academicJournals

Vol. 8(18), pp. 885-896, 16 May, 2013
DOI: 10.5897/IJPS2013.3842
ISSN 1992-1950 © 2013 Academic Journals
http://www.academicjournals.org/IJPS

International Journal of Physical Sciences

Full Length Research Paper

Feature subset selection using association rule mining and JRip classifier

Waseem Shahzad, Salman Asad and Muhammad Asif Khan

National University of Computer and Emerging Sciences Department of Computer Science, Sector H-11/4, Islamabad, Pakistan.

Accepted 29 March, 2013

Feature selection is an important task in many fields such as statistics and machine learning. It aims at preprocessing step that include removal of irrelevant and redundant features and the retention of useful features. Selecting the relevant features increases the accuracy and decreases the computational cost. Feature selection also helps to understand the relevant data, addressing the complexity of dimensionality. In this paper, we have proposed a technique that uses JRip classifier and association rule mining to select the most relevant features from a data set. JRip extracts the rules from a data set and then association rules mining technique is applied to rank the features. Twenty datasets are tested ranging from binary class problem to multi-class problem. Extensive experimentation is carried out and the proposed technique is evaluated against the performance of various familiar classifiers. Experimental results demonstrate that while employing less number of features the proposed method achieves higher classification accuracy as well as generates less number of rules.

Key words: Feature subset selection, association rules mining, JRip, J48, Ridor, PART.

INTRODUCTION

The speedy growth of data on daily basis have resulted the size of databases to Terabytes, where a lot of useful information is invisible. Discovering such hidden information is called data mining. Data mining can also be defined as a process that is used to analyze data to find the hidden patterns and relationships among data. It helps in making correct prediction. Techniques used in data mining are the combination of multiple fields like databases, machine learning, computational intelligence, statistics, pattern recognition and neural networks (Edelstein, 1999).

Data mining is used in businesses where large transactions are involved. For example, it can detect the characteristics of customer who buy products from store, which help to find what trends are being followed. It can also help in fraud detection if customer's transaction behavior is abnormal. Such predictions are based on historical data. Data mining techniques can be classified as supervised and unsupervised. Supervised technique takes training set to build a model and learn it. After a predictive model is built, test data is produced to find out the accuracy of the predictive model. In contrast, an unsupervised method has no target class. The unsupervised technique searches for patterns and similarity among the variables on the basis of which they are grouped. Clustering is one of the most common unsupervised data mining techniques.

Classification is one of the important techniques of supervised data mining. Its objective is to build a model by using the training set that can predict the class labels of test set. Applications of classification are fraud detection, weather prediction, customer behavior, risk management (Han and Kamber, 2006). Building classification model, starts with training set or training

*Corresponding author. E-mail: waseem.shahzad@nu.edu.pk, salman.asad87@gmail.com, masif.khan@nu.edu.pk.

data. The next step is to find the relationships and hidden patterns in that training data. Different classification algorithm uses different techniques to find the relationships and hidden patterns in training data (Liu et al., 2002; Martens et al., 2007; Parpinelli et al., 2002). These relationships and hidden patterns are then used to build a model. The model can then be used to predict the target value of unknown cases.

The existence of strong and hidden relationship in large databases is the major concern of most of the researchers today. Data mining resolve this issue by providing a technique called association rule mining (ARM). The rules created by association rule mining is in the form on IF – THEN statement. IF part is called the antecedent part of the rule while THEN part is called the consequent of the rule. Support and confidence are the two important factors, associated with the ARM. Collection of important information is also carried out with the help of a feature selection mechanism by choosing a subset of features also known as feature extraction, feature reduction, variable selection or attributes selection. The technique selects the set of most significant features that improve performance and increase the accuracy. Feature selection algorithms can be classified into two categories – feature ranking and feature subset selection. In feature ranking, features are ranked by a metric; while the latter one searches for the optimal subset of features from possible subsets (Wei and Billings, 2007; Yahang and Honavar, 1997; Zhou et al., 2007).

In this paper, for feature selection, we proposed a feature subset selection technique that uses JRipper classifier with association rules mining to select most valuable features of a dataset. First, we build rules using JRipper classifier, then assign ranks to features by using these rules based on support and confidence. Extensive experimentation on various datasets has been performed using the proposed approach and the results greatly showed the worth of proposed technique over other techniques.

RELATED WORK

Feature selection is an important stage of preprocessing. In literature it has been proven that feature selection is an effective mechanism in reducing the dimension to improve the computational efficiency (Kanan et al., 2007; Raymer et al., 2000). Feature subset selection technique can be classified into three categories – filtered based techniques, wrappers, and embedded technique. Filtered approach first searches the data and then use a filtered approach. In wrapper based methods any classification algorithm is used that can go through all data and find a set of relevant features. Wrappers based methods have high computation cost. Embedded technique is embedded in the model and it selects features by making a model.

Approaches used for feature subset selection have number of drawbacks. To avoid such drawbacks many hybrid approaches have been introduced like (Shahzad and Baig, 2010). Yang et al. (2005) concluded that there is no any specific criteria which can be used to pick any filtered algorithm. They introduced a hybrid algorithm by merging different filtered based algorithms and produced better results. The methodology lacks some helpful classification facts and information. It did not explain what difficulties would occur when each gene is treated distinctly. Wang et al. (2005) enhanced the classification process by merging hierarchical clustering, different filters and uniform gene selection classification method. A special filter algorithm is used in their approach for gene ranking and clustered hierarchically the top 50 to 100 genes. Hassan et al. (2009) and Su et al. (2007) suggested the multiple classifier concept for efficient system and better results. A GA-based classifier prototype proposed by Zhang et al. (2009) namely “Genetic Ensemble” uses GA in finding nonlinear association from genes and also in assessing gene sets by groups.

Classification is termed as grouping of similar objects. A number of classification algorithms exist such as decision trees, k-nearest neighbor classifiers, neural networks and support vector machines. Among many classification applications, few are fraud detection, weather prediction, customer behavior, risk management (Baesens et al., 2003; Han and Kamber, 2006). A Naïve Bayes classifier is a simple classifier based on probability (Friedman et al., 1997). It uses a formula of Bayes theorem to calculate probability by counting the frequency of values in the historical order.

Decision tree is a well known classification algorithm which uses conditional probability formula. Decision tree generate rules that can be understand and read easily by human. Decision trees can work for both binary and multi-class classification problems and can also contain categorical and numeric information (Edelstein, 1999). Decision tree finds strong relationships between the values of the data. When a set of values is found having strong relationship than others then that set is grouped and it becomes a branch. The algorithm is recursive and repeats to create more branches. After constructing a rule, instances covered by that rule are removed, so that the remaining dataset is considered for other rules to be constructed.

Ripper is also a well-known algorithm for supervised data classification. Its rule set is easy to understand and usually better than decision tree learners (Cohen, 1995). In ripper classifiers training data is randomly distributed into growing set and pruning set. Each rule keeps on growing until no information gain is possible further.

PROPOSED TECHNIQUE

The proposed technique uses JRipper algorithm with ARM to select

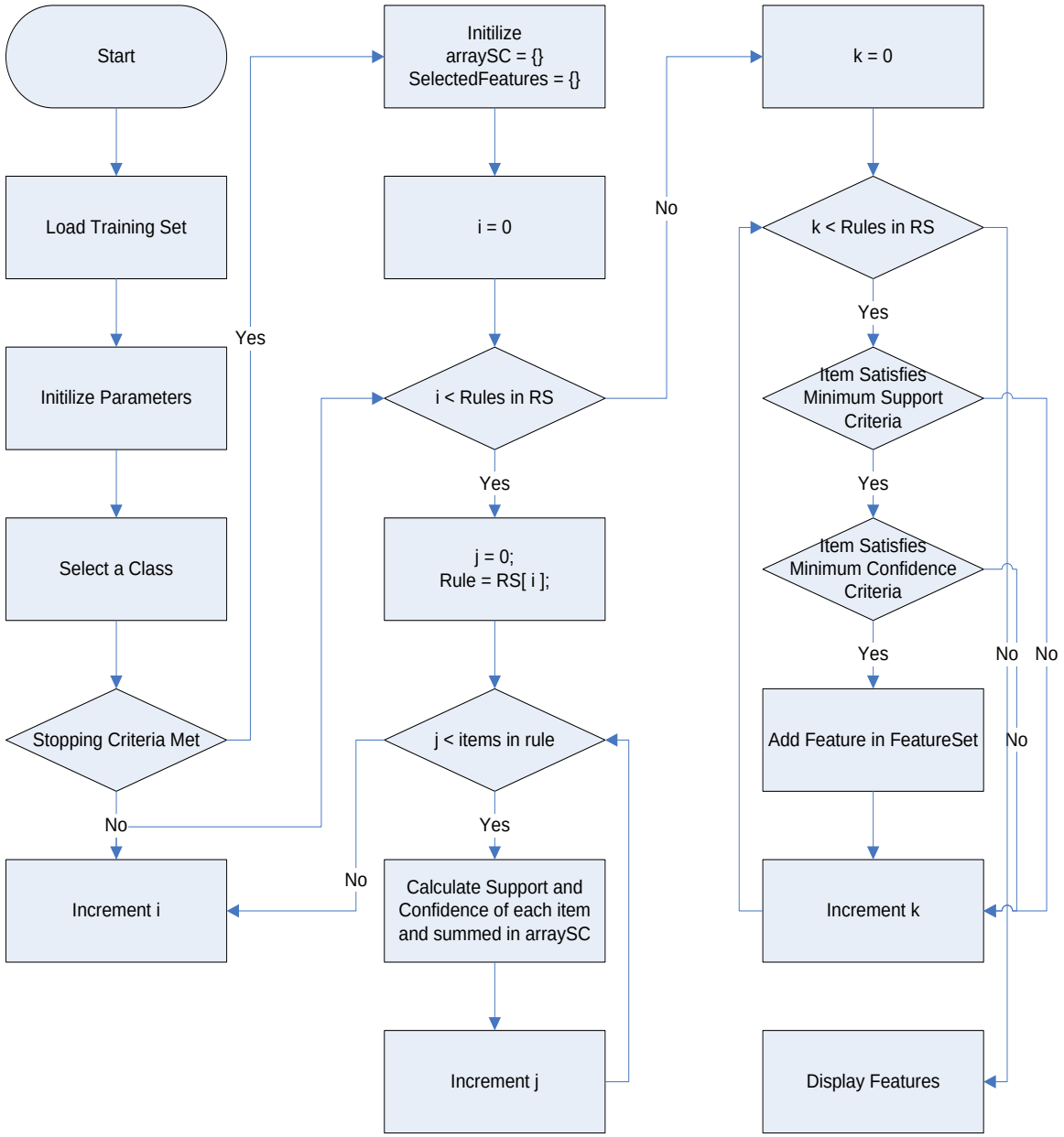

Figure 1. Flowchart of proposed approach.

features of a dataset. Ripper is one of the supervised classification approach where training data is distributed randomly into two sets-growing set and the pruning set. Each rule keeps on growing until no information gain is possible further. After this process, the rule is then passed through a pruning step where unnecessary terms are eliminated in order to maximize the following function given in Equation (1).

$$\frac{p + (N - n)}{P + N} \quad (1)$$

Where p is the number of positive examples covered by rule, N is the number of negative examples in prune set, n is the number of negative examples covered by rule, and P is the number of positive examples in Prune Set. Figure 1 shows the proposed flowchart of the proposed approach followed by the algorithm.

JRip algorithm

JRip is an optimized version of IREP (Cohen, 1995). It was introduced by William W. Cohen. With the repeated incremental pruning JRip produce error reduction.

Initialization

Initialize RS = {}, and from each class from the less frequent one to the most frequent one.

Building stage

Repeat the grow phase and prune phase until there are no positive examples or error rate >= 50%.

(i) Grow phase: Keep on greedily adding terms to the rule until the rule is perfect (100% accurate)

(ii) Prune phase: Each rule is incrementally pruned. Now any finishing sequences can get pruned. The pruning value can be measured using the formula 2p / (p + n) – 1 (Figure 1).

Algorithm: An algorithm for the proposed approach

```
load (TrainingSet);                /*rule set will be initialized as empty*/
Initialize RS = {};
minSupport, minConfidence;         /*this will be set by user*/
for each (class in TrainingSet)
   begin
      while (!positive_examples () && error_rate >= 50)
         begin
            Rule = buildRule();
            Rule = pruneRule();
            addRule(RS, Rule);
         end
   end
      optimizationRules();
      deleteRules(); /*After this stage we have the rules created by the JRip in ruleset*/
      arraySC[][] = {}; /*stores the support and confidence of each item*/
      SelectedFeatureList = {}; /*stores the array of selected features*/
for (int i=0; i<RS.count(); i++)
   begin
         Rule = RS[i];
         for (int j=0; j<rule.count(); j++)
            begin
               Item = rule[j];
               Item_index = findItemIndex(Item);
               Item_support = findItemSupport(Item);
               Item_confidence = findItemConfidence(Item);
               arraySC[Item_index][0] = Item_support;
               arraySC[Item_index][1] = Item_confidence;
               arraySC[Item_index][2]++;  /*Tell the frequency of item*/
            end
   end
for (int i=0; i<arraySC.rowCount(); i++)
   begin
      if (arraySC[i][0]/arraySC[i][2]>minSupport)
         begin
            if (arraySC[i][1]/arraySC[i][2]>minConfidence)
               begin
                  SelectedFeatureList.Add(i);
               end
         end
   end
```

Optimization stage

In this optimization stage of JRip algorithm first initial rule set is identified and is known as {Ri}. Now create two variants by using procedures of grow phase from each rule Ri using random data and prune these variants. Empty rule must be applied for generating one of the variants and second variant is generated by adding antecedents to the original rule. In this way, the pruning metric used is (TP+TN) / (P+N). Now, for each variant and the original rule, smallest possible DL is calculated. Final representative from the set R is taken on the basis of minimum DL. If there are residual positives after observing {Ri} rules, some more rules are identified using building stage again on the basis of residual positives.

Delete stage

All the rules that increase the DL get deleted, the left set is then added in the resultant set. This resultant set will be denoted as RS.

Association rule mining

Association rule mining (ARM) is one of the important technique of data mining. It is used to find the hidden and interesting patterns in the data. It only works on categorical data.

Rule structure

The rule generated using association rule mining has two parts. First part is called ANTECEDENT or IF part and the second part is the CONSEQUENT or THEN part. Antecedent part contains terms. For example:

$$IF\ term_1\ and\ term_2 \ldots term_n \quad (2)$$

The consequent part contains only the class label. Therefore, the complete rule will look like:

$$IF\ term_1\ and\ term_2 \ldots term_n\ THEN\ class \quad (3)$$

Each term consists of attribute name and value. The structure of a term is given below:

$$AttributeName = AttributeValue \quad (4)$$

For example the term “Outlook = Sunny” has an *Attribute name* “Outlook” and *Attribute value* “Sunny”.

Market basket analysis

Market basket analysis is a common example of association rule mining. It is a modeling technique based upon the theory that if an item from a certain group is bought then it is more or less possible that another group of item is bought. The set of items customer buys is referred as the item set. The market basket analysis finds interesting relationships between items. This is done using association rule mining.

Factors of association rule mining

There are two very important factors of association rule mining. These factors are:

(i) Support: Support is used to tell how frequently a rule occurs in the dataset. It’s actually the number of instances covered by the rule against total number of transactions/records. The formula for

Table 1. Ranking of features arranged on the basis of support and confidence.

Features with support and confidence			Features arranged on the basis of confidence			Features arranged on the basis of support		
Feature index	Confidence	Support	Feature index	Confidence	Support	Feature index	Confidence	Support
1	0.9	0.02	1	0.9	0.02	10	0.9	0.11
2	0.8	0.05	10	0.9	0.11	1	0.9	0.02
3	0.8	0.05	12	0.9	0.006	12	0.9	0.006
4	0.7	0.08	2	0.8	0.05	2	0.8	0.05
5	0.5	0.02	3	0.8	0.05	3	0.8	0.05
6	0.7	0.02	4	0.7	0.08	4	0.7	0.08
7	0.6	0.005	6	0.7	0.02	9	0.7	0.04
8	0.2	0.001	9	0.7	0.04	11	0.7	0.04
9	0.7	0.04	11	0.7	0.04	6	0.7	0.02
10	0.9	0.11	7	0.6	0.005	7	0.6	0.005
11	0.7	0.04	5	0.5	0.02	15	0.5	0.05
12	0.9	0.006	15	0.5	0.05	5	0.5	0.02
13	0.1	0.02	8	0.2	0.001	14	0.2	0.004
14	0.2	0.004	14	0.2	0.004	8	0.2	0.001
15	0.5	0.05	13	0.1	0.02	13	0.1	0.02

support of a rule X => Y is:

$$Support(X => Y) = P(X \cup Y) \quad (5)$$

Suppose a database with 10,000 transactions. Out of those 10,000 transactions 1000 include both items 1 and item 2 and 500 of these include item C. Therefore the support of the rule “If items 1 and 2 then item C” is 500/10000=0.05=5%.

(ii) Confidence

The formula for confidence of a rule X => Y is:

$$Confidence(X => Y) = P(Y|X) \quad (6)$$

Suppose a database with 10,000 transactions. Out of those 10,000 transactions 1000 include both item 1 and item 2 and 500 of these include item C. Therefore the confidence of the rule “If item 1 and item 2 then item C” is 500/1000=0.5=50%.

Feature selection

Feature selection an important technique of data mining. The main task of feature selection is to exclude or remove extra features from the dataset. Feature selection not only removes irrelevant features to reduce computational cost but also increases the accuracy. Feature selection is a very useful technique and has many advantages. The main advantage of feature selection is that as features are removed so learning process will be faster. Secondly, problem will get simpler as irrelevant and redundant features are removed. There is another advantage that accuracy of model will be increased. Feature selection therefore provides us information about the importance of features and this helps us to understand the data.

Feature weightage

After rules are built by JRip they are stored in the rule set. Each rule has antecedent and consequent part. Each antecedent part has at least one term and each term has an attribute in it. The support and confidence of each rule is summed and assigned to that attribute. Now each attribute has its support sum, its confidence sum and the number of times it appears in the rule list. The equation for assigning support to each feature is given below:

$$FS_i = \frac{\left(\sum_{j=0}^{k} x.Support_j\right)}{\left(\sum_{j=0}^{k} x\right)} \quad (7)$$

Where FS_i is the support of i[th] feature, $Support_j$ is the support of j[th] rule, k is the total number of rules and x will be 1 if the rule contains the i[th] feature else it will be 0. The equation for assigning the confidence to each feature is given below:

$$FC_i = \frac{\left(\sum_{j=0}^{k} x.Confidence_j\right)}{\left(\sum_{j=0}^{k} x\right)} \quad (8)$$

Where FC_i is the confidence of i[th] feature, $Confidence_j$ is the confidence of j[th] rule, k is the total number of rules and x will be 1 if the rule contains the i[th] feature else it will be 0.

Ranking of features

All features are sorted on the basis of support and confidence values. First the features are sorted with respect to confidence. After this the feature are again sorted on the basis of confidence. But this will only affect those features whose confidence value is same. For example, suppose we have 15 features in a dataset along with their support and confidence as given in Table 1. There are some features with the same confidence value and same support value.

It is important to note that features with same confidence are now grouped, but still their support value is not sorted. As a second step these values are sorted on the basis of their support. Hence, the

Table 2. Features and there ranking.

Feature index	Confidence	Support	Rank
10	0.9	0.11	1
1	0.9	0.02	2
12	0.9	0.006	3
2	0.8	0.05	4
3	0.8	0.05	5
4	0.7	0.08	6
9	0.7	0.04	7
11	0.7	0.04	8
6	0.7	0.02	9
7	0.6	0.005	10
15	0.5	0.05	11
5	0.5	0.02	12
14	0.2	0.004	13
8	0.2	0.001	14
13	0.1	0.02	15

Table 3. Selected features.

Feature index	Confidence	Support	Rank
10	0.9	0.11	1
1	0.9	0.02	2
2	0.8	0.05	3
3	0.8	0.05	4
4	0.7	0.08	5
9	0.7	0.04	6
11	0.7	0.04	7
6	0.7	0.02	8
15	0.5	0.05	9
5	0.5	0.02	10

features are arranged on the basis of both support and confidence as shown in Table 1.

Feature ranking

In feature ranking a metric (rank no.) is assigned to each feature based on their sorted position (as in Table 1). Feature ranking for all features is shown in Table 2.

Selection of features

Features that meet the minimum threshold are selected and the features that do not meet the minimum threshold criteria are removed. The parameters used are Support = 0.01 and Confidence = 0.5. The selected features using the above parameters are shown in Table 3.

RESULTS

Experimentation results are discussed here. The algorithm is tested on 20 different datasets. The datasets are taken from machine learning repository. Detail of all datasets used is given in Table 4.

Dataset are first discredited and also missing values are removed from it before performing experimentation. The comparisons are done using four different classifiers named JRip (Cohen, 1995), J48 (Quinlan, 1993), PART (Frank and Witten, 1998) and Ridor (Gaines and Compton, 1995). Table 5 also shows the features selected by proposed technique. In most cases the 60 to 70% features are removed and in some cases even 93% features are removed. But it is important that important features not get removed else there will be a fall in accuracy level. Therefore, removing a certain number of

Table 4. Datasets used in experimentation.

S/N	Dataset	Attributes	Instances	Classes	Selected features
1	Audiology	69	226	24	12
2	Autos	25	205	6	13
3	Breast Cancer	9	286	2	4
4	Colic	22	368	2	5
5	Credit –A	15	690	2	1
6	Credit –G	20	1000	2	7
7	Diabetes	8	768	2	3
8	Glass	9	214	6	7
9	Heart –C	13	303	2	6
10	Heart Statlog	13	270	2	7
11	Hepatitis	19	155	2	7
12	Ionosphere	34	351	2	11
13	Iris	4	150	3	2
14	Labor	16	57	2	3
15	Primary Tumor	17	399	21	10
16	Segment	19	2310	7	12
17	Sonar	60	208	2	16
18	Splice	60	3190	3	11
19	Vehicle	18	846	4	14
20	Waveform	40	5000	3	20

Table 5. Comparison between accuracies using different classifiers.

S/N	Dataset	JRip		PART		J48		Ridor	
		Accuracy before feature selection (%)	Accuracy after feature subset (%)	Accuracy before feature selection (%)	Accuracy after feature subset (%)	Accuracy before feature selection (%)	Accuracy after feature subset (%)	Accuracy before feature selection (%)	Accuracy after feature subset (%)
1	Audiology	70.79	73.45	79.64	76.10	78.31	77.43	74.33	76.10
2	Autos	65.36	66.34	69.75	69.26	75.12	78.53	64.39	63.90
3	Breast Cancer	70.27	74.47	69.58	74.12	75.52	75.87	72.02	74.82
4	Colic	86.41	87.50	81.52	84.51	85.05	85.59	82.60	86.14
5	Credit –A	85.21	85.50	84.92	85.50	85.79	85.50	85.36	85.50

Table 5. Contd.

6	Credit Germany	69.60	71.30	72.20	73.10	71.30	72.90	71.10	71.20
7	Diabetes	74.21	75.00	69.53	72.39	74.21	72.65	73.43	74.60
8	Glass	58.41	59.34	60.74	63.08	59.81	61.68	57.00	56.07
9	Heart –C	81.18	83.82	78.87	82.50	77.88	80.85	77.55	80.52
10	Heart Statlog	78.14	82.96	77.40	82.59	77.03	81.85	72.59	81.85
11	Hepatitis	80.64	84.51	81.95	82.58	81.29	81.29	79.35	80.64
12	Ionosphere	86.89	88.88	88.03	88.88	86.32	90.59	86.03	87.74
13	Iris	92.00	91.33	90.00	91.33	90.66	92.00	88.00	92.00
14	Labor	78.94	91.22	85.96	87.71	82.45	94.73	85.96	84.21
15	Primary Tumor	38.93	39.82	38.64	39.23	40.11	36.87	35.39	34.51
16	Segment	91.38	91.68	91.94	92.52	91.94	92.12	91.42	91.25
17	Sonar	73.55	75.96	75.00	77.88	67.78	69.71	69.23	69.71
18	Splice	94.13	95.17	92.50	92.79	94.35	94.51	92.10	94.01
19	Vehicle	59.21	60.16	65.36	65.60	65.48	65.13	64.18	61.34
20	Waveform	75.46	76.72	75.88	76.70	75.16	77.12	72.16	72.24

features will not be a good approach, as we can analyze that in some cases only a few features are removed and in those cases removing more features will cause fall in accuracy.

Figure 2 clearly analyzes the number of features selected. The figure shows the comparison between the original features and the extracted features. The blue bars show the features of original dataset while red bars show the number of features selected by the proposed approach. It can be clearly seen that in most cases 60 to 65% features are removed and in some cases like audiology, colic, credit–A, labor, sonar and splice 85 to 93% features are removed.

Comparison using different classifiers

Using JRip classifier, rules are generated for extracting features. First the accuracy of dataset is calculated with all features and then it is compared with the accuracy achieved by the selected features. The best performance is shown in boldface as given in Table 8. The accuracies of original dataset and the selected feature is calculated using the same classifier that is, JRipper. Even if the accuracy is not much increase but the computational cost has decreased. For example in Credit-A the accuracy is not much increased but the computational cost is very much decreased. The algorithm only picks 1 feature out of 15 features and ignores rest 14 features. 93% of the features are removed in this case and the problem is also very much simpler now. As features are removed so this helps user to understand the dataset more easily. Therefore even if the accuracy is not much increased we can say that two things can be achieved that is, the computational cost is decreased and the other is problem gets simpler and more understandable. If we analyze it closely then we can easily see that result of Labor dataset is coming very good. Only 3 features out of 16 features are selected which means 81% of the features are removed. But not only features are removed also its accuracy is much better now. There is a 12.28% increase in accuracy now.

With PART classifier (Frank and Witten, 1998), same procedure is followed that is, first the accuracy of dataset with all is computed and then compute accuracy of dataset with the features selected by proposed technique. Table 5 shows the comparison of accuracies of original features and selected features by the proposed technique.

The accuracy of original dataset and the selected feature is calculated using the same classifier that is, PART. In comparison with JRip some of the results are given more good results in selected features. For example on Colic dataset accuracy increased is 1.04% on JRip but in PART

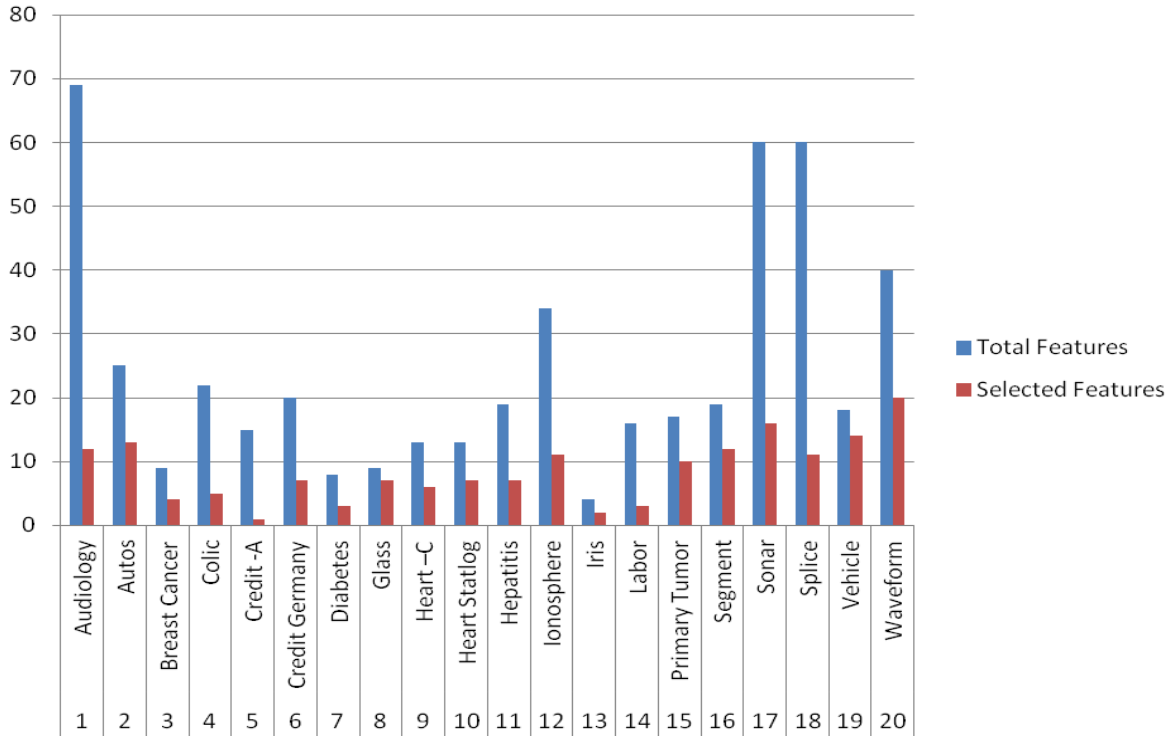

Figure 2. Comparison of feature reduction (total features: left bar, selected feature: right bar).

its 3%. Similarly Heart-C is also providing good results. The results in which accuracies are increased are highlighted in the table. Accuracies on almost all dataset have increased this shows that the features selected are good and it not only gives good result in JRip but also using PART its is giving good results. This shows the robustness of proposed technique.

In J48 classifier, first the accuracy of dataset with full features is calculated and then computes the accuracy of dataset with selected features. Table 5 shows the comparison of accuracies of original features and selected features by the proposed technique. The accuracy of original dataset and the selected feature is calculated using the same classifier that is, J48. The results in which accuracies are increased are highlighted in the table. It seems that the features selected are working fine. As most of the results are giving increases in accuracies. But like JRip and PART, J48 is not giving much good results but still its results are good.

Using Ridor classifier, first the accuracy of dataset with full features is calculated and then accuracy of the dataset with selected features. Table 5 shows the comparison of accuracies of original features and selected features by the proposed technique. The accuracy of original dataset and the selected feature is calculated using the same classifier that is, Ridor. The results in which accuracies are increased are highlighted as bold. After analyzing the results obtained by Ridor, reflects that features selected by the approach are good. Even if the accuracy is not increase in some cases but in those cases computational cost is decreased and not only computational cost is decreased but also the problem becomes more simple to understand and analyze.

Comparison of proposed approach with CFS using different classifiers

Correlation based feature subset selection (CFS) is a well-known technique that uses correlation based heuristic for feature selection (Hall and Smith, 1999). This algorithm is simple and fast. The algorithm maintains the correlation matrix of target class and another matrix in which every feature's correlation is maintained with every other feature. Using best first search it searches for the appropriate features. Best first is only used because it was giving better results than others were but in some cases hill climbing approach was giving better results.

Table 6 shows the selected features using the correlation based feature subset selection. The indices of selected features using proposed technique are also given. In Table 7, number of selected features of proposed technique and CFS is compared. In the third column number of selected features using CFS is shown and last column contains the number of selected features

Table 6. Comparison of number of features selected using CFS and proposed approach.

S/N	Dataset	CFS	Proposed approach
1	Audiology	15	12
2	Autos	7	13
3	Breast Cancer	5	4
4	Colic	5	5
5	Credit -A	4	1
6	Credit Germany	4	7
7	Diabetes	4	3
8	Glass	6	7
9	Heart –C	8	6
10	Heart Statlog	9	7
11	Hepatitis	9	7
12	Ionosphere	5	11
13	Iris	2	2
14	Labor	7	3
15	Primary Tumor	12	10
16	Segment	9	12
17	Sonar	12	16
18	Splice	22	11
19	Vehicle	8	14
20	Waveform	15	20

using the proposed algorithm. In most cases, the proposed algorithm selects fewer features than CFS. While in some cases it selects more, however, in such cases the accuracy of the proposed approach is far much better than CFS.

Table 7 shows the accuracy of selected features by CFS and proposed technique that clearly demonstrate that the proposed approach is better than CFS in terms of accuracy. The important point to note that despite in cases where more features are selected by the proposed approach, their accuracy still beats CFS. This means the proposed feature selection technique is much better than CFS. For example CFS selects seven features in Autos and the proposed approach selects thirteen features, but the accuracy is better than CFS. Same is the case with Credit Germany, Sonar, Vehicle and Waveform datasets. Only in case of Credit A the accuracy is 0.15% less than achieved by CFS, but in that case just 1 feature is selected and CFS have selected 4 features which reflects that both results seems to be almost equal.

Ionosphere is the only case in which our results are not good. We have selected more features than CFS and our accuracy is also not better than CFS. So Ionosphere is the only case in which CFS is better than the proposed technique. It makes clear that our results are better than CFS. The above results also reveal that there is much difference between accuracies. One of the examples is the big difference between the accuracies in dataset "vehicle". Other than that in most of the cases, the proposed outperforms the CFS. Moreover, Table 7 also shows that the proposed approach produced better results in most of the cases. In three cases - Autos, Vehicle and Labor, the proposed technique has achieved much more accuracies than the CFS. As another example, the proposed approach produced much better results on the datasets heart statlog, vehicle and heart-c as compared to the results produced by the CFS.

Conclusion

Feature selection is an important technique of data mining to exclude or remove insignificant features from the dataset. In other words, it is to choose a subset of significant features for building a learning model. Feature selection not only removes insignificant features to reduce computational cost but it also increases the accuracy. Removal of the most irrelevant and redundant features from the dataset are the main goal of feature selection. Feature selection is very useful technique and has many advantages. The main advantage of feature selection is that as features are removed learning process becomes faster. Secondly, problem will get simpler with the removal of irrelevant and redundant features. Another advantage is that accuracy of the model will increase. Feature selection, therefore, helps researchers to acquire better understanding about their data by telling them which one are the important

Table 7. Accuracies comparison between CFS and proposed approach using different classifiers.

S/N	Dataset	JRip		PART		J48		Ridor	
		Accuracy after feature selection using CFS (%)	Accuracy after feature selection using proposed approach (%)	Accuracy after feature selection using CFS (%)	Accuracy after feature selection using proposed approach (%)	Accuracy after feature selection using CFS (%)	Accuracy after feature selection using proposed approach (%)	Accuracy after feature selection using CFS (%)	Accuracy after feature selection using proposed approach (%)
1	Audiology	73.45	73.45	73.00	76.10	76.10	77.43	73.45	76.10
2	Autos	65.36	66.34	68.78	69.26	67.80	78.53	59.51	63.90
3	Breast Cancer	73.42	74.47	71.32	74.12	73.07	75.87	73.77	74.82
4	Colic	86.41	87.50	84.23	84.51	85.59	85.59	85.05	86.14
5	Credit –A	85.65	85.50	84.34	85.50	86.08	85.50	85.65	85.50
6	Credit Germany	70.70	71.30	72.00	73.10	72.10	72.90	71.00	71.20
7	Diabetes	74.08	75.00	75.26	72.39	75.13	72.65	74.73	74.60
8	Glass	59.81	59.34	60.28	63.08	62.14	61.68	60.28	56.07
9	Heart –C	79.20	83.82	80.52	82.50	79.53	80.85	78.21	80.52
10	Heart Statlog	79.25	82.96	81.48	82.59	79.62	81.85	74.44	81.85
11	Hepatitis	82.58	84.51	81.93	82.58	81.93	81.29	77.41	80.64
12	Ionosphere	89.45	88.88	90.02	88.88	88.31	90.59	87.17	87.74
13	Iris	91.33	91.33	91.33	91.33	92.00	92.00	92.00	92.00
14	Labor	80.70	91.22	82.45	87.71	78.94	94.73	82.45	84.21
15	Primary Tumor	38.93	39.82	41.00	39.23	40.11	36.87	35.39	34.51
16	Segment	88.83	91.68	89.43	92.52	90.34	92.12	89.09	91.25
17	Sonar	72.59	75.96	73.07	77.88	72.59	69.71	71.15	69.71
18	Splice	94.95	95.17	93.38	92.79	94.48	94.51	92.47	94.01
19	Vehicle	48.10	60.16	47.04	65.60	47.39	65.13	43.14	61.34
20	Waveform	76.56	76.72	77.16	76.70	77.26	77.12	73.50	72.24

features and how they are related with each other.

The proposed technique used for feature selection is very simple to understand and implement, and also produces better results. JRipper is used because it uses good features only to build concise rules. Furthermore, association rule mining is used to filter the features picked up by the JRipper classifier. The two main factors support and confidence are used. Experimentation results shows that features selected are of significant impact. Accuracy of selected features is measured using JRip, J48, PART and Ridor to validate that these features are giving at least same or much better results. In some cases, accuracy achieved is not much better while using Ridor, PART and J48. On the other hand, the computational cost is less, causing the solution to be simpler and with a negligible difference in accuracy.

In future, a hybrid approach using ensemble classifiers will be experimented. To decrease the curse of dimensionality and will make data more understandable, change in threshold such as the support and confidence will also be investigated thoroughly to analyze the impact of number of features selected that significantly improve accuracy.

REFERENCES

Baesens B, Gestel TV, Viaene S, Stepanova M, Suykens J, Vanthienen J (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. J. Oper. Res. Soc. 54(6):627-635.

Cohen WW (1995). Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning. pp. 115-123.

Edelstein HA (1999). Introduction to data mining and knowledge discovery. 3rd edn. Two Crows Corporation.

Frank E, Witten IH (1998). Generating Accurate rule sets without global optimization. In: Proceedings of 15th International Conference on Machine Learning. pp. 144-151.

Friedman N, Geiger D, Goldszmidt M (1997). Bayesian network classifiers. J. Mach. Learn. Res. 29:131-163.

Gaines BR, Compton P (1995). Induction of Ripple-Down rules applied to modeling large databases. J. Intell. Inf. Syst. 5(3):211-228.

Hall MA, Smith LA (1999). Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper. Comput. Inf. Sci. AAAI Press pp. 235:239.

Han J, Kamber M (2006). Data mining: concepts and techniques, 2nd edn. Morgan Kaufmann Publishers.

Hassan MR, Hossain MM, Bailey J, Macintyre G, Ho JW, Ramamohanarao K (2009). A voting approach to identify a small number of highly predictive genes using multiple classifiers. BMC Bioinformatics. 10(Suppl 1):S19.

Kanan HR, Faez K, Taheri SM (2007). Feature selection using ant colony optimization (ACO): a new method and comparative study in the application of face recognition system. In: Proceedings of the 7th industrial conference on Advances in data. pp. 63-76.

Liu B, Abbass HA, McKay B (2002). Density-based heuristic for rule discovery with ant-miner. In: Proceedings of 6th Australia-Japan Joint Workshop on Intelligent Evolutionary Systems, Canberra, Australia. pp. 180-184.

Martens D, De Backer M, Haesen R, Vanthienen J, Snoeck M, Baesens B (2007). Classification with ant colony optimization. IEEE Trans. Evol. Comput. 11(5):651-665.

Parpinelli RS, Lopes HS, Freitas AA (2002). Data mining with an ant colony optimization algorithm. IEEE Trans. Evol. Comput. 6(4):321-332.

Quinlan JR (1993). C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Mateo, CA. USA.

Raymer ML, Punch WF, Goodman ED, Kuhn LA, Jain AK (2000). Dimensionality reduction using genetic algorithms. IEEE Trans. Evol. Comput. 4(2):164-171.

Shahzad W, Baig AR (2010). A hybrid associative classification algorithm using ant colony optimization. IJICIC 7(12):6815-6826.

Su Z, Hong H, Perkins R, Shao X, Cai W, Tong W (2007). Consensus analysis of multiple classifiers using non-repetitive variables: Diagnostic application to microarray gene expression data. Comput. Biol. Chem. 31(1):48-56.

Wang Y, Makedon FS, Ford JC, Pearlman J (2005). HykGene: A hybrid approach for selecting marker genes for phenotype lassification using microarray gene expression data. Bioinformatics 21(8):1530-1537.

Wei HL, Billings SA (2007). Feature subset selection and ranking for data dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. 29(1):162-166.

Yahang J, Honavar V (1997). Feature subset selection using a genetic algorithm. Pattern Recogn. 13(2):380.

Yang YH, Xiao Y, Segal MR (2005). Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics 21(7):1084-1093.

Zhang Z, Yang P, Wu X, Zhang C (2009). An agent-based hybrid system for microarray data analysis. IEEE Intell. Syst. 24(5):53-63.

Zhou H, Wu J, Wang Y (2007). Wrapper approach for feature subset selection using genetic algorithm. In: Proceedings of International Symposium on Intelligent signal Processing and Communication Systems. pp. 188-191.