International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Fuzzy and Rough Set Theory Based Gene Selection Method

C. Kalaiselvi [1], Dr. G. M. Nasira[2]

[1]Research Scholar, Karpagam University, Coimbatore & Assistant professor, Dept of Computer Applications
Tiruppur Kumaran College for Women, Tirupur, Tamilnadu, India

[2]Ph.D, Assistant Professor, Department of Computer Applications, Chickanna Govt Arts College for Men, Tirupur, India

Abstract: ***The selection of genes from microarray gene expression datasets has become an important research in cancer classification because such data typically consist of a large number of genes and a small number of samples. In this work, Neighborhood mutual information is retrieved to evaluate the relevance between genes and is used to stop information loss. Firstly, an improved Relief Feature Selection algorithm is proposed to create candidate subsets of features. Based on the neighborhood mutual information the cohesion degree of neighborhood object and coupling degree between neighborhood objects have been defined. Furthermore, a new method of initialization for cluster centers in Fuzzy C-means (FCM) algorithm is proposed. FCM allows one piece of data that belong to two or more clusters. Neighborhood rough set is used for extraction and selection of features and is used in proposed FCM algorithm. Finally, to find the performance of the proposed approach, five gene expression datasets were taken. Experimental results show that the proposed approach can select genes effectively, and can obtain high and stable classification performance.***

Keywords: Fuzzy C-means, Neighborhood Rough Set, Relief Algorithm, Clustering, Gene Selection

1. Introduction

Microarray technology is the powerful technology has made it possible to simultaneously measure the expression levels of large numbers of genes in a short time [1, 2]. However, among the large amount of genes presented in microarray gene expression datasets, only a small fraction of them is effective for performing a certain diagnostic test. So, the curse of dimensionality caused by high dimensionality and small sample size of tumor dataset seriously challenges the tumor classification [3, 4]. How to select important gene subsets of thousands of genes in the gene expression profiles dataset to drastically reduce the dimensionality of tumor dataset is the key step to address this problem.

Many gene selection methods have been proposed for the analysis of gene expression datasets [5, 6]. Usually, the feature selection methods can be divided into three broad categories: filter, wrapper, and embedded methods [7, 8, and 9]. The filter method is used in feature selection and it is independent of a specific classification algorithm. Thus features that accurately present the original data set can be found.

The filter methods are of several types. Some of them are correlation-based feature selection [10], t-test, information gain, mutual information, and entropy-based methods [11]. However, they ignore feature dependencies, resulting in poor classification performance. Wrapper methods focus on improving classification accuracy of pattern recognition problems and typically perform better than filter methods. However, the wrapper methods are more time-consuming than filter methods [12]. Embedded techniques combine filter methods and wrapper methods. The advantage of the embedded algorithms is that they take the interactions with the classifiers into account.

Clustering analysis is an important technique in pattern recognition, which aims to divide a data set into several clusters [13]. The clustering algorithms can be broadly classified into several types. They are Hard, Fuzzy,

Possibilistic, and Probabilistic [14]. The ability of clustering methods is to extract groups of genes with similar functions from huge datasets according to the fact that genes with similar functions evince similar expression patterns of co-regulation [15, 16]. Intuitively, genes in a cluster are more correlated with each other, whereas genes in different clusters are less interdependent [17].

The K-means clustering algorithm is one of the popular algorithm which partitions data objects into k number of clusters where the number of clusters, k is calculated in prior based on the application purposes. However, hard clustering methods which assign each gene exactly to one cluster are poorly suited to the analysis of gene expression datasets because in such datasets the clusters of genes frequently overlap [4]. To overcome the limitations of these hard clustering methods, fuzzy clustering has been widely studied, in which a data point is associated with multiple clusters to different extents based on its membership values to these clusters [18, 19]. The proposed FCM algorithm is one of the most popular fuzzy clustering techniques because of its efficiency, implementation and straightforwardness.

2. Related work

2.1 Fuzzy C-means Clustering

The FCM clustering algorithm is developed by Dunn [20] and later refined by Bezdek [21], is an unsupervised fuzzy clustering algorithm with multiple applications, ranging from attribute analysis, to clustering and classifier design. Let the sample set be $X = \{x_1, x_2 \ldots x_n\}$ where n is the number of sample. FCM algorithm divide the sample set X into c ($2 \leq c$

≤ n) classes is U = $[u_{ij}]_{c\times n}$, where $u_{ij}(1 \le i \le c, 1 \le j \le n)$ is the fuzzy membership degree of the j^{th} sample x_j belongs to the i^{th} class, and u_{ij} should satisfy the following constraint

$$\sum_{i=1}^{c} u_{ij} = 1, 0 \leqslant u_{ij} \leqslant 1, 1 \leqslant i \leqslant c, 1 \leqslant j \leqslant n.$$

The objective function of FCM is termed as

$$minJ_m(U,P) = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^m d_{ij}^2,$$

where $d_{ij} = \|x_j - p_i\|$ is the distance between x_j and p_i, p_i is the center of the i^{th} class. The fuzziness of the membership is controlled by m which takes value higher than 1. The closer is the m value to 1, the more crisply the membership values are. As if the value of m increases progressively, the resulting membership function becomes fuzzier [21]. Pal and Bezdek advised that m should take value between 1.5 and 2.5 [22]. FCM algorithm is an process to minimize the objective function min Jm(U,W).

2.2 Neighborhood Rough Set

In order to effectively cope with continuous attributes, avoiding the information loss which caused by discretization, Hu et al. [24] the proposed neighborhood rough set based on classical rough sets and the concept of neighborhood. The basic concepts of neighborhood rough set are explained as follows. Given arbitrary xi ∈U and B ⊆ A, d > 0 is a constant, and then the neighborhood of sample xi is denoted by

$$\delta_B(x_i) = \{x \in U | \Delta_B(x, x_i) \le \delta\},$$

where D is a distance function on U. Let A = $\{a_1, a_2, \ldots, a_n\}$ be a discrete random variable. $P(a_i)$ is the probability of a_i, the entropy of A is denoted by

$$H(A) = -\sum_{i=1}^{n} p(a_i)\log p(a_i)$$

3. Efficient Gene Selection Algorithm

3.1 Neighborhood Mutual Information Measure

Euclidean distance, Pearson's correlation coefficient and mutual information are widely used as the measure to compute relevance between attributes. However, to measure the correlation between selected genes, the Euclidean distance measure which is not effective enough to describe functional similarity such as positive or negative correlation in values. Thus, Pearson's correlation coefficient [24] is put forward by some researchers.

Empirical studies have shown that it may assign a high similarity score to a pair of dissimilarity genes. There is a problem to employ mutual information in gene evaluation due to the difficulty in estimating probability density of genes. However, most methods are not able to effectively cope with continuous attributes, which is also a distinctive characteristic of gene expression datasets. 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 [20]. Hu et al. [25] proposed neighborhood mutual information to cope with continuous attributes, evaluate the relevance between attributes. The neighborhood mutual information combines the concept of neighborhood with information theory, and generalizes Shannon's entropy to numerical information.

3.2 Improved Relief Algorithm

Relief as a kind of attribute ordering algorithm has been widely applied in the field of feature selection. Its core idea is to distinguish similar samples as the standard of evaluation attribute importance, and thus gives the attribute weights in classification.

The advantage of this algorithm is less computational complexity, considering the correlation between attributes to a certain extent. For arbitrary sample, searching out two class neighbors which nearest to this sample, one kind with the same classes of groups (called nearest hit), and another kind is the category with its distinct groups (called nearest miss). Then the search process in a sample of nearest neighbors is to take the distance between the two samples as the standard.

In the Relief algorithm, all the attributes are involved in the distance calculation process. However, in gene expression datasets, only a small number of genes associated with the sample type, the vast majority of genes as noise properties exist. If use Relief algorithm to select the gene expression datasets directly, will make the noise drowned out the useful information, resulting in the classification weights calculated of genes deviate from the true value.

The RFE algorithm presented [26], firstly, computes the attribute classification weights and then removes the attribute with the value of minimum weight. Then the effect of noise can be reduced gradually. However, this algorithm did not take into account the relationship between features of each sample, which affects the accuracy of classification. In this paper, we improve the RFE Relief algorithm and propose an improved Relief algorithm to select sample classification genes, which uses the neighborhood mutual information to measure the correlation between genes.

3.3 Cohesion Degree of the Neighborhood of an Object and Coupling Degree between Neighborhoods of Objects

Formally, the structural data used for classification learning can be written as an information system, denoted by = <U,A,V, f >, where U is the is the nonempty set of samples $\{x_1, x_2, \ldots, x_m\}$, called a universe; A is a set of attributes $\{a_1, a_2, \ldots, a_n\}$ to characterize the samples; V is the union of all attribute domains, ie., V = ∪Va, where Va is the value domain of attribute a and V ⊂ R; f is a mapping called an information function such that for any x ∈U and a ∈ A, f (x,a) ∈ Va.

3.4. The Significance of Attributes based on Neighborhood Mutual Information

The significance of attributes can be used as heuristic information in greedy algorithm to compute a minimal attribute reduction. In this paper, the significance of an attribute is proposed based on neighborhood mutual information. If the neighborhood mutual information is larger, the two attribute sets are closely related. If the neighborhood mutual information becomes zero, the two attributes are independent.

3.5. Description of the Improved Gene Selection Algorithm

Firstly, an improved Relief feature selection algorithm is proposed to sequence the gene selection, and is used to generate candidate feature subsets. Then, FCM algorithm is used to cluster the centers are initialized based on the candidate feature gene subsets. Furthermore, the relevance's between attributes are evaluated by neighborhood mutual information, and the significance of attributes is defined. Finally, selection of the attribute to represents the cluster which has the highest significance within each cluster. The detailed processing steps in the proposed algorithm are illustrated in flow chart form in the Fig.1.

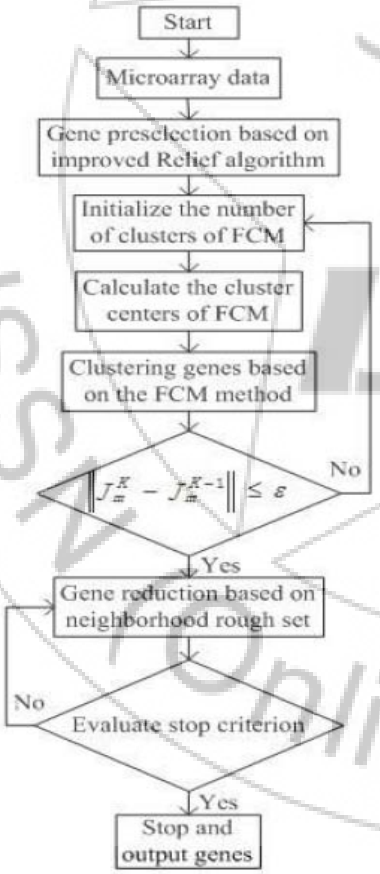


Figure 1

Table 1: Gene Expression Data Sets

Dataset	*Gene*	*Classes*	*Samples*
Leukemia	7129	3	72
Colon	2000	2	62
SRBTCM	2308	5	88
DLBCL	12582	3	72

Table 2: Accuracy of Genes by Genes Selection Algorithms (%)

Dataset	*Raw*	*Relieff*	*Cfs*	*Nrs*	*Fcm*
Leukemia	95.4± 6.5	98.6± 7.4	96.3± 5.4	83.7± 8.9	98.8 ± 3.2
Colon	84.6± 6.4	76.4± 9.8	82.6± 6.9	70.5± 5.3	88.6 ± 11.9
SRBTCM	82.4±8.2	79.8±8.2	86.0± 9.5	67.0± 7.6	89.0 ± 9.6
DLBCL	94.6±3.2	98.6±7.4	96.3±5.3	90.3±6.8	99.6 ± 5.8

4. Future Scope

The future scope of this work lies in analyzing limitations of this algorithm in comparison with other algorithms and significant improvements can be made in algorithms for better attribute selection process. A lot of future work can be done to improve the performance of gene selection process based on the selection of neighborhood rough set.

5. Conclusion

Since gene expression data sets have thousands of genes and only a small number of samples, feature selection is an essential step to perform cancer classification, which to predict classes and a relatively small number of samples. Gene identification is done by Clustering and Classification. In attribute selection process, the rough set theory has been widely used. While the gene expression data sets are always continuous, the classical rough set methods cannot handle this case directly. The Neighborhood rough set is used to deal with the continuous data in gene expression. In this paper, we introduce NMI to compute the relevance between genes and define the cohesion degree of the neighborhood of an object and coupling degree between neighborhoods of objects which based on neighborhood mutual information. Furthermore, the new initialization method of cluster centers for the Fuzzy C-means algorithm and the novel algorithm for gene selection based on Fuzzy C-means algorithm and neighborhood rough set are proposed. Compared with Relief, CFS, NRS; FCM gets good genes for cancer classification.

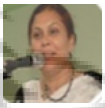
References

[1] J.C. Xu, L. Sun, Y.P. Gao, T.H. Xu, An ensemble feature selection technique for cancer recognition, Bio-Medical Materials and Engineering, 24, 1001-1008 (2014).
[2] S.Y. Kim, J.W. Lee, J.S. Bae, Effect of data normalizationon fuzzy clustering of DNA microarray data, BMC Bioinformatics, 7, 134-148 (2006).
[3] L. Sun, J.C. Xu, Feature selection using mutual information based uncertainty measures for tumor classification, Bio-Medical Materials and Engineering, 24, 763-770 (2014).
[4] J.J. Dai, L. Lieu, D. Rocke, Dimension reduction for classification with gene expression microarray data, Statistical Applications in Genetics and Molecular Biology,5, 1-21 (2006).
[5] L. Sun, J.C. Xu, A granular computing approach to gene selection, Bio-Medical Materials and Engineering, 24, 1307-1314 (2014).
[6] J.C. Xu, T.H. Xu, L. Sun, J.Y. Ren, An Improved Correlation Measure-based SOM Clustering Algorithm

for Gene Selection, Journal of Software, 8, 3082-3087 (2013).
[7] Z.X. Zhu, Y.S. Ong and M. Dash, Wrapper-filter feature selection algorithm using a memetic framework, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 37, 70-76 (2007).
[8] H.Y. Lin, Feature selection based on cluster and variability analyses for ordinal multi-class classification problems, Knowledge-Based Systems, 37, 94-104 (2013).
[9] L. Sun, J.C. Xu, T. Yun, Feature selection using rough entropy-based uncertainty measures in incomplete decision systems, Knowledge-Based Systems, 36, 206-216 (2012).
[10] M.A. Hall, Correlation-based feature subset selection for machine learning, Department of Computer Science, University of Waikato, Haimlton, New Zealand, (1999).
[11] J.C. Xu, L. Sun, Knowledge entropy and feature selection in incomplete decision systems, Applied Mathematics & Information Sciences, 7, 829-837 (2013).
[12] H. Liu, L. Yu, Toward integrating feature selection algorithm for classification and clustering, IEEE Transactions on Knowledge and Data Engineering, 17, 491-502 (2005).
[13] X. Li, H.S. Wong, S. Wu, A fuzzy minimax clustering model and its applications, Information Sciences, 186, 114- 125 (2012).
[14] R.J. Hathway, J.C. Bezdek, Optimization of clustering criteria by reformulation, IEEE transactions on Fuzzy Systems, 3, 241-245 (1995).
[15] P. Tamayo, D. Slonim, et al, Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, Proc Natl Acad Sci, 96, 2907-2912 (1999).
[16] P.T. Spellman, G. Sherlock, et al, Comprehensive identification of cell cycle-regulated genes of the yest Saccharomyces cerevisiae by microarray hydridization, Mol Biol Cell, 9, 3273-3279 (1998).
[17] J.C. Xu, Y.P. Gao, S.Q. Li, L. Sun, T.H. Xu, J.Y. Ren, A greedy correlation measure based attribute clustering algorithm for gene selection, Journal of Computers, 8, 951- 959 (2013).
[18] I.B. Aydilek, A. Arslan, A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm, Information Sciences, 233, 25-35 (2013).
[19] X. Li, H.S. Wong, S. Wu, A fuzzy minimax clustering model and its applications, Information Sciences, 186, 114- 125 (2012).
[20] Q.H. Hu, D.R. Yu, Z.X. Xie, Numerical Attribute Reduction Based on Neighborhood Granulation and Rough Approximation, Journal of Software, 3, 640-649 (2008).
[21] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, 3, 32-57 (1973).
[22] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, New York, Plenum Press, (1981).
[23] N.R. Pal, J.C. Bezdek, On cluster validity for the fuzzy Cmeans model, IEEE Transactions on Fuzzy systems, 3, 370- 379 (1995).
[24] C. Budayan, I. Dikmen, M.T. Birgonul, Comparing the performance of traditional cluster analysis, self-organizing maps and fuzzy C-means method for strategic grouping, Expert Systems with Application, 36, 11772-11781 (2009).
[25] Q.H. Hu, W. Pan, et al, An efficient gene selection technique for cancer recognition based on neighborhood mutual information, Int. J. Mach. Learn. & Cyber, 1, 63-74 (2010).
[26] J. Li, H. Su, et al, Optimal search-based gene subset selection for gene array cancer classification, IEEE Trans Inform Technol Biomed, 11, 398-405 (2007).

Author Profile

C. Kalaiselvi received her **M.C.A.,** degree from Mother Teresa Womens University, **M.Phil** degree from Manonmaniam Sundaranar University and SET Qualified. She is having 14 years of teaching experience and presently working as Associate Professor & Head in the dept. of Computer Applications. She is pursuing her PhD at Karpagam University, Coimbatore. Her area of interest includes Data Mining, Networks and communication security, Software engineering.



Dr. G.M. Nasira is working as Assistant Professor of Computer Application, Department of Computer Science, Chikkanna Government Arts College, Tiruppur,Tamilnadu, India. She has 17 years of teaching experience. Her Areas of Interest include Artificial Neural Networks, Data mining, Optimization, and Soft Computing. She has 102 publications and organized 21 Seminars / Conferences, Workshop / FDP / Orientation. She has been profiled in various organizations her academic contributions as member of board of studies. She is life member and senior member of ISTE, IACSIT etc. She is the recognized trainer associate for UGC's capacity building programs. She received best faculty award from various institutions.