A Comparative Study of Techniques for Data Classification based on Naïve Bayes

Antriksh Pandita
UG Student
Dr. D.Y.Patil School of Engineering and Technology
Savitribai Phule Pune University

Ajinkya Jadhav
UG Student
Dr. D.Y.Patil School of Engineering and Technology
Savitribai Phule Pune University

Vijay Singh
UG Student
Dr. D.Y.Patil School of Engineering and Technology
Savitribai Phule Pune University

Ashok Pawar
UG Student
Dr. D.Y.Patil School of Engineering and Technology
Savitribai Phule Pune University

Nilav Mukhopadhyay
Assistant Professor
Dr. D.Y.Patil School of Engineering and Technology
Savitribai Phule Pune University

ABSTRACT
The Naïve Bayes model is used for text classification and the data is considered by using the Naïve Bayes classifier and also the probabilistic based model. To define the discrete variable we use the multinomial distribution and for the numeric variable we use the Gaussian distribution.

In this research, graphical structure has been considered due to properties of Naïve Bayes classifier such as flexibility, energy efficient and high performance.

The main idea of classification has been introducedthat is the basic techniques for data classification which includes Naive Bayesian classifier.

General Terms
Naïve Bayes, Text representation, Multinomial distributionDiscretization,Gaussian distribution.

Keywords
I.5.3 clustering, similarity measure, H.3.1 Information Storage and Retrieval, G.1.6 Global Optimization.

1. INTRODUCTION
Nowadays various applications are developed like inspection of product quality, fraud debit card detection, spam identification and automatic data abstraction has large database therefore application of data mining is growing day by day. Data mining is the mining of large amount of data that is stored on database. Large collection of data or document from various sources like text message, chat message, blog article, books, and digital data are considered in data mining.

A significant part of the available information is stored on these text document database [8]. As amount of information or data is stored on a device in data electronic form and is used to increase the text database expand rapidly. These information are electronic mail, various paper or book electronic publication.

Most of the information of company form and government are stored electronically.Semi-structured data are stored in the most text database that is data stored on text database are neither structured nor unstructured [10].

In data mining different functions are used which mainly classifies data according to clustering, characteristic selection and their association rule for classification of data in a given classes. In classification when class labels are used to order the objects in the data collection then it is called as supervised classification. For classification it has training set in which all the objects of the databases are already associated with known class label. The classification algorithm builds a model by learning from training set.

2. Naïve Bayes Classifier
Naïve Bayes classifiers are probabilistic classifier that predicts set of probabilities by counting the various combinations of values in a particular class. It can also be termed as an independent feature model which deals with probabilistic classifier. The classification is based on strong independence assumption.

In Naïve Bayes the attribute value is independent to all other attributes irrespective to the class variable. This is also known as class conditional independence. It is used to simplify the computations. The class condition performs well in various supervised classification problems. It also has exhibited high accuracy and high speed [1].

Naïve Bayes algorithm is a probabilistic classifier based on Bayes Theorem. Therefore, let X be a data type which is described by measurements made on set of n attributes. Let H be some hypothesis such that data tuple X that is $P(H|X)$ is determined for classifications problems. Thus $P(H|X)$ is a posterior probability of H conditional X and $P(H|X)$ is considered as prior probability of X . [2]The posterior probability $P(H|X)$ is based on large information than prior probability $P(H)$ which is not dependent on X[4].

The probability of a document (D) containing the vector $V = (x_1, x_2, \ldots . x_n)$ belongs to the hypothesis H as follows

$$P(H|X) = \frac{P(X|H)P(H)}{P(X|H) + P(X|H_2)P(H_2)}$$

Where, $P(H|X)$ is considered as posterior probability and $P(H)$ is prior probability associated with hypothesis.

For n no of various hypothesis, we consider

$$P(X) = \sum_{j=1}^{n} P\left(X\middle|H_j\right) P(H_j)$$

Thus we have

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

3. Text in Naive Bayes

3.1 Naïve Bayes Text Representation

The sequence of characters is generally represented as a text where characters represent the expression of a written natural language of a text. A variety of methods for alter the character string represents a document as more manageable than that of statistical classification has developed. The feature extraction methods are used in speech recognition is similar to those used in image recognition but speech recognition is more complex compared to image recognition [5], [11].

In the formation of text forretrieval of information in a systems a different statistical and knowledge based technologies involving various amount of machine processing as well as manual processing. One of the drawback of this approach is that simple structure representation is produced without semantic domain knowledge and has been as effective as other[12], [14]. To overcome this make the common assumption that the pre-processing of the document produce a multiset of index terms which does not have internal structure.

3.2 Naive Bayes Text classifications

In Naïve Bayes text classification we use MNB model. MNB model is a probabilistic model for a probability of a text or document belongs to class is derived as

$$P(c|d) \propto p(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

Where the conditional probability is represented by $P(t_k|c)$ of term $t_k$incident in a document of a class. [3] The amount of data t_k in the correct class is measured by interpret$P(t_k|c)$. The prior probability of a document is represented by $p(c)$ that is occurring in the class.

If a document doesn't have class evidence in which class it belongs to then it chooses the one that have a prior probability.[13] The goal of the text classification in the Naïve Bayes is to find the best class for the document. In Naïve Bayes best class is the most likely.

4. PROBABILISTIC BASED-MODEL

In random way of modeling a data cluster we use the probabilistic approach to hide the variables randomly in the data set that is the value of the variables which is missed in all the records [4]. The variables that are hidden are referred as class variable.

This model is combination of different models. It consists the class variable which is hidden and is considered with its state which is processed to respond with the different clusters. For different values we have different models as we can see the multinomial distribution and Gaussian distribution these are used to discrete variables and numerical variables respectively. In this we can learn from data that is unlabeled and the EM algorithm is used for the learning which is carried out when the graphical structure and the EM structural both are fixed [6].

In this paper we are considering the fixed structure and that too in a simplest model, which leads us to the Naïve Bayes structure where the root variable makes the basic class and its attributes to the given class are conditionally independent. [7] If our graphical model is fixed, then the clustering is reduced and the data set of that instance and previous specified clusters are considered.

Let us consider the previous fixed number of clusters $(K)$and take each clusters distribution that is multinomial or Gaussian and the distribution between the clusters. The EM algorithm (expectation-maximization) is used to obtain the parameters and performs the function in the following manner

- Expectation: - In the expectation process we basically estimate the distributed hidden variable and that to based on the current setting of the parameter vector.
- Maximization: - In the maximization we consider the new distribution and maximize it to obtain new set of parameters that from the previously observed data.

Consider a scenario (fig.1) in which the text document which are generated by a combination of different models also called mixture models, parameterized by Θ.

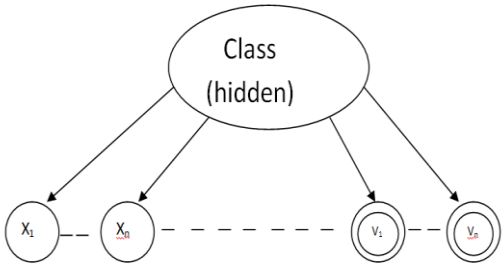

Fig.1 Graphical structure of the model

X_1... Xnrepresents discrete variables and V_1... Vn representsnumerical variable

The mixture model consists mixture of components as $c_j \in c = \{c_1, \ldots, c_{|c|}\}$every component is parameterized by a disjoint subset of Θ. Thus the document d_i is created by selecting a component according to the prior's$p\left(c_j\middle|\theta\right)$.Then having the mixture component generate a document according to its own parameter, with distribution$P\left(d_i\middle|c_j;\theta\right)$. We can characterize the document with a total probability is as follows.

$$P(d_i|\theta) = \sum_{j=1}^{|c|} P\left(c_j\middle|\theta\right)P\left(d_i\middle|c_j;\theta\right)$$

5. KERNEL DESTINY ESTIMATION

The Density Estimation is referred as an act of calculating a continuous density field from differently collected points extracted from that density field. In kernel Density estimation discretization is called as a set of cuts over domains of attributes that represents a vital pre-processing task for analysis of numeric data. When no class information is available it is called as domain. Continuous attributes must be considered in real world applications but the machine learning (ML) algorithms need a discrete feature space [9].

In this paper unsupervised discretization is used which is based on non-parametric density estimators that can automatically adapt sub interval dimension to data. The algorithm looks for next two sub-intervals to generate best cut point based on kernel density estimator for each particular sub interval.

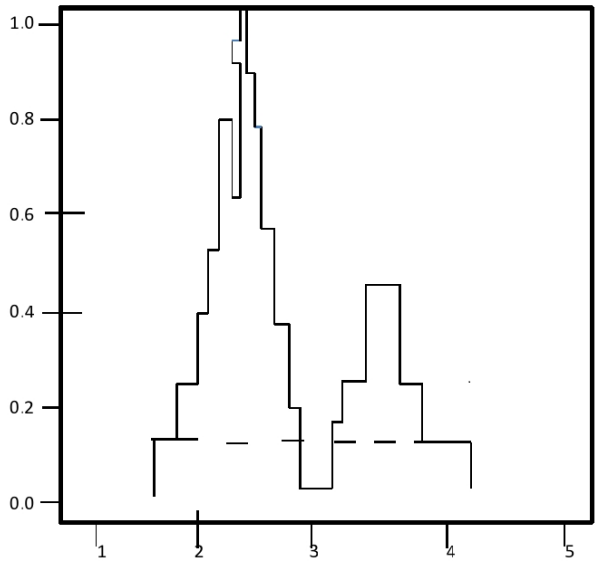

Fig.2 Placing a box for every instance in the interval and adding them up

Y-axis represents the probability density function and the X-axis represents the interval around.

5.1 Selection Criteria For Cut-Points

We can cut the middle points between instances values .As in supervised top-down discretization cut is exactly at point in main interval to separate that represents instances of data. A mentioned earlier cut is in middle point there is no need of deciding in which bin the cut will be included that is in right sub interval or left sub interval. This can be considered as the main advantage. The cut must be done based on objective of capturing the significant changes of density in various separated bins. Every sub interval generated by a cut has as averaged bin density which is completely differently from density estimated with kernel function [9].

5.2 Cut-points scoring function

In each particular step of discretization process, choice is from different sub intervals to split. Each sub intervals cut points can be defined as the middle points between instances. Score of the each cut point (T) can be computed as follows.

$$Score(T) = \sum_{i=1}^{k} (p(x_i) - f(x_i)) + \sum_{i=k+1}^{n} (p(x_i) - f(x_i))$$

Where i= 1,2 …. k refers to interval that goes in left sub interval and $i = k+1, k+2 \ldots\ldots n$ refer to interval that goes in right sub interval/bin. Where p is kernel density function and f is simple binning density function .These functions can be given as

$$f(x_i) = \frac{m}{w * N}$$

m is number of instances that goes in left or right bin. w is bandwidth and N is the number of instances in an interval that is being split. Therefore, the kernel density estimator is given as.

$$p(x_i) = \frac{1}{hN} \sum_{j=1}^{N} K(\frac{x_i - X_j}{h})$$

Where K is a kernel function.

6. PERFORMANCE ANALYSIS OF VARIOUS PREDICTION TECHNIQUES

Out of all the three classifiers discussed above, multinomial based model which basically works on probabilistic model is used to derive text .The main objective of this model is to search the best class for a text or a document.

In Probabilistic based model, modeling is performed on text clusters. This approach basically uses the discrete variables and the numeric variables .This approach is used for dataset having hidden variables.

In Kernel density the modeling is done underdiscretization, it is based on non parametric density estimators and provides the best cut on density. Methods like equal frequency or equal width binning are not well organized therefore there is need of discretization.

7. FUTURE SCOPE

We can use a model which is combination of probabilistic based model and the multinomial model in order to use both type of data that is using the multinomial model for the search of the best class for text and probabilistic model for forming the data clusters.

By this we can provide the efficiency to the model that is proposed by forming the combination of the two models. Since the probabilistic model provides the best efficiency for numeric data, this model can be used in RTI for various applications like rating of roads, buildings, quality of water supply, etc.

8. CONCLUSION

Thus we have compared all the three models used for prediction in unsupervised data and also studied the basic concepts of various models used for prediction in unsupervised data.

We also find out the advantages of all models and how these models can be used for different purposes.

9. REFERENCES

[1] An Effective Algorithm for Improving the Performance of Naive Bayes for Text Classification, GuoQiang Higher Vocational College Shanghai University of Engineering Science Shanghai, China

[2] Naive Bayes Classification of Uncertain Data,Jiangtao Ren∗, Sau Dan Lee†, Xianlu Chen∗, Ben Kao†, Reynold Cheng† and David Cheung† ∗Department of Computer

Science, Sun Yat-sen University, Guangzhou, 510275, China

[3] Clustering Unstructured Text Documents Using Naïve Bayesian Concept and Shape Pattern Matching. International Journals of Computer Application and Technology (IJCAT) 2012

[4] Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal of Computer Science and Applications 2013.

[5] Naïve (Bayes) at Forty: The independence assumption in information retrieval David .D.Lewis AT&T Labs – Research 180 park avenue Florham Park , NJ 07932-0971 USA

[6] Unsupervised naive Bayes for data clustering with mixtures of truncated exponentials. International Journal of Approximate Reasoning, 2012.

[7] Performance Comparison of Naive Bayes and J48 Classification Algorithms, Published in IJAER, Vol. 7, No. 11, 2012.

[8] Data mining, Introductory and Advanced Topics, Person education, 1st edition, 2006.

[9] Unsupervised Discretization Using Kernel Density EstimationMarenglen Biba, Floriana Esposito, Stefano Ferilli, Nicola Di Mauro, Teresa M.A BasileDepartment of Computer Science, University of BariVia Orabona 4, 70125 Bari, Italy

[10] J. Han, M. Kamber, "Data Mining: Concepts andTechniques", Second Edition, Elsevier Inc., Rajkamal Electric Press, 2006, pp. 1-628.

[11] L. Yanjun, L. Congnan, S. M. Chung, "Text Clustering with Feature Selection by Using Statistical Data", IEEE Transactions on Knowledge and Data Engineering, IEEEJournal, Volume 20, Issue 5, May 2008, pp. 641-652.

[12] L. Xinwu, "Research on Text Clustering Algorithm Based on *k*-means and SOM",International Symposium on Intelligent Information Technology Application Workshops 2008, IITAW 2008, 21-22 Dec. 2008, pp. 341-344.

[13] X. Liu, P. He, H. Wang, "The Research of Text Clustering Algorithms Based on Frequent Term Sets", Proc. 2005 International Conference on Machine Learning and Cybernetics 2005, Volume 4, 18-21 Aug., 2005, pp. 2352-2356.

[14] Gerard Salton and Michael J.McGill. Introduction to modern information retrieval. McGraw –Hill book company, NewYork, 1983.