Volume 5, Issue 10, October-2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Handling Uncertain Data for Naïve Bayesian Classifier Using Log Normal Distribution

B. V. S. T. Sai
Professor, Dept. of CSE,
STMW, Guntur, A.P, India

Dr. D. Nagaraju
Professor, Dept. of CSE
LBRCE, Vijayawada, A.P, India

Shaik Subhani
Asst. Professor, Dept. of CSE,
STMW, Guntur, A.P, India

Abstract*— *Many real time databases contain uncertain data. To overcome the uncertainty, the data must be integrated. This can be accomplished by using probabilistic approaches. Data uncertain is a generalized approach in many real world applications. The uncertain data can be controlled by using many statistical and soft computing techniques. In this paper, we are introducing a probabilistic technique for calculating the conditional probabilities. Here, we are proposing a naive Bayesian classifier for handling uncertain data using Log-Normal distribution. Our main objective is to determine uncertain of multiple attributes using the proposed approach (UNBC). Finally results shows that the proposed method classifies the uncertain data with high accuracy.

Keywords*— *Uncertainty, Lognormal Distribution, Naïve Bayesian classifier, conditional probability

I. INTRODUCTION

Data uncertainty arises in many real-world applications [13] such as sensor network, market analysis and medical diagnosis, where the precise values of data might be unknown due to imprecise measurement, outdated sources, or decision errors [1]. Uncertain data mining studies include clustering, classification, frequent item mining and outlier detection [2, 18]. The basic idea behind is that when computing the distance between two uncertain data objects, the probability distributions of objects are used to calculate the expected distance[3]. In recent years, many advanced technologies have been developed to store and record large quantities of data continuously. In most of the cases, the data partially may complete or contain errors [4]. Uncertainty may appear in numerical attributes. For example, vast amount of uncertain data are present in sensor network as a result of the imperfect hardware used for the collection process. For uncertain numerical data, an optimization mechanism is used to merge adjacent bins which have equal classifying class distribution. Uncertainty can also arise in categorical attributes. For instance, in cancer diagnosis, it is difficult for the doctor to accurately decide a tumor to be benign or malignant due to the experiment precision limitation. Traditional machine learning algorithms assume that data is exact or precise. This assumption may not hold in some situations because of data uncertainty arising from measurement errors, data staleness, and repeated measurements, etc. With uncertainty, the value of each data item is represented by a probability distribution function (pdf) [5].In many applications, data contains inherent uncertainty, such as environmental investigation, market analysis etc. [6]. Uncertain data in these applications are generally caused by data haphazardness and incompleteness, limitations of measuring equipment, data updates delays, etc. Due to the importance of those applications and the rapidly increasing amount of uncertain data collected and accumulated, analyzing huge collections of uncertain data has become an essential task [7].
In uncertain data management, data items are represented by using probability distributions rather than deterministic values. The following examples illustrates in which uncertain data management techniques are relevant

- Uncertainty may be arises due to the limitations of the underlying equipment (In case of sensor networks the output is uncertain because of the noise in sensor inputs or errors in wireless transmission).
- In demographic data sets, only incomplete aggregated data items are available because of privacy concerns. In some cases, probability density functions of the records may be available. Recent techniques construct privacy models, such that the output of the transformation approach is friendly to the use of uncertain data mining and management techniques.
- Data attributes are constructed using statistical methods (such as forecasting or imputation). In such cases, the underlying uncertainty in the derived data can be estimated accurately from the underlying methodology (e.g. missing data).
- In mobile applications, the curve of the objects may be unknown. In fact, many spatiotemporal applications are inherently uncertain, since the future behavior of the data can be predicted only approximately. The further into the future that the trajectories are extrapolated, the greater the uncertainty.

The problem of indexing uncertain data arises frequently in the context of several application domains such as moving trajectories or sensor data. In these cases, the data is updated only periodically in the index, and the current attribute values cannot be known exactly; they can only be estimated. There are many different kinds of questions which can be resolved with the use of index structures:

- **Range queries:** The aim is to find all the objects in a given range. Since the objects are uncertain, their exact positions cannot be known, and hence their membership in the range also cannot be known deterministically. Therefore, a probability value is associated for each object to belong to a range. All objects whose probability of membership lies above a certain threshold are retained.
- **Nearest neighbor queries**: we attempt to determine the objects with the least expected nearest neighbor distance to the target. An alternative way of formulating the probabilistic nearest neighbor query is in terms of the nonzero probability that a given object is the nearest neighbor to the target.
- **Aggregate queries**: The aim is to determine the aggregate statistics from queries such as the sum or the max. Aggregate queries are inherently more difficult than other kinds of queries such as range or nearest neighbor queries because one has to account for the interplay of different objects.

The value of a numerical attribute is uncertain, the attribute is called an uncertain numerical attribute (UNA), and is denoted by A_{ij}.Here we use $A_{ij}.U$ to denote the j^{th} instance of $A_i.U$. Cheng and S. Prabhakar have been introduced the concept of uncertain numerical attribute[14] . The probability distribution function (PDF) is denoted as $A_{ij}.f(x)$and it is defined as $\int_{Aij.min}^{Aij.max} Aij.f(x)\,dx$ Since data uncertainty is ubiquitous, it is important to develop data mining algorithms for uncertain datasets [9].

However, when data contains uncertainty – for example, when some numerical data are, instead of precise value, an interval with probability distribution function with that interval - these algorithms cannot process the uncertainty properly. We perform experiments on real datasets with both exponential and Gaussian distribution, and the experimental results shows that UNBC algorithm performs well even on highly uncertain data.

document is a template. An electronic copy can be downloaded from the Journal website. For questions on paper guidelines, please contact the journal publications committee as indicated on the journal website. Information about final paper submission is available from the conference website.

II. EXISTING WORK

Naive Bayes classifiers a commonly used classification technique based on Bayes theory with strong (naive) independence assumptions. That, means a naive Bayes classifier assumes that the presence / absence of a particular feature is independent to the presence/absence of any other feature, given the class variable. Based on class conditional density estimation and class prior probability, the posterior class probability of a test data point can be derived and the test data will be assigned to the class with the maximum posterior class probability. For some types of probability models, naive Bayes technique can be trained very efficiently in a supervised learning setting. In many practical applications, estimation of parameters for naive Bayes models uses the method of maximum likelihood. The main advantage of Naive Bayes is that it only requires a small amount of training data to estimate the parameters (such as means and variances of the variables) necessary for classification. The key problem in naive Bayes classifier is the estimation of class conditional density. Traditionally the class conditional density is estimated based on data objects. For uncertain classification problems, the class conditional density from uncertain data objects represented by probability distributions [8].

A. Bayes Theorem

Bayes' theorem is a mathematical formula used for calculating conditional probabilities. It can be seen as a way of understanding how the probability that a theory is true and affected by a new piece of evidence. Prior probabilities combined with Conditional Probabilities yields Posterior Probabilities.

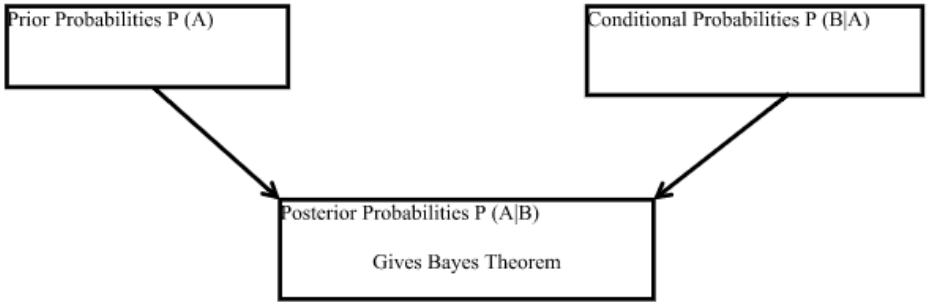

Fig 1 Bayes Theorem gives a way of calculating P (A|B) from knowledge of P (B|A).

B. Conditional Probability

The conditional Probability that an event 'X' will occur, given that 'Y' has occurred is denoted by the symbol P (A/B) and is defined as by [19]

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)} \quad ; P(Y) > 0 \qquad (1)$$

C. Multiplication Theorem of Probability

The Probability of the instantaneous occurrence of two events 'X' and ' Y' is equal to the product of probability of 'X' and the conditional Probability of 'Y' on the assumption that 'X' occurred.

i.e. $P(X \cap Y) = P(X).P(Y/X) = P(Y).P(X/Y)$ (2)

From (1) & (2)

$$P(X/Y) = \frac{P(X).P(Y/X)}{P(Y)} \quad (3)$$

D. Multiplication Theorem for Independent Events

If $(X \cap Y)$ and $(X \cap Y')$ are mutually exclusive events then by axiomatic definition

$$P(X) = P(Y).P(X/Y) + P(Y').P(X/Y') \quad (4)$$

From (3) & (4)

$$P(X/Y) = \frac{P(X).P(Y/X)}{P(Y).P\left(\frac{X}{Y}\right) + P(Y').P(X/Y')} \quad (5)$$

E. Total Probability

If $Y_1, Y_2, Y_3 \ldots\ldots\ldots\ldots Y_n$ are ' n' mutually exclusive events of which one of the event occur then,

$$P(X) = \sum_{i=1}^{i=n} P(Y_i).P(X/Y_i)$$

From (3)

$$P(X/Y) = \frac{P(X).P(Y/X)}{\sum_{i=1}^{i=n} P(Yi).P(X/Yi)} \quad (6)$$

Which is called Bayes theorem statement.

Bayes theorem provides a formula for finding the probability that the "effect" Y was "caused" by the event X [20].

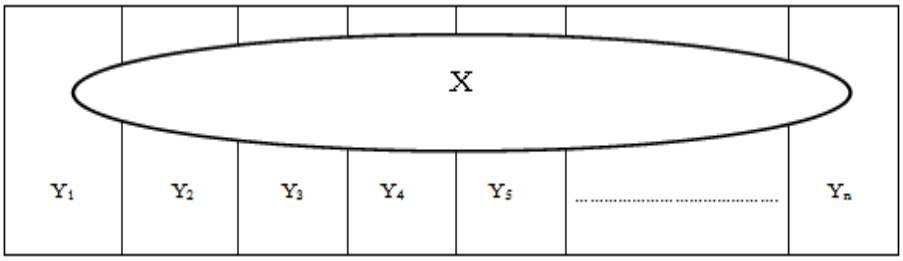

Fig 2 Venn diagram for occurrence of one the event from ' n' mutually exclusive events

The above Venn diagram shows Y1, $Y_2, Y_3 \ldots\ldots\ldots\ldots Y_n$ are ' n' mutually exclusive events of which one of the event 'X' occur.The following flow chart shows step-by-step process of calculating the posterior probabilities.

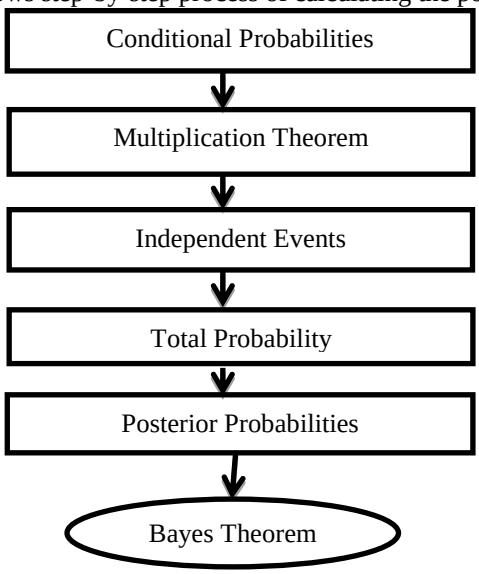

Fig 3 Process of calculating the posterior probabilities.

III. PROPOSED WORK

Our proposed approach contains two steps, which are used to find out the conditional probabilities for the uncertain numerical attributes and by calculating the mean and variance of the Log-Normal distribution.

A. Calculating Condition probabilities for uncertain numeric attributes

The probability density function of log- normal distribution is given by[17]

$$P(A/C_K) = \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}} ; \quad x>0, \sigma>0$$

Here μ and σ are the parameters of the distribution

The notation for lognormal distribution is $\log X \sim N(\mu, \sigma^2)$

Derivation for parameters of the distribution

The mean of the log normal distribution is

$u=\int_0^{+\infty} x\, f(x)\, dx$-- (1)

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

From eq (1)

$u=\int_0^{+\infty} x \frac{1}{m}\sum_{j=1}^{m} fj(x))\, dx$

$=\frac{1}{m}\sum_{j=1}^{m} \int_0^{\infty} x\, fj(x)\, dx$

$=\int_0^{+\infty} x \frac{1}{m}\sum_{j=1}^{m} \frac{1}{x\sigma\sqrt{2\pi}} e^{\frac{-(\ln x-\mu)2}{2\sigma 2}}\, dx$

$=\frac{1}{m}\sum_{j=1}^{m} \int_0^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\ln x-\mu)2}{2\sigma 2}}\, dx$

Substitute

Log x= t $\Rightarrow x=e^t$

$\frac{1}{x} dx = dt$

dx=x.dt

$dx=e^t dt$

$=\frac{1}{m}\sum_{j=1}^{m} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu)2}{2\sigma 2}} e^t dt$

$=\frac{1}{m}\sum_{j=1}^{m} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)2-4t\sigma 2} dt$

$=\frac{1}{m}\sum_{j=1}^{m} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu-\sigma 2)2-2\mu\sigma 2-\sigma 4} dt$

$= \frac{1}{m}\sum_{j=1}^{m} e^{\frac{2\mu\sigma 2+\sigma 4}{2\sigma 2}} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu-\sigma 2)2}{2\sigma 2}} dt$

$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu-\sigma 2)2}{2\sigma 2}} dt = 1$ as $\frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu-\sigma 2)2}$ is probability density of a normal variable with mean $\mu+\sigma^2$ and Standard deviation σ

$$=\frac{1}{m}\sum_{j=1}^{m} e^{\frac{2\mu\sigma 2+\sigma 4}{2\sigma 2}}$$

$$\frac{1}{m}\sum_{i=1}^{m} e^{\mu+\frac{\sigma 2}{2}}$$

$=\frac{1}{m}\sum_{i=0}^{m} e^{\frac{(aj+bj)+2(bj-aj)2}{2\sigma 2}}$

$= u \exp\frac{A+B}{2} {}_{+\Delta}1$

Where ${}_{\Delta}1 = \frac{(bj-aj)2}{\sigma 2}$

Similarly,

We can derive the sample variance

$s^2= E(X^2) - [E(X)]^2$

$=\int_{-\infty}^{+\infty} \frac{1}{m}\sum_{j=1}^{m} fj(x))\, x^2\, dx - (u)^2$

$E(X^2) =\int_0^{+\infty} f(x)\, dx\, .x^2$

$= \int_0^{+\infty} \frac{1}{m}\sum_{j=1}^{m} fj(x).x^2 dx$

$=\frac{1}{m}\sum_{j=1}^{m} \int_0^{\infty} fj(x).x^2 dx$

$=\int_0^{+\infty} x \frac{1}{m}\sum_{j=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\ln x-\mu)2}{2\sigma 2}}\, dx$

Substitute

Log x= t $\Rightarrow x=e^t$

$$\frac{1}{x} dx = dt$$

dx=x.dt

$dx=e^t dt$

$=\frac{1}{m}\sum_{j=1}^{m} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu)2}{2\sigma 2}} e^t e^t dt$

$=\frac{1}{m}\sum_{j=1}^{m} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(t-\mu)2-4t\sigma 2}{2\sigma 2}} dt$

The mean and variance with respective class Yes, for the uncertain numerical attribute Age are given by

$u= \frac{1}{m}\sum_{i=1}^{m} e^{\mu+\frac{\sigma 2}{2}}$

$u=\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{2}\right)+\left(\frac{bj-aj}{72}\right)2}$

$u=\frac{1}{6}\sum_{j=1}^{6} e^{\left(\frac{bj+aj}{2}\right)+\left(\frac{bj-aj}{72}\right)2}$

$u=\frac{1}{6}[e^{\left(\frac{b1+a1}{2}\right)+\left(\frac{b1-a1}{72}\right)2}+e^{\left(\frac{b2+a2}{2}\right)+\left(\frac{b2-a2}{72}\right)2}+\cdots+e^{\left(\frac{b6+a6}{2}\right)+\left(\frac{b6-a6}{72}\right)2}]$

$=\frac{1}{6}[e^{\left(\frac{27+20}{2}\right)+\left(\frac{27-20}{72}\right)2}+e^{\left(\frac{48+37}{2}\right)+\left(\frac{48-37}{72}\right)2}+\cdots+e^{\left(\frac{35+21}{2}\right)+\left(\frac{35-21}{72}\right)2}]$

$=\frac{1}{6}[e^{24.18}+e^{44.18}+e^{78.51}+e^{51.39}+e^{30.85}+e^{30.72}]$

$=\frac{1}{6}[3.17\ 10^{10}+1.5410^{9}+\ldots..+2.1910^{13}]$

$=1.40*10^{31}$

$s^2=\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{1}\right)+\left(\frac{bj-aj}{72}\right)2}-[\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{2}\right)+\left(\frac{bj-aj}{72}\right)2}]^2$

$s^2=\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{1}\right)+\left(\frac{bj-aj}{72}\right)2}-u^2$

$s^2=\frac{1}{6}\sum_{j=1}^{6} e^{\left(\frac{bj+aj}{1}\right)+\left(\frac{bj-aj}{18}\right)2}-u^2$

$s^2=\frac{1}{6}[e^{\left(\frac{b1+a1}{1}\right)+\left(\frac{b1-a1}{18}\right)2}+e^{\left(\frac{b2+a2}{1}\right)+\left(\frac{b2-a2}{18}\right)2}+\cdots+e^{\left(\frac{b6+a6}{1}\right)+\left(\frac{b6-a6}{18}\right)2}]$

$=\frac{1}{6}[e^{\left(\frac{27+20}{1}\right)+\left(\frac{27-20}{18}\right)2}+e^{\left(\frac{48+37}{1}\right)+\left(\frac{48-37}{18}\right)2}+\cdots+e^{\left(\frac{35+21}{1}\right)+\left(\frac{35-21}{18}\right)2}]$

$=\frac{1}{6}[e^{49.72}+e^{91.72}+e^{155.05}+e^{105.55}+e^{62.39}+e^{66.89}]$

$=\frac{1}{6}[3.9210^{21}+6.8110^{39}+\ldots..+1.2110^{29}]$

$=0.36*10^{67}$

In the similar way, we have calculated the mean and variance with respective class Yes, for the uncertain numerical attribute Income are

$u=\frac{1}{m}\sum_{i=1}^{m} e^{\mu+\frac{\sigma 2}{2}}$

$u=\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{2}\right)+\left(\frac{bj-aj}{72}\right)2}$

$u=\frac{1}{6}\sum_{j=1}^{6} e^{\left(\frac{bj+aj}{2}\right)+\left(\frac{bj-aj}{72}\right)2}$

$u=\frac{1}{6}[e^{\left(\frac{b1+a1}{2}\right)+\left(\frac{b1-a1}{72}\right)2}+e^{\left(\frac{b2+a2}{2}\right)+\left(\frac{b2-a2}{72}\right)2}+\cdots+e^{\left(\frac{b6+a6}{2}\right)+\left(\frac{b6-a6}{72}\right)2}]$

$=\frac{1}{6}[e^{\left(\frac{75+50}{2}\right)+\left(\frac{75-50}{72}\right)2}+e^{\left(\frac{220+175}{2}\right)+\left(\frac{220-175}{72}\right)2}+\cdots+e^{\left(\frac{140+120}{2}\right)+\left(\frac{140-120}{72}\right)2}]$

$=\frac{1}{6}[e^{71.18}+e^{225.62}+e^{70}+e^{67.51}+e^{167}+e^{135.55}]$

$=\frac{1}{6}[8.1910^{30}+9.6710^{97}+\ldots\ldots+7.3910^{58}]$

$=0.369*10^{69}$

The variance is given by

$s^2=\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{1}\right)+\left(\frac{bj-aj}{72}\right)2}-[\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{2}\right)+\left(\frac{bj-aj}{72}\right)2}2$

$s^2=\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{1}\right)+\left(\frac{bj-aj}{72}\right)2}-\hat{u}^2$

$s^2=\frac{1}{6}\sum_{j=1}^{6} e^{\left(\frac{bj+aj}{1}\right)+\left(\frac{bj-aj}{18}\right)2}-\hat{u}^2$

$s^2=\frac{1}{6}[e^{\left(\frac{b1+a1}{1}\right)+\left(\frac{b1-a1}{18}\right)2}+e^{\left(\frac{b2+a2}{1}\right)+\left(\frac{b2-a2}{18}\right)2}+\cdots+e^{\left(\frac{b6+a6}{1}\right)+\left(\frac{b6-a6}{18}\right)2}]$

$=\frac{1}{6}[e^{\left(\frac{75+50}{1}\right)+\left(\frac{75-50}{18}\right)2}+e^{\left(\frac{220+175}{1}\right)+\left(\frac{220-175}{18}\right)2}+\cdots+e^{\left(\frac{140+120}{1}\right)+\left(\frac{140-120}{18}\right)2}]$

$=\frac{1}{6}[e^{159.72}+e^{507.5}+e^{156}+e^{143.05}+e^{434}+e^{282.22}]$

$=\frac{1}{6}[2.3210^{69}+e^{507.5}+\ldots..+e^{282.22}]$

$=1.61*10^{97}$

The Characteristic function of log-normal distribution is given by

$M_X(t)=E[e^{Xt}]$

$=\int_0^{\infty} f(x)e^{xt}dx$

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

$=\int_0^{\infty}\frac{1}{m}\sum_{j=1}^{m} f\,j(x)e^{xt}dx$

$=\frac{1}{m}\sum_{j=1}^{m}\int_0^{\infty} fj(x)e^{xt}dx$

$=\frac{1}{m}\sum_{j=1}^{m}\int_0^{\infty}\frac{1}{x\sigma\sqrt{2\pi}}e^{\frac{-(\ln x-\mu)2}{2\sigma 2}}e^{xt}dx$ .....(1)

$M_X(t)=E[e^{Xt}]$

$X=\log y \Rightarrow Y=e^x$

$M_X(t)=E[e^{Xt}] \Rightarrow E[Y^t]$

Similarly

$M_X(t)=E[Y^n]$

The characteristic function, $E[e^{itX}]$, has a number of representations. The integral itself converges for $Im(t) \leq 0$. The simplest representation is obtained by Taylor expanding e^{itX} and using formula for moments below, giving

$$M_{X(t)}=\sum_{n=0}^{\infty}\frac{1}{n}(it)^n e^{n\mu+n2\sigma 2/2}$$

From eq^n(1)

$$M_{X(t)}=\frac{1}{m}\sum_{j=1}^{m}\sum_{n=0}^{\infty}\frac{1}{m}(it)^n e^{n\mu+n2\sigma 2/2}$$

B. The Proposed Naïve Bayesian Classifier for Uncertain Data

A novel algorithm was proposed, which is used to classify the uncertain data is shown below [12].

Algorithm:HandlingUncertain DataforNaïve Bayesian Classifier (UNBC)
Input: D, Dataset Contains set of Attributes and tuples
Output: Certain Data
Begin
For eachtuple in X in D
{
For each attribute A_{ij}
{
If (A_{ij} is uncentertain numerical)
{

$$u=\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{2}\right)+\left(\frac{bj-aj}{72}\right)2}$$

$$I_j=\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{1}\right)+\left(\frac{bj-aj}{72}\right)2}-\left[\frac{1}{m}\sum_{j=1}^{m} e^{\left(\frac{bj+aj}{2}\right)+\left(\frac{bj-aj}{72}\right)2}\right]^2$$

Log nor (μ_i, σ_i) =update lognormal (OJ, X, w);
}
Else
{
D: Set of tuples
Each tuple is an 'n' dimensional attribute vector X: (x_1, x2, x3,….xn)
'm' Classes: C_1, C2,C3,…..C_m
// Naïve Bayes classifier predicts X belongs to Class $C_i$iff
$P(C_i/X) > P(C_j/X)$ for 1<= j <=m, j<>i
// Maximum Posteriori Hypothesis
$P(C_i/X) = P(X/C_i)\, P(C_i) / P(X)$
Maximize $P(X/C_i)\, P(C_i)$ as P(X) is constant
// Naïve assumption of "Class Conditional independence"
$P(X/C_i) = \prod_{k=1}^{n} P(x_k/C_i)$
$P(X/C_i) = P(x_1/C_i) * P(x_2/C_i) * \ldots * P(x_n/C_i)$
}// end if loop
}// end for loop
For each weight w_j
{
$w_j = x \,.\, w_j$// weight increment
$w_j = w_j + x \,.\, w_j$// weight update
}// End for loop
} // end for the main for loop
For each uncertain Numerical attribute A_{ij}
{
$s^{2=} \beta^2_i + (I_j/w_j)$;
}// end for loop
END //end for UNBC

IV. RESULTS

In this paper, we have implemented the proposed Naïve Bayesian Classifier is used to classify uncertain data set. Naïve Bayesian Classifier has been implemented in Weka [11] for the chosen dataset. The test mode is 10- fold cross validation. The correctly classified instances are with 21.4286% accuracy and incorrectly classified instances with 78.5714% accuracy. When UNBC is applied on certain data, it works as the naïve Bayesian Classification (NB), which has very good results. After applying the UNBC on the dataset the correctly classified accuracy is has been increased. The computational time has been reduced.

To mark numerical attributes uncertain, we convert every numerical value to an uncertain interval with log-normal distribution [15, 16].The uncertain interval is generated near by the original value, which is the centre point of the

interval. In [10], every numerical value is converted into a set of sample points between the uncertain intervals [a_j, bj] with the associated value f(x), effectively approximating every f(x) by a discrete distribution.

V. CONCLUSIONS

In this paper, we proposed Uncertain Data for Naïve Bayesian Classifier (UNBC), which is used to classify and predicting the given uncertain data. Uncertain data are available in many applications such as sensor network, market analysis and medical diagnosis, imprecise measurement, outdated sources, or decision errors. Instead of trying to eliminate the uncertainty in one attribute, we have considered the total dataset as uncertain. Here, wearecombiningtheuncertaindatamodelwithBayestheoremandintroduceda new technique for calculating the conditional probabilities, as well as the mean, variance and characteristic function of the log-normal distribution.

Finallyexperimentalresultsshowsthattheclassifiersforuncertaindatacanbeclassifyefficientlyandalsopredicttheuncertaindata with high accuracy.

ACKNOWLEDGMENT

Authors are thankful to Dr. Devarakonda Nagaraju, Professor, Dept. of computer science & engineering, LBRCE; Vijayawada, A.P, India for giving continues support and encouragement to carry out this work. Authors are also thankful to the reviewer for critically going through the manuscript and giving valuable suggestions for the improvement of manuscript.

REFERENCES

[1] B. Qin, Y. Xia, S. Prabhakar and Y. Tu, *A Rule-based Classication Algorithm for Uncertain Data*, ICDE, pp. 1633{1640, 2009.

[2] *Naive Bayes Classifier for Positive Unlabeled Learning with Uncertainty* by Jiazhen Hey, Yang Zhangz, XueLix, Yong Wang.

[3] Biao Qin, Yuni Xia, Shan Wang, Xiaoyong Du: *A novel Bayesian classification for uncertain data*. Knowl.-Based Syst. 24(8): 1151-1158 (2011)

[4] *A Survey of Uncertain Data Algorithms and Applications* Charu C. Aggarwal, Senior Member, IEEE, and Philip S. Yu, Fellow, IEEE. *IEEE Transactions On Knowledge And Data Engineering*, Vol. 21, No. 5, May 2009.

[5] JiangtaoRen, Sau Dan Lee, Xianlu Chen, Ben Kao, Reynold Cheng and David Cheung "*Naive Bayes Classification of Uncertain Data*",In Wei Wang, HillolKargupta, Sanjay Ranka, Philip S. Yu, and Xindong Wu, editors,

[6] *Probabilistic Skylines on Uncertain Data* JianPeiBinJiangXuemin Li Simon Fraser University, CanadaYidongYuanthe University of New South Wales & NICTA, Australia.

[7] *Density-Based Clustering of Uncertain Data* Hans-Peter Kriegel Martin Pfeifle.

[8] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[9] Biao Qin, YuniXia, Fang Li "*A Bayesian Classifier for Uncertain Data*" Proceedings of SAC'10 March 22-26, 2010, Sierre, Switzerland.

[10] 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.

[11] I. H. Witten, E. Frank, *Datamining: practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufman Publishers, 2005.

[12] NagarajuDevarakonda ,NagamaniChippada ,ShaikSubhani, *Naive Bayesian Classifier for Uncertain Data using Exponential Distribution*, Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering -- CEEE 2013 Copyright © Institute of Research Engineers and Doctors. All rights reserved. ISBN: 978-981-07-6260-5 doi:10.3850/ 978-981-07-6260-5_18.

[13] Anish Das Sarma, "*Managing Uncertain Data" Ph.D Thesis*, Stanford University, Nov 2009

[14] R. Cheng, D. Kalashnikov, and S. Prabhakar. *Evaluating probabilistic queries over imprecise data*. In SIGMOD 2003, pages 551–562.

[15] H. H. Bock, E. Diday, *Analysis of symbolic data. Exploratory methods for extracting statistical information from complex data*, Springer Verlag, 2000.

[16] E.Diday, M.N.Fraiture, *Symbolic data analysis and the sodas software*, Wiley, 2008.

[17] Sergiostheodoridis, Konstantinoskoutroumbas, *Pattern recognition*, 4thEdition, Elsevier.

[18] Aggarwal, Philip Yu, *Outlier detection with uncertain dat*a, in: Proceedings of SDM08.

[19] J.pratap Reddy, S.V Padmakar Reddy, S.Bhoopal Reddy, *Statistics*, Telugu academy, Hyderabad.

[20] Richard A.Johnson, Millers and freund's *Probability and Statistics for engineers*, Pearson Education, 2006.

