Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

GRAPH BASED GENE/PROTEIN PREDICTION AND CLUSTERING OVER UNCERTAIN MEDICAL DATABASES

[1]SHAHANA BANO, [2]K.RAJASEKHARA RAO

[1]Research Scholar, Department of Computer Science and Engineering, KL University

[2]Professor, Department of Computer Science and Engineering, Sri Prakash College of Engineering

E-mail: [1]shahanabano_cse@kluniversity.in, [2] krr_it@yahoo.co.in

ABSTRACT

Clustering over protein or gene data is now a popular issue in biomedical databases. In general large set of gene tags are clustered using high computation techniques over gene or protein distributed data. Most of the traditional clustering techniques based on subspace, hierarchical and partitioning feature extraction. Various clustering techniques have been proposed in the literature with different cluster measures, but the performance is limited due to its spatial noise and uncertainty. In this paper, an improved graph based clustering technique was proposed to generate efficient gene or protein clusters over uncertain and noisy data. Proposed graph based visualization can effectively identify different types of genes or proteins along with relational attributes. Experimental results show proposed graph model effectively clusters the complex gene or protein data compare to conventional clustering approaches.

Keyword: *Biomedical, gene, protein, clustering, medline, pubmed, Synonym Identification*

1. INTRODUCTION

The application of clustering method is one among the standard computational results for understanding microarray gene expression data [1],[2],[3].In the gene or protein clustering, the order of different gene expressions across treatments, time and tissues is formed as a group into different clusters in which genes in the same group are assumed to be potentially linked or to be affected by a common factor. For instance, in order to recognize gene functions in the gene expression, chemical treatments, mutants are to be used as a systematic tool. Since drug or mutants aim to monitor similar profiles are probably to share cellular functions [4]. Gene mutations have an influence on the similar organelle or impact the homogeneous signaling in other gene expression profiles. Hierarchical clustering methods are based on the fixing of threshold to describe members of the specific cluster from non-members by processing the overall cluster objective.

Traditional graph based methods [2-5] doesn't provide the level of tree pruning or specified number of clusters. Main problem to recognize distance metric is to select a user specified threshold for structured information such as gene expression profiles. These methods do not give a clear information about the gene or protein clustering, processing it very hard to solve the expected clustering and to process comparisons among clustering's depends on shapes and number of clusters. In this study, we use numerical inference to surmount these drawbacks. Bayesian model based approaches can give precise results during data prediction. With these advances, there is no use of processing arbitrary selections on the number of clusters in the data, still after modeling ,one can have a chance to ask a question such as "how possible is it that two genes belong to the identical cluster".

This gene or protein prediction depends on the mathematical limit of an uncountable number of components in a normal finite mixture related to Dirichlet process prior [5],[6],[7]. There is no need to make an arbitrary choice about how various clusters are there in data for an infinite Gaussian mixture method. Astonishingly, it is possible to do inference in these vast Bayesian models effectively using Hidden Markov chain Monte Carlo methodology. Although there is infinite mixture model has a various number of parameters, the parameters of a fixed number of mixture components need to be shown explicitly. Moreover, this process is not largely known and various groups have independently generated values depend on DPMs [9].We also subsequently used the approach to the clustering of protein sequences [7] .

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Gene or Protein feature extraction refers to the detection of best relevant patterns and has been similarly applied to forward classification in protein or gene expression [10]. The cost of estimating relevant subset patterns from the million patterns becomes a major problem, because a critical protein or gene microarray records matches thousands of protein patterns. In addition, peptides were filtered during the feature extraction process may be important in separating inter or intra class sub-types. DR means a class of various techniques that change the high standard data into a concise subspace to show data in minimal dimensions. The reduced dimensional was kept in order along with the principal eigenvectors in a DR method or Principal Component Analysis (PCA) and linear methods. But one important note that unlike with feature extraction, the embedded samples never shows particular protein or gene expressions from the originated high dimensional space but sooner encloses data similarities in low dimensions. Even data objects in transforming space are separated from their original meaning, maintaining the order of patient records in low dimensional embedding space tends it to data representation and segregation .So if both of the patient samples for a particular disease are linked adjacently to each other in embedding space obtained from their respective high dimensional profiles, then it says that both patients have a similar type of disease.

The main objective and contribution of this research work is to optimize the inter and intra gene relationship among the attributes.Also, the uncertainty in the distributed medical data is efficiently handle using the proposed graph based model.

2. RESEARCH METHOD

Linear Discrimination Analysis [2] with Fisher Discriminate is another linear Discrimination Analysis scheme, which introduced gene or protein label information to search data extensions that segregate the information into various categories. Multidimensional Scaling minimizes attribute dimensions by using the statistical least squares distance method in Low dimensions. The famous method for gene or protein visualization for bio-informatics related issues have been a PCA implemented by Hotelling, PCA recognizes orthogonal eigenvectors along with the high variability.

Classification with Discrimination Analysis schemes for medical information was an ambivalence. Dawson et al. discovered that medical prediction of gene or protein, which are not visible with linear multidimensional scaling. Ye et al detects Linear Discrimination Analysis did not get fair results for differentiating gene or protein disease classes with nine gene patterns . The isomap technique estimates geo-desic distances, described as the distance between the nonlinear distances and points along the manifold with Euclidean distances utilized in linear approaches, by reducing the projected data into a low dimension. The geodesic distance between the data is linear or non-linear. Niyogi and BELKIN implemented the LEM technique by using Isomap , spectral clustering and LLE which finds local neighborhood points, but uses the Laplacian weights with low dimensional data. Graph based embedded approach, isomaps, LLE and LEM all dream to nonlinearly project the high dimensional data in the copy that two things x1 and x2 that lie side to each other on the manifold in the low dimensional embedding space, and in the same manner two things that are far from each other in the low dimensional space.

Gene pattern information is generally derived from two medical sources 1.microarrays, 2. Pattern sequencing. Due to loss of spatial component and imaging very few genes can be recognized with spatial relationship. In this whole process datasets are obtained from digital imaging approaches .The wide information about the gene pattern is gained from visualization and by using heatmap [6,9]. So we arrange in matrix samples are linked to columns and genes are linked to rows, each gene in the matrix was allocated from a color depend on gene pattern value [3]. An improvement of heatmap called a curve map, it utilizes a time period curve as base unit in a matrix, by visualizing the comparison of temporal data .This view introduced in pathline [7],it is a tool which is developed for functional genomics dataset. The curve map shows time-series gene or protein patterns as filled curve lines in a heat map –style matrix layout and involves overlap curve plots along the rows and columns to specify the discovery of trends .We introduce the curve map show in Multee Sum for spectating the temporal gene data from independent gene cells in the embryos. Both the systems are defined by using linked views in a data depend on refining queries , which are being highlighted subsets of the data .For a 3D scatter plots a wide variety of improvements were introduced by Piringer and Kosara.

Truntzer et al also discovered an application of PCA and Linear Discrimination Analysis for explaining the protein and gene

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

patterns of a scattered large B-cell lymphoma dataset from the classes presented to be non-linearly and cannot be segregated .The above survey represents to recommend that bio-medical dataset has a non-linear with random structure and that Discrimination Analysis methods did not impose linear or non-linear constraints in ascertaining the information projection might be more suitable compared to MDA, PCA and Linear Discrimination Analysis for visualization and division of data classes in gene and protein patterns data. Now a day's non-linear Discrimination Analysis methods such are Isometric mapping , Spectral Clustering and Locally Linear Embedding had strengthened to decrease the data dimensionality without adopting a Euclidean relationship between information samples in the high dimensional space.

3. PROPOSED ALGORITHM:

Medical Gene or Protein Prediction Algorithm (MGPPA)

Input:

D: Gene or Protein Database $A_1, A_2, ...A_n$ are the Protein or Gene Attributes list. $G(V_m, E_n)$ be the Prediction Graph with m number of vertices and n number of edges. $|V|$ Total number of genes or proteins. $|E|$ Total number of connected genes or proteins. w_{ij} is the weight between ith vertex to jth vertex. $p_1, p_2....p_n$ are the partitions of the data D.

// Each partition sub graph should be connected with at least one path.

$c_1, c_2....c_n$ be the partition clusters.

Procedure:

Find all gene or protein patterns in D

For each data point p in D

do

$gen_i = find(D(p_1, p_2....p_n), Gpat, geneDB);$

$pro_i = find(D(p_1, p_2....p_n), Ppat, geneDB);$

done

For each attribute a_i *in* A

Do

Find the attribute gene or protein categories using geneDB.

$Gcat_i = find(a_i, gen_i);$

$Pcat_i = find(a_i, pro_i);$

done

// Find the relational attributes based on user selected feature.

For each attribute a_i in A

Do

for each property v_j in a_i

Do

If($input \in Gcat_i$)

$sim(a_i, v_j) = prob(\text{input}/ Gcat_i) \cup prob(v_j / Pcat_i) / prob(v_j / A)$

Else

$sim(a_i, v_j) = prob(v_j / Gcat_i) \cup prob(\text{input}/ Pcat_i) / prob(v_j / A)$

Done

For each attribute similarity a_i in A

Do

// Graph construction with weights

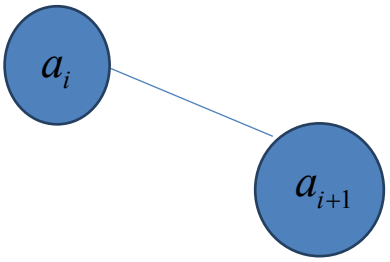
For each property v_j in a_i

Do

$w_{i,j} = sim(a_i, v_j)$

$w_{i+1,j} = sim(a_{i+1}, v_j)$

Construct a node with vertices a_i , a_{i+1} with weight e(a_i , a_{i+1})=($w_{i,j} + w_{i+1,j}$)/2.



Done

//Find the similarity path between user selected features with the gene or protein attributes.

For each node n in N

Do

$c_1, c_2....c_n$ =GetAllpath($n_i, n_{i+1}, w_{i,i+1}$)>simthreshold

Done

4. EXPERIMENTAL RESULTS

Proposed algorithm was implemented on large gene or protein databases by Loading Gene Ontology Dataset. Pattern identification is used on GeneDB to get the 5-Clustered, 3-Clustered, 6-Clustered Gene/Protein Result into the Feature Based Clustering Result. Proposed graph based visualization was executed on the XML dataset with multidimensional data. Experimental results show proposed approach generates cluster paths

Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

with attribute relationships, run the time performance analysis and Accuracy Comparison.

Fig 1: Loading Gene Ontology Dataset

Fig 1, describes the information about the gene data loading from the XML data file. After loading the gene dataset, the graph based model constructs the initial relationship between the genes.

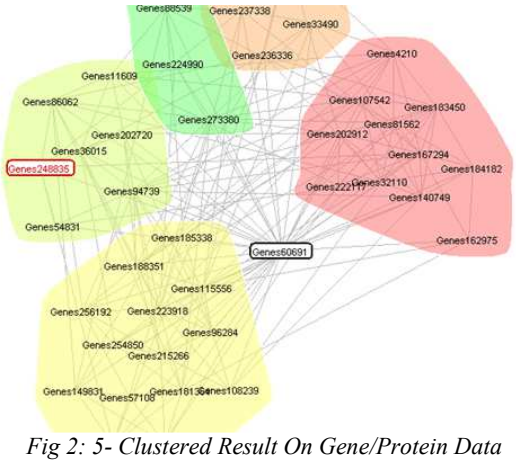


Fig 2: 5- Clustered Result On Gene/Protein Data

Fig 2 describes the gene/protein clustering result using the proposed graph based model.

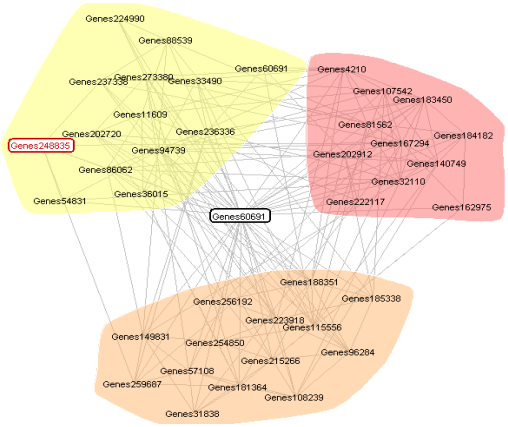


Fig 3: 3-Clustered Gene/Protein Result

Fig 3 , specifies the graph based clustering result, when the default clusters are specified as three. These clusters represent the intra cluster relationships among the gene/protein attributes.However, these clusters cannot specify the inter cluster relationship among the attributes.

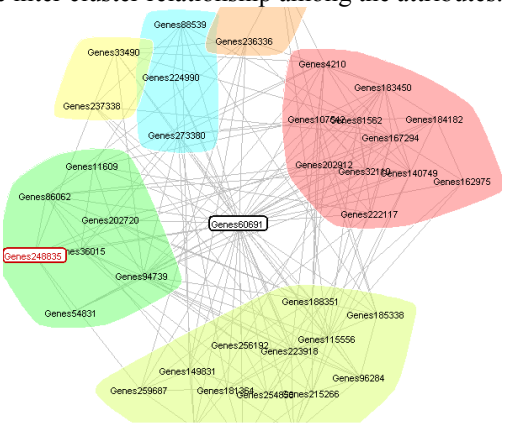


Fig 4: 6-Clustered Result

Fig 4 , specifies the graph based clustering result, when the default clusters are specified as six. These clusters represent the intra cluster relationships among the gene/protein attributes.However, these clusters cannot specify the inter cluster relationship among the attributes.

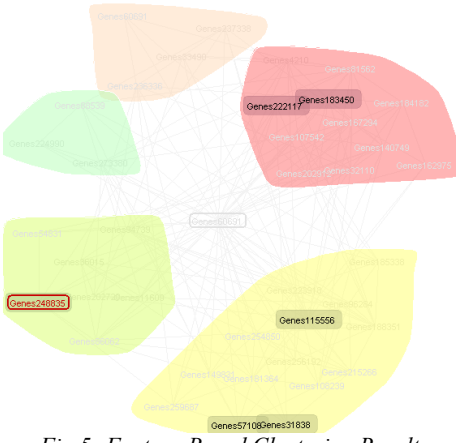


Fig 5: Feature Based Clustering Result

Fig 4 , specifies the graph based clustering result, when the default clusters are specified as five. These clusters highlight the intra cluster relationships among the gene/protein attributes.Also, these clusters specify the inter cluster relationship to the user selected feature.

Journal of Theoretical and Applied Information Technology
31st December 2015. Vol.82. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Runtime Performance Analysis

DataSize	PCA-LDA	BELKIN (LEM)	PROPOSED
2kb	45.64	46.23	36.77
5kb	67.33	73.55	62.14
10kb	125.46	154.33	113.23
20kb	373.12	397.53	307.44
100kb	1863.45	1933.23	1673.44

Table 1, describes the performance analysis of proposed model compare to the traditional models. In this table, proposed model has less execution time compared to traditional model in terms of memory size and time are concerned.

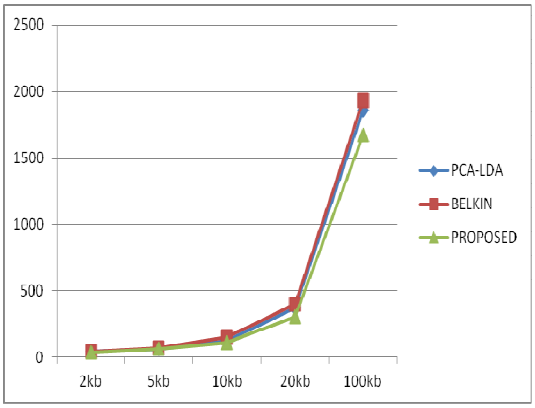


Fig 6: Runtime Comparison

Fig 6, describes the performance analysis of proposed model compare to the traditional models. In this table, proposed model has less execution time compared to traditional model in terms of memory size and time are concerned.

Table 2:Accuracy Comparison

DataSize	PCA-LDA	BELKIN	PROPOSED
2kb	73.56	74.66	93.45
5kb	74.23	75.64	94.67
10kb	78.35	79.31	93.56
20kb	76.33	79.44	95.67
100kb	79.33	81.44	94.78

Table 1, describes the performance analysis of proposed model compare to the traditional models. In this table, proposed model has high accuracy compared to traditional model in terms of recall and precision are concerned.In this table, as the true positive rate of the proposed model increases then the accuracy of the overall graph model increases.

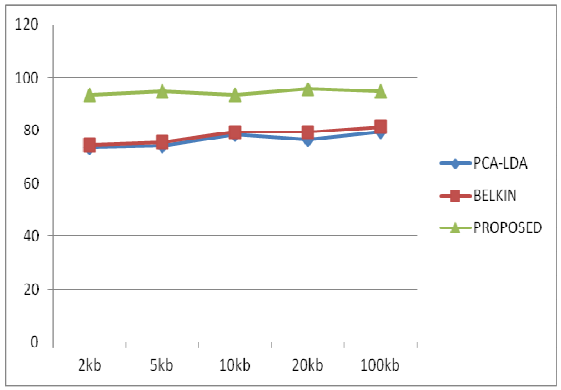


Fig 7, describes the performance analysis of proposed model compare to the traditional models. In this table, proposed model has high accuracy compared to traditional model in terms of recall and precision are concerned.In this table, as the true positive rate of the proposed model increases then the accuracy of the overall graph model increases.

5. CONCLUSION

In this proposed work graph based visualizing biomedical terms from abstracts using gene/protein features. Clustering each gene/ protein(s) based on protein/gene id, synonyms, name, category, patterns and its description. Easy to get all relations of gene/protein(s) using graph based techniques by using an improved gene or protein clustering algorithm on complex data. By using the user selection based clustering approach to medical data can be viewed in graphical format depends on the similarity measure. Experimental results show proposed graph model effectively clusters the complex gene or protein data compare to conventional clustering approaches.In future, this work can be extended to cloud based gene/protein similarity and classification computation on large distributed databases.

REFRENCES:

[1] L. Chen, X. Xu, Y. Chen, and P. He, "A Novel Ant Clustering Algorithm Based on Cellular Automata," in Proceedings – IEEE International Conference on Intelligent Agent Technology. IAT 2004, United States, 2004, pp. 148-154.

[2] X. Cui, J. Gao, and T. E. Potok, "A Flocking Based Algorithm for Document Clustering Analysis," Journal of System Architecture, no. Special issue on Nature Inspired Applied Systems July2006.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[3] S. Osinski and D. Weiss. Conceptual Clustering Using Lingo Algorithm: Evaluation on Open Directory Project Data, in Proc. Intelligent Information Processing and Web Mining Conf., Zakopane, Poland, Springer Physica-Verlag, 2004, pp. 369—378.

[4]. Y. Zhao, G. Karypis. Comparison of Agglomerative and Partitional Document Clustering Algorithms, The SIAM workshop on Clustering High-dimensional Data and Its Applications, Washington, DC, April 2002.

[5] Neeraj sahu, R. S. Thakur,G. S. Thakur, D. S. Rajput"Analysis of Social Networking sites using KMean clustering algorithm" 2012

[6] T. Theodosiou, N. Darzentas, L. Angelis, and C. Ouzounis, "PuReDMCL: A graph-based PubMed document clustering methodology," Bioinformatics, vol. 24, no. 17, pp. 1935–1941, Sep. 2008.

[7] D.Saravanan, Dr.S.Srinivasan, " A proposed New Algorithm for Hierarchical Clustering suitable for Video Data mining.", International journal of Data Mining and Knowledge Engineering", Volume 3.

[8] Bader Aljaber, Nicola Stoke, James Bailey and Jian Pei, "Document clustering of scientific texts using citation contexts," 2009.

[9] Anand Karandikar, "Clustering short status messages: A topic model based approach," 2010.

[10]S. Zhu, J. Zeng, and H. Mamitsuka,"Enhancing MEDLINE document clustering by incorporating mesh semantic similarity," Bioinformatics, vol. 25, no. 15, pp. 1944–1951, Aug. 2009.

[11] Shahana Bano and Dr. K. Rajasekhara Rao "Key Word Based Word Sense Extraction in A Index For TextFiles: Design Approach", CIIT International Journal Of Data Mining And Knowledge Engineering JAN '12.

[12] Shahana Bano and Dr. K. Rajasekhara Rao "Key Word Based Word Sense Extraction in Text: Design Approach", International Journal of Computer Science and Communication March'12. [13] Shahana Bano and Dr. K. Rajasekhara Rao "Pattern Based Gene/Protein Synonyms Identification from Biological Databases", International Journal of Applied Engineering Research (IJAER), Volume 9, Number 12 (2014).

[14] Shahana Bano and Dr. K. Rajasekhara Rao "Partial context similarity of gene/proteins in leukemia using context rank based hierarchical clustering algorithm" International Journal of Electrical and Computer Engineering vol 5(3), pp.483-490 (2015)

[15]Shahana Bano and Dr. K. Rajasekhara Rao " Context rank based hierarchical clustering algorithm on medical databases (Crbhca)" Journal of Theoretical and Applied Information Technology vol 75(2), pp. 199-211 (2015)