

A Novel Bayesian Classification Technique for Uncertain Data

Biao Qin^{1,2}*, Yuni Xia², Shan Wang¹, Xiaoyong Du¹

1. Dept. of Computer Science, Renmin University of China, P. R. China, 100872

2. Dept. of Computer and Information Science, Indiana University - Purdue University Indianapolis 46202

Abstract

Data uncertainty can be caused by numerous factors such as measurement precision limitations, network latency, data staleness and sampling errors. When mining knowledge from emerging applications such as sensor networks or location based services, data uncertainty should be handled cautiously to avoid erroneous results. In this paper, we apply probabilistic and statistical theory on uncertain data and develop a novel method to calculate conditional probabilities of Bayes theorem. Based on that, we propose a novel Bayesian classification algorithm for uncertain data. The experimental results show that the proposed method classifies uncertain data with potentially higher accuracies than the Naive Bayesian approach. It also has a more stable performance than the existing extended Naive Bayesian method.

Key words: Bayes theorem; Uncertain data; Classification

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

* corresponding author. Work was done when Qin was a postdoc at IUPUI

Email addresses: bqin@cs.iupui.edu (Biao Qin^{1,2}), yxia@cs.iupui.edu (Yuni Xia²), swang@ruc.edu.cn (Shan Wang¹), duyong@ruc.edu.cn (Xiaoyong Du¹).

generation and collection process, measurement and decision errors, unreliable data transmission and data staling. For example, in location based services, moving objects of interest are attached with locators and the information is periodically updated and streamed to the control center. However, those location data are typically inaccurate due to locator energy and precision limitation, network bandwidth constraint and latency. There are also massive uncertain data in sensor networks such as temperature, humidity and pressure.

When mining knowledge from these applications, data uncertainty needs to be handled with caution. Otherwise, unreliable or even wrong mining results would be obtained. In this paper, we focus on Naive Bayesian classification for uncertain data. Naive Bayesian classification is tremendously appealing because of its simplicity, elegance, and robustness. It is one of the oldest formal classifications, 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 [30]. It is widely used in areas such as text classification and spam filtering. Based on Naive Bayesian classification, we propose a novel method to directly classify and predict uncertain data in this paper. The main contributions of this paper are:

- Based on a new method to calculate conditional probabilities of Bayes theory, we extend Naive Bayesian classification so that it can process uncertain data.
- We prove through extensive experiments that the proposed classifier can be efficiently generated and it can classify uncertain data with potentially higher accuracies than Naive Bayesian classifier. Furthermore, the proposed classifier is more suitable for mining uncertain data than the previous work [24].

This paper is organized as follows. In the next section, we discuss related work. Section 3 introduces basic concepts of Naive Bayesian classification. Section 4 describes the techniques to calculate conditional probabilities for uncertain numerical data sets. Section 5 describes the Bayesian algorithm for uncertain data and its prediction. The experimental results are shown in Section 6. Section 7 concludes the paper.

2 Related Work

Uncertain data, also called symbolic data [5,14], has been studied for many years. Many works focus on clustering [8,7,17,21,12]. The key idea is that when computing the distance between two uncertain objects, the probability distributions of objects are used to calculate the expected distance. In [12], Cormode et al. show reductions to their corresponding weighted versions on data with uncertainties. In [31], Xia et al. introduce a new conceptual clustering algorithm for uncertain categorical data.

Classification is a well-studied area in data mining. Many methods have been proposed in the literature, such as decision tree [26], rule-based classifications [11], Bayesian classifications [18] and so on. In spite of the numerous methods, 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 [25,20]. Various strategies have been developed to predict or fill missing attribute values, for example, Dayanik [13] presented feature interval learning algorithms which represent multi-concept descriptions in the form of disjoint feature intervals. 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 [9].

Recently, Tsang et al [28] and Qin et al [22] independently developed decision tree classifications for uncertain data. Both adopt the technique of fractional tuple for splitting tuples into subsets when the domain of its PDF spans across the cut point. Tsang et al [28] converted every numerical value into a set of s sample points between the uncertain interval $[a_j, b_j]$ with the associated value $f(x)$, effectively approximating every $f(x)$ by a discrete distribution. Qin et al. also proposed a rule-based classification [23]. The key problem in learning rules is to efficiently identify the optimal cut points from training data. For uncertain numerical data, an optimization mechanism is proposed to merge adjacent bins which have equal classifying class distribution. In our earlier work, we proposed a Bayesian classification method for uncertain data[24]. It will be compared with our new approach in this paper.

There is also some research on OLAP computation on uncertain data [6], and on identifying frequent item sets and association mining [3,4,10,15,32] from uncertain data sets. In these work, the support of itemsets and confidence of association rules are integrated with the existential probability of transactions and items.

3 Background

The naive Bayesian classifier estimates the class-conditional probability by assuming that the attributes are conditionally independent, given the class label C_k . Suppose that there are n classes, C_1, C_2, \dots, C_n , the conditional independence assumption can be formally stated as follows:

$$P(X|C_k) = \prod_{i=1}^m P(X_i|C_k)$$

where every attribute set $X = \{X_1, X_2, \dots, X_m\}$ consists of m attributes.

With the conditional independent assumption, instead of computing the class-conditional probability for every combination of X , we only have to estimate the conditional probability of every X_i given C_k . The latter approach is more practical because it does not require a very large training set to obtain a good estimation of the probability [27]. To classify a test record, the naive Bayesian classifier computes the posterior probability for every class C_k as follows:

$$P(C_k|X) = \frac{P(C_k)\prod_{i=1}^m P(X_i|C_k)}{P(X)} \quad (1)$$

Since $P(X)$ is fixed for every C_k , it is sufficient to choose the class that maximizes the numerator term, $P(C_k)\prod_{i=1}^m P(X_i|C_k)$.

A common assumption, not intrinsic to the naive Bayesian classification but often made nevertheless, is within every class, the values of numeric attributes are normally distributed. One can represent such a distribution in terms of its mean and variance, and one can efficiently compute the probability of an observed value from such estimations. For continuous attributes we have

$$P(x_i \leq X_i \leq x_i + \Delta | Y = y_i) = \int_{x_i}^{x_i + \Delta} f(x; \mu, \sigma^2) dx$$

$$\approx f(x; \mu, \sigma^2) \times \Delta$$

where Δ is a small constant.

Table 1

Training Set for predicting borrowers who will default on loan payments

RowID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	110-120	No
2	No	Married	100-120	No
3	No	Single	60-85	No
4	Yes	Married	110-145	No
5	No	Divorced	110-120	Yes
6	No	Married	50-80	No
7	Yes	Divorced	170-250	No
8	No	Single	85-100	Yes
9	No	Married	80-100	No
10	No	Single	120-145	Yes
11	No	Divorced	105-125	Yes
12	No	Divorced	80-95	No

4 Conditional probabilities for uncertain numerical attributes

In this section, we describe the uncertain data model and the new approach for calculating conditional probabilities for uncertain numerical data. In this paper, we focus on the uncertainty in attributes and assume the class type is certain.

4.1 A model for uncertain numerical data

When the value of a numerical attribute is uncertain, the attribute is called an uncertain numerical attribute (UNA) [9]. We use A_{ij} to denote the j th instance of A_i . The value of A_i is represented as a range or interval [19,33,34] and the probability distribution function (PDF) over this range [23]. Note that A_i 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 every instance if every instance has different distributions.

Table 1 shows an example of UNA. The data in this table are used to predict whether borrowers will default on loan payments. Among all the attributes, the Annual Income is a UNA, whose precise value is not available. We only know the range of the Annual Income of every person and the PDF $f(x)$ over that range. The probability distribution function of the UNA attribute Annual Income is assumed to be normal and the PDF $f(x)$ is not shown in Table 1. For an uncertain numerical attribute value $[A_{ij}.a, A_{ij}.b]$, μ_j can be estimated as $(a+b)/2$ since there are no biased error; and σ_j can be estimated as $(b-a)/6$ since the probability that a value falls within the mean plus minus 3 standard deviation is more than 99% for a normal random variable [16].

Definition 1 An uncertain interval of A_{ij} , denoted by $A_{ij}.U$, is an interval $[A_{ij}.a, A_{ij}.b]$ where $A_{ij}.b \geq A_{ij}.a$.

Definition 2 An uncertain PDF of A_{ij} is denoted by $A_{ij}.f(x)$, such that $\int_{A_{ij}.a}^{A_{ij}.b} A_{ij}.f(x)dx = 1$, $\int_{-\infty}^{A_{ij}.a} A_{ij}.f(x)dx = 0$ and $\int_{A_{ij}.b}^{\infty} A_{ij}.f(x)dx = 0$.

Definition 3 Assume an interval random variable Z has m instances and every instance can be observed with equal probability $1/m$, the empirical density function of Z is

$$f(x) = \frac{1}{m} \sum_{j=1}^m f_j(x) \quad (2)$$

where $f_j(x)$ denotes the density function of the j th instance of Z .

4.2 Conditional probabilities of Bayes theorem

If an attribute A is numerical, it is often assumed to have a Gaussian distribution for naive Bayesian classification. If an attribute A is uncertain numerical, we can also assume it has a Gaussian distribution, i.e.,

$$P(A|C_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \quad (3)$$

The following sections discuss two methods to calculate μ_k and σ_k^2 for uncertain numerical data sets.

4.2.1 The proposed Bayes theorem for uncertain data

We assume that an interval-valued random variable Z is satisfied with Gaussian distribution and every instance z_j is also satisfied with Gaussian distribution in its own interval. The following theorem discusses the way to calculate the parameters of Gaussian distribution of the interval-based random variable Z .

Theorem 1 *For an interval-valued random variable Z , its observed values are $[a_j, b_j]$ for $j = 1, \dots, m$. Let A and B be random variables denoting the minimal and maximal values of the sample interval; then Z satisfies the distribution of*

$$P(Z|(A, B)) \sim N(\hat{\mu}_{\frac{A+B}{2}}, \hat{S}_{\frac{A+B}{2}}^2 + \Delta)$$

where $\Delta = \frac{1}{m} \sum_{j=1}^m \left(\frac{b_j - a_j}{6}\right)^2$.

Proof. We recall that the empirical mean $\hat{\mu}$ in terms of the empirical density function is

$$\hat{\mu} = \int_{-\infty}^{\infty} x f(x) dx$$

Substituting from Equation (2), we have

$$\begin{aligned}
\hat{\mu} &= \int_{-\infty}^{\infty} x \frac{1}{m} \sum_{j=1}^m f_j(x) dx \\
&= \frac{1}{m} \sum_{j=1}^m \int_{a_j}^{b_j} \frac{x}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \mu_j)^2 / \sigma_j^2\right) dx
\end{aligned}$$

We can use two well-known equations $\int_{-\infty}^{\infty} \exp(-\frac{1}{2}t^2) dt = \sqrt{2\pi}$ and $\int_{-\infty}^{\infty} x \cdot \exp(-\frac{1}{2}t^2) dt = 0$ to calculate the above integral. On substituting $t = (x - \mu_j) / \sigma_j$, we have $dx = \sigma_j dt$. Thus

$$\begin{aligned}
\hat{\mu} &= \frac{1}{m} \sum_{j=1}^m \frac{1}{\sigma_j \sqrt{2\pi}} \int_{a_j}^{b_j} \sigma_j (\mu_j + \sigma_j t) \exp\left(-\frac{1}{2}t^2\right) dt \\
&= \frac{1}{m \sqrt{2\pi}} \sum_{j=1}^m \left(\mu_j \int_{a_j}^{b_j} \exp\left(-\frac{1}{2}t^2\right) dt + \sigma_j \int_{a_j}^{b_j} t \exp\left(-\frac{1}{2}t^2\right) dt \right) \\
&= \frac{1}{m \sqrt{2\pi}} \sum_{j=1}^m (\mu_j \sqrt{2\pi} + 0) \\
&= \frac{1}{m} \sum_{j=1}^m \mu_j \\
&= \frac{1}{m} \sum_{j=1}^m \frac{a_j + b_j}{2} \\
&= \hat{\mu}_{\frac{A+B}{2}}
\end{aligned}$$

Similarly, we can derive the sample variance.

$$\begin{aligned}
\hat{S}^2 &= E(Z^2) - (E(Z))^2 \\
&= \int_{-\infty}^{\infty} x^2 \frac{1}{m} \sum_{j=1}^m f_j(x) dx - \left(\frac{1}{m} \sum_{j=1}^m \frac{a_j + b_j}{2} \right)^2 \\
&= \frac{1}{m} \sum_{j=1}^m \int_{a_j}^{b_j} x^2 f_j(x) dx - \left(\frac{1}{m} \sum_{j=1}^m \frac{a_j + b_j}{2} \right)^2
\end{aligned}$$

We first derive $y = \int_{a_j}^{b_j} x^2 f_j(x) dx$. Using, once again, the substitution $t = (x - \mu_j) / \sigma_j$.

$$\begin{aligned}
y &= \int_{a_j}^{b_j} x^2 f_j(x) dx \\
&= \frac{1}{\sigma_j \sqrt{2\pi}} \int_{a_j}^{b_j} \sigma_j (\mu_j + \sigma_j t)^2 \exp(-\frac{1}{2}t^2) dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{a_j}^{b_j} (\mu_j^2 + 2\mu_j \sigma_j t + (\sigma_j t)^2) \exp(-\frac{1}{2}t^2) dt \\
&= \frac{1}{\sqrt{2\pi}} \left(\mu_j^2 \int_{a_j}^{b_j} \exp(-\frac{1}{2}t^2) dt + 2\mu_j \sigma_j * \right. \\
&\quad \left. \int_{a_j}^{b_j} t * \exp(-\frac{1}{2}t^2) dt + \sigma_j^2 \int_{a_j}^{b_j} t^2 * \exp(-\frac{1}{2}t^2) dt \right) \\
&= \frac{1}{\sqrt{2\pi}} (\mu_j^2 \sqrt{2\pi} + 2\mu_j \sigma_j * 0 + \sigma_j^2 \sqrt{2\pi}) \\
&= \mu_j^2 + \sigma_j^2
\end{aligned}$$

Hence,

$$\begin{aligned}
\hat{S}^2 &= \frac{1}{m} \sum_{j=1}^m (\mu_j^2 + \sigma_j^2) - \left(\frac{1}{m} \sum_{j=1}^m \frac{a_j + b_j}{2} \right)^2 \\
&= \frac{1}{m} \sum_{j=1}^m \left(\left(\frac{a_j + b_j}{2} \right)^2 + \left(\frac{b_j - a_j}{6} \right)^2 \right) - \left(\frac{1}{m} \sum_{j=1}^m \frac{a_j + b_j}{2} \right)^2 \\
&= \frac{1}{m} \sum_{j=1}^m \left(\frac{a_j + b_j}{2} \right)^2 - \left(\frac{1}{m} \sum_{j=1}^m \frac{a_j + b_j}{2} \right)^2 + \frac{1}{m} \sum_{j=1}^m \left(\frac{b_j - a_j}{6} \right)^2 \\
&= \hat{S}_{\frac{A+B}{2}}^2 + \frac{1}{m} \sum_{j=1}^m \left(\frac{b_j - a_j}{6} \right)^2 \\
&= \hat{S}_{\frac{A+B}{2}}^2 + \Delta
\end{aligned}$$

Thus the theorem is proved. \square

This approach can also be applied to certain data. When a data instance is certain, it is a point instead of an interval; therefore, $A = B = Z$. Consequently, the mean of the whole data set is $\hat{\mu} = \hat{\mu}_{\frac{A+B}{2}} = \hat{\mu}_{\frac{Z+Z}{2}} = \mu_Z$. The variance of the dataset is

$$\begin{aligned}\hat{S}^2 &= \hat{S}_{\frac{A+B}{2}}^2 + \frac{1}{m} \sum_{j=1}^m \left(\frac{b_j - a_j}{6}\right)^2 \\ &= \hat{S}_{\frac{A+B}{2}}^2 = \hat{S}_Z^2\end{aligned}$$

This shows that for certain data, the mean is estimated to be $\hat{\mu}_Z$ and the variance is estimated to be \hat{S}_Z^2 , which is consistent with naive Bayesian classification. Therefore, $\hat{\mu} = \hat{\mu}_{\frac{A+B}{2}}$ and $\hat{S}^2 = \hat{S}_{\frac{A+B}{2}}^2 + \frac{1}{m} \sum_{j=1}^m \left(\frac{b_j - a_j}{6}\right)^2$ are general forms of mean and variance estimation for both uncertain and certain numerical data. Actually, certain data can be treated as a special case of uncertain data which has zero uncertainty. When data has zero uncertainty, its process automatically evolves to naive Bayesian classification.

Example 1 Refer to the data shown in Table 1. For the uncertain numerical attribute Annual Income, according to Theorem 1, the mean and variance with respect to class **No** are as follows.

$$\begin{aligned}\hat{u} &= \hat{u}_{\frac{A+B}{2}} \\ &= \frac{1}{8} \left(\frac{110 + 120}{2} + \frac{100 + 120}{2} + \frac{60 + 85}{2} + \frac{110 + 145}{2} \right. \\ &\quad \left. + \frac{50 + 80}{2} + \frac{170 + 250}{2} + \frac{80 + 100}{2} + \frac{80 + 95}{2} \right) \\ &= 109.7\end{aligned}$$

$$\begin{aligned}\hat{S}^2 &= \frac{1}{m} \sum_{j=1}^m \left(\frac{a_j + b_j}{2}\right)^2 - \left(\frac{1}{m} \sum_{j=1}^m \frac{a_j + b_j}{2}\right)^2 + \frac{1}{m} \sum_{j=1}^m \left(\frac{b_j - a_j}{6}\right)^2 \\ &= 13864.84 - 12031.35 + 35.67 \\ &= 1869.2\end{aligned}$$

4.2.2 The previous Bayes theorem for uncertain data

For every uncertain variable Z , its observed value is in the form of $[a_j, b_j]$. Suppose its true value is z_j , then $a_j \leq z_j \leq b_j$ as shown in Figure 1. We denote the left error (negative error) with ϵ_{1j} , the right error (positive error) with ϵ_{2j} , and the overall error with ϵ_j . Since common errors such as measurement imprecision can usually be approximated by Gaussian distributions, we assume the left error is a Gaussian distribution with mean μ and variance σ_1^2 , that is, $\epsilon_{1j} \sim (\mu, \sigma_1^2)$. Similarly, we assume the right error is a Gaussian

distribution with mean μ and variance σ_2^2 , $\epsilon_{2j} \sim (\mu, \sigma_2^2)$, and the overall error is a Gaussian distribution with mean 0 and variance σ^2 , $\epsilon \sim (0, \sigma^2)$. 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 μ , and the overall error has a mean 0. If the error is statistically negatively or positively biased, the assumptions and computations can be easily adjusted. The following theorem discusses the way to calculate the parameters of Gaussian distribution of the error-based random variable Z .

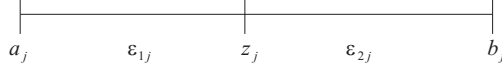


Fig. 1. An example uncertain data interval

Theorem 2 [24] *Assume an uncertain numerical attribute Z satisfies Gaussian distribution. Let A and B be the random variable denoting the minimal and maximal values of the sample interval respectively; ϵ_{1j} , ϵ_{2j} and ϵ denote the left error (negative error), the right error (positive error) and the overall error respectively. Assume $\epsilon_{1j} \sim (\mu, \sigma_1^2)$, $\epsilon_{2j} \sim (\mu, \sigma_2^2)$, $\epsilon \sim (0, \sigma^2)$, ϵ_{1j} and ϵ_{2j} are independent, then Z satisfies the distribution of*

$$P(Z|(A, B)) \sim N(\hat{\mu}_{\frac{A+B}{2}}, \hat{S}_{\frac{A+B}{2}}^2 - \Delta)$$

where $\Delta = \hat{S}_{\frac{B-A}{2}}^2$.

Proof. Because Z satisfies Gaussian distribution, assume $z_j = \mu_j + \epsilon_j$. Hereby, we need to estimate μ_j and ϵ_j . As shown in Figure 1, we have

$$\begin{aligned} a_j &= z_j - \epsilon_{1j} \\ b_j &= z_j + \epsilon_{2j} \end{aligned}$$

From the above equations, we get

$$\begin{aligned} a_j &= \mu_j + \epsilon_j - \epsilon_{1j} \\ b_j &= \mu_j + \epsilon_j + \epsilon_{2j} \\ b_j + a_j &= 2\mu_j + 2\epsilon_j + \epsilon_{2j} - \epsilon_{1j} \\ b_j - a_j &= \epsilon_{1j} + \epsilon_{2j} \end{aligned}$$

Gaussian distribution has the property that for two independent Gaussian random variables, their sum and difference are also normally distributed. That is, if $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are independent Gaussian random variables, then their sum is normally distributed with $X + Y \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ and their difference is normally distributed with $X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$. According to our assumption, ϵ_{1j} and ϵ_{2j} are independent Gaussian random variables. Since $b_j - a_j = \epsilon_{1j} + \epsilon_{2j}$, the variance of $b_j - a_j$, denoted as \hat{S}_{B-A}^2 , should be

$$\begin{aligned}\hat{S}_{B-A}^2 &= \hat{S}_{\epsilon_{1j} + \epsilon_{2j}}^2 \\ &= \hat{S}_{\epsilon_{1j}}^2 + \hat{S}_{\epsilon_{2j}}^2 \\ &= \sigma_1^2 + \sigma_2^2\end{aligned}$$

Since $a_j + b_j = 2\mu_j + 2\epsilon_j + \epsilon_{2j} - \epsilon_{1j}$, $2\mu_j$ is a constant whose variance is 0, ϵ_j , ϵ_{1j} and ϵ_{2j} are all Gaussian random variables, and the variance of $a_j + b_j$ is

$$\begin{aligned}\hat{S}_{B+A}^2 &= \hat{S}_{2\epsilon_j + \epsilon_{2j} - \epsilon_{1j}}^2 \\ &= \hat{S}_{2\epsilon_j}^2 + \hat{S}_{\epsilon_{2j}}^2 + (-1)^2 \hat{S}_{\epsilon_{1j}}^2 \\ &= (2\sigma)^2 + \sigma_2^2 + \sigma_1^2 \\ &= 4\sigma^2 + \sigma_2^2 + \sigma_1^2\end{aligned}$$

Therefore, $\hat{S}_{B+A}^2 - \hat{S}_{B-A}^2 = 4\sigma^2$, and subsequently

$$\begin{aligned}\sigma^2 &= \frac{\hat{S}_{A+B}^2 - \hat{S}_{B-A}^2}{4} \\ &= \hat{S}_{\frac{A+B}{2}}^2 - \hat{S}_{\frac{B-A}{2}}^2\end{aligned}$$

Further, since $a_j + b_j = 2\mu_j + 2\epsilon_j + \epsilon_{2j} - \epsilon_{1j}$, $E(A+B) = E(2\mu_j + 2\epsilon_j + \epsilon_{2j} - \epsilon_{1j}) = E(2\mu_j) + E(2\epsilon_j) + E(\epsilon_{2j}) - E(\epsilon_{1j}) = 2\mu_j + 0 + \mu - \mu = 2\mu_j$, thus

$$\begin{aligned}\mu_j &= \frac{E(A+B)}{2} \\ &= \frac{\hat{\mu}_{A+B}}{2} \\ &= \hat{\mu}_{\frac{A+B}{2}}\end{aligned}$$

So the Gaussian distribution of every uncertain numerical attribute is $P(Z|(A, B)) \sim N(\hat{\mu}_{\frac{A+B}{2}}, \hat{S}_{\frac{A+B}{2}}^2 - \hat{S}_{\frac{B-A}{2}}^2)$. Thus the theorem is proved. \square

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 the same as its maximal boundary, that is, $A = B = Z$. Therefore, the mean of the whole dataset is $\hat{\mu}_{\frac{A+B}{2}} = \hat{\mu}_Z$. The variance of the dataset is

$$\begin{aligned} \hat{S}_{\frac{A+B}{2}}^2 - \hat{S}_{\frac{B-A}{2}}^2 &= \hat{S}_{\frac{Z+Z}{2}}^2 - \hat{S}_{\frac{Z-Z}{2}}^2 \\ &= \hat{S}_Z^2. \end{aligned}$$

So for certain data, the mean is estimated to be $\hat{\mu}_Z$ and the variance is estimated to be \hat{S}_Z^2 , which is consistent with the naive Bayes classification algorithm. Therefore,

$$P(Z|(A, B)) \sim N(\hat{\mu}_{\frac{A+B}{2}}, \hat{S}_{\frac{A+B}{2}}^2 - \hat{S}_{\frac{B-A}{2}}^2)$$

is also a general form for mean and variance calculations. It applies to both uncertain and certain numerical data.

Example 2 Refer to the data shown in Table 1. For the uncertain numerical attribute Annual Income, according to Theorem 2, the mean and variance with respect to class **No** are as follows.

$$\begin{aligned} \hat{\mu} &= \hat{\mu}_{\frac{A+B}{2}} \\ &= 109.7 \end{aligned}$$

$$\begin{aligned} \hat{S}^2 &= \hat{S}_{\frac{A+B}{2}}^2 - \hat{S}_{\frac{B-A}{2}}^2 \\ &= \frac{1}{m} \sum_{j=1}^m \left(\frac{a_j + b_j}{2}\right)^2 - \left(\frac{1}{m} \sum_{j=1}^m \frac{a_j + b_j}{2}\right)^2 - \frac{1}{m} \sum_{j=1}^m \left(\frac{b_j - a_j}{2}\right)^2 + \left(\frac{1}{m} \sum_{j=1}^m \frac{b_j - a_j}{2}\right)^2 \\ &= (13864.84 - 12031.35) - (321.09 - 215.72) \\ &= 1728.1 \end{aligned}$$

Algorithm 1 NBU1(Dataset D)

begin

```
1: for (Every instance  $T_j \in D$ ) do
2:   for (Every attribute  $A_i$ ) do
3:     if ( $A_i$  is uncertain numerical) then
4:        $sum_i = (A_{ij}.max + A_{ij}.min)/2$ ;
5:        $diff_i+ = (A_{ij}.max - A_{ij}.min) * (A_{ij}.max - A_{ij}.min)/36$ ;
6:        $N(\mu_i, \sigma_i^2) = \text{updateGaus}(sum_i, T_j.w)$ ;
7:     else
8:       use naive Bayesian algorithm;
9:     end if;
10:  end for;
11:   $weight_i+ = T_j.w$ ;
12: end for;
13: for (Every uncertain numerical attribute  $A_i$ ) do
14:   $\hat{S}_i^2 = \sigma_i^2 + diff_i/weight_i$ ;
15: end for;
end
```

5 Uncertain Bayesian Classification and Prediction

Based on Theorem 1 and Theorem 2, this section discusses the techniques to construct the classifier for uncertain data and predict the class type of previous unseen data. If the classification is based on Theorem 1, we call it Naive Bayesian classification one and denote it by NBU1. The other classification is called Naive Bayesian classification two and denoted by NBU2 [24].

5.1 The Bayesian algorithms for uncertain data

In this section, we first present NBU1, which is shown in Algorithm 1. Its principal steps are as follows:

- For every uncertain numerical attribute instance A_i , let $sum_i = (A_{ij}.max + A_{ij}.min)/2$ and $diff_i+ = (A_{ij}.max - A_{ij}.min) * (A_{ij}.max - A_{ij}.min)/36$. Then we update the Gaussian distribution $N(\mu_i, \sigma_i^2)$ by Function update-

Gaus() (steps 3-6).

- For every certain attribute instance A_i , we use naive Bayesian algorithm (step 8).
- For every instance T_j , we update $weight_i$ (step 11).
- Finally, we modify the variance \hat{S}_i^2 for every uncertain numerical attribute A_i using Theorem 1 (steps 13-15).

NUB2 is shown in Algorithm 2 with major steps as follows:

- For every uncertain numerical attribute instance A_i , let $sum_i = (A_{ij}.max + A_{ij}.min)/2$ and $diff_i = (A_{ij}.max - A_{ij}.min)/2$. Then we update the Gaussian distribution $N(\mu_i^+, (\sigma_i^+)^2)$ and $N(\mu_i^-, (\sigma_i^-)^2)$ by Function updateGaus() (steps 3-7).
- For every certain attribute instance A_i , we use naive Bayesian algorithm (step 9).
- Finally, we modify the variance \hat{S}_i^2 for every uncertain numerical attribute A_i using Theorem 2 (steps 13-16).

Algorithm 2 NBU2(Dataset D)

begin

```

1: for (Every instance  $T_j \in D$  do) do
2:   for (Every attribute  $A_i$  do) do
3:     if ( $A_i$  is uncertain numerical) then
4:        $sum_i = (A_{ij}.max + A_{ij}.min)/2$ ;
5:        $diff_i = (A_{ij}.max - A_{ij}.min)/2$ ;
6:        $N(\mu_i^+, (\sigma_i^+)^2) = \text{updateGaussian}(sum_i, T_j.w)$ ;
7:        $N(\mu_i^-, (\sigma_i^-)^2) = \text{updateGaussian}(diff_i, T_j.w)$ ;
8:     else
9:       use naive Bayesian algorithm;
10:    end if;
11:  end for;
12: end for;
13: for (Every uncertain numerical attribute  $A_i$ ) do
14:    $\mu_i = \mu_i^+$ ;
15:    $\hat{S}_i^2 = (\sigma_i^+)^2 - (\sigma_i^-)^2$ ;
16: end for;

```

end

An important benefit of naive Bayesian classification is that it is incremental, which means that the model can evolve gradually when more training data becomes available. Many other classifications, on the contrary, require the whole classification to be rebuilt with newly added training data. For example, the decision tree is essentially non-incremental, with more training data, the cut point and tree structure can be completely different and it is better to rebuild it. Please note that our NBU1 and NBU2 preserve the incremental feature. For uncertain data, both μ_j and \hat{S}_j^2 can be incrementally updated. This is very important in data stream applications where new data constantly becomes available and the classification should be continuously adjusted.

5.2 Prediction with the proposed classifiers

Once the parameters of Bayesian classification are learned from the training data, it can be used to predict the class type of previously unseen data. Every instance T_j may be predicted in different classes with certain probabilities. We can compute a vector of every instance T_j .

Definition 4 *The class distributions of every instance T_j can be denoted by a vector $CDV(T_j, C) = (P(C_1|T_j), P(C_2|T_j), \dots, P(C_n|T_j))^T$, and we call it a Class Distribution Vector (CDV) of the instance T_j .*

Suppose T_j has m attributes, A_1, A_2, \dots, A_m , among them p attributes are uncertain numerical attributes and the rest are certain attributes. Then by Equation (1), $P(C_k|T_j)$, the probability T_j is in class C_k , should be:

$$P(C_k|T_j) = \frac{P(C_k) * \prod_{j=1}^p P(A_{ij}|C_k) * \prod_{j=p+1}^m P(A_{ij}|C_k)}{P(A_{i1}, A_{i2}, \dots, A_{im})}$$

We can compute $P(C_k|T_j)$ for all classes and predict T_j to be the class with the highest probability.

For every uncertain numerical attribute instance A_{ij} , we need compute the conditional probability that A_{ij} falls within the uncertain interval $[a_j, b_j]$. This probability can be computed based on the mean and variance of the uncertain numerical attribute, which have been computed using Theorem 1 or Theorem

2 based on training data. Suppose with respect to class C_k , the mean A_i is μ_{ik} and the variance of A_i is σ_{ik}^2 , then

$$P(A_{ij} = [a_j, b_j] | C_k) = \int_{a_j}^{b_j} \frac{1}{\sqrt{2\pi}\sigma_{ik}} \exp^{-\frac{(x-u_{ik})^2}{2\sigma_{ik}^2}} dx.$$

6 Experiments

Using Java, we implemented the proposed Bayesian classification to classify uncertain data sets. When NBU1 and NBU2 are applied on certain data, they work as the naive Bayesian classification (NB), which has been implemented in Weka [29]. In the following experiments, we use ten times ten-fold cross validation. For every ten-fold cross validation, data is split into 10 approximately equal partitions; each one is used in turn for testing while the rest is used for training, that is, 9/10 of data is used for training and 1/10 for test.

The experiments are executed on a PC with an Intel Pentium IV 3.2 GHZ CPU and 2.0 GB main memory. In this section, we present the experimental results of the proposed algorithm. As test domains we use 19 data sets, which are described in Table 2, from the UCI repository [1]. In Table 2, Attrs gives the number of numerical attributes out of the total number of attributes, column m is the number of classes, n is the total number of tuples in the domain, and \bar{V} is the average number of distinct values for an attribute.

To make numerical attributes uncertain, we convert every numerical value to an uncertain interval with normal distribution as in books [5,14] and papers [22,23]. The uncertain interval is generated around the original value, which is the center point of the interval. In [28], every numerical value is converted into a set of s sample points between the uncertain interval $[a_j, b_j]$ with the associated value $f(x)$, effectively approximating every $f(x)$ by a discrete distribution. In the following sections, the data set with 10% uncertainty is denoted by U10. For example, if an original value is 20, then its U10 has an interval [18, 22). Other notations have the similar meaning. We use U0 to denote accurate or certain data sets.

Table 2

Characteristic figures of the nineteen numerical data sets

Data set	Attrs	m	n	\bar{V}
Abalone	7/8	29	4,177	863.7
Blood	4/4	2	748	43.8
Breast	30/30	2	569	481.1
Diabetes	8/8	2	768	156.8
Ecoli	7/7	8	336	51.9
Glass	9/9	6	214	100.7
Ionosphere	34/34	2	351	239.4
Iris	4/4	3	150	30.8
Letter	16/16	26	20,000	16
Liver	6/6	2	345	54.7
Magic	10/10	2	19,020	14710.7
Page	10/10	5	5,473	909.2
Satellite	36/36	6	4,435	76.3
Segment	19/19	7	2,310	628.3
Sonar	60/60	2	208	187.6
Vehicle	18/18	4	846	79.4
Waveform	21/21	3	5,000	714
Wine	13/13	3	178	98.1
Yeast	8/8	10	1,484	51.5

6.1 Accuracy

As prediction accuracy is by far the most important measure for a classification method, we compared the prediction accuracy of NBU1 with that of NBU2. The results are shown in Table 3, where every cell is the average accuracy. We

make all numerical attributes uncertain for every data set.

Table 3 shows that the classifier built by NBU1 can be potentially more accurate than NB for most data sets. For instance, for the glass data set, the average accuracy is improved from 48.49% to 50.23%. NBU1 gives worse accuracies than NB for few data sets, for instance, the blood data set. Table 3 also shows that NBU2 can build more accurate classification than NB on some data sets. For instance, for the ionosphere data set, the average accuracy is improved from 82.4% to 85.47%, which is big improvement. NBU2 gives worse accuracies than NB for some data sets, for instance, the blood, diabetes, yeast and vehicle data sets. Further, we find two exceptions: The accuracies of NBU2 are 64.94% and 19.01% for U1 and U5 of the ecoli data sets respectively. The accuracies of NBU2 are 47.23% and 47.13% for U1 and U5 of the yeast data sets respectively.

Comparing NBU1 with NBU2, we find the following phenomena. First, the accuracy differences between NBU1 and NBU2 are small in most situations. The difference between Theorem 1 and Theorem 2 is in the way variance is computed. In Theorem1, variance is calculated as $\hat{S}_1^2 = \hat{S}_{\frac{A+B}{2}}^2 + \Delta_1$; while in Theorem 2, variance is calculated as $\hat{S}_2^2 = \hat{S}_{\frac{A+B}{2}}^2 - \Delta_2$. In most situations, both Δ_1 and Δ_2 are small. Thus the difference between NBU1 and NBU2 is slight. Second, we find \hat{S}_1^2 is a negative modification of $\hat{S}_{\frac{A+B}{2}}^2$ while \hat{S}_2^2 is a positive modification of $\hat{S}_{\frac{A+B}{2}}^2$. If \hat{S}_2^2 is too small, then NBU2 becomes unstable. The reason is that when the estimated variance \hat{S}_2^2 is too small, a small change in data value can result in great fluctuation in probability estimation, which makes the classifier's performance unstable. Therefore, Δ_1 has small influence on the NBU1 classification while Δ_2 may have significant influence on NBU2, which affects its robustness. Hence NBU1 is more stable than NBU2 in some cases.

6.2 Computation time

We investigate the time it takes to construct a classification. Table 4 depicts the absolute run time in seconds when all tuples of a data set are used to build a classification. When data is certain, the time it takes to construct a

classification is determined by the data set size in terms of both the number of instances and dimensionalities. For example, Letter is the largest one among all numerical data sets; therefore, it takes the longest time to build a classifier for the naive Bayesian classifier.

Comparing with constructing an NB classification, constructing an NBU1 (NBU2) classification introduces extra overhead according to Theorem 1 (Theorem 2). The workload increase is small. Thus we observe from Table 4 that if the number of instances of a data set is not large, the time of constructing an NBU1 (NBU2) classifier and an NB one is almost the same. However, according to the naive Bayesian algorithm implemented in Weka, we should calculate the numeric precision from differences between adjacent values of numerical attributes. So if the number of instances of a data set is large, the cost between constructing an NBU1 (NBU2) classification and constructing an NB classification can not be ignored. We further observe that although the letter data set has more instances than the magic data set, constructing an NBU1 (NBU2) classification takes more time from the magic data set than from the letter data set. The reason is that the magic data set has much more distinct values than the letter data set, as shown in Table 2. We also observe that the time it takes to construct an NBU1 and an NBU2 classifier are almost the same. This is reasonable since the complexity of Theorem 1 and Theorem 2 are similar.

7 Conclusions

In this paper, we propose a novel Bayesian classification for classifying and predicting uncertain data sets. Uncertain data are extensively presented in modern applications such as sensor databases and biometric information systems. Instead of trying to eliminate uncertainty and noise from data sets, this paper follows the new paradigm of directly mining uncertain data. We integrate the uncertain data model with Bayes theorem and propose new techniques to calculate conditional probabilities. Besides laying the theoretical foundations for enhancing naive Bayesian classification to process uncertain data sets, we show how to put these concepts into practice. Our experimental evaluation demonstrates that the classifiers for uncertain data can be efficiently constructed

and effectively classify and predict even highly uncertain data. Further, the proposed classification is more stable and more suitable for mining uncertain data than the previous work [22].

References

- [1] <http://archive.ics.uci.edu/ml/datasets.html>.
- [2] C. Aggarwal, Philip Yu, Outlier detection with uncertain data, in: Proceedings of SDM08.
- [3] C. Aggarwal, Y. Li, J. Wang, J. Wang, Frequent pattern mining with uncertain data, in: Proceedings of SIGKDD, 2009, pp.29-38.
- [4] T. Bernecker, H. Kriegel, M. Renz, F. Verhein, A. Zfle, Probabilistic frequent itemset mining in uncertain databases, in: Proceedings of SIGKDD, 2009, pp.119-128.
- [5] H. H. Bock, E. Diday, Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data, Springer Verlag, 2000.
- [6] D. Burdick, M. P. Deshpande, T. S. Jayram, R. Ramakrishnan, S. Vaithyanathan, OLAP over uncertain and imprecise data, the VLDB Journal 16(1)(2007) 123-144.
- [7] F. Carvalho, P. Brito, H. H. Bock, Dynamic clustering for interval data based on L2 distance, Computational Statistics, 21(2)(2006) 231-250.
- [8] M. Chavent, F. Carvalho, Y. Lechevallier, R. Verde, New clustering methods for interval data, Computational Statistics, 21(2)(2006) 211-229.
- [9] R. Cheng, D. Kalashnikov, S. Prabhakar, Evaluating probabilistic queries over imprecise data, in: Proceedings of the ACM SIGMOD, 2003, pp.551-562.
- [10] C. Chui, B. Kao, E. Hung, Mining frequent itemsets from uncertain data, in: Proceedings of the PAKDD, 2007, pp.47-58.
- [11] W. W. Cohen, Fast effective rule induction, in: Proceedings of ICML, 1995, pp.115-123.
- [12] G. Cormode, A. McGregor, Approximation algorithm for clustering uncertain data, in: Proceedings of the ACM PODS, 2008, pp.191-199.

- [13] A. Dayanik, Feature interval learning algorithms for classification, *Knowledge-Based Systems*, 23(5)(2010)402-417
- [14] E. Diday, M. N. Fraiture, *Symbolic data analysis and the sodas software*, Wiley, 2008.
- [15] C. Gao, J. Wang, Direct mining of discriminative patterns for classifying uncertain data, in: *Proceedings of the SIGKDD*, 2010, pp. 861-870.
- [16] N. L. Johnson, S. Kotz, N. Balakrishnan, *Continuous univariate distributions*. (2nd ed.), Wiley and Sons, 1994.
- [17] H. Kriegel, M. Pfeifle, Density-based clustering of uncertain data, In: *Proceedings of the SIGKDD*, 2005, pp.672-677.
- [18] P. Langley, W. Iba, K. Thompson, An analysis of bayesian classifiers, in: *Proceedings of the tenth National Conference on artificial intelligence*, 1992, pp.223-228.
- [19] P. Liu, F. Jin, X. Zhang, Y. Su, M. Wang, Research on the multi-attribute decision-making under risk with interval probability based on prospect theory and the uncertain linguistic variables, *Knowledge-Based Systems*, 2011, in press.
- [20] O. Lobo, M. Numao, Ordered estimation of missing values, in: *Proceedings of PAKDD*, 1999, pp.499-503.
- [21] W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, K. Y. Yip, Efficient clustering of uncertain Data, in: *Proceedings of ICDM*, 2006, pp.436-445.
- [22] B. Qin, Y. Xia, F. Li, DTU: a decision tree for uncertain data, in: *Proceedings of PAKDD*, 2009, pp.4-15.
- [23] B. Qin, Y. Xia, S. Prbahakar, Rule induction for uncertain data, *Knowledge Information Systems*, in press.
- [24] B. Qin, Y. Xia, Li F, A bayesian classifier for uncertain data, in: *Proceedings of SAC*, 2010.
- [25] J. R. Quinlan, *Probabilistic decision trees in machine learning: an artificial intelligence approach*, Morgan Kaufmann Publishers Inc. San Francisco, 1990.
- [26] J. R. Quinlan, *C4.5: programs for machine learning*, Morgan Kaufman Publishers, 1993.
- [27] P. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Addison Wesley, 2005.

- [28] S. Tsang, B. Kao, K. Y. Yip, W. S. Ho, S. D. Lee, Decision trees for uncertain data, *IEEE Transactions on Knowledge and Data Engineering*, 23(1)(2011) 64-78.
- [29] I. H. Witten, E. Frank, *Data mining: practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufman Publishers, 2005.
- [30] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg Top 10 algorithms in data mining, *Journal Knowledge and Information Systems*, 14(1)(2008) 1-37.
- [31] Y. Xia, B. Xi, Conceptual clustering categorical data with uncertainty, in: *Proceedings of international conference on tools with artificial intelligence*, 2007, pp.329-336.
- [32] Z. Yu, H. Wong, Mining uncertain data in low-dimensional subspace, in: *Proceedings of ICPR*, 2006, pp.748-751.
- [33] Z. Yue, An extended TOPSIS for determining weights of decision makers with interval numbers, *Knowledge-Based Systems*, 24(1)(2011)146153.
- [34] H. Zhang, W. Zhang, C. Mei, Entropy of interval-valued fuzzy sets based on distance and its relationship with similarity measure, *Knowledge-Based Systems*, 22(6)(2009)449-454.

Table 3

Accuracy of NBU1 vs. NBU2 over data sets with all numerical attributes uncertain

Data sets	U0	U1		U5		U10		U20	
	NB	NBU1	NBU2	NBU1	NBU2	NBU1	NBU2	NBU1	NBU2
Abalone	23.82	23.85	23.82	23.76	23.7	23.58	23.51	23.52	23.46
Blood	75.4	74.62	74.46	74.68	74.54	74.78	74.57	74.93	74.73
Breast	93.19	93.23	93.15	93.32	93.26	93.39	93.29	93.57	93.29
Diabetes	75.68	75.21	75.08	75.21	75.08	75.43	75.16	75.15	75.03
Ecoli	85.48	86.83	64.94	84.69	19.01	86.17	86.31	86.23	85.84
Glass	48.49	48.51	48.4	49.91	49.61	50.23	50.18	48.49	48.41
Iono.	82.4	83.54	85.47	83.49	83.76	83.43	83.65	82.97	82.85
Iris	95.47	95.52	95.2	95.39	95.13	95.78	95.4	96.12	95.93
Letter	64.13	63.92	63.8	64.17	63.62	63.45	63.92	63.75	63.6
Liver	54.8	55.94	56.01	55.31	55.55	54.89	54.68	52.75	53.33
Magic	72.68	72.73	72.63	72.75	72.65	72.77	72.66	72.91	72.76
Page	90.14	89.98	89.99	90.13	90.03	90.18	90.07	90.16	90.1
Satellite	79.58	79.72	79.61	79.64	79.59	79.59	79.51	79.46	79.35
Segment	80.06	80.18	80.19	80.21	80.23	80.07	80.06	79.59	79.65
Sonar	67.73	69.07	69.07	67.63	67.63	67.15	67.15	67.14	67.15
Vehicle	44.42	45.01	44.85	44.44	44.47	44	43.97	44.49	44.07
Waveform	80.97	80.95	80.95	81.06	80.96	80.95	80.96	81.03	81
Wine	97.51	97.53	97.52	97.18	97.29	96.84	96.84	95.93	95.93
Yeast	57.8	56.05	47.23	58.04	47.13	57.73	54.15	58.49	56.7

Table 4

Classifier construction time of uncertain numerical data sets

Data sets	U0	U1		U5		U10		U20	
	NB	NBU1	NBU2	NBU1	NBU2	NBU1	NBU2	NBU1	NBU2
Abalone	0.07	0.19	0.21	0.23	0.2	0.19	0.28	0.19	0.22
Blood	0.02	0.02	0.01	0.02	0.01	0.02	0.02	0.02	0.02
Breast	0.03	0.05	0.02	0.03	0.05	0.03	0.04	0.05	0.03
Diabetes	0.01	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.02
Ecoli	0.02	0.01	0.02	0.01	0.02	0.02	0.02	0.02	0.02
Glass	0.01	0.02	0.04	0.02	0.02	0.01	0.02	0.01	0.02
Iono.	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Iris	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Letter	0.4	0.2	0.26	0.19	0.24	0.21	0.23	0.22	0.27
Liver	0.01	0.01	0.02	0.01	0.02	0.01	0.02	0.01	0.02
Magic	0.26	21.06	21.2	23.39	23.69	23.84	25.89	25.49	26.09
page	0.1	0.25	0.27	0.25	0.33	0.25	0.26	0.25	0.26
Satellite	0.1	0.16	0.19	0.14	0.19	0.17	0.19	0.18	0.19
Segment	0.05	0.15	0.16	0.14	0.16	0.05	0.16	0.16	0.16
Sonar	0.03	0.05	0.03	0.03	0.03	0.04	0.03	0.1	0.03
Vehicle	0.14	0.02	0.02	0.06	0.02	0.03	0.02	0.03	0.02
Waveform	0.08	0.58	0.6	0.55	0.56	0.56	0.67	0.52	0.61
Wine	0.01	0.02	0.02	0.01	0.02	0.01	0.02	0.01	0.02
Yeast	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03