ABSTRACT

The study on gene expression profiling of tissues and cells has become a major tool for discovery in medicine. Identification of co-expressed genes and coherent patterns is the central goal in gene expression profiling and the important task in the field of bioinformatics research. Clustering is an important unsupervised learning technique for Gene Expression Profile Analysis. Many conventional clustering algorithms have been adapted or directly applied to gene expression data. Among them, Rough Point Symmetry (RoughPsym) and Rough Symmetry (Roughsym) based clustering is applied for recognizing symmetrical patterns of gene expression profiles. Rough-set theory helps in faster convergence and initial automatic optimal classification, thereby solving the problem of unknown knowledge of number of clusters in microarray data. In case of RoughPsym and Roughsym methods, efficiency or higher accuracy is not achieved because of the larger dataset samples. To solve this problem and to further enhance the clustering and thereby enabling the clustering results of large microarray data, in this article, a distributed time-efficient scalable Sequential Rough Parallel Bounded Symmetrical clustering (SeqRoughPBSym) is applied to rough set based approach.

1. INTRODUCTION

The GENE expression profiles exhibit the expression levels for many genes simultaneously under multiple biological processes [1]. Genes which has the similar expression patterns may be co-regulated or in same signal pathway. Clustering is an unsupervised pattern classification method based on maximum intra-class similarity and minimum inter-class similarity. Eisen [2] first classified coexpressed genes using hierarchical clustering. Since then numerous methods have been proposed for clustering microarray data, [3], like parallel K-means algorithms using Hadoop. The concept of lower and upper approximations of rough sets deals with uncertainty, vagueness, and incompleteness in class definition[4]. However, the membership function of rough sets also enables efficient handling of overlapping partitions. Therefore, Hirano[5]et al proposed indiscernibility based clustering method to handle relative proximity. The present study concentrates on the integration of automatic optimal classification using rough-set-theory and point symmetry-based distance norm for analyzing gene-expression datasets varying both time-courses and environmental conditions. Clusters are associated with indiscernibility classes over genes.
Among conventional clustering methods, an effective partitional clustering algorithm is the K-means clustering algorithm. Given a set of points in a -dimensional space, the problem is to determine in for disjoint clusters. This minimizes the following crisp partitioning metric, where denotes Euclidean distance of pattern from centroid. Symmetry is considered as an inherent feature for recognition of shapes hidden in any clusters. Su and Chou [6] have proposed a variation of K-means algorithm with a symmetry-based distance measure. However, it fails when inherent symmetry with respect to some intermediate points lies within asymmetrical intra-cluster.

To solve this problem, Bandyopadhyay et al. [7] proposed an efficient point symmetry-based distance measure with a quadratic time complexity. In microarray analysis, the expression levels of two genes may rise and fall synchronously varying magnitudes but with symmetrical patterns. This reveals the point-symmetry feature. Analysis of high throughput gene expression profiles is very time-consuming. Therefore, parallel algorithms with point-symmetry detection are more suitable to analyze these data to reveal gene clusters with relevant gene regulatory networks.

Based on these observations, memory mapping technique based implementations of both SeqRoughParallelSymmetry (seqroughpsym) algorithm and SeqRoughParallelBoundSymmetry algorithm has been proposed in this article. Automatic rough indiscernibility based initial classification in a distributed master-slave environment based on the point symmetry norm. Mapping binary rough set theory trees onto parallel in a way that allows one to access any rooted subtree with a bounded number of conflicts. The contribution of this article is faster and efficient discovery of symmetrical gene clusters in large microarray datasets by the reformulation of the rough set theory with memory mapping.

1.1 GENE EXPRESSION PROFILING IN CANCER

Cancer is a disease characterized by uncontrolled cell growth and proliferation. For cancer to develop genes regulating cell growth and differentiation must be altered; these mutations are then maintained through subsequent cell divisions and are thus present in all cancerous cells. Gene expression profiling is a technique used in molecular biology to query the expression of thousands of genes simultaneously. In the context of cancer, gene expression profiling has been used to more accurately classify tumors [18,19]. The information derived from gene expression profiling often has an impact on predicting the patient’s clinical outcome.

1.2 ROUGH SET THEORY

In computer science, a rough set is a formal approximation of a crisp set (i.e., conventional set) in terms of a pair of sets which give the lower and the upper approximation of the original set. 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”[8].

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[9].RST terminologies are as follows

In discernibility Relation

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

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]

\[
\text{Px (Low)} = \{ x / [x]p \subseteq X \}
\]

\[
\text{Px (Upp)} = \{ x / [x]p \cap X \neq 0 \}
\]

The figure 1 shows diagrammatic representation of upper and lower approximations of the Rough set.

![Diagrammatic representation of Rough set](image)

**Fig 1: Diagrammatic representation of Rough set**

2. ROUGH CLUSTERING

A rough cluster is defined to have a lower and upper approximation like rough set[10]. 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[11]. An appropriate distance measure should be used in rough clustering such that the strict requirement of indiscernibility relation used in normal clustering is relaxed[12]. Thus rough clustering allows for grouping of objects based on a notion of similarity relation rather than based on equivalence relation.

3. BOUNDED PARALLEL MAPPING

A natural basis for analyzing gene expression data using the Bounded parallel mapping based rough set symmetrical based rough set algorithm is to group
together genes with similar symmetrical patterns of expression. Rough-set theory helps in faster convergence and initial automatic optimal classification, thereby solving the problem of unknown knowledge of number of clusters in microarray data[13]. This new sequential rough parallel bounded collision symmetrical clustering also satisfies the linear speedup in timing on large microarray datasets.

3.1 Bounded parallel algorithm

Memory mapping problem via a graph-theoretic abstraction allows us to partition problem into two logical phases. In phase 1, find, for each of families of templates, a single structure/graph that “almost” contains as a subgraph every one. If “almost” were replaced by “exactly” in the preceding sentence, then the graph we find would be perfect-universal. The qualifier “almost” refers to our allowing a bounded number of nodes of a subgraph to map to the same node of the universal graph reflecting our allowing a bounded number of conflicting accesses to each module in our memory-access problem. In phase 2 find efficient storage schemes for perfect centroids values in universal graphs[14].

For any gene dataset samples D, let G_D be the graph whose adjacencies describe those of D. For any graph G=(V^G, E^G), Size G = |V^G|, the number of nodes of G. A template t for a gene dataset samples D is any subgraph of G_D. Each subgraph of G_D that is isomorphic to t is an instance of template t. Informally, a template describes a pattern for accessing the nodes of D. The templates that consider here are more “ambitious” than those found in most earlier sources. Specifically, when consider centroids calculation for trees, allow all rooted binary trees as templates, centroids calculation for any of three gene samples[15].

A graph G_c = (V_c, E_c) is c-perfect-universal for the family X_n if, for each X=(V^X, E^X) in the family, there exists a c-contraction I_X = (V, E) ∈ H^c that is a labeled subgraph of G_c. By this, for any template X ∈ X_n, the fact that I_X is a subgraph of G_c is witnessed by a mapping f : V → V_c for which each v ∈ V is a subset of V^X; cf. is shown in Fig.

![Fig. 2 An example of centroids mapping](image)

3.2 Sequential Rough Bounded Parallel Symmetry Based Clustering

New Sequential Bounded parallel Point Symmetry-Based K-means clustering method with automatic rough set based initial clustering, has been proposed. Bounded-collision parallel algorithm also has been developed to compare its results with those of algorithm. The implementations only differ in the symmetry-based centroid updation phase, with two different distance norms.

Algorithm: Both partitioning and clustering phases of and algorithms have been implemented in a distributed master-slave paradigm. Among M distributed nodes, M_0 acts as the master to ensure load balancing and other M-1 nodes act as slaves. The algorithm composed of 4 phases—initial horizontal partitioning of universal dataset, local rough set based automatic classification, Bounded parallely-
computed local centroids updation using K-mean
method and Bounded parallel point symmetry- based
fine-tuning. Initial rough set based automatic cluster
assignment puts N elements in clusters K=RL. Each
slave then performs centroids updation on partitioned
data locally and returns its local cluster assignment to
\( M_0 \). Subsequently merges them into global cluster
assignment, using the union-find data structure with
an average runtime of inverse Ackermann's function.
If optimization continues, continues with
redistribution of corrected cluster assignment.

**Steps of Sequential Rough Bounded Parallel
Symmetry Based Clustering algorithm**

**Step 1**: Initialization on master node \( M_0 \), tree
Horizontal partitioning of N elements to \( m = M - 1 \)
slave nodes:

\[
\text{SCATTER} \left( \text{GLOBAL\_DATA}, \text{LOCAL\_DATA}, \frac{N}{m}, \text{Size}(T_n) \right)
\]

**Step 2**: Local rough set based automatic classification

a. Compute B-upper approximations for
each decision variable d

\[
\text{Perf}_c(T_n) > \exp \left( \frac{n+1}{c+1} - \frac{11}{3} \right)
\]

c. Compute reduced set \( \mathcal{R}_\theta(x,d) \) \( \forall x \in N \)
and decision rule set \( \Delta_\theta \) over decision d
d. Update K=RL-minimal set of rules for
d relative reducts covering all patterns
with optimized heuristic search
e. Assign \( \bar{x}_i \in \text{cluster } C_k \) \( \forall k = 1, \ldots \text{RL} \)
such that cluster \( |C_R| \geq 1 \) \( \forall k = 1, \ldots \text{RL} \)

**Step 3**: Bounded parallel computation of centroids
towards convergence

a. In

\[
M_0; \text{BROADCAST} \left( \text{GLOBAL\_CLUSTERS}, N \right)
\]

and lower, upper bound computation
b. On M slave nodes:

Compute centroids \( \bar{c}_k \) \( \forall k = 1, \ldots K \)
Execute 1 pass of K-means method
on LOCAL\_DATA to update \( \bar{c}_k \)
c. In \( M_0 \)

\[
\text{ALL\_GATHER} \left( \text{LOCAL\_CLUSTERS}, \text{GLOBAL\_CLUSTERS}, \frac{N}{m} \right)
\]
d. If \( \bar{c}_k \) converge then go to step 4
Else repeat step 3

**Step 4**: Bounded Parallel fine tuning phase:

a. Parallel \( \theta \) computation with PS-distance and
lower, upper bound computation
b. On M slave nodes:

Compute \( \max P = \text{Maximum PS distance} \forall i = 1, \ldots \frac{N}{m} \),
\( \text{Size}(T_n) \leq n \) elements
c. In \( M_0 \)

\[
\text{ALL\_GATHER} \left( \text{LOCAL\_\max\_P}, \text{GLOBAL\_\max\_P}, m, \text{Size}(T_n) \leq n \right)
\]
Compute
\[ \theta = \text{maximum nearest neighbor distance } \forall i = 1, \ldots N, \text{Size}(\tilde{T}_n) \leq n \text{ from GLOBAL maxP} \]
(BROADCAST(\theta, 1))

d. In
\[ M_0 : \text{LOCAL}_i \text{CHANGES} \]
Compute new PS distance corrected centroids
\[ c_k(t + 1) = \frac{\sum_{i} x_i(t)}{S_k(t)} \forall k = 1, \ldots K, S_k(t) = \{ x_i | x_i \in \text{Cluster C}_k \text{ at time t}, \text{Size}(\tilde{T}_n) \leq n \}
\]
Continuation
e. In \[ M_k: \text{if K cluster centroids converge or no pattern changes cluster then stop else go to step 2} \]
f. If \[ \bar{c}_k \] converge then go to step 4
Else repeat step 3

4. EXPERIMENTAL FRAMEWORK

4.1 Dataset description

“Rat CNS”—This normalized dataset [13] contains 112 gene expression levels during rat central nervous system development at 9 time points. This dataset is available at:

Imbalanced Data Sets: Two chosen skewed cancer gene expression datasets from Kent Ridge Biomedical Dataset are Breast cancer[16], and thyroid samples with 1.82, and 1.6 imbalance ratios [19] respectively

4.2 Performance analysis

In this section measure the performance results of the various parallel K-means (PKM), Rough set Symmetry (Rough Sym), Sequential Rough Symmetry (Seq Rough Sym), Rough Set Parallel Symmetry (Rough PSym), Sequential Rough Parallel Symmetry (Seq Rough PSym), proposed methods such as Sequential Rough Bounded Symmetry (Seq Rough BSym), and Sequential Rough Parallel Bounded Symmetry (Seq Rough PBSym), using the metrics of accuracy and time. The comparison metrics are shown below

![Fig 3 Time comparison of the algorithms for all three dataset](image)

![Fig 4 Accuracy comparison of the algorithms for all three dataset](image)
5. CONCLUSION

Gene expression microarray is one latest breakthrough in molecular biology, which monitors the expression patterns of thousands of genes across multiple conditions simultaneously. Those high throughput patterns may be alike and symmetrical. The contribution of this article is faster and efficient discovery of symmetrical gene clusters in large microarray datasets by the reformulation of the Seqrough_Psym and Seqrough_sym algorithms to the time-efficient scalable parallel implementations. The global centroids updation without using All-to-All communication pattern shows linear speedup with parallel symmetrical convergence. The contribution of this article is the automatic rough indiscernibility based initial classification in a distributed master-slave environment based on the point symmetry norm. The bounded parallel mapping based rough set symmetrical clustering algorithm, the rough membership function is used to detect the memberships of one gene to many clusters. Studied a new, highly flexible strategy for mapping complex templates to a parallel memory system, while allowing efficient access to an arbitrarily complex repertoire of access patterns. The strategy derives from extending the symmetry based clustering work on “perfect” method to allow bounded constraints for centroids calculation. Illustrated the new strategy for centroids calculation.

6. Future scope

The area of future research is to analyze the multi-shaped scattered gene expression data with dimension reduction. With the wide application of the microarray technology, more genome-size gene expression experimental data for an organism are becoming available. Because of the curse of dimensionality, traditional clustering algorithms are not suitable for analyzing these high dimensional data sets. Subspace clustering, which tries to find density sub-“blocks” in the high-dimensional data, is not only appropriate for the purpose of mining high-dimensional large gene expression profiles but also consistent with the knowledge from biology that only part of the genes is expressed at certain experimental conditions[17]. Designing an efficient subspace clustering algorithm that can discover biologically meaningful clusters still is an important research direction.

7. REFERENCE


