Feature Selection for Large Scale Data by Combining Class Association Rule Mining and Information Gain: a Hybrid Approach

Appavu alias Balamurugan, Pramala, Rajalakshmi and Rajaram
Department of Information Technology,
Thiagarajar College of Engineering, Madurai, India

***Abstract*— Feature selection is a fundamental problem in data mining to select relevant features and cast away irrelevant and redundant features from an original feature set based on some evaluation criterion. In this paper, we propose a filter method to find associate attributes with respect to class and rank using information gain. Highly associated features are searched using class association rule mining. Information gain is used for removing redundancy and for further pruning.The efficiency and effectiveness of our method is demonstrated through extensive comparisons with other methods using real-world data of high dimensionality.**

***Index Terms*— Association mining, Data mining, Feature selection, Information gain**

I. INTRODUCTION

DATA MINING deals with the discovery of hidden knowledge, unexpected patterns and new rules from large databases. It often involves datasets with a large number of attributes. Many of the attributes in most real world data are redundant and/or simply irrelevant to the purposes of discovering interesting patterns.

Attribute reduction selects relevant attributes in the dataset prior to performing data mining. Attribute reduction is also known as feature selection. For example, high dimensional data (i.e., data sets with hundreds or thousands of features), can contain high degree of irrelevant and redundant information which may greatly degrade the performance of learning algorithms. Therefore, feature selection becomes very necessary for machine learning tasks when they face high dimensional data nowadays. Thus the dimensionality of the datasets has to be reduced. Selection of features is one of the most important tasks for designing a good classifier. Before we enter the learning phase, it is said that as many features as possible are to be collected for improving the performance. Yet irrelevant and correlated may degrade the performance of the classifier. Large number of features may yield to complex models which make interpretation difficult. This work aims to enhance feature selection methodologies and correspondingly the accuracy and performance of the algorithm.

II. RELATED WORK

Several feature selection techniques have been proposed in the literature, including some important surveys on feature selection algorithms such as Molina et al.[2] Guyon and Elisseeff [3]. Different feature selection methods can be categorized into the wrapper [4], filter [5-9] and the embedded model [1, 5]. The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subset. The most important point about this method is the higher computational cost [4]. The filter model separates feature selection from classifier learning and selects feature subsets that are independent of any learning algorithm. It relies on various measures of the general characteristics of training data such as distance, information dependency and consistency [15]. Filter methods are found to perform faster than wrapper methods and are therefore widely used to deal with high dimensional datasets. Several feature selection techniques have been proposed in the literature like Correlation-based feature selection (CFS), Principal Component Analysis (PCA),Gain Ratio Attribute Evaluation(GR) etc[17]. Most of these methods usually combine with some other method to find the appropriate number of attributes [18, 19]. There are some proposals on how to find optimal feature selection algorithms. They have comprehensively discussed about attribute selection based on entropy metrics [10, 11, 12]. Several new feature selection methods have been derived based on entropy metric such as symmetrical uncertainty, information gain and gain ratio [13] and mutual information [14]. The concept of two dimensional discriminant rules initially grew in association rules generation and also support for classification and regression [7]. The core of the operation is laid on the concept of x-monotone region optimization. The concern of correlation in numeric and discrete class has been discussed in detail in comparison to accuracy level is considered for single attributes .The need to accommodate pair wise attributes as a two dimensional feature is applied as a pattern [18].Other recent researchers also attempt to explore the pair wise attributes selection in the aspect of noise detection [19].

III. PROBLEM DEFINITION

The enormity in the number of instances causes serious problems to many machine learning algorithms. High dimensional data (i.e., data sets with hundreds or thousands of features) can contain high degree of irrelevant and redundant information which may greatly degrade the performance of learning algorithms. In general any learning algorithm faces the problem of selecting a subset of features upon which attention must be focused. Therefore, feature selection becomes very necessary. In this paper Feature selection is done using class association rule incorporated along with information gain. Our proposed method filters out the irrelevant attributes to target class by rule generation according to class based association rule [20] and then select features with relevant support and confidence value. Redundancy from the selected features is removed using information gain algorithm.

IV. PROPOSED METHODOLOGY

In this work, we propose a filter approach to select relevant features based on the associations among them. Such associations can be discovered using class based association rule mining. The proposed associative feature selection approach is based on the heuristics discussed above to separate relevant and irrelevant terms. The occurrence of terms in many association rules means that they are associated with many other terms. These terms should then be assigned with a high score so that they are considered as relevant terms.

On the contrary, terms that occur infrequently in association rules should be assigned with a low score of irrelevant terms. The assignment of scores to features comprises the following three steps:

(1) We first determine the constraints of the association rules;
(2) We search for association rules satisfying the constraints;
(3) Further reduce feature by setting threshold and removing redundant features using information gain by ranking criteria.

In the first step, we determine the constraint F for the association rules: F: ARs → Boolean. Here, ARs is the set of all association rules on the set of terms. Conventionally, the constraint for association rules is that the values of support and confidence should be greater than the minimal values (thresholds) of min_supp and min_conf. The constraint F can also include other measures for mining association rules.

The second step searches for association rules that satisfy the constraint F. The typical approach for searching association rules is the Apriori algorithm[16]. In this approach, the support measure is first used to filter out most of the rules that do not satisfy min_supp. The confidence measure (or other measures) is then applied to the remaining rules that satisfy min_conf.

Finally, the threshold is set for selected features by retrieving the percentage of selected features from the set of original features. Redundancy is also removed from those selected features having similar ranking.

Here pruning is done at two stages, One is to prune low score feature based on the confidence and support value. The other is to prune redundant attributes which are present within the threshold value.

PROPOSED ALGORITHM

Input:
S (a_1, a_2, a_3……a_N, C) // a training dataset
M instance, S classes
Min sup, min conf // minimum support and confidence
Output:
S_{best} //selected features

Find Relevant Features:

1 begin
2 c_k : candidate item set of size k;
3 L_k : frequent item set of size k;
4 L_1 : frequent item set1;
5 C_1 : class label;
6 L_2 : combine L_1 and C_1 and check with minsup and conf;
7 for k=2 to L_k! =null do begin
8 L_{k+1} : combine L_k and L_k;
9 c_{k+1} : candidates generated from $L_{k+1;}$
10 for each transaction t in database d do begin
11 count++; //all candidates in c_{k+1}
12 L_{k+1} : candidates in c_{K+1} with minsup and minconf; also in t
13 end;
14 end;
15 Sl_{ist} : selected feature from rule
16 Selected percentage = (selected feature/total feature) * 100;

Remove Redundant Features:

17 find gain value for S_{list} features using info gain algorithm;
18 $\text{Info}(D) = -\sum_{i=1}^{m} P_i \log_2(P_i)$ //m-no of distinct class
19 for each feature value
20 $\text{Info}_A(D) = \sum_{j=1}^{v} (|D_j| / |D|) * \text{Info}(D_j)$ //v-no of distinct value
21 $\text{Gain}(A) = \text{Info}(A) - \text{Info}_A(D)$
22 end
23 if two or more have same gain value
24 remove all redundant features leave one feature
25 end
26 No of feature select = (selected percentage/100)* selected feature
27 S_{best} : Final Selected feature ;
Where, Final Selected feature =No of feature select - feature selected from order of Info gain value

Let dataset have N features, M instances, S target classes. Towards the end of the proposed algorithm, association rules are generated, from which we select the corresponding feature. To remove redundancy from relevant features we go for information gain. Using information gain algorithm, the gain value for each attribute with class label is calculated. If two or more features have same information gain value, that is called as redundancy. Remove those redundant features. S_{best} is the final selected based on percentage calculation.

The above proposed algorithm is based on the following two concepts namely the class based and correlation based

association rule mining.

A. Class Based Association

Association rule mining has gained a great deal of attention. Formally, an association rule R is an implication X ⇨ Y, where X and Y are sets of items in a given dataset. The confidence of the rule confidence(R) is the percentage of records that contain Y among the total number of records containing X. The support of the rule 'support(R)' is the percentage of records containing X and Y with respect to the number of all records.

Let P[S] be the probability of an itemset S present in a certain transaction of the database. P[S] can be considered as the support of the itemset S as defined above,(also denoted as support(S)).The antecedence support is denoted by P[X] and the consequence support by P[Y] of the rule R. Assume that the database contains N records with the numbers of records that contain X, Y, and both X and Y are a, b, and c respectively. It can be implied from the definitions of support and confidence of association rules that support(R) = c/N and confidence(R) = c/a; and from the definition of support of an itemset that P[X] = a/N, P[Y] = b/N and P[X ∧ Y] = c/N. Thus, the values of supp(R) and conf(R) can be computed using P[X ∧ Y] and P[Y] as follows:

$$Support(R) = P[X \wedge Y] \quad (1)$$

$$Confidence(R) = \frac{P[X \wedge Y]}{P[X]} \quad (2)$$

The confidence of a rule R measures the implication relation of the antecedence (X) to the consequence (Y), which is the actual interestingness to the rule. It shows the prediction of Y when X occurs.

Association rules are proposed for resolving market basket problems on transactional data. However, when all rule targets are constrained by the class labels, association rules become class (or constraint) association rules and they can be used for classification purpose.

B. Information Gain

In our solution we adopt an approach based on the information-theoretical concept of entropy, a measure of the uncertainty of a random variable. . The entropy of a variable *X* is defined as

$$H(X) = \sum_i P(x_i) \log_2 (P(x_i)) \quad (3)$$

and the entropy of X after observing values of another variable Y is defined as

$$H(X/Y) = \sum_j P(y_j) \sum_i P(x_i/y_j) \log_2 (P(x_i/y_j)) \quad (4)$$

where P(xi) is the prior probability for all values of X, and P(xi/yj) is the posterior probability of X given the values of Y. The amount by which the entropy of *X* decreases reflects additional information about X provided by Y and is called Information gain given by

$$IG(X/Y) = H(X) - H(X/Y) \quad (5)$$

According to this measure, a feature Y is regarded to be more correlated to feature X than to feature Z, if IG(X/Y) > IG(Z/Y). Symmetry is a desired property for a measure of correlations between features. However, information gain is biased in favor of features with more values. Furthermore, the values have to be normalized to ensure that they are comparable and have the same aspect. If two or more features have the same value of information gain, they are removed from the dataset to reduce redundancy.

V. Implementation Of Proposed Algorithm

The proposed algorithm is implemented using Java net beans. The Oracle ODBC driver is first created, and then the 'Oracle ODBC driver' is selected in the 'Create new Data Source' window. The properties 'Data Source name', 'Description' and 'User ID' are given in the 'ODBC Setup window'. The user interface (GUI) is designed to be simple and user friendly. The user interface consists of three buttons. Button1 is 'Read Input from Arff Format'. On clicking Button1, a file chooser is opened to select the ARFF file and it converts the selected input arff file into oracle. Button2 is 'Features based on Class association. On clicking Button2 new window opens to get support and confidence values, a dataset undergoes attribute reduction using class association rule mining method. Class association rule mining will select the relevant features with minimum support and confidence values. Button3 is 'Features based on Info gain'. On clicking Button3, features selected from button2 is further undergoes attribute reduction using information gain. Information gain is used to reduce redundant attribute by calculating gain value. On clicking Button4 i.e. 'display features', features selected from Button2 will be displayed. Button5 is to drop tables. Links are provided appropriately for each button, so that each option opens in a separate window so as to make the user understand and feel free to access it.

VI. Performance Evaluation

In order to evaluate the efficiency and performance of our proposed methodology, we conduct tests on various datasets. Table 1 gives the list of datasets chosen for evaluation along with the number of features and instances present in each domain. These domains are then subject to various feature selection methods that already exist. The results obtained from existing methods and our proposed methodology has been tabulated in Table 2. The domains with the reduced number of features are then evaluated with classifiers like J48 and Naïve bayes. Their performance measures are shown in Table 3 and Table 4 respectively. It seen from the tables that our proposed method selects lesser number of features when compared to other methods. The time taken for classification reduces, consequently increasing the efficiency of the classification algorithms. Table 5 gives the performance of the entire set of features. Figure 1 depicts the improvement in accuracy when NB is applied to a set of selected features from a feature selection algorithm. The improvement in accuracy when

Information Gain, association rule mining and NB are applied sequentially to a set of selected features is shown in Figure2. A comparative analysis of accuracies of applying C4.5 and NB after information gain and ARM, over a set of selected features is shown in Figure 3.From the above observations, we show the following results.

1.Reduced (or equal) number of selected features when compared to other methods in datasets such as vote, spectheart, shuttle, monk, Hayes Roth, car evaluation, contact lenses, postoperative, weather, parity (refer Table 2) .

2. Improved performance (or equal) when compared to the performance of other feature selection methods,

A.With j48 classifier in datasets such as vote, spectheart, shuttle, monk, contact lenses, postoperative (refer Table 3).

B. With naïve bayes classifier in datasets such as vote, spectheart, shuttle, parity, weather, nursery, monk, contact lenses, postoperative. (refer Table 4)

3. Improved performance (or equal) when compared to the performance of with total features, (refer Table 5 & 7)

A. With naïve bayes classifier in datasets such as vote, shuttle, parity, weather, monk, contact lenses, tic-tac, car evaluation.

B. With j48 classifier in datasets such as shuttle, contact lenses, postoperative.

C. With ID3 classifier in datasets such as spectheart, shuttle, parity, monk, contact lenses.

VII. Conclusion And Future Work

The objective of this paper is to describe a novel approach for attribute reduction. In data-mining, most of the classification algorithms are less accurate due to the presence of some relevant and redundant attributes in the dataset. A new Attribute Reduction algorithm using class based association rule mining has been incorporated along with information gain and evaluated through extensive experiments. We used Apriori algorithm to identify the relevant attributes and remove irrelevant attributes from the dataset. We then remove redundant features using information gain. Our approach demonstrates its efficiency and effectiveness in dealing with high dimensional data for classification. Our future work, involves the study to combine this approach with feature discretization algorithms to smoothly handle data of different feature types.

References

[1] De Sousa, E.P.M., Traina, C., Traina, A.J.M., Wu, L.J., Faloutsos, C.: A fast and effective method to find correlations among attributes in databases. Data Mining and Knowledge Discovery 14 (2007) 367-407

[2] Molina, L.C., Belanche, L., Nebot, and A.: Attribute Selection Algorithms: A survey and experimental evaluation. Proceedings of 2nd IEEE's KDD 2002 (2002) 306-313

[3] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. Journal of Machine Learning Research 3 (2003) 1157-1182

[4] Applied Intelligence 9 (1998) 217-230 11. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97 (1997) 273-324

[5] Bhavani, S.D., Rani, T.S., Bapi, R.S.: Feature selection using correlation fractal dimension: Issues and applications in binary classification problems. Applied Soft Computing 8 (2008) 555-563

[6] Haindl, M., Somol, P., Ververidis, D., Kotropoulos, C.: Feature selection based on mutual correlation. Progress in Pattern Recognition, Image Analysis and Applications, Proceedings 4225 (2006) 569-577

[7] Liu, H., Motoda, H., Yu, L.: A selective sampling approach to active feature selection. Artificial Intelligence 159 (2004) 49-74

[8] Liu, H., Yu, L., Dash, M., Motoda, H.: Active feature selection using classes. Advances in Knowledge Discovery and Data Mining 2637 (2003) 474-485

[9] Yu, L., Liu, and H.: Feature Selection for High-Dimensional Data: A Fast Correlation-based Filter Solution. Proc. Int.Conference ICML2003 2003 (2003) 856-86310. Liu, H.A., Setiono, R.: Incremental feature selection

[10] Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis: An International Journal 1(1997) 131-156

[11] Koller, D., Sahami, and M.: Toward Optimal Feature Selection. Proc. Int.Conference ICML'96 (1996)170-178

[12] Bakus, J., Kamel, and M.S.: Higher order feature selection for text classification. Knowledge and Information Systems 9 (2006) 468-491

[13] Liu, H., Yu, and L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering 17 (2005) 491-502

[14] Peng, H.C., Long, F.H., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1226-1238

[15] Liu, H.A., Setiono, R.: Incremental feature selection.

[16] Liu, W.Z., White, A.P.: The Importance of Attribute Selection Measures in Decision Tree Induction. Machine Learning 15 (1994) 25-41

[17] H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kauffmann, San Francisco, 2nd Edition,2005.

[18] Harol, A., Lai, C., Pezkalska, E., Duin, and R.P.W.: Pairwise feature evaluation for constructing reduced representations. Pattern Analysis and Applications 10(2007) 55-68

[19] Van Hulse, J.D., Khoshgoftaar, T.M., Huang, H.Y.: The pair wise attribute noise detection algorithm. Knowledge and Information Systems 11 (2007)171-190

[20] Tin Dung Do, Siu Cheung Hui and Alvis C.M. Fong.: Associative Feature Selection for Text Mining. International Journal of Information Technology, Vol. 12 No.4 (2006).

APPENDIX

Table I
Dataset Description:

Dataset	Instances	Features	Classes
Vote	435	17	2
Spectheart	267	23	2
shuttle landing	15	7	2
Parity	100	11	2
Monk	124	6	2
Nursery	12960	9	3
Postoperative	57	9	3
Hayes Roth	132	6	3
tic-tac	958	10	2
Carevalution	1728	7	4
Contact lenses	24	5	3
Weather	14	5	2
Soybean(Small)	35	47	4

Table II
Number of selected features for each feature selection algorithm

Dataset	Cfs	Chi square	Gain	Info gain	Oneratt r	Symmetric	Relief	Association	Class Association & Infogain
Vote	4	6	7	7	6	7	8	3	2
Spectheart	12	8	10	9	16	9	8	11	6
Shuttle	2	6	3	5	4	4	4	6	1
Parity	3	6	6	6	6	6	6	5	3
Monk	2	2	2	2	2	2	2	4	2
Nursery	1	1	1	1	5	5	3	5	4
Postoperative	5	5	5	5	7	7	5	6	4
Hayes Roth	1	3	3	3	3	3	3	3	1
tic-tac	5	1	1	1	1	1	5	5	4
car evaluation	1	6	6	6	6	6	5	3	2
contact lenses	1	2	2	2	3	1	3	3	2
Weather	2	2	2	2	3	2	2	3	2
Soybean(Small)	22	35	35	35	35	35	35	40	28

Table III
Accuracy of C4.5 on selected features for each feature selection algorithm

Dataset	Cfs	Chi-square	Gain	Info gain	Oneattr	Symmetric	Relief	Association	Class Association& Info gain
Vote	96.092	95.6322	91.7241	91.7241	95.6322	91.7241	96.092	95.6322	95.6322
Spectheart	81.6479	79.4007	75.6554	75.6554	79.0262	757.655	79.4007	79.4007	79.4007
Shuttle	53.3333	53.3333	60	53.3333	60	53.333	53.3333	53.333	53.3333
Parity	44	40	40	40	50	40	48	50	46
Monk	91.9355	91.9355	91.9355	91.9355	91.9355	91.935	91.9355	90.322	91.9355
Nursery	70.9722	70.9722	70.9722	70.9722	90.7407	70.972	89.213	78.603	78.248
Postoperative	71.1864	71.1864	71.1864	71.1864	71.1864	71.186	71.1864	71.929	71.9298
Hayes Roth	37.8788	50	50	50	70.4545	70.454	70.4545	70.454	54.5455
Tic-tac	79.4363	69.9374	69.9374	69.9374	69.9374	69.937	79.4363	79.436	76.3048
car evaluation	70.0231	92.3611	92.3611	92.3611	92.3611	92.3611	93.2292	80.324	76.5625
contact lenses	70.8333	87.5	87.5	87.5	58.3333	70.833	83.3333	83.333	87.5
Weather	42.8571	42.8571	42.8571	42.8571	50	42.857	42.8571	64.285	42.8571
Soybean(Small)	96.31	98.15	98.15	98.15	98.15	98.15	98.15	91.33	98.23

Table IV
Accuracy of NB on selected features for each feature selection algorithm

Dataset	Cfs	Chi square	Gain	Info gain	Onerattr	Symmetric	Relief	Association	Class Association & info gain
Vote	96.092	91.0345	91.7241	91.7241	91.0345	91.7241	93.5633	92.8736	95.6322
Spectheart	82.0225	76.779	80.1498	79.0262	79.0262	79.0262	76.779	77.1536	77.9026
Shuttle	80	80	80	73.3333	73.3333	73.3333	73.3333	80	80
Parity	50	46	46	46	47	46	43	47	51
Monk	100	100	100	100	100	100	100	99.1935	100
Nursery	70.9722	70.9722	70.9722	70.9722	88.8426	70.9722	87.6543	77.2531	77.3848
Postoperative	72.8814	72.8814	72.8814	72.8814	67.966	72.8814	76.2712	71.9298	75.4386
Hayes Roth	37.8788	59.8485	59.8485	59.8485	81.8182	59.8485	81.8182	81.8182	54.5455
tic-tac	72.4426	69.9374	69.9374	69.9374	69.9374	69.9374	72.4426	72.4426	70.5637
Car evaluation	70.0231	85.5324	85.5324	85.5324	85.5324	85.5324	85.3588	79.5718	76.8519
Contact lenses	70.8333	87.5	87.5	87.5	54.1667	70.8333	83.3333	83.3333	87.5
Weather	78.5714	78.5714	78.5714	78.5714	71.4286	78.5714	78.5714	50	78.5714
Soybean(Small)	94.47	92.94	92.94	92.94	92.94	92.94	92.94	87.34	92.94

Figure 1: Accuracy of NB on selected features for each feature selection algorithm

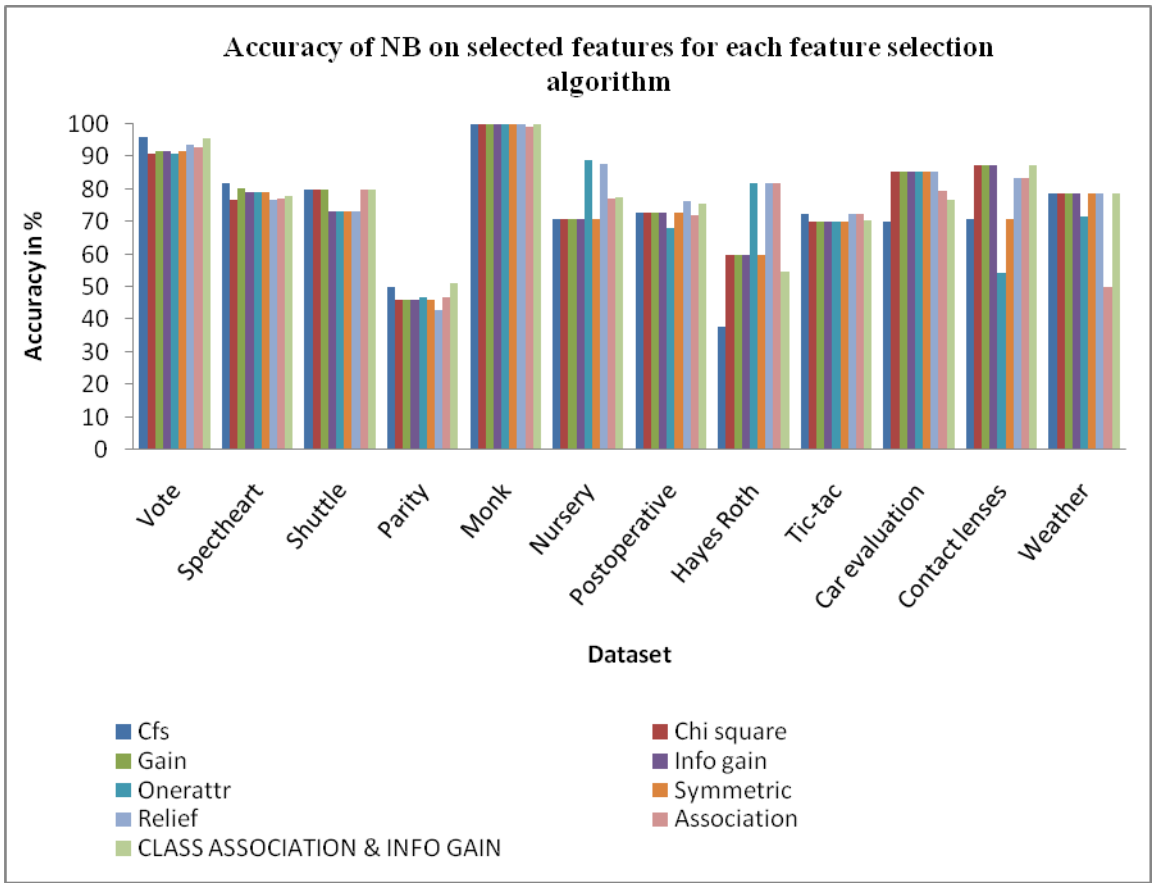

Table V
Accuracy of ID3, C4.5 and NB on full set of features

DATASET	TOTAL FEATURES	ID3 (%)	C4.5 (%)	NAIVE BAYES (%)
Vote	16	-	96.3218	90.1149
Spectheart	22	70.0375	80.8989	79.0262
Shuttle landing	6	60	53.3333	80
Parity	10	45	48	40
Monk	5	95.9677	90.3226	99.1935
nursery	8	98.1867	97.0525	90.3241
postoperative	8	69.4915	71.1864	-
Hayes Roth	5	0	72.7273	80.303
tic-tac	9	83.4029	85.0731	69.6242
Carevalution	6	89.3519	92.3611	85.5324
Contact lenses	4	70.8333	83.3333	70.8333
Weather	4	85.7143	50	57.1429
Soybean(Small)	35	91.50	92.09	92.97

Figure 2: Improvement in performance of Naive Bayes Classifier

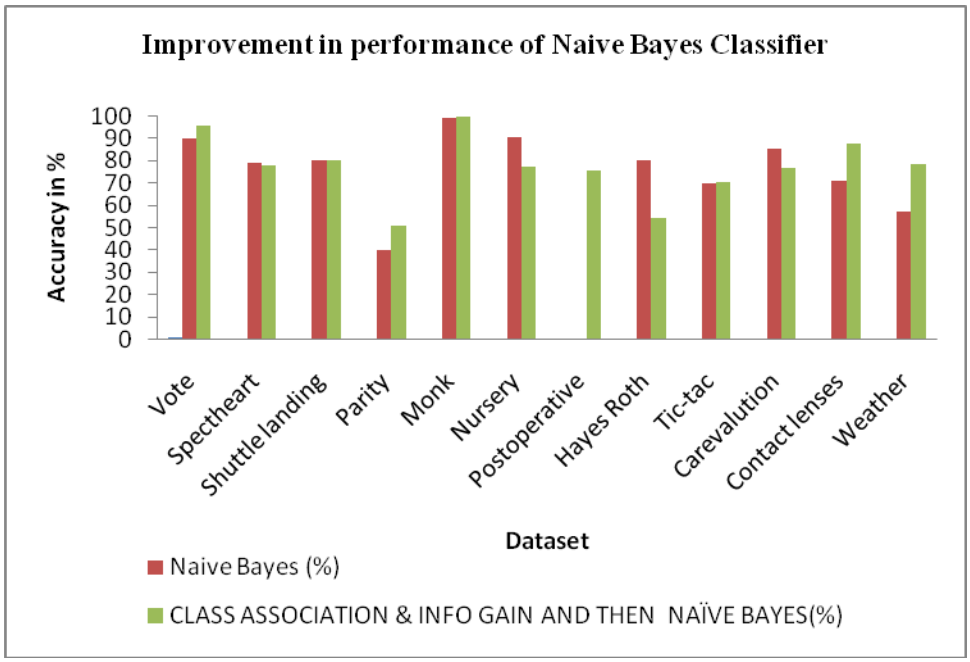

Figure 3: Improvement in performance of C4.5 and Naive Bayes Classifier