Medical datamining with a new algorithm for Feature Selection and Naïve Bayesian classifier

Ranjit Abraham,
Azhikakathu, Tripunithura, Cochin, INDIA.
ranjit.abraham@gmail.com

Jay B.Simha,
ABIBA Systems, Bangalore, INDIA.
jay.b.simha@abibasystems.com

Iyengar S.S
Dept. of Computer Science, Louisiana State University, Baton Rouge, USA.
iyengar@bit.cse.lsu.edu

Abstract

Much research work in datamining has gone into improving the predictive accuracy of statistical classifiers by applying the techniques of discretization and feature selection. As a probability-based statistical classification method, the Naïve Bayesian classifier has gained wide popularity despite its assumption that attributes are conditionally mutually independent given the class label. In this paper we propose a new feature selection algorithm to improve the classification accuracy of Naïve Bayes with respect to medical datasets. Our experimental results with 17 medical datasets suggest that on an average the new CHI-WSS algorithm gave best results. The proposed algorithm utilizes discretization and simplifies the' wrapper' approach based feature selection by reducing the feature dimensionality through the elimination of irrelevant and least relevant features using chi-square statistics. For our experiments we utilize two established measures to compare the performance of statistical classifiers namely; classification accuracy (or error rate) and the area under ROC to demonstrate that the proposed algorithm using generative Naïve Bayesian classifier on the average is more efficient than using discriminative models namely Logistic Regression and Support Vector Machine.

1. Introduction

Medical applications of data mining include prediction of the effectiveness of surgical procedures, medical tests and medications, and discovery of relationships among clinical and pathological data. In the last few years, the digital revolution has provided relatively inexpensive and available means to collect and store large amounts of patient data in databases containing rich medical information and made available through the Internet for Health services globally. Data mining techniques applied on these databases discover relationships and patterns that are helpful in studying the progression and the management of diseases [21].

Several computer programs developed to carry out optimal management of data for extraction of knowledge or patterns contained in the data include Expert systems, Artificial Intelligence and Decision support systems. One such program approach has been data classification with the goal of providing information such as if the patient is suffering from the illness or not from a case or collection of symptoms. Particularly, in the medical domain high classification accuracy is desirable.

Based on the theory of Bayesian networks, Naïve Bayes is a simple yet consistently performing probabilistic model. Data classification with naïve Bayes is the task of predicting the class of an instance from a set of attributes describing that instance and assumes that all the attributes are conditionally independent given the class. It has been shown that naïve Bayesian classifier is extremely effective in practice and difficult to improve upon [5].

Many factors affect the success of machine learning on medical datasets. The quality of the data is one such factor. If information is irrelevant or redundant or the data is noisy and unreliable then knowledge discovery during training is more difficult. Feature selection is the process of identifying and removing as much of the irrelevant and redundant information as possible [15,17]. Regardless of whether a learner attempts to select features itself or ignores the issue, feature selection prior to learning can be beneficial. Reducing the dimensionality of the data reduces the size of the hypothesis space and allows algorithms to operate faster and more effectively. The performance of the naïve Bayes classifier is a good candidate for analyzing feature selection algorithms since it does not perform implicit feature selection like decision trees.

In this paper, we show that it is possible to reliably improve the naïve Bayesian classifier by applying a new feature selection algorithm that is both simple and effective.

0-7695-3068-0/07 $25.00 © 2007 IEEE
DOI 10.1109/ICIT.2007.41

2. Naïve Bayes and NB Classifier

Naïve Bayes, a special form of Bayesian network has been widely used for data classification in that its predictive performance is competitive with state-of-the-art classifiers such as C4.5 [7]. As a classifier it learns from training data from the conditional probability of each attribute given the class label. Using Bayes rule to compute the probability of the classes given the particular instance of the attributes, prediction of the class is done by identifying the class with the highest posterior probability. Computation is made possible by making the assumption that all attributes are conditionally independent given the value of the class. Naïve Bayes as a standard classification method in machine learning stems partly because it is easy to program, its intuitive, it is fast to train and can easily deal with missing attributes. Research shows naïve Bayes still performs well in spite of strong dependencies among attributes [5].

Naïve Bayes is best understood from the perspective of Bayesian networks. Bayesian networks (BN) graphically represent the joint probability distribution of a set of random variables. A BN is an annotated directed acyclic graph that encodes a joint probability distribution over a set of attributes X. Formally a BN for X is a pair B= <G,θ>, where G represents the directed acyclic graph whose nodes represent the attributes X1, X2,..Xn and whose edges represent direct dependencies between the attributes. The BN can be used to compute the conditional probability of a node given values assigned to the other nodes. The BN can be used as a classifier where the learner attempts to construct a classifier from a given set of training examples with class labels. Here nodes represent dataset attributes.

Assuming that X_1, X_2,..X_n are the n attributes corresponding to the nodes of the BN and say an example E is represented by a vector x_1, x_2,..x_n where x_1 is the value of the attribute X_1. Let C represent the class variable and c its value corresponding to the class node in the Bayesian network, then the class c of the example E (c(E)) can be represented as a classifier by the BN [9] as

$$c(E) = \arg\max_{c \in C} p(c)\, p(x_1, x_2, ..x_n \mid c) \quad (1)$$

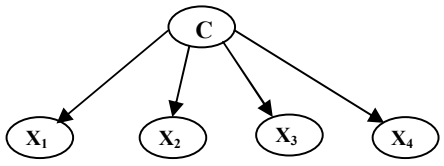

Fig. 1: Structure of naïve Bayes

Although Bayesian networks can represent arbitrary dependencies it is intractable to learn it from data. Hence learning restricted structures such as naïve Bayes is more practical. The naïve Bayesian classifier represented as a BN has the simplest structure. Here the assumption made is that all attributes are independent given the class and equation 1 takes the form

$$c(E) = \arg\max_{c \in C} p(c) \prod_{i=1}^{n} p(x_i \mid c) \quad (2)$$

The structure of naïve Bayes is graphically shown in Fig.1. Accordingly each attribute has a class node as its parent only. The most likely class of a test example can be easily estimated and surprisingly effective [5]. Comparing naïve Bayes to Bayesian networks, a much more powerful and flexible representation of probabilistic dependence generally did not lead to improvements in accuracy and in some cases reduced accuracy for some domains [19].

3. Implementing the NB Classifier

Considering that an attribute X has a large number of values, the probability of the value P(X=xi |C=c) from equation 2 can be infinitely small. Hence the probability density estimation is used assuming that X within the class c are drawn from a normal (Gaussian) distribution

$$\frac{1}{\sqrt{2\pi}\sigma_c} e^{-\frac{(x_i-\mu_c)^2}{2\sigma_c^2}}$$

where σ_c is the standard deviation and μ_c is the mean of the attribute values from the training set [6]. The major problem with this approach is that if the attribute data does not follow a normal distribution, as often is the case with real-world data, the estimation could be unreliable. Other methods suggested include the kernel density estimation approach [12]. But since this approach causes very high computational memory and time it does not suit the simplicity of naïve Bayes classification.

When there are no values for a class label as well as an attribute value, then the conditional probability P(x|c) will be also zero if frequency counts are considered. To circumvent this problem, a typical approach is to use the Laplace-m estimate [2]. Accordingly

$$P(C = c) = \frac{n_c + k}{N + n \times k}$$

where n_c = number of instances satisfying C=c, N = number of training instances, n= number of classes and k=1.

$$P(X = x_i | C = c) = \frac{n_{ci} + m \times P(X = x_i)}{n_c + m}$$

where n_{ci} = number of instances satisfying both $X=x_i$ and C=c, m=2 (a constant) and $P(X=x_i)$ estimated similarly as P(C=c) given above.

We also need to consider datasets that have a few unknowns among the attribute values. Although

unknowns can be given a separate value [4], we have chosen to ignore them in our experiments.

4. MDL discretized Naïve Bayes

Discretization is the process of transforming data containing a quantitative attribute so that the attribute in question is replaced by a qualitative attribute [25]. Data attributes are either numeric or categorical. While categorical attributes are discrete, numerical attributes are either discrete or continuous. Research study shows that naïve Bayes classification works best for discretized attributes and discretization effectively approximates a continuous variable [3].

The Minimum Description Length (MDL) discretization is Entropy based heuristic given by Fayyad and Irani [8]. The technique evaluates a candidate cut point between each successive pair of sorted values. For each candidate cut point, the data are discretized into two intervals and the class information entropy is calculated. The candidate cut point, which provides the minimum entropy is chosen as the cut point. The technique is applied recursively to the two sub-intervals until the criteria of the Minimum candidate cut point, the data are discretized into two intervals and the class information entropy is Description Length.

For a set of instances S, a feature A and a partition boundary T, the class information entropy of the partition induced by T is given by

$$\mathrm{E(A,T,S)} = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

and

$$Ent(S) = -\sum_{i=1}^{C} P(C_i, S)\log_2(C_i, S)$$

For the given feature the boundary Tmin that minimizes the class information entropy over the possible partitions is selected as the binary discretization boundary. The method is then applied recursively to both partitions induced by Tmin until the stopping criteria known as the Minimum Description Length (MDL) is met. The MDL principle ascertains that for accepting a partition T, the cost of encoding the partition and classes of the instances in the intervals induced by T should be less than the cost of encoding the instances before the splitting. The partition is accepted only when

$$Gain(A,T,S) \; \rangle \; \frac{\log_2(N-1)}{N} + \frac{\Delta(A,T,S)}{N}$$

where

$$\Delta(A,T,S) = \log_2\left(3^c - 2\right) - cEnt(S) - c_1 Ent(S_1) - c_2 Ent(S_2)$$

and

$$Gain(A,T,S) = Ent(S) - E(A,T,S)$$

N = number of instances, c,c1,c2 are number of distinct classes present in S, S1 and S2 respectively.

MDL discretized datasets show good classification accuracy performance with naïve Bayes [23].

5. Feature Selection for NB Classifier

Feature selection is often an essential data pre-processing step prior to applying a classification algorithm such as naïve Bayes. The term feature selection is taken to refer to algorithms that output a subset of the input feature set. One factor that plagues classification algorithms is the quality of the data. If information is irrelevant or redundant or the data is noisy and unreliable then knowledge discovery during training is more difficult [15,17]. Regardless of whether a learner attempts to select features itself or ignores the issue, feature selection prior to learning can be beneficial. Reducing the dimensionality of the data reduces the size of the hypothesis space and allows algorithms to operate faster and more effectively. In some cases accuracy on classification can be improved [15]. As a learning scheme naïve Bayes is simple, very robust with noisy data and easily implementable. We have chosen to analyze feature selection algorithms with respect to naïve Bayes method since it does not perform implicit feature selection like decision trees.

Algorithms that perform feature selection as a preprocessing step prior to learning can generally be placed into one of two broad categories [11]. One approach referred to as the 'wrapper' employs as a subroutine a statistical resampling technique such as cross validation using the actual target learning algorithm to estimate the accuracy of feature subsets. This approach has proved useful but is slow because the learning algorithm is called repeatedly. The other approach called the 'filter' operates independently of any learning algorithm. Undesirable features are filtered out of the data before induction commences. Although filters are suitable to large datasets they have not proved as effective as wrappers. While the filter approach is generally computationally more efficient than the wrapper approach, its major drawback is that an optimal selection of features may not be independent of the inductive and representational biases of the learning algorithm to be used to construct the classifier. The wrapper approach, on the other hand involves the computational overhead of evaluating candidate feature subsets by executing a selected learning algorithm on the dataset represented using each feature subset under consideration[22]. Hence we propose a new algorithm (CHI-WSS) that combines the advantages of the filter approach with the wrapper approach.

The CHI-WSS algorithm

1. Apply MDL (Fayyad and Irani) discretization to

all non-categorical features.

2. Identify and remove all irrelevant and least relevant features using Chi-square feature ranking. Irrelevant features pertain to those features whose chi-square average merit is zero-(features where MDL discretization levels equaled one). The least relevant features pertain to those features that satisfy the condition $100 \times \frac{avg\_merit_i \times \log(N^2)}{\Sigma avg\_merit \times N} \langle \delta$ where we set $\delta = 0.1$ to satisfy our criterion. avg_merit_i is the average merit for the feature in consideration and N is the total number of attributes.
3. From the remaining features of the dataset, identify the feature subsets using wrapper subset approach with naïve Bayes and feature subset search using BestFirst.
4. Apply Forward Sequential Search (FSS) using naïve Bayes to find the set of features that maximizes the classification accuracy.

6. Experimental Evaluation

We have used 17 natural Medical datasets obtained from public repositories for our experiments whose technical specifications are as shown in Table 1. All the chosen datasets had at least one or more attributes that

Table 1: Specifications for the Medical datasets

SL. No.	Medical Dataset	No. of Instances	Total no. of attributes	Number of Classes	Missing attr. status	Noisy attr. status
1	Wisconsin Breast Cancer [1]	699	10	2	Yes	No
2	Pima Diabetes [1]	768	9	2	No	No
3	Bupa Liver Disorders [1]	345	7	2	No	No
4	Cleveland Heart Disease [1]	303	14	2	Yes	No
5	Hepatitis [1]	155	20	2	Yes	No
6	Thyroid –new [1]	215	6	3	No	No
7	Thyroid (ann-train) [1]	3772	22	3	No	No
8	Statlog- heart [1]	270	14	2	No	No
9	Hepatobiliary disorders	536	10	4	No	No
10	Appendicitis [24]	106	9	2	Yes	No
11	Leisenring neo audiology [20]	3152	8 (7)	2	No	No
12	Norton neonatal audiology	5058	9 (7)	2	Yes	No
13	Laryngeal 1 [16]	213	17	2	No	No
14	RDS [16]	85	18	2	No	No
15	Voice_3 [16]	238	11	3	No	No
16	Voice_9 [16]	428	11	9 (2)	No	No
17	Weaning [16]	302	18	2	No	No

were continuous. For our experiments we have substituted all noisy data with unknowns. For datasets with redundant attributes and non-computational attributes (such as patient identification number), we have ignored them from our experiments. All missing attribute values were ignored.

We have used 10-fold cross validation test method to all the Medical datasets [13]. The dataset was divided into 10 parts of which 9 parts were used as training sets and the remaining one part as the testing set. The classification accuracy was taken as the average of the 10 predictive accuracy values.

Table 2 gives the results of feature selection using the proposed CHI-WSS algorithm. The wins given at the bottom of the table indicate that using the new algorithm, all the datasets saw improvement in dimensionality reduction and better classification accuracy.

Table 2: Classification accuracy before and after applying CHI-WSS algorithm

SL. No.	Medical Dataset	NB with original data	Applying CHI-WSS algorithm	
			Attributes removed using Chi-Square ranking criterion	Classification Accuracy - Forward Seq. Search
1	Wisconsin Breast Cancer	95.9943 (10)	none	97.4249 (6)
2	Pima Diabetes	76.3021 (9)	2	79.8177 (5)
3	Bupa Liver Disorders	55.3623 (7)	5	63.1884 (2)
4	Cleveland Heart Disease	83.8284 (14)	3	84.4884 (4)
5	Hepatitis	84.5161 (20)	7	88.3871 (6)
6	Thyroid –new	96.7442 (6)	none	97.6744 (5)
7	Thyroid (ann-train)	95.5196 (22)	15	98.1707 (4)
8	Statlog- heart	84.8148 (14)	3	84.8148 (4)
9	Hepatobiliary disorders	47.9478 (10)	1	69.403 (8)
10	Appendicitis	84.9057 (9)	2	90.566 (4)
11	Leisenring neo audiology	100 (7)	2	100 (2)
12	Norton neo audiology	96.0854 (7)	3	96.6983 (4)
13	Laryngeal 1	75.5869 (17)	1	87.3239 (13)
14	RDS	89.4118 (18)	6	96.4706 (7)
15	Voice_3	69.7479 (11)	1	82.3529 (3)
16	Voice_9	78.7383 (11)	none	89.7196 (5)
17	Weaning	89.7351 (18)	4	92.7152 (7)
Wins				17 / 17

Note: Given in brackets is the total number of attributes selected

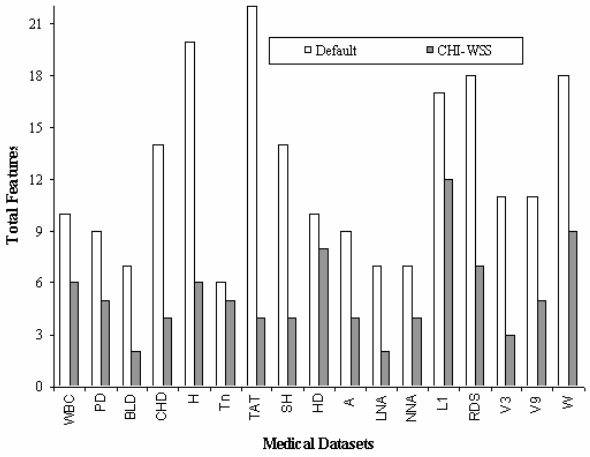

Fig. 2: Feature dimensionality reduction before and after using CHI-WSS algorithm

Fig.2 shows that feature dimensionality reduction was achieved for all the datasets using the CHI-WSS algorithm. Fig.3 shows the classification accuracy performance before and after the application of the CHI-WSS algorithm.

In order to compare the efficiency of our proposed new feature selection algorithm based on generative naïve Bayes model, we have used two established measures namely; classification accuracy (or error rate) and the area under ROC to compare the classifier performance with two other popular discriminative models such as Support Vector Machine (SVM) and Logistic regression (LR) that are used in Medical data mining.

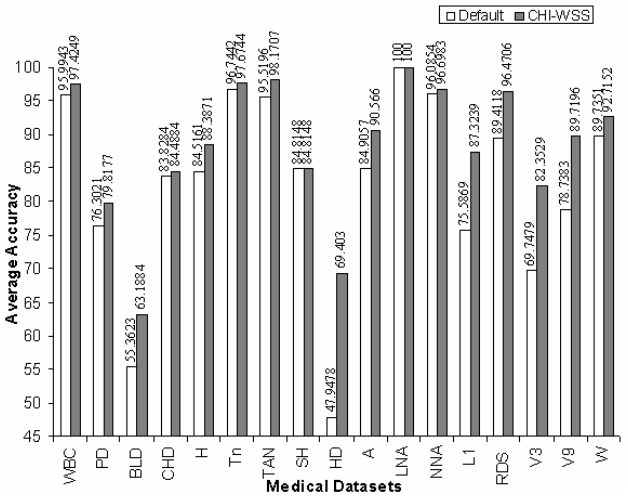

Fig. 3: Classification accuracy performance before and after using CHI-WSS algorithm

Table 3 provides the results of the classifier accuracy achieved by feature selection using our proposed CHI-WSS algorithm as well as those achieved by WrapperSubset based feature selection using discriminative models- LR and SVM. From the wins given at the bottom of the table we see that feature selection with our proposed algorithm on the average gives better performance than the discriminative methods.

Table 3: Comparative analysis of Feature selection based on Classification Accuracy

SL. No.	Medical Dataset	NB with original data	Classification Accuracy using		
			CHI-WSS with NB	Forward Seq. Search with WrapperSubset using LR *	Forward Seq. Serach with WrapperSubset using SVM*
1	Wisconsin Breast Cancer	95.9943 (10)	97.4249 (6)	96.8526 (5)	96.9957 (5)
2	Pima Diabetes	76.3021 (9)	79.8177 (5)	79.8177 (5)	78.3854 (6)
3	Bupa Liver Disorders	55.3623 (7)	63.1884 (2)	63.1884 (2)	63.1884 (2)
4	Cleveland Heart Disease	83.8284 (14)	84.4884 (4)	84.8185 (8)	84.8185 (8)
5	Hepatitis	84.5161 (20)	88.3871 (6)	85.8065 (4)	84.5161 (4)
6	Thyroid –new	96.7442 (6)	97.6744 (5)	97.6744 (4)	97.2093 (4)
7	Thyroid (ann-train)	95.5196 (22)	98.1707 (4)	98.1972 (4)	98.2503 (4)
8	Statlog- heart	84.8148 (14)	84.8148 (4)	85.1852 (5)	84.4444 (4)
9	Hepatobiliary disorders	47.9478 (10)	69.403 (8)	70.8955 (8)	69.5896 (6)
10	Appendicitis	84.9057 (9)	90.566 (4)	88.6792 (2)	88.6792 (3)
11	Leisenring neo audiology	100 (7)	100 (2)	100 (2)	100 (2)
12	Norton neo audiology	96.0854 (7)	96.6983 (4)	97.0542 (4)	97.0542 (4)
13	Laryngeal 1	75.5869 (17)	87.3239 (13)	86.385 (5)	84.507 (2)
14	RDS	89.4118 (18)	96.4706 (7)	92.9412 (6)	92.9412 (12)
15	Voice_3	69.7479 (11)	82.3529 (3)	83.6134 (4)	81.9328 (4)
16	Voice_9	78.7383 (11)	89.7196 (5)	91.5888 (7)	91.8224 (5)
17	Weaning	89.7351 (18)	92.7152 (7)	82.4503 (3)	90.7285 (11)
Wins		1 / 17	10 / 17	9 / 17	6 / 17

*all irrelevant and least relevant features were initially removed using our Chi-square feature ranking criterion

Further, in Table 4 using true positive rates given in terms of the area under ROC (AUROC), the proposed CHI-WSS feature selection with naïve Bayes gets more wins than the other methods as shown at the bottom of the table.

For the proposed feature selection algorithm (CHI-WSS), the computational overhead with naïve Bayes is much lower compared to using discriminative models such as SVM and LR.

Table 4: Comparative analysis of Feature selection based on AUROC (in percentage)

SL. No.	Medical Dataset	Classification AUROC (in %) using		
		CHI-WSS with Naïve Bayes	Forward Seq. Search with WrapperSubset using LR *	Forward Seq. Search with WrapperSubset using SVM *
1	Wisconsin Breast Cancer	99.21	99.36	96.92
2	Pima Diabetes	84.34	84.17	73.27
3	Bupa Liver Disorders	55.95	55.95	61.99
4	Cleveland Heart Disease	88.46	86.43	84.44
5	Hepatitis	82.61	76.27	69.44
6	Thyroid –new	99.32	99.2	96.69
7	Thyroid (ann-train)	99.95	99.95	99.93
8	Statlog- heart	87.99	86.01	83.92
9	Hepatobiliary disorders	86.67	87.57	81.59
10	Appendicitis	78.18	68.12	78.60
11	Leisenring neo audiology	100	100	100
12	Norton neo audiology	57.19	50.00	50.00
13	Laryngeal 1	93.79	90.05	83.72
14	RDS	98.78	97.28	92.92
15	Voice_3	90.97	88.90	91.70
16	Voice_9	93.44	96.14	90.15
17	Weaning	95.85	87.16	90.73
Wins		11 / 17	5 / 17	4 / 17

*all irrelevant and least relevant features were initially removed using our Chi-square feature ranking criterion

7. Conclusion

In this research work an attempt was made to evaluate feature selection with naïve Bayes classifier that could be used for medical data mining. Our experimental results indicate that, on an average, with the proposed CHI-WSS algorithm utilizing naïve Minimum Description Length (MDL) discretization, Chi-square feature selection ranking and wrapper approach, provides on the average better accuracy performance and feature dimensionality reduction. When compared to the use of discriminative models such as Logistic Regression and Support Vector Machines employing wrapper based approach for feature selection, our proposed algorithm on the average gives better performance with much reduced computational overhead. The new feature selection algorithm is simple as well as effective and augments the argument that simple methods are better in medical data mining.

8. References

[1] Blake C.L, Merz C.J., "UCI repository of machine learning databases". [http://www.ics.uci.edu/~mlearn/ML Repository.html], Department of Information and Computer Science, University of California, Irvine.

[2] Cestnik B., "Estimating probabilities: A crucial task in machine learning", *In Proceedings of the 9th European Conference on Artificial Intelligence,* pp. 147–149. 1990.

[3] Chun-Nan Hsu, Hung-Ju Huang, Tsu-Tsung Wong, "Why Discretization works for Naïve Bayesian Classifiers", *17th ICML*, pp 309-406, 2000.

[4] Domingos, P., Pazzani, M., "Beyond independence: conditions for the optimality of the simple Simple Bayesian Classifier", *Machine Learning Proceedings of the Thirteenth International Conference*, Morgan Kaufman, July 1996.

[5] Domingos, P., Pazzani, M., "On the Optimality of the Simple Bayesian Classifier under Zero-One loss", *Machine Learning*, 29(2/3): 103-130, November/December 1997.

[6] Dougherty J., Kohavi R., Sahami M., "Supervised and unsupervised discretization of continuous features", *In Proceedings of the 12th International Conference on Machine Learning*, pp. 194–202. 1995.

[7] Duda and Hart. "Pattern Classification and Scene Analysis", 1973, *John Wiley & Sons*, NY.

[8] Fayyad U. M., Irani K. B., "Multi-interval discretization of continuous-valued attributes for classification learning", *In Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 1022–1027, 1993.

[9] Harry Zhang, Charles X. Ling, "A Fundamental Issue of Naïve Bayes", *Canadian Conference on AI*: 591-595, 2003.

[10] Hayashi Y., "Neural expert system using fuzzy teaching input and its application to medical diagnosis", *Information Sciences Applications*, Vol. 1, pp. 47-58, 1994.

[11] John G. H., Kohavi R., Pfleger P., "Irrelevant Features and the Subset Selection Problem", *In Machine Learning, Proceedings of the Eleventh International Conference*, pp 121-129, Morgan Kaufmann Publishers, San Francisco, CA.

[12] John G. H., Langley P., "Estimating continuous distributions in Bayesian classifiers", *In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence* , pp. 338–345. 1995.

[13] Kohavi R., "A study of cross-validation and bootstrap for accuracy estimation and model selection". *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995, pp. 1137–1145.

[14] le Cessie S., van Houwelingen J., "Ridge estimators in logistic regression", *Applied Statistics*, Vol 41, no 1, pp. 191-201, 1992.

[15] Liu H., Sentino R, "Some issues on scalable Feature Selection, Expert Systems with Application", vol 15, pp 333-339, 1998.

[16] Ludmila I. Kuncheva, - School of Informatics, University of Wales, Bangor, Dean Street, Bangor Gwynedd LL57 1UT, UK. http://www.informatics.bangor.ac.uk/ ~kuncheva / activities/ real_data_full_set.htm.

[17] Mark A Hall, Lloyd A Smith, "Feature Selection for Machine Learning: Comparing a Correlation based Filter Approach to the Wrapper", *Proc Florida Artificial Intelligence Symposium*, pp-235-239, AAAI Press, 1999.

[18] Mitra S., "Fuzzy MLP based expert system for medical diagnosis", *Fuzzy Sets and Systems*, Vol. 65, pp. 285-296, 1994.

[19] Pearl J, "Probabilistic Reasoning in Intelligent Systems", *Morgan Kaufmann Publishers*, 1988.

[20] Pepe M.S., "The Statistical Evaluation of Medical Tests for Classification and P rediction", http://www.fhcrc.org /science/ labs/pepe/book/, Oxford Statistical Science Series, Oxford University Press. 2003.

[21] Prather J. C., Lobach D. F., Goodwin L. K., Hales J. W., Hage M. L., Edward Hammond W., *"Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse"*, 1997.

[22] Ranjit Abraham, Jay B. Simha, Iyengar S.S., "A comparative study of discretization and feature selection methods for Medical datamining using Naïve Bayesian classifier", *Indian Conference on Computational Intelligence and Information Security (ICCIIS '07)*, pp.196-201, 2007.

[23] Ranjit Abraham, Jay B.Simha, Iyengar S.S., "A comparative analysis of discretization methods for Medical datmining with Naïve Bayesian classifiers", *9th International Conference for Information Technology (ICIT2006)*, pp. 235-236, 2006.

[24] Shalom M. Weiss, (for the Medical dataset on Appendicitis), sholom@us.ibm.com.

[25] Ying Yang, Geoffrey I Webb, "A Comparative Study of Discretization Methods for Naïve Bayes Classifiers", In Proceedings of PKAW, Japan pp 159-173, 2002.