IJASRM
International Journal of Advanced Scientific Research and Management, Vol. 1 Issue 4, April 2016.
www.ijasrm.com
ISSN 2455-6378

Development of algorithm based on Bayesian model of Classification

Meenu Jat[1], Om Prakesh Sharma[2]

[1]PG Scholar, TIT-College,
Bhopal, MP, India

[2] CSE Department, TIT-College,
Bhopal, MP, India

Abstract

Data Mining is evolving to provide automated analysis solutions and is defined as the non- trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data and describing them in a concise and meaningful manner. This work consider two common interpretations of structured data: the occurrence of relations between categories of the units of analysis, that is, between the principal entities of a statistical study (categorization structure) or the occurrence of relations between the units of analysis and/or the units observation, that is, the secondary entities of the statistical study that are correlated with the units of analysis (unit structure).This work face and deeply investigate the problem of Naive Bayesian learning from these two forms of structured data. In particular, for the case of categorization structure, It propose a framework for the usage of Naive Bayes classifiers in the case of hierarchically related categories, while, for the case of unit structure, and resort to a multi-relational approach to Data Mining.

Keywords: *Data Mining, Naïve Bayes Classification, Classification, Structured Analysis, PCA, Learning.*

1. Introduction

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute – not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how "good" the algorithm is. For example, in a medical database the training set would have relevant patient information recorded previously, where the prediction attribute is whether or not the patient had a heart problem. Given data D , we can assert

$D_1 \rightarrow M_1 \ldots D_i \rightarrow M_i$

How do we learn these? In what other forms, knowledge could be extracted from the data given?

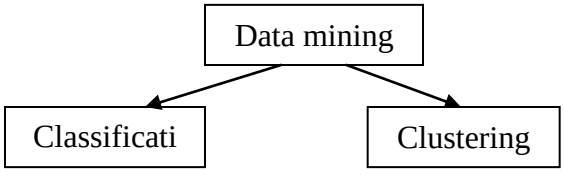

Fig.1Data mining

IJASRM International Journal of Advanced Scientific Research and Management, Vol. 1 Issue 4, April 2016.
www.ijasrm.com
ISSN 2455-6378

2. Background and Motivation

The Naive Bayesian Classifier

The approach to classification taken here is to calculate the probabilities of the different classes given some observed evidence. If the objective is to make as few classification errors as possible, the class with the highest probability should be selected as the classification result. This gives what is called the optimal Bayesian classification [Duda and Hart, 1973]. We thus want to find P(y j x), the probability of a class y 2 Y given an attribute x 2 X. (X is a random variable with the possible attribute values as outcomes, and Y is a random variable with the different classes as outcomes.) If there are few enough possible values of x, and a large enough set of training samples, it is of course possible to estimate this probability directly from the training data: for each value of x find all training samples with that value, and count how many of them belongs to each class. This requires that there are at least a few samples of each possible x. However, in most situations it is more natural to deal with P(x j y), the probability of the attribute value x of a certain class y. There may be some information on what each class \looks like", i. e. about the distribution of x for each class. To calculate P(y | x) Bayes theorem for conditional probabilities can then be used

P(y | x) = P(y) P(x | y) / P(x) ……. (1)

Suppose now that we are given the values of N input attributes, $x = \{x_1, x_2, \ldots x_N\}$, which can be considered independent both unconditionally and conditionally given y. This means that the probability of the joint outcome x can be written as a product,

$P(x) = P(x_1).P(x_2) \ldots \ldots P(x_N)$……..2

And so can the probability of x within each class y,

$P(x|y) = P(x_1|y).P(x_2|y) \ldots P$……..3

With the help of these it is possible to write

$P(y|x) = P(y)\ \frac{P(x|y)}{P(x)} = P(y)\prod_{i=1}^{N}\frac{P(x_i|y)}{P(x_i)}$..4

The basis for the *Naïve Bayesian Classifier* the designation naïve is due to the sometimes too simplistic assumption that different input attributes are independent.

$$\frac{P(x_i|y)}{P(x_i)} = \frac{P(x_i, y)}{P(y)P(x_i)} = \frac{P(y|x_i)}{P(y)}$$

The expression to the right can be interpreted as how many more times likely the class becomes if we get to know the feature xi. The middle expression shows that the contribution is symmetric; i. e. the contribution from a feature to a class in this sense is equally large as the contribution from the class to the feature. By taking the logarithm of Eq. (4), it may be written as a sum:

$$\log\ P(y \mid x) = \log\ P(y) + \sum_i \log\left(\frac{P(y, x_i)}{P(y)P(x_i)}\right) ..5$$

An advantage of this form is that it is a linear expression in the contribution from the attributes. This means that the naive Bayesian classifier can be implemented with linear discriminant functions [Minsky, 1961], and thus in the form a one-layer neural network.

3. Literature Survey

Michael L. Raymer, Leslie A. Kuhn, and William F. Punch [1] defines that a key element of many bioinformatics research problems is the extraction of meaningful information from large experimental data sets. Proposed methodology:

A. *Bayesian Discriminant Functions*
B. *Nonlinear Weighting of the Bayes Discriminant Function*
C. *Gaussian Smoothing*
D. *GA Optimization of the Nonlinear Discriminant Coefficients*
E. *Representation Issues and Maskin*

Ishtiaq Ahmed, Donghai Guan, and Tae Choong Chung [2] proposes a hybrid system of SMS classification to detect spam or ham, using Naïve Bayes classifier and Apriori algorithm. Though this technique is fully logic based, its performance will rely on statistical character of the database. There are several steps for text classification and each of them is discussed below:

A. *Loading Database*
B. *Feature Extraction*
C. *Vector Creation and Training*
D. *Running the Naïve Bayes System*

Ms S. Vijayarani1, Ms M. Muthulakshm[3] Data mining is the non-trivial extraction of implicit, earlier unknown and potentially useful information about data. There are several data mining techniques have been developed and used in data mining projects which includes classification, clustering, association rules, prediction, and sequential patterns.

Doreswamy, Hemanth. K. S [4] In this paper, naive Bayesian and C4.5 Decision Tree Classifiers (DTC)

IJASRM
International Journal of Advanced Scientific Research and Management, Vol. 1 Issue 4, April 2016.
www.ijasrm.com
ISSN 2455-6378

are successively applied in materials informatics to classify the engineering materials into different classes for the selection of materials that suit the input design specifications. Here, the classifiers are analyzed individually and their performance evaluation is analyzed with confusion matrix predictive parameters and standard measures, the classification results are analyzed on different class of materials.

4. Proposed Method

Our proposed method defines about various complexes

I. Data Sampling

Sampling is the process by which inference is made to the whole by examining a part. The purpose of sampling is to provide various types of statistical information of a qualitative or quantitative nature about the whole by examining a few selected units. The sampling method is the scientific procedure of selecting those sampling units which would provide the required estimates with associated margins of uncertainty, arising from examining only a part and not the whole.

II. Feature Selection

The primary purpose of feature selection is to reduce the dimensionality to decrease the computation time. This is particularly important concerning text categorization where the high dimensionality of the feature space is a problem. In many cases the number of features is in the tens of thousands. Then it is highly desirable to reduce this number, preferably without any loss in accuracy.

III. Rank

Evaluating the statistical significance of the ranking is important for understanding the results and for further investigations, but this question has not been well addressed for learning classification methods in existing works. Here, we address this problem by formulating it in the framework of hypothesis testing and propose a solution based on FCS method. Let for each value i g in G, we firstly calculate the mean μ_i^+ (resp. μ_i^-) and standard deviation σ_i^+ (resp. σ_i^-) which correspond to the gene g_i of samples labeled +1(-1), respectively. Then we calculate a feature score $F(gi) = |(\mu_i^+ - \mu_i^-) / (\sigma_i^+ + \sigma_i^-)$ for each $g_i \in G$, and rank the according to their score values. At last, we simply take the the highest $F(g_i)$ scores as our top-ranking values G_{top}, satisfying $|G_{top}| << |G|$. After selecting p values, we may obtain M_{mxp}

Algorithm

Step 1: Get data g_i in G.
Step 2: calculate the mean μ_i^+ (resp. μ_i^-)
Step 3: Calculate the standard deviation σi^+ (resp. σ_i^-)
Step 4: Calculate the feature score $F(gi) = (\mu_i^+ - \mu_i^-) / (\sigma_i^+ + \sigma_i^-)$, for all $g_i \in G$
Step 5: assign rank $F(g_i)$ according to their score value top-ranking genes G_{top}, satisfying $|G_{top}| << |G|$

IV. Naive Bayesian Classifier Method

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

5. Flow Diagram

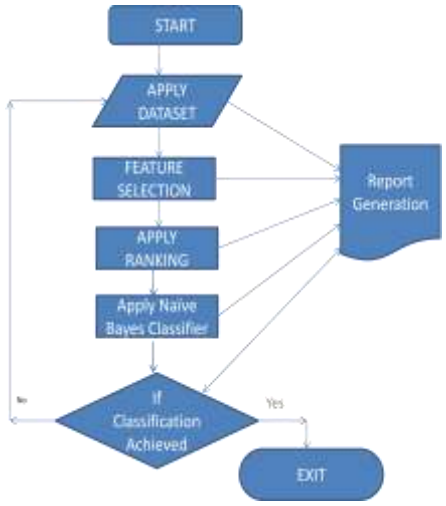

Fig 2 Flow Diagram

IJASRM
International Journal of Advanced Scientific Research and Management, Vol. 1 Issue 4, April 2016.
www.ijasrm.com
ISSN 2455-6378

6. Experiment Method

Experiment method for the implementation is follows,

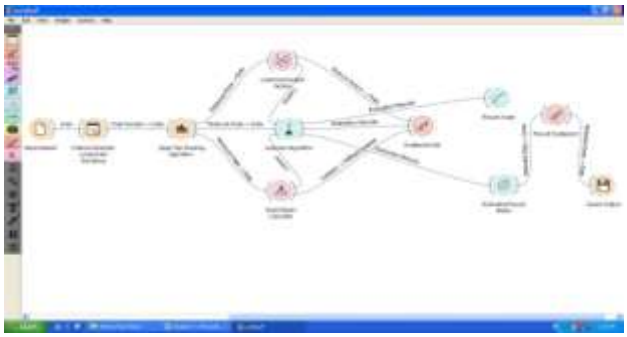

Fig 3 Experimental Model

Step1 Collection of samples: Samples are collected through various methods here in favor of our experiment we collect pre analyzed gene express profiled dataset from lab

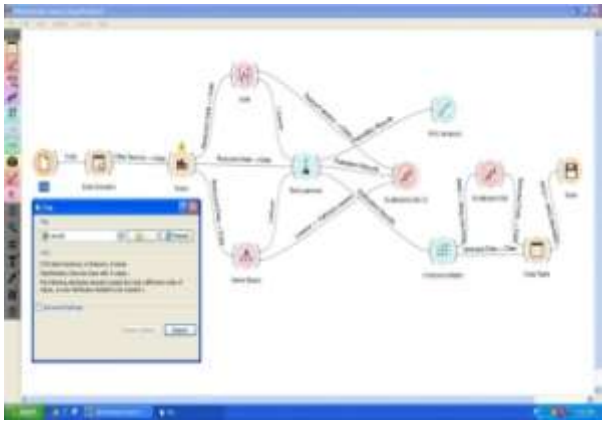

Fig 4 Applying Dataset

i. Feature Selection

A number of techniques have been developed to address the problem of dimensionality, including feature selection and feature extraction. The main purpose of feature selection is to reduce the number of features used in classification while maintaining acceptable classification accuracy.

ii. Feature Extraction

Feature extraction, a superset of feature selection, involves transforming the original set of features to provide a new set of features, where the transformed feature set usually consists of fewer features than the original set. While both linear and non-linear transformations have been explored, most of the classical feature extraction techniques involve linear transformations of the original features.

iii. Bayesian Discriminate Functions

The Bayesian classifier has a computational advantage over the previously-employed knn classifier in that the training data are summarized, rather than stored. The comparison of each test sample with every known training sample to find nearest neighbors during knn classification is a computationally expensive process, even when efficient search methods are employed. In contrast the Bayes decision rule is invariant to linear scaling of the feature space. In other words, multiplying the feature values for a given feature by a constant has no effect on the class-conditional probabilities considered by the classifier.

7. RESULTS

Table 1

Evaluated Result Matrix Sun Oct 18 15, 15:23:14					
Contents: Based on number of example for correct selections					
Learner: Naive Bayes					
Data: Number of examples					
Matrix					
	unacc	**acc**	**good**	**v-good**	
unacc	**101**	**11**	**0**	**0**	**112**
acc	**17**	**18**	**4**	**0**	**39**
good	**0**	**3**	**11**	**0**	**14**
v-good	**0**	**2**	**1**	**5**	**8**
	118	**34**	**16**	**5**	**173**
Note: columns represent predictions, row represent true classes					

IJASRM International Journal of Advanced Scientific Research and Management, Vol. 1 Issue 4, April 2016.
www.ijasrm.com
ISSN 2455-6378

Table 2

Evaluated Result Matrix Sun Oct 18 15, 15:23:25					
Contents: Based on number of example for misclassified selections					
Learner: Naive Bayes					
Data: Number of examples					
Matrix					
	unacc	acc	good	v-good	
unacc	101	11	0	0	112
acc	17	18	4	0	39
good	0	3	11	0	14
v-good	0	2	1	5	8
	118	34	16	5	173
Note: columns represent predictions, row represent true classes					

Following result defines about comparison between 2 previously used algorithms and our algorithm Train/Tune defines about training data sets Test tell about the testing methods Features contains classification features of applied data sets Prediction defines about outcomes after applying proposed methods

Learning	Naïve Bayesian	SVM	C 4.5
Train/Tune	12	15	14
Test	3	4	1
Features	113	150	457
Prediction	97.30%	94.22%	80%

8. Conclusion

The potential of applying learning techniques is very high for classification of complex datasets. We demonstrate the classification using hybrid method with the use of naïve bayesian classifiers. Automatic categorization is the task of assigning level of different categorization. In our work it's for complex and difficult decision dataset and to make this procedure in reality we have incorporated Naïve Bayes classification with the help of feature selection and ranking with little bit modification. Although this technique is logic based, but the result id depended with dataset. By applying our strategy we depicted significant improvement than the state of the art algorithm. Our supervised machine learning system for handling and organizing system and by performing our proposed strategy this technique have reached accuracy levels that can outperform even the state of the art algorithm. Further, the classification accuracy can be improved by employing noise removal techniques for eliminating outliers in data sets. We apply our method for biological predications like protein structuring and smart system decision making using machine learning. We will try our approach with particle swarm optimization methods to understand the behavior and properties of swarm domain.

References

[1] T. Bayes. *An essay towards solving a problem in the doctrine of chances*. Phil. Trans. Roy. Soc., 53, (1763).
[2] C. L. Blake and C. J. Merz. *UCI repository of machine learning databases. University of California, Irvine, Dept. of Information and Computer Sciences*, (1998). ttp://www.ics.uci.edu/_mlearn/MLRepository.html.
[3] T. M. Cover and J. M. V. Campenhout. *On the possible orderings in the measurement selection problem. IEEE Transactions on Systems, Man, and Cybernetics*, 7:657–661, (2011).
[4] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, IT- 13:21–27, (1999).
[5] Ashok Savasere, Edward Omiecinski, Shamkant Navathe ―*An Efficient Algorithm for Mining Association Rules in Large Databases*‖ College of Computing Georgia In&ute of Technology Atlanta, GA 30332
[6] Linyu Yang, Dwi H. Widyantoro, Thomas Ioerger, John Yen ―*An Entropy-based Adaptive Genetic Algorithm for Learning Classification Rules*‖, IEEE Transactions on Computer Science, PP 13:21–27, (2014).
[7] Fadila Bentayeb J´erˆome, Darmont C´edric Udr´ea France *Efficient Integration of Data Mining Techniques in Database Management Systems ERIC*, University of Lyon 2 5 avenue Pierre Mend`es-France 69676 Bron Cedex
[8] A. K. Jain and B. Chandrasekaran, ―*Dimensionality and sample size considerations in pattern recognition in practice*,‖ in *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, Eds. vol. 2, pp. 835–855, (1982) North-Holland.
[9] G. V. Trunk, ―*A problem of dimensionality: A simple example*,‖ *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, vol. 1, pp. 306–307, (1979).
[10] T. M. Cover and J. M. Van Campenhout, ―*On the possible orderings in the measurement selection problem*,‖ IEEE Transactions on Systems, Man, and Cybernetics, vol. 7, pp. 657–661, (1977).
[11] P. M. Narendra and K. Fukunaga, ―*A branch and bound algorithm for feature subset selection*,‖ IEEE Transactions on Computers, vol. C-26, pp. 917–922 (1977).
[12] A. Whitney, ―*A direct method of nonparametric measurement selection*,‖ IEEE Transactions on Computers, vol. 20, pp. 1100–1103, (1971).
[13] P. Pudil, J. Novovicova, and J. Kittler, *"Floating search methods in feature selection,*‖ Pattern Recognition Letters, vol. 15, pp. 1,119–1,125, Nov. (1994).
[14] A. K. Jain and D. Zongker, *"Feature selection: Evaluation, application, and small sample performance*,‖ IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 2, pp. 153–158, February (1997).
[15] J. Mao, K. Mohiuddin, and A. K. Jain, ―*Parsimonious network design and feature selection through node pruning*,‖ in Proc. of the Intl. Conf. on Pattern Recognition, Jerusalem, , pp. 622–624, October (1994).
[16] K. Nozaki, H. Ishibuchi, and H. Tanaka, ―*Adaptive fuzzy rulebased classification systems*,‖ IEEE Transactions on Fuzzy Systems, vol. 4, no. 3, pp. 238–250, August (1996).
[17] H. Ishibuchi, K. Nozaki, N. Yamamoto, and H. Tanaka, ―*Selecting fuzzy if-then rules for classification problems using genetic algorithms*,‖ IEEE Transactions on Fuzzy Systems, vol. 3, no. 3, pp. 260–270, August (1995).
[18] J. R. Quinlan, *"Induction of decision trees, Machine Learning*, vol. 1, pp. 81–106, (1986).
[19] J. R. Quinlan, ―*Simplifying decision trees*,‖ International Journal of Man-Machine Studies, pp. 221–234, (1987).