3rd International Conference on Bioinformatics and Biomedical Engineering, iCBBE 2009
11-13 June 2009, Page(s):1 - 4, Article number 5162877
DOI: http://dx.doi.org/10.1109/ICBBE.2009.5162877

Gene Expression Analysis Using Clustering

Kumar Dhiraj, Santanu Kumar Rath, and Abhishek Pandey
Dept of Computer science and Engineering,
National Institute of Technology Rourkela
Rourkela, Orissa-769008, INDIA
Email: kumardhiraj.nit.rourkela@gmail.com

***Abstracts*--Data Mining has become an important topic in effective analysis of gene expression data due to its wide application in the biomedical industry. In this paper, k-means clustering algorithm has been extensively studied for gene expression analysis. Since our purpose is to demonstrate the effectiveness of the k-means algorithm for a wide variety of data sets, we have chosen two pattern recognition data and thirteen microarray data sets with both overlapping and non-overlapping cluster boundaries, where the number of features/genes ranges from 4 to 7129 and number of sample ranges from 32 to 683. The number of clusters ranges from two to eleven. We use the clustering error rate (or, clustering accuracy) as evaluation metrics to measure the performance of k-means algorithm.**

Keywords-- Bio-informatics; Cancer-Genomics; Gene-expression; Clustering; Data-mining; Microarray.

I. INTRODUCTION

In recent years, the DNA microarray [17] has become an important and widely used technology since it enables the possibility of examining the expressions of thousands of genes simultaneously in a single experiment. The main algorithmic problem here is to cluster multi-conditions gene expression patterns. Basically, a cluster algorithm partitions entities into groups based on the given features of the entities, so that the clusters are homogeneous and well separated. A variety of clustering methods have been proposed for the mining of gene expression data [18], [19], [20], [22]. Although a number of clustering methods have been studied in the literature, they do not deal with clustering accuracy rather they generally deals with clustering validation metrics [21] to assess their performances.

The rest of the paper is organized as follows: In section II, K-means clustering is introduced. Experimental work conducted to evaluate the performance of the K-means clustering algorithm is presented in section III. Section IV deals with the result and discussion. Conclusions and future works are given in section V.

II. K-MEANS CLUSTERING ALGORITHM

The K-means algorithm [23], one of the most widely used clustering techniques. The steps of the K-means algorithm are described in brief as follows:

```
Step 1: Choose $K$ initial cluster centers $\{Z_{1,}Z_2,\ldots,Z_K\}$ randomly from the $n$ points $\{X_{1,}X_2,\ldots,X_n\}$.
Step 2: Assign point$\{X_i\}$, $i = \{1,2,\ldots,n\}$. to cluster $\{C_{j,}\}$.
  $j \in \{1,2,\ldots,K\},\ iff$   $\|X_i - Z\| < \|X_i - Z_p\|,\ p = \{1,2,\ldots,K\}, and\ j \neq p$.
  Ties are resolved arbitrarily.
Step 3: Compute new cluster centers $\{z_1^*, z_2^*, \ldots, z_k^*\}$ as follows:
  $z_i^* = \frac{1}{n}\sum_{x_j \in C_i} X_j$   $i = 1,2,\ldots,K,$
  Where $n_i$= is the number of elements belonging to cluster$C_{i.}$
  Step 4: If $z_i^* = Z_{i,}$ , i=1,…,K then terminate. Otherwise repeat from step 2.
```

Note that in case the process does not terminate at step 4 normally, then it is executed for a maximum "*fixed number of iterations*".

III. EXPERIMENTAL WORK

In this section, we describe the datasets used for assessing the performance the k-means algorithm, which are listed in Table 1. The first two datasets represent pattern recognition data, while the other represents gene-expression microarray data. For pattern recognition, we use IRIS [1], [2] and WBCD [2].[16] data and for microarray data we use serum data (Iyer et. al) [3], [4], [8], [14], yeast data (Cho et. al) [5], [6], [7], [8], [14], leukemia data (Golub et. al) [8], [14], [23], breast data (Golub et. al) [9], [14], Lymphoma data (Alizadeh et al.)[9], [10], [11], lung cancer (Bhattacharjee et. al)[9], [12], and St. Jude leukemia data (Yeoh et. al)[9], [13]. To identifying common subtypes in independent disease data we use four different types of breast data (Golub et. al) and four DLBCL (Diffused Large B-cell Lymphoma) data.

IV. RESULT AND DISCUSSIONS

Table 1 represents summary of k-means clustering algorithm result for fifteen datasets (fig. 1). It also consists some of the relevant characteristics, such as number of classes, number of features/genes and the number of item samples for each datasets.

Table 2 represents the result of IRIS data. We have achieved accuracy up to 88.67%. The total count error in this case is 17. Note that we have achieved 100% accuracy for cluster 1.

978-1-4244-2902-8/09/$25.00 ©2009 IEEE

Table 3 represents the result of WBCD data. 62.81% accuracy we have achieved in this case.

Table 4 represents the result of Serum data (Iyer data). We have achieved accuracy up to 51.84%.

Table 5 represents the result of Yeast data (Cho data). We have achieved accuracy up to 60.88%. Note that cluster 4 is having accuracy approx. 12%. The reason for this is the data belonging to this cluster are very overlapping in nature with other clusters.

Table 6 represents the result of Leukemia data. We have achieved accuracy up to 59.72%. The Golub et al.'s microarray data set is very challenging because the appearance of the two types of acute leukemia is highly similar in nature. This was the reason we have not achieved more accuracy in this case. One probable solution to deal with this problem is that we can use dimensionality reduction techniques to reduce the number of feature and then apply k-means clustering algorithm.

Table 7 represents the result of Lung Cancer. We have achieved accuracy up to 72.08%. In this, cluster 2 and cluster 4 are highly separable in nature compare to cluster 1 and cluster 3. We achieve approx. 95% accuracy for cluster 2 and cluster 4 whereas for cluster3 we got least accuracy (47%).

Table 8 represents the result of St. Jude Leukemia data. We have achieved accuracy up to 85.08%. The data in this case is highly separable in nature. We achieve 100% accuracy for cluster 2 and least one we got for cluster 5.

Due to constraint of space; we have not included the details result for subtype for microarray breast cancer and DLBCL (Diffused Large B-cell Lymphoma) data. Summary of result for these subtypes has been shown in Table 1. As far as microarray breast cancer data is concerned, we achieved maximum accuracy for Breast Multi data A (79.61%) whereas the least accuracy for Breast data B (53.06%). The reason of this could be probably Breast data B is more overlapping in nature and is having nonlinear structure.

AS far as subtypes of DLBCL are concerned, DLBCL D is of highly overlapping nature and that's why we have achieved least accuracy 42.64% in this case. The data of DLBCL B is of highly distinctively separated in nature compare to other DLBCL(A, C, D) and that is the reason we have achieved higher accuracy in case of DLBCL B.

V. CONCLUSIONS AND FUTURE WORK

Clustering is an efficient way of analyzing information from microarray data and K-means is a basic method for it. Although it is easy to implement and understand, K-means has some drawbacks [24], [25]. It is because of these drawbacks several approximate methods are developed to solve the underlying optimization problem. In the future, we plan to study K-Means data clustering with other heuristic based search methods like simulated annealing or some others.

REFERENCES

[1]. E. Anderson, "The IRISes of the Gaspe Penisula," Bulletin of the American IRIS society, vol. 59, pp. 2-5. 1939.

[2]. http://archive.ics.uci.edu/ml/datasets

[3]. http://www.sciencemag.org/feature/data/984559.shl

[4]. Iyer,V.R., Eisen,M.B., Ross,D.T., Schuler,G., Moore,T., Lee,J.C.F, Trent,J.M., Staudt,L.M., Hudson Jr,J., Bogosk,M.S. et al. The transcriptional program in the response of human fibroblast to serum. Science, 283, 83–87, 1999.

[5]. Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. and Davis,R.W. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell., 2, 65–73, 1998.

[6]. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. Systematic determination of genetic network architecture. Nat. Genet., 22, 281–285, 1999.

[7]. Doulaye Dembélé, and Philippe Kastner, "Fuzzy Cmeans method for clustering microarray data", Bioin**for**matics, Vol. 19, no. 8, Pages 973-980, 2003.

[8]. http://www.cse.buffalo.edu/faculty/azhang/Teaching/index.html

[9]. http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi

[10]. A. A. Alizadeh et al., "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," Nature, vol. 43, pp. 503-511, 2000.

[11]. Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, , and E. Lander: 'Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression'. Science 286(5439), 531–537, 1999.

[12]. Bhattacharjee, A., W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, and M. Meyerson, 'Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinomas Sub-classes'. Proceedings of the National Academy of Sciences 98(24), 13790–13795, 2001.

[13]. Yeoh, E.-J., M. E. Ross, S. A. Shurtleff, W. K. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, C. Cheng, D. Campana, D. Wilkins, X. Zhou, J. Li, H. Liu, C.-H. Pui, W. E. Evans, C. Naeve, L. Wong, and J. R. Downing, 'Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling'. Cancer Cell 1(2), 2002.

[14]. http://www-igbmc.u-strasbg.fr/projets/fcm.

[15]. Y Hoshida, JP Brunet, P Tamayo, TR Golub, JP Mesirov, "Subclass mapping: identifying common subtypes in independent disease data sets", PLoS ONE, Vol. 2, No. 11, 2007, 2002.

[16]. O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September, pp 1 & 18, 1990.

[17]. J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, Y. Chen, Y. A. Su, and J. M. Trent, "Use of a cDNA microarray to analyze gene expression patterns in human cancer," Nature Genetics, Vol. 14, pp. 457-460, 1996.

[18]. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," in Proceedings of National Academy of Science, Vol. 96, pp. 6745-6750, 1999.

[19]. A. Ben-Dor and Z. Yakhini, "Clustering gene expression patterns," Journal of Computational Biology, Vol. 6, pp. 281-297, 1999.

[20]. M. B. Eissen, P. T. Spellman, P. O. Brown, and D. Botstein, "Clustering analysis and display of genome wide expression patterns," in Proceedings of the National Academy of Sciences, Vol. 95, pp. 14863-14868, 1998.

[21]. D Jiang, C Tang, A Zhang," Cluster analysis for gene expression data: a survey", Knowledge and Data Engineering, IEEE Transactions on, Vol. 16, No. 11, pp. 1370-1386, 2004.

[22]. G. P. Shapiro, T. Khabaza and S. Ramaswamy, Capturing best practice for microarray gene expression data analysis'', SIGKDD '03, August 24-27, 2003.

[23]. A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[24]. S.Z. Selim, M.A. Ismail, K-means type algorithms: a generalized convergence theorem and characterization of local optimality, IEEE Trans. Pattern Anal. Mach. Intell. 6, 81-87, 1984.

[25]. H. Spath, Cluster Analysis Algorithms, Ellis Horwood, Chichester, UK, 1989.

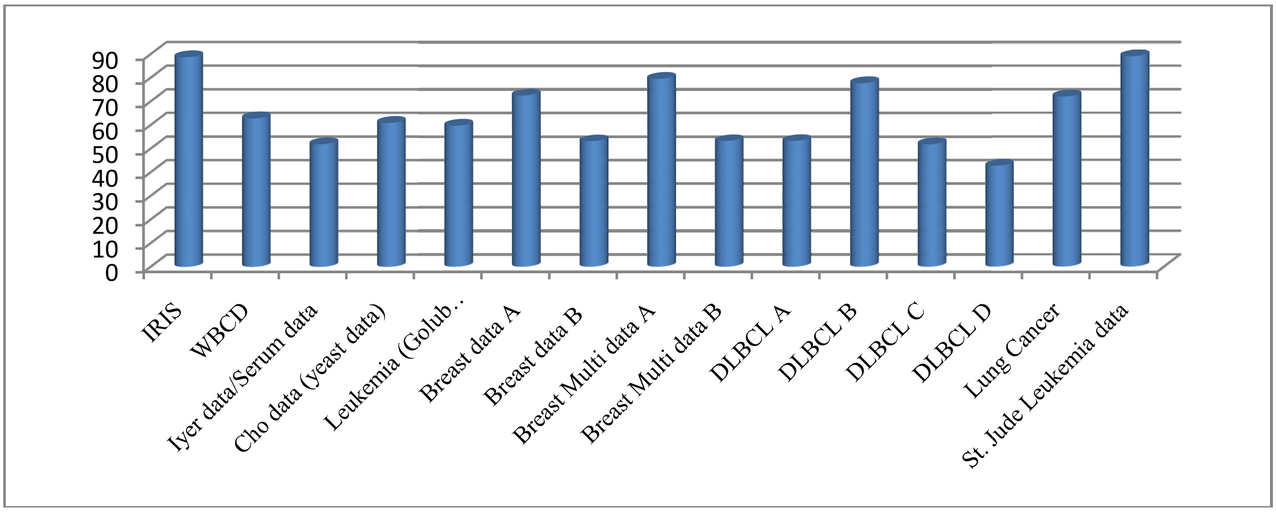

Fig. 1. Clustering accuracy for k-means for pattern recognition as well as microarray data, Horizontal axis represents different Datasets whereas vertical axis represents accuracy in percentage.

TABLE 1: COMPARISON OF RESULTS FOR ALL FIFTEEN DATASETS

Datasets	Primary source	Secondary source	Dimension	# of cluster	K-means			
					# Error	Error(%)	# correct	Accuracy(%)
IRIS	[1]	[2]	[150x4]	3	17	11.33	133	88.67
WBCD	[16]	[2]	[683x9]	2	254	37.19	429	62.81
Iyer data/Serum data	[4]	[3], [8], [14]	[517x12]	11	249	48.1624	268	51.84
Cho data (yeast data)	[5],[6]	[7], [8], [14]	[386x16]	5	151	39.1191	235	60.88
Leukemia Golub Experiment)	[23]	[8], [14]	[72x7129]	2	29	40.28	43	59.72
Lung Cancer	[12]	[9]	[197x581]	4	55	27.92	142	72.0812
St. Jude Leukemia data	[13]	[9]	[248x985]	6	37	14.91	211	85.08
Breast data A	[15]	[9]	[98x1213]	3	27	27.55	71	72.44
Breast data B	[15]	[9]	[49x1024]	4	23	46.93	26	53.0612
Breast Multi data A	[15]	[9]	[103x5565]	4	21	20.39	82	79.61
Breast Multi data B	[15]	[9]	[32x5565]	4	15	48.88	17	53.125
DLBCL A	[10],[11]	[9]	[141x661]	3	66	46.8085	75	53.191
DLBCL B	[10],[11]	[9]	[180x661]	3	40	22.22	140	77.78
DLBCL C	[10],[11]	[9]	[58x1772]	4	28	48.28	30	51.7241
DLBCL D	[10],[11]	[9]	[129x3795]	4	74	57.36	55	42.64

Note: Colored index(Secondary source) indicates the reference from where we have downloaded data.

TABLE 2: RESULTS FOR IRIS DATA

	Cluster 1	Cluster 2	Cluster 3	Total
The right number of data point	50	50	50	150
number of data point wrongly clustered	0	4	13	17
number of data point correctly clustered	50	46	37	133
Accuracy(%)	100	92	74	88.67

TABLE 3: RESULTS FOR WBCD DATA

	Cluster 1	Cluster 2	Total
The right number of data point	444	239	683
number of data point wrongly clustered	142	112	254
number of data point correctly clustered	302	127	37.19
Accuracy(%)	68.01	53.14	62.81

TABLE 4: RESULTS FOR SERUM DATA (IYER DATA)

	# data point	# data point wrongly clustered	# data point correctly clustered	Accuracy
Cluster1	33	17	16	48.48
Cluster2	100	84	16	16
Cluster3	145	7	138	95.17
Cluster4	34	34	0	0
Cluster5	43	31	12	27.91
Cluster6	7	6	1	14.29
Cluster7	34	17	17	50
Cluster8	14	1	13	92.86
Cluster9	63	28	35	55.56
Cluster10	19	12	7	36.84
Cluster11	25	12	13	52
Total	517	249	268	51.84

TABLE 5: RESULTS FOR CHO DATA

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
The right number of data point	67	135	75	54	55	386
number of data point wrongly clustered	30	25	28	47	21	151
number of data point correctly clustered	37	110	47	7	34	235
Accuracy(%)	55.2	81.5	62.7	12.9	61.8	61

TABLE 6: RESULTS FOR LEUKEMIA DATA (GOLUB EXPERIMENT)

	Cluster 1	Cluster 2	Total
The right number of data point	47	25	72
Number of data point wrongly clustered	15	14	29
Number of data point correctly clustered	32	11	43
Accuracy(%)	68.09	44	59.72

TABLE 7: RESULTS FOR LUNG CANCER

	# data point	# data point wrongly clustered	# data point correctly clustered	Accuracy
Cluster1	139	42	97	69.78
Cluster2	17	1	16	94.11
Cluster3	21	11	10	47.62
Cluster4	20	1	19	95
Total	197	55	142	72.08

TABLE 8: RESULTS FOR ST. JUDE LEUKEMIA DATA

	# data point	# data point wrongly clustered	# data point correctly clustered	Accuracy
Cluster1	15	15	0	0
Cluster2	27	0	27	100
Cluster3	64	3	61	95.31
Cluster4	20	4	16	80
Cluster5	43	14	29	67.44
Cluster6	79	1	78	98.73
Total	248	37	211	85.08