

# Performance Comparison based on Attribute Selection Tools for Data Mining

M. Sudha<sup>1\*</sup> and A. Kumaravel<sup>2</sup>

<sup>1</sup>Department of Mathematics, Amet University, Kanathur, Chennai-600112, India; seedinmenew@yahoo.com

<sup>2</sup>Department of Computer Science and Engineering, Bharath University, Selaiyur, Chennai-600073, India; drkumaravel@gmail.com

## Abstract

Recent years have seen wide efforts in attribute selection research. Attribute selection can efficiently reduce the hypothesis space by removing irrelevant and redundant attributes. Attribute reduction of an information system is a key problem in rough set theory and its applications. In this paper, we compare the performance of attribute selection using two technical tools namely WEKA 3.7 and ROSE2. Filter methods are used as an alternative measure instead of the error rate to score a feature subset. This measure was chosen to be fast to compute, at the same time as still capturing the usefulness of the feature set. Many filters provide a feature ranking rather than an explicit best feature subset, and the cutoff point in the ranking is chosen via cross-validation. We used search methods like Best first and Greedy stepwise to evaluate a subset of features as a group for suitability. We use the internet usage data set for this purpose and then comparison results are tabulated for various methods for searching the solution space to eliminate the irrelevant attribute. Results of this research show some interesting issues of attribute selection tools where we found better ways to have select irrelevant attributes. Comparing the tools of attributes reductions evidence some considerable differences between them.

**Keywords:** Classifications, Data Mining, Rough Set Explorer, Search Methods, Selected Attributes, WEKA

## 1. Introduction

Attribute set reduction is essentially a task to remove irrelevant and/or redundant features. Irrelevant attributes can be removed without affecting learning performance. Redundant attributes are a type of irrelevant attribute. The distinction is that a redundant attribute implies the co-presence of another attribute; individually each feature is relevant, but the removal of one of them will not affect learning performance<sup>1</sup>.

The fundamental goal is to predict whether the given attribute will bind to a target and it's how strongly gives the exact result of the survey. Broadly, the aim of Dimensionality Reduction (DR) is to make a more concise

and representative model than the initial dataset model representation. More specifically, DR aims to:

- (1) reduce the amount of data given by the input space;
- (2) retain only the relevant input data; and
- (3) improve the quality of data for predictive accuracy and algorithm time performance<sup>1</sup>.

In this paper we use data mining methods to improve the survey results and in the same hand we are comparing the tools rough set explorer ROSE 2.2 and WEKA. For that we apply classification models based on different data mining approaches that can help us to select appropriate attributes among more attributes<sup>2</sup>.

\*Author for correspondence

Classifications methods belong to supervised learning. The methods help us to identify attributes which might be classified as highly useful. The parameters describe how effective the internet is. It indicates how much a particular student is needed to restrict to use the web sites. Figure 1 defines process we are going to take for the comparison.

## 2. Methods and Materials

In this section, we describe the data set with various search methods and compare the accuracy obtained by the tools with whole dataset with the selected attributes. Rose and Weka tools are used to reduce dimensionality with various attribute select techniques and classification models<sup>4,7</sup>.

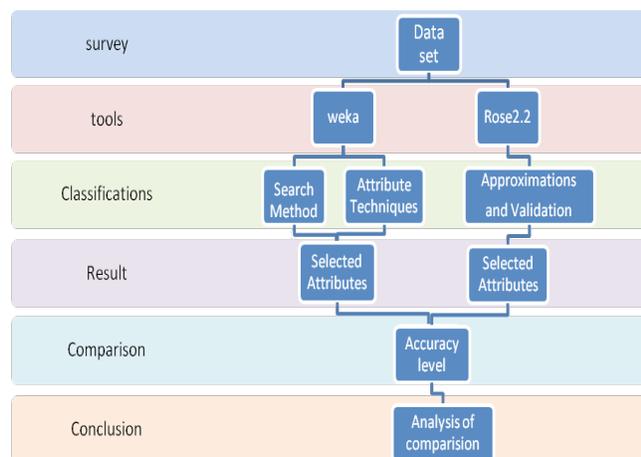
The data set for this analysis are from the survey taken among the students of the AMET university.

### 2.1 Dataset

#### 2.1.1 Data Set Description

The database of the survey conducted among students is considered as our experimental dataset. Number of instances in this data base is 250 and the number of attributes is 11 plus the class.

This internet user survey is focused on issues such as how students accessed the internet, what applications the used, how they managed their personnel information online and the usages of internet. Key findings from this survey cover a broad range of topics.



**Figure 1.** Process of optimum attributes selection tool searching.

#### 2.1.2 Attribute Description

The usage of internet attributes of students are enumerated below with their description:

- (1) Duration: How many times a student launch browser per day or week for a specific set of tasks or activities; how many hours per week they use the computer for fun/play or job or studies.
- (2) Effective: How effectively the students use the web effectively to interact with other students or parents and for developing and to reduce the red tape.
- (3) Like: Whether the students like the Internet a lot or sometimes for the purpose of advertising or to search for specific information or to convey the right impression to others or to have fun and explore.
- (4) Bookmarks: This make it easy to get back to your favorite places thereby we get data of internet usages. How many items a student contain in bookmarks? What problems they have in using bookmarks? Which bookmark activities often they perform?
- (5) Problems: To find the problems the students are facing when they using the internet like takes too long to view or costs too much or Getting errors from pages or Too many “junk” sites or to register or useless graphics or take too long to load.
- (6) Prime use: For what the students primarily use the web, education, shopping, entertainment, communication, information etc.
- (7) Essential: What internet technologies are essential for the student’s consideration? www, online, email, streaming audio and video, phone, fax etc..
- (8) New Sites: Students are using internet for so many purposes and searching so many sites. How they get to know about new sites. Through books, magazines, emails, newspapers, friends, hyperlinks, directories or advertisements etc..
- (9) Categories: How often the students have used the web for the categories like news, online chatting, commercial products, reference materials, projects, maps, medical information, job listings etc.
- (10) Activities: Web usage can take away time from other activities. How frequently have Students used the Web instead of the following activities like watching TV, talking in phone, sleeping, exercising, reading, socializing, doing works etc.
- (11) Class: Finally the survey finds on the usage of internet. Internet not a useful class is indicated by 1, useful class by 2 and very useful class by 3.

## 2.2 Methodology

The first step of our analysis was to reduce the high data dimensionality<sup>7</sup>. For this purpose we use Rose tool and Weka tool for attribute selection based on various search methods made in the attribute space as shown in Table 1 and Table 2. We use factors which are selected after preprocessing as new predictors<sup>8</sup>.

The information collected from the survey will be analyzed and evaluated in several stages in each tool.

### 2.2.1 Rose 2 Tool

All data analysis methods in ROSE system are grouped as follows:<sup>14</sup>

Preprocessing – methods meant for preliminary data analysis and modifications, like discretization, 2. Reducts – methods dealing with the reduction of attributes, 3. Rules – methods used to generate decision rules, 4. Classification – validation of decision rules, Similarity relation – methods using similarity relation approach. One method is excluded from all groups – rough set approximations.

It is used to discretizes attributes with continuous domains into the ones with discrete domains. All such attributes in the source data file will be processed. It is an

**Table 1.** Selection of attributes by various search methods of ROSE tool

| S. No | Method of Parameters                     | Selected Attributes  | Total No. of Attributes Removed | Deduction Rate |
|-------|--|--|---------------------------------|----------------|
| 1.    |  | Effective, Like, Bookmarks, Prime use, New sites, Categories, Activities           | 3                               | 70%            |
| 2.    |  | Duration, Effective, Like, Bookmarks, Problems, New Sites, Categories, Activities  | 2                               | 66%            |
| 3.    | Reduction Method Core and Lattice search | Effective, Like, Bookmarks, Problems, Essential, New Sites, Categories, Activities | 2                               | 69%            |
| 4.    |  | Effective, Like, Bookmarks, Prime Use, Essential, Activities                       | 4                               | 80%            |
| 5.    |  | Effective, Like, Bookmarks, Prime Use, New Sites, Categories, Activities           | 3                               | 71%            |
| 6.    |  | Effective, Like, Problems, Prime Use, New Sites, Categories, Activities            | 3                               | 74%            |

**Table 2.** Selection of attributes by various search methods of WEKA tool

| S. No | Method of Parameters | Selected Attributes | Total No. of Attributes Removed                                   | Deduction Rate |       |
|-------|----------------------|---------------------|---|----------------|-------|
| 1.    | CfsSubsetEval        | Greedy step wise    | Duration, bookmarks, new sites, categories                        | 6              | 81.3% |
|       |                      | Best first          | Duration, bookmarks, essential, new sites, categories, activities | 4              | 78.4% |
| 2.    | Info gain            | Ranker              | categories, new sites, essential, Duration, activities, bookmarks | 4              | 78.6% |
| 3.    | Gain ratio           | Ranker              | new sites, categories, essential, activities, Duration, bookmarks | 4              | 79%   |

entropy-based method. User must select one of the decision attributes and enter the name of the result file (there is a default filename). There are also some additional parameters: Additional stopping conditions :

1. Desired value of entropy  $<0, 1>$  – if it is reached the method doesn't generate additional subintervals,
2. Maximum number of intervals – method doesn't create more intervals than that.

Here the data preprocessed under local discretization without additional stopping conditions. The approximation viewer gave accuracy percentage of lower and upper approximations of each class with quality of classification 0.97.

Reduction method produced the quality of classification for a all conditional attributes from core attributes. Here the core attributes are  $A_2$ (effective),  $A_3$ (like) and  $A_{10}$ (Activities).

In ROSE system you are able to do the classification test using any of the available methods for rule generation. For that we selected Minimal covering validation method from the Classification group. ROSE system uses L-metric type classifier, which has several parameters:

Lattice search: This method is used to obtain all (or limited number) reducts of the information table. It implements optimized lattice search method. Due to time consuming nature of this process user may limit the

number of found reducts using Number of reducts input line. Decision attribute must be selected also<sup>2</sup>.

**Indiscernibility matrix:** This method is used to obtain all reducts of the information table. Its implementation is based on a structure called indiscernibility matrix. It is very fast method so it should be preferred above lattice search. User must select one of decision attributes<sup>2</sup>.

**Heuristic search:** This method is used to obtain approximate reducts when it is impossible to generate all the exact ones. This heuristic works by adding attributes to the existing core by selecting most promising of them all and trying different paths. One of the most important elements of data analysis is rule generation. We will use one of the available methods – minimal covering rules (LEM2)<sup>2</sup>. For that we selected **Minimal covering** method from the **Rules** group<sup>14</sup>.

It cross validate the 10 folds under the division random. If the division stratified, then distribution of examples in decision classes in each fold is the same as in the whole learning (input) file. It gave classification results with majority threshold 21% and minimum similarity 50%. It provides the confusion matrix and average accuracy percentages of correctly classified and incorrectly classified attributes. To get more accuracy we can classify using all partially matched rules or most similar which was supported by class of rule strength or rule strength x similarity.

The above process done to the data set and the results are shown in below Table 1.

### 2.2.2 Weka Tool

Weka supports several data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization and attribute selection. All of Weka's techniques are predicated on the assumption that the data available as a single flat file or relation, where each data point is described by a fixed number of attributes<sup>15,16</sup>.

Any learning algorithm in WEKA is derived from the abstract `weka.classifiers.Abstract Classifier` class. This, in turn, implements `weka.classifiers.Classifier`. Surprisingly little is needed for a basic classifier: a routine which generates a classifier model from a training dataset (= `buildClassifier`) and another routine which evaluates the generated model on an unseen test dataset (= `classifyInstance`), or generates a probability distribution for all classes (= `distributionForInstance`).

A classifier model is an arbitrary complex mapping from all-but-one dataset attributes to the class attribute. The specific form and creation of this mapping, or model, differs from classifier to classifier. For example, ZeroR's (= `weka.classifiers.rules.ZeroR`) model just consists of a single

value: the most common class, or the median of all numeric values in case of predicting a numeric value (= regression learning). ZeroR is a trivial classifier, but it gives a lower bound on the performance of a given dataset which should be significantly improved by more complex classifiers. As such it is a reasonable test on how well the class can be predicted without considering the other attributes<sup>11</sup>.

Let see how to interpret the output from classifiers in detail and focus on the Correctly Classified Instances in the section Stratified cross-validation and notice how it improves from ZeroR to J48:

A more elaborate method is cross-validation. Here, a number of folds  $n$  is specified. The dataset is randomly reordered and then split into  $n$  folds of equal size. In each iteration, one fold is used for testing and the other  $n-1$  folds are used for training the classifier. The test results are collected and averaged over all folds. This gives the cross-validation (cv) estimate of the accuracy. The folds can be purely random or slightly modified to create the same class distributions in each fold as in the complete dataset. In the latter case the cross-validation is called stratified. Leave-one-out (= loo) cross-validation signifies that  $n$  is equal to the number of examples. Out of necessity, loo cv has to be non-stratified, i.e. the class distributions in the test set are not related to those in the training data. Therefore loo cv tends to give less reliable results. However it is still quite useful in dealing with small datasets since it utilizes the greatest amount of training data from the dataset.

We processed the data in several classifier like Bayes-naivebayes, function-multilayer perception, rules-(JRip, ZeroR) and Trees- (J48, random forest)<sup>13</sup> and reduced the attributes by several attribute evaluators subset selection methods from `CfsSubsetEval` to wrapper `subeval` which processed by search methods called Greedy stepwise, Best first and Ranker. The following Table 2 shows the results of our search methods where we can the accuracy level of each method of selected attributes<sup>1</sup>.

## 3. Results and Comparison

In this experiment a comparative study of attribute reduction of rough set and classification technique of data mining technique for dataset on various parameters using missing attribute value in data set containing 10 condition attribute and 1 decision attribute<sup>13</sup>. During Rough set, the dataset is given as input to ROSE2 tool reduce the attribute based on consistency and inconsistency of a table was implemented for rule generation.

In data mining during simple classification, the training dataset is given as input to WEKA3.7 tool and the classification algorithm namely Random Tree was implemented.

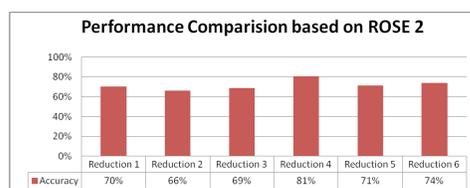
While assessing and comparing performance of one over the other, accuracy and error rate are among the common methods that are widely used. Other evaluation factors include speed, interpretability, ease of programmability and risk when errors are generalized<sup>3,13</sup>.

We used Weka (3.7) a learning machine tool to draw the comparative analysis. In this paper different combination of feature selection methods are tried and they include BestFirst + CfsSubsetEval, GeneticSearch + CfsSubsetEval, GreedyStepwise + CfsSubsetEval, Ranker + ChiSquaredAttributeEval, Ranker + InfoGainAttributeEval and Ranker + GainRatioAttributeEval<sup>15,16</sup>. The details of the combinations and the features selected by each combination and their visualization is described in Table 2 and Figure 3<sup>8</sup>.

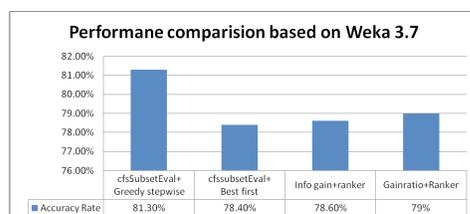
As discussed later, rough set theory approaches to attribute selection are presently included in this category; as are Decision Trees, the Naive Bayes classifier, and other Rule Learners<sup>9</sup>. Performance of Weka 3.7 and Rose 2 induction has been shown to improve after effective discretisation<sup>10</sup>. Accuracy level each selected attribute of conditional attribute of lattice search method of ROSE tool shown in the Table 1 and Figure 2

## 4. Conclusion

In this paper we have presented the results generated in two different tools WEKA 3.7 and ROSE2. This shows



**Figure 2.** Performance comparisons of various reductions on lattice search methods.



**Figure 3.** Performance comparison of various search methods on Bayes, JRip and J48.

comparable strength between ROSE2 with WEKA3.7. In Future the experiment can be extended to bigger data.

## 5. Acknowledgement

The first author would like to thank of AMET University for the support and the encouragement for this research work.

## 6. References

- Liu H, Motoda H, editors. Computational Methods of Feature Selection.
- Scully P, Jensen R. Investigating rough set feature selection for gene expression analysis dissertation BSc Computer Science. Aberystwyth University; 2011 May.
- Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufman; 2005.
- Ponce J, Karahoca A. Data Mining and Knowledge Discovery in Real Life Applications, chapter 3. Rough Set Theory: Fundamental Concepts, Principals, Data Extraction, and Applications; 2009 Jan.
- Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data; 1991.
- Jelonek J, Krawiec K, Slowinski R. Rough set reduction of attributes and their domains for neural networks. Comput Intell. 1995; 11:339–47.
- Jensen R, Shen Q. Semantics-preserving dimensionality reduction: Rough and fuzzyrough-based approaches. IEEE Trans Knowl Data Eng. 2004; 16:1457–71.
- Swiniarski RW, Skowron A. Rough set methods in feature selection and recognition. Pattern Recogn Lett. 2003; 24:833–49.
- Swiniarski RW, Skowron A. Rough set methods in feature selection and recognition. Pattern Recogn Lett. 2003; 24:833–49.
- Han J, Kamber M. Data Mining: Concepts and Techniques. Kauffman Press; 2002.
- Cunningham P, Delany SJ. KNearest Neighbour Classifiers; 2007.
- Grzymala-Busse JW, Hu M. A comparison of several approaches to missing attribute values in data mining. Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing, RSCTC-2000. Banff, Canada. 2000 Oct 16-19. 340–7.
- Breiman L. Random Forests. 2001.
- Available from: [http://rosecompiler.org/ROSE\\_UserManual/ROSE-UserManual.pdf](http://rosecompiler.org/ROSE_UserManual/ROSE-UserManual.pdf)
- Available from: <http://weka.wikispaces.com/Primer>.
- Available from: <http://weka.sourceforge.net/doc.dev/weka/attributeSelection>.