International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.5, No.7, July 2016 DOI:10.15693/ijaist/2016.v5i7.68-73

Threshold Clustering Technique for Gene Selection of T2DM

K.Vijayalakshmi
Research Scholar
Dept. of Computer Science
S.V.University, A.P, India

Prof.M.Padmavathamma
Dept. of Computer Science
S.V.University
A.P, India

ABSTRACT: Clustering is a threshold implementation methodology that has been proposing to reduce the dimensionality of the data in order to simplify subsequent data analysis and allow for summarization of the data in a cost-conscious manner. It is also becoming easy and done by microarray data analysis tool. For a typical the microarray data set, it is often difficult to compare the overall gene expression difference between observations from different groups or conduct the classification based on a very large number of genes. In this research, we propose a threshold gene selection methodology based on the strategy used by K-means Clustering. We demonstrate the effectiveness of this procedure using T2DM gene expression data set and compare it with different k values and consider ESS value where the local optimal value is minimized. It turns out that the proposed method selects the best gene subset for preserving the original data structure and also easy to make best relevant gene selection from cluster groups.

KEYWORDS: T2DM, Clustering, Microarray Gene Expression Data, K-means clustering

I. INTRODUCTION

Preparing data for a machine learning algorithm is a big challenge. This chapter aims to provide how to do analysis, summarization of gene expression using unsupervised learning technique. Generally, DNA microarray analysis monitors the expression of thousands of genes all together within in a single RNA sample. This analysis enables the measurement of thousands of genes at time. Researchers can analyze the global patterns of gene expression of an organism during specific physiological responses or developmental processes by coupling microarray analysis with the results from genome sequencing projects. In our work we collected microarray data and grouped into clusters to analyze the similarity between different genes. In section 2 we are describing about importance of microarray experiments, analysis of gene expression data obtained from it. In section 3 we are exploring the importance of Data Reduction and in section 4 briefly discussing about threshold methodology for identification of genes using unsupervised learning techniques. Section 5 explains how to choose the appropriate gene clusters by following Elbow Method based on SSE value and conclusions of this chapter are presented at section 6

2 RELATED WORK

Considerably voluminous datasets of information derived from variety of biological experiments and also analyzed by well designed genomics. Gene Expression Analysis is one type of experiment that is monitoring the expression levels of thousands of genes simultaneously under a particular condition. It is possible with Microarray technology where the quantity of data engendered from each experiment is profoundly and immensely massive. This technology has become one of the essential implements for biologists to monitor genome wide expression levels of various genes in a given organism. Analysis of gene expression data can be relegated into two different type's namely supervised or unsupervised learning. In the case of a supervised learning, we do utilize the annotation of either the gene or the sample, and engender clusters of genes or samples in order to identify patterns that are characteristic for the cluster. In the case of an unsupervised learning, the expression data are analyzed to identify patterns that can group genes or samples into clusters without the utilization of annotation. However, annotation information may be taken into account at a later stage to make consequential biological inferences. Set of genes or set of experimental conditions that has homogeneous expression profiles will be referred to as a cluster. Thus, a cluster consists of objects with homogeneous expression profiles, where an object may either refer to genes or samples.

3. IMPORTANCE OF DATA REDUCTION

In the modern world data is becoming progressively more dimensional. One of the widespread examples is in the field of Bioinformatics where the machine learning is in utilization to make prognostications on data which are capable to understand patient’s health quandary. The challenge that this presents to the machine learner is what to do with all this data? Data pre-processing is a considerable solution for the above challenge. Raw data is highly susceptible to noise, missing values, and eroticism. Data mining results are depending on the quality of the data. In order to amend the quality, raw data is pre-processed so as to ameliorate the efficiency and facilitate of the mining process. Transformation of the initial data set for mining process will be after Pre-processing only. This Data Preparation Process can be done by iteratively with many loops in different steps like Data Cleaning, Data Integration, Data Transformation and Data Reduction.

4. IDENTIFYING GENES USING THRESHOLD UNSUPERVISED TECHNIQUES

Data clustering is the process of grouping data elements predicated on some aspect of homogeneous attribute between the elements in the group. Clustering has many applications such as data mining, statistical data analysis and bioinformatics. It is additionally utilized for relegating substantial amount of data, which in turn is utilizable when analyzing data engendered from search engine queries, articles and texts, images etc. In our work we have taken into the consideration of Hierarchical clustering and Centroid-predicated clustering using K-means algorithm.

Datasets Used in Experiments

"T2D-Db" is a widespread database which presents in sequence about the feature involved in the pathogenesis of Type 2 Diabetes Mellitus (T2DM) in humans. Information that concerning to the Type 2 diabetes can additionally be probed in T2D-Db utilizing kenned risk factors for this disease. From the repository of GEO Datasets Microarray Expression data have been accumulated that cognate to the Type 2 diabetes. In our work we concentrated on the following list of genes microarray expression data and implemented clustering techniques. The below Data set is normalized by Min-max normalization and transform the data into a new range between [0, 1].

	GSM47885	GSM47886	GSM47887	GSM47888	GSM47889	GSM47890	GSM47891	GSM47878	GSM47879	GSM47880
ADRB3	43.9	40.5	40.7	41.7	28.3	40.2	44.3	53.2	44.3	24.5
APOA2	325.1	42.7	189.3	29.4	60.5	128.1	212.9	64	168.4	175.3
BDNF	2071.4	2168.9	3760.7	1415.4	1895.2	3076.4	967.2	2721.7	1757.9	3700
C4A	1074.8	1150	1534.2	903.1	1002.7	579.9	1285.5	1045.2	607.7	1048.1
CACNA1D	1260.6	1433.7	1216.7	1213.1	1245.1	968.8	1564.4	1096.3	1150	1145.2
CACNB3	1821.7	1103.8	1509.9	1417.5	1428.3	1577.9	1491.3	1751.6	1651.1	2022.4
CAMK2G	1973.3	2435.4	2486.1	2246.1	2252.1	2939.2	1709.2	2992.7	2400.5	2758.2
HFE	438.6	541.2	479.4	481.2	357.3	439.1	604	484.4	491	576.3
HIF1A	62255.4	64092.5	57803.9	58372.2	55793.4	56082.2	63822.8	56295.4	54205.2	51455.7
HK2	2851.3	4061.6	3384.9	3812.8	3741.6	3060.5	3358.7	3121.7	4923.3	3447
HNF4A	355.3	395	187.7	181.7	110.5	626.4	132.7	738.5	196.7	614.7
IGF2	85.4	53.9	88.7	61.8	57.2	98.8	66.2	66.2	48.5	51.3
KCNA3	42.2	100.7	91.8	106.5	105.6	110.1	129.6	102.5	62.6	149.8
KCNJ10	367.1	154.6	276	412.5	416.3	505.8	238.1	604.5	368.6	329.4
TCF7L2	18353.2	15124.7	17294.8	18109.9	12673	13819.7	18547.3	12240.3	17734.9	17498.3

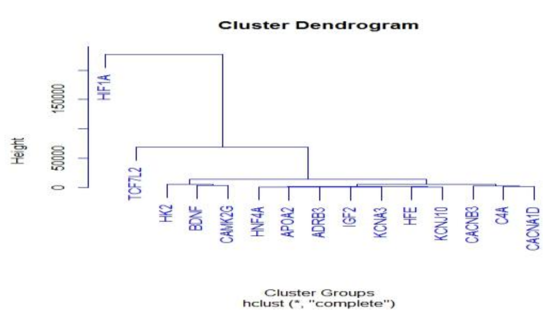
	GSM47881	GSM47882	GSM47883	GSM47884	GSM47899	GSM47900	GSM47901	GSM 47902	GSM47903	GSM47904
ADRB3	55.5	53.2	50.2	28.4	33.1	39.1	54.5	63.5	53.1	34.6
APOA2	163	243.2	13	146.8	322.4	153.8	47.1	68.1	42.6	189.3
BDNF	3511.6	1624.6	1544.7	3478.5	1582.4	1599	1738.1	2376.6	2223	2985.4
C4A	1352.7	1090.8	942.3	899	1127.1	995.1	1188.4	1217.1	1005.4	677.6
CACNA1D	1774.6	1580.9	1336.6	1171.5	1754.3	1151.7	1190.1	1629.1	990.2	1258.8
CACNB3	1339.5	2392.5	1221.4	1480	1636.7	1072.8	1907.6	844.3	1766.8	1463
CAMK2G	2775.7	1907.3	2093.1	2300.6	3121.8	1731.9	2714.6	1691.5	2013.7	2541.3
HFE	706.4	445.2	683.3	636.3	456.2	328.1	337.7	747.7	405.2	535.6

HIF1A	54533.7	54877.3	51544.8	55611.5	58194.8	56867.1	62570.2	64465.3	53494.5	54190.4
HK2	3074.5	4174.6	3635.6	3027.1	3576	2296.5	3353.5	2845	3044.2	2920.6
HNF4A	193.5	427.2	165.5	151.8	625.4	368.5	224.6	223.1	450.6	158.2
IGF2	119.6	59.1	74.4	60.4	109.9	62.4	61.2	104.4	49.7	98.2
KCNA3	106.3	96	85	105	46.6	86.9	97.5	163.7	65.1	109.3
KCNJ10	600.1	139.1	134.7	96.1	83.3	219.1	527.6	208.1	43.2	454.6
TCF7L2	23378.9	18864.7	12585.2	13915.8	19994.9	18502.9	20269.8	17844.3	19692	23670.6

	GSM47905	GSM47892	GSM47893	GSM47894	GSM47895	GSM47896	GSM47897	GSM47898
ADRB3	36.5	51.1	38.5	31.6	32.9	46.3	52.4	34
APOA2	66	163.7	51.9	53.6	93.5	47.7	404.6	180.9
BDNF	1284.8	1770	2383.3	2671	1376.4	2035.9	2414.6	2019.6
C4A	1119.1	1105.3	1129.2	1022	864.8	509.3	1094.5	990.4
CACNA1D	1338.2	860.8	1017.3	980.6	1230.9	1264	1629.6	1058
CACNB3	1546.2	1806.4	1434	833.6	1709.9	1569.3	1424.5	1709.7
CAMK2G	3055	1803.7	2063.1	2578.2	2158.1	3296.1	2805.5	2499.5
HFE	658.5	418.4	352.3	481.6	591.1	667.7	510.3	507.7
HIF1A	47733.1	50138.5	59426.8	61041.3	52481.1	51933.2	55741.3	49520.1
HK2	3001	2788.3	2546.4	3494	3676.2	2822.5	3743.5	3903.3
HNF4A	580.8	132.1	183.9	370.3	573.5	150.5	293.4	152.5
IGF2	55.6	85.1	63.7	91.9	56.2	67	119.2	52.8
KCNA3	105.6	94.4	129.2	84.2	122.9	111.3	181.3	98.5
KCNJ10	443.3	676.9	364.4	461.4	394.4	729	543.5	521.3
TCF7L2	12071.8	14759.1	19179.8	19008.8	16324.8	11517.8	16162.1	13963.8

4.1 GENE SELECTION USING HIERARCHICAL CLUSTERING

In our present work we implemented hierarchical clustering which is the represented in the Connectivity of the models and here we build model predicated on distance connectivity. We can define several different ways of quantifying distance or dissimilarity the between the rows or between the columns of the data matrix, depending on the quantification scale of the observations. In this work we shall consider a graphical representation of a matrix of distances which is perhaps the most facile to understand – a dendrogram, or tree – where the objects are joined together in a hierarchical fashion. By the implementing Hierarchical clustering, we get the following dendrogram as a result as.



4.2 GENE SELECTION USING K-MEANS CLUSTERING

We also implemented an idea of Centroid models by using k-means algorithm for the representation of each cluster of a single mean vector. The important idea of centroid based clustering is each data points belong to the cluster, which center is within closest distance of that data point. Also clusters are represented by a central vector which may not essentially be a portion of the data set. We use K-means clustering which is one of the simplest unsupervised learning algorithms. It is a simple, easy way to classify a given data set through a definite number of clusters by assuming a k value. It uses the squared Euclidean distance to assign objects to clusters. Because of this, the K-means algorithm proceeds to try to find a minimum of Error Sum of Squares (SSE). It is calculated by

$$\mathrm{ESS} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

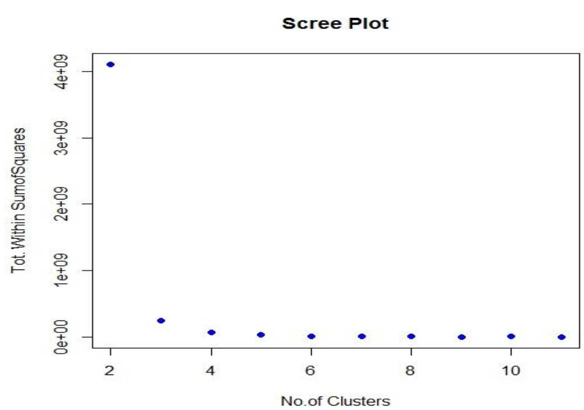
We repeated the above steps no. of times by changing centers=2, 3, 4….10 and tabulated obtained clusters sizes, clustering vectors and % of SumofSquares as follows in the given table...

Centers	Size of Clusters	% between_SS / total_SS	Clustering Vectors
2	14,1	91.6%	1{ADRB3, APOA2, BDNF , C4A , CACNA1D, CACNB3,CAMK2G,HFE,HK2,HNF4A,IGF2, KCNA3,KCNJ10,***TCF7L2***}; 2{HIF1A};
3	13,1,1	99.5%	1{ADRB3, APOA2, BDNF , C4A , CACNA1D, CACNB3,CAMK2G,HFE,HK2,HNF4A,IGF2, KCNA3,KCNJ10 }; 2{HIF1A}; ***3{ TCF7L2};***
4	10,1,1,3	99.9%	1{ADRB3, APOA2, C4A , CACNA1D, CACNB3, HFE, HNF4A,IGF2,KCNA3,KCNJ10 }; ***2{ TCF7L2};*** 2{HIF1A}; 4{ BDNF, CAMK2G, HK2,};
5	3,3,7,1,1	99.9%	1{C4A , CACNA1D, CACNB3}; 2{ BDNF,CAMK2G, HK2}; 3{ADRB3,APOA2,HFE,HNF4A,IGF2,KCNA3,KCNJ10}; ***4{ TCF7L2};*** 5{ HIF1A};
6	7,2,3,1,1,1	100%	1{ADRB3, APOA2, HFE,HNF4A,IGF2, KCNA3,KCNJ10 }; 2{ BDNF,CAMK2G }; 3{C4A , CACNA1D, CACNB3}; 4{HIF1A}; 5{HK2}; ***6{ TCF7L2};***
7	1,3,7,1,1,1,1	100%	***1{ TCF7L2};*** 2{ C4A , CACNA1D, CACNB3} 3{ ADRB3, APOA2, HFE, HNF4A,IGF2,KCNA3,KCNJ10} 4{HK2}; 5{HIF1A}; 6{ CAMK2G} 7{ BDNF}
8	2,1,1,1,1,5,2,2	100%	1{ HFE,KCNJ10}; 2{HIF1A}; 3{HK2}; ***4{TCF7L2};*** 5{CACNB3}; 6{ADRB3,APOA2,HNF4A,IGF2,KCNA3}; 7{CAMK2G}; 8{ C4A,CACNA1D};
9	5,1,1,1,2,1,2,1,1	100%	1{ADRB3,APOA2,HNF4A,IGF2,KCNA3}; 2{HIF4A}; 3{HK2}; 4{CACN1D}; 5{BDNF,CAMK2G}; ***6{TCF7L2};*** 7{HFE,KCNJ10}; 8{CACNB3}; 9{C4A};
10	2,1,1,1,2,4,1,1,1,1	100%	1{C4A,CACNA1D}; 2{HNF4A};3{HK2}; 4{CACNB3}; 5{BDNF,CAMK2G}; 6{ ADRB3,APOA2,IGF2,KCNA3}; 7{KCNJ10}; 8{HFE}; ***9{TCF7L2};*** 10{HIF4A};

The input data set run number of times by using K-means and recording the value of SSE at each time. It depends on the Local optimal value of SSE where SSE is smaller compared to any other possible solution. We have found the peak of a function in a small part of the space and therefore we have to believe a cluster with lowest SSE for usage.

5. SELECTION OF APPROPRIATE GENE CLUSTER BASED ON SSE VALUE

After repetition of K-means clustering algorithm we have chosen the appropriate cluster solution by comparing the Sum of Squared Error (SSE) for a number of cluster solutions. It defines value of SSE between each member of a cluster and its cluster centroid. Therefore, it can be considered as a global measure of error. In general, number of clusters increases, the SSE value should decrease because clusters are smaller in size. We plot the values of the SSE against a series of sequential cluster levels in a graphical way to choose an appropriate cluster. It produces an "elbow" in the plot of SSE against cluster solutions. As shown below, there is an "elbow" at the 3 cluster solution suggesting that solutions >3 do not have a substantial impact on the total SSE.



Accordingly, many times TCF7L2 appearance as a single cluster. And also it there in cluster=3 with clustering vector as 1{ADRB3, APOA2, BDNF,C4A,CACNA1D, CACNB3,CAMK2G,HFE,HK2,HNF4A,IGF2,KCNA3,KCNJ10}; 2{HIF1A}; 3{TCF7L2}; When we plot the scree plot also this is the cluster we have consider because of its optimal value of SEE. Also strongly by Review of Literature TCF7L2 is the most relevant suspicious gene which affects insulin secretion and glucose production

6. CONCLUSION

If data encountered has many features we may face many problems for analysis, for example `Genomic' data in Bioinformatics. This makes it difficult to process and comprehend as many features are often not relevant. So we used unsupervised learning techniques i.e. hierarchical clustering and k-means clustering method to reduce the number of attribute values. Using K-means clustering we make different data objects /clusters which are robust and easier to understand and produces relatively efficient clusters on each object. In our work we considered cluster size of 3 with local optima value and also obtained Elbow in the Scree Plot. Among the clusters we had chosen TCF7L2 from obtained clustering vector that is more relevant to T2DM.

REFERENCES

[1]. The Genetics Of Type 2 Diabetes Mellitus : A Review by Sunita Singh, Department of Zoology, Mahila ahavidyalaya, Banaras Hindu University, Varanasi.

[2]. k-means and Hierarchical Clustering by Andrew W. Moore.

[3]. Hierarchical Document Clustering by Benjamin C. M. Fung, Ke Wang and Martin Ester.

[4]. How to explain Hierarchical Clustering by S. P. Borgatti.

[5]. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html

[6]. BIRCH: An efficient data clustering method for very large Databases by T. Zhang, R. Ramakrishnan and M. Livny.

[7]. An Efficient k-means Clustering Algorithm: Analysis and Implementation by Tapas Kanungo,DavidM.Mount,Nathan S.Netanyahu, Christine D. Piatko, Ruth Silverman and AngelaY.Wu.

[8]. Research issues on K-means Algorithm: An Experimental Trial Using Matlab by Joaquin Perez Ortega, Ma. Del Rocio BooneRojas and Maria J. Somodevilla Garcia.

[9]. The k-means algorithm - Notes by Tan, Steinbach, KumarGhosh.

[10]. k-means clustering by ke chen.

[11]. http://t2ddb.ibab.ac.in/microarray.shtml#browse

[12]. http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/chapter-final.pdf

[13]. "Identifying Variant Positions of TCF7L2 Gene related to T2DM: by K.Vijayalakshmi, Prof.M.Padmavathamma published in International Journal of Innovative Research in Science,Engineering and Technology (IJIRSET) Vol.4, Issue 10, October 2015. ISSN: 2319-8753 (Online).

[14]. DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly Department of Computer Science, BITS Pilani Dubai, United Arab Emirates, IJDKP) Vol.5, No.1, January 2015.

Authors Profile

Smt K.Vijayalakshmi received MCA degree from Sri Padmavathi Mahila Visvavidyalayam,Tirupati. Presently working as Asst.Professor in the Dept.of Computer Science, SVUniversity,Tirupati. Pursuing her Ph.D(part-time) from the same Dept. Her research interest includes Expert Systems, Data Mining and Software Engineering.

Prof.M.Padmavathamma Head, Dept.of Computer Science and Research Supervisor. Her research interest includes Cryptography & Network Security, DM. She published nearly 142 articles. She published 4 books. She received State Best Teacher Award (Inter University Level in Engineering Subject), announced by APSCHE, Hyderabad for the year 2014.

Details of Conference Publication