Enhancing Classifier Performance Via Hybrid Feature Selection and Numeric Class Handling- A Comparative Study

S. Vijayasankari*
Assistant Professor in Computer Applications
EMG Yadava Women's College
Madurai – 625014, India
*Corresponding Author

K. Ramar
Professor of Computer Science Engineering
Principal, Einstein College of Engineering,
Tirunelveli – 627 012, India

ABSTRACT

Classification is a supervised machine learning procedure in which the effective model is constructed for prediction. The accuracy of classification mainly depends on the type of features and the characteristics of the dataset. Feature selection is an efficient approach in searching the most descriptive features which would contribute to the increase in the performance of the inductive algorithm by reducing dimensionality and processing time. In the present work a hybrid embedded feature selection algorithm with class label refining and handled numeric class problem in classifier are implemented. A novel feature selection algorithm based on ranker search optimization method and ensemble genetic search for selecting the appropriate features and class label refining for correcting misclassified instances from the dataset have been done. By modelling this approach, it reaches a near global optimal solution set of features. Hybrid feature selection with enhanced C4.5 and naïve bayes classification can handle numeric class to achieve better performance. The efficiency of this method is demonstrated by comparing with the other existing methods in terms of accuracy, number of features selected and ability to handle numerical class values. Experimental results on datasets reveals that the proposed algorithm increases the classifier accuracy with less error rate and the quality of results are comparable.

Keywords

Data mining, Hybrid feature selection, Classification, Decision tree, Accuracy

1. INTRODUCTION

Data mining (DM) deals with the problem of discovering novel and interesting knowledge from large amount of data. This problem is often performed heuristically when the extraction of patterns is difficult using standard query mechanisms or classical statistical methods. A comparison with the results achieved by other techniques on a classical benchmark set is carried out. Furthermore, some of the obtained rules are shown and the most discriminating variables are evidenced.. The core of this process is the application of machine learning based algorithms to databases. There are two basic ways of performing data mining and they are supervised and unsupervised learning. The former exploits known cases that show or imply well defined patterns to find new patterns by means of which generalizations are formed. Experts lead search towards some features which are supposed to be of prominent interest, and the relationships between those features and the remaining ones are sought. In unsupervised learning, data patterns are found from some logical characterization of the regularities in a set of data. In this case, no pre-assumptions are made about the forms of relations among attributes. Data classification represents the most commonly applied supervised data mining technique.

Data mining sometimes referred as Knowledge Discovery in Database (KDD), is a systematic approach to find the underlying patterns, trend and relationships buried in data. Basically, the researches regarding DM can be classified into two categories such as methodologies and technologies. The technology part of DM consists of techniques such as statistical methods, neural networks, decision trees, genetic algorithms, and non-parametric methods. In the proposed work a novel Feature Selection (FS) algorithm based on Ranker Search (RS) optimization method and Ensemble Genetic Search (EGS) is chosen for selecting the appropriate features and also class label refining for correcting misclassified instances from the dataset. The anomaly detection methods for mobile ad hoc networks to detect the intrusions are used with genetic algorithm technique this audit data is reduced by means of feature selection technique [14].

The remainder of this paper is organized as follows. Section 2 discusses literature review related to this work. Section 3 describes the proposed methodology includes the data cleaning, filter approach, enhancing the C4.5, naïve bayes classifier and building the classifier Section 4 reports the results of experiments evaluating the performance of the system. Finally, Section 5 presents conclusions.

2. LITERATURE REVIEW

Data mining uses a broad family of computational methods that include statistical analysis, decision trees, neural networks, rule induction and refinement, and graphics visualization. Although, data mining tools have been available for a long time, the advances in computer hardware and software, particularly exploratory tools like data visualization and neural networks, have made data mining more attractive and practical.

Classification is an important theme in data mining [4]. Rough sets and neural networks are two common techniques applied to data mining problems and integrating the advantages of two, approach a hybrid system to extract classification rules efficiently from decision table. The classification problem is a two - step process, where the first is to build a classification model by analyzing the training sample set described by attributes and the second is to use this model to classify the future sample for which the class label is not known. Among these approaches, the approach of decision tree classifier is probably the most popular and widespread, for its strength such as generation of understandable rules, classification without requiring much computation, ability to handle both continuous and categorical variables and to providing a clear indication of which attributes are the most important for prediction or classification.

In the field of knowledge discovery and data mining, many techniques have been suggested to perform classification. However, most of this work primarily focuses on developing models with high predictive accuracy without trying to explain how the classifications are being made. The decision trees generated by J48 can be used for classification. J48 builds decision trees from a set of labelled training data using the concept of information entropy. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 examines the normalized information gain that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then the algorithm recurs on the smaller subsets. The splitting procedure stops if all the instances in a subset belong to the same class. Then a leaf node is created in the decision tree telling to choose that class. But it can also happen that none of the features give any information gain. In this case J48 creates a decision node higher up in the tree using the expected value of the class.

Data-mining algorithms were used in many classification problems [8]. Among them, the decision tree (DT), back-propagation network (BPN), and support vector machine (SVM) are popular and can be applied to various areas**.** The global pattern mining step in existing pattern-based hierarchical clustering algorithms may result in an unpredictable number of patterns [10].

ID6NB is a novel algorithm for building simple and reasonably accurate classification model [1]. In that, the unhandled exceptions of the decision tree induction algorithm and improved its performance by fusing data cleaning, dimensionality reduction, and data smoothening with the algorithm ID6NB. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. Generalized Discriminant Analysis (GDA) has provided an extremely powerful approach to extracting non-linear features [12].

Intrusion Detection Systems (IDS) examine all data features to detect intrusion or misuse patterns [11]. They were also investigated new techniques for intrusion detection and performed data reduction and evaluated their performance on the benchmark intrusion data and explored general Bayesian Network (BN) classifier and Classification and Regression Trees (CART) as intrusion detection models. Filter and wrapper models are used for selecting and protecting network and its resources from illegal penetration [13].

From a theoretical perspective, it can be shown that optimal feature selection for supervised learning problems requires an exhaustive search of all possible subsets of features of the chosen cardinality. If large numbers of features are available, this is impractical. For practical supervised learning algorithms, the search is for a satisfactory set of features instead of an optimal set. Feature selection algorithms typically fall into two categories such as feature ranking and subset selection.

Integration of classifiers is currently an active research area in the machine learning and neural networks communities [5]. Feature extraction is a special form of dimensionality reduction. In statistics, dimensionality reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction.

A novel hybrid method is used for feature selection in micro array data analysis. This method uses a genetic algorithm with dynamic parameter setting (GADP) to generate a number of subsets of genes and to rank the genes according to their occurrence frequencies in the gene subsets [7]. To improve computational efficiency, feature selection technique, feature selection via supervised model construction (FSSMC), an optimisation of relief-F, was used to rank the important attributes affecting diabetic control [9].

3. PROPOSED METHODOLOGY

3.1 Training and Building the classifier

The trained enhanced C4.5 algorithm classifier is used for efficient prediction. The newly refined dataset is used to train and build the enhanced C4.5 and NB classifier and then it is used for prediction purpose as per the flow chart shown in the figure 1.

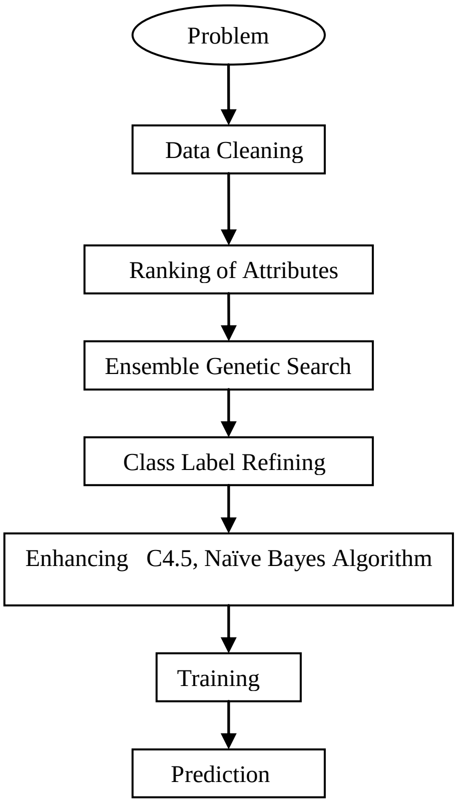

Figure 1 General flow diagram of the proposed system

3.2 Data Cleaning

There are a number of data pre-processing techniques. Data cleaning can be applied to remove noise and inconsistencies in the data. These data processing techniques, when applied prior to mining, can substantially improve the overall data mining results. Data cleaning routines work to clean the data by identifying redundant or duplicate data and removing them, and resolving inconsistencies. If the users believe that the data are dirty, they are unlikely to trust the results of any data mining that has been applied to it. Furthermore, dirty data can cause confusion in the mining procedure, resulting in an unreliable output. Although, most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust. Instead, they may concentrate on avoiding over fitting the data to the function being modelled. Therefore, a useful pre-processing step is to run your data through the following data cleaning routines.

3.2.1. Replacing Missing Values

Missing value in the dataset can be replaced by using ignore the tuple, fill in the missing value manually, use a global constant to fill in the missing values, use the attribute mean to fill in the missing value, use the attribute mean, use the most probable value to fill in the missing value.

3.2.2. Removing Redundant data

Having a large amount of redundant data may slow down or confuse the knowledge discovery process. The removal of redundant data may be seen as a form of data cleaning, as well as data reduction. To remove redundant data, we search the entire dataset in sequential manner to check whether a tuple is duplicated i.e. whether a tuple is repeated one or more times.

3.2.3. Handling Conflicting data

When instances have same set of attribute values but different class labels, then they are said to be conflicting instances. The presence of conflicting instances in dataset makes the learning process very complex. Thus the conflicting instances must be identified and it should be rectified with correct class label.

3.3. Integrated Feature Selection and Filter Approach (IFSFA)

The wrapper approach is usually associated with a considerable computational effort since it requires the rerunning of an induction algorithm multiple times. The filter methods, on the other hand, are computationally cheaper, but, there is a danger that the features selected by a filter method will not allow a classification algorithm to fully exploit its potential. Feature selection and filter approach can be done by the following methods

3.3.1. Discretization of Attributes

Data discretization is a procedure that takes a data set and converts all continuous attributes to categorical. Many algorithms developed in the machine learning community focus on learning in nominal feature spaces. However, many real-world classification tasks involve continuous features where such algorithms could not be applied unless the continuous features are discretized. Supervised discretization method is used here since majority of datasets contains class labels.

3.3.2. Ranking Attributes

Features are ranked based on their efficiency to partition the dataset. Ranking is a crucial part of feature selection. It is able to compute sorted score when features are given. Depending on applications the scores may represent the degrees of relevance, preference, or importance. In this present work, without loss of generality, ranking in relevance, search was done. Traditionally only a small number of strong features were used to represent relevance and to rank documents. The following attribute evaluations are used such as Information Gain (IG), gain ratio, symmetrical uncertainty, relief-F, one-R and chi-squared. In this work the evaluation of the practical usefulness of IG attribute is considered. Entropy is a common measure used in the information theory, which characterizes the purity of an arbitrary collection of examples. It is in the foundation of the IG attribute ranking methods.

3.3.3. Filter Method

Pre-process computes score for each feature and then select feature according to the score or the threshold value. All filter methods use heuristics based on general characteristics of the data rather than a learning algorithm to evaluate the merit of feature subsets. As a consequence, filter methods are generally much faster than wrapper methods, and are more practical for use on data of high dimensionality. The filter method is used in the proposed method for feature selection in correlation based approach. The central hypothesis is that good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other.

3.3.4. Wrapper Method

The wrapper utilizes learning as a black box to score subset features. Wrapper strategies for feature selection use an induction algorithm to estimate the merit of feature subsets. The rationale for wrapper approaches is that the induction method that will ultimately use the feature subset should provide a better estimate of accuracy than a separate measure that has an entirely different inductive bias. Interacting features are that whose values are dependent on the values of other features and also provides further information about the class. On the other hand, redundant features are those whose values are dependent on the values of other features irrespective of the class.

3.3.5. Embedded Method

Feature selection is performed both in filter and wrapper method. In this proposed method the embedded method for feature selection using genetic search is applied. The goal of wrapper-filter feature selection algorithm (WFFSA) is to improve the classification performance and accelerate the search to identify important feature subsets. In particular, the filter method fine-tunes the population of GA solutions by adding or deleting features based on univariate feature ranking information. Hence, the focus is on filter methods that are able to assess the goodness or ranking of the individual features. In the first step, Using C4.5 classifier each feature subset produced is evaluated for their accuracy and correct set of features is produced as a result.

3.3.6. Genetic Algorithm

As such they represent an intelligent exploitation of a random search they are used to solve optimization problems. Although randomized, GA are by no means random, instead they exploit historical information to direct the search into the region of better performance within the search space. The basic techniques of the GA are designed to simulate processes in natural systems necessary for evolution especially those follow the principles laid down on Charles Darwin of survival of the fittest.

3.3.7. Class Label Refining

In this phase correct class label for the instances are fixed. This is done using supervised learning method. Supervised learning is a machine learning technique for deducing a function or class label from training data. The training data consist of pairs of input objects, and desired outputs. The parallel task in human and animal psychology is often referred as concept learning.

3.4. Enhancing C4.5 and Naïve Bayes Classifier

The decision trees generated by C4.5 was used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy [3].

3.4.1. Creating New Class Label

In order to rectify the problem in these algorithms, a new class label for the datasets is created. Since the dataset with numerical class value cannot be discretized it is necessary to create a virtual class label for the dataset and discretize all other attribute including the original class label of dataset. Here a new class label is created by applying a mathematical expression to existing attributes.

3.4.2. Discretization of Attributes

Data discriminator is a procedure that takes a data set and converts all continuous attributes to categorical. Many algorithms developed in the machine learning community focus on learning in nominal feature spaces. However, many real-world classification tasks exist that involve continuous features where such algorithms could not be applied unless the continuous features are discretized. Recently continuous variable discretization has received significant attention in the machine learning community.

3.4.3. Deleting newly formed class label

Now all the attributes in the dataset are converted into nominal values after discretization. Here the attributes are removed and in which it has been formed using expression to facilitate discretization. Refined dataset can now be fed into any classifier without numerical class label problems.

3.5. Proposed Algorithm for Enhancing C4.5 and NB Classifier

```
BEGIN
If Class attribute is Numeric
begin
For i in 1 to total no of instance
begin
Find a new class value by using expression (^,+)
End For
Create a new temporary class attribute using newly formed
   values
Discretize the dataset
Delete the new created temporary class attribute
End If
END
```

4. RESULT AND DISCUSSION

4.1 Dataset Description

The performance of the proposed algorithm was evaluated with 10 publicly available datasets. Descriptions of these 10 datasets are as shown in Table 1. All these datasets are posted in the UCI Repository of Machine Learning Databases [2] and widely used by the data mining community for evaluating learning algorithms [6]. The dataset used here are having huge dimensions with maximum attributes of 2000 and maximum instances of 800.

Table 1 Description of Datasets

Name	Default Task	Attribute Types	Instances	Attributes	Year
Anneal	Classification	Categorical, Integer, Real	798	38	1995
Balance Scale	Classification	Categorical	625	4	1994
Breast Cancer	Classification	Categorical	286	9	1988
Bridge	Classification	Categorical, Integer, Real	107	13	1998
C.N System	Classification	Real	60	74	1994
Colic	Classification	Integer	62	2001	1994
Colon Tumor	Classification	Integer, Real	368	28	1998
Credit	Classification	Categorical	512	39	1995
Cylinder Bands	Classification	Integer	503	40	2010
Ecoli	Classification	Real	336	8	1996

Table 2 Classifier accuracy before and after applying proposed method

Dataset	Logistic		Decision Table		J48		KStar		OneR		Random Tree	
	B	A	B	A	B	A	B	A	B	A	B	A
Anneal	83.63	83.52	98.89	98.55	98.44	98.33	95.77	97.80	83.63	83.52	96.44	98.44
Balance Scale	85.86	86.88	77.26	83.82	77.41	85.57	87.76	90.20	59.18	61.81	78.43	87.76
Breast Cancer	96.63	97.22	96.05	96.34	95.02	97.80	81.70	98.70	91.80	95.75	94.44	98.39
Bridges	15.23	65.71	54.29	78.09	56.19	82.86	61.90	85.70	15.24	65.71	51.43	80.95
Central Nervous System	63.33	81.67	76.67	88.33	63.33	83.33	50.00	85.00	63.33	81.67	80.00	86.67
Colic	80.98	86.14	81.25	88.86	85.33	88.32	76.63	86.30	81.52	81.25	70.65	91.03
Colon Tumor	69.35	79.03	74.19	85.48	82.26	88.71	35.48	91.90	69.35	79.03	50.00	85.48
Credit	85.51	86.23	85.07	84.64	86.09	89.86	78.99	92.50	85.51	86.23	74.93	90.43
Cylinder Bands	78.70	89.44	67.59	84.63	57.78	61.67	100.0	90.70	49.63	77.59	65.56	85.00
Ecoli	86.31	87.20	76.49	85.71	84.23	87.50	80.95	89.30	63.69	66.67	78.27	88.39
Average	**74.553**	**84.304**	**78.775**	**87.445**	**78.608**	**86.395**	**74.918**	**90.81**	**66.288**	**77.923**	**74.015**	**89.254**

4.2 Performance Analysis Based on Classifiers Accuracy

The Table 2 shows classifier accuracy before (B) and after (A) the application of proposed IFSFA method for Logistic, Decision Table, C4.5, KStar, OneR and Random Tree. From the result it is found that the classifier accuracy increases after applying the proposed method.

Table 3 Classifier Accuracy on data set

Dataset	Original Dataset	After Ranking	After Ensem ble GA	After Class Label Refining
Anneal	99.77	99.77	99.44	**99.42**
Balance Scale	86.73	86.73	86.73	**87.46**
Breast Cancer	97.95	97.95	97.95	**98.15**
Bridges	65.71	65.71	66.03	**70.74**
Central Nervous System	96.66	96.66	96.11	**96.42**
Colic	85.87	85.87	85.32	**85.75**
Colon Tumor	98.39	98.39	98.39	**98.62**
Credit	88.55	89.63	88.33	**89.13**
Cylinder Bands	57.78	57.78	57.78	**58.33**
Ecoli	92.26	92.26	92.26	**92.73**
Average	86.967	87.075	86.834	**87.675**

The Table 3 shows classifier accuracy in each phase for the dataset. This shows how the dataset is refined step by step in each phase. The following table 4 shows the classifier accuracy for dataset having numerical class value. It is done by using modified C4.5 and NB algorithm which is made to handle numeric class value. Here 9 datasets from UCI repository having numeric class is taken and passed to modified C4.5 and NB for handling numeric class exception.

Table 4 Classifier Accuracy for numeric class label dataset using J48

Dataset	MC4.5	MNB
Autos	84.39	64.39
Blood	76.17	76.31
CMC	49.29	50.71
Cont	49.56	50.85
CPU	88.04	90.91
Houses	70.13	62.93
Movement Libras	50.42	66.29
Slump Test	34.95	35.92
Wine	32.58	35.96

Figure 2 shows how the proposed method producing higher accuracy when compared with existing feature selection methods like Genetic search, Greedy stepwise search, Ranker, Rank Search. Table 5 shows the tabulated result of comparing various feature selection methods and Table 6 shows the number of features selected by various feature selection methods.

Table 5 Accuracy of various Feature Selection Methods

Dataset	Hybrid	Genetic	Greedy Step wise	Ranker	Ranker Search
Anneal	99.42	98.21	96.88	98.44	97.88
Balance Scale	87.46	77.41	77.41	77.55	77.41
Breast Cancer	98.15	95.17	95.17	95.17	95.17
Bridges	70.74	55.24	57.14	56.19	55.24
Central Nervous System	96.42	73.33	65.00	63.33	68.33
Colon Tumor	98.61	80.64	87.09	87.09	90.32
Colic	66.45	66.30	66.30	66.30	66.30
Credit	89.13	85.07	85.07	86.09	85.07
Cylinder Bands	58.33	56.67	56.67	57.78	56.67
Ecoli	92.68	84.23	84.23	84.23	84.23
Average	85.739	77.227	77.096	77.217	77.662

Table 6 Number of Features selected by various Feature Selection Methods

Dataset	Attribute	Instance	Hybrid	Genetic	Greedy	Ranker	Rank Search
Anneal	39	898	12	13	13	39	19
Balance Scale	5	686	4	4	4	4	4
Breast Cancer	11	683	9	9	9	10	9
Bridges	13	107	7	7	2	12	8
Central Nervous System	74	60	40	46	39	73	46
Colon Tumor	2001	62	61	574	87	2000	47
Colic	28	368	2	2	2	27	2
Credit	16	690	7	7	7	15	7
Cylinder Bands	40	540	9	9	6	39	6
Ecoli	8	336	6	6	6	7	6
Average			15.7	67.7	17.5	222.6	15.4

4.3 Comparative Analysis of Various Feature Selection Methods Using Classifier Accuracy, Precision, Recall and Time required for building the model.

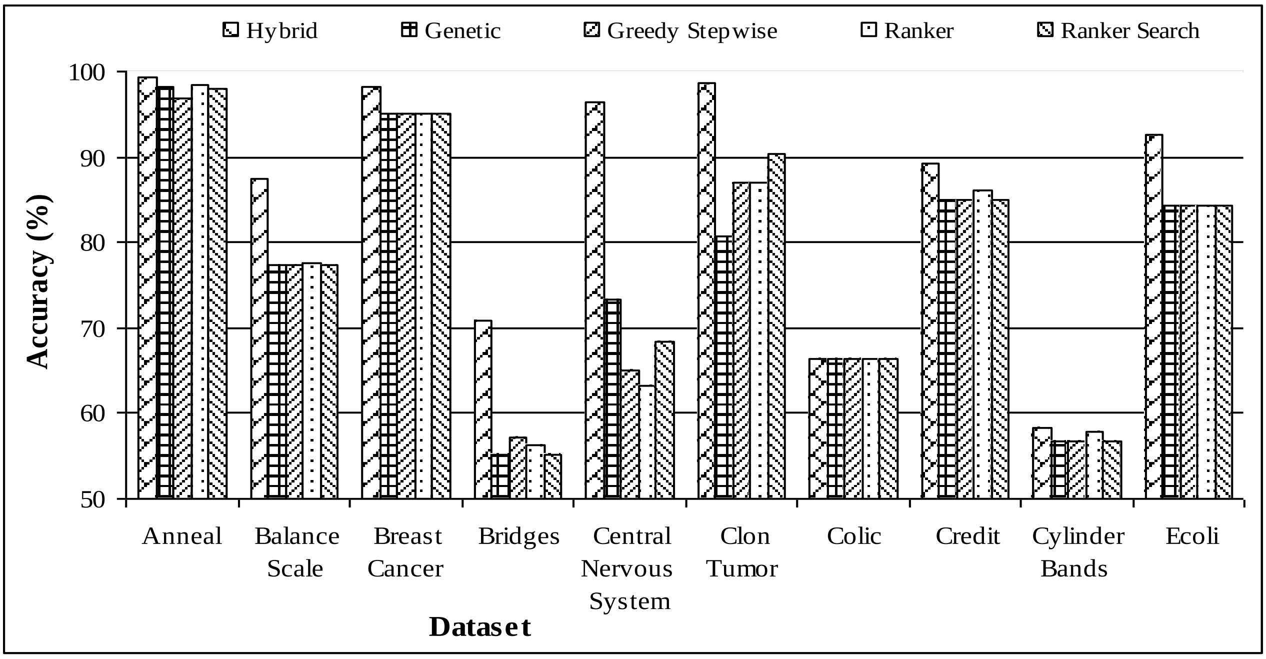

Figure 2 Comparison of performances of various feature selection methods

Table 5 and Figure 2 shows the comparative study of various feature selection algorithm using precision, recall value. From the results it was found that the proposed method is having high precision, recall value comparing with other methods. Table 7 present the precision (P), recall (R) values of various feature selection methods. Figure 3 list outs the comparison of various feature selection algorithm using precision and recall and the Table 8 shows the time required for processing various feature selection methods.

Table 7 Precision, Recall values of various feature selection methods

Dataset	Genetic		Greedy		Rank Search		Ranker		Hybrid	
	P	R	P	R	P	R	P	R	P	R
Anneal	0.982	0.982	0.968	0.969	0.979	0.979	0.984	0.984	**0.983**	**0.983**
Balance	0.745	0.774	0.745	0.774	0.745	0.774	0.745	0.776	**0.838**	**0.856**
Breast Cancer	0.952	0.952	0.952	0.952	0.952	0.952	0.952	0.952	**0.978**	**0.978**
Bridge	0.515	0.552	0.36	0.571	0.537	0.552	0.535	0.562	**0.852**	**0.829**
Central Nervous System	0.626	0.633	0.67	0.683	0.639	0.65	0.724	0.733	**0.833**	**0.833**
Colic	0.44	0.663	0.44	0.663	0.44	0.663	0.44	0.663	**0.454**	**0.674**
Colon Tumor	0.871	0.871	0.903	0.903	0.873	0.871	0.804	0.806	**0.886**	**0.887**
Credit	0.308	0.555	0.308	0.555	0.308	0.555	0.308	0.555	**0.899**	**0.899**
Cylinder Bands	0.334	0.575	0.334	0.575	0.334	0.575	0.334	0.575	**0.38**	**0.617**
Ecoli	0.181	0.426	0.181	0.426	0.181	0.426	0.181	0.426	**0.874**	**0.872**
Average	0.5954	0.6983	0.5861	0.7071	0.5988	0.6997	0.6007	0.7032	0.7977	**0.8428**

Table 8 Time (in seconds) required for processing various feature selection methods

Dataset	Genetic	Greedy	Rank Search	Ranker	Hybrid
Anneal	0.06	0.06	0.16	0.24	0.08
Balance	0.05	0.06	0.09	0.03	0.05
Breast Cancer	0.03	0.03	0.03	0.06	0.03
Bridge	0	0	0	0	0.01
Central Nervous System	0.02	0.02	0.02	0.03	0.02
Colic	0	0	0	0.02	0
Colon Tumor	0.24	0.01	0.02	0.75	0.02
Credit	0	0	0	0	0.03
Cylinder Bands	0	0	0	0	0.02
Ecoli	0	0	0	0	0.06
Average	0.04	0.018	0.032	0.113	0.032

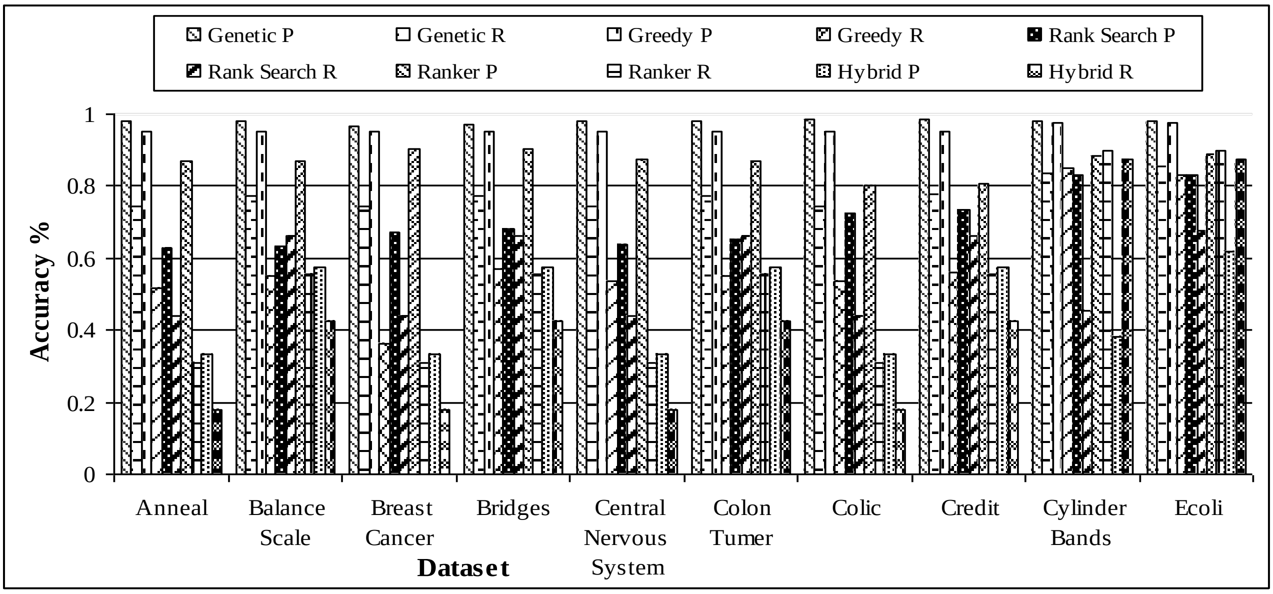

Figure 3 Comparing feature selection algorithms using precision and recall

5. CONCLUSION

Classification is one of the important processes in data mining which is used to train and build a classifier or derive a set of rules based upon the given dataset. In this present work a novel hybrid feature selection method is used to select relevant features, cast away irrelevant and redundant features from the original feature set and solution to handle numeric class label for C4.5 and NB classifier is proposed. The efficiency and effectiveness of the proposed method is demonstrated through extensive comparisons with other methods using real world data of high dimensionality. This generated classifier can be used to predict or analyze the unknown class labels. The efficiency of the classifier mainly depends upon the quality of dataset i.e. the features and instances. We have also proposed a new framework of efficient feature selection via relevance and redundancy analysis, and a correlation-based feature selection method. A new feature selection algorithm of integrated feature selection and filter approach is implemented and evaluated through extensive experiments comparing with six representative feature selection algorithms. In addition, the inability of C4.5 and naïve bayes classifier to handle numeric class label is rectified. The feature selection results are further verified by six different learning algorithms.

6. REFERENCES

[1] Appavu S and Rajaram R. 2009. Knowledge - based system for text classification using ID6NB algorithm, Knowledge-Based Systems, Vol. 22. pp. 1–7.

[2] Blake C.L. and Merz C.J. UCI Repository of Machine Learning Databases,

http://www.ics.uci.edu/~mlearn/MLRepository.html.

[3] Ong Y. S. and Keane A.J. 2004, Meta-Lamarckian in Memetic Algorithm, IEEE Trans. Evolutionary Computation, Vol. 8, pp. 99 -110.

[4] Renpu Li, Zheng - ou Wanga. 2004, Computing, Artificial Intelligence and Information Technology Mining classification rules using rough sets and neural networks, European Journal of operational Research, Vol. 157, pp. 439 - 448.

[5] Umamaheswari G, Ramar K, Manimegalai D and Gomathi V. 2011, An adaptive region based color texture segmentation using fuzzified distance metric, Applied soft computing, Vol. 11, pp.2916 – 2924.

[6] Waikato Environment for Knowledge analysis (WEKA), Machine learning algorithms in java.

[7] Chien-Pang Lee and Yungho Leu. 2011, A novel hybrid feature selection method for micro array data analysis, Applied Soft Computing, Vol. 11, pp. 208-213.

[8] Shih-Chieh Chen, Shih-Wei Lin and Shuo-Yan Chou. 2011, Enhancing the classification accuracy by scatter-search-based ensemble approach, Applied Soft Computing, Vol. 11 pp. 1021-1028.

[9] Yue Huang, Paul McCullagh, Norman Black and Roy Harper, 2007, Feature selection and classification model construction on type 2 diabetic patients' data, Artificial Intelligence in Medicine, Vol. 41, pp. 251-262.

[10] Hassan H. Malik, John R. Kender · Dmitriy Fradkin, and Fabian Moerchen, 2010, Hierarchical document clustering using local patterns, Data Mining and Knowledge Discovery, Vol. 21, pp. 153-185.

[11] Srilatha Chebrolu, Ajith Abraham and Johnson P Thomas, 2005, Feature deduction and ensemble design of intrusion detection system, Computers and Security, Vol.24, pp.295-307

[12] Shailendra Singh and Sanjay Silakari, 2009, Generalized Discriminant Analysis algorithm for feature reduction in Cyber Attack Detection System', International Journal of Computer Science and Information Security, Vol. 6, pp - 173-180.

[13] Pratik Neelakantan N and Nagesh C, 2011, Role of Feature Selection in Intrusion Detection Systems for 802.11 Networks, International Journal of Smart Sensors and Ad Hoc Networks, Vol.1, pp: 98-101.

[14] Nallusamy R, Jayarajan K, and Duraiswamy K. 2009, Intrusion Detection In Mobile Ad Hoc Networks Using GA Based Feature Selection, Computer Science and Telecommunications, Vol.5, pp-28-35.