An Effective Attribute Clustering Approach for Feature Selection and Replacement

Tzung-Pei Hon, Po-Cheng Wang and Yeong-Chyi Lee

Department of Computer Science and Information Engineering, National University of Kaohsiung Kaohsiung, Taiwan.
Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan
Department of Information Management Cheng Shiu University, Kaohsiung, Taiwan
tphong@nuk.edu.tw, m963040041@student.nsysu.edu.tw, yeongchyi@csu.edu.tw

Abstract

Feature selection is an important pre-processing step in mining and learning. A good set of features can not only improve the accuracy of classification, but also reduce the time to derive rules. It is executed especially when the amount of attributes in a given training data is very large. In this paper, an attribute clustering method based on genetic algorithms is proposed for feature selection and feature replacement. It combines both the average accuracy of attribute substitution in clusters and the cluster balance as the fitness function. Experimental comparison with the k-means clustering approach and with all combinations of attributes also shows the proposed approach can get a good trade-off between accuracy and time complexity. Besides, after feature selection, the rules derived from only the selected features may usually be hard to use if some values of the selected features cannot be obtained in current environments. This problem can be easily solved in our proposed approach. The attributes with missing values can be replaced by other attributes in the same clusters. The proposed approach is thus more flexible than the previous feature-selection techniques.

Keywords: feature selection, feature clustering, genetic algorithms, k-means, reduct.
1. **Introduction**

Feature selection is a very important technique to the problems of classification and retrieval. One of the famous approaches is the concept of “reduct”, which came from the rough set theory (Pawlak 1982, Pawlak 1996). Finding reducts is, however, an NP-hard problem (Blum and Rivest 1992, Skowron and Rauszer 1984). Thus some methods for finding out approximate reducts are proposed. For example, Al-Radaideh et al. used the discernibility matrix and a weighting strategy to derive reducts with a greedy method (Al-Radaideh et al. 2005). Sun and Xiong proposed an approach based on genetic algorithms to deal with the incomplete information systems (Sun and Xiong 2003). Besides, Gao et al. proposed a feature ranking strategy (similar to the attribute weighting) with a sampling process included (Gao et al. 2005).

Reducts provide conveniences and efficiencies in many applications. They may have the advantage of taking less cost for the classification process than the whole set of features. However, there are some drawbacks resulting from reducts. For example, only the features in a reduct are usually considered for classification or for retrieval in applications. But unexpected or uncertain events may happen occasionally. For example, the machines or the sensors may handle or detect an object with noises or missing values, which may happen to be from the attributes in the reduct. The reduct is then not useful in this situation for these objects since some values are unavailable.
In the past, Hong and Liou proposed a feature selection approach based on the concept of feature clustering (Hong and Liou 2007). It could not only find out an approximate reduct for classification but also cluster the attributes with high similarity into the same cluster. Its advantage was that attributes could be easily replaced by others in the same cluster if some attribute values were missed.

Furthermore, genetic algorithms (GAs) (Goldberg 1989, Grefenstette 1986) have become increasingly important for researchers in solving difficult problems since they could provide feasible solutions in a limited amount of time (Homaifar et al. 1993). They are adaptive heuristic search algorithms derived from the evolutionary ideas of natural selection and genetics. Wróblewski used the genetic algorithm to find approximate reducts (Wróblewski 1995), thus spending less time than that to find correct reducts. In this paper, we thus propose a new attribute clustering approach based on genetic algorithms for feature selection. Hong and Liou’s previous approach considered the dependency relation between attributes as the criterion of clustering. The proposed GA-based approach, however, would like to cluster attributes into the same group if they have high degrees of substitution. We used the accuracy of subsets of attributes on training data for measuring the degree of substitution. The proposed approach will thus remain a better clustering result from the viewpoint of classification accuracy than Hong and Liou’s since the clustering based on genetic
algorithms will tend to find a (nearly) optimal solution according to the criteria of the adopted fitness function, which is accuracy in the paper. The proposed feature clustering approach is based on genetic algorithms rather than on $k$-means because the latter is sensitive to the selection of the initial partition and easily converge to a local optimum if the initial partition is not properly chosen.

The remaining parts of this thesis are organized as follows. Some related researches are reviewed in Chapter 2. The proposed method based on genetic algorithms for attribute clustering and feature selection is described in Chapter 3. The experimental results with discussion are then given in Chapter 4. The conclusions and future works are stated in Chapter 5.

2. Review of Related Works

In this chapter, we review some related researches about this paper. They are the $k$-means clustering approach, the reduct, and genetic algorithms.

2.1 The $K$-means clustering

The $k$-means clustering approach is the simplest and most commonly used algorithm among the clustering techniques (MacQueen et al. 1967, Selim and Ismail 1984). The $k$-means algorithm consists of four steps. It first randomly chooses $k$ objects to represent the $k$ centers of the $k$ groups, where $k$ is predefined. It then assigns
each data or object into the cluster with the minimum distance from it among all the $k$
groups. The center of each cluster is then recomputed by averaging the objects in the
cluster. At last, the second and the third steps are repeated until the convergence
criterion is met.

2.2 The reduct

Given a dataset, a subset of attributes may be identified as equivalent to another
subset of attributes if the two subsets have the same classification performance on the
dataset. In this case, only one subset from the two needs to be used to represent the
entire data set for saving computation time but getting the same results. The other
subset is then redundant and can be removed since they cannot make the classification
better. There sometimes exist several subsets of attributes which can represent the
original data set with the same classification accuracy. The one among those subsets
with a minimal number of attributes are called reducts or minimal reducts (Han et al.
2005, Li et al. 2006, Wróblewski 1995). Finding a minimal reduct is an NP-hard
problem. This means that it is a non-trivial task and cannot be solved simply by an
increase of computational resources. This is the main bottleneck of the rough-set
methodology.
2.3 Genetic Algorithms

Since the 1960s, there has been much interest in developing powerful algorithms for difficult optimization problems. Some nature-inspired approaches were then proposed. An approach now in common usage is evolutionary computation, in which the genetic algorithms are a well-known technique.

Genetic Algorithms (GAs) (Holland 1975) have become increasingly important for researchers in solving difficult problems since they could provide feasible solutions in a limited amount of time (Homaifar et al. 1993). They have been successfully applied to the fields of optimization, machine learning (Michalewicz 1994), neural network (Mitchell 1996), fuzzy logic controllers (Sanchez et al. 1997), and so on. GAs are developed mainly based on the ideas and the techniques from genetic and evolutionary theory (Grefenstette 1986). According to the principle of survival of the fittest, they generate the next population by several operations, with each individual in the population representing a possible solution. In general, a genetic algorithm has five basic components, as summarized by Michalewicz (Michalewicz 1996):

1. a genetic representation of solutions to the problem,
2. a way for generating the initial population,
3. an evaluation function for measuring goodness of solutions, and
4. genetic operators that alter the genetic composition of children, and

5. parameter values.

3. The Proposed GA-Based Clustering Method

In this chapter, the GA-based clustering algorithm is proposed. In this algorithm, we encoded the result of clustering into a chromosome, and used GA to drive the best one. The fitness of each chromosome is evaluated by both of the average accuracy of attribute substitution in clusters and the cluster balance.

3.1 Chromosome representation

In this paper, each possible clustering solution is encoded as a chromosome and handled as an individual by the representation scheme of positive integer numbers. Assume there are $N$ attributes for classification. Each chromosome then consists of a sequence of $N$ positive integer numbers, with the value in the $i$-th position representing the cluster number of the $i$-th attribute. If the attributes will be divided into $K$ clusters, each integer thus lies among 1 to $K$. The value $k$ for a gene represents the $k$-th cluster which the corresponding attribute belongs to. Below, a simple example is given to illustrate the above idea.
3.2. Initial population

A genetic algorithm requires a population of feasible solutions to be initialized and then updated during the evolution process. As mentioned above, each individual within a population is a possible clustering result. An integer number among 1 to \(K\) (the given cluster number) is then randomly assigned to each gene of a chromosome. The chromosome generated in this way is thus feasible. This process is then repeated to generate each chromosome in the population.

3.3. Fitness and selection

In order to develop a good result of attribute clustering from an initial population, the proposed algorithm selects *parent* chromosomes with high fitness values for mating. A good evaluation (fitness) function is thus needed to achieve the purpose. The proposed fitness function consists of two factors, cluster accuracy and cluster balance. They are described as follows.

The cluster accuracy is used to evaluate the accuracy of a possible clustering result on the given training data. Since one purpose of the proposed attribute-clustering approach is to reduce the adopted attribute number in approximate reasoning, a reasonable criterion for clustering results is thus the average accuracy of each attribute combination, which is composed of an attribute from each cluster. Therefore for \(K\) clusters, one attribute from each cluster is selected and all the \(K\)
selected attributes gathered together form an attribute subset \( S \). If the accuracy of the set \( S \) is high, then \( S \) is suitable for classification. Consider all the possible combinations from a chromosome. If all the combinations are of high accuracy, then any attribute in each cluster can be chosen to form the attribute set. That means the clustering result is good for classification. In this case, if an attribute value of a data is missing or unavailable due to cost, it can be easily replaced with another attribute in the same cluster for classification. The replaced attribute subset can be expected to have a high accuracy for classification as well if the clustering result is good. The average accuracy of all the possible attribute combinations in a chromosome thus provides a reasonable measure for its goodness. Formally, the accuracy of a chromosome \( C_i \) is defined as follows:

\[
\text{accuracy}(C_i) = \sum_{p=1}^{NC} \frac{\text{accuracy}(S_p)}{NC},
\]

where \( NC \) is the number of attribute combinations resulting from \( C_i \) and \( \text{accuracy}(S_p) \) is the accuracy of the \( p \)-th possible attribute combination \( S_p \) for the given training dataset.

Another evaluation criterion for the goodness of a clustering result is the cluster balance. Assume in a classification problem, an attribute subset with only less than \( K \) (say \( R \)) attributes can classify the dataset with a high accuracy. If each of the \( R \) attributes is assigned to a cluster and the remaining attributes are assigned to the other
$K-R$ clusters, then it will cause the highest accuracy among all the possible combinations. The clustering result in this way is, however, unbalanced although it has high accuracy. In this case, a new object with missing values may not be classified since no other alternative attributes can be used in the single-attribute clusters.

The factor of cluster balance is thus important and designed here to avoid the above unbalanced situation. Formally, the factor of cluster balance for a chromosome $C_i$ is defined as follows:

$$balance(C_i) = \sum_{i=1}^{K} -\frac{|\text{cluster}_i|}{N} \log \frac{|\text{cluster}_i|}{N},$$

where $|\text{cluster}|$ represents the attribute number in the $i$-th cluster. The measure of balance is mainly based on the principle of entropy. If a clustering result is more balanced, then its value will be larger. When the balance measure is considered in the fitness function, the unsuitable clustering result can thus be sieved out from the population.

According to the above discussion, the fitness function thus takes the two factors into consideration. The first factor is the average accuracy of a chromosome and the second one is the cluster balance. It is then defined as follows:

$$f(C_i) = \text{accuracy}(C_i) \cdot [balance(C_i)]^\alpha,$$

where the parameter $\alpha$ is used to control the trade-off between the two factors. When $\alpha$ is set at 0, only the average accuracy is considered in determining the clustering
results. On the other hand, if $\alpha$ is set at a value larger than 0, then the cluster balance is considered. It can be defined by users according to the requirements of applications.

3.4. Genetic Operators

The two commonly-seen genetic operators, the *uniform crossover* and the *multi-point mutation*, are used in the proposed genetic attribute-clustering approach. The *uniform crossover* operator defines a binary mask sequence to decide which genes in the two parent chromosomes will be changed. The binary value 1 in the mask represent the values of the two parent chromosomes in the corresponding position are exchanged and the value 0 represents they do not. The value of each binary bit in a mask sequence is randomly set to 1 according to a given probability. The adopted multi-point mutation operator will exchange a gene value (representing the cluster ID of the corresponding attribute) into a new one among 1 to $K$ with a mutation probability.

4. Experimental Results

In this section, the experiments for showing the performance of the proposed algorithm for clustering attributes were described. They were done for different cluster numbers. One virtual dataset and one real-world data set, the Single Proton Emission Computed Tomography (SPECT) (Kurgan et al. 2001), were used in the
experiments. Experiments were implemented in C++ on an Intel Centrino Duo Core T2450 laptop with 2GHz CPU and 2GB RAM.

Experiments were made to show the accuracy of the classification by our proposed clustering approach. Firstly, the virtual data set was used for our experiments. The database contained 569 objects, 29 attributes and 1 decision attribute with 10 classes. The accuracy was evaluated for different group numbers. The initial population size $P$ was set at 20, the crossover rate $p_c$ was set at 0.5, and the mutation rate $p_m$ was set at 0.01. Besides, the best clustering results by $k$-means were found for comparison.

The highest accuracy from our approach and from the $k$-mean approach for the group number set at 3, 4 and 5 was shown in Figure 1. The highest accuracy from all the possible combinations of attributes for the group number set at 3, 4 and 5 was also found as the reference value.

![Figure 1. The comparison of the highest accuracy for the virtual data set.](image)

In the experiments, when the numbers of clusters were set at 3 and 4, there were
no reducts, but the best subset with 5 attributes was the complete reduct. In other words, the number of attributes for the complete reduct for the given data set was 5.

Figure 2 then showed the average accuracy obtained by our method, $k$-means and all possible combinations.

![Figure 2](image)

**Figure 2.** The comparison of the average accuracy for the virtual data set.

For $K = 3$, the average accuracy by all of the possible combinations was 69.4%, the average accuracy from our proposed approach was 72.8%, and the average accuracy from $k$-means was 70.7%. When the number of clusters was set at 4 or 5, the average accuracy and the highest accuracy by our proposed method were also higher than those from all of the possible combinations. It meant that if the attributes are randomly chosen, the accuracy from our method was usually better than that from randomly chosen attribute subsets. Figure 3 then showed the experimental results for the worst attribute subsets from the three approaches.
Figure 3. The comparison of the worst accuracy for the virtual data set.

For $K = 5$, our method had the worst accuracy of 87.6% for classification, better than $k$-means had (74.9%). The worst accuracy from all the combinations of the attributes was 36.9%. It meant that most of the worse combinations of the attributes for classifying are effectively filtered out by the proposed clustering algorithm. It was showed from the figures that by the proposed approach, when an attribute was replaced by another attribute in the same cluster, the accuracy of the new attribute subset for classifying the objects still remained a good accuracy.

The second data set used in the experiments was the Single Proton Emission Computed Tomography (SPECT). The database contained 267 objects, 44 attributes and 1 decision attribute with 2 classes. The results for the comparison of the highest, the average and the worst accuracy obtained by our proposed method, by $k$-means, and by all possible combinations were shown in Figures 4 to 6.
For the data set, the reduct could be found to contain three attributes since its highest accuracy was 100% when all possible combinations were considered. For $K = 3$, the best accuracy obtained by our proposed method was also 100%, meaning the corresponding subset was a reduct as well. In other words, the reduct could be found...
by our proposed algorithm. The other results are similar to those for the virtual data set.

5. Conclusions

In this paper, an attribute clustering method based on genetic algorithm is proposed for feature selection and feature replacement. The contributions of this paper can be summarized into the following four points.

1. A new method for feature clustering and feature selection is proposed.

2. After the attributes are partitioned into several clusters, any attribute with a missing value can be replaced by an attribute within the same group.

3. It provides a real-time processing mechanism for classification even if there are some unavailable attributes or attribute values.

4. Through attribute clustering, it can also help guess missing values of attributes.

The experimental results also show the above results. Therefore, the proposed approach provides another alternative for feature selection.

References

Blum, A. L. and Rivest, R. L. “Training a 3-node neural networks is NP-complete,”


"Knowledge Discovery Approach to Automated Cardiac SPECT Diagnosis,”


Addison Wesley, 1989.

Grefenstette, J. J. “Optimization of control parameters for genetic algorithms,” IEEE

Gao, K., Liu, M., Chen, K., Zhou, N. and Chen, J. “Sampling-based tasks scheduling
in dynamic grid environment,” The Fifth WSEAS International Conference on

Holland. J. H. Adaptation in Natural and Artificial Systems, University of Michigan

Homaifar, A., Guan, S. and Liepins, G. E. “A new approach on the traveling salesman
problem by genetic algorithms,” The Fifth International Conference on Genetic

Han, J., Hu, X. and Lin, T. Y. “Feature selection based on rough set and information
entropy,” The IEEE International Conference on Granular Computing 1 (2005):
153-158.

Hong, T. P. and Liou, Y. L. "Attribute Clustering in High Dimensional Feature
Spaces," The International Conference on Conference on Machine Learning

in building CBR classifiers,” IEEE Transactions on Knowledge and Data


