International Journal of Computational Intelligence Research.
ISSN 0974-1259 Vol.4, No.X (2008), pp. XXX–XXX

Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical datamining

Ranjit Abraham [1], Jay B. Simha [2] and S. Sitharama Iyengar [3]

[1] Research Scholar,
Dr. MGR University, Chennai, INDIA.
ranjit.abraham@gmail.com

[2] Abiba Systems, Bangalore, INDIA.
jbsimha@gmail.com

[3] Department of Computer Science,
Louisiana State University, Baton Rouge, USA
iyengar@bit.csc.lsu.edu

Abstract*: 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. Improving the predictive accuracy and achieving dimensionality reduction for statistical classifiers has been an active research area in datamining. Our experimental results suggest that on an average, with Minimum Description Length (MDL) discretization the Naïve Bayes Classifier seems to be the best performer compared to popular variants of Naïve Bayes as well as some popular non-Naïve Bayesian statistical classifiers. We propose a Hybrid feature selection algorithm (CHI-WSS) that helps in achieving dimensionality reduction by removing irrelevant data, increasing learning accuracy and improving result comprehensibility. Experimental results suggest that on an average the Hybrid Feature Selector gave best results compared to individual techniques with popular filter as well as wrapper based feature selection methods. The proposed algorithm which is a multi-step process utilizes discretization, filters out irrelevant and least relevant features and finally uses a greedy algorithm such as best first search or wrapper subset selector. For experimental validation we have utilized two established measures to compare the performance of statistical classifiers namely; classification accuracy (or error rate) and the area under ROC. Our work demonstrates 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. This work based on empirical evaluation on publicly available datasets validates our hypothesis of development of parsimonious models from our generalized approach.

***Keywords*:** Naive Bayesian classifier, discretization, Minimum description length, feature selection, chi-square statistics**.**

I. Introduction

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 [38]. For a Physician who is guided by empirical observation and clinical trials, this data becomes appropriate if it is provided in terms of generalized knowledge such as information pertaining to patient history, diseases, medications, and clinical reports.

Several computer programs have been developed to carry out optimal management of data for extraction of knowledge or patterns contained in the data. These 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. Data classification using Naïve Bayes (NB) has gained much prominence because of its simplicity and comparable accuracy with other classifiers. Research study shows that Naïve Bayesian classification works best for discretized attributes [4], [10].

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. This assumption grossly violates real-world problems and much effort has been focused in the name of Naïve Bayes variants by relaxing the independence assumptions to improve classification accuracy. It has been shown that Naïve Bayesian classifier is extremely effective in practice and difficult to improve upon [9].

Research work show that Naïve Bayes (NB) classification works best for discretized attributes and the application of

Fayyad and Irani's Minimum Discretization Length (MDL) discretization gives on the average best classification accuracy performance [41]. In this paper we compare the accuracy performance of non-discretized NB with MDL discretized NB, popular variants of NB and with state-of-the-art classifiers such as k-Nearest Neighbor, Decision Trees, Logistic Regression, Neural Networks and Support Vector Machines.

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 [29], [34]. 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.

The motivation for this work comes from studies utilizing the combination of machine learning techniques in literature. They include the use of 3-NN for selecting best examples for 1-NN [45], application of decision trees to identify the features for indexing in case based reasoning and selection of the examples [7], and instance based learning using specific instances [6]. The idea behind our general approach is to reduce the space complexity at each phase of the process so that greedy algorithms at the final step of the process have to deal with relatively smaller subset of features than the original.

The approach we have adopted is a three phased framework. First, the continuous variables are discretized to reduce the effect of distribution imbalance. (Naïve Bayes works well with categorical attributes). In the second phase, irrelevant attributes are removed to minimize the feature count for the model. Even though Naïve Bayes gracefully handles irrelevant attributes, we are removing the irrelevant attributes to bring parsimony to the model structure. In the third phase, a greedy search algorithm is applied to search the best feature subset.

Through this paper we propose a Hybrid feature selection algorithm which is a multi-step process. In the first step the data is discretized. During the second step the discretized data is filtered by removing irrelevant and least relevant features using chi-square feature selection. In the third step, a greedy algorithm like Wrapper Subset or Best First search is used to identify the best feature set. We experimentally compare the accuracy performance with individual techniques drawn from popular filter and wrapper based approaches. The experimental results with our proposed Hybrid feature selection algorithm show that it achieves on the average better dimensionality reduction and increased learning accuracy by reducing the space complexity at each phase of the process. Our experimental study shows that it is possible to reliably develop parsimonious models by applying the Hybrid feature selection algorithm that is both simple and effective.

II. 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 [12]. 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 [9].

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 x1 is the value of the attribute X1. 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 [12] as

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

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 Figure 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 [9]. 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 [36].

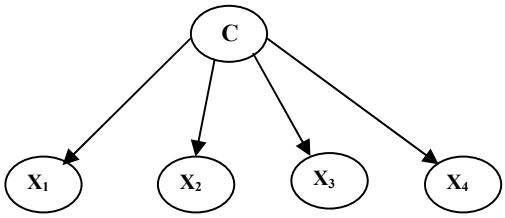

Figure 1. Structure of Naïve Bayes

III. Implementing the NB Classifier

Considering that an attribute X has a large number of values, the probability of the value $P(X=x_i|C=c)$ 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 where σ_c is the standard

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

deviation and μ_c is the mean of the attribute values from the training set [10]. 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 [22]. 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 [3]. Accordingly

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

where nc = 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=xi and C=c, m=2 (a constant) and P(X=xi) 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 [8], we have chosen to ignore them in our experiments.

IV. Discretization for NB Classifier

Data discretization is the process of transforming data containing a quantitative attribute so that the attribute in question is replaced by a qualitative attribute [46]. 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 [4].

Discretization involves dividing an attribute's values into a number of intervals (min_i .. max_i) so that each interval can be treated as one value of a discrete attribute. The choice of the intervals can be determined by a domain expert or with the help of an automatic procedure that makes the task easy. For Naïve Bayes, computational time complexity is only linear with respect to the size of the training data. This is much more efficient than the exponential complexity of Non-Naïve Bayesian approaches [47]. They are also space efficient. With discretization, the learning complexity of the Naïve Bayes classifier should get reduced. Although several discretization methods have been developed for Naïve Bayes classifiers, we have chosen 2 unsupervised (Equal Width and Equal Frequency discretization) as well as the popular supervised Fayyad and Irani's Minimum Description Length (MDL) [14], [42] methods for our experiments.

V. Equal Width & Frequency discretization

Both Equal Width and Equal Frequency discretization are unsupervised direct methods and have been used because of their simplicity and reasonable effectiveness [4]. In Equal Width Discretization (EWD) an attribute's values are divided between x_{min} and x_{max} into k equal intervals such that each cut point is x_{min}+ m x ((x_{max} - x_{min}) / k); where m takes on the value from 0..(k-1). In Equal Frequency Discretization (EFD) each interval in k between x_{min} and x_{max} has approximately the same number of the sorted values of the attribute. Both EWD and EFW suffer from possible attribute loss on account of the pre-determined value of k. For our experiments we have chosen k to be 10.

VI. MDL discretized Naïve Bayes

The Minimum Description Length (MDL) discretization is Entropy based heuristic given by Fayyad and Irani [13]. 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 (MDL).

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

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

and

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

For the given feature the boundary T_{min} 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 T_{min} 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.

VII. Variants of Naïve Bayes Classifier

Real-world problems rarely show the conditional independence assumption used in Naïve Bayes. Extending the structure was adopted as a direct way to possibly overcome the limitation posed by Naïve Bayes (NB) resulting in various NB variants. Briefly described are 4 popular NB variants that were used in our experiments.

The Tree Augmented Naïve Bayes (TAN) is an extended NB [15] where with a less restricted structure in which the class node directly points to all attribute nodes and an attribute node can have only one parent attribute node. TAN is a special case of Augmented Naïve Bayes (ANB), which is equivalent to learning an optimal BN, which is N-P hard. TAN has shown to maintain NB robustness and computational complexity and at the same time displaying better accuracy. The structure of TAN is shown in Figure 2.

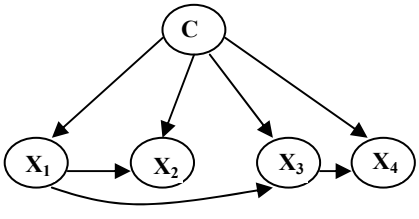

Figure 2. Structural representation of Tree Augmented Naïve Bayes (TAN)

Boosting involves learning a series of classifiers, where each classifier in the series learns more attention to the examples that have been misclassified by its predecessors. Hence each next classifier learns from the reweighed examples. The final boosted classifier outputs a weighted sum of the outputs of each individual classifier series with each weighted according to its accuracy on its training set. Boosting requires only linear time and constant space and hidden nodes are learned incrementally starting with the most important [13]. A graphical representation for Boosted Naïve Bayes (BAN) is shown in Figure 3. The hidden nodes ψ correspond to the outputs of the NB classifier after each iteration of boosting. With sample datasets BAN shows comparable accuracy with TAN.

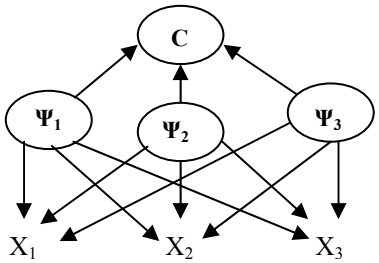

Figure 3. Structural representation for the Boosted Augmented Naïve Bayes (BAN)

The Forest augmented Naïve Bayes (FAN) represents an Augmented Bayes Network defined by a Class variable as parent to every attribute and an attribute can have at most one other attribute as its parent [23],[43]. By applying the algorithm [17] incorporating Kruskal's Maximum Spanning Tree algorithms an optimal Augmented Bayes Network can be found. A graphical structural representation for the Forest augmented NB is shown in Figure 4.

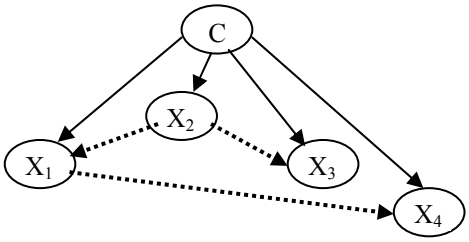

Figure 4. Structural representation for Forest augmented Naïve Bayes (FAN)

The Selective Naïve Bayesian classifier (SNB) uses only a subset of the given attributes in making the prediction [28]. The model enables to exclude redundant, irrelevant variables so that they do not reflect any differences for classification purposes. Experiments with sample datasets reveal that SNB appears to overcome the weakness of NB classifier. An example structural representation for SNB is shown in Figure 5.

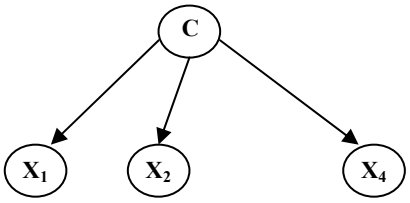

Figure 5. Structural representation for Selective Naïve Bayes (SNB)

For the above given model, and an example given by E=< x1, x2, X3, x4 >, will be assigned to the class

$$c(E) = \arg\max_{c \in C} p(c)\; p(x_1 \mid c)\; P(x_2 \mid c) P(x_4 \mid c)$$

VIII. Popular non-NB statistical classifiers

Here we briefly describe the 5 non-Naïve Bayesian Statistical classifiers we have used in our experiments. The idea of a Decision Tree (DT) [39] is to partition the input space into small segments, and label these small segments with one of the various output categories. A DT is a k-ary tree where each of the internal nodes specifies a test on some attributes from the input feature set used to represent the data. Each branch descending from a node corresponds to one of the possible values of the feature specified at that node. Each test results in branches, which represent different outcomes of the test. The basic algorithm for DT induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The class probability of an example is estimated by the proportion of the examples of that class in the leaf into which the example falls. For our experiments we have used the J48 class of C4.5 decision trees provided in the Weka machine learning environment. The J48 tree classifier forms rules from pruned partial decision trees built using C4.5's heuristics. The J48 classifier parameters in Weka were set as follows: Confidence Factor is 0.25 (sets the threshold for the InformationGainRatio measure used by J48), minimum number of Instances per leaf is 2.0, number of folds is 3 (1 for pruning and 2 for growing the tree) and sub tree pruning enabled.

The k-NN is a supervised learning algorithm where the result of new instance query is classified based on majority of k-Nearest Neighbor category [6]. The purpose of this algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and only based on memory. Given a query point, we find k number of objects or (training points) closest to the query point. The classification is using majority vote among the classification of the k objects. Any ties can be broken at random. k-NN algorithm uses neighborhood classification as the prediction value of the new query instance. k-nearest neighborhood may be influenced by the density of the neighboring data points. We have used Weka's IBk implementation of the k-nearest neighbor classifier [6] and set the classifier parameters as - number of nearest neighbors (k) as 1, the windowSize is 0 (indicating that there was no limit on the number of training instances) and disabled distance weighting method, cross validation for selecting the best k value and attribute normalization.

Logistic regression (LR) is part of a category of statistical models called generalized linear models. LR allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these [28]. LR is often referred to as a discriminative classifier unlike NB which is referred to as a generative classifier. To cater to Logistic Regression for more than 2 discrete outcomes, we have used Weka's Multinominal Logistic Regression algorithm with ridge estimator in our experiments. The implementation of the Multinominal Logistic Regression with ridge estimator is believed to be suitable for small training sets [48]. The classifier parameters were configured for Weka's default values as follows: the ridge is 1.0E-8 (to enable log–likelihood) and the maximum number of iterations to be performed is -1.

Artificial neural networks (NN) are relatively crude electronic networks of "neurons" based on the neural structure of the brain. They process records one at a time, and "learn" by comparing their classification of the record (which, at the outset, is largely arbitrary) with the known actual classification of the record. The errors from the initial classification of the first record is fed back into the network, and used to modify the networks algorithm the second time around, and so on for many iterations [19]. For the Neural Network classifier, we have used Weka's Multilayer Perceptron algorithm. This network uses a sigmoid function as its activation function and back propagation as its learning algorithm. The classifier parameters for the multilayer perceptron in Weka were set to the default values as follows – Hidden Layers is 'a' [the wildcard a = (attributes + classes) / 2], Momentum is 0.2, Learning rate is 0.3, Number of Epochs is 500, Random seed for Weights is 0, validationSetSize is 0 and the validation threshold is 20.

Support Vector Machines (SVMs), are one of the most powerful methods in machine learning for solving binary classification problems, based on the idea of identifying hyperplanes that maximizes the margins between the two classes [2]. The concept of decision planes that define decision boundaries are used in SVM. A decision plane is one that separates between a set of objects having different class memberships. This approach constructs hyper planes in a multidimensional space that separates cases of different class labels. SVM can handle multiple continuous and categorical variables [5]. For our experiments we have used Weka's SMO classifier- which implements John C. Platt's sequential minimal optimization algorithm for training a support vector classifier using polynomial kernels [49][50]. We have chosen Weka's default values as follows: complexity parameter c is 1.0, gamma is 1.0, kernel cache size is 250007, use of the Polynomial kernel with exponent is 1.0 and the values (not to be changed) 1.0E-8 and 0.0010 for Epsilon and toleranceParameter respectively.

IX. 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. 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 [29]. 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 namely filter and wrapper based approaches [21]. For our experimental study we have considered 3 popular filter based approaches namely Chi-squared, Gain Ration and ReliefF and 3 popular wrapper based approaches namely Correlation feature selection (CFS), WrapperSubset feature selection and Consistency-based subset feature selection.

X. Filter based feature selection

The Filter based feature selection methods operate 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. Discussed below are 3 popular filter approaches used for our experiments.

The Chi-squared feature selection algorithm evaluates the worth of a feature by computing the value of the chi-squared statistic with respect to the class. The Chi-Squared (χ^2) method [30] is built on the top of the entropy method. The χ^2 method evaluates features individually by measuring their chi-squared statistic with respect to the classes.

Information Gain is one of the most popular feature selection algorithms. It uses the measure of information entropy of one variable (or feature) before and after observing another variable, the difference of which is the information gain [30]. Gain Ratio is a modified version of the Information Gain measure, and tells us the amount of information gain of the variable relative to the entropy of the class. The Gain Ratio can be termed as a modification of the information gain that reduces its bias towards attributes with more states. However a drawback of Gain Ratio is that it may overcompensate, i.e. choose an attribute just because its intrinsic information is very low.

The Relief algorithm [27] assigns “relevance” weights to each feature which indicates the relevance to the target concept [29]. The method works by randomly sampling an instance from the data and locating its nearest neighbor from the same and opposite class. The values of the attributes of the nearest neighbor are compared to the sampled instance and used to update relevance scores for each attribute. The process is repeated for a user specified number of instances m. Here the useful attribute differentiates between instances from different classes and have the same value from the same class. This method of Relief originally intended for two-class problems has been extended using ReliefF to handle noisy and multi-class datasets. An advantage of ReliefF algorithm is that it can deal with noisy and incomplete datasets [29].

XI. Wrapper based feature selection

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 wrapper approach 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 [40]. Wrapper methods are widely recognized as a superior alternative in supervised learning problems, but on account of the number of executions that the search process requires results in a high computational cost than filters methods. We briefly describe 3 popular wrapper methods that were used in our experiments.

Correlation Feature Selection (CFS) evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred [34]. The method employs a heuristic to select a subset taking into account its usefulness for predicting the class along with the level of inter-correlation among them. While irrelevant features will be poor predictors, redundant attributes will be highly correlated to one or more of the other features. If expanding a subset results in no improvement, the search drops back to the next best unexpanded subset and continues from there. Given enough time a best first search will explore the entire feature subset space, so it is common to limit the number of subsets expanded that result in no improvement. The best subset found is returned when the search terminates. CFS uses a stopping criterion of five consecutive fully expanded non- improving subsets. The greatest limitation of CFS is its failure to select features that have locally predictive values when they are overshadowed by strong, globally predictive features.

The Wrapper attribute selection uses a target learning algorithm to estimate the worth of attribute subsets. In this method, selection is made on a subset of original features of the dataset such that the induction algorithm (Naïve Bayes in our case) that is run on the data containing only those features generates a classifier with the highest possible accuracy [26]. Cross validation (we have used 5 CV for our experiments) is used to provide an estimate of the accuracy of the NB classifier when using only the attributes in a given subset. The forward selection search is used to produce a list of attributes ranked according to their overall contribution to the accuracy of the attribute set with respect to the target learning algorithm.

Consistency-based subset evaluates the worth of a subset of features by the level of consistency in the class values when the training instances are projected onto the subset of features. For this feature selection approach, combinations of attributes whose values divide the data into subsets containing a strong single class majority is looked for. This approach is biased towards small feature subsets with a high-class consistency. Here we use Liu and Sentiono’s consistency metric [31]. The forward selection search is used

to produce a list of attributes ranked according to their overall contribution to the consistency of the attribute set.

XII. CHI-WSS feature selection algorithm

Our proposed feature selection algorithm (CHI-WSS) combines the filter approach with a greedy subset search approach such as wrapper subset selector. The reason for using both filter based and wrapper based approach is to reduce the search space in each phase. Specifically wrapper based approach will not remove irrelevant features and filter algorithms do not greedily search the feature space. The hypothesis of our research is to find the effectiveness of combining these two approaches to reduce the search space and build a parsimonious model. Our approach can be viewed in terms of 3 distinct phases as shown in Figure 6.

In the Discretization phase all non-categorical features of the dataset are discretized. All irrelevant and least relevant features are removed in the Filter phase. During the Subset search phase, feature subsets are identified using a greedy algorithm to find the set of features that maximizes classification accuracy.

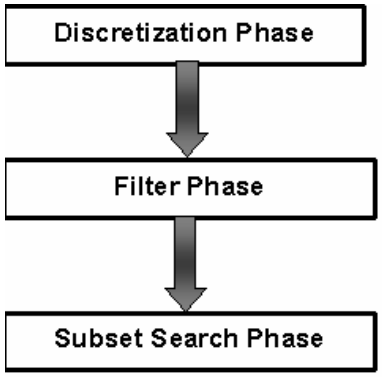

Figure 6. CHI-WSS feature selection Phases

The CHI-WSS algorithm in Steps

Step 1: Given the Dataset, apply MDL (Fayyad and Irani) discretization to all non-categorical features.

Step 2: Compute chi-square feature ranking. Remove all irrelevant features from the dataset if their chi-square average merit equals zero. Next, remove all least relevant features from the dataset that satisfy the condition

$$100 \times \frac{avg\_merit_i \times \log(N^2)}{\Sigma avg\_merit \times N} \langle \delta$$

where we set δ= 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.

Step 3: Identify the feature subsets using a greedy algorithm such as Best First Search or Wrapper Subset selector.

The MDL discretization is carried out in the first step because greedy subset search methods like Wrapper subset do not do data discretization. Through the second step, by removing irrelevant and least relevant features, we reduce the computational overhead of the greedy feature search. Further, our approach is a generalized one as any suitable greedy search method such as Best First search or Wrapper subset search selector may be employed in the final step.

The chi-square feature ranking computes the average of the 10 Cross Validation chi-square statistics with respect to the class – called average merit and given as avg_merit in the above given mathematical formula which is used for identification of least relevant features. The mathematical formula was empirically obtained through experimentation with various publicly available datasets. The wrapper methods are widely recognized as a superior alternative in supervised learning problems, since by employing the inductive algorithm to evaluate alternatives they have into account the particular biases of the algorithm. However, even for algorithms that exhibits a moderate complexity, the number of executions that the search process requires results in a high computational cost [33]. The CHI-WSS algorithm helps to reduce the space complexity at each phase so that greedy algorithms such as the Wrapper subset selector used at the final step have to deal with relatively smaller feature subsets than the original. This in turn validates the hypothesis of the development of parsimonious models from our generalized approach.

XIII. Experimental Evaluation

We have used 17 natural Medical datasets for our experiments whose technical specifications are as shown in Table 1. All the chosen datasets had at least one or more attributes that were continuous. The main software package used in our experiments is Weka version 3.4.8 (Waikato Environment for Knowledge Analysis), developed at the University of Waikato in New Zealand [51]. 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.

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 [20,35]	536	10	4	No	No
10	Appendicitis [44]	106	9	2	Yes	No
11	Leisenring neo audiology [37]	3152	8 (7)	2	No	No
12	Norton neonatal audiology [37]	5058	9 (7)	2	Yes	No
13	Laryngeal 1 [32]	213	17	2	No	No
14	RDS [32]	85	18	2	No	No
15	Voice_3 [32]	238	11	3	No	No
16	Voice_9 [32]	428	11	9 (2)	No	No
17	Weaning [32]	302	18	2	No	No

We have used 10-fold cross validation test method to all the medical datasets [25]. 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 provides the experimental evaluation of discretization techniques employed. The wins for Fayyad and Irani's MDL discretization indicates on the average improved classification accuracy compared to that of Equal Frequency and Equal Width discretization. We argue that MDL discretization does better on account of using the class information entropy after discretization and EWD and EFD discretization levels are not optimized.

Table 2: Naïve Bayes Classification Accuracy with and without Discretization

SL. No.	Medical Dataset	NB without discretization	Naïve Bayes with discretization		
			EWD	EFD	MDL
1	Wisconsin Breast Cancer	95.9943	97.2818	97.2818	96.9957
2	Pima Diabetes	76.3021	75.3906	75	77.8646
3	Bupa Liver Disorders	55.3623	64.9275	62.3188	63.1884
4	Cleveland Heart Disease	83.8284	83.4983	83.4983	83.8284
5	Hepatitis	84.5161	84.5161	83.2258	84.5161
6	Thyroid -new	96.7442	92.093	95.814	96.2791
7	Thyroid (ann-train)	95.5196	93.7169	96.315	98.807
8	Statlog- heart	84.8148	84.0741	82.2222	83.3333
9	Hepatobiliary disorders	47.9478	50.5597	65.6716	68.4701
10	Appendicitis	84.9057	81.1321	84.9057	88.6792
11	Leisenring neo audiology	100	100	100	100
12	Norton neonatal audiology	96.0854	96.5204	96.9751	96.6983
13	Laryngeal 1	75.5869	80.2817	82.6291	86.8545
14	RDS	89.4118	94.1176	94.1176	95.2941
15	Voice_3	69.7479	71.4286	72.6891	76.8908
16	Voice_9	78.7383	82.9439	83.8785	84.5794
17	Wearing	89.7351	88.7417	89.404	91.0596
Wins		5/17	3/17	3/17	11/17

Table 3 shows the accuracy results for non-discretized NB, MDL discretized NB and variants of NB. The 4 variants of Naïve Bayes chosen for our experiments are Selective Naïve Bayes (SNB), Boosted Naïve Bayes (BNB), Tree Augmented Naïve Bayes (TAN) and Forest Augmented Naïve Bayes (FAN).

Table 3: Classification Accuracy with Naïve Bayes (NB), MDL discretized NB and variants of NB.

Sl No.	Medical Dataset	NB	NB (MDL)	Variants of NB			
				SNB	BNB	TAN	FAN
1	Wisconsin Breast Cancer	95.9943	96.9957	96.7096	95.5651	96.7096	95.5651
2	Pima Diabetes	76.3021	77.8646	77.0833	74.349	74.6094	73.9583
3	Bupa Liver Disorders	55.3623	63.1884	61.4493	66.087	56.2319	68.9855
4	Cleveland Heart Disease	83.8284	83.8284	84.4884	83.4983	83.4983	83.4983
5	Hepatitis	84.5161	84.5161	87.0968	82.5806	83.2258	83.2258
6	Thyroid-new	96.7442	96.2791	97.6744	95.814	94.4186	95.3488
7	Thyroid (ann-train)	95.5196	98.807	95.6257	93.0806	99.3107	99.3637
8	Statlog- heart	84.8148	83.3333	84.8148	80.7407	80.7407	80.3704
9	Hepatobiliary disorders	47.9478	68.4701	49.0672	45.1493	65.4851	79.1045
10	Appendicitis	84.9057	88.6792	88.6792	84.9057	87.7358	86.7925
11	Leisenring neo audiology	100	100	100	100	100	100
12	Norton neonatal audiology	96.0854	96.6983	96.0854	97.0542	96.9751	96.9751
13	Laryngeal 1	75.5869	86.8545	85.446	82.1596	84.0376	84.0376
14	RDS	89.4118	95.2941	91.7647	89.4118	90.5882	92.9412
15	Voice_3	69.7479	76.8908	80.2521	71.8487	74.3697	73.1092
16	Voice_9	78.7383	84.5794	89.2523	86.6822	88.0841	94.8598
17	Wearing	89.7351	91.0596	89.0728	87.7483	89.7351	87.4172
Wins		1/17	7/17	7/17	2/17	1/17	5/17

Abbreviations Used: NB- Naïve Bayes, NB (MDL) – Naïve Bayes with MDL discretization, SNB – Selective Naïve Bayes, BNB- Boosted Naïve Bayes, TAN- Tree Augmented Naïve Bayes, FAN – Forest Augmented Naïve Bayes

Table 4 shows the accuracy performance with non-discretized NB, MDL discretized NB and some popular non-NB classifiers. The 5 popular non-NB statistical classifiers are Decision Tree (DT), k –Nearest Neighbor (k- NN), Logistic Regression (LR), Neural Network (NN) and Support Vector Machine (SVM). The wins at the bottom of Table 3 and Table 4 provides the ratio of medical datasets where the accuracy is highest among the considered classifiers to the total number of datasets used for our experiments. In both tables the MDL discretized NB on the average gave best results.

Table 4: Classification Accuracy with Naïve Bayes (NB), MDL discretized NB and non-NB classifiers

Sl No.	Medical Dataset	NB	NB (MDL)	Popular non-NB Classifiers				
				DT	k-NN	LR	NN	SVM
1	Wisconsin Breast Cancer	95.9943	96.9957	94.5637	94.9928	96.5665	95.279	96.9957
2	Pima Diabetes	76.3021	77.8646	73.8281	70.1823	77.2135	75.1302	77.3438
3	Bupa Liver Disorders	55.3623	63.1884	68.6957	62.8986	68.1159	71.5942	58.2609
4	Cleveland Heart Disease	83.8284	83.8284	75.9076	75.9076	84.8185	80.8581	85.1485
5	Hepatitis	84.5161	84.5161	83.871	80.6452	82.5806	81.9355	85.1613
6	Thyroid-new	96.7442	96.2791	92.093	97.2093	96.7442	96.7442	89.7674
7	Thyroid (ann-train)	95.5196	98.807	99.7084	92.1262	96.8717	96.2354	93.7964
8	Statlog- heart	84.8148	83.3333	76.2963	75.5556	83.3333	83.3333	82.963
9	Hepatobiliary disorders	47.9478	68.4701	71.0821	73.3209	59.3284	60.8209	42.3507
10	Appendicitis	84.9057	88.6792	86.7925	83.0189	87.7358	87.7358	86.7925
11	Leisenring neo audiology	100	100	100	100	100	100	100
12	Norton neonatal audiology	96.0854	96.6983	97.0739	94.6619	97.0542	96.9553	97.0542
13	Laryngeal 1	75.5869	86.8545	78.4038	79.8122	84.507	82.1596	84.0376
14	RDS	89.4118	95.2941	84.7095	82.3529	87.0588	87.0588	88.2353
15	Voice_3	69.7479	76.8908	74.7899	71.8487	78.1513	76.4706	78.5714
16	Voice_9	78.7383	84.5794	91.1215	94.8598	87.1495	89.486	85.2804
17	Wearing	89.7351	91.0596	82.4503	78.4768	81.7889	84.4371	82.4503
Wins		2/17	6/17	4/17	3/17	1/17	1/17	5/17

Abbreviations Used: NB- Naïve Bayes, NB (MDL) – Naïve Bayes with MDL discretization, DT – Decision Tree, k-NN - k -Nearest Neighbor, LR- Logistic Regression, NN-Neural Network, SVM – Support Vector Machine

To further substantiate the results obtained in Table 3 and 4, we have tabled the results for the Area under the Receiver Operator Characteristics (AUROC) in Table 5 and 6 for the above mentioned statistical classifiers. Clearly the wins

Table 5: AUROC (in percentage) with Naïve Bayes (NB), MDL discretized NB and variants of NB.

Sl No.	Medical Dataset	NB	NB (MDL)	Variants of NB			
				SNB	BNB	TAN	FAN
1	Wisconsin Breast Cancer	98.75	99.2	99.11	97.62	98.94	98.68
2	Pima Diabetes	81.86	84.64	82.79	80.08	80.75	78.42
3	Bupa Liver Disorders	64.01	55.95	61.78	68.41	51.36	73.67
4	Cleveland Heart Disease	90.71	91.27	88.46	89.44	90.92	90.92
5	Hepatitis	85.95	86.94	85.82	85.15	87.70	86.43
6	Thyroid - new	99.27	99.69	99.45	99.54	98.79	98.46
7	Thyroid (ann-train)	99.73	99.94	99.72	99.88	99.95	99.97
8	Statlog- heart	90.83	91	87.99	87.86	90.20	87.11
9	Hepatobiliary disorders	74.35	87.48	71.34	54.85	85.53	91.14
10	Appendicitis	79.33	79.94	78.38	80.81	78.85	85.60
11	Leisenring neo audiology	100	100	100	100	100	100
12	Norton neonatal audiology	60.91	58.44	60.91	61.04	57.83	56.01
13	Laryngeal 1	90.24	93.62	91.02	89.82	89.97	91.19
14	RDS	95.39	98.78	90.78	96.11	97.33	96.22
15	Voice_3	89.9	95.01	90.67	82.95	92.17	88.87
16	Voice_9	90.91	95.36	91.75	92.28	94.50	98.29
17	Wearing	95.95	96.75	97.24	94.29	97.24	93.16
Wins		1/17	9/17	2/17	2/17	3/17	6/17

Abbreviations Used: NB- Naïve Bayes, NB (MDL) – Naïve Bayes with MDL discretization, SNB – Selective Naïve Bayes, BNB- Boosted Naïve Bayes, TAN- Tree Augmented Naïve Bayes, FAN – Forest Augmented Naïve Bayes

obtained by MDL discretized NB classifier proves that it is the best performer.

Table 7 provides the results of feature selection using the proposed CHI-WSS algorithm. From the wins given at the bottom of the table; applying the proposed hybrid feature selector, all the 17 datasets saw improvement in dimensionality reduction with comparable classification accuracy.

Table 6: AUROC (in percentage) with Naïve Bayes (NB), MDL discretized NB and non-NB classifiers

Sl No.	Medical Dataset	NB	NB (MDL)	Popular non-NB Classifiers				
				DT	k-NN	LR	NN	SVM
1	Wisconsin Breast Cancer	98.75	99.2	95.47	97.31	99.33	98.61	96.82
2	Pima Diabetes	81.86	84.64	75.14	65.01	83.18	79.09	71.95
3	Bupa Liver Disorders	64.01	55.95	66.5	62.96	71.76	74.16	50.34
4	Cleveland Heart Disease	90.71	91.27	78.28	78.28	90.86	88.30	84.75
5	Hepatitis	85.95	86.94	70.82	65.35	80.26	81.63	75.62
6	Thyroid -new	99.27	99.69	89.12	96.44	98.82	99.46	83.08
7	Thyroid (ann-train)	99.73	99.94	99.97	82.50	97.56	99.24	84.99
8	Statlog- heart	90.83	91	76.73	75.42	90.44	88.55	82.33
9	Hepatobiliary disorders	74.35	87.48	78.95	80.07	77.37	80.69	75.90
10	Appendicitis	79.33	79.94	63.81	78.54	79.10	77.20	72.04
11	Leisenring neo audiology	100	100	100	100	100	100	100
12	Norton neonatal audiology	60.91	58.44	51.68	51.41	62.50	59.12	50.00
13	Laryngeal 1	90.24	93.62	75.23	77.84	90.83	87.07	83.54
14	RDS	95.39	98.78	89.28	79.44	92.31	94.56	88.06
15	Voice_3	89.9	95.01	68.46	77.87	90.42	90.78	84.39
16	Voice_9	90.91	95.36	89.75	92.63	91.74	93.11	77.54
17	Weaning	95.95	96.75	91.46	90.98	96.18	93.23	89.07
Wins		1/17	14/17	2/17	1/17	2/17	2/17	1/17

Abbreviations Used: NB- Naïve Bayes, NB (MDL) – Naïve Bayes with MDL discretization, DT – Decision Tree, k-NN- k -Nearest Neighbor, LR- Logistic Regression, NN-Neural Network, SVM – Support Vector Machine

Table 7: 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

Figure 7 shows that feature dimensionality reduction was achieved for all the 17 datasets using the CHI-WSS algorithm. Figure 8 depicts the classification accuracy performance before and after the application of the CHI-WSS feature selection algorithm.

In Table 8a and Table 8b we compare the performance accuracy of the NB classifier using CHI-WSS algorithm with popular filter based algorithms. For our study we have considered 3 popular filter based approaches namely Chi-squared, Gain Ration and ReliefF. While Table 8a shows results for the datasets without any discretization, Table 8b gives the results for MDL discretized datasets carried out at the pre-processing stage. From the wins at the bottom of both tables, naïve Bayesian classification using the CHI-WSS algorithm gives best accuracy results.

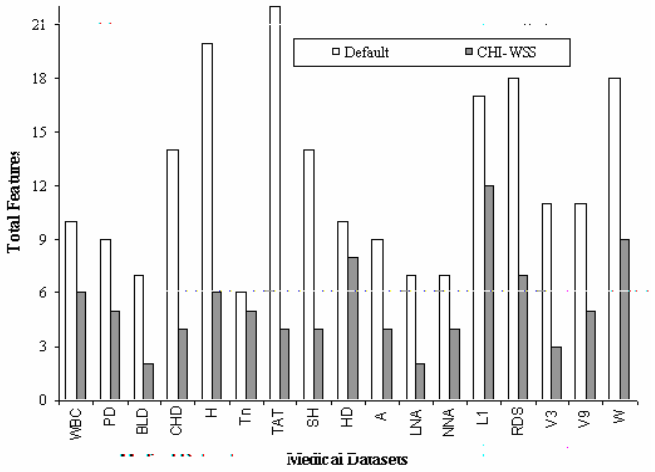

Figure 7. Feature dimensionality reduction before and after using CHI-WSS algorithm

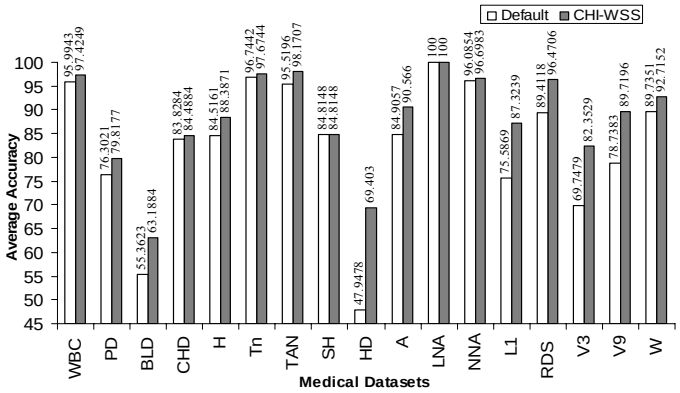

Figure 8. Classification accuracy performance before and after using CHI-WSS algorithm

Table 8a: Classification accuracy of NB with Filter based Feature Selection Algorithms

SL. No.	Medical Dataset	naïve Bayes Classification	Feature Selection Applied (Forward Sequential Search)			
			Filter Approaches			CHI-WSS Algorithm
			ChiSquare	Gain Ratio	ReliefF	
1	Wisconsin Breast Cancer	95.9943 (10)	96.1373 (7)	96.1373 (8)	96.1373 (4)	97.4249 (6)
2	Pima Diabetes	76.3021 (9)	75.0000 (2)	76.4323 (4)	76.4323 (3)	79.8177 (5)
3	Bupa Liver Disorders	55.3623 (7)	60.5797 (4)	60.5797 (4)	56.2319 (2)	63.1884 (2)
4	Cleveland Heart Disease	83.8284 (14)	76.2376 (2)	76.2376 (2)	75.9076 (2)	84.4884 (4)
5	Hepatitis	84.5161 (20)	85.1613 (3)	84.5161 (2)	79.3548 (2)	88.3871 (6)
6	Thyroid -new	96.7442 (6)	97.6744 (4)	96.7442 (6)	83.2258 (3)	97.6744 (5)
7	Thyroid (ann-train)	95.5196 (22)	95.4931 (8)	95.9726 (10)	93.4252 (3)	98.1707 (4)
8	Statlog- heart	84.8148 (14)	74.8148 (2)	84.8148 (7)	74.0741 (2)	84.8148 (4)
9	Hepatobiliary disorders	47.9478 (10)	40.1119 (4)	40.1119 (4)	48.6940 (7)	69.4030 (8)
10	Appendicitis	84.9057 (9)	87.7358 (2)	86.7925 (3)	87.7358 (2)	90.566 (4)
11	Leisenring neo audiology	100.00 (7)	100.00 (2)	100.00 (2)	100.00 (2)	100 (2)
12	Norton neonatal audiology	96.0854 (7)	97.0542 (2)	97.0542 (2)	97.0542 (3)	96.6983 (4)
13	Laryngeal 1	75.5869 (17)	82.6291 (2)	79.8122 (5)	81.2207 (7)	87.3239 (13)
14	RDS	89.4118 (18)	92.9412 (4)	92.9412 (4)	91.7647 (3)	96.4706 (7)
15	Voice_3	69.7479 (11)	77.7311 (2)	77.7311 (2)	70.5882 (2)	82.3529 (3)
16	Voice_9	78.7383 (11)	82.4766 (4)	83.8785 (2)	76.6355 (2)	89.7196 (5)
17	Weaning	89.7351 (18)	86.7550 (5)	89.7351 (7)	90.3974 (8)	92.7152 (7)
Wins		2 / 17	3 / 17	3 / 17	2 / 17	16 / 17

Table 9a and Table 9b compare the performance accuracy

of the NB classifier using CHI-WSS algorithm with popular wrapper based algorithms. In our experimental study we have considered 3 popular wrapper based approaches namely Correlation feature selection (CFS), WrapperSubset feature selection and Consistency-based subset feature selection. From the wins at the bottom of both tables, Naïve Bayesian classification with the CHI-WSS algorithm gave on the average best accuracy results comparable to the computationally intensive Wrapper Subset approach. The results also show that by using the CHI-WSS algorithm we achieve on the average better dimensionality reduction compared to the widely recognized Wrapper Subset feature selection method.

Table 8b: Classification accuracy of NB with (MDL discretized) Filter based Feature Selection Algorithms

SL. No.	Medical Dataset	naïve Bayes Classification (MDL disc)	MDL disc -Feature Selection Applied (FSS)			
			Filter Approaches			CHI-WSS Algorithm
			ChiSquare	Gain Ratio	ReliefF	
1	Wisconsin Breast Cancer	96.9957 (10)	97.4249 (9)	97.4249 (9)	97.4249 (6)	97.4249 (6)
2	Pima Diabetes	77.8646 (9)	78.776 (6)	78.776 (6)	76.0417 (3)	79.8177 (5)
3	Bupa Liver Disorders	63.1884 (7)	63.1884 (7)	63.1884 (7)	63.1884 (7)	63.1884 (2)
4	Cleveland Heart Disease	83.8284 (14)	76.2376 (2)	76.2376 (2)	75.9076 (2)	84.4884 (4)
5	Hepatitis	84.5161 (20)	85.8065 (3)	84.5161 (2)	79.3548 (2)	88.3871 (6)
6	Thyroid -new	96.2791 (6)	96.2791 (6)	97.6744 (5)	96.2791 (6)	97.6744 (5)
7	Thyroid (ann-train)	98.807 (22)	98.0912 (4)	98.0912 (4)	98.0912 (4)	98.1707 (4)
8	Statlog- heart	83.3333 (14)	74.8148 (2)	84.0741 (7)	75.1852 (2)	84.8148 (4)
9	Hepatobiliary disorders	68.4701 (10)	56.903 (3)	56.903 (3)	65.1119 (6)	69.403 (8)
10	Appendicitis	84.9057 (9)	88.6792 (9)	88.6792 (9)	89.6226 (4)	90.566 (4)
11	Leisenring neo audiology	100 (7)	100 (2)	100 (2)	100 (2)	100 (2)
12	Norton neonatal audiology	96.6983 (7)	97.0542 (2)	97.0542 (2)	97.0542 (3)	96.6983 (4)
13	Laryngeal 1	86.8545 (17)	86.8545 (5)	86.385 (3)	86.8545 (4)	87.3239 (13)
14	RDS	95.2941 (18)	95.2951 (4)	95.2951 (4)	95.2951 (4)	96.4706 (7)
15	Voice_3	76.8908 (11)	80.2521 (3)	80.2521 (3)	81.9328 (3)	82.3529 (3)
16	Voice_9	84.5794 (11)	86.215 (3)	87.8505 (3)	86.215 (3)	89.7196 (5)
17	Weaning	91.0596 (18)	82.4503 (3)	82.4503 (4)	82.4503 (3)	92.7152 (7)
Wins		3 / 17	4 / 17	4 / 17	4 / 17	15 / 17

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

Table 9a: Classification accuracy of NB with Wrapper based Feature Selection Algorithms

SL. No.	Medical Dataset	naïve Bayes Classification	Feature Selection Applied (Forward Sequential Search)			
			Wrapper Approaches			CHI-WSS Algorithm
			CFS	Wrapper Subset	Consistency Subset	
1	Wisconsin Breast Cancer	95.9943 (10)	95.9943 (10)	96.7096 (8)	96.2804 (8)	97.4249 (6)
2	Pima Diabetes	76.3021 (9)	77.4740 (5)	77.0833 (6)	76.4323 (8)	79.8177 (5)
3	Bupa Liver Disorders	55.3623 (7)	56.5217 (2)	61.4493 (4)	56.5217 (2)	63.1884 (2)
4	Cleveland Heart Disease	83.8284 (14)	84.4884 (7)	84.4884 (4)	83.8284 (14)	84.4884 (4)
5	Hepatitis	84.5161 (20)	87.7419 (11)	87.0968 (9)	85.8065 (4)	88.3871 (6)
6	Thyroid -new	96.7442 (6)	96.7442 (6)	84.5161 (2)	96.7442 (6)	97.6744 (5)
7	Thyroid (ann-train)	95.5196 (22)	95.1485 (5)	95.6257 (9)	95.5196 (22)	98.1707 (4)
8	Statlog- heart	84.8148 (14)	85.9259 (8)	84.8148 (4)	84.8148 (14)	84.8148 (4)
9	Hepatobiliary disorders	47.9478 (10)	47.9478 (10)	49.0672 (3)	47.9478 (10)	69.4030 (8)
10	Appendicitis	84.9057 (9)	86.7925 (5)	88.6792 (3)	87.7358 (6)	90.566 (4)
11	Leisenring neo audiology	100.00 (7)	100.00 (5)	100.00 (2)	100.00 (2)	100 (2)
12	Norton neonatal audiology	96.0854 (7)	96.6983 (4)	96.0854 (7)	96.7378 (2)	96.6983 (4)
13	Laryngeal 1	75.5869 (17)	78.8732 (10)	85.4460 (5)	78.8732 (5)	87.3239 (13)
14	RDS	89.4118 (18)	92.9412 (4)	91.7647 (14)	92.9412 (10)	96.4706 (7)
15	Voice_3	69.7479 (11)	78.5714 (4)	80.2521 (3)	71.8487 (5)	82.3529 (3)
16	Voice_9	78.7383 (11)	80.8411 (5)	89.2523 (5)	81.5421 (8)	89.7196 (5)
17	Wearing	89.7351 (18)	92.0530 (14)	92.0530 (12)	92.0530 (13)	92.7152 (7)
Wins		1 / 17	3 / 17	2 / 17	2 / 17	15 / 17

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

Figure 9 shows a comparison between the feature dimensionality reduction achieved using the hybrid feature selection algorithm (CHI-WSS) to the widely recognized Wrapper Subset Feature Selection. The WIN-LOSS-TIE for the CHI-WSS feature selection method with respect to the widely recognized Wrapper Subset Feature selection is 8-4-5 clearly demonstrating that on the average the proposed hybrid feature selection (CHI-WSS) method achieves better dimensionality reduction compared to Wrapper Subset feature selector.

Table 9b: Classification accuracy of NB with (MDL discretized) Wrapper based Feature Selection Algorithms

SL. No.	Medical Dataset	naïve Bayes Classification (MDL disc)	Feature Selection Applied (Forward Sequential Search)			
			Wrapper Approaches			CHI-WSS Algorithm
			CFS	Wrapper Subset	Consistency Subset	
1	Wisconsin Breast Cancer	96.9957 (10)	96.9957 (10)	97.4249 (6)	97.2818 (9)	97.4249 (6)
2	Pima Diabetes	77.8646 (9)	77.8646 (9)	79.8177 (5)	77.8646 (9)	79.8177 (5)
3	Bupa Liver Disorders	63.1884 (7)	63.1884 (7)	63.1884 (7)	63.1884 (7)	63.1884 (2)
4	Cleveland Heart Disease	83.8284 (14)	84.1584 (7)	84.4884 (4)	82.8383 (11)	84.4884 (4)
5	Hepatitis	84.5161 (20)	85.8065 (9)	88.3871 (6)	84.5161 (7)	88.3871 (6)
6	Thyroid -new	96.2791 (6)	97.6744 (5)	97.6744 (5)	96.2791 (6)	97.6744 (5)
7	Thyroid (ann-train)	98.807 (22)	98.807 (22)	99.1782 (5)	98.8865 (7)	98.1707 (4)
8	Statlog- heart	83.3333 (14)	84.0741 (7)	84.8148 (5)	83.3333 (14)	84.8148 (4)
9	Hepatobiliary disorders	68.4701 (10)	68.4701 (10)	68.4701 (10)	68.4701 (10)	69.403 (8)
10	Appendicitis	84.9057 (9)	88.6792 (9)	90.566 (4)	91.5094 (4)	90.566 (4)
11	Leisenring neo audiology	100 (7)	100 (2)	100 (2)	100 (2)	100 (2)
12	Norton neonatal audiology	96.6983 (7)	97.0542 (3)	96.6983 (7)	96.6983 (7)	96.6983 (4)
13	Laryngeal 1	86.8545 (17)	87.3239 (6)	85.9155 (3)	87.7934 (13)	87.3239 (13)
14	RDS	95.2941 (18)	95.2951 (4)	91.7647 (2)	96.4706 (9)	96.4706 (7)
15	Voice_3	76.8908 (11)	79.8319 (5)	82.3529 (3)	78.1513 (9)	82.3529 (3)
16	Voice_9	84.5794 (11)	86.9159 (7)	89.7196 (5)	86.4486 (8)	89.7196 (5)
17	Wearing	91.0596 (18)	92.3841 (10)	92.7152 (7)	93.3775 (9)	92.7152 (7)
Wins		2 / 17	3 / 17	12 / 17	5 / 17	13 / 17

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

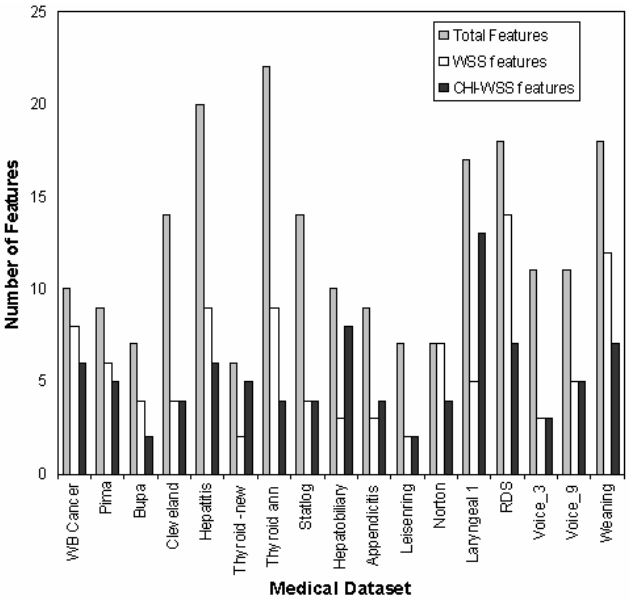

Figure 9. Number of features of Original set, Wrapper using Naïve Bayes and CHI-WSS using Naïve Bayes

Table 10 provides the results of the classifier accuracy using our proposed CHI-WSS algorithm as well as those achieved by WrapperSubset based feature selection using discriminative models- Logistic Regression and Support Vector Machine. 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 other non-generative methods.

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 SVM and Logistic Regression that are used in Medical data mining.

Table 10: 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 11 using true positive rates given in terms of the area under ROC (AUROC), the proposed hybrid feature selector (CHI-WSS) with Naïve Bayes gets more wins than the other methods as shown at the bottom of the table. For the proposed hybrid feature selection algorithm (CHI-WSS), the computational overhead with Naïve Bayes is much lower compared to using discriminative models such as Support Vector Machine (SVM) and Logistic Regression (LR).

Table 11: 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	83.93
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

XIV. Conclusions

In this work an attempt was made to show how the Naïve Bayesian classification accuracy and dimensionality reduction could be achieved with discretization methods and with the proposed hybrid feature selector (CHI-WSS) algorithm for Medical datamining.

Our experimental results indicate that with Medical datasets, on an average, Naïve Bayes with Fayyad and Irani's Minimum Description Length (MDL) discretization seems to be the best performer compared to the 4 popular variants of Naïve Bayes and the 5 popular non-Naïve Bayesian statistical classifiers. Since most of the state of the art classifiers are performing well on these datasets, it is clear that the data transformation is more important than the classifier itself.

The experimental results with the proposed hybrid feature selector (CHI-WSS) indicate that, utilizing Naïve Minimum Description Length (MDL) discretization, filtering out irrelevant and least relevant features using Chi-square feature selection ranking and finally using a greedy algorithm like Wrapper subset selector to identify the best feature set, we could achieve effective feature dimensionality reduction and increased learning accuracy compared to using individual techniques – popular filter as well as wrapper based methods.

Comparing 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 hybrid feature selection algorithm helps in reducing the space complexity through its process steps enabling greedy algorithms in the final step to deal with relatively smaller subset of features than the original. We validate our method for the development of parsimonious models from the generalized approach. We also propose that the Naïve Bayesian classifier with the proposed hybrid feature selector (CHI-WSS) could be set as a benchmark for statistical classifiers.

References

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

[2] B.E. Boser, I.M. Guyon, and V. N. Vapnik. "A training algorithm for optimal margin classifiers", In Fifth Annual Workshop on Computational Learning Theory , ACM., pages 144–152, Pittsburgh, 1992.

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

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

[5] Cortes C., Vapnik V., "Support Vector Networks", Machine Learning, 20(3), pp. 273-297, 1995

[6] David W. Aha, Dennis Kibler, Mark C Albert, “Instance-Based learning algorithms”, Machine Learning, 6, pp. 37-66, 1991.

[7] David B. Skalak, Edwina L. Rissland, “Inductive Learning in a Mixed Paradigm Setting”, AAAI , 840-847, 1990

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

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

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

[11] M. Dash and H. Liu, "Feature Selection for Classification," Intelligent Data Analysis: An Int'l J., vol. 1, no. 3, pp. 131-156, 1997.

[12] Duda and Hart. “Pattern Classification and Scene Analysis”, 1973, John Wiley & Sons, NY.

[13] C. Elkan, “Boosting and Naive Bayesian learning”, (Technical Report) University of California, San Diego, 1997.

[14] U.M. Fayyad, K.B Irani, “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.

[15] N. Friedman, D. Geiger, M. Goldszmidt, “Bayesian network classifiers”, Machine Learning, vol. 29, pages 131-163, 1997.

[16] Hall M., Smith L., “Practical Feature Subset Selection for Machine Learning” Proceedings of the 21st Australian Computer Science Conference. Springer. pp 181-191, 1998.

[17] V. Hamine, P. Helman, “Learning Optimal Augmented Bayes Networks”, Tech Report TR-CS-2004-11, Computer Science Department, University of New Mexico, 2004.

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

[19] Herve Abdi, “A Neural Network Primer”, Journal of Biological Systems, Vol 2(3), pp. 247-283, 1994.

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

[21] G.H. John, R. Kohavi , P. Pfleger , “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.

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

[23] E.J. Keogh, M.J. Pazzani, “Learning Augmented Bayesian Classifiers: A Comparison of Distribution-based and Classification-based Approaches”, Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics: 225-230, 1999.

[24] Kira K., Rendell L., “A practical approach to feature selection”, in Proceedings of the Ninth International Conference on Machine Learning, pp. 249-256, Morgan Kauffmann, 1992.

[25] R. Kohavi, “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.

[26] Kohavi R., John G.H., “Wrappers for feature subset selection”, Artificial Intelligence, vol 97, pp. 273-324, 1997

[27] Kononenko I., “Estimating attributes: Analysis and extensions of Relief”, in Proceedings of the Seventh European Conference on Machine Learning, pp. 171-182, Springer-Verlag, 1994.

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

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

[30] Liu, H. and Setiono, R., “Chi2: Feature selection and discretization of numeric attributes”, Proc. IEEE 7th International Conference on Tools with Artificial Intelligence, pp. 338-391, 1995.

[31] Liu H., Sentino R, “Feature Selection via discretization of Numeric attributes”, IEEE Trans. Knowledge and Data Engineering, vol 9, No 4, pp 642-645, 1997.

[32] I. Ludmila 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.

[33] Luis Talavera, “An evaluation of filter and wrapper methods for feature selection in categorical clustering”, IDA, pp 440-451, 2005.

[34] A. Mark 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.

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

[36] J. Pearl, “Probabilistic Reasoning in Intelligent Systems”, Morgan Kaufmann Publishers, 1988.

[37] M.S. Pepe, “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.

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

[39] J.R. Quinlan, “C4.5, Programs for MachineLearning”, Morgan Kaufmann, San Mateo, Ca, 1993.

[40] 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.

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

[42] I. Rissanen, "Minimum Description Length principle", Encyclopedia of Statistical Sciences, 5:523-527, 1987.

[43] J.P. Saha, http://jbnc.sourceforge.net/JP_Sacha_PhD _ dissertation. pdf".

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

[45] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data", IEEE Transactions on Systems, Man and Cybernetics, 2:408--420, 1972.

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

[47] H. Zhang, C.X Ling and Z Zhao, "The learnability of Naïve Bayes", In Proceedings of Canadian Artificial Intelligence Conference, pp. 432–441, 2000.

[48] George Forman and Ira Cohen, "Learning from Little: Comparison of Classifiers Given Little Training", 15th European Conference on Machine Learning, 2004.

[49] J. Platt , "Fast Training of Support Vector Machines using Sequential Minimal Optimization.", Advances in Kernel Methods - Support Vector Learning, B. Schlkopf, C. Burges, and A. Smola, eds., MIT Press, 1998.

[50] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya and K.R.K. Murthy, "Improvements to Platt's SMO Algorithm for SVM Classifier Design", Technical Report CD-99-14. Control Division, Dept of Mechanical and Production Engineering, National University of Singapore, 1999.

[51] I.H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques," 2nd Edition, Morgan Kaufmann, San Francisco, 2005

Author Biographies

Ranjit Abraham is pursuing his PhD programme with Dr. MGR University, Chennai, India. He received his BSc and MSc in Physics and then MTech in Computer & Information Sciences from Kerala, Kurukshetra and Cochin University of Science & Technology in 1980, 1982 and 1995 respectively. He was a Senior Scientist in Defence Research & Development Organization till 1997 and worked as Senior Computer Consultant to Citibank, New York and is presently COO of Armia Systems Ltd, Kochi.

Jay B.Simha currently heads the analytics group at Abiba systems. He received M.Tech degree from Mysore univserity in Maintenance Engineering (1994), M.Phil. degree in Computer science from MS university (2003) and Ph.D degree from Bangalore university in data mining and decision support (2004). He worked with Alpha systems, Siemens and other companies in software development and analysis. His areas of interest are decision support, predictive analytics and visualization. He has published over 25 papers in refereed journals and conferences.

S. Sitharama Iyengar is the chairman and Roy Paul Daniels Chaired Professor of Computer Science at Louisiana State University, Baton Rouge, and is also the Satish Dhawan Chaired Professor at the Indian Institute of Science, Bangalore. His research interests include high-performance algorithms, data structures, sensor fusion, data mining, and intelligent systems. He is a world class expert in computational aspects of sensor networks. His publications include six textbooks, five edited books, and more than 380 research papers. He is a recipient of IEEE awards, best paper awards, the Distinguished Alumnus Award from the Indian Institute of Science, Bangalore, and other awards. He has served as the editor of several IEEE journals and is the founding editor-in-chief of the International Journal of Distributed Sensor Networks. He has served on the US National Science Foundation and other agencies.