Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering -- CEEE 2013

ISBN: 978-981-07-6260-5

# Naive Bayesian Classifier for Uncertain Data using Exponential Distribution

Nagaraju Devarakonda Nagamani Chippada ShaikSubhani

**Abstract: Most real databases contain data whose correctness is uncertain. In order to work with such data, there is a need to quantify the integrity of the data. This is achieved by using probabilistic databases. Data uncertain is common in real world applications. The uncertainty can be controlled very prudently. In this paper, we are using probabilistic models on uncertain data and develop a novel method to calculate conditional probabilities for uncertain numerical attributes. Based on that, we propose a Naive Bayesian classifier algorithm for uncertain data(NBCU) using exponential distribution. The ultimate aim is determine the uncertainty of multiple attributes using our proposed approach (NBCU).The experimental results show that the proposed method classifies uncertain data with potentially higher accuracy.**

**Keywords: Uncertainity, Exponential Distribution, Naïve Bayesian classifier, conditional probability**

## 1. Introduction

There are many examples of uncertain data arising in real-world applications, such as information extraction, privacy, duplication, data integration, other sources of inherent uncertainty etc[1]. Data uncertainty arises naturally in many applications due to various reasons. For example, data obtained from measurements by physical devices are often imprecise due to measurement errors[4].The uncertain data models for both numerical and categorical attribute, which are the most common types of data come across in data mining applications. When the value of a numerical attribute is uncertain,[3] 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 [2]. 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. 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 NBCU algorithm performs well even on highly uncertain data.

## 2. Existing Work

In Probability theory, Bayes theorem relates to conditional and marginal probabilities of two random events. It is often used to compute posterior probabilities given observations. Let $x=(x^1,x^2,...,x^d)$ be a dimensional instance which has no class label, and our goal is to build a classifier to predict its unknown class label based on Bayes theorem[4]. Let $C=\{C_1,C_2,...,C_K\}$ be the set of the class labels. $P(C_k)$is the prior probability of $C_k$ (k=1,2,...,K) that are inferred before new evidence; $P(x|C_k)$ be the conditional probability of seeing the evidence x if the hypothesis $C_k$ is true [11]. A technique for constructing such classifiers to employ Bayes' theorem to obtain:

$$P(C_k|x)= \frac{P(x|Ck)P(Ck)}{\sum_{k'} P(x|ck')P(ck')}$$

A naive Bayes classifier assumes that the value of a particular feature of a class is unrelated to the value of any other feature, so that

$$P(x|Ck)=\prod_{j=1}^{d} P(x^j|C_k)$$

## 3. 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 exponential distribution.

Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering -- CEEE 2013

ISBN: 978-981-07-6260-5

## 3.1 Calculating Condition probabilities for uncertain numeric attributes

Our motivation is to describe the uncertain data and calculating conditional probabilities for the uncertain numerical data. So we emphasize uncertainty in multiple attributes and assume the class type is certain.

$$P(A/C_K) = 1/\beta_k \, e^{-(x-\mu_k)/\beta_k}$$

Let $A_{ij}.U$ be uncertain interval of attribute $A_{ij}$.

i.e. $A_{ij} = [A_{ij}.min, A_{ij}.max]$

Where $A_{ij}.max \geq A_{ij}.min$ and $A_{ij}.f(x)$ be uncertain P.D.F of $A_{ij}$

$$\text{i.e.} \int_{Aij.min}^{Aij.max} Aij.f(x)dx = 1$$

## 3.2 Calculating the mean and variance of Exponential distribution

The probability density function of exponential function is

$$f(x) = 1/\beta \, e^{-(x-\mu)/\beta}; \; X>0$$

Here μ and β are the parameters of the distribution

The notation for exponential distribution is X~ exp(β)

### *3.2.1 Derivation for parameters of the distribution*

The mean $\hat{u} = \int_{-\infty}^{+\infty} f(x)dx$ ---------(1)

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

From equation (1)

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

$u = \frac{1}{m}\sum_{j=1}^{m} \int_0^{\infty} (x/\beta_j) e^{-(x-\mu_j)/\beta_j} dx$ ------(2)

Substitute

$-(x-\mu_j)/\beta_j = t \Rightarrow x-\mu = \beta_j t$

$x = \mu + \beta_j t$

$dx = \beta_j \, dt$

From equation (2)

$u = \frac{1}{m}\sum_{j=1}^{m} \int_0^{\infty} (\mu + \beta_j t) / \beta_j \, e^{-t} \beta_j dt$

$= 1/m\sum_{j=1}^{m} \int_0^{\infty} (\mu + \beta_j t)\, e^{-t} dt$

$= \frac{1}{m}\sum_{j=1}^{m} \mu_j \int_0^{\infty} e^{-t}\, dt + \beta_j \int_0^{\infty} te^{-t} dt$

$= \frac{1}{m}\sum_{j=1}^{m} \mu_j(1) + \beta_j \, (1)$

$u = \frac{1}{m}\sum_{j=1}^{m} \mu_j + \beta_j$ ------ (3)

$u = \frac{1}{m}\sum_{j=1}^{m} \frac{bj+aj}{2} + \frac{bj-aj}{6}$

$u = u_{(A+B)/2} + \Delta_1$

Where

$\Delta_1 = \frac{1}{m}\sum_{j=1}^{m} \frac{bj-aj}{6}$

The variance $\hat{s}^2$ is given by

$s^2 = E(x^2) - E(x)$ -------- (4)

$E(x^2) = \int_{-\infty}^{+\infty} x^2 f(x)$ -------- (5)

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

From the equation (5)

$E(x^2) = \int_{-\infty}^{\infty} x^2 1/m \sum_{j=1}^{m} f_j(x)\, dx$

$= \frac{1}{m}\sum_{j=1}^{m} \int x^2 f_j(x)\, dx$

$= \frac{1}{m}\sum_{j=1}^{m} \int_0^{\infty} (x^2/\beta_j)\, e^{-(x-\mu_j)/\beta_j} dx$

$X-\mu_j/\beta_j = t \Rightarrow x-\mu = \beta_j t$

$x = \mu + \beta_j t$

$dx = \beta_j \, dt$

From the above equations

$= \frac{1}{m}\sum_{j=1}^{m} \int_0^{\infty} ((\mu_j+\beta_j t)^2/\beta_j) e^{-t}) \beta_j dt$

$= \frac{1}{m}\sum_{j=1}^{m} \int_0^{\infty} (\mu_j+\beta_j t)^2 e^{-t} dt$

$= \frac{1}{m}\sum_{j-1}^{m} \int_0^{\infty} (\mu_j^2 + t^2\beta_j^2 + 2t\mu_j\beta_j)\, e^{-t} dt$

$= \frac{1}{m}\sum_{j-1}^{m} (\mu_j^2 \int_0^{\infty} e^{-t}dt + \beta_j^2 \int_0^{\infty} t^2 e^{-t}dt + 2\mu_j \beta_j \int_0^{\infty} te^{-t}dt$

$= \frac{1}{m}\sum_{j=1}^{m} (\mu^2_j\,(1) + \beta_j^2(2) + 2\mu_j\beta_j(1))$

$E(X^2) = \frac{1}{m}\sum_{j=1}^{m} (\mu_j^2 + 2\beta_j^2 + 2\mu_j\beta_j)$

From equation (4)

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

$= E(X^2) - \hat{u}^2$

$s^2 = \frac{1}{m}\sum_{j=1}^{m} (\mu^2_j + 2\beta^2_j + 2\mu_j\beta_j) - (\frac{1}{m}\sum_{j=1}^{m} \frac{bj+aj}{2} + \frac{bj-aj}{6})^2$

$= \frac{1}{m}\sum_{j=1}^{m} (\frac{bj+aj}{2})^2 + 2(\frac{bj-aj}{6})^2 + 2(\frac{bj+aj}{2})(\frac{bj-aj}{6}) - (\frac{1}{m}\sum_{j=1}^{m} \frac{bj+aj}{2} + \frac{bj-aj}{6})^2$

$s^2 = s^2_{(A+B)/2} + \Delta_2$

Where

$\Delta_2 = \frac{2}{m}\sum_{j=1}^{m} (\beta_j^2 + \mu_j\beta_j) - (\frac{1}{m}\sum_{j=1}^{m} (\beta_j))^2 - 2\,(\frac{1}{m}\sum_{j=1}^{m} (\mu_j))\,(\frac{1}{m}\sum_{j=1}^{m} (\beta_j))$

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

$u = \frac{1}{m}\sum_{j=1}^{m} \frac{bj+aj}{2} + \frac{bj-aj}{6}$

Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering -- CEEE 2013

ISBN: 978-981-07-6260-5

$u= \frac{1}{6}\sum_{j=1}^{6} \frac{bj+aj}{2} + \frac{bj-aj}{6}$

$u= \frac{1}{6}[ (\frac{27+20}{2} + \frac{27-20}{6}) + (\frac{48+37}{2} + \frac{48-37}{6}) + (\frac{78+61}{2} + \frac{78-61}{6}) + (\frac{55+45}{2} + \frac{55-45}{6}) + (\frac{33+28}{2} + \frac{33-28}{6}) + (\frac{35+21}{2} + \frac{35-21}{6})$

=42.445

$s^2 = \frac{1}{m}\sum_{j=1}^{m}(\frac{bj+aj}{2})^2 + 2(\frac{bj-aj}{6})^2 + 2(\frac{bj+aj}{2})(\frac{bj-aj}{6}) - (\frac{1}{m}\sum_{j=1}^{m} \frac{bj+aj}{2} + \frac{bj-aj}{6})^2$

$= \frac{1}{6}\sum_{j=1}^{6}(\frac{bj+aj}{2})^2 + 2(\frac{bj-aj}{6})^2 + 2(\frac{bj+aj}{2})(\frac{bj-aj}{6}) - (\frac{1}{m}\sum_{j=1}^{6} \frac{bj+aj}{2} + \frac{bj-aj}{6})^2$

$= \frac{1}{6}[ (\sum_{j=1}^{6}(\frac{27+20}{2})^2 + 2(\frac{27-20}{6})^2 + 2(\frac{27+20}{2})(\frac{27-20}{6})) + \ldots. +$

$(\sum_{j=1}^{6}(\frac{35+21}{2})^2 + 2(\frac{35-21}{6})^2 + 2(\frac{35+21}{2})(\frac{35-21}{6}))]$

=1583.23433

In the similar way, we have calculated the mean and variance with respective class Yes, for the uncertain numerical attribute Income are $\hat{u}$=110.7225 and $\hat{s}^2$ =12846.989

Table 1. Training dataset for the class buys-computer fromhe All Electronics customer database

| Row ID | Age | Income | Student | Credit_ Rating | Buys_ Computer |
|---|---|---|---|---|---|
| 1 | 20-27 | 50-75 | Yes | Excellent | Yes |
| 2 | 18-24 | 62-80 | Yes | Fair | No |
| 3 | 37-48 | 175-220 | Yes | Fair | Yes |
| 4 | 64-70 | 90-146 | No | Fair | No |
| 5 | 61-78 | 50-74 | No | Fair | Yes |
| 6 | 74-80 | 65-80 | No | Excellent | NO |
| 7 | 45-55 | 55-72 | No | Excellent | Yes |
| 8 | 28-33 | 87-147 | Yes | Fair | Yes |
| 9 | 22-29 | 54-74 | Yes | Fair | NO |
| 10 | 65-79 | 100-150 | No | Fair | NO |
| 11 | 21-35 | 120-140 | Yes | Excellent | Yes |
| 12 | 49-57 | 94-130 | No | Excellent | NO |
| 13 | 38-49 | 165-220 | Yes | Fair | No |
| 14 | 63-75 | 87-143 | No | Excellent | No |

## 3.3 The Proposed Naïve Bayesian Classifier for Uncertain Data

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

Algorithm : Naïve Bayesian Classifier for Uncertain Data( NBCU)

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)

{

$u= \frac{1}{m}\sum_{j=1}^{m} \frac{2bj+aj}{3}$

$I_j= \frac{1}{m}\sum_{j=1}^{m}(2(\frac{bj-aj}{6})^2 + 2(\frac{bj+aj}{2})(\frac{bj-aj}{6})) - (\frac{1}{m}\sum_{j=1}^{m}((\frac{bj-aj}{6})))^2 - 2(\frac{1}{m}\sum_{j=1}^{m}(\frac{bj+aj}{2}))(\frac{1}{m}\sum_{j=1}^{m}((\frac{bj-aj}{6})))$

Exp ($\mu_i$,$\beta_i$)=update exponential( $O_j$, X,w);

}

Proc. of the Second Intl. Conf. on Advances in Computer, Electronics and Electrical Engineering -- CEEE 2013

ISBN: 978-981-07-6260-5

Else

{

D: Set of tuples

Each tuple is an 'n' dimensional attribute vector X : ($x_1$,$x_2$,$x_3$,……$x_n$)

'm' Classes: $C_1$,$C_2$,$C_3$,…..$C_m$

// Naïve Bayes classifier predicts

X belongs to Class $C_i$

If $P(C_i/X) > P(C_j/X)$ 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 NBCU

# 4. Results

In this paper, we have implemented the proposed Naïve Bayesian Classifier is used to classify uncertain dataset. Naïve Bayesian Classifier has been implemented in Weka[10] 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 NBCU is applied on certain data, it works as the naïve Bayesian classification (NB), which has very good results. After applying the NBCU 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 normal distribution [5,6,7,8]. The uncertain interval is generated near by the original value, which is the center point of the interval. In [9], every numerical value is converted in to 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.

# 5. Conclusion

In this paper, we propose a novel Naïve Bayesian Classifier for Uncertain Data ( NBCU), which is used to classify and predict the given uncertain dataset. Uncertain data are widely obtainable in modern applications such as sensor network databases, biometric information systems, data integration, data extraction and risk management. Instead of trying to eliminate uncertainty and noise from datasets, in this paper follows the novel approach of directly mining the uncertain data. We integrate the uncertain data model with Bayes theorem and propose new techniques to calculate conditional probabilities for the uncertain numerical attributes and to calculate the mean and variance of the exponential distribution. Our experimental evaluation shows that the classifiers for uncertain data can be efficiently classify and predict for highly uncertain data also.

# 6. References

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

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

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

[4] 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, *The 2009 edition of the IEEE International Conference on Data Mining*

ISBN: 978-981-07-6260-5

*series (ICDM 2009)*, pages 944–949, Miami, FL, USA, 6–9 December 2009. IEEE Computer Society DOI:10.1109/ICDM.2009.90. ISBN 978-0-7695-3895-2.

[5]H. H. Bock, E. Diday, Analysisof symbolic data.Exploratorymethodsfor extracting statistical informationfromcomplexdata,Springer Verlag,2000.

[6]E.Diday,M.N.Fraiture,Symbolicdataanalysisandthesoda ssoftware,Wiley,2008.

[7]B.Qin,Y.Xia,F.Li,DTU:adecisiontreeforuncertaindata,i n:Proceedings ofPAKDD, 2009,pp.4-15.

[8] B.Qin, Y.Xia,S.Prbahakar,Rule induction for uncertain data, Knowledge Information Systems, inpress.

[9] 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.

[10] I. H. Witten, E. Frank, Data mining: practical machine learning tools and techniques, 2ndEdition, MorganKaufman Publishers,2005.

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

About Author (s):

**Nagaraju Devarakonda** received his B.Tech from Sri Venkateswara University, Tirupathi, in 2002. M.Tech from Jawaharlal Nehru University(JNU), New Delhi in 2005.PhD (Thesis Submitted) from Jawaharlal Nehru Technological University ,JNTU Hyderabad, 2013.Area of Interest : Intrusion Detection Systems, Data mining, Soft computing Techniques. Present Working as Assistant Professor in Dept of CSE, University College of Engineering Technology, Acharya Nagarjuna University, Ngarjuna Nagar, Guntur, India.

Nagamani Chippada received her B.Tech from Jawaharlal Nehru Technological University ,JNTU Kakinada, in 2006. M.Tech from Acharya Nagarjuna University , Nagarjuna Nagar in 2010. Present Working as Assistant Professor in Dept of CSE, Vikas College of Engineering Technology, Nunna, Vijayawadar, India. Her area of interest is DataMining, Network Security.

**SHAIK SUBHANI received** first-class-first honors Bachelor of Technology from the Acharya Nagarjuna University of Guntur, Andhra Pradesh in 2011 and right now he is doing his M.Tech in computer Science and Engineering from the same university. His research areas include Digital Image Processing, Data Mining, Computer Network and Design and analysis of algorithm. Currently he is pursuing his Research Dissertation from Andhra Pradesh State Remote Sensing application Centre (APSRAC), Hyderabad, AP, under the supervision of C V S Sandilia, Senior Scientific Officer.