International Journal of Automation and Computing 06(1), February 2009, 62-71
DOI: 10.1007/s11633-009-0062-2

Effective and Efficient Feature Selection for Large-scale Data Using Bayes′ Theorem

Subramanian Appavu Alias Balamurugan[1,*] Ramasamy Rajaram[2]

[1]Department of Information Technology, Thiagarajar College of Engineering, Madurai, India

[2]Department of Computer Science and Information Technology, Thiagarajar College of Engineering, Madurai, India

Abstract: This paper proposes one method of feature selection by using Bayes′ theorem. The purpose of the proposed method is to reduce the computational complexity and increase the classification accuracy of the selected feature subsets. The dependence between two attributes (binary) is determined based on the probabilities of their joint values that contribute to positive and negative classification decisions. If opposing sets of attribute values do not lead to opposing classification decisions (zero probability), then the two attributes are considered independent of each other, otherwise dependent, and one of them can be removed and thus the number of attributes is reduced. The process must be repeated on all combinations of attributes. The paper also evaluates the approach by comparing it with existing feature selection algorithms over 8 datasets from University of California, Irvine (UCI) machine learning databases. The proposed method shows better results in terms of number of selected features, classification accuracy, and running time than most existing algorithms.

Keywords: Data mining, classification, feature selection, dimensionality reduction, Bayes′ theorem.

1 Introduction

As computer and database technologies develop rapidly, data accumulates in a speed unmatchable by human capacity of data processing. Data mining[1−4] as a multidisciplinary joint effort from databases, machine learning, and statistics, is championing in turning mountains of data into nuggets. Researchers and practitioners realize that in order to use data mining tools effectively, data preprocessing is essential to successful data mining[5, 6]. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining[7, 8]. It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications, speeding up a data mining algorithm, improving mining performance such as predictive accuracy and result comprehensibility. Feature selection is a fertile field of research and development in statistical pattern recognition[9−13], machine learning[7, 14−16], and data mining[17−19] since the 1970s, and widely applied to many fields such as text categorization[20−22], image retrieval[23, 24], customer relationship management[25], intrusion detection[26], and genomic analysis[27]. Feature selection is a process that selects a subset of original features. The optimality of a feature subset is measured by an evaluation criterion.

As the dimensionality of a domain expands, the features N increases in number. Finding an optimal feature subset is usually intractable[16] and many problems related to feature selection have been shown to be NP-hard[28]. A typical feature selection process consists of four basic steps namely, subset generation, subset evaluation, stopping criterion, and result validation[18]. Subset generation is a search procedure[5, 29] that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation criterion. If the new subset turns out to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Then, the selected best subset usually needs to be validated by prior knowledge or different tests via synthetic and/or real world datasets. Feature selection can be found in many areas of data mining such as classification, clustering, association rules, and regression. For example, feature selection is called subset or variable selection in statistics[30]. A number of approaches to variable selection and coefficient shrinkage for regression are summarized in [31]. In this survey, we focus on feature selection algorithms for classification and clustering. Early research efforts mainly focus on feature selection for classification with labeled data[12, 18, 32] (supervised feature selection) where class information is available. The latest developments, however, show that the above general procedure can be well adopted to feature selection for clustering with unlabeled data[33−36], (or unsupervised feature selection) where data is unlabeled.

Feature selection algorithms designed with different evaluation criteria broadly fall into three categories: the filter model[17, 37−39], the wrapper model[16, 19, 35, 40], and the hybrid model[27, 41, 42]. The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance, but it also tends to be more computationally expensive than the filter model[16, 29]. The hybrid model attempts to take advantage of the two models by exploiting their different evaluation criteria in different search stages.

In this paper, we give an overview of the popularly used feature selection algorithms under a unified framework.

Manuscript received April 21, 2008; revised October 6, 2008
*Corresponding author. E-mail address: sbit@tce.edu

Moreover, we propose a novel feature selection algorithm based on the Bayes′ theorem for determining the dependent attributes in a dataset and removing those dependent attributes, thereby reducing the attribute set to increase the classification accuracy and reduce the computational time. Experiments on real world datasets show that the proposed method is favorable in terms of its effectiveness and efficiency when compared with other state-of-art algorithms.

2 Related work

Feature selection is a mature area of research. We will present a brief overview of the different feature selection methods.

Blum and Langley[7] classified the feature selection techniques into three basic approaches. In the first approach, known as the embedded approach, a basic induction method is used to add or remove features from the concept description in response to prediction errors on new instances. The second approach is known as the filtering approach, in which, various subsets of features are explored to find an optimal subset, which preserves the classification. The third approach is known as wrapper methods which evaluate alternative feature sets by running some induction algorithm on the training data and using the estimated accuracy of the resulting classifier as its metric.

Quinlans iterative dichotomiser 3 (ID3)[43] and classifier 4.5 (C4.5)[44], classification and regression trees (CART) proposed by Breiman et al.[45] are some of the most successful supervised learning algorithms. These algorithms use a greedy search through the space of decision trees, at each stage using an evaluation function to select the attribute that has the best ability to discriminate among the classes. Michalski[46] proposed the algorithm quasi-optimal (AQ) learning algorithm, which uses positive and negative examples of a class along with a user defined criterion function, to identify a disjunctive feature set that can maximize the positive events and minimize the negative events. Narendra and Fukunaga[47] presented a branch-and-bound algorithm for finding the optimal feature set that uses a top-down approach with back-tracking. Pudil et al.[48] proposed a set of suboptimal algorithms called the floating search methods that do not require the fulfillment of monotonicity condition for feature selection criterion function. Somol et al.[49] provided a modified and efficient branch-and-bound algorithm for feature selection. Though computationally less expensive than the branch-and-bound algorithms, there exists no theoretical upper bound on the computational costs of the algorithms because of their heuristic nature.

Kohari and John et al.[16] proposed another feature selection framework known as the wrapper technique. The wrapper methods evaluate alternative feature sets by running some induction algorithm on the training data and using the estimated accuracy of the resulting classifier as its metric. The major disadvantage of the wrapper approach is that it requires much computation time.

A number of feature selection techniques based on the evolutionary approaches have also been proposed. Casillas et al.[50] presented a genetic feature selection technique which is integrated into a multi-stage genetic learning process to obtain a fuzzy rule based classification system (FRBCS). In the first phase of this method, a filtering approach is used to determine an optimal feature subset for a specific classification problem using class-separability measures. This feature subset along with expert opinion is used to obtain the adequate feature subset cardinality in the second phase, which is used as the chromosome length. Xiong[51] proposed a hybrid approach to input selection, which distinguished itself from existing filter and wrapper-based techniques, but utilized the advantages of both. This process uses case based reasoning to select candidate subsets of features which are termed as "hypothesis". The performance of case-based reasoning under a hypothesis is estimated using training data on a "leave-one-out" procedure. The error estimate is then combined with the subset of selected attributes to provide an evaluation function for the genetic algorithm to find the optimal hypothesis. Kuncheva and Bezdek[52] proposed a genetic algorithm for simultaneous editing and feature selection to design 1-nn classifiers. They had posed the problem as bi-criteria combinatorial optimization problem having an NP-hard search space. Ho et al.[53] proposed the design of an optimal nearest neighbor classifier using intelligent genetic algorithm. Thawonmas and Abe[54] suggested a feature selection technique to eliminate irrelevant features, based on analysis of class regions generated by a fuzzy classifier. The degree of overlaps in a class region is used to define exception ratio, and the features that have the lowest sum of exception ratios are the relevant ones. Irrelevant features are eliminated using a backward selection search technique.

Kira and Rendell[55] proposed a different approach to feature selection and the filter based feature ranking algorithm (RELIEF) also proposed by them assigns a weight to each feature based on the ability of the feature to distinguish among the classes, and then selects those features whose weights exceed a user defined threshold as relevant features. The weight computation is based on the probability of the nearest neighbors from two different classes having different values for an attribute and the probability of two nearest neighbors of the same class having the same value of the attribute. The higher the difference between these two probabilities, the more significant is the attribute. Inherently, the measure is defined for a two-class problem which can be extended to handle multiple classes, by splitting the problem into a series of two-class problems. Kononenko[56] suggested to use k-nearest neighbours to increase the reliability of the probability approximation. It also suggested how RELIEF can be extended to work with multiple sets more efficiently. Weighting schemes are easier to implement and are preferred for their efficiency.

Learning to classify objects is an inherently difficult problem for which several approaches like instance-based learning or nearest neighbor-based algorithms are used. However, the nearest neighbor algorithms need some kind of distance measure. Cost and Salzberg[57] emphasized the need to select appropriate metrics for symbolic values. Stanfill and Waltz[58] proposed the value difference metric (VDM) which measures the distance between values of symbolic features. It takes into account the overall similarity of classification of all instances for each possible value of each feature. Based on this, Cost and Salzberg[57] proposed the modified value distance metric (MVDM) which

is symmetric, and satisfies all the metric properties. They showed that nearest neighbour algorithms perform well even for symbolic data using this metric. It is observed that distance-values are similar if the pairs occur with the same relative frequency for all classes. Zhao and Tsang[59] proposed an attribute reduction with fuzzy approximation operators. Sharma and Paliwal[60] proposed a rotational linear discrimination analysis technique for dimensionality reduction which is a supervised learning technique that finds a linear transformation such that the overlap between the classes is minimum for the projected feature vectors in the reduced feature space.

3 Proposed work

In this paper, we introduce a novel approach for feature selection in high dimensional data using Bayes′ theorem. The dependent attributes are identified and are removed from the dataset. The dependent attributes are the attributes, in which an attribute depends on the other attribute in deciding the value of the class attribute. Dependency between attributes are calculated by first grouping them and then by calculating the probabilities of their values in deciding the value of class attribute using Bayes′ theorem. The difference between the probabilities is then calculated. The procedure is repeated for all possible combinations of attributes, and the dependencies between the whole attribute set in a dataset are found. Find the predictive accuracy with classifiers and place the most accurate attribute in the reduced attribute set. The paper also evaluates the approach by comparing it with existing feature selection algorithms over 8 datasets from University of California, Irvine (UCI) machine learning databases[61]. The proposed method shows better results in terms of number of selected features, classification accuracy, and running time than most existing algorithms.

3.1 Bayes′ theorem and the proposed method for feature selection

Bayes′ theorem[62] describes how the conditional probability of a set of possible causes for a given observed event can be computed from knowledge of the probability of each cause and the conditional probability of the outcome of each cause. It relates the conditional and marginal probabilities of stochastic events A and B Bayes′ theorem is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1}$$

Each term in Bayes′ theorem has a conventional name $P(A)$ is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B. $P(A|B)$ is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B. $P(B|A)$ is the conditional probability of B, given A. $P(B)$ is the prior or marginal probability of B, and acts as a normalizing constant.

In its most general form, Bayes′ theorem states that

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots + P(B|A_n)P(A_n)} \tag{2}$$

where i is any number between 1 and n.

The general structure of a training set is shown in Table 1. The predictor attribute a_1 can take values $\{a_{11}, a_{12}, \cdots, a_{1n}\}$, a_2 can take values $\{a_{21}, a_{22}, \cdots, a_{2n}\}$, $\cdots$, a_n can take values $\{a_{n1}, a_{n2}, \cdots, a_{nn}\}$, and the class attribute c can take the values $\{c_1, c_2, \cdots, c_n\}$.

Table 1 Structure of training dataset

a_1	a_2	$\cdots$	a_n	c
a_{11}	a_{21}		a_{n1}	c_1
a_{12}	a_{22}		a_{n2}	c_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
a_{1n}	a_{2n}		a_{nn}	c_n

This paper proposes an algorithm for feature selection. The basic idea of the algorithm is to test the dependency of all pairs of attributes in deciding the value of the class attribute. The dependency of two attributes is measured by the conditional probabilities of the class attribute given the values of the attributes, which can be easily computed by Bayes′ theorem. Two attributes are defined to be dependent if the difference between their corresponding probabilities satisfies a predefined threshold. If two attributes are identified to be dependent, either of them can be removed to achieve the attribute reduction.

3.2 Proposed algorithm for feature selection

The main steps of the proposed algorithm are given below.

1) Let $A = \{a_1, a_2, a_3, \cdots, a_n\}$ be the initial set of attributes and $a_1 = \{a_{11}, a_{12}, \cdots, a_{1n}\}, \cdots, a_n = \{a_{n1}, a_{n2}, \cdots, a_{nn}\}$.

2) Group attributes in set A into an attribute set of pairs.

3) For each attribute value a_{ij} in an attribute a_i, find the dependency between attributes a_i with respect to attribute values in set A using Bayes theorem.

if dependency exists **then**
  store the attribute a_i into another set B;
else increase i.

4) Use set B to find more dependent attributes and remove those attributes from the set A.

5) Find the predictive accuracy with classifiers and place the most accurate attribute in set A.

The proposed algorithm is enumerated as follows:

Step 1. Let the initial set of predictor attributes be $A = \{a_1, a_2, a_3, \cdots, a_n\}$, where $a_1 = \{a_{11}, a_{12}, \cdots, a_{1n}\}$, $a_2 = \{a_{21}, a_{22}, \cdots, a_{2n}\}$, $\cdots$, $a_n = \{a_{n1}, a_{n2}, \cdots, a_{nn}\}$ and class attribute $c = \{c_1, c_2, \cdots, c_n\}$

Step 2. Use Bayes′ theorem to calculate the probabilities between attribute values.

When $i = 0$,

$$A_0 = \{a_1, a_2\}$$
$$P(a_{11}, a_{21} \,|\, c_1) = 0$$
$$P(a_{12}, a_{21} \,|\, c_2) = 0$$
$$\cdots$$
$$P(a_{11}, a_{22} \,|\, c_1) = 0$$

$$\cdots$$

$$P(a_{1n}, a_{2n} \mid c_n) = 0.$$

No dependency exists. Increase i.
When $i = 1$,

$$A_1 = \{a_1, a_3\}$$
$$P(a_{11}, a_{31} \mid c_1) = 0$$
$$P(a_{12}, a_{31} \mid c_2) = 0$$
$$\cdots$$
$$P(a_{11}, a_{32} \mid c_1) = 0$$
$$\cdots$$
$$P(a_{1n}, a_{3n} \mid c_n) = 0.$$

No dependency exists. Increase i.

$$\cdots$$

When $i = n - 2$,

$$A_i = \{a_1, a_n\}$$
$$P(a_{11}, a_{n1} \mid c_1) = 0$$
$$P(a_{12}, a_{n1} \mid c_2) = 0$$
$$\cdots$$
$$P(a_{11}, a_{n2} \mid c_1) = 0$$
$$\cdots$$
$$P(a_{1n}, a_{nn} \mid c_n) = 0.$$

No dependency exists. Increase i.

$$\cdots$$

$$A_i = \{a_{n-1}, a_n\}$$
$$P((a_{n-1})_1, a_{n1} \mid c_1) = 0$$
$$P((a_{n-1})_2, a_{n1} \mid c_2) = 0$$
$$\cdots$$
$$P((a_{n-1})_1, a_{n2} \mid c_1) = 0$$
$$\cdots$$
$$P((a_{n-1})_n, a_{nn} \mid c_n) = 0.$$

No dependency exists.

If there is dependency between any of the predictor attributes a_i and a_j. Then, store them in set B.

Now,

$$B = \{a_i, a_j\}.$$

Loop terminates.

Step 3. $B = \{a_i, a_j\}$. From Set B, we found that there is a relationship between the attributes a_i and a_j. Find the predictive accuracy with classifiers and place the most accurate attribute in set A.

4 System implementation

The proposed algorithm is implemented using Java. The stepwise approach is as follows.

The input to the system is given as an attribute-relation file format (ARFF) file. A table is created in Oracle using the name specified in "@relation". The attributes specified under "@attribute" and instances specified under "@data" are retrieved from the ARFF file and then they are added to the created table. This procedure is followed for providing the training set as well as test set. The created table acts as the dataset and is given as the input to the proposed algorithm.

The number of predictor attributes and its distinct values and number of distinct values in class attribute are calculated, and these values are used for the calculation of probabilities. The probabilities of attribute are calculated using Bayes′ theorem and the values are listed. The attributes in the dataset are numbered from $1, 2, \cdots, n$. The relationship between two attributes is denoted by X-> Y, where X attribute 1 related to Y attribute 2 and their probabilities with respect to class attribute are calculated using Bayes′ theorem and listed under it.

The combination of attribute value should occur at least once in the dataset, because while finding the dependency between attribute values if a combination of attribute value did not occur once, then it will lead to alternate zeros resulting in zero probability and dependency that cannot be found. Thus, the above condition is checked before a combination of attribute value is given to the proposed method. The probabilities are calculated for the given input. Based on the probabilities, the dependent attributes are identified.

5 Experimental results and discussion

The feature selection using Bayes′ theorem is applied to many datasets, and the performance evaluation is done. We presented the performance evaluation on 8 datasets.

All together 8 datasets are selected from the UCI machine learning repository and the UCI knowledge discovery in databases (KDD) archive[61]. A summary of datasets is presented in Table 2. For each dataset, we run all nine feature-selection algorithms, Bayes′ theorem, wrapper subset eval, consistency subset eval, InfoGain attribute eval, GainRatio attribute eval, OneR attribute eval, ChiSquared attribute eval, principal components, classifier subset eval, respectively, and record the running time and the number of selected features for each algorithm. We then apply naive Bayes, decision tree (ID 3 & J 48), neural-network and support vector machine on the original dataset as well as each newly obtained dataset containing only the selected features from each algorithm and recorded the overall accuracy by 10 fold cross validation.

From Table 3, it is found that for all the datasets, some feature selection algorithms (wrapper subset eval, consistency subset eval, principal components, and classifier subset eval) will select the attributes that are fewer than the number of attributes selected by the proposed method. However, when the selected attributes by the above specified existing feature selection methods are used for the classification, the classification accuracy decreases gradu-

ally. Fig. 1 shows the performance results of the proposed method.

Table 2 Details description of dataset used in the experiment

Data sets	Features	Instances	Classes
Molecular biology (Promoter gene sequences)	57	106	2
Connect 4	42	67557	3
Soybean(Small)	35	47	4
Zoo	17	101	7
Balloon	4	16	2
Mushroom	22	8124	2
Lenses	4	24	3
Fictional	4	14	2

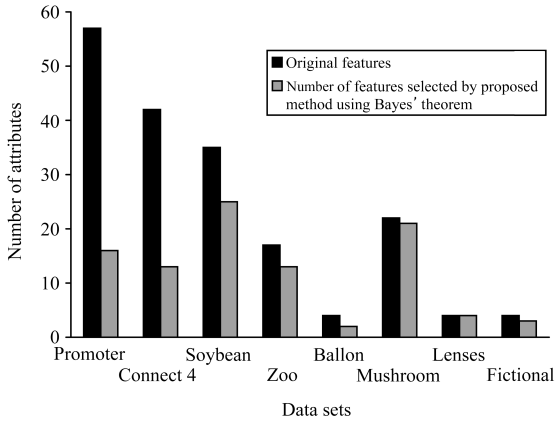


Fig. 1 Number of features selected by the proposed method

The proposed method is fully based on the probabilities (Bayes′ theorem). The main idea in the proposed method is finding the dependency between the attributes in deciding the class attribute value, and also the probabilities will decide the dependency between a set of attributes. Therefore, the proposed method removes the dependent attributes and identifies the perfect attributes which are sufficient for the classification of the datasets and also improve the classification accuracy.

From Table 3, the other existing feature selection algorithms (InfoGain attribute eval, GainRatio attribute eval, OneR attribute eval, ChiSquared attribute eval) suggest that all the attributes are important for the classification of datasets. Because of that, the classification accuracy remains constant for naive Bayes (NB) classifier as shown in Table 4. We conclude that the attributes selected by the proposed method are perfect in the classification of various datasets.

We infer from Table 4 that for the promoter, connect 4, zoo, mushroom, and fictional datasets, the attribute reduction by Bayes′ theorem shows superior results over the original datasets with the initial number of attributes. However, for other datasets small soybean, balloon, and lenses, the classification accuracy is maintained and it shows that only the specified attributes by our novel method is sufficient for the classification.

From Table 3, it is also clear that feature selection by Bayes′ theorem achieves the highest level of dimensionality reduction by selecting a fewer number of features.

Table 3 Number of selected features for each feature selection algorithm

Data sets	Bayes′ theorem	Wrapper subset	Consistency subset	Info-Gain	Gain-Ratio	OneR	Chi-Squared	Principal components	Classifier subset
Molecular biology (promoter)	16	1	4	57	57	57	57	57	1
Connect 4	13	1	19	42	42	42	42	42	1
Soybean	25	1	6	35	35	35	35	20	1
Zoo	13	1	5	17	17	17	17	15	1
Balloon	2	3	3	4	4	4	4	4	1
Mushroom	21	1	9	22	22	22	22	22	1
Lenses	4	1	3	4	4	4	4	4	1
Fictional	3	3	3	4	4	4	4	4	3

Table 4 Accuracy of naive Bayes on selected features for each feature selection algorithm

Data sets	Classification accuracy (%)									
	Full set	Bayes′ theorem	Wrapper subset	Consistency subset	Info-Gain	Gain-Ratio	OneR	Chi-Squared	Principal components	Classifier subset
Molecular biology (Promoter)	90.56	93.33	75.23	92.38	90.56	90.56	90.56	90.56	90.56	75.23
Connect 4	55.40	59.43	44.33	55.09	55.40	55.40	55.40	55.40	55.40	44.33
Soybean	100.00	100.00	57.44	97.87	100.00	100.00	100.00	100.00	97.87	57.44
Zoo	93.06	99.00	54.45	94.05	93.06	93.06	93.06	93.06	92.07	54.45
Balloon	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	78.54
Mushroom	93.83	94.25	50.35	93.56	93.83	93.83	93.83	93.83	93.83	50.35
Lenses	95.83	95.83	66.66	83.33	95.83	95.83	95.83	95.83	95.83	66.66
Fictional	50.00	70.00	40.00	50.00	50.00	50.00	50.00	50.00	50.00	40.00
Average	**84.835**	**88.99**	**61.08**	**83.285**	**84.835**	**84.835**	**84.835**	**84.835**	**84.445**	**58.375**

In Table 5, the running time over different datasets with the proposed feature selection using Bayes′ theorem is given. We found that the average time taken for attribute reduction with Bayes′ theorem is much lower than the other traditional algorithms average. We can also observe that for the proposed method, the running times over different datasets are consistent, which verifies the proposed method′s superior computational efficiency.

Table 5 Running time of the proposed method

Data sets	Bayes′ theorem
Molecular biology (promoter)	20 ms
Connect 4	980 ms
Soybean	13 ms
Zoo	17 ms
Balloon	5 ms
Mushroom	170 ms
Lenses	8 ms
Fictional	5 ms
Average	**152.25** ms

Table 4 and Fig. 2 show the learning accuracy of NB classifier on different feature sets. From the averaged accuracy over all datasets, we observe that attribute reduction by Bayes′ theorem improves the accuracy of the NB classifier. For the promoter, connect 4, zoo, mushroom, and fictional datasets, the novel approach "attribute reduction by Bayes′ theorem" shows superior results over the feature selection algorithms (wrapper subset eval, consistency subset eval, InfoGain attribute eval, GainRatio attribute eval, OneR attribute eval, ChiSquared attributeb eval, principal components, classifier subset eval) (see Fig. 2). However, for the balloon dataset, the classification accuracy remains the same as that of the existing feature selection methods. For the small soybean dataset, attribute reduction by Bayes′ theorem shows superior results over the feature selection algorithms (wrapper subset eval, consistency subset eval, principal components, and classifier subset eval). However, the classification accuracy remains constant in comparison with the other feature selection methods (InfoGain attribute eval, GainRatio attribute eval, OneR attribute eval, ChiSquared attribute eval). For the lenses data set, attribute reduction by Bayes′ theorem shows superior results over the feature selection algorithms wrapper subset eval, consistency subset eval, and classifier subset eval. However, the classification accuracy remains constant in comparison with the other feature selection methods (InfoGain attribute eval, GainRatio attribute eval, OneR attribute eval, principal components, and ChiSquared attribute eval).

It is important that when we calculate the average value of accuracy for each reduction algorithm, our proposed algorithm gives a higher value than others. We conclude that after "attribute reduction by Bayes′ theorem" is applied to the original data set, the classification accuracy either increases or remains the same, showing that only the specified attributes by the new method are sufficient for classification. From Table 6 and Fig. 3, we can observe that running support vector machine (SVM) with selected features is significantly faster than SVM with original features. Thus, we conclude that both the factors "classification accuracy" and "running time" are taken, and the "attribute reduction by Bayes′ theorem" shows superior results over the existing feature selection methods.

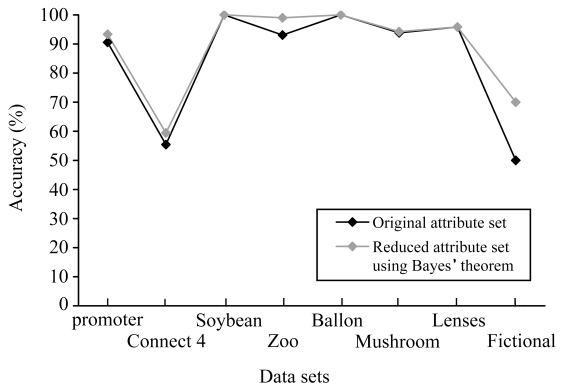


Fig. 2 Effect of the proposed feature selection method on naive Bayes′ classifier performance

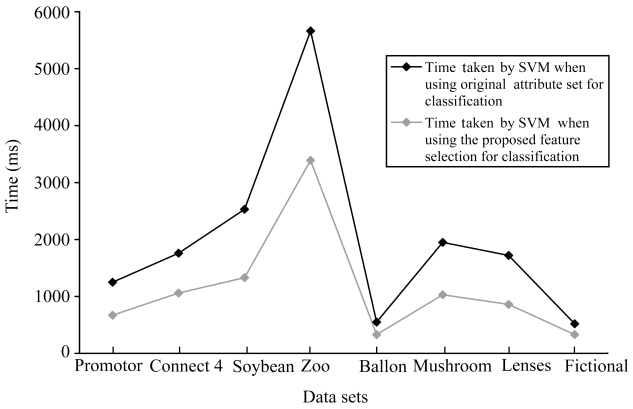


Fig. 3 Time comparison of SVM classifiers

Table 6 Running time of SVM on selected features for each feature selection algorithm

Data sets	Running time (ms)									
	Full set	Bayes′ theorem	Wrapper subset	Consistency subset	Info-Gain	Gain-Ratio	OneR	Chi-Squared	Principal components	Classifier subset
Molecular biology (Promoter)	670	580	5900	2590	670	670	670	670	670	5900
Connect 4	1060	700	830	3880	1060	1060	1060	1060	1060	830
Soybean	1330	1200	1580	1080	1330	1330	1330	1330	1630	1580
Zoo	3390	2272	4780	4030	3390	3390	3390	3390	4740	4780
Balloon	330	220	170	170	330	330	330	330	330	170
Mushroom	1030	920	360	6380	1030	1030	1030	1030	1030	360
Lenses	860	860	610	890	860	860	860	860	860	610
Fictional	330	190	170	220	330	330	330	330	330	170

We can see the learning accuracy of ID3, C4.5, NB, SVM, and neural network (NN), respectively on different feature sets. From the averaged accuracy over all datasets, we observe that in general, attribute reduction by Bayes′ theorem improves the accuracy of NB, ID3, and NN. From individual accuracy values, we also observe that for most of the datasets, attribute reduction by Bayes′ theorem can maintain or even increase the accuracy. The above experimental results suggest that Bayes′ theorem is practical for feature selection for classification of high dimensional data. It can efficiently achieve a high degree of dimensionality reduction and enhance classification accuracy with predominant features.

6 Conclusions

This paper proposes a feature selection algorithm based on Bayes′ theorem. The algorithm can remove redundancy from the original dataset. The main idea provided is to find the dependent attributes and remove the redundant ones among them. The technology to obtain the dependency needed is based on Bayes′ theorem. A new attribute reduction algorithm of using Bayes′ theorem is implemented and evaluated through extensive experiments via comparison with related attribute reduction algorithms. In this paper, we consider the task of feature selection and investigate the performance of nine feature selection algorithms. Our findings can be summarized as follows:

1) In feature selection approach, we have shown that Bayes′ theorem is a promising approach for automatic feature selection. It outperforms most existing algorithms in terms of number of selected features, classification accuracy, and running time. Well-established algorithms, such as wrapper subset eval, consistency subset eval, InfoGain attribute eval, GainRatio attribute eval, OneR attribute eval, ChiSquared attribute eval, principal components, classifier subset eval, are also more complex than Bayes feature selection. Bayes′ theorem based feature selection runs very efficiently on large datasets, which makes it very attractive for feature selection in high dimensional data.

2) We have implemented a new feature selector using Bayes′ theorem and found that it performs better than the popular and computationally expensive traditional algorithms.

3) We compared the performance of a number of algorithms on the UCI machine learning repository datasets.

Appendix

Tables A1–A5 in Appendix show the accuracy of ID3, J48, SVM, and NN on selected features for each feature selection algorithm and time taken by ID3, J48, NB, and NN on selected features.

Table A1 Accuracy of ID3 on selected features for each feature selection algorithm

Data sets	Classification accuracy (%)									
	Full set	Bayes′ theorem	Wrapper subset	Consistency subset	Info-Gain	Gain-Ratio	OneR	Chi-Squared	Principal components	Classifier subset
Molecular biology (Promoter)	76.41	77.35	75.33	76.41	76.41	76.41	76.41	76.41	76.41	75.23
Connect 4	59.11	72.95	44.33	59.11	59.11	59.11	59.11	59.11	59.11	44.33
Soybean	95.94	100.00	57.44	87.23	95.74	95.74	95.74	95.74	95.74	57.44
Zoo	98.01	99.00	54.45	95.04	98.01	98.01	98.01	98.01	98.01	54.44
Balloon	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	78.54
Mushroom	99.79	100.00	97.38	99.79	99.79	99.79	99.79	99.79	99.79	97.38
Lenses	100.00	100.00	66.66	83.33	100.00	100.00	100.00	100.00	100.00	66.66
Fictional	40.00	70.00	30.00	50.00	40.00	40.00	40.00	40.00	40.00	30.00
Average	**83.66**	**89.91**	**65.70**	**81.36**	**83.63**	**83.63**	**83.63**	**83.63**	**83.63**	**63.00**

Table A2 Accuracy of J48 on selected features for each feature selection algorithm

Data sets	Classification accuracy (%)									
	Full set	Bayes′ theorem	Wrapper subset	Consistency subset	Info-Gain	Gain-Ratio	OneR	Chi-Squared	Principal components	Classifier subset
Molecular biology (Promoter)	81.13	83.80	75.23	63.80	81.13	81.13	81.13	81.13	81.13	81.13
Connect4	66.03	66.98	44.33	62.01	66.03	66.03	66.03	66.03	66.03	44.33
Soybean	97.87	100.00	57.44	95.74	97.87	97.87	97.87	97.87	95.74	57.44
Zoo	92.07	99.00	54.45	91.08	92.07	92.07	92.07	92.07	92.07	54.45
Balloon	100	100.00	100	100	100	100	100	100	100	78.54
Mushroom	99.89	100.00	97.38	99.58	99.89	99.89	99.89	99.89	99.89	97.38
Lenses	91.66	91.66	66.66	83.33	91.66	91.66	91.66	91.66	91.66	66.66
Fictional	20.00	20.00	30	30	20	20	20	20	20	30
Average	**81.08**	**82.68**	**65.69**	**78.19**	**81.08**	**81.08**	**81.08**	**81.08**	**80.70**	**63.74**

Table A3 Accuracy of SVM on selected features for each feature selection algorithm

Data sets	Classification accuracy (%)									
	Full set	Bayes′ theorem	Wrapper subset	Consistency subset	Info-Gain	Gain-Ratio	OneR	Chi-Squared	Principal components	Classifier subset
Promoter	93.39	93.39	78.09	85.71	93.39	93.39	93.39	93.39	93.39	78.09
Connect4	62.01	63.83	44.33	60.75	62.01	62.01	62.01	62.01	62.01	44.33
Soybean	100	100	57.44	97.87	100	100	100	100	97.87	57.44
Zoo	96.03	97.02	54.44	95.04	96.03	96.03	96.03	96.03	96.03	54.44
Balloon	100	100	100	100	100	100	100	100	100	78.54
Mushroom	99.89	100	97.38	99.58	99.89	99.89	99.89	99.89	99.89	97.38
Lenses	83.33	83.33	66.66	70.83	83.33	83.33	83.33	83.33	83.33	66.66
Fictional	100	92	30	60.00	100	100	100	100	100	30
Average	**91.83**	**91.20**	**66.04**	**83.72**	**91.83**	**91.83**	**91.83**	**91.83**	**91.57**	**63.37**

Table A4 Accuracy of NN on selected features for each feature selection algorithm

Data sets	Classification accuracy (%)									
	Full set	Bayes′ theorem	Wrapper subset	Consistency subset	Info-Gain	Gain-Ratio	OneR	Chi-Squared	Principal components	Classifier subset
Promoter	100	100	80	100	100	100	100	100	100	80
Connect 4	69.37	72.94	44.65	69.37	69.37	69.37	69.37	69.37	69.37	44.65
Soybean	100	100	57.44	100	100	100	100	100	100	57.44
Zoo	100	100	54.45	99.00	100	100	100	100	100	54.45
Balloon	100	100	100	100	100	100	100	100	100	78.54
Mushroom	100	100	97.38	100	100	100	100	100	100	97.38
Lenses	100	100	66.66	91.66	100	100	100	100	100	66.66
Fictional	60	80	40	70	60	60	60	60	60	40
Average	**91.17**	**94.12**	**67.57**	**91.25**	**91.17**	**91.17**	**91.17**	**91.171**	**91.17**	**64.89**

Table A5 Running time on selected features for each classification algorithm

	Running time (ms)			
	ID3	J48	NB	NN
Full set	220	260	50	59249
Bayes′ theorem	80	100	5	41390
Wrapper subset	40	230	20	6400
Consistency subset	100	70	160	87830
InfoGain	220	260	50	59249
GainRatio	220	280	50	59249
OneR	220	260	50	59249
ChiSquared	220	260	50	59249
Principal components	210	190	70	54109
Classifier subset	40	230	20	6400

References

[1] R. Agrawal, T. Imielinski, A. Swami. Database Mining: A Performance Perspective. *IEEE Transactions on Knowledge and Data Engineering*, vol. 5, no. 6, pp. 914–925, 1993.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From Data Mining to Knowledge Discovery: An Overview. *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.), pp. 495–515, AAAI Press/MIT Press, Menlo Park, CA, USA, 1996.

[3] J. Han Y. Fu. Attribute-oriented Induction in Data Mining. *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (eds.), pp. 399–421, AAAI Press/MIT Press, Menlo Park, CA, USA, 1996.

[4] J. Han, M. Kamber. *Data Mining: Concepts and Techniques*, Morgan Kaufman, 2005.

[5] H. Liu, H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic, Boston, USA, 1998.

[6] D. Pyle. *Data Preparation for Data Mining*, Morgan Kaufmann, 1999.

[7] A. L. Blum, P. Langley. Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence*, vol. 97, no. 1–2, pp. 245–271, 1997.

[8] H. Liu, H. Motoda. *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Kluwer Academic, Boston, USA, 1998, 2nd printing, 2001.

[9] M. Ben-Bassat. Pattern Recognition and Reduction of Dimensionality. *Handbook of Statistics II*, P. R. Krishnaiah, L. N. Kanal (eds.), North Holland, pp. 773–791, 1982.

[10] A. Jain, D. Zongker. Feature Selection: Evaluation, Application, and Small Sample Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, 153–158, 1997.

[11] P. Mitra, C. A. Murthy, S. K. Pal. Unsupervised Feature Selection Using Feature Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 301–312, 2002.

[12] W. Siedlecki, J. Sklansky. On Automatic Feature Selection. *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 2, no. 2, pp. 197–220, 1988.

[13] N. Wyse, R. Dubes, A. K. Jain. A Critical Evaluation of Intrinsic Dimensionality Algorithms. *Pattern Recognition in Practice*, E. S. Gelsema, L. N. Kanal (eds.), pp. 415–425, Morgan Kaufmann, 1980.

[14] G. H. John, R. Kohavi, K. Pfleger. Irrelevant Feature and the Subset Selection Problem. In *Proceedings of the 11th International Conference on Machine Learning*, Morgan Kaufmann, New Brunswick, New Jersey, USA, pp. 121–129, 1994.

[15] K. Kira, L. A. Rendell. The Feature Selection Problem: Traditional Methods and a New Algorithm. In *Proceedings of the 10th National Conference on Artificial Intelligence*, MIT Press, San Jose, California, USA, pp. 129–134, 1992.

[16] R. Kohavi, G. H. John. Wrappers for Feature Subset Selection. *Artificial Intelligence,* vol. 97, no. 1–2, pp. 273–324, 1997.

[17] M. Dash, K. Choi, P. Scheuermann, H. Liu. Feature Selection for Clustering – A Filter Solution. In *Proceedings of the 2nd International Conference on Data mining*, IEEE Computer Society Press, Maebashi City, Japan, pp. 115–122, 2002.

[18] M. Dash, H. Liu. Feature Selection for Classification. *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[19] Y. Kim, W. N. Street, F. Menczer. Feature Selection for Unsupervised Learning via Evolutionary Search. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, Boston, MA, USA, pp. 365–369, 2000.

[20] E. Leopold, J. Kindermann. Text Categorization with Support Vector Machines: How to Represent Texts in Input Space? *Machine Learning*, vol. 46, no. 1, pp. 423–444, 2002.

[21] K. Nigam, A. K. Mccallum, S. Thrun, T. Mitchell. Text Classification from Labeled and Unlabeled Documents Using EM. *Machine Learning,* vol. 39, no. 2, pp. 103–134, 2000.

[22] Y. Yang, J. O. Pederson. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the 14th International Conference on Machine Learning*, Morgan Kaufmann, Nashville, Tennessee, USA, pp. 412–420, 1997.

[23] Y. Rui, T. S. F. Huang, S. Chang. Image Retrieval: Current Techniques, Promising Directions and Open Issues. *Journal of Visual Communication and Image Representation*, vol. 10, no. 1, pp. 39–62, 1999.

[24] D. L. Swets, J. J. Weng. Efficient Content-based Image Retrieval Using Automatic Feature Selection. In *Proceedings of IEEE International Symposium on Computer Vision*, IEEE Computer Society Press, pp. 85–90, 1995.

[25] K. S. Ng, H. Liu. Customer Retention via Data Mining. *Artificial Intelligence Review*, vol. 14, no. 6, pp. 569–590, 2000.

[26] W. Lee, S. J. Stolfo, K. W. Mok. Adaptive Intrusion Detection: A Data Mining Approach. *Artificial Intelligence Review*, vol. 14, no. 6, pp. 533–567, 2000.

[27] E. Xing, M. I. Jordan, R. M. Karp. Feature Selection for High-dimensional Genomic Microarray Data. In *Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann, Madison, Wisconson, USA, pp. 601–608, 2001.

[28] A. L. Blum, R. L. Rivest. Training a 3-Node Neural Networks is NP-Complete. *Neural Networks*, vol. 5, no. 1, pp. 117–127, 1992.

[29] P. Langley. Selection of Relevant Features in Machine Learning. In *Proceedings of AAAI Fall Symposium on Relevance*, AAAI Press, Menlo Park, California, USA, pp. 140–144, 1994.

[30] A. J. Miller. *Subset Selection in Regression*, 2nd Edition, Chapman & Hall/CRC, 2002.

[31] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning*, Springer, 2001.

[32] J. Doak. An Evaluation of Feature Selection Methods and Their Application to Computer Security, Technical Report, Department of Computer Science, University of California at Davis, USA, 1992.

[33] M. Dash, H. Liu. Handling Large Unsupervised Data via Dimensionality Reduction. In *Proceedings of SIGMOD Research Issues in Data Mining and Knowledge Discovery Workshop*, 1999.

[34] M. Dash, H. Liu, J. Yao. Dimensionality Reduction of Unsupervised Data. In *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, IEEE Press, Newport Beach, CA, USA, pp. 532–539, 1997.

[35] J. G. Dy, C. E. Brodley. Feature Subset Selection and Order Identification for Unsupervised Learning. In *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, USA, pp. 247–254, 2000.

[36] L. Talavera. Feature Selection as a Preprocessing Step for Hierarchical Clustering. In *Proceedings of the 16th International Conference on Machine Learning*, Morgan Kaufmann, Bled, Slovenia, pp. 389–397, 1999.

[37] M. A. Hall. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, Stanford University, USA, pp. 359–366, 2000.

[38] H. Liu, R. Setiono. A Probabilistic Approach to Feature Selection – A Filter Solution. In *Proceedings of the 13th International Conference on Machine Learning*, Morgan Kaufmann Publishers, Bari, Italy, pp. 319–327, 1996.

[39] L. Yu, H. Liu. Feature Selection for High-dimensional Data: A Fast Correlation-based Filter Solution. In *Proceedings of the 20th International Conference on Machine Learning*, AAAI Press, Washington DC, USA, pp. 856–863, 2003.

[40] R. Caruana, D. Freitag. Greedy Attribute Selection. In *Proceedings of the 11th International Conference of Machine Learning*, Morgan Kaufmann, New Jersey, USA, pp. 28–36, 1994.

[41] S. Das. Filters, Wrappers and a Boosting-based Hybrid for Feature Selection. In *Proceedings of the 18th International Conference on Machine Learning*, Morgan Kaufmann, Williams College, Williamstown, MA, USA, pp. 74–81, 2001.

[42] A. Y. Ng. On Feature Selection: Learning with Exponentially Many Irrelevant Features as Training Examples. In *Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann, Madison, Wisconson, USA, pp. 404–412, 1998.

[43] J. R. Quinlan. Induction of Decision Trees. *Machine Learning,* vol. 1, no. 1, pp. 81–106, 1986.

[44] J. R. Quinlan. *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco, 1993.

[45] L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen. *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1984.

[46] R. S. Michalski. Pattern Recognition as Rule-guided Inductive Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 4, pp. 349–361, 1980.

[47] P. M. Narendra, K. Fukunaga. A Branch and Bound Algorithm for Feature Subset Selection. *IEEE Transactions on Computers*, vol. 26, no. 9, pp. 917–922, 1977.

[48] P. Pudil, J. Novovicova, J. Kittler. Floating Search Methods in Feature Selection. *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125,1994.

[49] P. Somol, P. Pudil, J. Kittler. Fast Branch and Bound Algorithm in Feature Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 900–912, 2000.

[50] J. Casillas, O. Cordon, M. J. Del Jesus, F. Herrera. Genetic Feature Selection in a Fuzzy Rule-based Classification System Learning Process for High-dimensional Problems. *Information Sciences*, vol. 136, no. 1–4, pp. 135–157, 2001.

[51] N. Xiong. A Hybrid Approach to Input Selection for Complex Processes. *IEEE Transactions on Systems, Man, and Cybernetics – Part A*, vol. 32, no. 4, pp. 532–536, 2002.

[52] L. I. Kuncheva, J. C. Bezdek. Nearest Prototype Classification: Clustering, Genetic Algorithms or Random Search. *IEEE Transactions on Systems, Man, and Cybernetics – Part C*, vol. 28, no. 1, pp. 160–164, 1998.

[53] S. Y. Ho, C. C. Liu, S. Liu. Design of an Optimal Nearest Neighbor Classifier Using an Intelligent Genetic Algorithm. *Pattern Recognition Letters*, vol. 23, no. 13, pp. 1495–1503, 2002.

[54] R. Thawonmas, S. Abe. A Novel Approach to Feature Selection Based on Analysis of Class Regions. *IEEE Transactions on Systems, Man, and Cybernetics – Part B*, vol. 27, no. 2, pp. 196–207, 1997.

[55] K. Kira, L. A. Rendell. A Practical Approach to Feature Selection. In *Proceedings of the 9th International Conference on Machine Learning*, Morgan Kaufmann, Aberdeen, Scotland, pp. 249–256, 1992.

[56] I. Kononenko. Estimating Attributes: Analysis and Extensions of RELIEF. In *Proceedings of Europe International Conference on Machine Learning*, Springer-Verlag, New York, USA, pp. 171–182, 1994.

[57] S. Cost, S. Salzberg. A Weighted Nearest Algorithm with Symbolic Features. *Machine Learning*, vol. 10, no. 1, pp. 57–78, 1993.

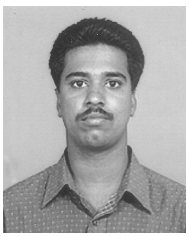
[58] C. Stanfill, D. Waltz. Towards Memory Based Reasoning. *Communications of the ACM*, vol. 29, no. 12, pp. 1213–1228, 1986.

[59] S. Zhao, E. C. C. Tsang. On Fuzzy Approximation Operators in Attribute Reduction with Fuzzy Rough Sets. *Information Sciences*, vol. 178, no. 16, pp. 3163–3176, 2008.

[60] A. Sharma, K. K. Paliwal. Rotational Linear Discriminate Analysis Technique for Dimensionality Reduction. *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 10, pp. 1336–1347, 2008.

[61] C. L. Blake, C. J. Merz. UCI Repository of Machine Learning Databases, Department of Information and Computer Science, Universitry of California, Irvine, USA, [Online], Available: http://www.ics.uci.edu/m̃learn, 1998.

[62] J. Joyce. Bayes′ Theorem. *Standford Encyclopedia of Philosophy*, E. N. Zalta (ed.), The Metaphysics Research Lab, Stanford University, USA, 2003.



Subramanian Appavu Alias Balamurugan is a Ph. D. candidate at the Department of Information and Communication Engineering, Anna University, Chennai, India. He is also an faculty at Thiagarajar College of Engineering, Madurai, India.

His research interests include data mining and text mining.

Ramasamy Rajaram received the Ph. D. degree from Madurai Kamaraj University, India. He is a professor of Department of Computer Science and Information Technology at Thiagarajar College of Engineering, Madurai, India.

His research interests include data mining and information security.