AN IMPROVED CORRELATION-BASED ALGORITHM WITH DISCRETIZATION FOR ATTRIBUTE REDUCTION IN DATA CLUSTERING

S. Senthamarai Kannan[1*] and Dr N. Ramaraj[2]

*[*1] Department of Information Technology, Thiagarajar College of Engineering, Madurai, India*
Email: sskit@tce.edu
[2]Principal, G. K. M. Engineering College, Chennai, India

ABSTRACT

Attribute reduction aims to reduce the dimensionality of large scale data without losing useful information and is an important topic of knowledge discovery, data clustering, and classification. In this paper, we aim to solve the current problem that a continuous attribute in a clustering or classification algorithm must be made discrete. We propose a new algorithm of data reduction based on a correlation model with data discretization. It deals with selection of continuous attributes from a very large set of attributes. The proposed algorithm is an extended version of the Fast Correlation-based filter algorithm and is named FCBF$^+$. The FCBF$^+$ algorithm performs the discretization of continuous attributes in an efficient manner. Then it selects the relevant attributes from a very large set of attributes. Performance evaluation is done on clustering accuracy for all the features, and a reduced set of features is obtained using FCBF$^+$. It is found that the proposed FCBF$^+$ algorithm improves the clustering accuracy of various clustering algorithms.

Keywords: Clustering, Attribute reduction, Data discretization, Correlation-based model, Knowledge discovery, Data mining

1 INTRODUCTION

Data mining or knowledge discovery in databases (KDD) (Han & Kamber, 2005) simply means the non-trivial extraction of implicit, previously unknown, and potentially useful information from data. It deals with the discovery of hidden knowledge, unexpected patterns, and new rules, especially from large databases. KDD is the process of identifying a valid, potentially useful, and ultimately understandable structure in data.

Feature selection (Liu & Motoda, 1998; Pyle, 1999; Blum & Langley, 1997; Liu & Motoda, 1998; John, Kohavi, & Pfleger, 1994; Kira & Rendell. 1992; Kohavi & John, 1997) is a fundamental problem in data mining that selects out relevant features and casts away irrelevant and redundant features from an original feature set based on certain evaluation criteria. Feature selection improves mining performance such as predictive accuracy and result comprehensibility.

Feature selection algorithms fall into two broad categories (Witten & Frank, 2000): the filter model or the wrapper model. The filter model (Dash, Choi, Scheuermann, & Liu, 2002) relies on general characteristics of a set of training data to select some features without involving any learning algorithm. The wrapper model (Kohavi & John, 1997) requires one predetermined learning algorithm for feature selection and uses its performance to evaluate and determine which features are selected. In these algorithms, first, a goodness measure of feature subsets based on data characteristics is used to choose best subsets for a given cardinality. Then, cross validation is exploited to identify the best subset across different cardinalities. These algorithms mainly focus on combining filter and wrapper algorithms to achieve the best possible performance for a particular learning algorithm with a similar time complexity of filter algorithms alone. In this work, we focus on the filter model and aim to develop a new feature selection algorithm that can effectively remove both irrelevant and redundant features.

Early research efforts mainly focused on feature selection for classification with labeled data (supervised feature selection) where class information is available (Jimenez & Landgrebe, 1997; Dash & Liu. 1997; Ng. 1998; W.

Siedlecki & J. Sklansky, 1988; Jain & Zongker, 1997; Han & Fu, 1996; J. Novovicova, Malik, & Pudil, 2004; Hastie, Tibshirani, & Friedman, 2000). The latest developments, however, show that the above general procedure can be well adopted to feature selection for clustering with unlabeled data (or unsupervised feature selection) where data is unlabeled (Mitra, Murthy, & Pal. 2002; Kim, Street, & Menczer, 2000; Dash & Liu. 1999; Dash, Liu, & Yao, 1997; Dy & Brodley, 2000; Yu & Liu. 2003)

2 PROBLEM DEFINITION

The set of techniques that can be employed for attribute reduction can be partitioned into two important types: techniques that apply to supervised or unsupervised learning and techniques that entail feature selection or feature extraction.

Feature extraction can be viewed as a preprocessing step that removes distracting variance from a dataset so that downstream classifiers or regression estimators perform better (Liu & Motoda, 1998). The area where feature extraction ends and classification or regression begins is necessarily murky: an ideal feature extractor would simply map the data to its class labels for the classification task.

Variable selection is a search problem with each state in the search space specifying a subset of the possible attributes of the task (Liu, Motoda, & Yu, 2002). Exhaustive evaluation of all variable subsets is usually intractable. Genetic algorithms, population-based learning, and related Bayesian methods have been commonly used as search engines for the variable selection process (Jimenez & Landgrebe, 1997). Particularly for SVMs, a variable selection method was introduced based on finding the variables that minimize bounds on the leave-one-out error for classification (Hastie, Tibshirani, & Friedman, 2000). The search of variable subsets can be efficiently performed by a gradient descent algorithm. The method, however, was limited to separable classification problems and thus is not directly applicable to the regression problems. Liu et al. (2002) proposed another variable selection method for classification by recursively eliminating the input variables that decrease the margin the least.

Existing discretization algorithms can classified as top-down or bottom-up, Top-down methods can be further divided into unsupervised or supervised (Liu, Hussain, Tan, & Dash, 2002). Experiments by Kurgan and Cios (2004) showed that the Class-Attribute Interdependence Maximization (CAIM) discretization algorithm is superior to other top-down discretization algorithms because its schemes can generally maintain the highest interdependence between target class and discretized attributes, resulting in the least number of generated rules and attaining the highest classification accuracy. Fast Class-Attribute Interdependence Maximization (FCAIM), which is an extension of the CAIM algorithm, has been proposed to speed up CAIM (Kurgan & Cios, 2003). The main framework, including the discretization criterion and the stopping criterion, as well as the time complexity between CAIM and F-CAIM, are all the same. The only difference is the initialization of the boundary point between the two algorithms. Compared to CAIM, F-CAIM is faster and had a similar C5.0 accuracy but obtained a slightly worse Class-Attribute Interdependence Redundancy (CAIR) value. As the main goal of our approach is to reach a higher CAIR value and attain an improvement in the accuracy of classification, we compared our approach to CAIM instead of F-CAIM in our experiments. Of course, Class-Attribute Contingency Coefficient (CACC) can be easily extended to Fast Class-Attribute Contingency Coefficient (F-CACC) with the same considerations as in F-CAIM if faster discretization is considered more important than the quality of a discretization scheme.

The computational complexity of bottom-up methods is usually larger than top-down ones, as they start with a complete list of all continuous values of the attribute as cut-points and then remove some of them by merging intervals in each step. Another common characteristic of these methods is in the use of the significance test to check if two adjacent intervals should be merged. ChiMerge (Kerber, 1992) is the most typical bottom-up algorithm. In addition to the problem of high computational complexity, the other main drawback of ChiMerge is that users have to provide several parameters during the application of this algorithm that include the significance level as well as the maximal and minimal intervals. Hence, Chi2(Liu, Huan and Setiono, Rudy,1995) was proposed, based on the ChiMerge. Chi2 improved ChiMerge by automatically calculating the value of the significance level. However, Chi2 still requires users to provide an inconsistency rate to stop the merging procedure and does not consider the freedom that would have an important impact on the discretization schemes. Thereafter, a Modified Chi2(Francis E.H. Tay, Lixiang Shen,2002) takes the freedom into account and replaces the inconsistency checking in Chi2 by the quality of approximation after each step of discretization. Such a mechanism makes Modified Chi2 a completely automated method to attain a better predictive accuracy than Chi2. After ModifiedChi2, Extended Chi2(Su, C. & Hsu, J. 2005) was developed to take into consideration that the classes of instances often overlap in the real world. Extended Chi2 determines the predefined misclassification rate from the data itself and considers the effect of variance in two adjacent intervals. With these modifications, Extended Chi2 can handle an uncertain dataset. Experiments on these bottom-up approaches by using C5.0 also showed that the Extended Chi2 outperformed other bottom-up discretization algorithms as its discretization scheme, on average, can reach the highest accuracy (Su & Hsu, 2005).

In recent years, datasets have become increasingly larger in both number of instances and number of features in many applications. This enormity of data may cause serious problems for many machine learning algorithms with respect to scalability and learning performance. For example, high dimensional data (i.e., data sets with hundreds or thousands of features) can contain a 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 facing high dimensional data. However, this trend of enormity of both size and dimensionality also poses severe challenges to feature selection algorithms. Some recent research efforts in feature selection have been focused on these challenges caused by handling a very large number of instances to deal with high dimensional data. Our work is concerned with feature selection for high dimensional data along with discretization of continuous values in an efficient manner. Feature selection is done using an improved Fast Correlation-based filter algorithm (FCBF$^+$).

3 RELATED WORK

Blum and Langley (1997) classified the feature selection techniques into three basic approaches. In the first approach, known as the embedded approach, a basic induction method is used to add or remove features from the concept description in response to prediction errors on new instances. The second approach is known as the filtering approach, in which various subsets of features are explored to find an optimal subset that preserves the classification. The third approach is known as wrapper methods, which evaluate alternative feature sets by running an induction algorithm on the training data and using the estimated accuracy of the resulting classifier as its metric. ID3 (Quinlans, 1986), C4.5 (Quinlan, 1993), and CART (Breiman, Friedman, Olshen, & Stone, 1984) are some of the most successful supervised learning algorithms. These algorithms use a greedy search through the space of decision trees, at each stage using an evaluation function to select the attribute that has the best ability to discriminate among the classes. Michalski (1980) proposed the AQ learning algorithm, which uses positive and negative examples of a class along with a user-defined criterion function to identify a disjunctive feature set that can maximize the positive events and minimize the negative events. Narendra and Fukunaga (1977) presented a Branch and Bound algorithm for finding the optimal feature set that uses a top-down approach with back-tracking. Pudil et al. (1994) proposed a set of suboptimal algorithms called the floating search methods that do not require the fulfillment of a monotonicity condition for feature selection criterion function. Somol et al. (2000) provided a modified and efficient branch and bound algorithm for feature selection. Though computationally less expensive than the branch-and bound algorithms, no theoretical upper bound exists on the computational costs of the algorithms because of their heuristic nature.

John et al. (1994) proposed another feature selection framework known as the wrapper technique. The wrapper methods evaluate alternative feature sets by running an induction algorithm on the training data and using the estimated accuracy of the resulting classifier as its metric. The major disadvantage of the wrapper methods is in the computational cost involved in running the induction algorithm repeatedly for each feature set considered.

A number of feature selection techniques based on evolutionary approaches have also been proposed. Casillas et al. (2001) present a genetic feature selection technique that is integrated into a multi-stage genetic learning process to obtain a Fuzzy Rule Based Classification system (FRBCS). In the first phase of this method, a filtering approach is used to determine an optimal feature subset for a specific classification problem using class-separability measures. This feature subset, along with expert opinion, is used to obtain the adequate feature subset cardinality in the second phase, which is used as the chromosome length. Xiong (2002) proposed a hybrid approach to input selection, which distinguishes itself from existing filter and wrapper-based techniques but utilizes the advantages of both. This process uses case-based reasoning to select candidate subsets of features, which are called ‘hypothesis.’ The performance of case-based reasoning under a hypothesis is estimated using training data on a “leave-one-out” procedure. The error estimate is then combined with the subset of selected attributes to provide an evaluation function for a genetic algorithm (GA) to find the optimal hypothesis. Kuncheva and Bezdek (1998) proposed a genetic algorithm for simultaneously editing and doing feature selection to design 1-nn classifiers. They had posed the problem as a bi-criteria combinatorial optimization problem having an NP-hard search space. Ho et al. (2002)

proposed the design of an optimal nearest neighbor classifier using an intelligent genetic algorithm. Thawonmas and Abe (1997) suggested a feature selection technique to eliminate irrelevant features based on an analysis of class regions generated by a fuzzy classifier. The degree of overlap in a class region is used to define the exception ratio, and the features that have the lowest sum of exception ratios are the relevant ones. Irrelevant features are eliminated using a backward selection search technique.

Kira and Rendell (1992) proposed a different approach to feature selection, and their proposed RELIEF algorithm assigns a weight to each feature based on the ability of the feature to distinguish among the classes, then selects those features whose weights exceed a user defined threshold as relevant. The weight computation is based on the probability of the nearest neighbors from two different classes having different values for an attribute and the probability of two nearest neighbors of the same class having the same value of the attribute. The higher the difference between these two probabilities, the more significant is the attribute. Inherently, the measure is defined for a two-class problem, which can be extended to handle multiple classes by splitting the problem into a series of two-class problems. Kononenko (1994) suggests the use of k-nearest neighbours to increase the reliability of the probability approximation. He also suggests how RELIEF can be extended to work with multiple sets more efficiently. Weighting schemes are easier to implement and are preferred for their efficiency.

Learning to classify objects is an inherently difficult problem for which several approaches, such as instance-based learning or nearest neighbor-based algorithms, are used. However, the nearest neighbor algorithms need some kind of distance measure. Cost and Salzberg (1993) emphasized the need to select appropriate metrics for symbolic values. Stanfill and Waltz (1984) proposed the Value Difference Metric (VDM), which measures distance between values of symbolic features. It takes into account the overall similarity of classification of all instances for each possible value of each feature. Based on this, they proposed the Modified Value Distance Metric (MVDM), which is symmetric and satisfies all the metric properties. They have shown that nearest neighbour algorithms perform well even for symbolic data using this metric. It is observed that distance-values are similar if the pairs occur with the same relative frequency for all classes. Zhao and Tsang (2008) proposed an attribute reduction with fuzzy approximation operators. Sharma and Paliwal (2008) proposed a rotational linear discriminate analysis technique for dimensionality reduction, which is a supervised learning technique that finds a linear transformation such that the overlap between the classes is minimum for the projected feature vectors in the reduced feature space.

4 PROPOSED WORK

This section presents a new algorithm named FCBF$^+$ that combines a feature selection algorithm named Fast Correlation Based Filer (FCBF) algorithm (Yu & Liu, 2003) with feature discretization (Tsai, Lee, & Yang, 2008) because discrete values are not handled in the earlier FCBF algorithm. Discretization algorithms have played an important role in data mining and knowledge discovery. They not only produce a concise summarization of continuous attributes to help experts understand the data more easily but also make learning more accurate and faster. Thus a static, global, incremental, supervised, and top-down discretization algorithm based on a Class-Attribute Contingency Coefficient is proposed. Evaluation of this algorithm on a real dataset shows that the proposed algorithm could generate a better discretization scheme that improves the accuracy of classification and clustering.

A correlation measure based on the information-theoretical concept of entropy is a measure of the uncertainty of a random variable. The entropy of a variable X is defined as

$$H(X) = -\sum_i P(x_i)\log_2(P(x_i)),$$

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

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)\log_2(P(x_i|y_j)),$$

where $P(x_i)$ is the prior probability for all values of X. $P(x_i|y_i)$ is the posterior probability of X, given a value of Y. The amount by which the entropy of X decreases reflects additional information about X provided by Y and is called the information gain, as given by

$$IG(X|Y) = H(X) - H(X|Y).$$

According to this measure, a feature Y is regarded as being 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 they are comparable and have the same aspect. Therefore, we choose a symmetrical uncertainty defined as follows.

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X)+H(Y)}\right]$$

This formula compensates for a bias of information gain toward features with more values and normalizes its values to the range [0; 1] with the value 1 indicating that knowledge of either value completely predicts the value of the other and the value 0 indicating that X and Y are independent.

Using symmetrical uncertainty (SU) as the goodness measure, a procedure is developed to select good features for classification based on correlation analysis of features.

```
1 Input: Dataset($A_1$,$A_2$,…$A_{N,}$ C) with i continuous attributes, M examples and S  target classes;
2 Begin
3 For each continuous attribute $A_i$
    4 Find the maximum dn and the minimum d0 values of $A_i$;
    5 Form a set of all distinct values of A in ascending order;
    6 Initialize all possible interval boundaries B with the minimum and maximum
    7 Calculate the midpoints of all the adjacent pairs in the set;
    8 Set the initial discretization scheme as D: {[d0,dn]} and  Globalcacc = 0;
    9 Initialize k = 1;
        10 For each inner boundary B which is not already in scheme D;
        11 Add it into D;
        12 Calculate the corresponding cacc value;
        13 Pick up the scheme D' with the highest cacc value;
        14 If cacc > Globalcacc, then
            15 Replace D with D';
            16 Globalcacc = cacc;
            17 k = k + 1;
            18 Goto Line 10;
        19 Else
            20 D' = D;
        21 End If
        22 for i = 1 to N do begin
            23 calculate $SU_{i,c}$ for Fi;
            24 append $F_i$ to $S_{list}'$;
```

```
      end;
  order $S_{list}'$ in descending $SU_{i,c}$ value;
  $F_P$ = getFirstElement($S_{list}'$’ $F_p$);
  do begin
    $F_P$ = getNextElement($S_{list}'$’ $F_p$);
    if ($F_q$ <> NULL)
      do begin
      $F_q'$ = $F_q$;
      if ($SU_{p,q}$ >= $SU_{q,c}$)
        remove F from $S_{list}'$;
        $F_q$ = getNextElement($S_{list}$‘, $F_q'$);
      else $F_q$ = getNextElement($S_{list}'$, $F_q$);
   end until ($F_q$ == NULL);
   $F_q$ = getNextElement($S_{list}$‘, $F_p$);
  end until ($F_p$ == NULL);
$S_{best}$ = $S_{list}'$;
end
```

Figure1. Proposed FCBF$^+$ algorithm: an updated procedure with data discretization

Thus the algorithm selects the features after discretizing the continuous attributes, in contrast to the FCBF algorithm in which continuous values are not handled. Given a dataset with i continuous attributes, M examples, and S target classes, for each attribute A_i, the algorithm first finds the maximum dn and minimum d0 and then sorts them in ascending order. The midpoints of all the adjacent boundaries are obtained. In the kth loop, cacc is computed for all possible cutting points to find the one with the maximum value and then this attribute is partitioned accordingly into k + 1 intervals. In other words, for every loop, cacc not only finds the best division point but also records a Globalcacc value. If the generated cacc value in loop k + 1 is less than the Globalcacc obtained in loop k, cacc would terminate and output the discretization scheme. The algorithm further processes the ordered list S_{list}' to remove redundant features and only keeps predominant ones among all the selected relevant features. The correlation between a feature F_i and the class C is predominant if there exists no F_j such that $SU_{j,i}$ >= $SU_{i,c}$. If there exists such F_j to a feature F_i, we call it a redundant peer to F_i. The iteration starts from the first element in S_{list}' and continues as follows. For all the remaining features, if F_p happens to be a redundant peer to a feature F_q, F_q will be removed from S_{list}'. After one round of filtering features based on F_p, the algorithm will put the current remaining feature right next to F_p as the new reference to repeat the filtering process. The algorithm stops until there are no more features to be removed from S_{list}'.

$$Y = [\sum_{i=1}^{s} \sum_{r=1}^{n} (Q_{ir}^{2}/M_{i+}M_{+r})-1]/\log(M)$$

$$cacc=\sqrt{y/y+M}$$

M is the total number of samples, n is the number of intervals, q_{ir} is the number of samples with class i (i = 1,2,. . . S, and r = 1,2,. . . n) in the interval (d_{r-1},d_r). M_{i+} is the total number of samples with class i, and M_{+r} is the total number of samples in the interval r-1 to r.

5 PERFORMANCE EVALUATION

In our experiments, we chose three representative feature selection algorithms for comparison with FCBF$^+$. One is a feature weighting algorithm, ReliefF (an extension to Relief), which searches for several nearest neighbors, is robust to noise, and handles multiple classes (Kononenko, 1994). In addition to feature selection algorithms, we also selected three different learning algorithms, EM, Farthest First, and K-Means algorithm, to evaluate the accuracy on selected features.

The experiments were conducted using Weka's implementation of all these algorithms, and FCBF$^+$ is also implemented in Weka environment (Witten & Frank, 2000). All together five data sets were selected from the UCI Machine Learning Repository (Blake & Merz, 1998). A summary of the data sets is presented in Table 1.

For each data set, we executed all four feature selection algorithms, FCBF$^+$, Relief, Chisquared, Principal components respectively, and recorded the number of selected features for each algorithm. We then applied EM, Farthest First, and K-Means algorithm on the original data set as well as with a reduced dataset containing only the selected features from FCBF$^+$ algorithm and recorded overall accuracy by 10-fold cross-validation.

DATASET	No of continuous attributes	No of attributes	No of classes	No of Instances
Diabetes	8	8	2	108
Iris	4	4	3	150
Chess	0	36	2	3196
Soybean	35	36	4	47
Liver	6	7	2	345

Table 1. Summary of bench-mark data sets

Feature Selection Algorithms	Chisquared Attribute evaluation	Relief Attribute Evaluation	Principal Components	FCBF$^+$
Diabetes	8	8	7	3
Iris	4	4	2	4
Chess	36	36	31	6
Soybean	35	35	18	5
Liver	6	6	5	2
Average	18	18	13	4

Table 2. Number of selected features for each feature selection algorithm

Table 2 summarizes the number of selected features for each feature selection algorithm. From the averaged values in the last row of Table 2, it is clear that FCBF$^+$ achieves the highest level of dimensionality reduction by selecting the least number of features, which is consistent with our theoretical analysis about FCBF$^+$s ability to identify redundant features.

6 PERFORMANCE COMPARISON

Performance evaluation is done using a software package called WEKA, which includes clustering algorithms. The clustering accuracy is determined by using the entire dataset as input for the package. Then the reduced set of features is tested with the same set of algorithms, and it is found that the clustering accuracy increases for the reduced set of features.

The farthest-point heuristic starts with an arbitrary point S1. Pick a point S2 that is as far from S1 as possible. Pick Si to maximize the distance to the nearest of all centroids picked so far. That is, maximize the min {dist (Si, S1), dist (Si, S2), ...}.

The k-means algorithm is an algorithm to cluster n objects based on attributes into k partitions, k < n. It is similar to the expectation maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in data. It assumes that the object attributes form a vector space. Its objective is to minimize total intra-cluster variance or the squared error function.

CLUSTERING ALGORITHMS	EM		FARTHEST FIRST		KMEANS	
	ALL	REDUCED	ALL	REDUCED	ALL	REDUCED
Diabetes	31.49	59.26	69	70.38	33	34
Iris	60	76	48	55	67	62
Soybean	78.73	100	57.45	57.45	57.45	57.45
Liver	45.51	49.57	56.24	58.48	54.21	54.5
Chess	47.32	51.54	53.79	58.48	53.32	56.79

Table 3. Clustering Accuracy of Various Algorithms (In %)

The performance improvement of the EM, Farthest First, and Kmeans algorithms is tabulated for two cases. The first case represents the complete full set of features, and the second case represents the reduced set of features. It is obvious from Table 3 that the performance of the algorithms with reduced subset doubled with certain datasets like Diabetes in particular using EM algorithm. In other cases also the performance improvement is considerable for various algorithms.

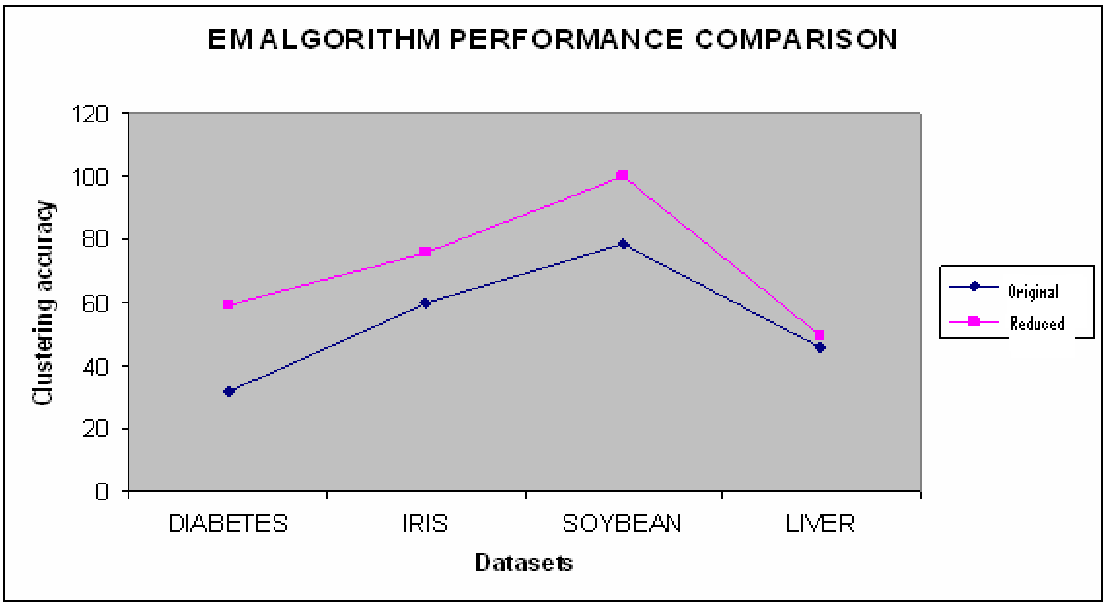

Figure 2. EM algorithm performance comparison

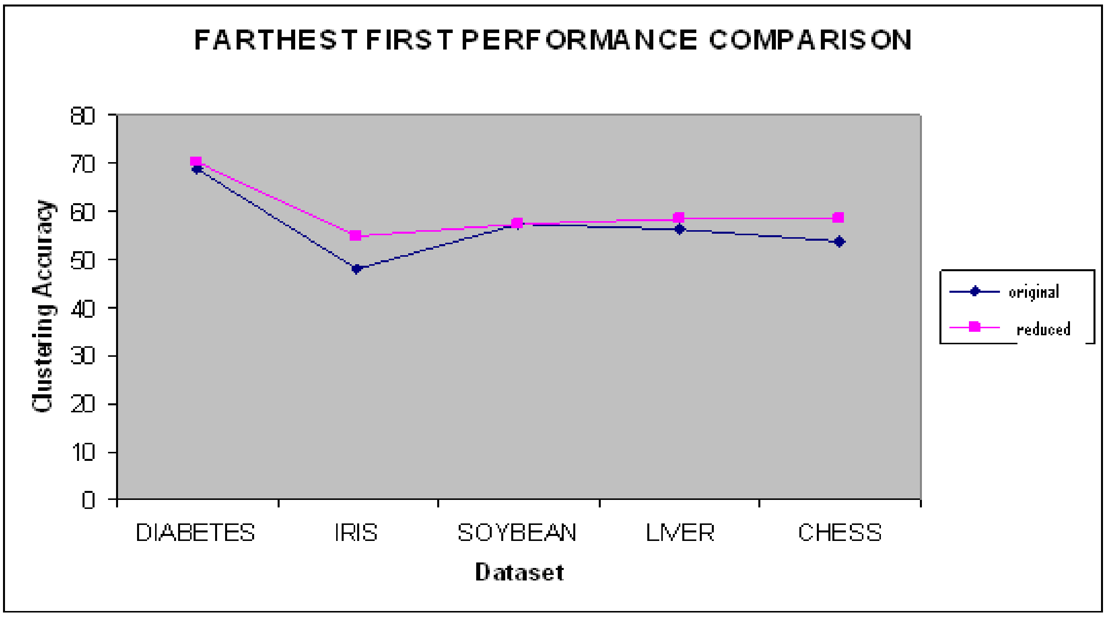

Figure 3. Farthest First algorithm performance comparison

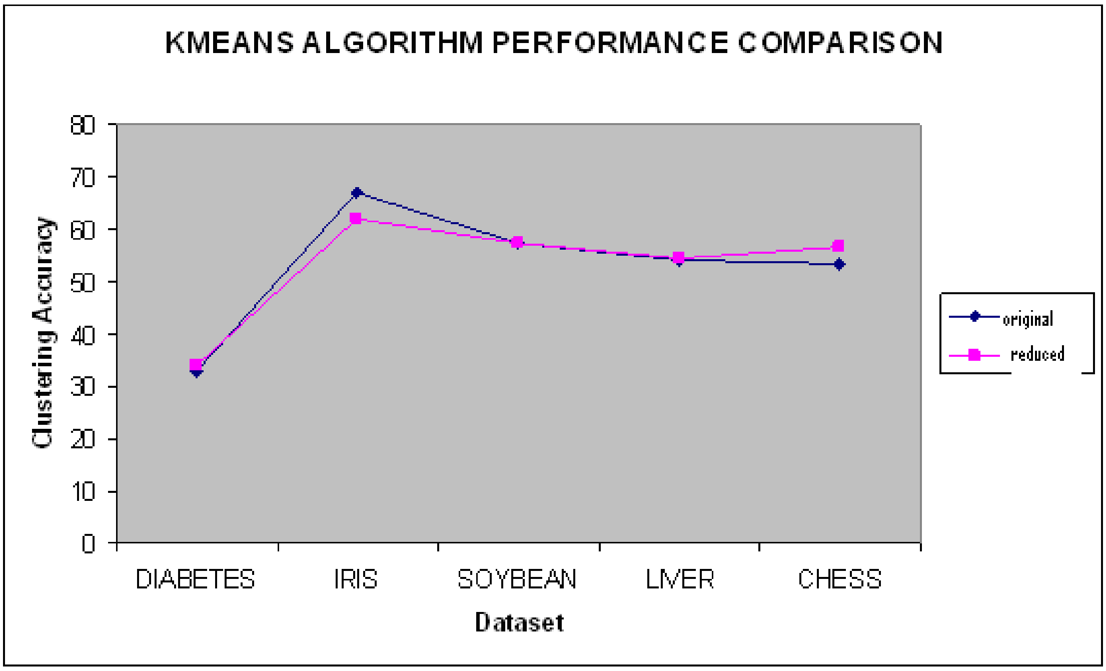

Figure 4. Kmeans algorithm performance comparison

The graphical performance of the various algorithms is illustrated through Figures 2 - 4. It is obvious from these graphs that the clustering accuracy with reduced feature subset is substantially high compared with the original full features set. It is also obvious from Figure 2 that FCBF$^+$ improves the clustering accuracy dramatically when the EM algorithm is used. Also the clustering accuracy is improved for Farthest First and K Means algorithms as shown in Figures 3 and 4. These results show that FCBF$^+$ is suitable for classification of high dimensional data.

The above experimental results suggest that FCBF$^+$ is practical for feature selection for classification of high dimensional data. It can efficiently achieve a high degree of dimensionality reduction and enhance classification accuracy with predominant features.

7 CONCLUSION

Because clustering results are dependent only on the data, the data measure, and the computing method, the data measure has to be chosen carefully. In many applications, correlation measures are preferred because of their favorable invariance properties, adding a constant offset to components of a data sample or applying a multiplicative factor does not affect correlation.

The work proposed here focuses on predominant correlation and introduces an efficient way of analyzing feature redundancy known as the Fast Correlation-Based Filter$^+$ approach for feature selection. It is evaluated through extensive experiments by comparison to related feature selection algorithms. The feature selection results are further verified by applying different clustering algorithms to data with and without feature selection. The FCBF$^+$ approach demonstrates its efficiency and effectiveness in dealing with high dimensional data for clustering with feature discretization.

8 REFERENCES

Ben-Bassat, M. (1982) Pattern Recognition and Reduction of Dimensionality. In *Handbook of Statistics-II*, Krishnaiah, P. & Kanal, L. (eds.), pp. 773-791. North Holland.

Blake & Merz (1998) UCI Repository of Machine Learning Databases. School of Information and Computer Sciences, Univ. of California, Irvine. Retrieved from the WWW, January 19, 2009: http://www.ics.uci.edu/~mlearn.

Blum, A. & Langley, P. (1997) Selection of Relevant Features and Examples in Machine Learning. *Artificial Intelligence, vol. 97*, pp. 245-271.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984) *Classification and Regression Trees*. Wadsworth: Belmont, CA.

Caruana, R. & Freitag, D. (1994) Greedy Attribute Selection. *Proc.11th Int'l Conf. Machine Learning*. pp. 28-36

Casillas, J., Cordon, O., Del Jesus, M., & Herrera, F. (2001) Genetic feature selection in a fuzzy rule-based classification system learning process for high-dimensional problems.*Informance Sci. 136 (August),* pp.135–157.

Cost, S., & Salzberg, S. (1993) In: *A Weighted Nearest Algorithm with Symbolic Features. Machine Learning, vol. 10,* pp. 57–78. Kluwer Publishers: Boston, MA.

Das, S. (2001) Filters, Wrappers and a Boosting-Based Hybrid for feature selection. *Proc. 18th Int'l Conf. Machine Learning,* pp. 74- 81.

Dash, M., Choi, K., Scheuermann, P., & Liu, H. (2002) Feature Selection for Clustering-a Filter Solution. *Proc. Second Int'l Conf. Data Mining,* pp. 115-122.

Dash, M. & Liu, H. (1997) Feature Selection for Classification. *Intelligent Data Analysis: An Int'l J., vol. 1, no. 3,* pp. 131-156.

Dash, M. & Liu, H. (1999) Handling Large Unsupervised Data via Dimensionality Reduction. *Proc. SIGMOD Research Issues in Data Mining and Knowledge Discovery Workshop*.

Dash, M., Liu, H., & Yao, J. (1997) Dimensionality Reduction of Unsupervised Data. *Proc. Ninth IEEE Int'l Conf. Tools with AI (ICTAI '97)*, pp. 532-539.

Doak, J. (1992) *An Evaluation of Feature Selection Methods and Their Application to Computer Security*. Technical report, Univ. of California at Davis, Dept. Computer Science.

Dy, J. & Brodley, C. (2000) Feature Subset Selection and Order Identification for Unsupervised Learning. *Proc. 17th Int'l Conf. Machine Learning*, pp. 247-254.

Francis E.H. Tay, Lixiang Shen(2002) A Modified Chi2 Algorithm for Discretization. IEEE Transactions on Knowledge and Data Engineering, Volume 14, Issue 3,pp.666 - 670

Hall, M. (2000) Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning. *Proc. 17th Int'l Conf. Machine Learning,* pp. 359-366.

Han, J. & Fu, Y. (1996) Attribute-Oriented Induction in Data Mining. In *Advances in Knowledge Discovery and Data Mining.* Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (eds.), pp. 399-421, AAAI Press/The MIT Press.

Han, J. & Kamber, M. (2005) *Data Mining: Concepts and Techniques*. Morgan Kaufman.

Hastie, T., Tibshirani, R., & Friedman, J. (2000) *The Elements of Statistical Learning Data Mining, Inference, and Prediction,Springer Series in Statistics*. Springer Verlag, New York.

Ho, S., Liu, C., & Liu, S. (2002) Design of an optimal nearest neighbor classifier using an intelligent genetic algorithm. *Pattern Recognition Lett. 23 (13*), pp.1495–1503.

Jain, A. & Zongker, D. (1997) Feature Selection: Evaluation, Application, and Small Sample Performance. *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 2*, 153-158.

Jimenez, L. & Landgrebe, D. (1997) Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man and Cybernetics, 28(1)*:39–54.

John, R., Kohavi, R., & Pfleger, K. (1994) Irrelevant Feature and the Subset Selection Problem. *Proc. 11th Int'l Conf. Machine Learning*, pp. 121-129.

Kerber, R. (1992) ChiMerge: discretization of numeric attributes. *Proceedings of the Ninth International Conference on Artificial Intelligence*, pp. 123–128.

Kim, Y., Street, W., & Menczer, F. (2000) Feature Selection for Unsupervised Learning via Evolutionary Search. *Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, pp. 365-369.

Kira, K. & Rendell, L. (1992) A practical approach to feature selection. In: *Proc. Ninth Internat. Conf. on Machine Learning.* Aberdeen, Scotland, pp. 249–256.

Kira, K. & Rendell, L. (1992) The Feature Selection Problem: Traditional Methods and a New Algorithm. *Proc. 10th Nat'l Conf. Artificial Intelligence*, pp. 129-134.

Kohavi, R. & John, G. (1997) Wrappers for Feature Subset Selection. *Artificial Intelligence, vol. 97, nos. 1-2,* pp. 273-324.

Kononenko, I. (1994) Estimating attributes: Analysis and extensions of RELIEF. In: *Proc. of Eur. Conf. on Machine Learning.*

Kuncheva, L. & Bezdek, J. (1998) Nearest prototype classification: Clustering, genetic algorithms or random search. *IEEE Trans. Systems Man Cybernet. C 28 (1),* pp.160–164.

Kurgan, L. & Cios, K. (2003) Fast Class-Attribute Interdependence Maximization (CAIM) Discretization Algorithm. In: *Proceedings of the International Conference on Machine Learning and Applications*, pp. 30–36.

Kurgan, L. & Cios, K. (2004) CAIM discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering 16 (2)* 145–153.

Liu, H., Hussain, F., Tan, C., & Dash, M. (2002) Discretization: an enabling technique. *Journal of Data Mining and Knowledge Discovery 6(4)* 393–423.

Liu, H. & Motoda, H. (1998) *Feature Extraction, Construction and Selection: A Data Mining Perspective, 2nd ed.* Boston: Kluwer Academic.

Liu, H. & Motoda, H. (1998) *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic.

Liu, H., Motoda, H., & Yu, L. (2002) Feature selection with selective sampling. *Proceedings of the Nineteenth International Conference on Machine Learning,* pp. 395 - 402.

Liu, H. & Setiono, R. (1996) A Probabilistic Approach to Feature Selection-A Filter Solution. *Proc. 13th Int'l Conf. Machine Learning,* pp. 319-327

Liu, Huan and Setiono, Rudy (1995) Chi2: Feature Selection and Discretization of Numeric Attributes**.** In Proceedings of the 7th IEEE International Conference on Tools with Artificial Intelligence

Michalski, R. (1980) Pattern recognition as rule-guided inductive learning. *IEEE Trans. Pattern Anal. Machine Intell. 2 (4),* pp. 349–361.

Miller, A. (2002) *Subset Selection in Regression, 2nd ed*. Chapman & Hall/CRC.

Mitra, P., Murthy, C., & Pal, S. (2002) Unsupervised Feature Selection Using Feature Similarity. *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 3*, pp. 301-312.

Narendra, P. & Fukunaga, K. (1977) A branch and bound algorithm for feature subset selection. *IEEE Trans. Comput .c-26 (9)*, pp.917–922.

Ng, A. (1998) On Feature Selection: Learning with Exponentially Many Irrelevant Features as Training Examples. *Proc. 15th Int'l Conf. Machine Learning*, pp. 404-412.

Novovicova, J., Malik, A., & Pudil, P. (2004) Feature selection using improved mutual information for text classification. Volume 3138 of *Lecture Notes in Computer Science*, pp. 1010 -1017. Springer.

Pudil, P., Novovicova, J., & Kittler, J. (1994) Floating search methods in feature selection. *Pattern Recognition Letters. 15 (11)*, pp.1119–1125.

Pyle, D. (1999) *Data Preparation for Data Mining*. Morgan Kaufmann Publishers.

Quinlan, J. (1986) Induction of decision trees. *Machine Learning 1,* pp.81–106.

Quinlan, J. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann: San Francisco.

Raymer, J., Punch, W., Goodman, E., Kuhn, L., and Jain, A. (2000) Dimensionality reduction using genetic algorithms. *IEEE Transactions on Evolutionary Computation, 4*:164–171.

Sharma, A. & Paliwal, K. (2008) Rotational Linear Discriminate Analysis Technique for Dimensionality Reduction *IEEE Transactions on Knowledge and Data Engineering,Vol.20, No.10*, pp.1336-1347.

Siedlecki, W. & Sklansky, J. (1988) On Automatic Feature Selection. *Int'l J. Pattern Recognition and Artificial Intelligence, vol. 2,* pp. 197- 220.

Somol P., Pudil P., Ferri F. J., Kittler J.: Fast Branch & Bound algorithm in feature selection. In: *Proceedings of SCI 2000. The 4th World Multiconference on Systemics, Cybernetics and Informatics.* (Sanchez B., Pineda M. J., Wolfmann J. eds.). IIIS, Orlando 2000, pp. 646-651.

Stanfill, C. & Waltz, D. (1986) Towards memory based reasoning. *Comm. ACM 29 (12),* pp. 1213–1228.

Su, C. & Hsu, J. (2005) An extended chi2 algorithm for discretization of real value attributes. *IEEE Transactions on Knowledge and Data Engineering 17 (3)* 437–441.

Swets, D. & Weng, J. (1995) Efficient Content-Based Image Retrieval Using Automatic Feature Selection. *IEEE Int'l Symp. Computer Vision,* pp. 85-90.

Talavera, L. (1999) Feature Selection as a Preprocessing Step for Hierarchical Clustering. *Proc. Int'l Conf. Machine Learning,* pp. 389-397.

Thawonmas, R. & Abe, S. (1997) A novel approach to feature selection based on analysis of class regions. *IEEE Trans Systems Man Cybernet. 27 (2*), pp. 196–207.

Witten, I. & Frank, E. (2000) *Data Mining: A Practical Machine Learning Tool with Java Implementation.* Morgan Kaufmann Publishers: San Francisco, California.

Wyse, N., Dubes, R., & Jain, A. (1980) A Critical Evaluation of Intrinsic Dimensionality Algorithms. *Pattern Recognition in Practice*, Gelsema, E. & Kanal, L. (eds.), pp. 415-425, Morgan Kaufmann, Inc.

Xing, E., Jordan, M., & Karp, R. (2001) Feature Selection for High-Dimensional Genomic Microarray Data. Proc. 15th Int'l Conf. Machine Learning, pp. 601-608.

Xiong, N. (2002) A hybrid approach to input selection for complex processes. *IEEE Trans. Systems Man Cybernet. Part A 32 (4)*, pp. 532–536.

Yu, L. & Liu, H. (2003) Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. *Proc. 20th Int'l Conf. Machine Learning*, pp. 856-863.

Yu, L. & Liu, H. (2004) Redundancy Based Feature Selection for Microarray Data. *Proc. 10th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*.

Zhao, S. & Tsang, E. (2008) On fuzzy approximation operators in attribute reduction with fuzzy rough sets, *Information Sciences, Volume 178, Issue 16*, pp. 3163-3176.

C. J. Tsai, C. I. Lee, and W. P. Yang,(2008) A Discretization Algorithm Based on Class-Attribute Contingency Coefficient. *Information Sciences*, accepted and will appear in vol. 178, pp. 714-731.

(Article history: Received 15 October 2008, Accepted 7 February 2009, Available online 24 April 2009)