

# A Bayesian Classifier for Uncertain Data

Biao Qin, Yuni Xia  
Department of Computer Science  
Indiana University - Purdue University  
Indianapolis, USA  
{biaoqin, yxia}@cs.iupui.edu

Fang Li  
Department of Mathematical Sciences  
Indiana University - Purdue University  
Indianapolis, USA  
fli@math.iupui.edu

## ABSTRACT

Data uncertainty is widespread in a variety of applications. This paper proposes a new Bayesian classification algorithm for classifying uncertain data. In the paper, we apply probability and statistics theory on uncertain data model, and provide solutions for model parameter estimation for both uncertain numerical data and uncertain categorical data. We also prove the correctness of the solutions. The experimental results demonstrate the proposed uncertain Bayesian classifier can be efficiently constructed, and it significantly outperforms the traditional Bayesian classifier in prediction accuracy when data is highly uncertain.

## 1. INTRODUCTION

In many applications, data contains inherent uncertainty. A number of factors contribute to the uncertainty, such as the random nature of the physical data generation and collection process, measurement and decision errors, unreliable data transmission and data staling. In this paper, we focus on Bayesian classification for uncertain data. Data classification is one of the most important data mining problem. Bayesian classification algorithm is tremendously appealing because of its simplicity, elegance, and robustness. It is one of the oldest formal classification algorithms, and it is often surprisingly effective. A large number of modifications have been introduced by the statistical, data mining, machine learning, and pattern recognition communities in an attempt to make it more flexible [4]. It is widely used in various areas including text classification and spam filtering.

In the paper, we extend the Bayesian algorithm to classify and predict uncertain data. We integrate the uncertain data model with Bayesian classifier and extend the traditional Bayesian classification algorithm so that it can process highly uncertain data. For both uncertain numerical data and uncertain categorical data, we present solutions for classification model parameter estimation. We also extend the traditional Bayesian prediction algorithm to predict data class based on uncertain attributes. We show through

experiments that the proposed NBU classifier can be efficiently generated, and it has a distinctly higher prediction accuracy than the traditional Bayesian classifier on uncertain data.

Classification is a well-studied area in data mining. Many classification algorithms have been proposed in the literature, such as decision tree classifiers [9], rule-based classifiers [3], Bayesian classifiers, support vector machines (SVM), artificial neural networks and ensemble methods. In spite of numerous classification algorithms, building classification based on uncertain data remains a great challenge. There is early work performed on developing decision trees when data contains missing or noisy values [8, 7, 5]. Various strategies have been developed to predict or fill missing attribute values. However, the problem studied in this paper is different from before - instead of assuming part of the data has missing or noisy values, we allow the whole dataset to be uncertain, and the uncertainty is not shown as missing or erroneous values, but represented as uncertain intervals with probability distribution functions [2] or  $x$ -tuples [10].

## 2. THE UNCERTAIN DATA MODELS

In this section, we will discuss the uncertain data models for both numerical and categorical attribute, which are the most common types of data encountered in data mining applications.

When the value of a numerical attribute is uncertain, the attribute is called an uncertain numerical attribute (UNA), denoted by  $A_i^{un}$ . Further, we use  $A_{ij}^{un}$  to denote the  $j$ th instance of  $A_i^{un}$ . The concept of UNA has been introduced in [2]. The value of  $A_i^{un}$  is represented as a range or interval and an optional probability distribution function (PDF) over this range. Note that  $A_i^{un}$  is treated as a continuous random variable. The PDF  $f(x)$  can be related to an attribute if all instances have the same distribution, or related to each instance if each instance has a different distribution.

An uncertain interval of  $A_i^{un}$ , denoted by  $A_i^{un}.U$ , is an interval  $[A_i^{un}.l, A_i^{un}.r]$  where  $A_i^{un}.l, A_i^{un}.r \in \mathcal{R}$  and  $A_i^{un}.r \geq A_i^{un}.l$ . An uncertain probability distribution function (PDF) of  $A_{ij}^{un}$ , denoted by  $A_{ij}^{un}.f(x)$ , is a PDF of  $A_{ij}^{un}$ , such that  $\int_{A_{ij}^{un}.l}^{A_{ij}^{un}.r} A_{ij}^{un}.f(x)dx = 1$  and  $\int_{A_{ij}^{un}.l}^{A_{ij}^{un}.r} A_{ij}^{un}.f(x)dx = 0$  if  $x \notin A_{ij}^{un}.U$ .  $\square$

The uncertain  $x$ -tuple model concept has been proposed in

database systems such as [10]. Each  $x$ -tuple  $T_j$  includes a number of *items* as its alternatives which are associated with probabilities, representing a discrete probability distribution of these alternatives being selected. Independence is assumed among the  $x$ -tuples.

Given a categorical domain  $Dom = \{v_1, \dots, v_n\}$ , an uncertain categorical attribute (UCA)  $A_i^{uc}$  is characterized by probability distribution over  $Dom$ . It can be represented by the probability vector  $\mathbf{P} = \{p_{j1}, \dots, p_{jn}\}$  such that  $P(A_i^{uc} = v_k) = p_{jk}$  and  $\sum_{k=1}^n p_{jk} = 1$  ( $1 \leq k \leq n$ ).

### 3. BACKGROUND

Our classifier is developed based on the naive Bayes classifier. We will first briefly explain the naive Bayes classifier. The probability model for a classifier is a conditional model  $p(C|A_1, \dots, A_n)$  over a dependent class variable  $C$  with a small number of outcomes or classes, conditional on several feature variables  $A_1$  through  $A_n$ . Using Bayes' theorem,  $p(C|A_1, \dots, A_n) = \frac{p(C) p(A_1, \dots, A_n|C)}{p(A_1, \dots, A_n)}$ . In another word, the above equation can be written as posterior =  $\frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$ . In practice we are only interested in the numerator of that fraction, since the denominator does not depend on  $C$  and the values of the features  $A_i$  are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model  $p(C, A_1, \dots, A_n)$  which can be rewritten as follows, using repeated applications of the definition of conditional probability:

$$\begin{aligned} p(C, A_1, \dots, A_n) &= p(C) p(A_1, \dots, A_n|C) \\ &= p(C) p(A_1|C) p(A_2, \dots, A_n|C, A_1) \\ &= p(C) p(A_1|C) p(A_2|C, A_1) p(A_3, \dots, A_n|C, A_1, A_2) \end{aligned}$$

and so forth. The naive Bayes classifier assumes that each feature  $A_i$  is conditionally independent of every other feature  $A_j$  for  $j \neq i$ . This means that  $p(A_i|C, A_j) = p(A_i|C)$ . Therefore, the joint model can be expressed as

$$\begin{aligned} p(C, A_1, \dots, A_n) &= p(C) p(A_1|C) p(A_2|C) p(A_3|C) \dots \\ &= p(C) \prod_{i=1}^n p(A_i|C). \end{aligned}$$

All model parameters (such as class priors and feature probability distributions) can be approximated with relative frequencies from the training set. For example, in order to estimate  $p(A_i|C)$ , we often assume  $p(A_i|C)$  follows a Gaussian distribution  $N(\mu, \sigma^2)$ , and we can compute the mean ( $\mu$ ) and variance ( $\sigma$ ) of  $p(A_i|C)$  based on the training data. Therefore, for a new test instance, it is easy to estimate  $p(A_i|C)$  according to the probability density function.

Data uncertainty bring unique challenges to the parameter estimation. Traditional maximum likelihood based parameter estimation needs to be extended to handle data uncertainty. In the next section, we will present our approaches for parameter estimation for both uncertain numerical data

and uncertain categorical data.

## 4. PARAMETER ESTIMATION

### 4.1 Uncertain numerical Attributes

As described earlier, the value of an uncertain numerical attribute is an interval with an optional associated PDF. Each uncertain value has a maximal value and a minimal value. For each uncertain numerical data, the observed value is in the form of  $[L_i, R_i]$ . Suppose the true value is  $X_i$ , then  $L_i \leq X_i \leq R_i$ , as shown in Figure 1.

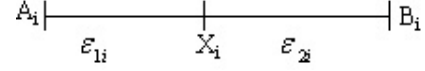


Figure 1: An example uncertain data interval

We denote the left error (negative error) with  $\epsilon_{iL}$ , the right error (positive error) with  $\epsilon_{iR}$ , and the overall error with  $\epsilon_i$ . We assume the left error is a positive distribution with mean  $\mu$  and variance  $\sigma_L^2$ , that is,  $\epsilon_{iL} \sim (\mu, \sigma_L^2)$ , and the right error is a positive distribution with mean  $\mu$  and variance  $\sigma_R^2$ ,  $\epsilon_{iR} \sim (\mu, \sigma_R^2)$ , and the overall error is a Gaussian distribution with mean 0 and variance  $\sigma^2$ ,  $\epsilon \sim N(0, \sigma^2)$ . Here we focus on statistically random error, which is the most common type of error in practice; therefore we assume that the left error and the right error have the same mean  $\mu$ , and the overall error has a mean 0. If the error is biased, the assumptions and computations can be easily adjusted.

Below, we give a theorem for Gaussian distribution parameters estimation based on uncertain numerical data.

**THEOREM 1.** *Assume an uncertain numerical attribute  $X_i$  satisfies Gaussian distribution. Let  $L$  and  $R$  be the random variable denoting the minimal and maximal values of the sample interval;  $\epsilon_{iL}$ ,  $\epsilon_{iR}$  and  $\epsilon$  denote the left error (negative error), the right error (positive error) and the overall error, respectively. Assume  $\epsilon_{iL} \sim (\mu, \sigma_L^2)$ ,  $\epsilon_{iR} \sim (\mu, \sigma_R^2)$ ,  $\epsilon \sim N(0, \sigma^2)$ ,  $\epsilon_{iL}$  and  $\epsilon_{iR}$  are independent, then  $X_i$  satisfies the distribution of*

$$P(X_i|(L, R)) \sim N\left(\frac{\mu_L + \mu_R}{2}, \frac{\sigma_{(L+R)}^2 - \sigma_{(R-L)}^2}{4}\right).$$

**PROOF.** Assume  $X_i = \mu_i + \epsilon_i$ , hereby, we need to estimate  $\mu_i$  and  $\epsilon_i$ . As shown in Figure 1,

$$\begin{aligned} L_i &= X_i - \epsilon_{iL} \\ R_i &= X_i + \epsilon_{iR} \end{aligned}$$

From the above equations, we know

$$\begin{aligned} L_i &= \mu_i + \epsilon_i - \epsilon_{iL} \\ R_i &= \mu_i + \epsilon_i + \epsilon_{iR} \\ R_i + L_i &= 2\mu_i + 2\epsilon_i + \epsilon_{iR} - \epsilon_{iL} \\ R_i - L_i &= \epsilon_{iL} + \epsilon_{iR} \end{aligned}$$

According to the assumption,  $\epsilon_{iL}$  and  $\epsilon_{iR}$  are both random variables.  $\epsilon_{iL} \sim (\mu, \sigma_L^2)$ ,  $\epsilon_{iR} \sim (\mu, \sigma_R^2)$ ,  $\epsilon_{iL}$  and  $\epsilon_{iR}$  are

independent. Since  $R_i - L_i = \epsilon_{iL} + \epsilon_{iR}$ , the variance of  $R_i - L_i$ , denoted as  $\sigma_{(R-L)}^2$ , should be

$$\begin{aligned}\sigma_{(R-L)}^2 &= \sigma_{\epsilon_{iL} + \epsilon_{iR}}^2 \\ &= \sigma_{\epsilon_{iL}}^2 + \sigma_{\epsilon_{iR}}^2 \\ &= \sigma_L^2 + \sigma_R^2\end{aligned}$$

Since  $L_i + R_i = 2\mu_i + 2\epsilon_i + \epsilon_{iR} - \epsilon_{iL}$ , here  $2\mu_i$  is a constant whose variance is 0, the variance of  $R_i + L_i$ , denoted as  $\sigma_{(R+L)}^2$ , should be

$$\begin{aligned}\sigma_{(R+L)}^2 &= \sigma_{2\epsilon_i + \epsilon_{iR} - \epsilon_{iL}}^2 \\ &= \sigma_{2\epsilon_i}^2 + \sigma_{\epsilon_{iR}}^2 + (-1)^2 \sigma_{\epsilon_{iL}}^2 \\ &= (2\sigma)^2 + \sigma_R^2 + \sigma_L^2 \\ &= 4\sigma^2 + \sigma_R^2 + \sigma_L^2\end{aligned}$$

Therefore,  $\sigma_{(R+L)}^2 - \sigma_{(R-L)}^2 = 4\sigma^2$ , and

$$\sigma^2 = \frac{\sigma_{(L+R)}^2 - \sigma_{(R-L)}^2}{4} \quad (1)$$

Further, since  $L_i + R_i = 2\mu_i + 2\epsilon_i + \epsilon_{iR} - \epsilon_{iL}$ , therefore,  $E(L_i + R_i) = E(2\mu_i + 2\epsilon_i + \epsilon_{iR} - \epsilon_{iL}) = E(2\mu_i) + E(2\epsilon_i) + E(\epsilon_{iR}) - E(\epsilon_{iL}) = 2\mu_i + 0 + \mu - \mu = 2\mu_i$ , from which we obtain:

$$\mu_i = \frac{E(L + R)}{2} = \frac{\mu_{L+R}}{2} = \frac{\mu_L + \mu_R}{2}$$

Therefore the distribution of  $X_i$  can be estimated as

$$P(X_i|(L, R)) \sim N\left(\frac{\mu_L + \mu_R}{2}, \frac{\sigma_{(L+R)}^2 - \sigma_{(R-L)}^2}{4}\right). \quad \square$$

This shows that for uncertain numerical data which are represented as intervals  $[L, R]$ , the mean can be estimated as  $\frac{\mu_L + \mu_R}{2}$  and the variance can be estimated as  $\frac{\sigma_{(L+R)}^2 - \sigma_{(R-L)}^2}{4}$ . Please note that this approach also applies to certain data. When a data instance is certain, it is a point instead of an interval; therefore, the minimal boundary is equal to its maximal boundary, that is,  $L = R = X$ . Therefore, the mean of the whole dataset is

$$\frac{\mu_L + \mu_R}{2} = \frac{\mu_X + \mu_X}{2} = \mu_X.$$

The variance of the dataset is

$$\begin{aligned}\frac{\sigma_{(L+R)}^2 - \sigma_{(R-L)}^2}{4} &= \frac{\sigma_{(2L)}^2 - \sigma_{(0)}^2}{4} = \frac{\sigma_{(2L)}^2}{4} \\ &= \frac{4\sigma_{(L)}^2}{4} = \sigma_{(L)}^2 = \sigma_{(X)}^2.\end{aligned}$$

That shows that for certain data, the mean is estimated to be  $\mu_X$  and the variance is estimated to be  $\sigma_{(X)}^2$ , which is consistent with the Naive Bayes classification algorithm. Therefore,

$$P(X|(L, R)) \sim N\left(\frac{\mu_L + \mu_R}{2}, \frac{\sigma_{(L+R)}^2 - \sigma_{(R-L)}^2}{4}\right)$$

is a general form for mean and variance estimation. It applies to both uncertain and certain numerical data. Actually, certain data can be treated as a special case of uncertain data with zero uncertainty. When data has zero uncer-

tainty, this process automatically evolves to the traditional Bayesian classification algorithm.

## 4.2 Uncertain categorical attributes

An uncertain categorical attribute (UCA)  $A_i^{uc}$  is characterized by probability distribution over its domain  $Dom$ . As mentioned earlier, it can be represented by the probability vector  $\mathbf{P} = \{p_{j1}, \dots, p_{jn}\}$  such that  $P(A_{ij}^{uc} = v_k) = p_{jk}$  ( $1 \leq i \leq n$ ).

Before discussing parameter estimation for uncertain categorical data, we first introduce the concept of probabilistic cardinality. The probabilistic cardinality for class  $C_k$  of the dataset is the sum of the probability of each instance  $T_j$  belonging to class  $C_k$ . That is,  $PC(C_k) = \sum_{j=1}^{|D|} P(C_{T_j} = C_k)$ , where  $C_{T_j}$  denotes the class label of instance  $T_j$ . Similarly, the probabilistic cardinality of the dataset over  $v_k$  of attribute  $A_i^{uc}$  is the sum of the probability of each instance whose corresponding UCA equals  $v_k$ . That is,  $PC(v_k) = \sum_{j=1}^{|D|} P(v_k \in T_j) = \sum_{j=1}^{|D|} p_{jk}$ . The probabilistic cardinality for class  $C_i$  of the dataset over  $v_k$  of attribute  $A_i^{uc}$  is the sum of the probability of each instance in class  $C_i$  whose corresponding UCA equals to  $v_k$ . That is,  $PC(v_k, C_i) = \sum_{j=1}^{|D|} P(v_k \in T_j \wedge C_{T_j} = C_i)$ .

The class distribution of each value of uncertain categorical attributes can be denoted by a vector, which we call the Class Distribution Vector (CDV).  $CDV(v_j, C)$  is  $(PC(v_j, C_1), PC(v_j, C_2), \dots, PC(v_j, C_n))^T$ , in which  $PC(v_j, C_i)$  is the probabilistic cardinality of instances in class  $C_i$  with attribute value  $v_j$ .

For an uncertain categorical attribute instance  $A_{ij}^{uc}$ , the conditional probability  $P(A_{ij}^{uc} = v_k | C_i)$  is estimated according to the fraction of the probabilistic cardinality of instances in class  $C_i$  that takes on a particular attribute value  $v_k$  over the total probabilistic cardinality of instances in class  $C_i$ , that is:

$$P(A_{ij}^{uc} = v_k | C_i) = \frac{PC(v_k, C_i)}{PC(C_i)}.$$

$P(A_{ij}^{uc})$  gives the uncertain categorical data distribution for each class, which will be used for Bayesian prediction.

## 5. ALGORITHM DESCRIPTION

In this section, we will present our uncertain Bayesian classification algorithm, which is shown in Algorithm 1. The principle procedure is as follows: 1. For each uncertain numerical attribute instance  $A_{ij}^{un}$ , we update the Gaussian distribution parameters  $\mu_{A_i}^+$ ,  $\sigma_{A_i}^+$ ,  $\mu_{A_i}^-$  and  $\sigma_{A_i}^-$  according to  $A_{ij}.R$ ,  $A_{ij}.L$  by Function `updateGaussian()` as shown in step3-5.

2. For each uncertain categorical attribute instance  $A_{ij}^{uc}$ , we update its CDV by the weight of the instance  $T_j.w$  and  $p_{jk}$  using Function `updateHistogram()` (steps 6 - 9).

3. For each numerical attribute instance  $A_{ij}^{un}$ , we update the Gaussian distribution parameter  $(\mu, \sigma)$  with Function `updateGaussian()` (steps 10-11).

4. For each categorical attribute instance  $A_{ij}^{uc}$ , we update

its CDV by the weight of the instance  $T_j.w$  using Function `updateHistogram()` (steps 12-13).

5. For each instance  $T_j$ , we update the Probabilistic Cardinality of its class  $PC(T_j.class)$  by the weight of the instance  $T_j.w$  using Function `updateProbabilisticCardinality()` (steps 16).

6. Finally, we compute the mean  $\mu$  and standard deviation  $\sigma$  for each uncertain numerical attribute  $A_i^{un}$  using Theorem 1 (steps 18 - 21).

---

**Algorithm 1** NBU(Dataset  $D$ )

---

```

begin
1: for (Each instance  $T_j \in D$  do) do
2:   for (each attribute  $A_i$  do) do
3:     if ( $A_i$  is uncertain numerical) then
4:        $(\mu_{A_i}^+, \sigma_{A_i}^+) = \text{updateGaussian}(A_{ij}.R + A_{ij}.L,$ 
          $T_j.w)$ ;
5:        $(\mu_{A_i}^-, \sigma_{A_i}^-) = \text{updateGaussian}(A_{ij}.R - A_{ij}.L,$ 
          $T_j.w)$ ;
6:     else if ( $A_i$  is uncertain categorical) then
7:       for (each  $v_k \in A_i$ ) do
8:          $PC(v_k, T_j.Class) = \text{updateHistogram}(T_j.w * p_{jk})$ ;
9:       end for;
10:    else if ( $A_i$  is numerical) then
11:       $(\mu, \sigma) = \text{updateGaussian}(A_{ij}, T_j.w)$ ;
12:    else if ( $A_i$  is categorical) then
13:       $PC(v_k, T_j.Class) = \text{updateHistogram}(T_j.w)$ ;
14:    end if;
15:  end for;
16:   $PC(T_j.Class) = \text{updateProbabilisticCardinality}(T_j.w)$ ;
17: end for;
18: for (each uncertain numerical attribute  $A_i$ ) do
19:    $\mu_{A_i} = (\mu_{A_i}^+ + \mu_{A_i}^-)/2$ ;
20:    $\sigma_{A_i} = \sqrt{(\sigma_{A_i}^+)^2 - (\sigma_{A_i}^-)^2}/2$ ;
21: end for;
end

```

---

An important benefit of Bayesian classification is that it is incremental, which means that model can evolve gradually when more training data become available. Many other classification methods, on the contrary, require the whole classification model to be rebuilt from scratch with newly added training data. For example, the decision tree is essentially non-incremental, with more training data, the splitting point and tree structure can be completely different and it is better to rebuild it. Please note that our NBU algorithm preserve the incremental feature. For uncertain data,  $\mu_{A_i}$ ,  $\sigma_{A_i}$  and  $PC(v_k, T_j.class)$  can all be incrementally updated. This is very important in data stream application where new data constantly become available and the classification model should be continuously adjusted.

## 6. EXPERIMENTS

In this section, we present the experimental results of the proposed Uncertain Bayesian Classifier algorithm. We implemented Uncertain Bayesian Classifier classification and predication algorithm. All the experiments presented in this section are executed on a PC with an Intel Pentium IV 3.2 GHz CPU and 2.0 GB main memory. A collection containing

10 real-world benchmark datasets were assembled from the UCI Repository [1]. We try to cover the spectrum of properties such as size, attribute numbers and types, number of classes and class distributions. Among these 10 datasets, 5 of them, namely Diabetes, Glass, Iris, Segment and Sonar, contain mainly numerical attributes. The other 5 datasets, namely Balance, Bridge, Mushroom, Promote and Soybean, have mostly categorical attributes.

**Data Uncertainty.** Data are made uncertain in the following way: 1. To make numerical attributes uncertain, we convert each numerical value to an uncertain interval. For each numerical attribute, we scan all of its value and get its maximum value  $X_{max}$  and minimum value  $X_{min}$ , respectively. For each attribute instance  $x$ , its uncertain interval is  $[x - (x - X_{min}) * rand1, x + (X_{max} - x) * rand2]$ , where  $rand1$  and  $rand2$  denote two the random numbers. If they range between 0 to X, we denote such dataset as UX. For example, U0.25 stands for the dataset generated with  $rand1$  and  $rand2$  between 0 to 0.25.

2. We make categorical attributes uncertain by converting them into probability vectors. For example, a categorical attribute  $L_i$  may have  $k$  possible values  $v_m (1 \leq m \leq k)$ . For each attribute  $A_{ij}$ , we first convert it into a probability vector  $P = (p_{j1}, p_{j2}, \dots, p_{ji}, \dots, p_{jk})$ , while  $p_{jl}$  is the probability for  $A_{ij}^{uc}$  to be equal to  $v_l$ , that is,  $P(A_{ij}^{uc} = v_l) = p_{jl}$ . If the original value of  $A_{ij}$  is equal to  $v_l$ , we set  $p_{jl}$  to be a value less than 1, and evenly distribute the rest probability  $1 - p_{jl}$  to all other values, that is,  $\sum_{k=1 \wedge k \neq j}^n p_{kl} = 1 - p_{jl}$ . We randomly select part of instances, for example 25% of them, and convert them uncertain, then we call such dataset U0.25. If all instances in a categorical dataset are made uncertain, it is represented as U1.00.

We use U0 to denote accurate or certain original datasets. When Uncertain Bayesian Classifier works on certain datasets U0, it is the same as the traditional Bayesian classifier.

We study the accuracy of Uncertain Bayesian Classifier algorithm. In our experiments, all attributes in the datasets are made uncertain except the ID attribute. We compare the proposed Uncertain Bayesian Classifier with the traditional Bayesian algorithm. Since traditional Bayesian classification algorithm cannot work uncertain data intervals directly, we convert uncertain datasets to certain. For uncertain numerical data, we use the center point of the uncertain interval; for uncertain categorical data, we choose the value with the highest probability. After this conversion, an uncertain dataset becomes a regular certain dataset and the traditional Bayesian classification algorithm can be applied on it. We examine the Prediction Accuracy Ratio(PAR), which is the accuracy of the Uncertain Bayesian Classifier classifier on uncertain data to the accuracy of the traditional Bayesian classifier on the corresponding converted certain data.

Figure 2 shows the experimental results on uncertain numerical datasets. For each dataset, we generate four uncertain datasets with different degree of uncertainty. The uncertain datasets are represented as U0.25, U0.5, U0.75 and U1.00, with uncertainty increasing. From this figure, we found that the Uncertain Bayesian Classifier classifier almost always has the same or higher accuracy than the regular Bayesian clas-

sifier when data is uncertain. On the Iris, Segment and Sonar dataset, Uncertain Bayesian Classifier clearly outperforms traditional Bayesian classifier. Furthermore, the difference in accuracy between Uncertain Bayesian Classifier and traditional Bayesian grows as data become more uncertain. For the U0.5 datasets, the performance gain of Uncertain Bayesian Classifier is mostly within 5%, for the U0.75 and U1.00 datasets, the accuracy gain of Uncertain Bayesian Classifier reaches over 10% or even 15%. The reason is that traditional Bayesian classifier works with the center points of uncertain intervals and compute the data distribution parameters solely based on the center points, and ignore other valuable information. Uncertain Bayesian Classifier, on the other hand, use a more sophisticated model which considers not only the centers, but also the left and right boundaries and the interval length. It enables more precise data distribution parameter estimation, which lead to more systematic modeling and more accurate prediction.

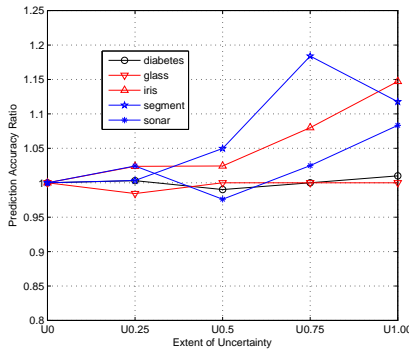


Figure 2: PAR on numerical datasets

We also studied the Uncertain Bayesian Classifier classifier accuracy on uncertain categorical data. The result is shown in Figure 3. Here, we still investigate the prediction accuracy ratio of Uncertain Bayesian Classifier over traditional Bayesian as uncertainty varies. The trend is similar to the results on uncertain numerical data. The benefit of Uncertain Bayesian Classifier is more distinct in this experiment. For the U0.50 datasets, the accuracy improvements are mostly within 15%. When uncertainty increases to U0.75, for 3 out of the 5 datasets (balance, bridge and promote datasets), the performance gain reaches 40-50%. For all the five U1.00 datasets, Uncertain Bayesian Classifier is 1.9 to 5 times better than traditional Bayesian. On the soybean dataset, Uncertain Bayesian Classifier has a slightly worse performance than Bayesian when the uncertainty is low, as uncertainty increase it starts to show its advantage. For the U0.75 dataset, it outperforms the traditional Bayesian, and on the U1.00 dataset, its performance is 5 times better. This is because the traditional Bayesian classifier work on certain datasets, which only keep the value of the highest probability for categorical attribute, while probability distribution in other values are neglected. Uncertain Bayesian Classifier utilizes all the probability distribution information available and builds more accurate classifier model.

## 7. CONCLUSIONS

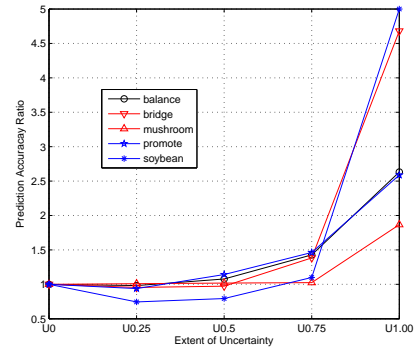


Figure 3: PAR on uncertain categorical datasets

In this paper, we propose an uncertain Bayesian algorithm for classifying and predicting uncertain datasets. Uncertain data are extensively present in modern applications including sensor network, moving object databases and biological databases. Instead of trying to eliminate uncertainty and noise from datasets, this paper follows the new paradigm of directly mining uncertain data. We integrate the uncertain data model with Bayesian theorem and propose new methods for model parameter estimation. The new methods allow us to derive more precise model based on uncertain data and attain higher prediction accuracy. Our experimental evaluation demonstrates that Uncertain Bayesian Classifier achieves higher prediction accuracy comparing to the traditional Bayesian classifier when working on uncertain data.

## 8. REFERENCES

- [1] <http://archive.ics.uci.edu/ml/datasets.html>.
- [2] R. Cheng, D. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In *SIGMOD 2003*, pages 551–562.
- [3] W. W. Cohen. Fast effective rule induction. In *Proceedings of the 12th International conference on machine learning*, pages 115–123.
- [4] X. W. et al. Top 10 algorithms in data mining. *Journal Knowledge and Information Systems*, 14(1):1–37, Jan. 2008.
- [5] L. Hawarah, A. Simonet, and M. Simonet. Dealing with missing values in a probabilistic decision tree during classification. In *The Second International Workshop on Mining Complex Data*, pages 325–329, 2006.
- [6] O. Lobo and M. Numao. Ordered estimation of missing values. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 499–503, 1999.
- [7] J. R. Quinlan. *Probabilistic decision trees, in Machine Learning: an Artificial Intelligence Approach*. Morgan Kaufmann Publishers Inc. San Francisco, 1990.
- [8] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman Publishers, 1993.
- [9] J. Widom. Trio: A system for integrated management of data, accuracy, and lineage. In *ICDR 2005*, pages 262–276.