ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Performance Enhancement of Robust Rough Fuzzy Clustering using Silhouette Index

K.Vengatesan[1] , S.Sadhana[2]

Associate Professor, Dept. of CSE, Muthayammal Engineering College, Rasipuram, India[1]

M.E Student, Dept. of CSE, Muthayammal Engineering College, Rasipuram, India[2]

ABSTRACT: The Clustering is a very important task in data processing, the practical analysis of sequence clump investigation will be performed by victimization numerous algorithms. The clump techniques are helpful to understand sequence functions, cellular method, sequence regulation and subtypes of cells. The various techniques are wont to measure the performance of the sequence overlapping like CLICK, SOM, and rRFCM. The projected ErRFCM increase the chance membership of the clusters and conjointly handle the overlapping sequence clusters effectively. It is conjointly helpful in addressing probabilistic lower approximation and risk lower approximation. The projected strategies are wont to establish the sturdy cluster of Co expressed genes and manufactures the most effective result. The sequence clusters created are HCM, FCM, RFCM, SOM, CLICK and rRFCM algorithms, and pictured by Tree read software package for Microarray dataset.

KEYWORDS*:* Clustering, Silhoutte Index, Micro array

I. INTRODUCTION

In biological domain, the gene based on the special characteristics of genes, clustering has several new challenges, due to special characteristics of genes. The retrieval natural data structure of the gene and data distribution, carried by gene expression data by clustering. The clustering finds natural group present in the gene set. It divides the genes into two category either same cluster or different cluster. The similar cellular function, can be cluster with similar functions of the gene, are called as co-expressed gene also understand the functionality of many genes. The co-expressed gene has strong correlation between the pair of genes. Regularity motifs used to search a common DNA sequence in promoter regions of genes within the same cluster.

A microarray gene expression data set can be represented by an expression table, where each row corresponds to one particular gene, each column to a sample or time point, and each entry of the matrix is the measured expression level of a particular gene in a sample or time point, respectively. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. A first step toward addressing this challenge is the use of clustering techniques, which is essential in pattern recognition process to reveal natural structures and identify interesting patterns in the underlying data. Cluster analysis is a technique for finding natural groups present in the gene set. It divides a given gene set into a set of clusters in such a way that two genes from the same cluster are as similar as possible and the genes from different clusters are as dissimilar as possible.

The co expressed genes are cluster using the gene clustering techniques also consider the co functions and co regulation. The gene expression data are grouped using clustering algorithms such as, model based method, Graph theoretic methods and soft computing density based methods, hierarchical methods and partitioning methods. The rough clusters are defined as similar to rough set using lower and upper boundary approximation, another important distinction of rough cluster is one cluster are overlapped with another cluster. The rough clustering allows grouping of related object based on the similarity relation equivalence. The Fuzzy set clustering same like Fuzzy C-mean of that data object belong to multiple clusters according to degree of membership. Rough set based clustering proves to provide a solution that is less restrictive than the traditional clustering algorithms like k-means and less descriptive than the fuzzy clustering methods.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

II. RELATED WORK

The major part of the project development sector considers and fully survey all the required needs for developing the project. For every project Literature survey is the most important sector in software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary software are needed to proceed with the next step such as developing the tools, and the associated operations.

The prediction of functional sites in proteins is an important issue in protein function studies and drug design. To apply the kernel based pattern recognition algorithms such as support vector machines for protein functional sites prediction, a new string kernel function, termed as the modified bio-basis function, is proposed recently. The bio-basis strings for the new kernel function are selected by an efficient method that integrates the Fisher ratio and the concept of degree of resemblance. In this regard, this paper introduces some quantitative indices for evaluating the quality of selected bio-basis strings. Moreover, the effectiveness of the new string kernel function and bio-basis string selection method, along with a comparison with existing bio-basis function and related bio-basis string selection methods, is demonstrated on different protein data sets using the proposed quantitative indices and support vector machines [1].

An important application of microarray data in functional genomics is to classify samples according to their gene expression profiles such as to classify cancer versus normal samples or to classify different types or subtypes of cancer. One of the major tasks with gene expression data is to find co-regulated gene groups whose collective expression is strongly associated with sample categories. In this regard, a gene clustering algorithm is proposed to group genes from microarray data. It directly incorporates the information of sample categories in the grouping process for finding groups of co-regulated genes with strong association to the sample categories, yielding a supervised gene clustering algorithm. The average expression of the genes from each cluster acts as its representative. Some significant representatives are taken to form the reduced feature set to build the classifiers for cancer classification. The mutual information is used to compute both gene-gene redundancy and gene-class relevance. The performance of the proposed method, along with a comparison with existing methods, is studied on six cancer microarray data sets using the predictive accuracy of naive Bayes classifier, K-nearest neighbor rule, and support vector machine. An important finding is that the proposed algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability [2].

Quantitative structure activity relationship (QSAR) is one of the important disciplines of computer-aided drug design that deals with the predictive modeling of properties of a molecule. In general, each QSAR dataset is small in size with large number of features or descriptors. Among the large amount of descriptors presented in the QSAR dataset, only a small fraction of them is effective for performing the predictive modeling task. In this paper, a new feature selection algorithm is presented, based on rough set theory, to select a set of effective molecular descriptors from a given QSAR dataset. The proposed algorithm selects the set of molecular descriptors by maximizing both relevance and significance of the descriptors. An important finding is that the proposed feature selection algorithm is shown to be effective in selecting relevant and significant molecular descriptors from the QSAR dataset for predictive modeling. The performance of the proposed algorithm is studied using R2 statistic of support vector regression method. The effectiveness of the proposed algorithm, along with a comparison with existing algorithms, is demonstrated on three QSAR datasets [3].

The selection of nonredundant and relevant features of real-valued data sets is a highly challenging problem. A novel feature selection method is presented here based on fuzzy-rough sets by maximizing the relevance and minimizing the redundancy of the selected features. By introducing the fuzzy equivalence partition matrix, a novel representation of Shannon's entropy for fuzzy approximation spaces is proposed to measure the relevance and redundancy of features suitable for real-valued data sets. The fuzzy equivalence partition matrix also offers an efficient way to calculate many more information measures, termed as f-information measures. Several f-information measures are shown to be effective for selecting nonredundant and relevant features of real-valued data sets. This paper compares the performance of different f-information measures for feature selection in fuzzy approximation spaces. Some quantitative indexes are introduced based on fuzzy-rough sets for evaluating the performance of proposed method. The effectiveness of the proposed method, along with a comparison with other methods, is demonstrated on a set of real-life data sets [4].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Among the great amount of genes presented in microarray gene expression data, only a small fraction is effective for performing a certain diagnostic test. In this regard, mutual information has been shown to be successful for selecting a set of relevant and nonredundant genes from microarray data. However, information theory offers many more measures such as the f-information measures that may be suitable for selection of genes from microarray gene expression data. This paper presents different f-information measures as the evaluation criteria for gene selection problem. To compute the gene-gene redundancy (respectively, gene-class relevance), these information measures calculate the divergence of the joint distribution of two genes' expression values (respectively, the expression values of a gene and the class labels of samples) from the joint distribution when two genes (respectively, the gene and class label) are considered to be completely independent. The performance of different f-information measures is compared with that of the mutual information based on the predictive accuracy of naive Bayes classifier, K -nearest neighbor rule, and support vector machine. An important finding is that some f-information measures are shown to be effective for selecting relevant and nonredundant genes from microarray data. The effectiveness of different f-information measures, along with a comparison with mutual information, is demonstrated on breast cancer, leukemia, and colon cancer datasets. While some f -information measures provide 100% prediction accuracy for all three microarray datasets, mutual information attains this accuracy only for breast cancer dataset, and 98.6% and 93.6% for leukemia and colon cancer datasets, respectively [5].

A generalized hybrid unsupervised learning algorithm, which is termed as rough-fuzzy possibility C-means (RFPCM), is proposed in this paper. It comprises a judicious integration of the principles of rough and fuzzy sets. While the concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in class definition, the membership function of fuzzy sets enables efficient handling of overlapping partitions. It incorporates both probabilistic and possibility memberships simultaneously to avoid the problems of noise sensitivity of fuzzy C-means and the coincident clusters of PCM. The concept of crisp lower bound and fuzzy boundary of a class, which is introduced in the RFPCM, enables efficient selection of cluster prototypes. The algorithm is generalized in the sense that all existing variants of C-means algorithms can be derived from the proposed algorithm as a special case. Several quantitative indices are introduced based on rough sets for the evaluation of performance of the proposed C-means algorithm. The effectiveness of the algorithm, along with a comparison with other algorithms, has been demonstrated both qualitatively and quantitatively on a set of real-life data sets [6].

III. PROPOSED SYSTEM

To understand gene function, gene regulation, cellular processes, and subtypes of cells, clustering techniques have proven to be helpful. The co expressed genes, that is, genes with similar expression patterns, can be clustered together with similar cellular functions. This approach may further understanding of the functions of many genes for which information has not been previously available. Furthermore, co expressed genes in the same cluster are likely to be involved in the same cellular processes and a strong correlation of expression patterns between those genes indicates co regulation. Searching for common DNA sequences at the promoter regions of genes within the same cluster allows regulatory motifs specific to each gene cluster to be identified and cisregulatory elements to be proposed. The inference of regulation through gene expression data clustering also gives rise to hypotheses regarding the mechanism of transcriptional regulatory network. The purpose of gene clustering is to group together co expressed genes which indicate co function and co regulation. Due to the special characteristics of gene expression data, and the particular requirements from the biological domain, gene clustering presents several new challenges and is still an open problem. The cluster analysis is typically the first step in data mining and knowledge discovery. The purpose of clustering gene expression data is to reveal the natural data structures and gain some initial insights regarding data distribution. Therefore, a good clustering algorithm should depend as little as possible on prior knowledge, which is usually not available before cluster analysis.

The gene selection for real value data set is a highly challenging task, the novel feature selection method is based on the fuzzy rough set while they maximize the relevance and minimizing the repeated features. The shannons entropy based method is a fuzzy equivalence partition matrix for gene selection, which is suitable for real value data set also increased the performance of the gene cluster using f-information measures. Even though the large amount of genes are represented as microarray data, only a small portion of data is effectively handled and applied for various test to enhance the performance of the genetic similarity. Consider two genes that are completely independent of each other, the joint distribution function is used to calculate the relationship between the genes, and also various predictive

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

accuracy measuring techniques such as support vector machine, k-nearest neighbor rule and naïve Bayes classifiers are applied. The f-information measures are efficient for breast cancer data, leukemia data sets.

3.1 Fuzzy C-Mean (FCM)

The rough fuzzy, c-mean is technically used to handle the rough set and fuzzy set, which effectively work on lower and upper approximation values and efficient to applicable for overlapped partitions. It also avoids the noise sensitivity problem occurred in fuzzy, c-mean algorithm. The second order fuzzy measure and weighted co-occurrence matrix are another method used to measure the gene co-occurrence based on the threshold value. The efficiency is more when compared to fuzzy entropy, fuzzy correction methods and local information are more accurately measured. Integrating the merits of fuzzy sets and rough sets, different rough-fuzzy clustering algorithms such as the rough fuzzy, c-mean, rough fuzzy possibility c-mean and rough possibility c-mean in which each cluster is represented by a cluster prototype with possibility boundary and lower approximation.

Various clustering algorithms are used in microarray to calculate the Co expression of gene expression data sets, from that crisp lower approximation and fuzzy boundary are generally assumed the spherical in shape, which find the arbitrary shape of gene clustering

The strong associated Co expressed genes are calculated using the fuzzy rough supervised gene clustering algorithm. The rough fuzzy, c-mean was derived from a rough fuzzy clustering algorithm, which is efficient to handle the micro array gene expression data even though of overlapped partition and noisy data. There are three basic parameters needed to form the clusters namely, possibility lower approximation, probability boundary and cluster prototype or centroid. The cluster centroid depends on the weighting average of the probabilistic boundary and possibility lower approximation. A main problem of an existing method is to detect the efficient method to find the prototype of the gene at initial stage. The effectiveness of the algorithm is compared with existing algorithm along to microarray data set.

3.2 Robust Rough fuzzy C-Mean Algorithm

It is an efficient method to handle the cluster in both possibility and probabilistic fuzzy sets, and upper and lower approximation of rough sets into C-Mean algorithm, while integrating both techniques, it will handle the overlapping cluster in noisy environments, also deal with vagueness, incompleteness uncertainty in cluster definition. The objective function is let Y= $\{y_1 \ldots . y_j \ldots . y_n\}$ be a set of n objects and C= $\{c_1 \ldots . c_i \ldots . c_c\}$ be the set of centroid, where $y_j \varepsilon R^m$ and $v_j \varepsilon R^m$. Each of the clusters β_i is represented by a cluster center υ_i which follows both lower and upper approximation of the cluster β_i. The minimization function of the proposed C cluster is written as

$$J=\begin{cases} wA_1 + (1-w)\beta_1 & if\ A(\beta) \neq \theta, B(\beta_i) \neq \theta \\ A_1 & if\ A(\beta) \neq \theta, B(\beta_i) \neq \theta \\ B_1 & if\ A(\beta) \neq \theta, B(\beta_i) \neq \theta \end{cases} \quad (1)$$

In which $A(\beta)$ is the lower approximation and $B(\beta_i)$ is the probability boundary where A_1, B_1 are represented as

$$A_1 = \sum_{i=1}^{c} \sum_{x_j \varepsilon\ A(B_i)} (v_{ij})^{m_2} \left\| x_j - v_i \right\|^2 + \sum_{i=1}^{c} \eta_i \sum_{x_j \varepsilon\ A(B_i)} (1-v)^{m_2}$$

$$B_1 = \sum_{i=1}^{c} \sum_{x_j \varepsilon\ B(B_i)} (\mu_{ij})^{m_1} \left\| x_j - v_i \right\|^2$$

The relative important of lower boundary is represented by the parameter w and (1-w), while $1<=m_1<\infty$ and $1<=m_2<\infty$ are the probability functions. The centroids of the cluster should be independent to the lower approximation along with memberships of the objects. The membership function between the object is represented as following equation

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

$$\mu_{ij} = \left[\sum_{k=1}^{c} \left(\frac{\|x_j - v_i\|^2}{\|x_j - v_k\|^2} \right)^{\frac{1}{m_1 - 1}} \right]^{-1}$$

$$v_{ij} = \left[1 + \left\{ \frac{\|x_j - v_i\|^2}{\eta_i} \right\}^{\frac{1}{m_2 - 1}} \right]^{-1}$$

$$\eta_i = K. \frac{\sum_{j=1}^{n} (v_{ij})^{m_2} \|x_j - v_i\|^2}{\sum_{j=1}^{n} (v_{ij})^{m_2}}$$

Where the scale parameter which represents the size of the cluster B. The centroid of the cluster are calculated based on the weighting average of the probabilistic boundary and possibility lower approximation.

The **table1** show the performance comparison of different algorithms using Silhouette index validation measures, consider the micro array dataset GDS608 the values of the HCM is 0.08, FCM is 0.01, RFCM is 0.15, rRFCM is 0.27 and ErRFCM is 0.77, similarly consider another micro array dataset GDS2003 the values of the HCM is 0.19, FCM is 0.17, RFCM is 0.31, rRFCM is 0.61 and ErRFCM is 1.11 the graphical representation is illustrate in **Figure 1**.

Table1: Performance Analysis of Silhouette Index

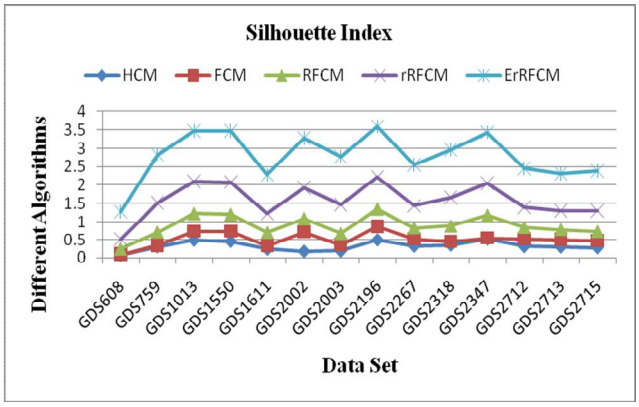

Figure 1: Performance Analysis of Silhouette Index with Yeast Microarray Data Sets

Micro array Data Sets	Silhouette Index				
	HCM	FCM	RFCM	rRFCM	ErRFCM
GDS608	0.08	0.01	0.15	0.27	0.77
GDS759	0.3	0.04	0.37	0.81	1.31
GDS1013	0.48	0.25	0.49	0.87	1.37
GDS1550	0.45	0.28	0.47	0.88	1.38
GDS1611	0.24	0.09	0.37	0.54	1.04
GDS2002	0.18	0.53	0.37	0.85	1.35
GDS2003	0.19	0.17	0.31	0.8	1.3
GDS2196	0.49	0.38	0.48	0.87	1.37
GDS2267	0.32	0.2	0.31	0.61	1.11
GDS2318	0.35	0.09	0.44	0.79	1.29
GDS2347	0.51	0.03	0.63	0.88	1.38
GDS2712	0.31	0.21	0.32	0.56	1.06
GDS2713	0.3	0.2	0.28	0.52	1.02
GDS2715	0.28	0.19	0.26	0.58	1.08

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

IV. CONCLUSIONS

The contribution of the system lies in developing a brand new clustering algorithm Enhanced robust rough fuzzy clustering algorithm, that integrates judiciously c-means formula, rough sets, and probabilistic and possibility memberships of fuzzy sets produce better result for Silhouette Index. This formulation is intermeshed toward maximizing the utility of each rough sets and fuzzy sets with regard to data discovery tasks. The effectiveness of the projected formula is incontestable, in conjunction with a comparison with different connected algorithms; on fourteen yeast microarray organic phenomenon information sets exploitation some commonplace cluster validity indices and clustering algorithm metaphysics.Moreover, the projected ErRFCM performs considerably higher than different ways, regardless of the microarray information sets and quantitative indices used, and provide biologically important and relevant clustering algorithm.

REFERENCES

[1] H. Causton, J. Quackenbush, and A. Brazma, Microarray Gene Expression Data Analysis: A Beginner's Guide. Wiley-Blackwell, 2003.
[2] E. Domany, "Cluster Analysis of Gene Expression Data," J. Statistical Physics, vol. 110, nos. 3-6, pp. 1117-1139, 2003.
[3] P. Maji and S.K. Pal, Rough-Fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging. John Wiley & Sons, Inc., 2012.
[4] D. Jiang, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey," IEEE Trans. Knowledge and Data Eng., vol. 16, no. 11, pp. 1370-1386, Nov. 2004.
[5] M.B. Eisen, P.T. Spellman, O. Patrick, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," Proc. Nat'l Academy of Sciences USA, vol. 95, no. 25, pp. 14-863-14-868, 1998.
[6] S. Tavazoie, D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church, "Systematic Determination of Genetic Network Architecture," Nature Genetics, vol. 22, no. 3, pp. 281-285, 1999.
[7] A. Brazma and J. Vilo, "Minireview: Gene Expression Data Analysis," Federation of European Biochemical Societies Letters, vol. 480, no. 1, pp. 17-24, 2000.
[8] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data," Proc. Second Int'l Workshop Information Processing in Cells and Tissues, pp. 203-212, 1998.
[9] P. Maji and C. Das, Protein Functional Sites Prediction Using Modified Bio-Basis Function and Quantitative Indices, IEEE Transactions on NanoBioscience, 9(4), pp. 250--257, December 2010.
[10] P. Maji and C. Das, Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification, IEEE Transactions on NanoBioscience, 11(2), pp. 161--168, June 2012.
[11] A.Brazma and J.Vilo,"Minireview:Gene Expression Data Analysis",Federation of European Biochemical Societies letters,vol.480,no.1,pp.17-24,2000.
[12] L.Heyer, S.Kruglyak,and S.Yooseph,"Exploring Expression Data:Identification and analysis of Coexpressed genes", Genome Research,vol.0,no.11,pp. 1106-1115,1999.
[13] E.Hartuv and R.Shamir, "A Clustering Algorithm based on Graph Expression Patterns",J. Computational Biology,vol.6,nos ¾,pp.281-297,1999.
[14] R.Sharmir, and R.Sharan,"Click :A Clustering Algorithm for Gene Expression Analysis",Proc. Eight Int'lConf. Intelligent System for Molecular Biology,2000.
[15] D.Ghosh and A.M.Chinnaiyan,"Mixture Modelling of gene Expression Data from Microarray Experiments",Bioinformatics,vol.18,no.2,pp.275-286,2002.
[16] D.Jiang,J.Pei and A.Zhang, "DHC:A Density Based Hierarchical Clustering Methods for Time Series Gene Expression Data," Proc,IEEE Third Int'l Symp, Bioinformatics and BioEng,pp.393-400,2003,
[17] J.C.Bezdek,Pattern Recognition with fuzzy objective function Algorithm,Plnum,1981.