy of r is calculated by $accuracy(r) = \frac{|C \cap [x]|}{|[x]|}$. A pair of thresholds α and β is used to compare with $Certainty_x$ to define the three email categories. An email x is classified into the non-spam category if $Certainty_x \geq \alpha$, is classified into the suspicious category if $\beta \leq Certainty_x < \alpha$. Otherwise, it will be classified into the spam category.

There are a few problems with Zhao et al.'s approach to spam filtering. First, when more than one rule matches a given email x, features on the left-hand side of rules in $RUL(x)$ may have overlaps. Simply adding up the accuracy of each rule may cause repeated consideration of some features which will lead to biased classification results. Second, accuracy is the only measure used to evaluate the strength of a rule. It may not be able to provide reliable indications for some data sets. For example, suppose we have a data set that contains important evidences of emails being spam, but it may not contain all evidences of emails being legitimate. Even if we have equally distributed positive and negative examples, it does not mean that the probability of an email being spam is 50%. In other words, using $|[x]|$ as a denominator to measure the accuracy of a rule may mislead the classification results. Other forms of rule evaluations should also be considered. Third, the two required thresholds are arbitrarily defined with one simple constraint, that is, $\alpha \in (\frac{1}{2}, 1]$ and $\beta = 1 - \alpha$. Finally, the time to learn accurate rule sets from data is higher than the statistical methods.

The Ensemble Methods

Siersdorfer et al. [88] proposed a restrictive meta method for eliminating junk documents, in which junk documents are defined as the class of documents that does not appear in the training set, but appears in the testing set. In their approach, restrictions are first made to a set of binary classifiers, these classifiers are then ensembled to make a ternary decision on a newly seen document: they can accept the document for the topic, reject it for the topic, or abstain if there is neither sufficient evidence for acceptance nor sufficient evidence for rejection.

The method they used to extend binary classifiers is similar to the ideas of ternary email spam filtering. That is, each classifier is required to accept or reject documents if their values above some threshold, and abstain otherwise. More specifically, given a set $V = \{v_1, v_2, ..., v_k\}$ of k binary classifiers with results $R(v_j, x)$ in $\{+1, -1, 0\}$ for a document x. The value of $R(v_j, x)$ is $+1$ if x is accepted for the given topic by v_j, -1 if x is rejected, and 0 if v_j abstains. These results are combined into a meta result that makes a unanimous decision as follows:

$$Meta(x) = \begin{cases} +1 & \text{if } \sum_{j=1}^{k} R(v_j, x) \cdot weight(v_j) > \alpha, \\ -1 & \text{if } \sum_{j=1}^{k} R(v_j, x) \cdot weight(v_j) < \beta, \\ 0 & \text{otherwise,} \end{cases}$$

where α and β are two thresholds with $\alpha > \beta$, and $weight(v_j)$ is a weight factor for each v_j. If all classifiers give the same result (either $+1$ or -1), $Meta(x)$ returns this result, and returns 0 otherwise. The author did not provide information on how to acquire the required thresholds.

Siersdorfer's approach can be applied to three-way email spam filtering. The main difference is that in three-way email spam filtering, there are only positive and

negative examples in the testing set, but we make a ternary classification on a binary scale. Thus, Siersdorfer's approach returns a 3×3 contingency table that contains the classification results, whereas the ternary classification methods returns a 2×3 contingency table, which represents different types of misclassification errors and costs. Similar ideas of the ensemble method can be found in [112], where multiple email filters are performed using different disjoint subsets of the training data. An email is classified based upon the level of disagreements between these filters. However, both of these approaches did not provide evidence to show that the ensembled method is better than applying a single filter. Siersdorfer's experimental results showed that for some datasets, the meta classifier outperformed the single classifier, but for some other datesets, the single classifier performed better.

4.4 Comments on Existing Approaches

There are two key issues in the existing studies on ternary email spam filtering. The first issue is the estimation of the hamminess or spamminess of an email. For content-based statistical filtering, a model generally produces a real-valued number that indicates the hamminess or spamminess of an email. For instance, the naive Bayes classifier uses a posteriori probability to represent the possibility of an email being legitimate given its feature descriptions. For rule-based approach, this estimation is done by evaluating strength of the matching rules through some quantitative measures, such as accuracy and coverage. The second issue is the interpretation and computation of required thresholds to define the three email categories. The existing

ternary email spam filtering methods choose the thresholds fairly arbitrarily based on an intuitive understanding of the levels of tolerance for errors. For example, SpamBayes uses 0.9 to determine between spam and unsure, while 0.2 separates unsure and legitimate class. Bogofilter uses 0.99-0.45 for the unsure range.

The existing ternary email spam filtering methods focus on the first issue, that is, working towards functions that give better estimations of hamminess or spamminess of emails. After testing on the statistically based and rule-based filtering, Graham [31] pointed out that although the rule-based approach is easy to begin with, but it gets very hard to catch the last few percentage of spam, the statistically based filtering is a better way to stop spam. On the other hand, estimations of required thresholds have not received much attention. Little analysis has been done to determine the optimal thresholds. There is a need for inferring these thresholds from a theoretical and practical basis. Moreover, email spam filtering is a typical cost-sensitive task [22, 145, 146]. Misclassifying a legitimate email to spam is usually considered more costly than misclassifying a spam to legitimate. Such characteristics have not been explicitly reflected and made clear in the existing ternary email spam filtering methods.

4.5 A New Formulation

In this section, after reviewing the basic formulations of the commonly used binary naive Bayesian spam filtering process, I propose a cost-sensitive three-way email spam filter based on the unified framework introduced in Chapter 3 to address the above mentioned issues.

4.5.1 Binary Naive Bayesian Spam Filtering

The naive Bayesian spam filtering is a probabilistic classification technique of email filtering [61, 79]. It is based on Bayes' theorem with naive (strong) independence assumptions [29, 58].

Given an email x described by its feature vector $Des(x) = (x_1, x_2, ..., x_n)$. The a posteriori odds can be computed as:

$$\frac{Pr(C \mid [x])}{Pr(C^c \mid [x])} = \prod_{i=1}^{n} \frac{Pr(x_i \mid C)}{Pr(x_i \mid C^c)} \frac{Pr(C)}{Pr(C^c)}. \tag{4.6}$$

The detailed computation of likelihood function based on naive independence assumption can be found in equation (3.30) from Section 3.4.2. Thus, an incoming email will be classified as C (i.e., legitimate) if $\frac{Pr(C|[x])}{Pr(C^c|[x])}$ (i.e., the a posteriori odds) exceeds a threshold, otherwise it is spam.

Suppose that an email x is being classified as C_i when its true class label is C_j, a classification error or cost λ_{ij} will be incurred. Since $Pr(C_j \mid [x])$ is the probability that the true class label is C_j, the expected cost or conditional risk of deciding class C_i is defined as:

$$R(C_i \mid d) = \sum_{j=1}^{m} \lambda_{ij} Pr(C_j \mid [x]). \tag{4.7}$$

The document is assigned to the category with the minimum expected cost or conditional risk. For binary classifications, we have two possible conditional risk:

$$\begin{aligned} R(C_1 \mid [x]) &= \lambda_{11} Pr(C_1 \mid [x]) + \lambda_{12} Pr(C_2 \mid [x]), \\ R(C_2 \mid [x]) &= \lambda_{21} Pr(C_1 \mid [x]) + \lambda_{22} Pr(C_2 \mid [x]), \end{aligned} \tag{4.8}$$

where λ_{12} denote the cost incurred for deciding C_1 when the true class label is C_2, and λ_{21} denote the cost incurred for deciding C_2 when the true class label is C_1. Thus, to minimize the risk of misclassification, the following rule must hold: decide C_1 if $R(C_1 \mid [x]) < R(C_2 \mid [x])$, that is,

$$(\lambda_{21} - \lambda_{11})Pr(C_1 \mid [x]) > (\lambda_{12} - \lambda_{22})Pr(C_2 \mid [x]). \tag{4.9}$$

Generally speaking, the cost incurred for making an error is greater than the cost incurred for being correct, and both of the factors $\lambda_{21} - \lambda_{11}$ and $\lambda_{12} - \lambda_{22}$ are positive. Thus, the classification decision is determined by the more likely class label, although the a posteriori probabilities must be scaled by the appropriated cost differences. By the definitions of the Bayes' theorem (see equation (3.25)), we can replace the a posteriori probabilities by the a priori probabilities and the likelihoods. Equation (4.9) can be rewritten as:

$$(\lambda_{21} - \lambda_{11})Pr([x] \mid C_1)Pr(C_1) > (\lambda_{12} - \lambda_{22})Pr([x] \mid C_2)Pr(C_2). \tag{4.10}$$

Under the reasonable assumption that $\lambda_{21} > \lambda_{11}$, equation (4.10) can be rewritten as:

$$\frac{Pr([x] \mid C_1)}{Pr([x] \mid C_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{Pr(C_2)}{Pr(C_1)}, \tag{4.11}$$

where $\frac{Pr([x]|C_1)}{Pr([x]|C_2)}$ is the likelihood ratio. Thus, the classification decision can be made of deciding C_1 if the likelihood ratio exceeds a threshold.

For simplicity, a special kind of classification error or cost, called a symmetrical or zero-one loss function, is considered in many existing spam filters. It is defined as:

$$\lambda_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

where $i, j = 1, \ldots, m$. This loss function assigns no loss to a correct decision, and assigns a unit loss to any error. Thus, all errors are equally costly. The risk corresponding to this loss function is the average probability of error, the conditional risk can be rewritten as:

$$R(C_i \mid [x]) = \sum_{j=1}^{m} \lambda_{ij} Pr(C_j \mid [x]) = \sum_{i \neq j} Pr(C_j \mid [x]) = 1 - Pr(C_i \mid [x]). \quad (4.12)$$

To minimize the average probability of error, we should select the C_i that maximizes the a posteriori probability $Pr(C_i \mid [x])$. In other words, for minimum error rate: decide C_i if $Pr(C_i \mid [x]) > Pr(C_j \mid [x])$ for all $j \neq i$.

4.5.2 Cost-Sensitive Three-Way Spam Filtering

In medical three-way decision making, a doctor needs to decide whether to treat the patient, not treat the patient, or further test on the patient. Similar ideas can be used for email spam filtering, where we can accept an email, reject an email, or put it into a suspected folder for further examination.

Based on notations in Section 3.2, there are two states (classes) regarding an incoming email: C (P) denoting a Legitimate email and C^c (N) denoting a Spam. There are three actions: a_P for accepting the email, a_B for making a deferred decision (i.e., neither accept nor reject the email due to insufficient information), and a_N for rejecting the email.

As shown in Figure 4.2, three email folders are produced instead of two, the *Inbox* folder contains accepted emails, the *Spam* folder contains rejected emails, and the *Suspected* folder contains suspicious emails that need to be further examined. This

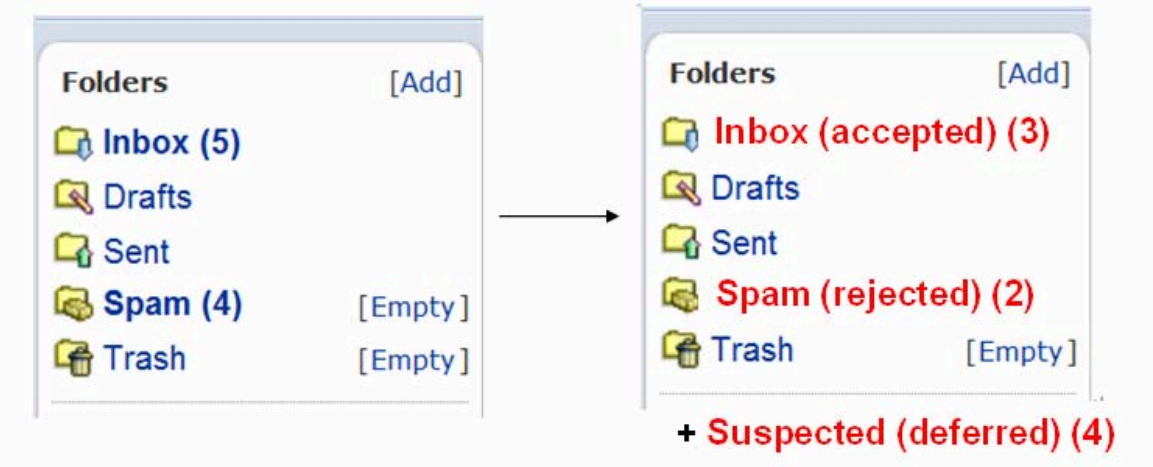

Figure 4.2: From binary spam filtering to ternary spam filtering

is a useful option when users view their emails under a time constraint. They may view the *Inbox* folder immediately, delete the *Spam* folder without viewing, and delay the processing of *Suspected* folders. The emails in the suspected folder can be ranked based on their probability of being spam or legitimate to help the user make decisions later on.

A loss function is interpreted as the costs of taking the corresponding actions. Generally speaking, a higher cost occurs when misclassifying a legitimate email as a spam; it could result in losing vital information for a user. On the other hand, misclassifying a spam to be a legitimate email brings unnecessary costs of processing the spam. Costs also occur when a delayed decision is made. The costs depend very much on a particular user's subjective evaluation about various actions and the tolerance of different types of errors. Different users may give different values depending on how critical it is for them to process a spam in the *Inbox* folder, to

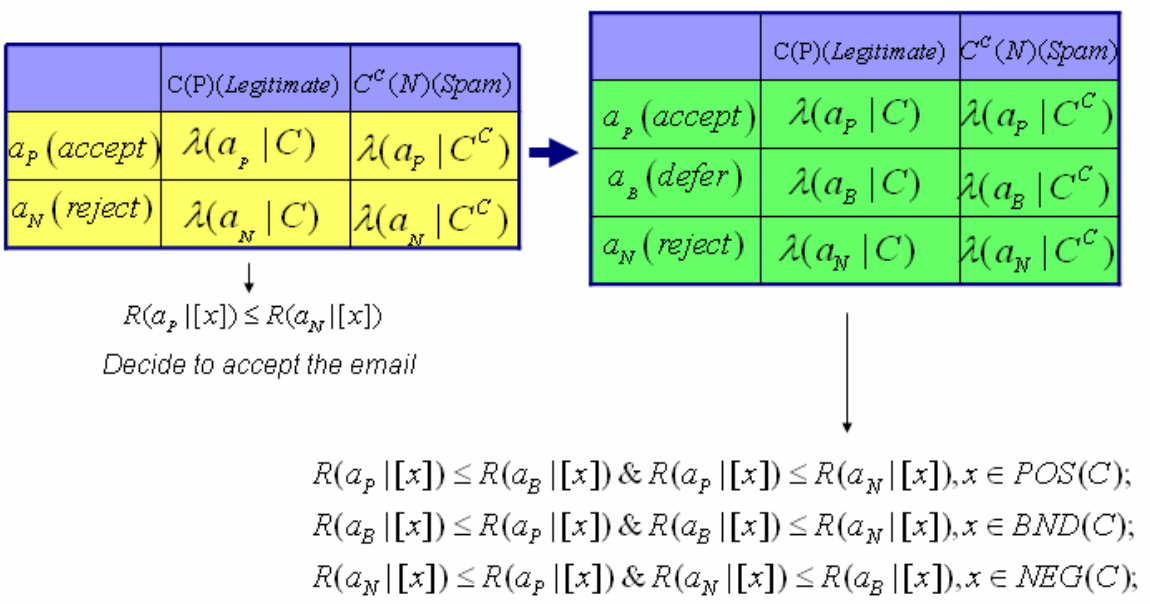

Figure 4.3: From binary cost matrix to ternary cost matrix

delete the *Spam* folder, and to delay processing of the *Suspected* folder.

Similar to the cost matrix given in Chapter 3 for the medical example, a similar cost matrix can be given for the email spam filtering. As shown in Figure 4.3, there are four types of loss functions in the binary cost matrix. According to the minimum risk decision rules in Bayesian decision theory, we accept an email if the expected risk is smaller than rejecting the email, that is, $R(a_P|[x]) \leq R(a_N|[x])$. In the cost matrix for ternary classification, there are six types of loss functions. In order to make a decision, we need to do a pairwise comparison of the three expected risks.

As explained in Section 3.2, the estimation of the required thresholds α and β is illustrated in Figure 4.4. It maybe reasonable to impose the constraint $\alpha > \beta$ so that the boundary region may be non-empty, and the loss incurred for deferment should

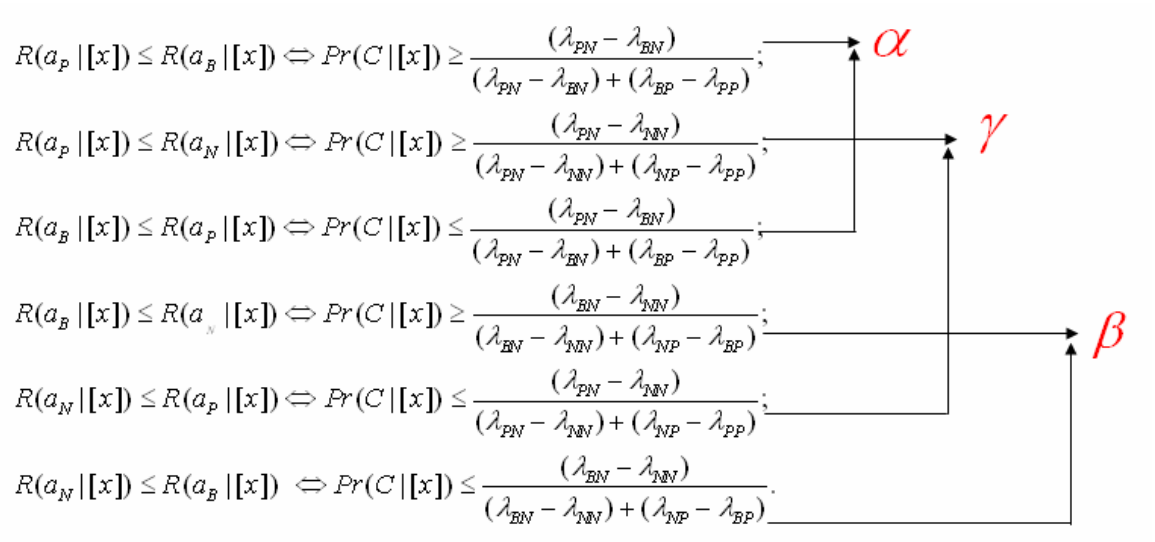

Figure 4.4: Estimating thresholds

be in between of accept and reject, therefore, we get $1 \geq \alpha > \gamma > \beta \geq 0$. After tie-breaking, γ is no longer needed.

In short, a three-way decision can be made based on the pairwise comparison of the three expected risks: $R(a_P|[x])$, $R(a_B|[x])$ and $R(a_N|[x])$. Each of these comparisons can be translated into the comparison between the a posteriori probability $Pr(C \mid [x])$ and a pair of thresholds α and β, where $Pr(C \mid [x])$ can be calculated based on Bayesian inference as shown in Section 3.4, and (α, β) can be calculated based on the loss functions as shown in Section 3.2.

Example 4 *Table 4.1 and Table 4.2 give the loss functions of two users, User 1 and User 2, respectively. It can be seen that User 1 is more concerned about losing a legitimate email and at the same time about processing a spam. In comparison, User 2 is not so much concerned.*

Table 4.1: Loss function of User 1

	C (P)(Legitimate)	C^c (N)(Spam)
a_P(Accept)	$\lambda^1_{PP} = 0$	$\lambda^1_{PN} = 10$
a_B(Defer)	$\lambda^1_{BP} = 5$	$\lambda^1_{BN} = 5$
a_N(Reject)	$\lambda^1_{NP} = 90$	$\lambda^1_{NN} = 0$

Table 4.2: Loss function of User 2

	C (P)(Legitimate)	C^c (N)(Spam)
a_P(Accept)	$\lambda^2_{PP} = 0$	$\lambda^2_{PN} = 8$
a_B(Defer)	$\lambda^2_{BP} = 5$	$\lambda^2_{BN} = 5$
a_N(Reject)	$\lambda^2_{NP} = 15$	$\lambda^2_{NN} = 0$

The pair of thresholds α^1 and β^1 for User 1 is calculated according to equation (3.6) as:

$$\begin{aligned}\alpha^1 &= \frac{(\lambda^1_{PN} - \lambda^1_{BN})}{(\lambda^1_{PN} - \lambda^1_{BN}) + (\lambda^1_{BP} - \lambda^1_{PP})} = \frac{10-5}{(10-5)+(5-0)} = 0.50,\\ \beta^1 &= \frac{(\lambda^1_{BN} - \lambda^1_{NN})}{(\lambda^1_{BN} - \lambda^1_{NN}) + (\lambda^1_{NP} - \lambda^1_{BP})} = \frac{5-0}{(5-0)+(90-5)} = 0.06.\end{aligned}$$

The pair of thresholds α^2 and β^2 for User 2 is calculated as:

$$\begin{aligned}\alpha^2 &= \frac{(\lambda^2_{PN} - \lambda^2_{BN})}{(\lambda^2_{PN} - \lambda^2_{BN}) + (\lambda^2_{BP} - \lambda^2_{PP})} = \frac{8-5}{(8-5)+(5-0)} = 0.38,\\ \beta^2 &= \frac{(\lambda^2_{BN} - \lambda^2_{NN})}{(\lambda^2_{BN} - \lambda^2_{NN}) + (\lambda^2_{NP} - \lambda^2_{BP})} = \frac{5-0}{(5-0)+(15-5)} = 0.33.\end{aligned}$$

It follows that $\beta^1 < \beta^2 < \alpha^2 < \alpha^1$. As expected, the thresholds of User 2 are

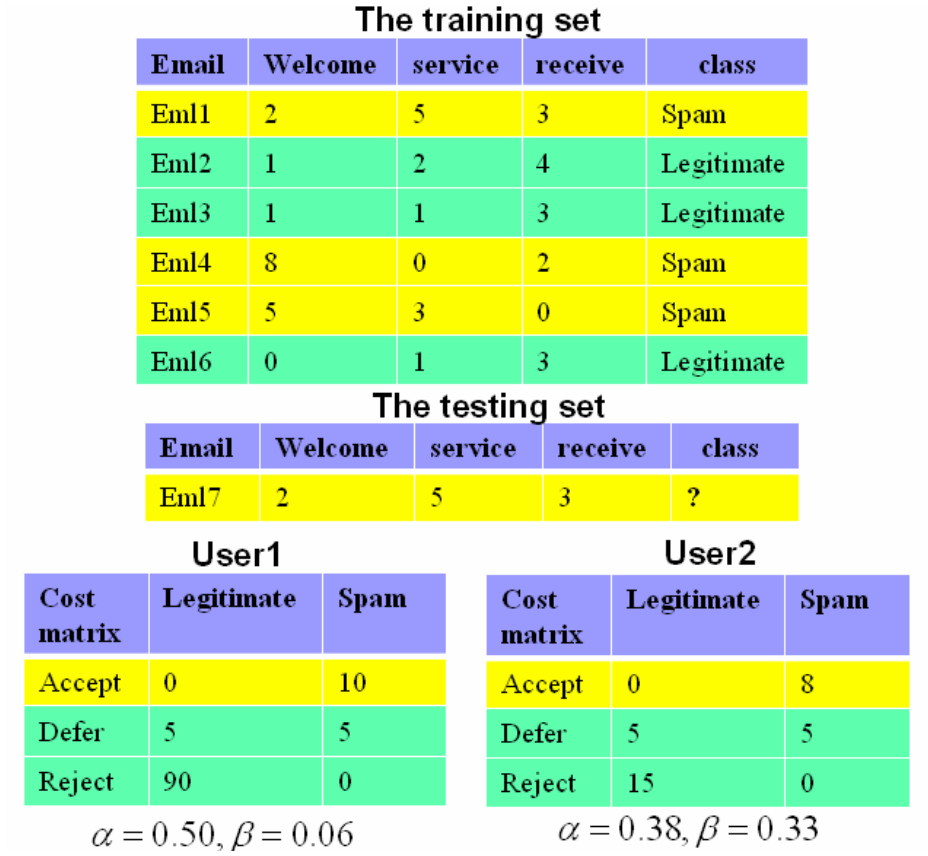

Figure 4.5: An example

within the thresholds of User 1, which shows that User 1 is much critical than User 2 regarding both incorrect acceptance and rejection. Consequently, User 1 would have smaller accepted and rejected folders but a large deferred folder. In contract, User 2 would have larger accepted and rejected folders but a smaller deferred folder.

For a new email $Eml7$, suppose its probability of being legitimate can be calculated from the training set as shown in Figure 4.5, that is, $Pr(Legitimate|[x]) = 0.3$. For User 1, $Eml7$ will be classified into the deferment folder for further examination because $0.06 < Pr(Legitimate|[x]) < 0.50$, but for User 2, $Eml7$ will be rejected because

$Pr(Legitimate|[x]) \geq 0.38$. Different filtering options are tailored to meet individual requirements in terms of minimum overall cost based on our unified framework.

4.6 Summary

In this chapter, I demonstrate the usefulness of the unified framework introduced in Chapter 3 through a real world application, namely, email spam filtering. After reviewing the existing work on ternary email spam filtering, I point out two basic issues, that is, the computation and interpretation of the required thresholds and the cost-sensitive characteristic. A cost-sensitive three-way email spam filtering architecture is proposed to address the above issues.

Chapter 5

EXPERIMENTS AND EVALUATIONS

In this chapter, I compare and evaluate the performance of proposed approach with traditional naive Bayesian spam filter and two existing ternary spam filters. Three benchmark corpora are used in the experiments. The results are evaluated in both cost-sensitive and non-cost-sensitive settings.

5.1 Dataset Preparations

The experiments were performed on three different datasets, the PU1 corpus [2], the Ling-Spam corpus [2], and a spambase data set from UCI Machine Learning Repository [56]. My goal is to compare the cost-sensitive three-way decision approach and the original naive Bayesian spam filter, Robinson's approach and the rough set approach on the three different datasets.

For the PU1 corpus, I selected the emails under the *bare* folder as the dataset, there were 1099 emails, 481 were spam, 618 were legitimate. The corpus was divided into 10 parts, and 10-fold cross validation was used with one part reserved for testing set at each repetition. After scanning the training set, I removed the tokens that appear less than 5% and more than 95% in all the emails. Information gain was used as the feature selection method and the top 300 attributes were selected as the feature set. For the Ling-Spam corpus, I selected the emails under the *lemm_stop* folder as the dataset, there were 2,412 legitimate emails from a linguistic mailing list and 481 spam ones. Again, 10-fold cross validation was used and the same feature selection strategy was used, the top 150 attributes were selected as the feature set based on their information gain. The UCI spambase data set consisted of 4601 instances, with 1813 instances as spam, and 2788 instances as legitimate, each instance was described by 58 attributes. I split the data set into a training set of 3681 instances, and a testing set of 920 instances. Entropy-MDL [23] is used as the discretization method applied to both the training and testing data sets. The best-first search is used for attribute selection and 15 attributes are selected. For the rough set approach, the set of decision rules is induced by using the genetic algorithm in the rough set tool kit, Rosetta [78].

5.2 Evaluation Measures

I use both the traditional non-cost-sensitive evaluation methods [27, 130, 138] and the cost-sensitive evaluation methods to compare the performances of the new approach

and other existing works.

Spam recall and *spam precision* have been used as indicators for measuring email spam filter performance [2, 79, 140]. They are defined as:

$$spam\ recall = \frac{n_{S\to S}}{n_{S\to S} + n_{S\to L}},$$

$$spam\ precision = \frac{n_{S\to S}}{n_{S\to S} + n_{L\to S}},$$

where $n_{S\to S}$ denotes the number of emails classified as spam which truly are, $n_{L\to S}$ denotes the number of legitimate emails classified as spam, and $n_{S\to L}$ denotes the number of spam emails classified as legitimate. Since misclassifying an legitimate email to spam is more costly than misclassifying a spam to legitimate, I consider *spam precision* as the main indicator for the non-cost-sensitive evaluation, the ternary email spam filtering should increase the *spam precision* in comparison with the original binary naive Bayes classifier (i.e., $n_{L\to S}$ decreases). The misclassification error discussed in previous chapters will be used in evaluations considering cost.

For cost-sensitive evaluations, I assume that misclassifying a legitimate email as spam is w times more costly than misclassifying a spam email as legitimate. Three different w values ($w = 1$, $w = 3$, and $w = 9$) were used for the original naive Bayesian spam filter. Three sets of loss functions for the three-way decision approach were provided accordingly with the same cost ratios. For instance, when we use $w = 9$ for the naive Bayesian spam filter, $\lambda_{NP}/\lambda_{PN} = 9$ was used in the three-way decision approach. Three pairs of thresholds (α, β) were calculated based on these three sets of loss functions, respectively, to distinguish the three email categories in ternary email spam filtering. The *weighted accuracy*, *weighted error rate* and *total cost ratio*

suggested by Androutsopoulos et al. [2], the *cost* measure suggested by Yao [122], and the *cost curve* suggested by Drummond et al. [18, 19] were used as the cost-sensitive evaluation measures.

The *weighted accuracy* is defined as [2]:

$$WAcc = \frac{w \cdot n_{L\to L} + n_{S\to S}}{w \cdot N_L + N_S},$$

where $n_{L\to L}$ denotes the number of emails classified as legitimate which truly are, N_L and N_S are the number of legitimate and spam emails to be classified by the spam filter. Similarly, we can define the *weighted error rate* as follows [2]:

$$WErr = \frac{w \cdot n_{L\to S} + n_{S\to L}}{w \cdot N_L + N_S}.$$

It is important to compare the *weighted accuracy* and *weighted error rate* to a baseline approach to avoid misinterpreting the often high accuracy and low error rate scores. A baseline is defined as the case where all legitimate emails are never blocked, and all the spam emails always pass the filter. The *weighted error rate* of the baseline is defined as [2]:

$$WErr^b = \frac{N_S}{w \cdot N_L + N_S}.$$

The *total cost ratio* is defined as [2]:

$$TCR = \frac{WErr^b}{WErr} = \frac{N_S}{w \cdot n_{L\to S} + n_{S\to L}}.$$

Greater TCR indicates better performance. If cost is proportional to wasted time, TCR measures how much time is wasted to delete manually all spam emails when no filter is used, compared to the time wasted to delete manually any spam emails

that pass the filter plus the time needed to recover from mistakenly blocked legitimate emails. The ternary email spam filtering introduce two additional types of misclassification errors besides the original incorrect acceptance and incorrect rejection, namely, deferment of positive and deferment of negative. In other words, the misclassification rate is reduced by the deferment errors. Therefore, the ternary email spam filtering should decrease the *weighted error rate* (i.e., its numerator decreases) and increase the *TCR* (i.e., its denominator decreases).

Suppose the classification results of the ternary email spam filtering are represented by the following 3× 2 contingency table:

	C (P): positive	C^c (N): negative
a_P: accept	n^t_{PP}	n^t_{PN}
a_B: defer	n^t_{BP}	n^t_{BN}
a_N: reject	n^t_{NP}	n^t_{NN}

The *cost* measure for ternary classification models is defined as:

$$\begin{aligned} Cost^t &= \frac{1}{U}[(\lambda_{PP} n^t_{PP} + \lambda_{PN} n^t_{PN}) + (\lambda_{BP} n^t_{BP} + \lambda_{BN} n^t_{BN}) \\ &\quad +(\lambda_{NP} n^t_{NP} + \lambda_{NN} n^t_{NN})], \end{aligned}$$

where U denotes the number of examples in the testing set, λ_{PP}, λ_{PN}, λ_{BP}, λ_{BN}, λ_{NP} and λ_{NN} are the loss functions defined in Chapter 3. Assume the classification results of the binary models are represented by the following 2× 2 contingency table:

	C (P): positive	C^c (N): negative
a_P: accept	n^b_{PP}	n^b_{PN}
a_N: reject	n^b_{NP}	n^b_{NN}

The *cost* measure for binary models is defined as:

$$Cost^b = \frac{1}{U}[(\lambda_{PP} n^b_{PP} + \lambda_{PN} n^b_{PN}) + (\lambda_{NP} n^b_{NP} + \lambda_{NN} n^b_{NN})],$$

Note that n^b_{NP} and n^t_{NP} correspond to $n_{L\to S}$ in the *weighted error rate* and *TCR* measures, and n^b_{PN} and n^t_{PN} correspond to $n_{S\to L}$. We used the symbols from the original paper [122] for the consistency with the subscripts of the loss functions. As it has been proved by the author [122], the associated costs of a ternary classification model are always less than the binary model. The main advantage of the *cost* measure is that it takes deferment costs and errors into consideration, while the other measures only consider the acceptance and rejection errors.

Cost curve is an alternative to *ROC curve* in which the expected cost of a classifier is represented explicitly [18, 19]. The x-axis in a cost curve is the probability-cost function for positive examples, which is defined as:

$$PCF(C) = \frac{Pr(C)\lambda_{NP}}{Pr(C)\lambda_{NP} + Pr(C^c)\lambda_{PN}}.$$

The y-axis is the expected cost normalized with respect to the cost incurred when every example is incorrectly classified, which is defined as:

$$NE(\lambda) = (1 - TP - FP) * PCF(C) + FP,$$

where TP is the true positive rate, and FP is the false positive rate. If one classifier

Table 5.1: Comparison results on PU1 corpus

Thresholds	Approaches	cost	WAcc	WErr	TCR	Spam	
						precision	recall
	NB	2.73	81.81%	18.18%	2.25	80.49%	73.33%
$w = 1$	Three-way	2.07	74.55%	10.00%	4.09	89.66%	57.78 %
$\alpha = 0.75$	Robinson	3.72	0.90%	0.00%	0.00	100.00%	2.22%
$\beta = 0.25$	RS	3.99	50.91%	19.10%	2.14	58.62%	37.78%
	NB	3.82	88.33%	11.67%	1.61	85.37%	77.78%
$w = 3$	Three-way	2.97	83.75%	7.08%	2.65	90.91%	66.67%
$\alpha = 0.65$	Robinson	5.36	0.42%	0.00%	0.00	100.00%	2.22%
$\beta = 0.15$	RS	8.43	50.42%	19.58%	0.96	57.58%	42.22%
	NB	3.13	91.59%	8.41%	0.85	88.10%	82.22%
$w = 9$	Three-way	1.79	86.35%	5.40%	1.32	91.18%	68.89%
$\alpha = 0.55$	Robinson	2.18	0.16%	0.00%	0.00	100.00%	2.22%
$\beta = 0.05$	RS	7.54	39.21%	23.17%	0.31	57.89%	48.89%

is lower in expected cost across the whole range of the probability-cost function, it dominates the other.

5.3 Results and Analysis

Table 5.1 to Table 5.3 show the evaluation results on three different datasets. We can see that Robinson's approach has better performance in non-cost-sensitive evaluations,

Table 5.2: Comparison results on Ling-Spam corpus

Thresholds	Approaches	cost	WAcc	WErr	TCR	Spam	
						precision	recall
	NB	1.71	88.58%	11.42%	1.61	72.73%	60.38%
$w = 1$	Three-way	1.63	88.58%	10.73%	1.71	74.42%	60.38%
$\alpha = 0.75$	Robinson	3.54	6.57%	0.35%	53.00	95.00%	35.85%
$\beta = 0.25$	RS	1.99	83.39%	12.11%	1.51	82.35%	26.42%
	NB	2.80	92.90%	7.10%	0.98	74.42%	60.38%
$w = 3$	Three-way	2.53	92.51%	6.31%	1.10	74.42%	60.38%
$\alpha = 0.65$	Robinson	4.43	3.15%	0.79%	8.83	92.31%	45.28%
$\beta = 0.15$	RS	2.32	89.09%	4.47%	1.56	80.00%	30.19%
	NB	1.80	95.22%	4.78%	0.51	76.92%	56.60%
$w = 9$	Three-way	1.66	93.94%	4.27%	0.57	78.38%	54.72%
$\alpha = 0.55$	Robinson	2.28	1.29%	0.83%	2.94	93.33%	52.83%
$\beta = 0.05$	RS	1.03	88.84%	1.88%	1.29	85.00%	32.08%

but provides poor performance in cost-sensitive settings on all three datasets (e.g., low *WAcc* and TCR). The rough set approach produces poor results on PU1 corpus for both the cost-sensitive and non-cost-sensitive evaluations (e.g., low *spam precision* and *WAcc*, high *WErr* and *cost*). On average, the three-way decision approach provides better results than the original naive Bayes classifier, and outperforms the other two ternary spam filtering methods for the cost-sensitivity aspect. That is, lower *cost* and

Table 5.3: Comparison results on UCI spambase dataset

Thresholds	Approaches	cost	WAcc	WErr	TCR	Spam	
						precision	recall
	NB	1.79	88.04%	11.96%	3.34	88.13%	80.93%
$w = 1$	Three-way	1.07	84.89%	4.46%	8.95	94.70%	77.93%
$\alpha = 0.75$	Robinson	0.08	97.93%	0.00%	0.00	100.00%	98.09%
$\beta = 0.25$	RS	2.29	82.61%	14.57%	2.74	97.38%	60.76%
	NB	2.11	93.63%	6.37%	2.84	94.70%	77.93%
$w = 3$	Three-way	1.72	88.75%	4.05%	4.48	97.56%	76.29%
$\alpha = 0.65$	Robinson	0.17	95.21%	0.00%	0.00	100.00%	98.91%
$\beta = 0.15$	RS	2.55	90.72%	7.26%	2.50	97.01%	61.85%
	NB	0.91	96.86%	3.14%	2.18	97.99%	66.49%
$w = 9$	Three-way	0.84	89.88%	1.78%	3.86	98.55%	55.59%
$\alpha = 0.55$	Robinson	0.56	61.19%	0.00%	0.00	100.00%	98.91%
$\beta = 0.05$	RS	1.08	94.70%	3.54%	1.94	97.02%	62.13%

weighted error, higher *TCR*. When w increases, *spam precision* increases for both the naive Bayes classifier and the three-way decision approach, and the *weighted error* and *TCR* decrease. On the other hand, Robinson's approach and the rough set approach are not sensitive to different cost settings. Their cost-sensitive evaluation results are poor. For instance, the rough set approach has higher *weighted error*, higher *cost*, and lower *TCR* comparing to other spam filters, and Robinson's approach has lower *TCR*

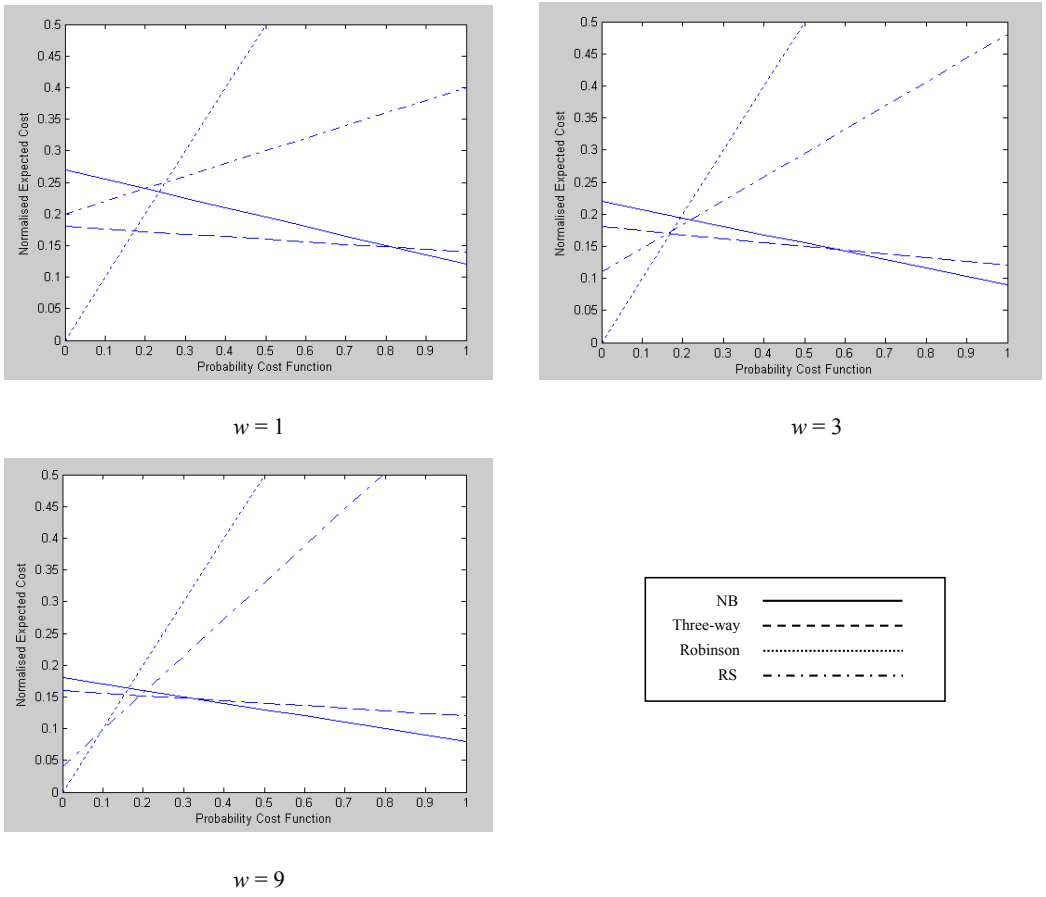

Figure 5.1: Comparison results of cost curves on PU1 corpus for different cost settings

and higher *cost* than the three-way decision approach.

Figure 5.1 shows the *cost curves* of four email spam filters on PU1 corpus under three different cost settings. As we can see, Robinson's approach and the rough set approach have high expected cost under all three settings. The three-way decision approach has lower expected cost than the naive Bayes classifier across the most range of the probability-cost function when $w = 1$ and $w = 3$, and has a close expected cost with the naive Bayes classifier when $w = 9$. Figure 5.2 shows the *cost curves* on Ling-

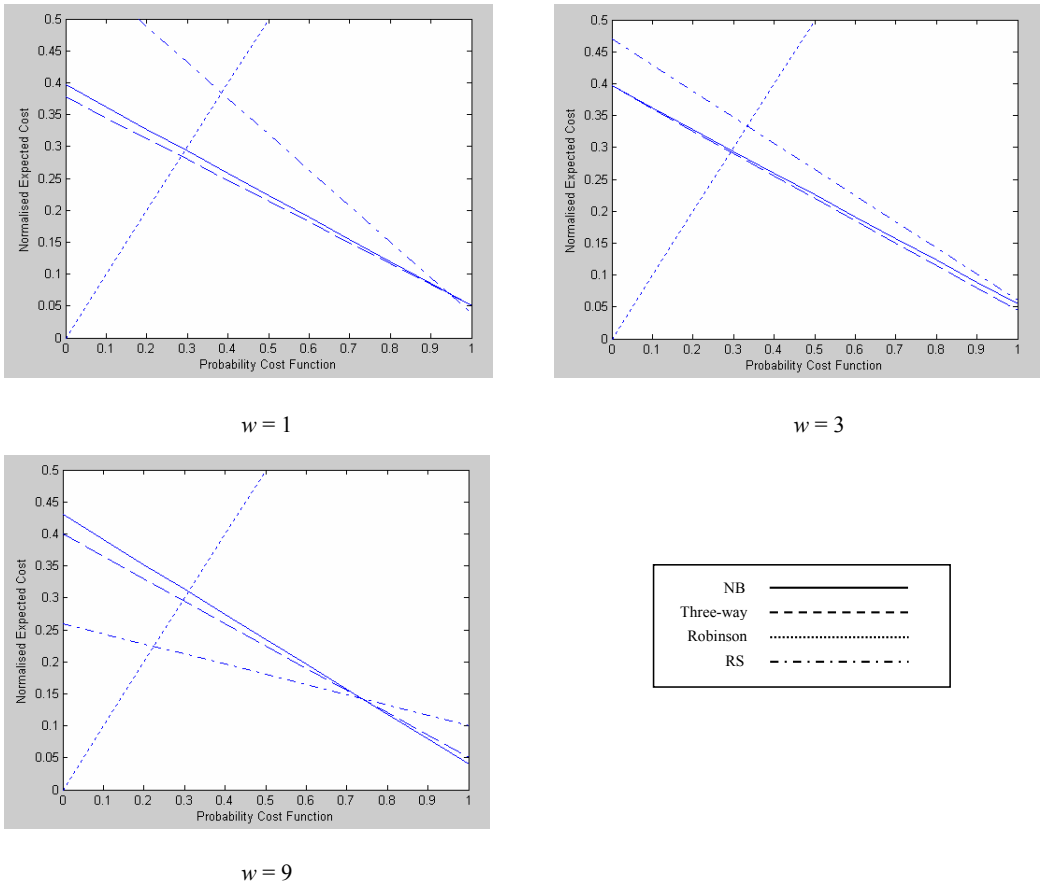

Figure 5.2: Comparison results of cost curves on Ling-Spam corpus for different cost settings

Spam corpus. Robinson's approach has high expected cost under all three settings. The rough set approach dominates the other three spam filters when $w = 9$, but has higher expected costs when $w = 1$ and $w = 3$. The three-way decision approach dominates the naive Bayes classifier under all three settings. Figure 5.3 shows the *cost curves* on the UCI dataset. The rough set approach has high expected costs under all three settings. Robinson's approach dominates the other three spam filters when $w = 1$ and $w = 3$, but has a higher expected cost when $w = 9$. The cost-sensitive

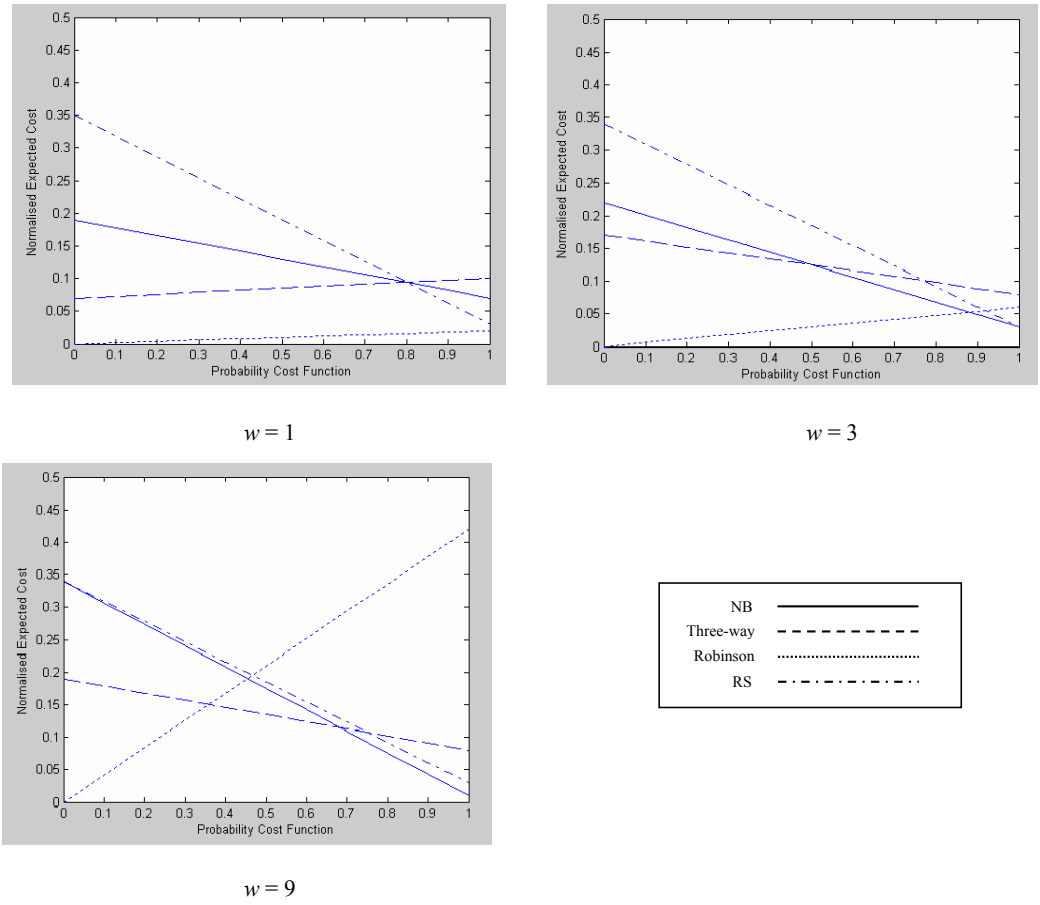

Figure 5.3: Comparison results of cost curves on UCI dataset for different cost settings

three-way decision approach has lower expected cost than the naive Bayes classifier across the most range of the probability-cost function under all three settings.

Overall, for the cost-sensitive evaluations, the three-way decision approach has lowest *cost* comparing to other three existing spam filters. It is sensitive to different cost settings and consistently performs better than other spam filters on all three datasets. Robinson's approach has better *spam precision* for the non-cost-sensitivity aspect, but provides poor performance in the cost-sensitive evaluations. Among the three ternary

spam filters, the three-way decision approach is the only one that outperforms the original naive Bayesian spam filter in both the cost-sensitive and non-cost-sensitive settings on all three datasets.

Chapter 6

EXTENSIONS OF THE BASIC MODEL

In this Chapter, I investigate two extensions of the proposed model. One is multi-class classification and the other is sequential decision-making.

6.1 Multi-Class Classification

The three-way decision approach to spam filtering makes a ternary classification of two classes, that is, there are only two decision classes for an email (e.g., legitimate or spam). A natural extension of this approach is to consider the multi-class case. For instance, phishing (password fishing) has become a severe problem in recent years. Different to unsolicited but harmless spam emails, phishing contains various criminal activities which try to fraudulently acquire sensitive data from internet users, and therefore it is usually considered as a separate email class in addition to legitimate

and spam.

6.1.1 Probabilistic Approximations of Multi-Class Classification

In Pawlak rough set theory [64] and its probabilistic generalizations [40,41,68,124,125, 147], all decision classes are treated as the same in the interpretation and applications of approximations and three regions. In other words, the same pair of thresholds are used to define the positive, negative and boundary regions. As a natural extension to these studies, rough set approximations for multi-class decision problems using different pairs of thresholds have been discussed in several studies. Ślęzak [89] suggests an approach to define the three probabilistic regions based on pair-wise comparisons of categories. A matrix of thresholds is used, with a pair of thresholds on the likelihood ratio of each pair of categories. Although the approach is mathematically appealing and sound, it suffers from a lack of guidelines and systematic methods on how to determining the required thresholds, one may have difficulties in estimating all thresholds. Yao [127] suggests an alternative method to change an m-class classification problem into m two-class classification problems based on the framework of decision-theoretic rough set model (DTRS) [124, 125]. The results from two-class classification can be immediately applied. In addition, m pairs of thresholds can be systematically calculated based on the well established Bayesian decision theory, and interpreted in terms of more practically operable notions such as cost, risk, benefit etc. Some practice issues when applying Yao's idea to classify new objects are further discussed by Liu et al. [53] with the aid from Bayesian decision procedure. However, their work as-

sumes that the losses incurred for misclassifying an object into any classes are the same. This assumption does not always hold in many real world applications. For example, in a medical example, misclassifying a patient with flu to cancer costs more than misclassifying this patient to have pneumonia.

In the multi-class solution given in the original Bayesian decision theory, one has the option of assigning the object to one of the $m(m \geq 2)$ classes. In other words, each class is associated with an action of accepting the object to be a member of that class. Lingras et al. [52] proposed a rough multi-class decision theoretic framework based on DTRS by building a similarity relation between each object and an action corresponding to a subset of categories. The final classification in both of these approaches is made by choosing the action with the minimum expected loss. However, one has to make an immediate decision to either accept or reject the object to be a member of one of the classes, and the losses incurred for making rejections to different classes are not considered.

In this section, a probabilistic approximation for multi-class decision problems based on the three-way decision approach is introduced. Instead of making an immediate acceptance or rejection decision, a third option of making a deferred decision is added to each class. This gives the user the flexibility of refusing to make a decision under certain situations. For example, if the doctor can not diagnose between a few different types of flu based on a patient's symptoms, a series of diagnostic tests can be performed to gather more information to help the doctor making the decision. Moreover, the losses incurred for misclassifying an object into different substitution classes are treated differently, and the losses incurred for making deferred and rejective

decisions to different classes are considered.

6.1.2 Existing Work

Bayesian Decision Theory for Multi-Class Classification

In Bayesian decision theory [20], let $\Omega = \{C_1, C_2, \ldots, C_m\}$ denote a finite set of m classes and $\mathcal{A} = \{a_1, a_2, \ldots, a_m\}$ a finite set of m possible actions. The loss function $\lambda(a_i|C_j)$ is given by a $m \times m$ matrix:

	C_1	C_2	$\cdots$	C_j	$\cdots$	C_m
a_1	$\lambda_{11} = \lambda(a_1\|C_1)$	$\lambda_{12} = \lambda(a_1\|C_2)$	$\cdots$	$\lambda_{1j} = \lambda(a_1\|C_j)$	$\cdots$	$\lambda_{1m} = \lambda(a_1\|C_m)$
a_2	$\lambda_{21} = \lambda(a_2\|C_1)$	$\lambda_{22} = \lambda(a_2\|C_2)$	$\cdots$	$\lambda_{2j} = \lambda(a_2\|C_j)$	$\cdots$	$\lambda_{2m} = \lambda(a_2\|C_m)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	$\lambda_{i1} = \lambda(a_i\|C_1)$	$\lambda_{i2} = \lambda(a_i\|C_2)$	$\cdots$	$\lambda_{ij} = \lambda(a_i\|C_j)$	$\cdots$	$\lambda_{im} = \lambda(a_i\|C_m)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_m	$\lambda_{m1} = \lambda(a_m\|C_1)$	$\lambda_{m2} = \lambda(a_m\|C_2)$	$\cdots$	$\lambda_{mj} = \lambda(a_m\|C_j)$	$\cdots$	$\lambda_{mm} = \lambda(a_m\|C_m)$

In general, the selection of the values of the loss function can be represented as:

$$\lambda(a_i|C_j) = \begin{cases} 0, & i = j, \\ \lambda_{ij}, & i \neq j, \end{cases}$$

where λ_{ij} $(i, j = 1, \ldots, m)$ denotes the loss incurred for deciding class C_i when the true class is C_j. The expected cost associated with taking action a_i is:

$$R(a_i|[x]) = \sum_{j=1}^{j=m} \lambda(a_i|C_j) Pr(C_j \mid [x]). \tag{6.1}$$

The classifier is said to assign an object to class C_i if

$$R(a_i|[x]) < R(a_j|[x]), \qquad \text{for all } j \neq i.$$

Yao's Approach

Inspired by the above multi-class classification solution, Yao [127] suggested changing an m-class classification problem into m two-class classification problems, and making a three-way decision for each class. For the finite set of m classes $\Omega = \{C_1, C_2, \ldots, C_m\}$, the C_j's form a family of pair-wise disjoint subsets of U, namely, $C_i \cap C_j = \emptyset$ for $i \neq j$, and $\cup C_j = U$. For each C_j, a two-class classification $\{C, C^c\}$ can be defined, where $C = C_j$ and $C^c = C_j^c = \cup_{i \neq j} C_i$. The loss function for each C_j is given by a $3 \times m$ matrix:

	C_1	C_2	$\cdots$	C_j	$\cdots$	C_m
a_{P_i}	$\lambda_{P1} = \lambda(a_{P_i}\|C_j)$	$\lambda_{P2} = \lambda(a_{P_i}\|C_j)$	$\cdots$	$\lambda_{Pi} = \lambda(a_{P_i}\|C_j)$	$\cdots$	$\lambda_{Pm} = \lambda(a_{P_i}\|C_j)$
a_{B_i}	$\lambda_{B1} = \lambda(a_{B_i}\|C_j)$	$\lambda_{B2} = \lambda(a_{B_i}\|C_j)$	$\cdots$	$\lambda_{Bi} = \lambda(a_{B_i}\|C_j)$	$\cdots$	$\lambda_{Bm} = \lambda(a_{B_i}\|C_j)$
a_{N_i}	$\lambda_{N1} = \lambda(a_{N_i}\|C_j)$	$\lambda_{N2} = \lambda(a_{N_i}\|C_j)$	$\cdots$	$\lambda_{Ni} = \lambda(a_{N_i}\|C_j)$	$\cdots$	$\lambda_{Nm} = \lambda(a_{N_i}\|C_j)$

The m pairs of thresholds can be systematically calculated based on the given loss functions. The results from DTRS can be immediately applied. However, the losses incurred for making any substitution errors are considered as the same in this approach. That is,

$$\begin{cases} \lambda_{P1} = \lambda_{P2} = \ldots = \lambda_{Pm} = \lambda(a_{Pi}|C_j), & \text{for all } i \neq j, \\ \lambda_{B1} = \lambda_{B2} = \ldots = \lambda_{Bm} = \lambda(a_{Bi}|C_j), & \text{for all } i \neq j, \\ \lambda_{N1} = \lambda_{N2} = \ldots = \lambda_{Nm} = \lambda(a_{Ni}|C_j), & \text{for all } i \neq j. \end{cases}$$

This assumption does not always hold in many real world scenarios, which makes it not practice in real applications.

Ślęzak's Approach

Ślęzak [89] introduced a rough Bayesian model (RB) and applied it to multi-class classifications based on pair-wise comparisons of decision classes. For an information table with a set of decision classes $\{0, \ldots, m-1\}$, a matrix of thresholds is given as:

$$\varepsilon = \begin{bmatrix} * & \varepsilon_1^0 & \cdots & \varepsilon_{m-1}^0 \\ \varepsilon_0^1 & * & & \vdots \\ \vdots & & * & \varepsilon_{m-1}^{m-2} \\ \varepsilon_0^{m-1} & \cdots & \varepsilon_{m-2}^{m-1} & * \end{bmatrix}.$$

where $\varepsilon_i^j \in [0, 1)$ for $i \neq j$ is called a significance threshold that expresses whether the degree of belief is strong enough for class C_i with respect to any other class C_j. The three probabilistic regions are defined as follows:

$$\begin{aligned} \mathrm{POS}_{(\varepsilon)}(C_i) &= \{x \in U \mid \forall_{j:j\neq i} Pr([x] \mid C_j) \leq \varepsilon_i^j Pr([x] \mid C_i)\}, \\ \mathrm{BND}_{(\varepsilon)}(C_i) &= \{x \in U \mid \exists_{j:j\neq i} Pr([x] \mid C_j) > \varepsilon_i^j Pr([x] \mid C_i) \wedge \\ &\qquad \forall_{j:j\neq i} Pr([x] \mid C_i) > \varepsilon_j^i Pr([x] \mid C_j)\}, \\ \mathrm{NEG}_{(\varepsilon)}(C_i) &= \{x \in U \mid \exists_{j:j\neq i} Pr([x] \mid C_i) \leq \varepsilon_j^i Pr([x] \mid C_j)\}. \end{aligned} \tag{6.2}$$

An object x belongs to $\mathrm{POS}_{(\varepsilon)}(C_i)$ if and only if the attribute values used to describe x (i.e., $Des(x)$ or $[x]$) are significantly more likely to occur under C_i than under any other class C_j, $j \neq i$. Object x belongs to $\mathrm{BND}_{(\varepsilon)}(C_i)$ if and only if $[x]$ is not significantly more likely under C_i than under all other C_j but there is also no alternative class, which makes $[x]$ significantly more likely than C_i does. object x belongs to $\mathrm{NEG}_{(\varepsilon)}(C_i)$ if and only if there is an alternative class C_j, which makes $[x]$

significantly more likely than C_i does. However, the selection of significance thresholds can be a subjective and difficult task without systematic guidelines.

6.1.3 Cost-Sensitive Multi-Class Classification

Similar to Yao's idea, for the finite set of m classes $\Omega = \{C_1, C_2, \ldots, C_m\}$, we make a three-way decision to each class C_i, that is, each C_i is associated with a set of three actions $\mathcal{A} = \{a_{P_i}, a_{B_i}, a_{N_i}\}$, where a_{P_i}, a_{B_i}, and a_{N_i} represent the three actions in deciding $x \in \text{POS}(C_i)$, $x \in \text{BND}(C_i)$, and $x \in \text{NEG}(C_i)$, respectively. The loss function is given by a $3 \times m$ matrix for each C_i [143]:

	C_1	C_2	$\cdots$	C_i	$\cdots$	C_m
a_{P_i}	$\lambda_{P1} = \lambda(a_{P_i}\|C_1)$	$\lambda_{P2} = \lambda(a_{P_i}\|C_2)$	$\cdots$	$\lambda_{Pi} = \lambda(a_{P_i}\|C_i)$	$\cdots$	$\lambda_{Pm} = \lambda(a_{P_i}\|C_m)$
a_{B_i}	$\lambda_{B1} = \lambda(a_{B_i}\|C_1)$	$\lambda_{B2} = \lambda(a_{B_i}\|C_2)$	$\cdots$	$\lambda_{Bi} = \lambda(a_{B_i}\|C_i)$	$\cdots$	$\lambda_{Bm} = \lambda(a_{B_i}\|C_m)$
a_{N_i}	$\lambda_{N1} = \lambda(a_{N_i}\|C_1)$	$\lambda_{N2} = \lambda(a_{N_i}\|C_2)$	$\cdots$	$\lambda_{Ni} = \lambda(a_{N_i}\|C_i)$	$\cdots$	$\lambda_{Nm} = \lambda(a_{N_i}\|C_m)$

Different from Yao's approach, the losses incurred for making substitution errors are considered differently. That is,

$$\begin{cases} \lambda_{Pi} = \lambda(a_{Pi}|C_j), & \text{for all } i \neq j, \\ \lambda_{Bi} = \lambda(a_{Bi}|C_j), & \text{for all } i \neq j, \\ \lambda_{Ni} = \lambda(a_{Ni}|C_j), & \text{for all } i \neq j, \end{cases}$$

The expected losses associated with taking different actions for objects in $[x]$ can

be expressed as:

$$\begin{aligned} R(a_{P_i}|[x]) &= \sum_{j=1}^{j=m} Pr(C_j \mid [x])\lambda(a_{P_i} \mid C_j), \\ R(a_{B_i}|[x]) &= \sum_{j=1}^{j=m} Pr(C_j \mid [x])\lambda(a_{B_i} \mid C_j), \\ R(a_{N_i}|[x]) &= \sum_{j=1}^{j=m} Pr(C_j \mid [x])\lambda(a_{N_i} \mid C_j). \end{aligned} \tag{6.3}$$

Recall the minimum-risk decision rules suggested by Bayesian decision procedure are as follows:

(P) If $R(a_{P_i}|[x]) \leq R(a_{B_i}|[x])$ and $R(a_{P_i}|[x]) \leq R(a_{N_i}|[x])$, decide $x \in \text{POS}(C_i)$;

(B) If $R(a_{B_i}|[x]) \leq R(a_{P_i}|[x])$ and $R(a_{B_i}|[x]) \leq R(a_{N_i}|[x])$, decide $x \in \text{BND}(C_i)$;

(N) If $R(a_{N_i}|[x]) \leq R(a_{P_i}|[x])$ and $R(a_{N_i}|[x]) \leq R(a_{B_i}|[x])$, decide $x \in \text{NEG}(C_i)$.

Ordinarily, the loss incurred for making an error is greater than the loss incurred for being correct, and the loss incurred for making a deferment decision is in between. Consider a special kind of loss functions with:

$$\begin{aligned} \text{(c2).} \quad & \lambda(a_{P_i}|C_i) \leq \lambda(a_{B_i}|C_i) < \lambda(a_{N_i}|C_i), \\ & \lambda(a_{N_i}|C_j) \leq \lambda(a_{B_i}|C_j) < \lambda(a_{P_i}|C_j), \text{ for all } j, j \neq i. \end{aligned}$$

That is, the loss of classifying an object x belonging to C_i into the positive region $\text{POS}(C_i)$ is less than or equal to the loss of classifying x into the boundary region $\text{BND}(C_i)$, and both of these losses are strictly less than the loss of classifying x into the negative region $\text{NEG}(C_i)$. The reverse order of losses is used for classifying an object not in C_i. Under condition (c2), we can simplify decision rules (P)-(N) as follows. For the rule (P), the first condition can be expressed as:

$$
\begin{aligned}
& R(a_{P_i}|[x]) \le R(a_{B_i}|[x]) \\
\iff & \sum_{j=1}^{j=m} Pr(C_j|[x])\lambda(a_{P_i}|C_j) \le \sum_{j=1}^{j=m} Pr(C_j|[x])\lambda(a_{B_i}|C_j) \\
\iff & Pr(C_i|[x])\lambda(a_{P_i}|C_i) + \sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])\lambda(a_{P_i}|C_j) \le Pr(C_i|[x])\lambda(a_{B_i}|C_i) + \sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])\lambda(a_{B_i}|C_j) \\
\iff & Pr(C_i|[x]) \ge \frac{\sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])(\lambda(a_{P_i}|C_j) - \lambda(a_{B_i}|C_j))}{\lambda(a_{B_i}|C_i) - \lambda(a_{P_i}|C_i)}.
\end{aligned}
$$

Similarly, other conditions of the three rules can be expressed as:

$$
\begin{aligned}
R(a_{P_i}|[x]) \le R(a_{N_i}|[x]) & \iff Pr(C_i|[x]) \ge \frac{\sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])(\lambda(a_{P_i}|C_j) - \lambda(a_{N_i}|C_j))}{\lambda(a_{N_i}|C_i) - \lambda(a_{P_i}|C_i)}, \\
R(a_{B_i}|[x]) \le R(a_{P_i}|[x]) & \iff Pr(C_i|[x]) \le \frac{\sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])(\lambda(a_{P_i}|C_j) - \lambda(a_{B_i}|C_j))}{\lambda(a_{B_i}|C_i) - \lambda(a_{P_i}|C_i)}, \\
R(a_{B_i}|[x]) \le R(a_{N_i}|[x]) & \iff Pr(C_i|[x]) \ge \frac{\sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])(\lambda(a_{B_i}|C_j) - \lambda(a_{N_i}|C_j))}{\lambda(a_{N_i}|C_i) - \lambda(a_{B_i}|C_i)}, \\
R(a_{N_i}|[x]) \le R(a_{P_i}|[x]) & \iff Pr(C_i|[x]) \le \frac{\sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])(\lambda(a_{P_i}|C_j) - \lambda(a_{N_i}|C_j))}{\lambda(a_{N_i}|C_i) - \lambda(a_{P_i}|C_i)}, \\
R(a_{N_i}|[x]) \le R(a_{B_i}|[x]) & \iff Pr(C_i|[x]) \le \frac{\sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])(\lambda(a_{B_i}|C_j) - \lambda(a_{N_i}|C_j))}{\lambda(a_{N_i}|C_i) - \lambda(a_{B_i}|C_i)}.
\end{aligned}
$$

By introducing three parameters:

$$
\begin{aligned}
\alpha_i &= \frac{\sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])(\lambda(a_{P_i}|C_j) - \lambda(a_{B_i}|C_j))}{\lambda(a_{B_i}|C_i) - \lambda(a_{P_i}|C_i)}, \\
\beta_i &= \frac{\sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])(\lambda(a_{B_i}|C_j) - \lambda(a_{N_i}|C_j))}{\lambda(a_{N_i}|C_i) - \lambda(a_{B_i}|C_i)}, \\
\gamma_i &= \frac{\sum_{j=1,j\neq i}^{j=m} Pr(C_j|[x])(\lambda(a_{P_i}|C_j) - \lambda(a_{N_i}|C_j))}{\lambda(a_{N_i}|C_i) - \lambda(a_{P_i}|C_i)}.
\end{aligned} \tag{6.4}
$$

We can express concisely the decision rules (P)-(N) as:

(P) If $Pr(C_i|[x]) \ge \alpha_i$ and $Pr(C_i|[x]) \ge \gamma_i$, decide $x \in \text{POS}(C_i)$;

(B) If $Pr(C_i|[x]) \le \alpha_i$ and $Pr(C_i|[x]) \ge \beta_i$, decide $x \in \text{BND}(C_i)$;

(N) If $Pr(C_i|[x]) \le \beta_i$ and $Pr(C_i|[x]) \le \gamma_i$, decide $x \in \text{NEG}(C_i)$.

Each rule is defined by two out of the three parameters.

The conditions of rule (B) suggest that it maybe reasonable to impose the constraint $\alpha_i > \beta_i$ so that the boundary region may be non-empty. We can add a sufficient condition on the loss function to ensure $\alpha_i > \beta_i$ as follow:

$$\text{(c3).} \qquad \frac{\lambda(a_{N_i}|C_i) - \lambda(a_{B_i}|C_i)}{\lambda(a_{B_i}|C_j) - \lambda(a_{N_i}|C_j)} > \frac{\lambda(a_{B_i}|C_i) - \lambda(a_{P_i}|C_i)}{\lambda(a_{P_i}|C_j) - \lambda(a_{B_i}|C_j)}. \tag{6.5}$$

The condition (c2) and (c3) imply that $\alpha_i > \gamma_i > \beta_i \geq 0$. After tie-breaking, the following simplified rules are obtained:

$$\begin{aligned}
&\text{(P)} && \text{If } Pr(C_i|[x]) \geq \alpha_i, \text{ decide } x \in \text{POS}(C_i);\\
&\text{(B)} && \text{If } \beta_i < Pr(C_i|[x]) < \alpha_i, \text{ decide } x \in \text{BND}(C_i);\\
&\text{(N)} && \text{If } Pr(C_i|[x]) \leq \beta_i, \text{ decide } x \in \text{NEG}(C_i).
\end{aligned}$$

From the rules (P), (B), and (N), the (α_i, β_i)-probabilistic positive, negative and boundary regions are given, respectively, by:

$$\begin{aligned}
\text{POS}_{(\alpha_i,\beta_i)}(C_i) &= \{x \in U \mid Pr(C_i|[x]) \geq \alpha_i\},\\
\text{BND}_{(\alpha_i,\beta_i)}(C_i) &= \{x \in U \mid \beta_i < Pr(C_i|[x]) < \alpha_i\},\\
\text{NEG}_{(\alpha_i,\beta_i)}(C_i) &= \{x \in U \mid Pr(C_i|[x]) \leq \beta_i\}.
\end{aligned} \tag{6.6}$$

We can extend the probabilistic approximations and regions of a single class to a partition. Let π_D be a partition of the universe U, defined by the decision attribute D. The three regions of the partition π_D can be defined as:

$$\begin{aligned}
\text{POS}_{(\alpha,\beta)}(\pi_D) &= \bigcup_{1\leq i\leq m} \text{POS}_{(\alpha_i,\beta_i)}(C_i),\\
\text{BND}_{(\alpha,\beta)}(\pi_D) &= \bigcup_{1\leq i\leq m} \text{BND}_{(\alpha_i,\beta_i)}(C_i),\\
\text{NEG}_{(\alpha,\beta)}(\pi_D) &= U - \text{POS}_{(\alpha,\beta)}(\pi_D) \cup \text{BND}_{(\alpha,\beta)}(\pi_D).
\end{aligned} \tag{6.7}$$

Table 6.1: A loss function table

	C_1	C_2	C_3	C_4
a_{P_1}	0	2	3	8
a_{B_1}	5	1	1	3
a_{N_1}	10	0	0	0
a_{P_2}	10	0	20	30
a_{B_2}	5	7	10	15
a_{N_2}	0	15	0	0
a_{P_3}	28	25	0	30
a_{B_3}	12	11	10	13
a_{N_3}	0	0	20	0
a_{P_4}	9	5	3	0
a_{B_4}	4	2	1	15
a_{N_4}	0	0	0	30

It is necessary to have a further study on the three probabilistic regions of a classification, as well as the associated rules. In general, one has to consider the problem of rule conflict resolution in order to make effective acceptance, rejection, and abstaining decisions.

6.1.4 An Example

Consider a medical diagnostic example, there is a set of four types of diseases $\Omega = \{C_1, C_2, C_3, C_4\}$. Each disease C_i associated with three actions $\mathcal{A} = \{a_{P_i}, a_{B_i}, a_{N_i}\}$, where a_{P_i} indicates that the doctor decides to treat the patient, a_{N_i} indicates that the doctor decides not to treat the patient, and a_{B_i} indicates that the doctor neither treat nor not treat the patient, further tests are needed. Suppose the symptoms of a new patient are described by $[x]$, the conditional probabilities of the four diseases can be derived from hospital historical data as follows:

$$Pr(C_1|[x]) = 0.4, \quad Pr(C_2|[x]) = 0.2, \quad Pr(C_3|[x]) = 0.15, \quad Pr(C_4|[x]) = 0.25.$$

The loss functions for the four diseases are represented in Table 6.1. Suppose the four diseases are listed in the increasing order of their severity levels, we can see that the loss incurred for misdiagnosing a patient having disease C_1 to C_4 is higher than misdiagnosing a patient having disease C_1 to C_2 or to C_3.

Based on equation (6.4) and Table 6.1, we can compute the pairs of thresholds (α_i, β_i) for the four diseases as:

$$\alpha_1 = 0.35, \quad \beta_1 = 0.22;$$

$$\alpha_2 = 0.94, \quad \beta_2 = 0.78;$$

$$\alpha_3 = 0.83, \quad \beta_3 = 0.65;$$

$$\alpha_4 = 0.19, \quad \beta_4 = 0.14.$$

Now we can compare the conditional probabilities with the corresponding thresholds. Since $Pr(C_1|[x]) = 0.4 > \alpha_1$ and $Pr(C_4|[x]) = 0.25 > \alpha_4$, class C_1 and C_4 are in the

positive region. The new patient could have both disease C_1 and C_4, or either one of them. Rule conflict resolution should be added at this point to further distinguish which disease that the patient is more likely to have.

6.2 Classification of the Deferred Examples

The cost-sensitive three-way decision approach introduced in this thesis allows a classifier to make a deferment decision when it is difficult to decide the class of an example, rather than being forced to make an immediate decision. The deferred examples must be reexamined by collecting further information. A question that remains is how to determine the classification of the deferred examples. One obvious solution is to reexamine and classify them manually. A better solution is to automatically classify these examples by a learning method.

In this section, I propose a solution to this problem based on the idea of granular computing (GrC) [141, 144]. Granular computing is an area of study that explores different levels of granularity in human-centered perception, problem solving, and information processing, as well as their implications and applications in the design and implementation of knowledge-intensive intelligent systems [4, 51, 70, 114, 117, 120]. An decision tree learning method is introduced that further classifies the deferred examples by searching for effective granularity. A decision tree is constructed based on three-way decisions with DTRS. At each level, I sequentially choose the attributes that provide the most suitable granularity. A subtree is added if the conditional probability lies in between of the two thresholds. A branch reaches its leaf node when

the conditional probability is above or equal to the first threshold value, or is below or equal to the second threshold value. More specifically, I start from the bigger granule at the top level of the tree, if the classification decision can not be made based on this granularity, I then search for the smaller granules by adding more attribute as inner nodes, until all the examples are correctly classified or certain condition is met. If the granularity at the current level is sufficient for classification, a finer level of granularity may not be needed at all, this ensures the generated decision tree to be "almost minimal".

6.2.1 Representation of Granules

A central notion in GrC is information and knowledge granularity. Following the classical interpretation of a concept as a pair of an extension and an intension [16,57], a granule can be interpreted as a pair of a set of objects and a logic formula describing the granule. Different sized granules form different levels of granularity. The detailed formulations are introduced as follows.

Recall that an information table can be represented in the form of $S = (U, At = A \cup \{D\}, \{V_a\}, \{I_a\})$, where A is a set of condition attributes describing the objects, and D is a decision attribute that indicates the classes of objects. Different attribute subsets will give different equivalence classes. For example, Table 6.2 is a simple information table, the column labeled by Class denotes an expert's classification of the objects. In Table 6.2, if attribute $A = \{Eyes\}$ is chosen, we can obtain the following family of equivalence classes, or a partition of U:

$$[x]_{\{Eyes\}} = \{\{o_1, o_3, o_4, o_6, o_8\}, \{o_2, o_5, o_7, o_9, o_{10}, o_{11}\}\}.$$

Table 6.2: A simple information table

Object	Weight	Hair	Eyes	Class
o_1	normal	red	blue	+
o_2	low	dark	brown	+
o_3	low	grey	blue	+
o_4	high	red	blue	+
o_5	low	blond	brown	-
o_6	high	dark	blue	-
o_7	low	red	brown	+
o_8	low	blond	blue	+
o_9	low	grey	brown	-
o_{10}	normal	dark	brown	+
o_{11}	high	dark	brown	-

If we consider attribute $A = \{Eyes, Weight\}$, the family of equivalence classes is:

$$[x]_{\{Eyes,Weight\}} = \{\{o_1\}, \{o_2, o_5, o_7, o_9\}, \{o_3, o_8\}, \{o_4, o_6\}, \{o_{10}\}, \{o_{11}\}\}.$$

If we consider each equivalence class as a granule, by choosing different set of attributes from an information table, different granularities can be produced. For certain applications, we may only need to look at granularities of a certain level.

Traditionally, a concept is interpreted as a pair of intension and extension. The intension of a concept is given by a set of properties. In order to formally define

intensions of concepts, we adopt the decision logic language $\mathcal{L}$ used and studied by Pawlak [64] and Zhou and Yao [129, 137, 139]. Formulas of $\mathcal{L}$ are constructed recursively based on a set of atomic formulas corresponding to some basic concepts. The set of atomic formulas are constructed from an attribute-value pair. With respect to an attribute $a \in At$ and an attribute value $v \in V_a$, an atomic formula of the language $\mathcal{L}$ is denoted by $(a = v)$. An object $x \in U$ satisfies an atomic formula $(a = v)$ if the value of x on attribute a equals to value v, written as:

$$x \models (a = v) \quad \text{iff} \quad I_a(x) \ = \ v.$$

From atomic formulas, we can construct other formulas by applying the logic connectives $\neg$, $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$. Each formula represents an intension of a concept. For two formulas ϕ and ψ, we say that ϕ is more specific than ψ, and ψ is more general than ϕ, if and only if $\phi \rightarrow \psi$, namely, ψ logically follows from ϕ. In other words, the formula $\phi \rightarrow \psi$ is satisfied by all objects with respect to any universe U and any information function I_a. If ϕ is more specific than ψ, we write $\phi \preceq \psi$, and call ϕ a sub-concept of ψ, and ψ a super-concept of ϕ.

In inductive learning and concept formation, extensions of concepts are normally defined with respect to a particular training set of examples. If ϕ is a formula, the set $m(\phi)$ defined by:

$$m(\phi) = \{x \in U \mid x \models \phi\}, \tag{6.8}$$

is called the meaning of the formula ϕ in M. The meaning of a formula ϕ is therefore the set of all objects having the property expressed by the formula ϕ. In other words, ϕ can be viewed as the description of the set of objects $m(\phi)$. Thus, a connection

between formulas and subsets of U is established. For example, in Table 6.2, if attribute $A = \{Eyes\}$ is chosen, the two equivalence classes can be written as:

$$m(Eyes = blue) = \{o_1, o_3, o_4, o_6, o_8\},$$
$$m(Eyes = brown) = \{o_2, o_5, o_7, o_9, o_{10}, o_{11}\},$$

where $Eyes = blue$ and $Eyes = brown$ are the intensions of the concepts described by the formulas of the language $\mathcal{L}$.

With the introduction of language $\mathcal{L}$, we have a formal description of concepts. A concept definable in a model S is a pair $(\phi, m(\phi))$, where $\phi \in \mathcal{L}$. More specifically, ϕ is a description of $m(\phi)$ in S, the intension of concept $(\phi, m(\phi))$, and $m(\phi)$ is the set of objects satisfying ϕ, the extension of concept $(\phi, m(\phi))$. A concept $(\phi, m(\phi))$ is said to be a sub-concept of another concept $(\psi, m(\psi))$, or $(\psi, m(\psi))$ a super-concept of $(\phi, m(\phi))$, in an information table if $m(\phi) \subseteq m(\psi)$. A concept $(\phi, m(\phi))$ is said to be a smallest non-empty concept in S if there does not exist another non-empty proper sub-concept of $(\phi, m(\phi))$.

Concept learning, to a large extent, depends on the structures of concept space and the target concepts. In general, one may not be able to obtain an effective and efficient learning algorithm, if no restrictions are imposed on the concept space. For this reason, each learning algorithm typically focuses on a specific type of concept.

6.2.2 A Learning Algorithm based on GrC

For classification problems, ID3 [71] is a well-known algorithm used to generate a decision tree by sequentially choosing the attribute that gives the most information about

the class label, the leaf node is added to a branch if all examples have the same class labels. In this subsection, I introduce a learning method to construct a decision tree for classification based on three-way decisions with DTRS [121]. A subtree is added recursively if the conditional probability lies in between of the two threshold values α and β. Otherwise, a leaf node is added if the conditional probability $Pr(C|[x]) \geq \alpha$, or $Pr(C|[x]) \leq \beta$, or if the granule meets certain condition. A general scheme of this classification process is shown as follows.

GrCL

Use the entire training set as the unlabeled root node of a decision tree;

while *there is an unlabeled leaf node in the tree* **do**
Choose an unlabeled leaf node;

if *the a posteriori probability is above or equal to* α **then**
| Change the node to a labeled leaf node with label = "accept";

else if *the a posteriori probability is below or equal to* β **then**
| Change the node to a labeled leaf node with label = "reject";

else if *the granule meets certain conditions* **then**
| Change the node to a labeled leaf node with label = "deferment";

else
| Replace the unlabeled node with an attribute with each branch corresponds to an attribute value, divide the granule into unlabeled nonempty sub-granules based on the attribute value

end

end

return a ternary classification tree.

Instead of generating the decision tree based on the information gain, in this approach, the attributes of each inner node of the decision tree are sequentially selected by searching for the most suitable granularity at each level. The actual learning process of building the decision tree is a little bit more complicated. There are a few important steps involved in the learning method, which can be described as follows.

Step 1: At the top most level, the attribute has the least attribute values is selected as the root node a, which will give us the largest granulation. If there is more than one attribute satisfies this condition, we choose the attribute that has has the minimum number of objects in its corresponding boundary region of C.

Step 2, A new branch is added for each possible value v_i of a. Estimate the conditional probability $Pr(C|[x])$ of each branch with respect to a. If $Pr(C|[x]) \geq \alpha$, objects of this branch belong to the positive region, add a leaf node labeled as *accept*; if $Pr(C|[x]) \leq \beta$, objects of this branch belong to the negative region, add a leaf node labeled as *reject*; if granule defined at the current level meet our predefined condition, add a leaf node labeled as *deferment*. Otherwise, the classification of this branch can not be determined, we then search for the next suitable granularity.

Step 3, At the next level, choose an attribute b that has the least attribute values from $(At - \{a\})$ as the child node for the branch. This time, $Des[x] = (a = v_i \wedge b = v_j)$, where v_j represents the possible values of b. Similarly, if there is more than one attribute satisfies this condition, we choose the attribute that has the minimum number of objects in its corresponding boundary regions of C. Repeat Step 2. until we can find a leaf node for each branch.

This three-way classification process is illustrated in the following example.

6.2.3 An Example

Table 6.2 is an information table. At the top level, attribute *Eyes* is chosen as the root node since it has the least attribute values. Two branches are added corresponding to two possible values *Eyes=blue* and *Eyes=brown*, which divides the data set into two granules:

$$m(Eyes = blue) = \{o_1, o_3, o_4, o_6, o_8\},$$
$$m(Eyes = brown) = \{o_2, o_5, o_7, o_9, o_{10}, o_{11}\}.$$

And since

$$m(Class = +) = \{o_1, o_2, o_3, o_4, o_7, o_8, o_{10}\}.$$

The conditional probability of each equivalence class can be calculated as follows:

$$\begin{aligned} Pr(+ \mid [o_1]_{\{Eyes\}}) &= \frac{|\{o_1, o_3, o_4, o_6, o_8\} \cap \{o_1, o_2, o_3, o_4, o_7, o_8, o_{10}\}|}{|\{o_1, o_3, o_4, o_6, o_8\}|} \\ &= \frac{4}{5}, \\ Pr(+ \mid [o_2]_{\{Eyes\}}) &= \frac{|\{o_2, o_5, o_7, o_9, o_{10}, o_{11}\} \cap \{o_1, o_2, o_3, o_4, o_7, o_8, o_{10}\}|}{|\{o_2, o_5, o_7, o_9, o_{10}, o_{11}\}|} \\ &= \frac{1}{2}. \end{aligned}$$

Assume the two thresholds calculated from the loss functions are $\alpha = 0.6$ and $\beta = 0.4$. We have $Pr(+ \mid [o_1]_{\{Eyes\}}) \geq \alpha$, objects $\{o_1, o_3, o_4, o_6, o_8\}$ belong to the positive region of class +, a leaf node can be added to this branch with the class label +. We also have $\beta < Pr(+ \mid [o_2]_{\{Eyes\}}) < \alpha$, objects $\{o_2, o_5, o_7, o_9, o_{10}, o_{11}\}$ belong to the boundary region of class + and need to be further analyzed.

At the second level, attribute *Weight* is chosen since it has less attribute values than attribute *Hair*. A subtree is added to the *Eyes=brown* branch, with three new

branches corresponding to three possible values, that is, *Weight=normal*, *Weight=high*, and *Weight=low*, which divide the data set into three granules:

$$m(Eyes = brown \wedge Weight = normal) = \{o_{10}\},$$
$$m(Eyes = brown \wedge Weight = high) = \{o_{11}\},$$
$$m(Eyes = brown \wedge Weight = low) = \{o_2, o_5, o_7, o_9\}.$$

The conditional probability of each equivalence class can be calculated as follows:

$$\begin{aligned} Pr(+ \mid [o_{10}]_{\{Eyes,Weight\}}) &= \frac{|\{o_{10}\} \cap \{o_1, o_2, o_3, o_4, o_7, o_8, o_{10}\}|}{|\{o_{10}\}|} \\ &= 1, \\ Pr(+ \mid [[o_{11}]_{\{Eyes,Weight\}}) &= \frac{|\{o_{11}\} \cap \{o_1, o_2, o_3, o_4, o_7, o_8, o_{10}\}|}{|\{o_{11}\}|} \\ &= 0, \\ Pr(+ \mid [o_2]_{\{Eyes,Weight\}}) &= \frac{|\{o_2, o_5, o_7, o_9\} \cap \{o_1, o_2, o_3, o_4, o_7, o_8, o_{10}\}|}{|\{o_2, o_5, o_7, o_9\}|} \\ &= \frac{1}{2}. \end{aligned}$$

Assume that we use the same α and β. We have $Pr(+ \mid [o_{10}]_{\{Eyes,Weight\}}) \geq \alpha$, objects $\{o_{10}\}$ belong to the positive region of class +, a leaf node can be added to this branch with the class label +. We also have $Pr(+ \mid [o_{11}]_{\{Eyes,Weight\}}) \leq \beta$, objects $\{o_{11}\}$ belong to the negative region of class +, a leaf node can be added to this branch with the class label -. Finally, $\beta < Pr(+ \mid [o_2]_{\{Eyes,Weight\}}) < \alpha$, objects $\{o_2, o_5, o_7, o_9\}$ belong to the boundary region of class + and need to be further analyzed.

At the third level, attribute *Hair* is chosen. A subtree is added to the *Eyes = brown* $\wedge$*Weight = low* branch, with four new branches corresponding to four possible value, that is, *Hair = red*, *Hair = blond*, *Hair = grey*, and *Hair = dark*, which divide the

data set into four granules:

$$m(Eyes = brown \wedge Weight = low \wedge Hair = red) = \{o_7\},$$

$$m(Eyes = brown \wedge Weight = low \wedge Hair = blond) = \{o_5\},$$

$$m(Eyes = brown \wedge Weight = low \wedge Hair = grey) = \{o_9\},$$

$$m(Eyes = brown \wedge Weight = low \wedge Hair = dark) = \{o_2\}.$$

The conditional probability of each equivalence class can be calculated as follows:

$$\begin{aligned}
Pr(+ \mid [o_7]_{\{Eyes,Weight,Hair\}}) &= \frac{|\{o_7\} \cap \{o_1, o_2, o_3, o_4, o_7, o_8, o_{10}\}|}{|\{o_7\}|} \\
&= 1, \\
Pr(+ \mid [o_5]_{\{Eyes,Weight,Hair\}}) &= \frac{|\{o_5\} \cap \{o_1, o_2, o_3, o_4, o_7, o_8, o_{10}\}|}{|\{o_5\}|} \\
&= 0, \\
Pr(+ \mid [o_9]_{\{Eyes,Weight,Hair\}}) &= \frac{|\{o_9\} \cap \{o_1, o_2, o_3, o_4, o_7, o_8, o_{10}\}|}{|\{o_9\}|} \\
&= 0, \\
Pr(+ \mid [o_2]_{\{Eyes,Weight,Hair\}}) &= \frac{|\{o_2\} \cap \{o_1, o_2, o_3, o_4, o_7, o_8, o_{10}\}|}{|\{o_2\}|} \\
&= 1.
\end{aligned}$$

Assume that we use the same thresholds α and β. We have $\{o_7\}$ and $\{o_2\}$ in the positive region of class +, a leaf node can be added to both of these branches with class label +. We also have $\{o_9\}$ and $\{o_5\}$ in the negative region of class +, a leaf node can be added to both of these branches with class label -. Up to this point, all the objects in the data set have been classified, the decision tree is complete. This classification process in search of effective granularity based on three-way decisions

is illustrated in Figure 6.1. Note that the decision tree can be complete at certain level if a finner granularity is not needed. Compare to other decision tree algorithms, such as ID3 and C4.5, the learning algorithm based on GrC does not allow to have different order of attributes on different branches. It may produce a deeper tree for certain datasets.

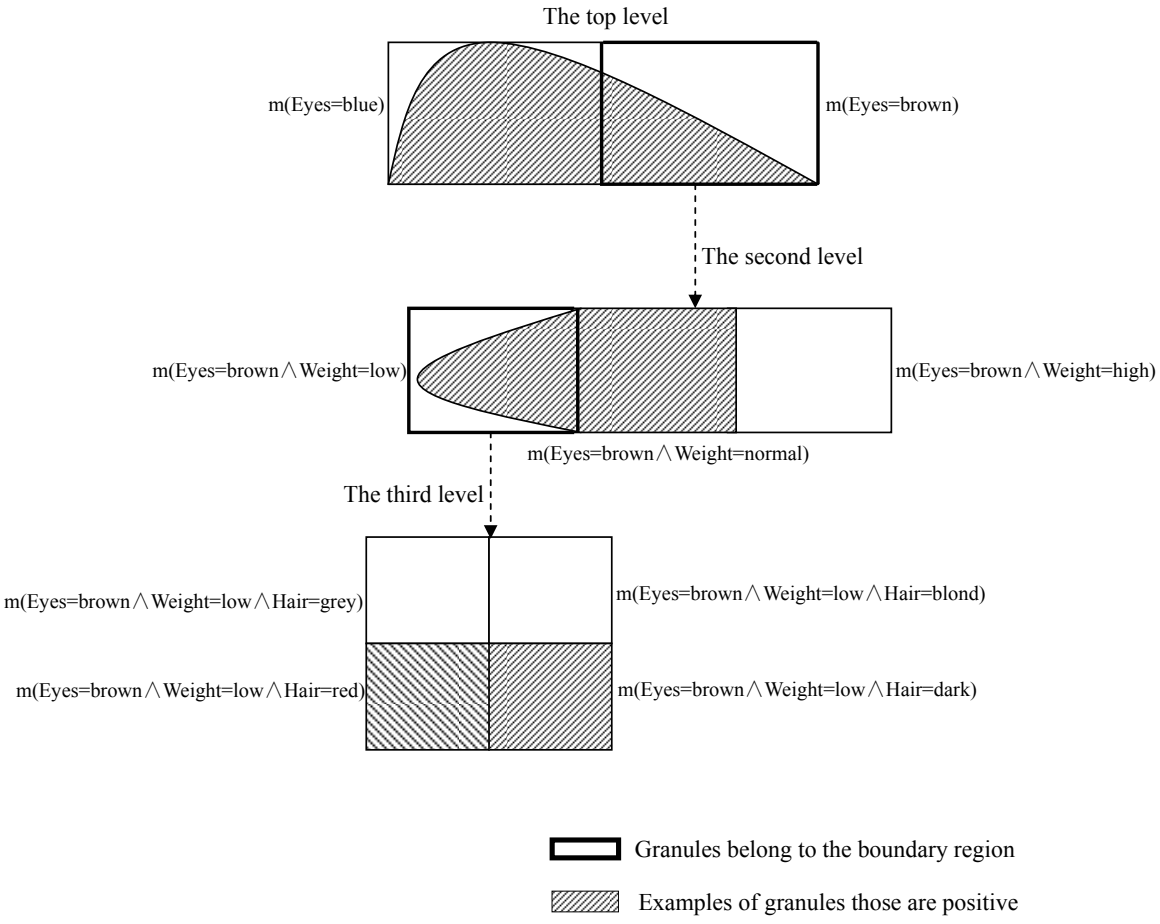

Figure 6.1: The classification process in search of effective granularity based on three-way decisions

Chapter 7

CONCLUSION AND FUTURE RESEARCH

In this chapter, I conclude the presented work, summarize contributions of the thesis, and outline some possible future research directions.

7.1 Summary

The contributions of this thesis are fourfold: a unified framework of cost-sensitive three-way decisions for building ternary classifiers, a complete model of Bayesian rough sets, a practical email spam filtering architecture based on this model, and an investigation of two extensions of the basic model. I detail each of the contributions as follows.

A cost-sensitive approach to ternary classification is presented in this thesis. Compared to the most commonly used binary classification, a third choice is added for

making classification decisions, which gives the user the flexibility of refusing to make an immediate decision under certain conditions. The theoretical foundation of the approach is based on a unified framework that connections two powerful data analysis tools, namely, Bayesian inference and rough set theory.

In the existing studies on Bayesian approaches to rough sets attention has been mainly paid to mathematical constructions and formal properties of various notions. The semantics of these models have not been explicitly studied and made clear. Through a careful examination of these existing work, I divide them into two classes, namely, decision-theoretic rough set models and confirmation-theoretic rough set models. Although these two classes share many similarities in terms of making use of Bayes' theorem and a pair of thresholds to produce three regions, their semantic interpretations and hence intended applications are different. An understanding of the differences behind these two applications enables me to focus on the classification task and propose a framework of three-way decisions for building ternary classifiers.

By investigating the main results of existing probabilistic rough set models, I summarize the salient features of each model and its unsolved problems. Existing studies show that a probabilistic rough set model must address at least three issues, that is, interpretation and computation of thresholds, estimation of conditional probability, and interpretation and applications of three regions in data analysis. Such an understanding enables me to show that there is a lack of a well-developed Bayesian rough set model and hence a complete model of Bayesian rough sets is introduced in this thesis by investigating these basic issues. The Bayesian rough set models proposed in this thesis are different from other Bayesian approaches to rough sets.

Email spam filtering is used as a real world application to show the usefulness of the three-way decision approach. The main difference between my approach and other existing ternary email spam filtering methods is the computation of the required thresholds. Instead of supplying them based on intuitive understandings or trail and error, I provide a systematically calculation based on the decision-theoretic rough set model. The cost associated with each decision is given by a loss function from the well established Bayesian decision theory. The cost-sensitive characteristic of email spam filtering is reflected by varying the values of loss functions. The experiment results on several benchmark datasets show that the new approach outperforms other existing spam filters in different cost settings and has the lowest overall cost.

Two extensions of the three-way decision approach are discussed. The first extension is the multi-class classifications. By changing an m-category classification problem into m two-category classification problems, the three-way decision approach for two-category classification can be immediately applied. This approach can be considered as a straightforward generalization of the three-way classification. The second extension is the classification of the deferred examples. This can be done by automatically searching for effective granularity. An adaptive learning algorithm is proposed, in which a decision tree is generated by sequentially searching the attributes that provide the most appropriate granularity. I analyze the differences between the new approach and other existing works, and illustrative examples are given as demonstrations.

7.2 Future Research

There are at least four interesting research directions arising from the investigation of this thesis.

1. One can easily apply the methodology in this thesis to develop a different version of a Bayesian rough set model with respect to any other definitions of three regions (e.g., confirmation regions of confirmation-theoretic models), by focusing on the same three basic issues. With further investigations, it may be possible to produce a more grand Bayesian rough set model that encompasses these models as its special cases.

2. Bayesian inference leads to two types of applications. My main focus in this thesis is on the first application, that is, classifying objects based on their satisfiability of the hypothesis. The second application, namely, evaluating the quality of different pieces of evidence, has not been fully explored. In future research, I will have a further study on the second type of application and develop related methodologies for feature selections.

3. Email spam filtering is used as a real world application to illustrate the usefulness of proposed framework. In general, the three-way decision approach can also be applied to junk eliminations of other types of web documents such as web pages, news groups and blogs with slight modifications.

4. For three-way decisions on multi-class classification, one may have a further study on three-way decision rules generated from different classes and the associated rule conflict resolutions for real classification applications. For the classification of the

deferred examples, I will continue working on the implementation of the algorithm and compare the effects and efficiencies with some other related studies.

5. The class imbalance problem has been recognized as a crucial problem in machine learning and data mining because in certain cases it causes seriously negative effect on the performance of learning methods. It has been indicated that learning from imbalanced data sets and learning when costs are unequal can be handled in a similar manner and cost-sensitive learning is a good solution to the class imbalance problem. One of my future research is to apply the proposed method in this thesis to class imbalance problems.

Bibliography

[1] Aizerman, M.A. Braverman, E.M., and Rozonoer, L.I. Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 25, pp. 821-837, 1964.

[2] Androutsopoulos, I. Koutsias, J. Chandrinos, K.V. Paliouras, G., and Spyropoulos, C.D. An evaluation of naive Bayesian anti-Spam filtering. In Proceeding of the workshop on Machine Learning in the New Information Age, 2000.

[3] Apte, C., Damerau, F., and Weiss, S.M. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, 12(3), pp. 233-251, 1994.

[4] Bargiela, A., and Pedrycz W. *Granular Computing: An Introduction*, Kluwer Academic Publishers, Boston, 2002.

[5] Barracuda Spam firewall, from http://www.barracudanetworks.com.

[6] Bayes, T., and Price, R. An essay towards solving a problem in the doctrine of chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. Philosophical Transactions of the Royal Society of London 53 (0), pp. 370-418, 1763.

[7] Bogofilter, from http://bogofilter.sourceforge.net.

[8] Box, G.E.P., and Tiao, G.C. *Bayesian Inference in Statistical Analysis*, Wiley, ISBN 0-471-57428-7, 1973.

[9] Brown, G., Pocock, A., Zhao, M.J., and Lujan, M. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. Journal of Machine Learning Research (JMLR), 2012.

[10] Clark, P., and Niblett, T. The CN2 induction algorithm. Machine Learning, 3(4), pp. 261-283, 1989.

[11] Cohen, W. Learning rules that classify email, Advances in Inductive Logic Programming, 1996.

[12] Cortes, C., and Vapnik, V.N. Support-vector networks, Machine Learning, 20, 1995.

[13] Cox, E. *The Fuzzy Systems Handbook: a Practitioners Guide to Building, Using and Maintaining Fuzzy Systems.* Academic Press, Inc., 1994.

[14] Cristianini, N., and Shawe-Taylor, J. *An Introduction to Support vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[15] Dembczyński, K., Greco, S., Kotlowski, W., and Slowiński, R. Statistical model for rough set approach to multicriteria classification. Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp.164-175, 2007.

[16] Demri, S., and Orlowska, E. Logical analysis of indiscernibility, in: Incomplete Information: Rough Set Analysis, Orl owska, E. (Ed.), Physica Verlag, Heidelberg, pp. 347-380, 1997.

[17] Domingos, P., and Pazzani, M. Beyond independence: conditions for the optimality of the simple Bayesian classifier. Proceedings of the 13th International Conference on Machine Learning, pp. 105-112, 1996.

[18] Drummond, C., and Holte, R.C. Explicitly representing expected cost: an alternative to ROC representation. KDD 2000, pp. 198-207, 2000.

[19] Drummond, C., and Holte, R.C. Cost curves: An improved method for visualizing classifier performance. Machine Learning, 65(1), pp. 95-130, 2006.

[20] Duda, R.O., and Hart, P.E. *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[21] Earman, J. Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory. The MIT Press, 1992.

[22] Elkan, C. The foundations of cost-senstive learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence, pp. 973-978, 2001.

[23] Fayyad, U.M., and Irani, K.B. Multi-interval discretization of continuous-valued attributes for classification learning. Proceedings of the 13th International Joint Conference on Artificial Intelligence, pp. 1022-1029, 1993.

[24] Festa, R. *Bayesian Confirmation.* In: M. Galavotti and A. Pagnini (Eds.), Experience, Reality, and Scientific Explanation, Dordrecht: Kluwer Academic Publishers, pp. 55-87, 1999.

[25] Fitelson, B. Studies in Bayesian Confirmation Theory. Ph.D. Dissertation, University of Wisconsin, http://fitelson.org/thesis.pdf, 2001.

[26] French, S. *Decision Theory: An Introduction to the Mathematics of Rationality.* Halsted Press, New York, 1988.

[27] Geng, L., and Hamilton, H.J. Interestingness measures for data mining: a survey, ACM Computing Surveys, 38(3), 2006.

[28] GFI MailEssentials, from http://www.gfi.com/.

[29] Good, I.J. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* M.I.T. Press. 1965.

[30] Goudey, R. Do statistical inferences allowing three alternative decision give better feedback for environmentally precautionary decision-making. Journal of Environmental Management, Vol. 85, pp. 338-344, 2007.

[31] Graham, P. Http://www.paulgraham.com/spam.html. A Plan for Spam, 2002.

[32] Greco, S., Pawlak, Z., and Slowiński, R. Bayesian confirmation measures within rough set approach. Rough Sets and Current Trends in Computing, pp. 264-273, 2004.

[33] Greco, S., Pawlak, Z., and Slowiński, R. Can Bayesian confirmation measures be useful for rough set decision rules? Engineering Applications of Arificial Intelligence, Vol. 17, pp. 345-361, 2004.

[34] Greco, S., Matarazzo, B., and Slowinski, R. Rough membership and Bayesian confirmation measures for parameterized rough sets. RSFDGrC (1), pp. 314-324, 2005.

[35] Greco, S., Matarazzo, B., and Słowiński, R. Parameterized rough set model using rough membership and Bayesian confirmation measures. International Journal of Approximate Reasoning, Vol. 49, pp. 285-300, 2009.

[36] Grzymala-Busse, J. Knowledge acquisition under uncertainty - a rough set approach. Journal of Intelligent and Robotic Systems, Vol. 1, pp. 3-16, 1988.

[37] Grzymala-Busse, J., Marepally, S.R., and Yao, Y.Y. An empirical comparison of rule sets induced by LERS and probabilistic rough classification. RSCTC'10, pp. 590-599, 2010.

[38] Guzella, T. S., and Caminhas, W. M., A review of machine learning approaches to spam filtering. Expert Systems with Applications, 36(7), 2009.

[39] Hall, M. Correlation-based Feature Selection for Machine Learning, 1999.

[40] Herbert, J.P., and Yao, J.T. Game-theoretic risk analysis in decision-theoretic rough sets. Proceedings of RSKT'08, LNAI 5009, pp. 132-139, 2008.

[41] Herbert, J.P., and Yao, J.T. Game-theoretic rough sets, Fundamenta Informaticae, 2009.

[42] Holte, R.C. Very simple classification rules perform well on most commonly used datasets. Machine Learning, 11, pp 63-90, 1993.

[43] Kalt, T. A new probabilistic model of text classification and retrieval TITLE2:. Technical Report UMCS- 1998-018, University of Massachusetts, Amherst, Computer Science, March, 1998.

[44] Kasabov, N.K. *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering.* The MIT Press, 1996.

[45] Katzberh, J.D., and Ziarko, W. Variable precision extension of rough sets. Foundamenta Informaticae. 27(2,3), pp. 155-169, 1996.

Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. The MIT Press, 1996.

[46] Komorowski, J., Pawlak, Z., Polkowski, L., and Skowron, A. Rough Sets: A Tutorial, in: S.K. Pal and A. Skowron (eds.), Rough fuzzy hybridization: A new trend in decision-making, Springer-Verlag, Singapore, pp. 3-98. 1999.

[47] Langley, P., Wayne, I., and Thompson, K. An analysis of Bayesian classifiers. Proceedings of the 10th National Conference on Artificial Intelligence, pp. 223-228, 1992.

[48] Leung, Y., Wu, W.Z., and Zhang, W.X. Knowledge acquisition in incomplete information systems: A rough set approach. European Journal of Operational Research, Vol. 168, pp. 164-180, 2006.

[49] Li, Y., Zhang, C., and Swanb, J. R. Rough set based model in information retrieval and filtering. Proceedings of the Fifth International Conference on Information Systems Analysis and Synthesis, pp. 398-403, 1999.

[50] Li, X., Hamilton, H.J., Karimi, K., and Geng, L. The multi-tree cubinga algorithm for computing iceberg cubes, Journal of Intelligent Information Systems, 33(2), pp. 179-208, 2009.

[51] Liang, J.Y., and Qian, Y.H. Information granules and entropy theory in information systems. Science in China (Series F). 51(9), 2008.

[52] Lingras, P., Chen, M., and Miao, D.Q. Rough multi-category decision theoretic framework. RSKT, pp. 676-683, 2008.

[53] Liu, D., Li, T.R., Hu, P., and Li, H.X. Multiple-category classification with decision-theoretic rough sets. RSKT, pp. 703-710, 2010.

[54] Lurie, J.D., and Sox, H.C. Principles of medical decision making. Spine 24, pp. 493-498, 1999.

[55] Masand, B. Linoff, G., and Waltz, D. Classifying news stories using memory based reasoning. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 59-65, 1992.

[56] http://www.ics.uci.edu/mlearn/MLRepository.html.

[57] Michalski, R.S., Carbonell, J.G., and Mitchell, T.M. (Eds.), *Machine Learning, an Artificial Intelligence Approach*, Morgan Kaufmann Publishers, Inc., Los Altos, California, 1983.

[58] Mitchell, T.M. *Machine Learning*, McGraw-Hill, New York, 1997.

[59] Moulinier, I. A framework for comparing text categorization approaches. In Proceedings of the 1996 AAAI Spring Symposium on Machine Learning in Information Access, 1996.

[60] Moukas, A., and Maes, P.A. An evolving multiagent information filtering and discovery system for the www. autonomous agents and multi-agent systems 1, 1998.

[61] Pantel, P., and Lin, D.K. SpamCop - a spam classification & organization program. In Proceedings of AAAI Workshop on Learning for Text Categorization. pp. 95-98. Madison, WI, 1998.

[62] Pauker, S.G., and Kassirer, J.P. The threshold approach to clinical decision making. New England Journal of Medicine, Vol. 302, pp. 1109-1117, 1980.

[63] Pawlak, Z. Rough sets, International Journal of Computer and Information Sciences, Vol. 11, pp. 341-356, 1982.

[64] Pawlak, Z. Rough Sets, *Theoretical Aspects of Reasoning about Data*, Dordrecht: Kluwer Academic Publishers, 1991.

[65] Pawlak, Z. Decision rules, Bayes' rule and rough sets. Proceedings of 7th International Workshop: New Directions in Rough Sets, Data Mining, and Granular -Soft Computing (RSFDGSC'99), LNAI 1711, pp. 1-9, 1999.

[66] Pawlak, Z. Rough sets, decision algorithms and Bayes theorem. European Journal of Operational Research, Vol. 136, pp. 181-189, 2002.

[67] Pawlak, Z., and Skowron, A. Rough membership functions. In: Yager, R.R., Fedrizzi, M. and Kacprzyk, J., Eds., Advances in the Dempster-Shafer Theory of Evidence, John Wiley and Sons, New York, pp. 251-271, 1994.

[68] Pawlak, Z., Wong, S.K.M., and Ziarko, W. Rough sets: probabilistic versus deterministic approach. International Journal of Man-Machine Studies, Vol. 29, pp. 81-95, 1988.

[69] Polkowski, L., and Skowron, A. (eds.) *Rough Sets in Knowledge Discovery: Methodology and Applications*, Vol. 1,2, Physica-verlag, 1998.

[70] Qian, Y.H., Liang, J.Y., and Dang, C.Y. Knowledge structure, knowledge granulation and knowledge distance in a knowledge base. International Journal of Approximate Reasoning, 50(1), pp. 174-188, 2009.

[71] Quinlan, J.R. Learning efficient classification procedures and their application to chess endgames, in: Michalski, J.S., Carbonell, J.G., and Michell, T.M. (Eds), *Machine Learning: An Artificial Intelligence Approach*, vol. 1, Morgan Kaufmann, Palo Alto, CA, pp. 463-482, 1983.

[72] Quinlan, J. R. Induction of decision trees. Machine Learning. 1(1), pp. 81-106.1986.

[73] Quinlan, J. R. C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[74] Quinlan, J. R. http://www.rulequest.com.

[75] Rennie, J. “ifile”. http://people.csail.mit.edu/jrennie/ifile/, 1996.

[76] Riloff, E., and Lehnert, W. Information extraction as a basis for high-precision text classification. ACM Transactions on Information Systems, 12(3), pp. 296-333, July 1994.

[77] Robinson, G. A Statistical Approach to the Spam Problem, Spam Detection. Why Chi? Motivations for the use of Fishers Inverse Chi-Square Procedure in Spam Classification, Handling Redundancy in Email Token Probabilities, 2004.

[78] http://rosetta.lcb.uu.se.

[79] Sahami, M. Dumais, S. Heckerman, D., and Horvitz, E. A Bayesian approach to filtering junk e-mail. AAAI Workshop on Learning for Text Categorization, Madison, Wisconsin. AAAI Technical Report WS-98-05, 1998.

[80] Salton, G. Wong, A., and Yang, C. S. A vector space model for automatic indexing, Communications of the ACM, 18(11), pp. 613-620, 1975.

[81] Savage, L.J. The Foundation of Statistics. 2nd edition. Dover. 1972.

[82] Schapire, E., and Singer, Y. BoosTexter: a boosting-based system for text categorization. Machine Learning, 39(2/3), pp. 135-168, 2000.

[83] Schechter, C.B. Sequential analysis in a Bayesian model of diastolic blood pressure measurement. Medical Decision Making 8, pp. 191-196, 1988.

[84] Sebastiani, F. Machine learning in automated text categorization. ACM Computing Surveys, 34(1), pp. 1-47, 2002.

[85] Shen, Q., and Chouchoulas, A. Combining rough sets and data-driven fuzzy learning. Pattern Recognition, 32(12), pp. 2073-2076, 1999.

[86] Sherif, M., and Hovland, C.I. Social Judgment: Assimilation and Contrast Effects in Communication and Attitude Change. Yale University Press, New Haven, 1961.

[87] Siegelmann, H.T., and Sontag, E.D. Turing computability with neural nets. Appl. Math. Lett, 4 (6), pp. 77-80, 1991.

[88] Siersdorfer, S., and G. Weikum. Using restrictive classification and meta classification for junk elimination. In Proceedings of ECIR'2005, pp. 287-299, 2005.

[89] Ślęzak, D. Rough sets and Bayes factor. Transactions on Rough Sets III, LNCS 3400, pp. 202-229, 2005.

[90] Ślęzak, D., and Ziarko, W. Bayesian rough set model. Proceedings of FDM'2002, pp. 131-135. 2002.

[91] Ślęzak, D., and Ziarko, W. The investigation of the Bayesian rough set model. International Journal of Approximate Reasoning, Vol. 40, pp. 81-91, 2005.

[92] Ślęzak, D., Wróblewski, J., Eastwood, V., and Synak, P. Brighthouse: an analytic data warehouse for ad-hoc queries. Proceedings of the VLDB Endowment, pp. 1337-1345, 2008.

[93] Smith, E.E. Concepts and induction, in Posner, M.I. (Ed.), *Foundations of Cognitive Science*, The MIT Press, Cambridge, Massachusetts, pp. 501-526, 1989.

[94] Soergel, D. *Organizing Information: Principles of Database and Retrieval Systems.* Orlando, FL: Academic Press. 1985.

[95] Sowa, J.F. *Conceptual Structures, Information Processing in Mind and Machine*, Addison-Wesley, Reading, Massachusetts, 1984.

[96] http://spamassassin.apache.org/.

[97] http://spambayes.sourceforge.net/.

[98] Steel, D. Bayesian confirmation theory and the likelihood principle. Synthese-dordrecht, 156, pp. 53-77, 2007.

[99] Talbot, W. Bayesian Epistemology. Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/archives/fall2001/entries/epistemology-bayesian, 2001.

[100] Tan, P.N., Steinbach, M., and Kumar, V. *Introduction to Data Mining.* Pearson Addison Wesley, 2006.

[101] Triola, M.F. *Elementary Statistics*, Addison Wesley. 2005.

[102] Tsumoto, S. Accuracy and coverage in rough set rule induction. Proceedings of RSCTC'02, LNAI 2475, pp. 373-380, 2002.

[103] van Mechelen, I., Hampton, J., Michalski, R.S., and Theuns, P. (Eds.), *Categories and Concepts, Theoretical Views and Inductive Data Analysis*, Academic Press, New York, 1993.

[104] van Rijsbergen, C. J. *Information Retrieval*. Butterworths, London, United Kingdom, 1979.

[105] Wald, A. Sequential tests of statistical hypotheses. The Annals of Mathematical Statistics, Vol. 16, pp. 117-186, 1945.

[106] Weller, A.C. Editorial Peer Review: Its Strengths and Weaknesses. Information Today, Inc., Medford, NJ. 2001.

[107] Wong, S.K.M., and Ziarko, W. A probabilistic model of approximate classification and decision rules with uncertainty in inductive learning, Technical Report CS-85-23, Department of Computer Science, University of Regina, 1985.

[108] Wong, S.K.M., and Ziarko, W. Algorithm for inductive learning, Bulletin of the Polish Academy of Science Technical Science. Vol. 34, pp. 271-276, 1986.

[109] Woodward, P.W., and Naylor, J.C. An application of Bayesian methods in SPC. The Statistician, Vol. 42, pp. 461-469, 1993.

[110] Yang, Y.M., and Pedersen, J.O. A comparative study on feature selection in text categorization. Proceedings of ICML-97, 14th International conference on Machine Learning. 1997.

[111] Yerazunis, W.S. Sparse binary polynomial hashing and the CRM114 discriminator. In Proceedings of the MIT Spam Conference, 2003.

[112] Yih, W. McCann, R., and Kolcz, A. Improving spam filtering by detecting gray mail. In Proceedings of the 4th Conference on E-mail and Anti-Spam (CEAS07), 2007.

[113] Yao, Y.Y. Probabilistic approaches to rough sets, Expert Systems, Vol. 20, pp. 287-297, 2003.

[114] Yao, Y.Y. Granular computing, *Computer Science (Ji Suan Ji Ke Xue)*, 31, pp. 1-5,2004.

[115] Yao, Y.Y. A note on definability and approximations. Transactions on Rough Sets VII, LNCS 4400, pp. 274-282, 2007.

[116] Yao, Y.Y. Decision-theoretic rough set models. Proceedings of RSKT'07, LNAI 4481, pp. 1-12, 2007.

[117] Yao, Y.Y., The art of granular computing, Proceedings of the International Conference on Rough Sets and Emerging Intelligent Systems Paradigms, LNAI 4585, pp. 101-112, 2007.

[118] Yao, Y.Y. Probabilistic rough set approximations, International Journal of Approximation Reasoning, Vol. 49, pp. 255-271, 2008.

[119] Yao, Y.Y. Three-way decision: an interpretation of rules in rough set theory. Proceedings of RSKT'09, LNAI 5589, pp. 642-649, 2009.

[120] Yao, Y.Y., Interpreting concept learning in cognitive informatics and granular computing, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 39, No. 4, pp. 855-866, 2009.

[121] Yao, Y.Y. Three-way decisions with probabilistic rough sets. Information Sciences, Vol. 180, pp. 341-353, 2010.

[122] Yao, Y.Y. The superiority of three-way decisions in probabilistic rough set models. Information Sciences, Vol. 181, pp. 1080-1096, 2011.

[123] Yao, Y.Y., An Outline of a Theory of Three-way Decisions, Manuscript.

[124] Yao, Y.Y., and Wong, S.K.M. A decision theoretic framework for approximating concepts. International Journal of Man-machine Studies, Vol. 37, pp. 793-809, 1992.

[125] Yao, Y.Y., Wong, S.K.M., and Lingras, P. A decision-theoretic rough set model. Methodologies for Intelligent Systems 5, Z.W. Ras, M. Zemankova and M.L. Emrich (Eds.), New York, North-Holland, pp. 17-24, 1990.

[126] Yao, Y.Y., Zhong, N. An analysis of quantitative measures associated with rules. Proceedings of PAKDD99, LNAI 1974, pp. 479-488, 1999.

[127] Yao, Y.Y., Zhao, Y. Attribute reduction in decision-theoretic rough set models, Information Sciences, 178(17), pp. 3356-3373, Elsevier B.V., 2008.

[128] Yao, Y.Y., Zhou, B., and Chen Y.H. Interpreting low and high order rules: a granular computing approach. Proceedings of International Conference on

Rough Sets and Emerging Intelligent System Paradigms (RSEISP'07), LNAI 4585, pp. 371-380, 2007.

[129] Yao, Y.Y., Zhou, B. A logic language of granular computing. Proceedings of 6th IEEE International Conference on Cognitive Informatics (ICCI07), pp. 178-185, 2007.

[130] Yao, Y.Y., and Zhou, B. Micro and macro evaluation of classification rules. Proceedings of the 7th IEEE International Conference on Cognitive Informatics (ICCI'08), pp. 441-448, 2008.

[131] Yao, Y.Y., Zhou, B. Naive Bayesian Rough Sets. RSKT'10, pp. 719-726, 2010.

[132] Yao, Y.Y., Zhou, B. Bayesian rough set models. European Journal of Operational Research, under second round review (EJOR-D-11-00505).

[133] Zadeh, L.A. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems, 19, pp. 111-127, 1997.

[134] Zhang, H. Exploring conditions for the optimality of naive Bayes. International Journal of Pattern Recognition and Artificial Intelligence, Vol. 19, 2005.

[135] Zhang, X.H., Zhou, B., and Li, P. A general frame for intuitionistic fuzzy rough sets. Information Science, to appear.

[136] Zhao, W., and Zhang, Z. An email classification model based on rough set theory. In Proceedings of the International Conference on Active Media Technology, pp. 403-408, 2005.

[137] Zhou, B., and Yao, Y.Y. A logic approach to granular computing. The International Journal of Cognitive Informatics & Natural Intelligence (IJCiNi). 2(2), pp. 63-79, 2007.

[138] Zhou, B., and Yao, Y.Y. Evaluating information retrieval performance based on multi-grade relevance. The 17th International Symposium on Methodologies for Intelligent Systems (ISMIS08). LNAI 4994, pp. 424-433, 2008.

[139] Zhou, B., and Yao, Y.Y. Unifying rough set analysis and formal concept analysis based on a logic approach to granular computing. Discoveries and Breakthroughs in Cognitive Informatics and Natural Intelligence, pp. 325-349, 2009.

[140] Zhou, B., and Yao, Y.Y. Evaluating information retrieval system performance based on user preference. Journal of Intelligent Information Systems (JIIS). 34(3), pp. 227-248, 2010.

[141] Zhou, B., and Yao, Y.Y. In search for effective granularity with DTRS. Proceedings of the 9th IEEE International Conference on Cognitive Informatics (ICCI'10), pp. 464-470, 2010.

[142] Zhou, B., Yao, Y.Y., and Luo, J.G. A three-way decision approach to email spam filtering. Proceedings of the 23th Canadian Conference on Artificial Intelligence (AI 2010). LNAI 6085, pp. 28-39, 2010.

[143] Zhou, B. A new formulation of multi-category decision-theoretic rough sets. The Sixth International Conference on Rough Sets and Knowledge Technology (RSKT'11), LNAI 6954, pp. 514-522, 2011.

[144] Zhou, B., and Yao, Y.Y. In search for effective granularity with DTRS for Ternary Classification. The International Journal of Cognitive Informatics & Natural Intelligence (IJCiNi), 5(3), pp. 47-60, 2011.

[145] Zhou, Z.H., and Liu, X.Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 18(1), pp. 63-77, 2006.

[146] Zhou, Z.H., and Liu, X.Y. On multi-class cost-sensitive learning. Computational Intelligence, 26(3), pp. 232-257, 2010.

[147] Ziarko, W. Variable precision rough sets model. Journal of Computer and Systems Sciences, Vol. 46, pp. 39-59, 1993.

[148] Ziarko, W. Set approximation quality measures in the variable precision rough set model. Proceedings of the 2nd International Conference on Hybrid Intelligent Systems (HIS'02), pp. 442-452, 2002.