Naive Bayesian Rough Sets*

Yiyu Yao and Bing Zhou

Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2
{yyao,zhou200b}@cs.uregina.ca

Abstract. A naive Bayesian classifier is a probabilistic classifier based on Bayesian decision theory with naive independence assumptions, which is often used for ranking or constructing a binary classifier. The theory of rough sets provides a ternary classification method by approximating a set into positive, negative and boundary regions based on an equivalence relation on the universe. In this paper, we propose a naive Bayesian decision-theoretic rough set model, or simply a naive Bayesian rough set (NBRS) model, to integrate these two classification techniques. The conditional probability is estimated based on the Bayes' theorem and the naive probabilistic independence assumption. A discriminant function is defined as a monotonically increasing function of the conditional probability, which leads to analytical and computational simplifications.

Key words: three-way decisions, naive Bayesian classification, Bayesian decision theory, cost-sensitive classification

1 Introduction

Naive Bayesian classifier and rough set classification are two useful techniques for classification problems. A naive Bayesian classifier is a probabilistic classifier based on Bayesian decision theory with naive independence assumptions [1, 2]. As a fundamental statistical approach, Bayesian decision theory is often used for binary classification problems, i.e., each class is associated with a yes/no decision. The Pawlak rough set theory provides a ternary classification method by approximating a set by positive, negative and boundary regions based on an equivalence relation of the universe [7, 16].

The qualitative categorization of Pawlak three regions may be too restrictive to be practically useful. This has led to the extension of rough sets by allowing some tolerance of uncertainty. Probabilistic rough set models were proposed [3, 5, 8, 10–12, 14, 17–19], in which the degrees of overlap between equivalence classes and a set to be approximated are considered. A conditional probability is used to state the degree of overlapping and a pair of threshold values α and β are used to defined three probabilistic regions. Elements whose probability is above the first threshold α are put into the positive region, between α and the second threshold

* Yiyu Yao and Bing Zhou, Naive Bayesian Rough Sets, Proceedings of RSKT 2010, LNAI 6401, pp. 719-726, 2010.

β in the boundary region, and below β is the negative region. The three regions correspond to a three-way decision of acceptance, deferment, and rejection [16]. The decision-theoretic rough set (DTRS) model provides a systematic way to calculate the two threshold values based on the well established Bayesian decision theory, with the aid of more practically operable notions such as cost, risk, benefit etc. [14, 17, 18].

On the other hand, the estimation of the conditional probability has not received much attention. The rough membership function is perhaps the only commonly discussed way [9]. It is necessary to consider other methods for estimating the probability more accurately. For this purpose, we introduce a naive Bayesian decision-theoretic rough set model, or simply a naive Bayesian rough set (NBRS) model. The conditional probability is estimated based on the Bayes' theorem and the naive probabilistic independence assumption. A discriminant function is defined as a monotonically increasing function of the conditional probability, which leads to analytical and computational simplifications.

2 Contributions of the Naive Bayesian Rough Set Model

The proposed naive Bayesian rough set model is related to several existing studies, but contributes in its unique way. In the Bayesian decision theory, one may identify three important components, namely, the interpretation and computation of the required threshold value when constructing a classifier, the use of Bayes' theorem that connects, based on the likelihood, the *a priori* probability of a class to the *a posteriori* probability of the class after observing a piece of evidence, and the estimation of required probabilities. These three components enable us to show clearly the current status of various probabilistic models of rough sets and the contributions of the naive Bayesian rough set model.

The decision-theoretic rough set model [14, 15, 17, 18] focuses on the first issue, namely, the interpretation and computation of a pair of threshold values on the *a posteriori* probability of class for building a ternary classifier. The later proposed variable precision rough set (VPRS) model [19] uses a pair of threshold values on a measure of set-inclusion to define rough set approximations, which is indeed equivalent to the result of a special case of the DTRS model [14, 20]. The more recent parameterized rough set model [3] uses a pair of thresholds on a Bayesian confirmation measure, in addition to a pair thresholds on probability. In contrast to the DTRS model, the last two models suffers from a lack of guidelines and systematic methods on how to determining the required threshold values.

The Bayesian rough set (BRM) model [11, 12] is an attempt to resolve the above problem by using the *a priori* probability of the class as a threshold for defining probabilistic regions, i.e., one compares the *a posteriori* probability and the *a priori* probability of the class. Based on the Bayes' theorem, one can show that this is equivalent to comparing two likelihoods [12]. The rough Bayesian (RM) model [10] further explores the second issue of the Bayesian decision theory. A pair of threshold values on a Bayes factor, namely, a likelihood ratio, is used to define probabilistic regions. The Bayesian rough set model, in

fact, uses a threshold of 0 on the difference between the *a posteriori* and the *a priori* probabilities, or a threshold of 1 on the likelihood ration; the rough Bayesian model uses a pair of arbitrary threshold values. However, the latter model does not address the problem of how to setting the threshold values. Recently, the Bayes' theorem is introduced into the decision-theoretic rough set model to address this problem [16].

All these probabilistic models do not address the third issue of the Bayesian decision theory, namely, the estimation of the required probabilities. The full implications of Bayesian decision theory and Bayesian inference have not been fully explored, even though the phrases, rough Bayesian model and Bayesian rough sets, have been used. In this paper, we propose a Bayesian decision-theoretic rough set model, or simply a Bayesian rough set model, to cover all three issues of the Bayesian decision theory, and a naive Bayesian rough set model, in particular, to adopt the naive independence assumption in probability estimation. Since the first issue, namely, interpretation and computation of the thresholds, has been extensively discussed in other papers [14–17], we will concentrate on the contributions of the naive Bayesian rough sets with respect to the other two issues, namely, application of Bayes' theorem and probability estimation.

3 Basic Formulation of Bayesian Rough Sets

We review the basic formulations of probabilistic rough set and Bayesian rough set models in the following subsections.

3.1 Decision-theoretic rough sets

Let $E \subseteq U \times U$ be an equivalence relation on U, i.e., E is reflexive, symmetric, and transitive. Two objects in U satisfy E if and only if they have the same values on all attributes. The pair $apr = (U, E)$ is called an approximation space. The equivalence relation E induces a partition of U, denoted by U/E. The basic building blocks of rough set theory are the equivalence classes of E. For an object $x \in U$, the equivalence class containing x is given by $[x] = \{y \in U \mid xEy\}$. For a subset $C \subseteq U$, one can divide the universe U into three disjoint regions, the positive region $\mathrm{POS}(C)$, the boundary region $\mathrm{BND}(C)$, and the negative region $\mathrm{NEG}(C)$ [6]:

$$\begin{aligned} \mathrm{POS}(C) &= \{x \in U \mid [x] \subseteq C\}, \\ \mathrm{BND}(C) &= \{x \in U \mid [x] \cap C \neq \emptyset \wedge [x] \not\subseteq C\}, \\ \mathrm{NEG}(C) &= \{x \in U \mid [x] \cap C = \emptyset\}. \end{aligned} \tag{1}$$

One can say with *certainty* that any object $x \in \mathrm{POS}(C)$ belongs to C, and that any object $x \in \mathrm{NEG}(C)$ does not belong to C. One cannot decide with certainty whether or not an object $x \in \mathrm{BND}(C)$ belongs to C.

The qualitative categorization in the Pawlak rough set model may be too restrictive to be practically useful. Probabilistic rough set model is proposed to

enable some tolerance of uncertainty, in which the Pawlak rough set model is generalized by considering degrees of overlap between equivalence classes and a set to be approximated, i.e., $[x]$ and C in equation (1),

$$Pr(C|[x]) = \frac{|C \cap [x]|}{|[x]|}, \tag{2}$$

where $|\cdot|$ denotes the cardinality of a set, and $Pr(C|[x])$ is the conditional probability of an object belongs to C given that the object is in $[x]$, estimated by using the cardinalities of sets. Pawlak and Skowron [9] suggested to call the conditional probability a rough membership function. According to the above definitions, the three regions can be equivalently defined by:

$$\begin{aligned} \text{POS}(C) &= \{x \in U \mid Pr(C|[x]) = 1\}, \\ \text{BND}(C) &= \{x \in U \mid 0 < Pr(C|[x]) < 1\}, \\ \text{NEG}(C) &= \{x \in U \mid Pr(C|[x]) = 0\}. \end{aligned} \tag{3}$$

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 $Pr(C|[x])$ is not taken into account.

A main result of decision-theoretic rough set model is parameterized probabilistic approximations. This can be done by replacing the values 1 and 0 in equation (3) by a pair of threshold values α and β with $\alpha > \beta$. The (α, β)-probabilistic positive, boundary and negative regions are defined by:

$$\begin{aligned} \text{POS}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C|[x]) \geq \alpha\}, \\ \text{BND}_{(\alpha,\beta)}(C) &= \{x \in U \mid \beta < Pr(C|[x]) < \alpha\}, \\ \text{NEG}_{(\alpha,\beta)}(C) &= \{x \in U \mid Pr(C|[x]) \leq \beta\}. \end{aligned} \tag{4}$$

The three probabilistic regions lead to three-way decisions [16]. We accept an object x to be a member of C if the probability is greater than α. We reject x to be a member of C if the probability is less than β. We neither accept or reject x to be a member of C if the probability is in between of α and β, instead, we make a decision of deferment.

The threshold values α and β can be interpreted in terms of cost or risk of the three-way classification. They can be systematically computed based on minimizing the overall risk of classification. The details can be found in papers on decision-theoretic rough sets [14, 15, 17, 18].

3.2 Classification based on Bayes' theorem

The conditional 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,

$$Pr(C|[x]) = \frac{Pr(C)Pr([x]|C)}{Pr([x])}, \tag{5}$$

where

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

$Pr(C|[x])$ is the *a posteriori* probability of class C given $[x]$, $Pr(C)$ is the *a priori* probability of class C, and $Pr([x]|C)$ the likelihood of $[x]$ with respect to C. The Bayes' theorem enable us to infer the *a posteriori* probability $Pr(C|[x])$, which is difficulty to estimate, from the *a priori* probability $Pr(C)$ through the likelihood $Pr([x]|C)$, which is easy to estimate.

One may define monotonically increasing functions of the conditional probability to construct an equivalent classifier. This observation can lead to significant analytical and computational simplifications.The probability $Pr([x])$ in equation (5) can be eliminated by taking the odds form of Bayes' theorem, that is,

$$O(Pr(C|[x])) = \frac{Pr(C|[x])}{Pr(C^c|[x])} = \frac{Pr([x]|C)}{Pr([x]|C^c)} \cdot \frac{Pr(C)}{Pr(C^c)} = \frac{Pr([x]|C)}{Pr([x]|C^c)} O(Pr(C)). \tag{6}$$

A threshold value on the probability can indeed be interpreted as another threshold value on the odds. For the positive region, we have:

$$\begin{aligned} Pr(C|[x]) \geq \alpha &\Longleftrightarrow \frac{Pr(C|[x])}{Pr(C^c|[x])} \geq \frac{\alpha}{1-\alpha} \\ &\Longleftrightarrow \frac{Pr([x]|C)}{Pr([x]|C^c)} \cdot \frac{Pr(C)}{Pr(C^c)} \geq \frac{\alpha}{1-\alpha}. \end{aligned} \tag{7}$$

By applying logarithms to both sides of the equation, we get

$$\log \frac{Pr([x]|C)}{Pr([x]|C^c)} + \log \frac{Pr(C)}{Pr(C^c)} \geq \log \frac{\alpha}{1-\alpha}. \tag{8}$$

Similar expressions can be obtained for the negative and boundary regions. Thus, the three regions can now be written as:

$$\begin{aligned} \mathrm{POS}^B_{(\alpha',\beta')}(C) &= \{x \in U \mid \log \frac{Pr([x]|C)}{Pr([x]|C^c)} \geq \alpha'\}, \\ \mathrm{BND}^B_{(\alpha',\beta')}(C) &= \{x \in U \mid \beta' < \log \frac{Pr([x]|C)}{Pr([x]|C^c)} < \alpha'\}, \\ \mathrm{NEG}^B_{(\alpha',\beta')}(C) &= \{x \in U \mid \log \frac{Pr([x]|C)}{Pr([x]|C^c)} \leq \beta'\}, \end{aligned} \tag{9}$$

where

$$\begin{aligned} \alpha' &= \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\alpha}{1-\alpha}, \\ \beta' &= \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\beta}{1-\beta}. \end{aligned} \tag{10}$$

This interpretation simplifies the calculation by eliminating $Pr([x])$. The detailed estimations of related probabilities need to be further addressed.

3.3 Naive Bayesian model for estimating probabilities

The naive Bayesian rough set model provides a practical way to estimate the conditional probability based on the naive Bayesian classification [1, 2]. In the Pawlak rough set model [7], information about a set of objects are represented in an information table with a finite set of attributes [6]. 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 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, $I_a(x) \in V_a$. Each object x is described by a logic formula $\bigwedge_{a \in At} a = I_a(x)$, where $v_a \in V_a$, and the atomic formula $a = I_a(x)$ indicates that the value of an object on attribute a is $I_a(x)$. For simplicity, we express the description of $[x]$ as a feature vector, namely, $Des([x]) = (v_1, v_2, ..., v_n)$ with respect to the set of attributes $\{a_1, a_2, ..., a_n\}$ where $I_{a_i}(x) = v_i$. For simplicity, we write $Des([x])$ as $[x]$.

Recall that the conditional probability $Pr(C|[x])$ can be reexpressed by the *prior* probability $Pr(C)$, the *likelihood* $Pr([x]|C)$, and the probability $Pr([x])$, where $Pr([x]|C)$ is a joint probabilities of $Pr(v_1, v_2, ..., v_n|C)$, and $Pr([x])$ is a joint probability of $Pr(v_1, v_2, ..., v_n)$. 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 [2]. That is, we assume each component v_i of $[x]$ to be conditionally independent of every other component v_j for $j \neq i$.

For the Bayesian interpretation of three regions based on equation (8), we can add the following naive conditional independence assumptions:

$$\begin{aligned} Pr([x]|C) &= Pr(v_1, v_2, ..., v_n|C) = \prod_{i=1}^{n} Pr(v_i|C), \\ Pr([x]|C^c) &= Pr(v_1, v_2, ..., v_n|C^c) = \prod_{i=1}^{n} Pr(v_i|C^c). \end{aligned} \tag{11}$$

Thus, equation (7) can be re-expressed as:

$$\begin{aligned} &\log \frac{Pr([x]|C)}{Pr([x]|C^c)} \geq \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\alpha}{1-\alpha} \\ \Longleftrightarrow &\sum_{i=1}^{n} \log \frac{Pr(v_i|C)}{Pr(v_i|C^c)} \geq \log \frac{Pr(C^c)}{Pr(C)} + \log \frac{\alpha}{1-\alpha}. \end{aligned} \tag{12}$$

where $Pr(C)$ and $Pr(v_i|C)$ can be easily estimated from the frequencies of the training data by putting:

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

where $m(a_i, v_i)$ is called the meaning set. It is defined as $m(a_i, v_i) = \{x \in U | I_{a_i}(x) = v_i\}$, that is, the set of objects whose attribute value equal to v_i with regard to attribute a_i. Similarly, we can estimate $Pr(C^c)$ and $Pr(v_i|C^c)$. We can then rewrite equation (8) as:

$$\begin{aligned}
\mathrm{POS}^B_{(\alpha',\beta')}(C) &= \{x \in U \mid \sum_{i=1}^{n} \log \frac{Pr(v_i|C)}{Pr(v_i|C^c)} \geq \alpha'\}, \\
\mathrm{BND}^B_{(\alpha',\beta')}(C) &= \{x \in U \mid \beta' < \sum_{i=1}^{n} \log \frac{Pr(v_i|C)}{Pr(v_i|C^c)} < \alpha'\}, \\
\mathrm{NEG}^B_{(\alpha',\beta')}(C) &= \{x \in U \mid \sum_{i=1}^{n} \log \frac{Pr(v_i|C)}{Pr(v_i|C^c)} \leq \beta'\}. \qquad (13)
\end{aligned}$$

All the related factors in the above equations are easily derivable from data for real applications.

4 Conclusion

This paper proposes a naive Bayesian rough set model to intergrade two classification techniques, namely, naive Bayesian classifier and the theory of rough sets. The conditional probability in the definition three regions in rough sets is interpreted by using the probability terms in naive Bayesian classification. A discriminant function is defined as a monotonically increasing function of the conditional probability, which leads to analytical and computational simplifications. The integration provides a practical solution for applying naive Bayesian classifier to ternary classification problems. Two threshold values instead of one are used, which can be systematically calculated based on loss functions stating how costly each action is.

Acknowledgements

The first author is partially supported by an NSERC Canada Discovery grant. The second author is supported by an NSERC Alexander Graham Bell Canada Graduate Scholarship.

References

1. Duda, R.O. and Hart, P.E. *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.
2. Good, I.J.,The Estimation of Probabilities: An Essay on Modern Bayesian Methods. M.I.T. Press. 1965.
3. Greco, S., Matarazzo, B. and Słowiński, R. Parameterized rough set model using rough membership and Bayesian confirmation measures, *International Journal of Approximate Reasoning*, **49**, 285-300, 2009.
4. Herbert, J.P. and Yao, J.T. Game-theoretic risk analysis in decision-theoretic rough sets, *Proceedings of RSKT'08*, LNAI 5009, 132-139, 2008.
5. Herbert, J.P. and Yao, J.T. Game-theoretic rough sets, *Fundamenta Informaticae*, 2009.
6. Pawlak, Z. Rough sets, *International Journal of Computer and Information Sciences*, **11**, 341-356, 1982.
7. Pawlak, Z. *Rough Sets, Theoretical Aspects of Reasoning about Data*, Dordrecht: Kluwer Academic Publishers, 1991.
8. Pawlak, Z., Wong, S.K.M. and Ziarko, W. Rough sets: probabilistic versus deterministic approach, *International Journal of Man-Machine Studies*, **29**, 81-95, 1988.
9. 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, 251-271, 1994.
10. Ślęzak, D. Rough sets and Bayes factor, *LNCS Transactions on Rough Sets III*, LNCS 3400, 202-229, 2005.
11. Ślęzak, D., Ziarko, W., Bayesian rough set model. In: Procedings of FDM'2002. December 9, Maebashi, Japan. pp. 131-135. 2002.
12. Ślęzak, D. and Ziarko, W. The investigation of the Bayesian rough set model, *International Journal of Approximate Reasoning*, **40**, 81-91, 2005.
13. Yao, Y.Y. Probabilistic approaches to rough sets, *Expert Systems*, **20**, 287-297, 2003.
14. Yao, Y.Y. Decision-theoretic rough set models, *Proceedings of RSKT 2007*, LNAI 4481, 1-12, 2007.
15. Yao, Y.Y. Probabilistic rough set approximations, *International Journal of Approximation Reasoning*, **49**, 255-271, 2008.
16. Yao, Y.Y. Three-way decisions with probabilistic rough sets, Information Sciences, Vol. 180, No. 3, pp. 341-353, 2010.
17. Yao, Y.Y. and Wong, S.K.M. A decision theoretic framework for approximating concepts, *International Journal of Man-machine Studies*, **37**, 793-809, 1992.
18. Yao, Y.Y., Wong, S.K.M. and Lingras, P. A decision-theoretic rough set model, in: *Methodologies for Intelligent Systems 5*, Z.W. Ras, M. Zemankova and M.L. Emrich (Eds.), New York, North-Holland, 17-24, 1990.
19. Ziarko, W. Variable precision rough sets model, *Journal of Computer and Systems Sciences*, **46**, 39-59, 1993.
20. Ziarko, W. Probabilistic approach to rough sets, *International Journal of Approximate Reasoning*, **49**, 272-284, 2008.