Sigmis: A Feature Selection Algorithm Using Correlation Based Method

E. Chandra Blessie[1,*] and E. Karthikeyan[2]
[1]Department of Computer Science
D. J. Academy for Managerial Excellence,
Coimbatore – 641 032, Tamil Nadu, India.
[2]Department of Computer Science, Govt. Arts College,
Udumalpet, Tamil Nadu, India.

Received: 01/03/2011; Accepted: 26/04/2012

ABSTRACT
Feature Selection is one of the preprocessing steps in machine learning tasks. Feature Selection is effective in reducing the dimensionality, removing irrelevant and redundant feature. In this paper, we propose a new feature selection algorithm (Sigmis) based on Correlation method for handling the continuous features and the missing data. Empirical comparison with three existing feature selection algorithms using UCI data sets shows that the proposed system is very effective and efficient in selecting the feature set.

Keywords: Feature selection, Dimensionality, Correlation and missing data

1. INTRODUCTION

Feature Selection is one of the prominent preprocessing steps in many of the machine learning applications. It is the process of reducing the feature set by choosing the relevant features from the original feature set according to an evaluation criterion and also removing the redundant features from the entire feature set.

Let D be the feature set with larger number of features $\{F_1, F_2, F_3 \ldots\ldots F_k\}$ where k is the number of features in the data set. Feature selection F_i is defined as the process of selecting d, most discriminatory features out of $D \geq d$ [1]. Feature selection methods involve generation of the subset, evaluation of each subset, criteria for stopping the search and validation procedures.

[*] Corresponding author. chandra_blessie@yahoo.co.in

Researchers have studied various aspects of feature selections. One of the key aspects is to measure the goodness of a feature selection in determining an optimal one. Different feature selection methods can be broadly categorized into the wrapper model [2], the filter model [3] and the hybrid model [4]. The wrapper model uses the predictive accuracy of a predetermined learning algorithm to determine the goodness of the selected subsets. These methods are computationally expensive for data with a large number of features. 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 the training data such as distance, information, dependency, and consistency.

To combine the advantages of both models, algorithms in a hybrid model have recently been proposed to deal with high dimensional data. In these algorithms, first, a goodness measure of feature subsets based on data characteristics is used to choose best subsets for a given cardinality, and then, cross validation is exploited to decide the best subset. These algorithms mainly focus on combining filter and wrapper algorithms to achieve best possible performance with a particular learning algorithm with similar time complexity of filter algorithms. Search is another key problem in feature selection.

One of the measures used for feature selection is dependency measures. Many dependency based methods have been proposed. The main measure is Correlation based method. Pearson's Correlation method is used for finding the association between the continuous features and the class feature. In [5], Symmetric uncertainty measures are used for finding the association between the discrete feature and the class feature. Also, it is used for finding the feature-feature correlation to remove redundant feature.

Correlation method is used for finding the association between the features for sample data. If the entire population is taken, (as there is a rapid growth in data size), Correlation Coefficient may not yield goodness of result. This motivated us to use the t-test of Correlation Coefficient to examine whether the association between the features is statistically significant.

Also in [6], the author had proposed to take the average of the data to fill the missing values of continuous features and the most common value to fill the missing value of the discrete feature. This will lead to the increase of misclassification rate which can affect the classification accuracy. So, instead of filling it with average value and most common value, two new methods are proposed. In the first method, the features which have many number of missing values are removed as these features will not give much information for learning purpose. In the second method, the missing values are found out by finite difference method which will lead to efficient selection of feature set.

The rest of this paper is organized as follows. Section 2 discusses about the related works on correlation based feature selection methods. Section 3 describes the proposed algorithm and the experimental result is given in section 4. In section 5, we conclude our work with some possible extension in the future.

2. RELATED WORKS

Various evaluation measures such as Information Theory, Consistency based measures [21], Chi-Squared based measures, Information Gain measures, Symmetric Uncertainty measures and Correlation-based measures are used for removing the irrelevant and redundant features.

2.1. Correlation based measures

Correlation is a well-known similarity measures between two features. If two features are linearly dependent, then their correlation coefficient is ±1. If the features are uncorrelated, the correlation coefficient is 0.

The association between the features is found out by using the correlation method. There are two broad categories that can be used to measure the correlation between two random variables. One is based on classical linear correlation and the other is based on information theory. Out of these two, the most familiar measure is linear correlation coefficient. As per the standard literature, for a pair of variables (X, Y), the linear correlation coefficient '*r*' is given by:

$$r = \frac{\sum (X_i - \overline{X_i})(Y_i - \overline{Y_i})}{\sqrt{\sum (X_i - \overline{X_i})^2}\sqrt{\sum (Y_i - \overline{Y_i})^2}} \tag{1}$$

Lei Yu and Huan Liu [5], have proposed FCBF (Fast Correlation-based Filter) algorithm. In this, the correlation between the continuous feature and class feature is found out by using symmetric uncertainty measure. If the value is higher than the threshold value (say 0.5), then the feature will be selected. The selection stops when the number of features equals *nlogn*. All the selected features are ranked according to the decreasing order. Then the feature-feature correlation is done to remove the redundant feature.

Jacek and Duch [7] have proposed Kolmogorov-Smirnov Correlation based filter method. In this filter method, The features are ranked by using F-score. The threshold for selecting the features is taken from frapper approach. The feature redundancy removal is based on the Kolmogorov-Smirnov correlation-based filter method.

Jacek and Duch [8], have proposed PRBF (Pearson's Redundancy Based Filter) algorithm. This algorithm uses Symmetric Uncertainty measure as relevancy measure to rank the features in decreasing order of their relevance. Pearson Chi-Square test is used to remove the redundant features. The authors [1–20], uses absolute mutual correlation method for removing irrelevant and redundant features by discarding the features having the largest value.

Mark and Smith [2], have proposed CFS (Correlation-based Feature Selection) algorithm. First the training data set is discretized using the method of Fayyad and Irani [9] and then passed to CFS. CFS calculates feature-class and feature-feature correlations using symmetric uncertainty and then searches the feature subset space.

Mark [2], has proposed an algorithm for continuous and discrete features. This algorithm calculates correlation coefficient between continuous & class features using heuristic measures. Selecting the best subset is done using best first search. BFS starts with empty feature and evaluates. Search stops when 5 consecutive fully expanded non-improving subsets arrive. For both continuous features, it uses correlation formula. For one continuous and one discrete feature, it uses weighted Pearson's formula. For both discrete features, it uses Symmetric Uncertainty. The author proposed to find the average of continuous features to fill the missing values and for discrete features, the most common value is used.

Lei Yu and Huan Liu [10], uses Markov blanket for finding the irrelevant features. Zheng Zhao and Huan Liu [11], have proposed INTERACT algorithm. For relevancy analysis, Symmettric Uncertainty [17–18] is used and the list is arranged in descending order. Then the correlation between the last feature and the remaining features are found out using C-Contribution measure. If the value is less than the threshold, it is removed.

The authors [12], uses CACC discretization method for continuous features and then removes irrelevant and redundant features. Babu and Reddy in [13], uses coefficient of dispersion for redundancy feature removal. Patricia and Andries [19] uses decision rule based categories to remove the irrelevant features.

3. PROPOSED CONCEPT

In this section, we describe the methodologies used for selecting the relevant features and for handling the missing values.

3.1. Methodologies and Pseudo Code

a) T-test for Correlation Coefficient

In the study, only sample correlation coefficient is used to find out the correlation between the continuous feature and the class feature. Due to the rapid growth of the data size, the correlation coefficient which is high for a sample may not yield good result for the entire population. So, it must be determined whether there is significant association between the features, while taking the entire population. The most frequently used test to examine whether the two features are statistically correlated is the t-test. The methodology used in the proposed algorithm is to use t-test for selecting the most significant features from the features set.

The simplest formula for computing the appropriate *t* value to test significance of a correlation coefficient employs the t distribution. The *t* value can be calculated by assuming that there is no correlation between them ($\rho = 0$).

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{2}$$

Where *n* is the number of instances and *r* is the correlation coefficient for sample data. The significance of the relationship is expressed in probability levels: p (e.g., significant at p = 05). The degrees of freedom for entering the t-distribution is $n - 2$. If the *t* value is greater than the critical value at 0.05 significant level, then the feature is significant and it is selected.

<u>Pseudo Code for Feature Selection Algorithm (Sigmis)</u>

```
Input : $S(F_1, F_2, \ldots\ldots, F_k, F_c)$ // a training data set
Output : $S_{best}$                 // the selected feature set

Step 1 : begin
Step 2 : for $i = 1$ to $k$ do begin
                 $r$ = calculate correcoeff($F_i$, $F_c$);
         end;
// let p = 0.05 significant level;
Step 3 : let $\rho = 0$ // assuming there is no significant correlation between $F_i$
         and $F_c$;
Step 4 : for $i = 1$ to $k$ do begin
                 $t$ = calculate signi($r$, $\rho$) for $F_i$; // using t-test value from eqn (2)
                 if $t$ > CV // Critical value
```

```
                $S_{best} = S_{list}$;
          end;
Step 5 :  return $S_{best}$;
          end;
```

b) Handling of missing values

To handle the missing values in a feature, CFS [3] replaces the missing values by taking the average value for continuous features and the most common value for discrete features. This may lead to the increase in misclassification rate. To overcome this problem, two methods are introduced. First method is, if there are many missing values greater than the threshold (user-specific), the feature can be removed, since very less information can be extracted during the learning process. Secondly, the missing value can be found out by using Finite Differences measure. Each missing value can be filled up with the correct value and then the correlation measures can be used which will give an effective and efficient way for selecting the feature.

Finite Difference measure

When one or more of the F_i values are missing, symbolic finite difference operators E and Δ are used. Let $V_1, V_2, \ldots.., V_n$ be the values of feature F_i. The difference between the previous two values of the feature F_i are

$V_2 – V_1, V_3 – V_2, \ldots\ldots, V_n – V_{n-1}$

This is called as the first order differences and it is denoted by Δ operator called as forward difference operator. The higher order differences can be found out recursively, till the missing value is found out.

Pseudo code for handling missing values

Let F be the set of features with $\{F_1, F_2, F_3, \ldots\ldots., F_k, F_c\}$ where k is the number of features and F_c is the class feature. Let S_{list} be the training data set with n instances and S_{miss} be the set of missing values $\{m_1, m_2, \ldots..m_s\}$, where s is the number of missing values. Let $V = \{v_1, v_2\ldots., v_n\}$ be the instances. Take a missing value from S_{miss} and find its value with S_{list}. Assume the fixed value for the missing value (say 10) to find the correct missing value. Repeat the process for all the missing values. To do this, first sort the data set along with the class feature.

```
Input : $S_{list}(F_1, F_2, \ldots\ldots., F_k, F_c)$ // a training data set
Output : $S_{new}$                    // the changed data set
Step 1 : begin
Step 2 : for $i = 1$ to $k$ do begin
```

```
                Sort the feature $F_i$ in $S$ along with the class feature $F_c$;
            end;
Step 3 : for $i$ = 1 to $n$ do begin
                If ($v_i$ = = NULL) // if there is a missing value
                $S_{miss} = S_{list}$ ;
                break;
            end;
Step 4 : for $i$ = 1 to length($S_{miss}$) do begin
                append($S_{list}$) = getFirstElement($S_{miss}$)
              $S_{miss} = v_i - v_{i-1}$ // Find the missing value by finite difference
              method;
            end;
        replace the missing value by using the formula
            miss_value = abs($S_{miss}$ − fix)
Step 5 : repeat Step 3 and 4 for the remaining features.
Step 6 : return $S_{new}$
Step 7 : end;
```

Let D = [2 4 6 ? 10 13 ? 23 31]. In the first run, S_{miss} = [2 4 6 ?]. Replace ? by 10. So the set is [2 4 6 10]. Find the difference [2 2 4] as in step 4. Again we get [0 2]. In the next process, we get 2. Do step 4, miss_value = abs(2–10) = 8. The second run uses [10 13 ?] data and so on.

4. EXPERIMENTAL ANALYSIS

The objective of this section is to evaluate the algorithm in terms of removing the irrelevant and redundant features.

4.1. Data Set

The data sets used for experimental analysis are selected from UCI Machine Learning repository. The data sets are listed in table 1.

Table 1. Data sets used.

Data Set	Number of Continuous Features	Number of instances	Number of Classes
Lung cancer	57	32	3
Diabetics	9	768	2
Breast cancer	10	699	2

4.2. Performance Evaluation

In our experiment, we choose the following three feature evaluation algorithms.

1. CFS Subset Evaluation
2. Consistency subset Evaluation
3. FCBF algorithm

The experiments are conducted using Weka's implementation of all these existing algorithms and the proposed algorithm is implemented in MatLab for comparison. The result for feature subset evaluation is shown in table 2.

Table 2 summarizes the number of selected features for each feature selection algorithm. From the result given above, it is clear that the proposed algorithm achieves the highest level of dimensionality reduction by selecting the least number of features. Fig. 1 depicts the number of features selected using various algorithms.

Table 2. Number of features selected.

Data set\ Algorithms	CFS	Consistency	FCBF	Proposed
Lung cancer	9	56	9	5
Breast cancer	7	7	4	4
Diabetes	4	8	4	6

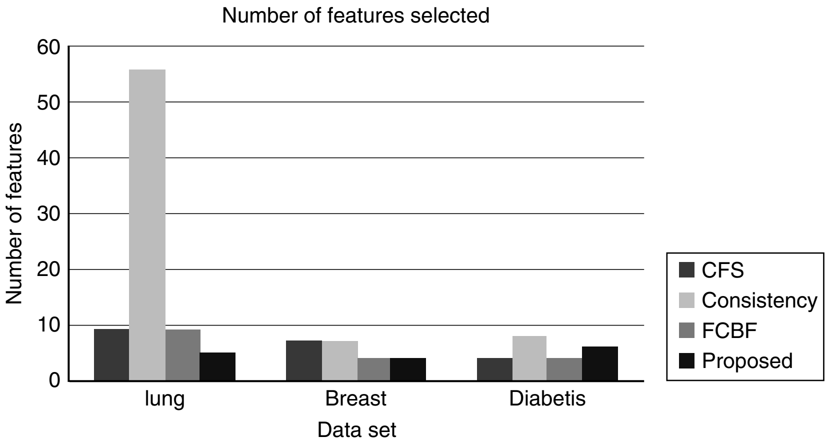

Figure 1. Number of features selected.

5. CONCLUSION AND FUTURE WORK

In this paper, a new algorithm is proposed for dealing with the missing values and for selecting the best feature set. Instead of choosing the correlation coefficient, t-test on correlation coefficient is used to find out the significant of features. The approach demonstrates its efficiency and effectiveness in dealing with the value. In future, sampling methods can be used to reduce the misclassification error and to increase the predictive accuracy of the model. Handling of discrete features in feature selection will be studied in future.

REFERENCES

[1] Haindl, M., Somol, P., Ververidis, D. and Kotropoulos, C., Feature Selection Based on Mutual Correlation, Progress in Pattern Recognition, Image Analysis and Applications, Springer Heidelberg, 2006, 4225, 569–577.

[2] Hall, M.A. and Smith, L.A., *Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper*, American Association of Artificial Intelligence, 1998.

[3] Sebban, M., On Feature Selection: a New Filter Model, *Proceedings of the twelfth International FLAIRS Conference*, 1999, 230–234.

[4] Li-Yeh Chuang, Chao-Hsuan Ke, and Cheng-Hong Yang, A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification, *Proceedings of the International MultiConference of Engineers and Computer Scientists*, IMECS, Hong Kong, 2008, I, 19–21.

[5] Yu, L. and Liu H., Feature selection for high-dimensional data: A fast correlation-based filter solution, *ICML-03 Proceedings of the twelfth International Conference on Machine Learning*, Morgan Kaufmann, Washington, D.C., San Francisco, CA, USA, 2003, 856–863.

[6] Hall, M.A., Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning, *Proceedings of the Seventeenth International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, USA, 2000, pages 359–366.

[7] Jacek Biesiada and Włodzisław Duch, A Kolmogorov-Smirnov Correlation-Based Filter for Microarray Data, *Neural Information Processing*, Springer Heidelberg, 2008, 4985, 285–294.

[8] Jacek Biesiada and Włodzisław Duch, Feature Selection for High-Dimensional Data: A Pearson Redundancy Based Filter, *Advances in Soft Computing*, Springer Heidelberg, 2007, 45, 242–249.

[9] Fayyad, U. and Irani, K.B., Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the thirteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Washington, D.C., San Francisco, CA, USA, 1993, 1022–1027.

[10] Lei Yu and Huan Liu, Efficient Feature Selection via Analysis of Relevance and Redundancy, *Journal of Machine Learning Research*, 2004, 5, 1205–1224.

[11] Zheng Zhao and Huan Liu, Searching for Interacting Features, *Proceedings of the twentyth International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, Washington, D.C., San Francisco, CA, USA, 2007, 1156–1161.

[12] Senthamarai Kannan, S. and Ramaraj, N., An Improved Correlation-based Algorithm with Discretization for Attribute Reduction in Data Clustering, *Data Science Journal*, 2009, 8(2), 125–138.

[13] Babu Reddy, M. and Reddy, L.S.S., Dimensionality Reduction: An Empirical Study on the Usability of IFE-CF (Independent Feature Elimination- by C-Correlation and F-Correlation) Measures, *IJCSI International Journal of Computer Science*, 2010, 7(1).

[14] Guyon, I. and Elisseeff A., An introduction to variable and feature selection, *Journal of Machine Learning Research*, 2003, (3), 1157–1182.

[15] Hall, M. A., *Correlation-based feature selection for machine learning*, PhD thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999.

[16] Liu, H. and Motoda H., *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[17] Kohavi, R. and John G.H., Wrappers for feature subset selection, *Artificial Intelligence*, 1997, (1–2), 273–324.

[18] Xing, E., Jordan, M. and Karp, R., Feature selection for high-dimensional genomic microarray data. *Proceedings of the Eighteenth International Conference on Machine Learning*, Morgan Kaufmann, Washington, D.C., San Francisco, CA, USA, 2001, 601–608.

[19] Patricia, E.N. Lutu, and Engelbrecht, A.P., A decision rule-based method for feature selection in predictive data mining, *Expert Systems with Applications*, 2010, 37, 602–609.

[20] Molina, L.C., Belanche, L. and Nebot, A., Feature Selection algorithms: a survey and experimental evaluation, *ICDM- Proceedings of IEEE International Conference on Data Mining*, IEEE Xplore, 2002, 306–313.

[21] Dash, M., Liu H. and Motoda H., Consistency based feature selection. *Proceedings of the fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, Springer-Verlag, London, UK, 2000, 98–109.