IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Volume-3, Issue-9 E-ISSN: 2347-2693

Rough Based Clustering For Gene Expression Data –A Survey

C. Udhaya Bharathy[1*] and C. Rathika[2]

[2]*Dept. Of Computer Science, Bharathiar University, India*

www.ijcseonline.org

Received: Aug/22/2015 Revised: Aug/29/2015 Accepted: Sep/24/2015 Published: Sep/30/2015

Abstract— Microarray technology has made it possible to simultaneously monitor the expression levels of thousands of genes during important biological processes and across collections of related samples. But the high dimensionality property of gene expression data makes it difficult to be analyzed. Clustering associated with the concept of rough set theory is very effective in such situations. This paper gives a briefly introduction about the concepts of RST, clustering, gene expression, microarray technology and discuss the basic elements of clustering on gene expression data. It also explain why rough clustering is preferred over other conventional methods by presenting a survey on few clustering algorithms based on rough set theory for gene expression data. Finally it concludes by stating that this area proves to be potential research field for the research community.

Keywords— Microarray technology, Rough Set, gene expression, rough clustering.

I. INTRODUCTION

Data mining is the exploration and analysis of data in order to discover a valid, novel, potentially useful and ultimately understandable patterns in data. The important techniques involved in data mining are classification, clustering, and association rules.

A. CLUSTERING

Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters. Grouping is done by finding similarities between data according to characteristics found in the actual data. The groups are called clusters. Cluster Analysis technique as a field grew very quickly with the goal of grouping data objects, based on information found in data and describing the relationships inside the data. The purpose is to separate the objects into groups, with the objects related (similar) together and unrelated with another group of objects. Clustering is an important unsupervised classification Technique. Cluster analysis is a difficult problem due to a variety of ways of measuring the similarity and dissimilarity concepts, which do not have a universal definition. It is being applied in variety of science disciplines and has been studied in research communities.

B. ROUGH SET THEORY

The data from the real world are often uncertain, vague or incomplete because of complications associated with the record or report of any natural phenomena or events that are under study. Some approaches are well known to handle such issues, mainly the Fuzzy Set theory, the Dempster-Shafer theory, and the possibility theory. In 1980, another theory emerged for treating such kind of data called the Rough Set Theory-RST. It is an extension of the set theory that deals with data uncertainty by means of an equivalence relation known as indiscernibility.

Two elements of a given set are considered as indiscernible if they present the same properties, according to a defined set of features, attributes or variables. Rough set theory, is a good mathematical tool for imperfect data analysis. The ideas of Rough Set proposed by Pawlak in 1980 and he is known to be 'Father of Rough Set Theory'. Methodology of RST is concerned analysis of missing attribute values, uncertain or incomplete information systems and knowledge, and it is considered one of the first non-statistical approaches in data analysis. Any subset defined by its upper and lower approximation is called "Rough Set"[1].

Lower approximation contains all the objects belong to the set but upper approximation contains the objects that may belong to the set. The differences between these lower and upper approximations define the boundary region of the rough set. The lower and the upper approaches are two basic functions in the rough sets theory.

C. TERMINOLOGIES ON ROUGHSET THEORY

In discernibility Relation

With any $P \subseteq A$, there is an associated Equivalence Relation $IND(P) = \{ (x , y) \in U$
$X U / \forall a \in P, a(x) = a(y)\}$.

Lower and Upper Approximation

Let $X \subseteq U$ can by approximated using only the information contained within P, by constructing P-Lower

and P-Upper approximations of a classical crisp set X are given by[3]

$$Px\ (Low) = \{\ x / [x]p \subseteq X\ \}$$
$$Px\ (Upp) = \{\ x / [x]p \cap X \neq \square\ \}$$

The figure 1 shows diagrammatic representation of Rough set.

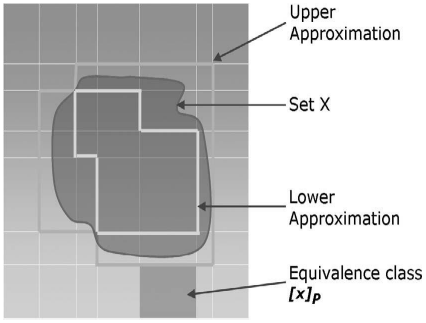

Fig 1: Diagrammatic representation of Rough set

D. GENE EXPRESSION DATA

Gene is the fundamental unit of storage of hereditary information in living beings. A gene is a segment of DNA, which contains the formula for the chemical composition of one particular protein. Genetic information is encoded in the linear sequence in which the bases on the two strands are ordered along the DNA molecule. Information from a gene is used in the synthesis of functional gene products like proteins and functional RNAs for non-protein coding genes. This process of synthesis is called Gene Expression.

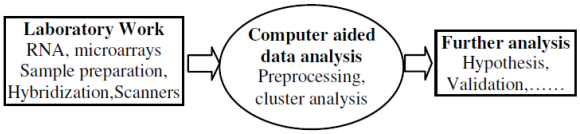

Fig 2: Gene Expression analysis pipeline

Gene expression is the process by which genetic instructions are used to synthesize gene products. These products are usually proteins, which go on to perform essential functions as enzymes, hormones and receptors, for example. Genes that do not code for proteins such as ribosomal RNA or transfer RNA code for functional RNA product.

The functional analysis of genes is a way to find what roles the genes play in the living organism. It is important to understand what proteins the genes code, and where the genes are expressed (in tissues or organs) and when they are expressed[2]. The methods for studying the expression of genes generally involve the expression at the transcription level. The important and effective method that can be used at the transcription level to analyze gene expression is the microarray technology.

E. MICROARRAY TECHNOLOGY

Microarray is a powerful tool for gene function analysis. Some of the applications of microarrays can be disease diagnosis, gene discovery, drug discovery, and toxicological research. Microarray technology has made it possible to monitor simultaneously the expression levels for 1000 of genes during important biological process and across collections of related samples. It has empowered the scientific community to understand the fundamental aspects underlining the growth and development of life as well as to explore the genetic causes of anomalies occurring in the functioning of the human body.

Microarray data consist of an array of color coded signals indicating the up and down regulation of genes under various cellular conditions. The image can be further converted into numerical data as pixel intensity. Micro arrays make possible the speedy and quantitative genomic scale analysis of gene expression patterns. It is often an important task to find genes with similar expression patterns (co-expressed genes) from microarray data[4].

II. ROUGH BASED CLUSTERING

A. ROUGH CLUSTERING

A rough cluster is defined to have a lower and upper approximation like rough set. The lower approximation of a rough cluster contains objects that only belong to that cluster. The upper approximation of a rough cluster contains objects in the cluster which are also members of other clusters. An important distinction between rough clustering and other conventional clustering approaches is that, with rough clustering, an object can belong to more than one cluster thereby allowing overlapping of clusters[4]. An appropriate distance measure should be used in rough clustering such that the strict requirement of indiscernibility relation used in normal clustering is relaxed. Thus rough clustering allows for grouping of objects based on a notion of similarity relation rather than based on equivalence relation

B. ADAPTATION OF ROUGH SET THEORY FOR CLUSTERING

The present study is about a generalized view of rough sets over clustering. The concept of LOWER and UPPER APPROXIMATIONS (bounds) which belongs to the rough set theory are used in clustering in order to deliver a better clustering algorithm. Let us consider a hypothetical classification scheme[6].

$$U/P = \{C1, C2, \ldots, Ck\} \quad (1)$$

The scheme partitions the set U based on an equivalence relation P. consider that due to insufficient knowledge that it is not possible to precisely describe the sets Ci, $1 \leq i \leq k$, in the partition. Based on the available information, however, it is possible to define each set $Ci \in U/P$ with its lower $A(Ci)$ and upper $A(Ci)$ bounds. We are considering the upper and lower bounds of only a few subsets of U. Therefore, it is not possible to verify all the properties of the rough sets [14].However, the family of upper and lower bounds of **ci** $\in$ **U/P** are required to follow some of the basic rough set properties[6] such as:

1. An object **x** can be part of at most one lower bound
2. **x** $\in A(\mathbf{c}i) \Rightarrow$ **x** $\in A(\mathbf{c}i)$
3. An object **x** is not part of any lower bound $\Leftrightarrow$ **x** belongs to two or more upper bounds.

First Property states the fact that a lower bound is included in a set. If two sets are mutually exclusive, their lower bounds should not overlap.

Second Property confirms the fact that the lower bound is contained in the upper bound.

Third Property is applicable to the objects in the boundary regions, which are defined as the differences between upper and lower bounds. The exact membership of objects in the boundary region is ambiguous. Therefore, third property states that an object cannot belong to only a single boundary region.

III. ROUGH BASED CLUSTERING IN GENE EXPRESSION

Information from a gene is used in the synthesis of functional gene products like proteins and functional RNAs for non-protein coding genes. This process of synthesis is called Gene Expression, by which genotype gives rise to phenotype. This Gene Expression Data is generally very huge in size and to search for useful patterns within this data, genes have to be grouped into "clusters" on the basis of similar features. The Gene Expression data is in the form of a 2-Dimensional matrix of Gene Names and the corresponding expression levels for features exhibited by the genes. Clustering of Gene Expression data has been done by various algorithms

J. Jeba Emilyn and K. Ramar,[13] Proposed an algorithm named as Rough Clustering of Gene Expression Data (RCGED), clusters genes based on rough set theory. The main advantage of our method is that it does not restrict a gene to one cluster. Genes can get expressed in two are more clusters ie Overlapping of genes are possible. It also finds the lower and upper approximation of the clusters. RCGED algorithm is designed to be intelligent in the sense that it itself detects the optimum number of clusters. This algorithm uses a similarity measure based on correlation coefficient.

Anasua Sarkar & Ujjwal Maulik[14] proposed a rough set based hybrid approach for modified symmetry-based clustering algorithm. A natural basis for analyzing gene expression data using the symmetry-based algorithm, is to group together genes with similar symmetrical patterns of microarray expressions. Rough-set theory helps in faster convergence and initial automatic optimal classification, thereby solving the problem of unknown knowledge of number of clusters in gene expression measurement data. For rough-set-theoretic decision rule generation, each cluster is classified using heuristically searched optimal reducts to overcome overlapping cluster problem. The rough modified symmetry-based clustering algorithm is compared with another newly implemented rough-improved symmetry-based clustering algorithm and existing K-Means algorithm over five benchmark cancer gene expression data sets, to demonstrate its superiority in terms of validity.

Adhikary, K.,Das, S. & Roy, S.[10], proposed a new gene expression clustering method, termed as A Novel and Efficient Rough Set Based Clustering Technique for Gene Expression Data (NRSBCGE), based on the Rough set theory. This method is designed intelligently as it itself detects the optimum number of clusters. This clustering method provides an efficient way of finding the unique gene expression patterns. The method was experimented with two publicly available cancer datasets and the results were compared with two existing methods of clustering.

Lijun[7] used a new method combining correlation based clustering and rough sets attribute reduction together for gene selection from gene expression data is proposed. Correlation based clustering is used as a filter to eliminate the redundant attributes, and then the minimal reduct of the filtered attribute set is reduced by rough sets. A successful gene selection method based on rough sets theory is presented. The experimental results indicate that rough sets based method has the potential to become a useful tool in bioinformatics.

Jung-Hsien [8] presents a novel rough-based feature selection method for gene expression data analysis. The method(RBFNN) finds the relevant features without requiring the number of clusters to be known *a priori* and identify the centers that approximate to the correct ones. For each cluster, the algorithm finds the number of data points in the upper bound and the lower bound. The method introduces a scheme that combines the rough-based feature

selection method with radial basis function neural network. This method also proves to produces high accurate results.

Pradipta Maji [9] proposed a new clustering algorithm, termed as fuzzy–rough supervised attribute clustering (FRSAC), to find groups of coregulated genes whose collective expression is strongly associated with sample categories. The proposed algorithm is based on the theory of fuzzy–rough sets, which directly incorporates the information of sample categories into the gene clustering process. The effectiveness of this algorithm is compared with other existing supervised and unsupervised gene selection and clustering algorithms and proves to be better. The better performance of the proposed FRSAC algorithm is achieved due to the fact that it uses the fuzzy–rough supervised similarity measure to generate co-regulated gene clusters with strong association to the class labels[10]. The fuzzy-rough property makes it possible to deal with uncertainty, vagueness, and incompleteness in the class definition.

Anasua Sarkar & Ujjwal Maulik[12],proposed a algorithm named as parallel rough set based hybrid approach for point symmetry-based clustering algorithm. It is used to enable fast automatic clustering of large microarray data. It is a distributed time-efficient scalable method. This algorithm is compared with parallel symmetry-based K-means and parallel version of existing K-means over four artificial and benchmark microarray datasets. This method includes initial automatic optimal rough set based classification using best decision rules and point symmetry based distance norm to compute nearest neighbors among genes.

Ruizhi Wang , Miao, Duoqian , Gang Li & Hongyun Zhang[5] presented a novel approach named as Rough Overlapping Biclustering (ROB).This technique is used to find potentially overlapping biclusters in the framework of generalized rough sets. Our scheme mainly consists of two phases. First, we generate a set of highly coherent seeds (original biclusters) based on two way rough kmeans clustering. And then, the seeds are iteratively adjusted (enlarged or degenerated) by adding or removing genes and conditions based on a proposed criterion. They have illustrated the method on yeast gene expression data. The experiments demonstrate the effectiveness of this approach.

Dhanalakshmi.K & Hannah Inbarani. H [15] proposed an algorithm called Fuzzy Soft Rough K-Means algorithm. The algorithm is developed based on Fuzzy Soft sets and Rough sets. Comparative analysis of the proposed work is made with bench mark algorithms like K-Means and Rough K-Means. The efficiency of the Fuzzy Soft Rough K-Means algorithm is illustrated by using various cluster validity measures such as DB index and Xie-Beni index.

Rudra Kalyan Nayak ,Debahuti Mishra, Kailash Shaw & Sashikala Mishra[16] proposed an algorithm called Rough Set based attribute Clustering for Sample Classification (RSCSC). RSCSC is an efficient algorithm for finding out the meaning patterns of gene expression data from high dimensional datasets. This algorithm is compared with the traditional algorithm such as K-Mediods, K-Means and Hierarchical Clustering algorithms.

IV. CONCLUSION

This paper presented a survey of the clustering methods for gene expression data that are based on rough set theory. Apart from gene expression rough set based Clustering algorithms helps in identifying hidden pattern and providing enhanced understanding of the functional genomics in a better way. Many other domains of applications like web mining, text mining and collaborative filtering are open to be explored using rough set based clustering algorithms

REFERENCES

[1]Z. Pawlak, "Rough sets", International Journal of Computer and Information Sciences, **11:341-356** (**1982**).
[2] R.E. Kent, Rough Concept Analysis: a synthesis of rough sets and formal concept analysis, Fundamenta Informaticae , pp. **169–181, 1996.**
[3] Lin T.Y. and Cercone N. "Rough Sets and Data Mining - Analysis of Imperfect Data", Kluwer Academic Publishers, Boston, London, Dordrecht, P.**430**,**1997**.
[4] Thabet Slimani, "Application of Rough Set Theory in Data Mining ",**2010**.
[5] Ruizhi Wang , Tongji Univ, Shanghai , Miao, Duoqian ,Gang Li, Hongyun Zhang," Rough Overlapping Biclustering of Gene Expression Data", Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference,Oct. **2007.**
[6] Sajid Nagi, Dhruba K. Bhattacharyya, Jugal K. Kalita," Gene Expression Data Clustering Analysis: A Survey",**2008**.
[7] Lijun Sun Duoqian Miao Hongyun Zhang,(2007) Gene Selection with Rough Sets for Cancer Classification, Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)IEEE **2007**.
[8] Jung-Hsien Chiang and Shing-Hua Ho," A Combination of Rough-Based Feature Selection and RBF Neural Network for Classification Using Gene Expression Data", IEEE Transactions On Nanobioscience, Vol. **7**, No. **1**,March **2008.**
[9] Pradipta Maji,"Fuzzy–Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data", IEEE Transactions On Systems, Man, And Cybernetics.
[10] Adhikary, K. ,Das, S. Roy, S, " A Novel and Efficient Rough Set Based Clustering Technique for Gene Expression Data", 2nd International Conference on Business and Information Management (ICBIM), **2014**.

[11] Maji, P. ; Pal, S. "Clustering Functionally Similar Genes from Microarray Data", WileyIEEE Press **2012,** ISBN :**9781118119723.**

[12]Anasua Sarkar and Ujjwal Maulik, "Rough Based Symmetrical Clustering for Gene Expression Profile Analysis" IEEE transactions on nanobioscience, vol. **14,** no. **4,** june **2015.**

[13] J. Jeba Emilyn and K. Ramar " A Rough Set based Gene Expression Clustering Algorithm**",** Journal of Computer Science 7 (7): **986-990, 2011,ISSN 1549-3636**, Science Publications.

[14] Anasua Sarkar & Ujjwal Maulik,"Cancer Gene Expression Data Analysis using Rough Based Symmetrical Clustering", 2013.

[15] Dhanalakshmi.K & Hannah Inbarani, "Fuzzy Soft Rough K-means Clustering for Gene Expression Data",**2011.**

[16] Rudra Kalyan Nayak ,Debahuti Mishra, Kailash Shaw & Sashikala Mishra,"Rough Set Based Attribute Clustering For Sample Classification Of Gene Expression Data",**ICMOC-**ScienceDirect.**2012**.