A Greedy Correlation Measure Based Attribute Clustering Algorithm for Gene Selection

Jiucheng Xu, Yunpeng Gao*, Shuangqun Li, Lin Sun, Tianhe Xu, Jinyu Ren
College of Computer and Information Technology, Henan Normal University, Xinxiang, China
*Corresponding author email: setfire1@163.com

***Abstract*—This paper proposes an attribute clustering algorithm for grouping attributes into clusters so as to obtain meaningful modes from microarray data. First the problem of attribute clustering is analyzed and neighborhood mutual information is introduced to solve it. Furthermore, an attribute clustering algorithm is presented for grouping attributes into clusters through optimizing a criterion function which is derived from an information measure that reflects the correlation between attributes. Then, by applying this method to gene expression data, meaningful clusters are discovered which assists to capture aspects of gene association patterns. Thus, significant genes containing useful information for gene classification and identification are selected. In the following, the proposed algorithm is employed to six gene expression data sets and a comparison is made with several well-known gene selection methods. Experiments show that the greedy correlation measure based attribute clustering algorithm, noted as GCMACA, is more capable of discovering meaningful clusters of genes. Through selecting a subset of genes which have a high significant multiple correlation value with others within clusters, informative genes can be acquired and gene expression of different categories can be identified as well.**

***Index Terms*—attribute clustering, gene selection, neighborhood mutual information, correlation, significant multiple correlation**

I. INTRODUCTION

In the last few years, the study of the transcriptome has made great progress thanks to the development of microarray technology. Today the number of scientific projects that include studies based on this possibility to measure simultaneously thousands of gene expressions across collections of samples is increasing dramatically [1-9,32]. The task of sample classification in the context of microarray is a major challenge.

Clustering is a main task of explorative data mining and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics, etc [2]. When applied to gene expression data, conventional clustering algorithms may encounter a problem which is a huge number of genes (attributes) versus a small number of samples [3]. Intuitively, attributes in a cluster are more correlated with each other, whereas attributes in different clusters are less interdependent. Attribute clustering is capable of reducing the search dimension of data mining algorithm, so as to discover interesting relationships or construct models in a tightly correlated subset of attributes instead of entire attribute space, which makes the algorithm more effective. After that, a smaller number of attributes are selected for further study.

Classification is an important task in gene expression data mining. Classification is concerned with assigning memberships to samples based on expression patterns and refining existing ones [4]. The dimension problems are tricky issue for clustering or patterns recognition. Typically, gene expression data sets consist of a large amount of genes but a small number of samples. Many data mining algorithms (e.g., classification [2,3,6,10-20,27-29], association rule mining [16], pattern discovery [15,21-24,32], and linguistic summaries [25], are developed and optimized to be scalable with respect to the number of tuples, so as to not handle a large number of attributes.

Various algorithms have been used for applying existing clustering algorithms to genes. Well-known examples are: k-means algorithms [17], Kohonen's self-organizing maps (SOM) [26], and various hierarchical clustering algorithms [4,15]. As for distance measures, Euclidean distance and Pearson's correlation coefficient are widely used for clustering genes [4].

A gene expression data set derived from a microarray can be described by an expression table $T = \left\{ t_{ij} \mid i = 1, \cdots, n; j = 1, \cdots, p \right\}$, where $t_{ij} \in \Re$ is the measured expression level of gene g_j in the sample s_i. Each row in the table corresponds to one sample and each column to a gene. A gene data set is typically composed of a large amount of genes, but a small number of samples. The distinctive characteristic of gene expression data allows clustering both genes and samples [4,5]. Generally speaking, Euclidean distance and Pearson's correlation coefficient are widely used as the correlation measure for clustering. However, for measuring the correlation between genes, Euclidean distance is not effective enough to describe functional similarity such as positive or negative correlation in values. Thus, Pearson's correlation coefficient [30-32] is put forward by some researchers. Empirical studies have shown that it may assign a high similarity score to a pair of dissimilarity genes. Au et al. constructed an attribute clustering

doi:10.4304/jcp.8.4.951-959

method grouping attributes with an information measure which obtained fairly good results [6]. However, most clustering methods are not able to effectively cope with continuous attributes, which is also a distinctive characteristic of gene expression data. When applied to the continuous attributes, conventional methods commonly discretize the continuous data into a finite number of intervals for data mining. But discretization may lead to information loss [7]. Furthermore, having so many genes related to so few samples is likely to result in the discovery of irrelevant patterns [8]. A useful technique to deal with it is to select a small number of the most promising genes and use them solely to build modes. To select genes, t-value is widely used [9]. However, it is necessary to note that t-value can only be used when the samples are preclassified. If no class information is provided, it cannot be used for gene selection [6]. In this paper, we propose an algorithm which incorporates correlation to obtain both superior classification and better performance.

The remainder of the paper is organized as follows. Section Ⅱ reviews some related work. In section Ⅲ, we define the problem of attribute clustering and then present GCMACA to address it. To evaluate our proposed algorithm's performance, we apply it to six gene expression data sets. The experimental results are presented in section Ⅳ, which validate the efficiency of the proposed approach. In section Ⅴ, we conclude this paper and discuss the future work.

II. Related Work

A. *The Attribute Correlation Measures*

As for attribute correlation measures, Euclidean distance and Pearson's correlation coefficient are widely used for clustering genes.

Given two genes A_i and A_j, $i, j \in \{1,\cdots, p\}$, $i \neq j$, the Euclidean distance between A_i and A_j is given by:

$$d_E(A_i, A_j) = \sqrt{\sum_{k=1}^{n} (t_{ik} - t_{jk})^2}, \quad (1)$$

where $t \in \Re$ is the measured expression level.

d_E measures the difference in the individual magnitudes of each genes. The genes regarded as similar by Euclidean distance may be very dissimilar in terms of their shapes or vice versa. For example, consider the two genes, which have an identical shape but only differ from each other by a large scaling factor. Their Euclidean distance is large although they have an identical shape. However, for gene expression data, the overall shapes of genes are of primary interest [29]. It is for this reason that Euclidean distance may not be able to yield a good proximity measurement of genes.

The Pearson's correlation coefficient between genes A_i and A_j is defined as:

$$d_C = \frac{\sum_{k=1}^{n} (t_{ik} - \bar{t_i})(t_{jk} - \bar{t_j})}{\sqrt{\sum_{k=1}^{n} (t_{ik} - \bar{t_i})^2} \sqrt{\sum_{k=1}^{n} (t_{jk} - \bar{t_j})^2}}, \quad (2)$$

where $\bar{t_i}$ and $\bar{t_j}$ are the means of t_{ik} and t_{jk}, $k = 1,\cdots, n,$ respectively.

It considers each gene as a random variable with n observations and measures the similarity between the two genes by calculating the linear relationship between the distributions of the two corresponding random variables. Empirical studies have shown that correlation coefficient is not robust to outliers and it may assign high similarity score to a pair of dissimilar genes.

Au et al. presented an information measure to evaluate the correlation between attributes. It is called the interdependence redundancy measure [6] between two attributes, A_i and A_j, $i, j \in \{1,\cdots, p\}$, which is defined as:

$$R(A_i : A_j) = \frac{I(A_i : A_j)}{H(A_i, A_j)}, \quad (3)$$

where $I(A_i : A_j)$ is the mutual information between A_i and A_j, and $H(A_i, A_j)$ is the joint entropy of A_i and A_j.

All the methods above and most of the literatures are seeking to discretize the continuous data into a finite number of intervals for data mining. But discretization leads to information loss.

Instead of using the Euclidean distance, Pearson's correlation coefficient, and interdependence redundancy measure, our proposed approach employs neighborhood mutual information to evaluate the interdependence of genes and groups genes that are dependent on each other into clusters. The use of this information measure allows GCMACA to discover meaningful clusters of genes reflecting similarity, both positive and negative correlation between expressions among genes. The detail of the neighborhood mutual information and its significance in gene expression correlation are given later in section Ⅲ.

Gene selection is another important step to further narrowing down the attribute number prior to data mining. A good number of algorithms have been developed for this purpose [24,27]. To select genes, the t-value is widely used in the literature. Assuming that there are two classes of samples in a gene expression data set, the t-value $t(A_i)$ for gene A_i is given by:

$$t(A_i) = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}, \quad (4)$$

where μ_r and σ_r are the mean and the standard deviation of the expression levels gene A_i for class r, respectively, and n_r is the number of samples in class r for $r = 1, 2$.

The top genes ranked by the t-value can then be selected for data mining. When there are multiple classes of samples, the t-value is typically computed for one class versus all the other classes.

A weakness of using the t-value to select genes is the redundancy among the selected genes [18,20]. To solve

this problem, methods that can handle both gene-class relevance and the gene-gene redundancy have been proposed [18,20,24,29]. These methods typically use some metric to measure the gene-class relevance (e.g., mutual information [11], the F-test value [18], information gain [15], symmetrical uncertainty [22], etc.) and employ the same or a different metric to measure the gene-gene redundancy (e.g., the L_1 distance [19], Pearson's correlation coefficient, etc.) To find a subset of relevant but nonredundant genes, they usually use a method called redundant cover to eliminate redundant genes with respect to a subset of genes selected according to the metric for measuring the gene-class relevance and gene-gene redundancy [18,20]. Another approach to doing so combines the metric for measuring the gene-class redundancy and that for measuring the gene-gene redundancy into a single criterion function is optimized [24].

It is important to note that both t-value and methods that handle the gene-class relevance and the gene-redundancy can only be used to select genes when the samples are preclassified.

B. Neighborhood Mutual Information Measure

In 2010, Hu et al. proposed neighborhood mutual information to cope with continuous gene data, evaluating the relevance between attributes [11].

There is a problem to employ mutual information in gene evaluation due to the difficulty in estimating probability density of genes. So neighborhood mutual information combines the concept of neighborhood with information theory, and generalizes Shannon's entropy to numerical information.

Training samples are usually given as vectors of attribute values and the attributes are numerical, as shown in Table Ⅰ, where A1 and A2 are two attributes, while C is the decision label of samples.

Let $U = \{x_1, x_2, \cdots, x_n\}$ be a set of samples described with gene set F, $x_i \in R^N$, Δ is a distance function on U, $\delta \geq 0$ is a constant, then define the neighborhood of sample x by:

$$\delta(x) = \{x_i \mid \Delta(x, x_i) \leq \delta\}. \quad (5)$$

Given $S \subseteq F$ is a subset of genes, the neighborhood of sample x_i in S is denoted by $\delta_S(x_i)$. The neighborhood uncertainty of x_i is defined as:

TABLE I.
CLASSIFICATION TASK DESCRIBED BY NUMERICAL FEATURE

Sample ID	A1	A2	C
1	0.52	0.36	1
2	0.28	0.00	1
3	0.50	0.24	1
4	0.18	0.73	1
5	0.42	0.48	2
6	0.01	0.58	2
7	0.30	0.71	2
8	0.49	0.04	2
9	0.34	0.36	3
10	0.64	0.35	3

$$NH_\delta^{x_i}(S) = -\log \frac{\| \delta_S(x_i) \|}{n}, \quad (6)$$

and the average uncertainty of the set of samples is computed as:

$$NH_\delta(S) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\| \delta_S(x_i) \|}{n}, \quad (7)$$

where $\| X \|$ is the cardinality of set X.

Assume $\delta = 0.2$, then $\delta_S(x_1) = \{x_1, x_3, x_5, x_9, x_{10}\}$, $NH_\delta^{x_1}(S) = -\log(\frac{5}{10}) = 1$, where $S = \{A_1, A_2\}$.

Given $R, S \subseteq F$ two subsets of genes, the neighborhood of sample x_i in gene subspace $S \cup R$ is denoted as $\delta_{S \cup R}(x_i)$, then the joint neighborhood entropy of $S \cup R$ is computed as :

$$NH_\delta(R,S) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\| \delta_{S \cup R}(x_i) \|}{n}. \quad (8)$$

Let $R, S \subseteq F$ be two subsets of genes, then the neighborhood mutual information of R and S is defined as:

$$NMI_\delta(R;S) = -\frac{1}{n}\sum_{i=1}^{n} \log \frac{\| \delta_R(x_i) \| \cdot \| \delta_s(x_i) \|}{n \| \delta_{S \cup R}(x_i) \|}. \quad (9)$$

III. AN GREEDY CORRELATION BASED ATTRIBUTE CLUSTERING ALGORITHM FOR GENE SELECTION

A. The Novel Correlation Measure

Our proposed measure is based on the following ideas: 1) any individual measure in the literature still fails to obtain good result in the recent years; 2) the proposed measure should be suitable to deal with continuous data analysis.

The correlation between two attributes, A_i and A_j, $i, j \in \{1,\cdots, p\}$, $i \neq j$, is defined as:

$$NR_\delta(A_i; A_j) = \frac{NMI_\delta(A_i; A_j)}{NH_\delta(A_i, A_j)}, \quad (10)$$

where $NMI_\delta(A_i; A_j)$ is the neighborhood mutual information between A_i and A_j, and $NH_\delta(A_i, A_j)$ is the joint neighborhood entropy of A_i and A_j.

$NMI_\delta(A_i; A_j)$ measures the average reduction in uncertainty about A_i that results from learning the value of A_j. If $NMI_\delta(A_i; A_j) > NMI_\delta(A_i; A_h)$, $h \in \{1,\cdots, p\}$, $h \neq i \neq j$, the relevance of A_i on A_j is greater than that of A_i on A_h. $NR_\delta(A_i; A_j)$ reflects the degree of deviation from correlation between A_i and A_j. If $NR_\delta(A_i; A_j) = 1$, then A_i and A_j are strictly correlated; If $NR_\delta(A_i; A_j) = 0$, then they are statistically independent; If $0 < NR_\delta(A_i; A_j) < 1$, then A_i and A_j are partially correlated. The definition of the correlation shows that it is independent of the composition of A_i and A_j. This implies that the number of attribute values does not affect the correlation between A_i and A_j. The properties of the correlation clearly render an ideal candidate to measure the relevance between different attributes. If two attributes are correlated to each other, they are more correlated to each other when compared to two more independent attributes.

In order to investigate the correlation of an attribute with all the others within a group, we introduce the concept of significant multiple correlation.

B. The Significant Multiple Correlation Measure

The significant multiple correlation measure of an attribute A_i within an attribute cluster, $C = \{A_j \mid 1,\cdots, p\}$ is defined as:

$$MNR_\delta(A_i) = \sum_{j=1}^{p} NR_\delta(A_i; A_j), \quad (11)$$

where $NR_\delta(A_i; A_j)$ is the correlation between A_i and A_j.

Based on the concept of $MNR_\delta(A_i)$, we introduce the concept of the "mode", which is an attribute with the highest multiple correlation in an attribute cluster.

The mode of an attribute cluster, $C = \{A_j \mid 1,\cdots, p\}$, denoted by $\eta(C)$ is an attribute, say A_i, in that cluster such that $MNR_\delta(A_i) \geq MNR_\delta(A_j)$, for all $j \in \{1,\cdots, p\}$.

C. The Description of The Greedy Correlation Based Attribute Clustering Algorithm

To group attributes $A_1,\cdots, A_p$ into clusters, we build our attribute clustering algorithm: 1) convert the concept of the term "mean", which represents the center of a cluster of entities, into the concept of mode, which is the attribute with the highest multiple correlation within an attribute group and 2) use correlation to evaluate the relevance between attributes. Then we can formulate the algorithm in the following.

1. Initialization. Let us assume that the number of clusters, k, where k is an integer greater than or equal to 2 is given. Of the p attributes, we randomly select k attributes, each of which represents a candidate for a mode η_r, $r \in \{1,\cdots, k\}$. Formally, we have $\eta_r = A_i$, $r \in \{1,\cdots, k\}$, $i \in \{1,\cdots, p\}$, to be the mode of C_r, and $\eta_r \neq \eta_s$ for all $s \in \{1,\cdots, k\} - \{r\}$.
2. Assignment of each attribute to one of the clusters. For each attribute, A_i, $i \in \{1,\cdots, p\}$, and each cluster mode η_r, $r \in \{1,\cdots, k\}$, we can calculate the correlation between A_i and η_r, $NR_\delta(A_i; \eta_r)$. We assign A_i to C_r if $NR_\delta(A_i; \eta_r) \geq NR_\delta(A_i; \eta_s)$ for all $s \in \{1,\cdots, k\} - \{r\}$.
3. Computation of mode for each attribute cluster. For each cluster C_r, $r \in \{1,\cdots, k\}$, we set $\eta_r = A_i$, if $MNR_\delta(A_i) \geq MNR_\delta(A_j)$ for all $A_i, A_j \in C_r$, $i \neq j$.
4. Termination. Steps 2 and 3 are repeated until the η_r for the clusters does not change. Alternatively, the algorithm also terminates when the prespecified number of iteration is reached.

It is important to note that the number of clusters, k, is fed to GCMACA as an input parameter. To find the best choice for k, we use the sum of the significant multiple correlation measure, $\sum_{r=1}^{k} \sum_{A_i \in C_r} NR_\delta(A_i; \eta_r)$, to evaluate the overall performance of each clustering. With this measure, we can run GCMACA for all $k \in \{2,\cdots, p\}$ and select the value k that maximizes the sum of the significant multiple correlation measure over all the clusters as the number of clusters. That is,

$$k = \arg \max_{k \in \{2,\cdots, p\}} \sum_{r=1}^{k} \sum_{A_i \in C_r} NR_\delta(A_i; \eta_r). \quad (12)$$

To investigate the complexity of our algorithm, we consider a gene expression table, which is composed of n samples such that each sample is characterized by p gene expression levels. The k-modes algorithm requires $O(np)$ operations to assign each gene to a cluster (Step 2). It then performs $O(np^2)$ operations to compute the mode for each cluster (Step 3). Let t be the number of iterations, the computational complexity of the k-modes algorithm is given by:

$$O(GCMACA) = O(k(np + np^2)t).$$
$$= O(knp^2t)$$

This kind of task is able to be completed in a reasonable amount of time by any modern off-the-shelf single-processor machine. Furthermore, the k-modes algorithm can easily be parallelized to run on clusters of processors because the calculation of the correlation is an independent task.

D. Assigning Quality to Cluster

Once we have done the cluster we know that genes in a cluster show similar expression profiles and might be involved in the same pathway. Since we want to have as many as possible involved in our list of significant genes, we would like to sample from each cluster/pathway. But it would not be fair to treat each cluster and gene equally. The size of the clusters as well as the quality of a cluster plays a rule. If a cluster is very tight and dense it can be assumed that the members are very similar. On the other hand, if a cluster has wide dispersion the members of the cluster are more heterogeneous. To capture the biggest possible variety of genes, it would therefore be favorable to take more genes from a cluster of bad quality than from a cluster with good quality. In this paper, we introduce overall similarity to assess the quality of clusters. From the analysis above, we define the overall similarity as:

$$OS(C_r) = \frac{\sum_{A_i \in C_r} NR_\delta(A_i; \eta_r)}{\| C_r \|}, \qquad (13)$$

where C_r is r-th cluster, $r \in \{1, \cdots, k\}$, $OS(C_r)$ is the overall similarity of C_r, and η_r is the mode of C_r, $i \in \{1, \cdots, p\}$.

A high cluster quality means low dispersion, and the closer the quality gets to 0, the more scattered the cluster becomes. In our selection algorithm we decide that no matter how bad the quality and how small the size of the cluster we should get at least one element from each other. The drawback is that a cluster might represent a pathway that is totally unrelated to the discrimination we look for. If the cluster then has a bad quality we might pick a lot of genes from that cluster even though they are not informative. To counteract this problem we implemented the possibility to mask out and exclude clusters that have an average bad test statistic p-value. Lastly we want to have genes that have a high discriminatory power. This can be achieved by using an appropriate test statistic.

IV. Experimental Analysis

A. Experimental Results on A Synthetic Dataset

To evaluate the clusters of attributes formed by GCMACA, we first applied it to a synthetic data set. Each tuple in the synthetic data set is composed of 20 continuous attributes and is preclassified into one of the three classes: C_1, C_2, and C_3. Let us denote the attributes as $A_1, \cdots, A_{20}$. In the designed experiment, attribute values of A_1 and A_2 alone can determine the class membership of a tuple. Values of the other attributes (i.e., $A_3, \cdots, A_{20}$) in the tuples are randomly generated in the following manner:

$A_3 - A_6$: uniformly distributed from 0 to 0.5 if the value of $A_1 < 0.5$; uniformly distributed from 0.5 to 1, otherwise.

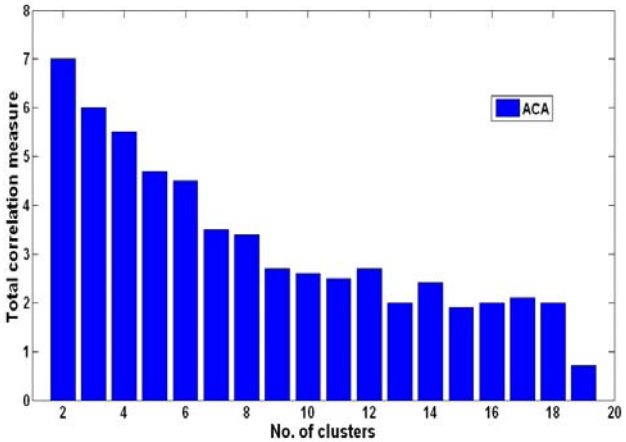
$A_7 - A_{11}$: uniformly distributed from 0 to 0.5 if the value of $A_1 \geq 0.5$; uniformly distributed from 0.5 to 1, otherwise.

$A_{12} - A_{15}$: uniformly distributed from 0 to 0.5 if the value of $A_2 < 0.5$; uniformly distributed from 0.5 to 1, otherwise.

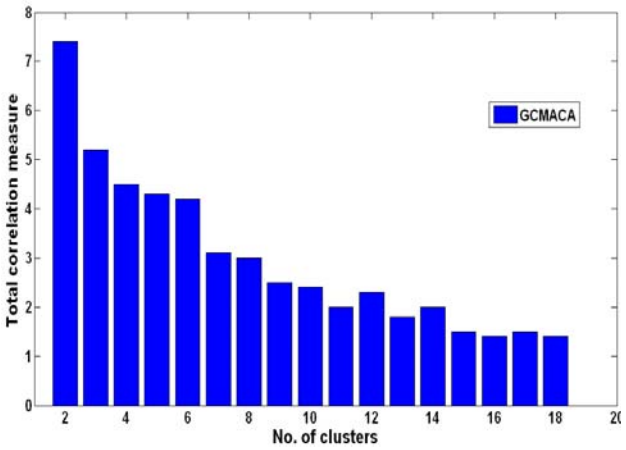
$A_{16} - A_{20}$: uniformly distributed from 0 to 0.5 if the value of $A_2 \geq 0.5$; uniformly distributed from 0.5 to 1, otherwise.

It is obvious that $A_3, \cdots, A_{11}$ are correlated with A_1, whereas $A_{12}, \cdots, A_{20}$ are correlated with A_2. For an attribute algorithm to be effective, it should be able to reveal such correlations. In our experiments, we generate 600 tuples in the synthetic data set and add noises to the data set by replacing the attribute values of $A_3, \cdots, A_{20}$ in 25 percent of the tuples with a random real number between 0 and 1.

For comparison, we show the performance of ACA [6] and GCMACA in Fig. 1. We set the $\delta = 0.1$, which is the best parameter verified through our experiment. Fig. 1 demonstrates the total correlation over all the clusters found in the synthetic data set.



(a)ACA



(b)GCMACA

Figure 1. The total correlation over all the clusters found in the synthetic data set

Fig. 1shows that both ACA and GCMACA are able to find that the optimal number of cluster is two and identify two clusters of attributes: $\{A_1, A_3 \cdots, A_{11}\}$ and $\{A_2, A_{12} \cdots, A_{20}\}$. A_1 is the mode of the former cluster, whereas A_2 is the mode of the latter. Furthermore, compared with ACA, the decision is more strongly supported by GCMACA, since the difference between 2 clusters and 3 clusters is bigger than that using ACA.

To compare with other algorithms, we applied the k-means algorithm [17], Kohonen's SOM [26], and the biclustering algorithm [29] to the synthetic data set. We can see from TABLE Ⅱ that the cluster configuration obtained by the k-means algorithm, SOM and the biclustering algorithm are not able to represent the correlations between attributes hidden in the data.

TABLE II.
CLUSTER NUMBER AND ITS CONFIGURATION

algorithm	cluster number	cluster
ACA	2	$\{A_1,A_3,\cdots,A_{11}\}$, $\{A_2,A_{12},\cdots,A_{20}\}$
GCMACA	2	$\{A_1,A_3,\cdots,A_{11}\}$, $\{A_2,A_{12},\cdots,A_{20}\}$
k-means	2	$\{A_1,A_3,\cdots A_6,$ $A_{17},\cdots A_{20}\}$, $\{A_2,A_7,\cdots,A_{16}\}$
biclustering	2	$\{A_1,A_3,A_8,A_9,A_{10},$ $A_{13},A_{14},A_{16},A_{17},A_{20}\}$, $\{A_2,A_4,\cdots,A_7,$ $A_{11},A_{12},A_{15},A_{18},A_{19}\}$
SOM	7	$\{A_2,A_{16}\}$, $\{A_9\}$ $\{A_4,A_6\}$, $\{A_1,A_3,A_5\}$ $\{A_7,A_8,A_{10},A_{11}\}$ $\{A_2,A_{13},A_{14},A_{15}\}$, $\{A_{17},\cdots,A_{20}\}$

TABLE III.
THE CLASSIFICATION PERFORMANCE OF SOME KINDS OF ALGORITHM

Algorithm	Decision tree configuration (leaf/nonleaf nodes)	Selected genes	Misclassified gene number
ACA	5/4	A_1,A_2	0
GCMACA	5/4	A_1,A_2	0
k-means	6/5	A_2,A_6	23
biclustering	5/4	A_2,A_{14}	72
SOM	9/8	A_2,A_4, A_5,A_8,A_9 ,A_{12},A_{19}	0

After clusters of attributes were obtained, we selected the top attribute in each cluster for classification. The selected attributes were fed to decision tree-based classification model. The experimental results are shown in the TABLE Ⅲ.

The experimental results on the synthetic data show that GCMACA is a very promising and robust technique 1) to group attributes into clusters, 2) to select a subset of attributes from the clusters formed, and 3) to allow classification algorithms to build accurate classification models.

B. Experimental Results on Gene Expression Data Sets

In order to test the our proposed algorithm, six cancer recognition data sets are collected. A review of these sets is given in TABLE Ⅳ.

To evaluate the attribute clustering results is difficult because we know too little about how genes actually associate among themselves. Hence, to have an objective and meaningful evaluation of GCMACA and others, we have to use what we know about the data to devise an evaluation scheme. From TABLE Ⅳ, we can see the detail of the gene expression data sets. Taking the preclassified knowledge as ground truth, we can devise an evaluation scheme as follows: Since the task objective of the proposed method is clustering and selection, we would like to ask how meaningful the clusters obtained are and what most useful information they contain. In view of this, we should

TABLE IV.
GENE EXPRESSION DATASETS

Data	Genes	Classes	Samples
Breast	9,216	5	84
DLBCL	4,026	6	88
Leukemia1	7,129	3	72
Leukemia2	12,582	3	72
Lung	7,129	3	96
SRBCT	2,308	5	88

TABLE V.
THE ACCURACY OF USING RAW DATA

Data	LSVM	CART	KNN
Breast	95.4±8.3	65.8±14.7	66.7±13.1
DLBCL	97.3±5.8	77.8±15.5	94.0±5.2
Leukemia1	94.8±6.8	78.5±14.9	77.4±17.1
Leukemia2	94.6±3.1	89.9±15.2	86.5±12.3
Lung	82.6±0.3	65.1±20.9	78.3±6.9
SRBCT	82.1±8.4	65.3±20.1	58.5±23.7
Average	91.1	73.7	76.9

TABLE VI.
THE NUMBER OF GENES USING ACA

Data	LSVM	CART	KNN
Breast	18	2	17
DLBCL	20	11	14
Leukemia1	19	4	10
Leukemia2	12	9	12
Lung	18	4	2
SRBCT	18	6	15
Average	18	6	12

TABLE VII.
CLASSIFICATION ACCURACY USING ACA

Data	LSVM	CART	KNN
Breast	96.3±6.0	80.8±13.4	95.0±6.2
DLBCL	97.3±5.7	88.9±10.7	99.0±5.3
Leukemia1	98.6±4.5	95.0±8.6	97.5±4.9
Leukemia2	99.0±3.1	93.3±18.0	100.0±0.0
Lung	80.2±3.4	82.8±7.8	87.4±5.9
SRBCT	79.9±28.5	77.5±17.7	76.6±27.2
Average	91.9	86.3	92.6

TABLE VIII.
THE NUMBER OF GENES USING GCMACA

Data	LSVM	CART	KNN
Breast	18	4	15
DLBCL	11	9	9
Leukemia1	11	2	16
Leukemia2	15	17	15
Lung	14	4	9
SRBCT	9	4	14
Average	13	7	13

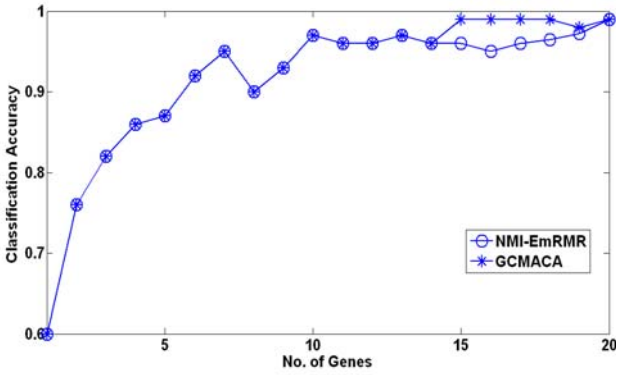
TABLE IX.
CLASSIFICATION ACCURACY USING GCMACA

Data	LSVM	CART	KNN
Breast	100.0±0.0	80.9±10.4	98.9±4.0
DLBCL	98.2±5.8	91.4±10.3	99.2±3.2
Leukemia1	98.7±4.5	95.3 ±7.3	99.1±4.5
Leukemia2	100.0±0.0	97.2 ±6.0	99.1±4.5
Lung	85.6±6.9	82.4±12.2	90.2±7.4
SRBCT	87.3±22.3	77.3±15.8	84.1±22.1
Average	95.0	87.4	95.1

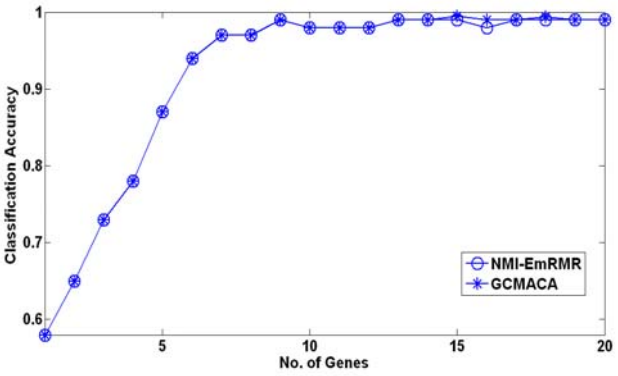
first examine the cluster configuration and infer by observation, which one reveals more information about the data, and gene grouping obtained. Next, we would like to get significant and insightful information from each cluster by selecting a subset of most representative genes and examining their patterns. Finally, we could use this extracted information for classification to see how the results obtained are backed by the ground truth.

Three of the most popular classifiers, LSVM, KNN, and CART were used to show the effectiveness of our proposed algorithm. Comparing TABLE Ⅴ~Ⅸ, we can see that after gene clustering by ACA，and GCMACA, the classification accuracy of selected genes outperform raw data which validate that raw data contain more complete information though, they are not suitable for classification. The cluster configurations are much the same by using ACA, GCMACA respectively, however, the classification accuracy of GCMACA is much better than ACA.

Fig. 2 shows the classification accuracy of the 20 top-ranked genes by NMI-EmRMR [11], GCMACA, respectively. Classification accuracy of the two methods are almost the same, but GCMACA is a little better than NMI-EmRMR through KNN, LSVM, respectively.



(a)KNN



(b)LSVM

Figure 2. The variation of classification accuracy with the number of selected genes.

V. Conclusions And Future Work

Gene clustering and classification are key task of gene identification. In virtue of the continuous attributes, our proposed greedy correlation measure based attribute clustering algorithm for gene selection can directly deal with the continuous data and acquire high accuracy. Experimental results show the algorithm outperforms other approaches. In the recent research, the cluster configuration is studied by using qualitative analysis. In order to get thorough understanding about gene expression profiles, and extend our model to improve its generation, we have to further investigate the cluster quality, which refers to its shape, size and distribution, etc. That is what we want to be involved with in the further concern.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 60873104, No. 61040037), the Key Project of Science and Technology Department of Henan Province (No. 112102210194), and the Science and Technology Research Key Project of Educational Department of Henan Province (No. 12A520027).

References

[1] J. Jaeger, R. Sengupta, and W.L. Ruzzo, “Improved Gene Selection for Classification of Microarrays,” Pacific Symposium on Biocomputing, 8, pp. 53-64, 2003.

[2] Yuan Hong, Jaideep, vaidya, Haibing Lu, “Searching Engine Query Clustering using top-k Search Results,” IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 1, pp. 112-119, 2011.

[3] Martin Hopfensitz, Christoph Mussel, Christian Wawra, “Multiscale Binarization of Gene Expression Data for Reconstructing Boolean Networks,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2, vol. 9, pp. 487-498, 2012.

[4] E. Domay, “Cluster Analysis of Gene Expression Data,” J. Statistical Physics, Vol. 110, pp. 1117-1139, 2003.

[5] D. Jiang, C. Tang, A. Zhang, “Cluster Analysis for Gene Expression Data: A survey,” IEEE Trans. Knowledge and Data Eng, 11, vol. 16, pp. 1370-1386, 2004.

[6] Wai-Ho Au, Keith C. C. Chan, Andrew K. C. Wong, Yang Wang, “Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2, vol. 2, pp. 83-101, 2005.

[7] Qinghua Hu, Daren Yu, Zongxia Xie, “Numerical Attribute Reduction Based on Neighborhood Granulation and Rough Approximation,” Journal of Software. 3, vol. 19, pp. 640-649, 2008.

[8] Kyriakos Kentzoglanakis, Matthew Poole, “A Swarm Intelligence Framework for Reconstructing Gene Networks: Searching for Biologicalling plausible Architectures,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9, vol. 2, pp. 358-371, 2012.

[9] G. Piatetsky-Shapiro, T. Khabaza, S. Ramaswamy, “Capturing Best Practice for Microarray Gene Expression Data Analysis,” Proc. Ninth ACM SIGKDD int’l Conf, Knowledge Discovery and Data Mining, pp. 407-415, 2003.

[10] L. J. Heyer, S. Kruglyak, S. Yooseph. “Exploring Expression Data: Identification and Analysis of Coexpressed Genes,” Genome Research, Vol. 286, pp. 531-537, 1999.

[11] Qinghua Hu, Wei Pan, Shuang An, Peijun Ma, Jinmao Wei, “An efficient gene selection technique for cancer recognition based on neighborhood mutual information,” Int. J. Mach. Learn. & Cyber, 1, pp. 63-74, 2010.

[12] K. C. C Chan, A. K. C Wong. “APACS: A System for the Automatic Analysis and Classification of Conceptual Patterns,” Computational Intelligence, 3, vol. 6, pp.119-131, 1990.

[13] K.C.C Chan, A.K.C Wong. “A Statistical Technique for Extracting Classificatory Knowledge from Databases,” Knowledge Discovery in Databases. G.Piatetsky-Shapiro, W.J. Frawley, Cambridge, Mass: AAAI/MIT Press. pp. 107-123, 1991.

[14] Y. Cheng, G. M. Church, “Biclustering of Expression Data,” Proc. Eighth Int’l Conf. Intelligent System for Molecular Biology, pp. 93-103, 2000.

[15] D. K. Y. Chiu, A. K. C. Wong, “Multiple Pattern Association for Interpreting Structural and Functional Characteristic of Biomolecules,” Information Sciences, vol. 167, pp.23-39, 2004.

[16] M. Delgado, N.Marin, D. Sanchez, M.-A. Vila, “Fuzzy Association Rules: General Model and Applications,” IEEE Trans. Fuzzy Systems. 2, vol. 11, pp.214-225, 2003.

[17] F. De Smet, J. Mathys, K. Marchal, G. Thijs, B. De Moor, Y. Moreau, “Adaptive Quality-Based Clustering of Gene Expression Profiles,” Bioinformatics, 5, vol. 18, pp.735-746, 2002.

[18] C. Ding, H. Peng, “Minimum Redundancy Feature Selection from Microarray Gene Expression Data,” Proc. IEEE Computational Systems Bioinformatics Conf, pp.523-528, 2003.

[19] E. Domany, “Cluster Analysis of Gene Expression Data,” J. Statistical Physics, vol. 110, pp. 1117-1139, 2003.

[20] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, “Cluster Analysis and Display of Genome-Wide Expression Patterns,” Proc. Nat’l Academy of Sciences of the United States of Am, 25, vol. 95, pp. 77-87, 2002.

[21] L. J. Heyer, S. Kruglyak, S. Yooseph, “Exploring Expression Data: Identification and Analysis of Coexpressed Genes,” Genome Research, vol. 9, pp. 1106-1115, 1999.

[22] K. Hirota, W. Pedrycz, "Fuzzy computing for Data Mining," Proc. IEEE. 9, vol. 87, pp. 1575-1600, 1999.
[23] C. Z. Janikow, "Fuzzy Decision Trees: Issues and Methods," IEEE Trans. Systems, Man, Cybernetics-Part B: Cybernetics, 1, vol. 28, pp. 1-14, 1998.
[24] D. Jiang, C. Tang, A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," IEEE Trans. Knowledge and Data Eng, 11, vol. 16, pp. 1370-1386, Nov.2004.
[25] J. Kacprzyk, S. Zadrozny, "On Linguistic Approaches in Flexible Query and Mining of Association Rules," Flexible Query Answering Systems: Recent Advances, H.L. Larsen, J. Kacprzyk, S. Zadrozny, T. Andreasen, H. Christiansen, eds, pp. 475-484, Physica-Verlag, 2001.
[26] T. Kohonen, Self-Organizing Maps, third ed. Berlin: Springer-Verlag, 2001.
[27] L. Liu, A. K. C. Wong, Y. Wang, "A Global Optimal Algorithm for Class-Dependent Discretization of Continuous Data," Intelligent Data Analysis, 2, vol. 8, pp. 151-170, 2004.
[28] J. B. McQueen, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. Fifth Berkeley Symp, Math, Statistics and Probability, pp. 281-297, 1967.
[29] S. C. Madeira, A. L. Oliveira, "Biclustering Algorithm for Biological Data Analysis: A Survey," IEEE/ACM Trans. Computational Biology and Bioinformatics, 1, vol. 1, pp. 24-45, Jan.-Mar. 2004.
[30] S. N. Mukherjee, P. Sykacek, S. J. Roberts, S. J. Gurr, "Gene Ranking Using Bootstrapped P-Values," SIGKDD Explorations, 2, vol. 5, pp.16-22, 2003.
[31] W. Pan, "A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments," Bioinformatics, vol. 18, pp. 546-554, 2002.
[32] A. K. C. Wong, Y. Wang, "Pattern Discovery: A Data Driven Approach to Decision Support," IEEE Trans. Systems, Man, and Cybernetics-Part C: Applications and Rev. 1, vol. 33, pp. 114-124, 2003.