International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 4 (2015) pp. 10783-10793

http://www.ripublication.com

Prediction Of Genetic Disorder By Using Phenotypic And Genotypic Information Using Clustering Approach

Bipin nair B J, Bhuvana P and Priyanka G

Amrita Vishwa Vidyapeetham, Mysore Campus, Karnataka, India
bipin.bj.nair@gmail.com, pbhuvanaram@gmail.com, priyankaharitsa@gmail.com

Abstract

A method for predicting the genetic disorder from the phenotypical appearance in human body is presented. The information is contain in the dataset and the dataset will be preprocessed to reduce noise and redundancy then the information is clustered by using K-means algorithm according to their phenotypic similarity. Each cluster will consist of phenotypically similar information which belongs to a particular genetic disease (alpha-1 antitrypsin). The main advantage is it is more suitable for large dataset. This proposed method will reduce the risk of misconception in predicting of genetic diseases.

Keywords— Phenotype , K-means,Genotype.

I. INTRODUCTION

Predicting genetic diseases by genomic information is an important research topic in bioinformatics. Bioinformatics is an application to the management the biological information by using computer technology. A genetic disorder is an illness caused by one or more abnormalities in the DNA which will show up by birth or in the development stage of the body. We can identify the genetic disorder by using genotypic and phenotypic information. Genotypic contains the hereditary information of an organism and phenotype is a description of your actual physical characteristics. This information is used in making of dataset and analysis process.

Many researches on the prediction of genetic disorder has showed that the particular disorder can be identified by analyzing the gene expression data, analyzing the single point protein mutations, single amino acid polymorphism and so on. In this paper we are predicting the genetic disorder by analyzing the phenotypic information by using K-means algorithm and predict how many patients are affected by the diseases and how many are not.

K-means is a partitioning algorithm which is used to divide the data into small number of clusters. In this paper we are mainly concentrating on lungs genetic disease like alpha1-antitripsin so in general we have to make n data points that have to be partitions in k cluster. Clustering data may contain much information like etiology, tissues, inheritance etc. Fig [1] tells the working process.

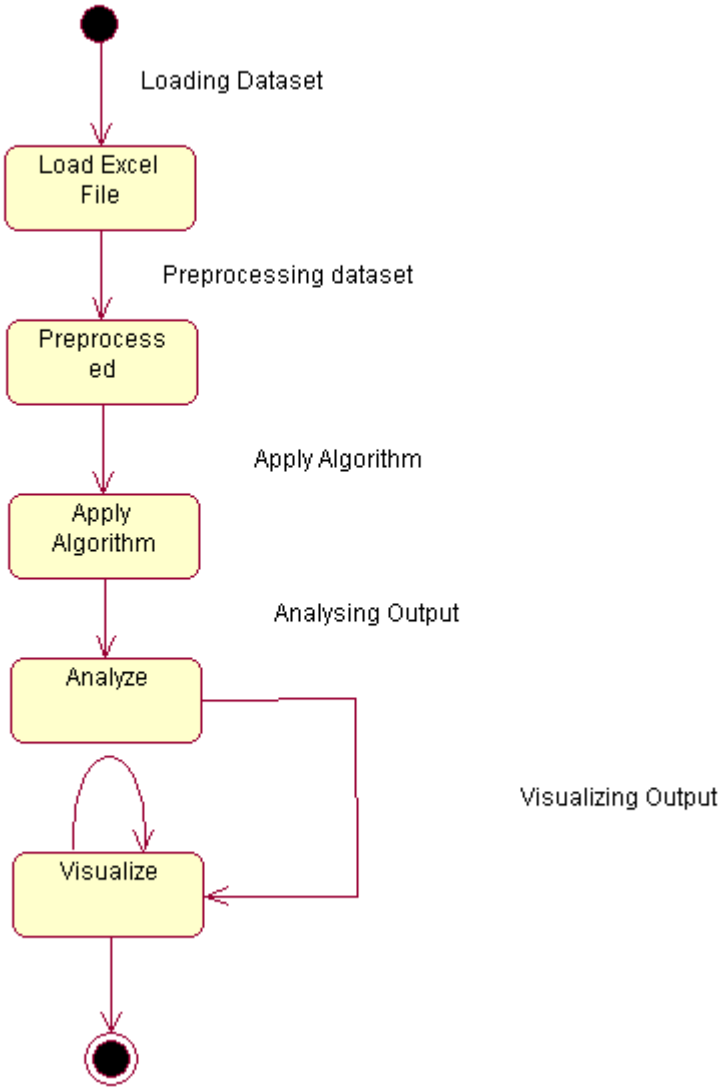

Fig (1) Working of Algorithm

- First we have to upload excel dataset
- Then we have to pre-process the data to reduce noise.
- And apply algorithm to cluster the data
- Analyze the output and visualize the result.

In Section 2, we present some Literature survey on algorithms, Section 3 presents some of the clustering techniques and Section 4 contain existing system and section 5 presents our proposed K-means algorithm. Section 6 reports the experimental analysis and finally result then we presents the conclusion.

II. LITERATURE SURVEY

The present studies show that [5] used k-means algorithm which comes under a new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets resulting in Perform better cluster discovery on sample with informative gene. And [3] suggested a clustering technique (GenClus) for gene expression data which can also handle incremental data resulting in the Improvement of the cluster quality by identifying sub-clusters within big clusters. And [6] presented support vector machines method which approximately can reach more than 74% accuracy in the specific task of predicting whether a single point mutation can be disease related or no. [7] proposed a method to find initial centroids for k-means and we have used similarity measure to find the informative genes. Comes under new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets The result of our clustering approach is to perform better cluster discovery on sample with informative gene. Method: In unsupervised sample based clustering, once informative genes have been identified, then it is relatively easy to use conventional clustering algorithms to cluster samples. The standard k-means can be used for partition. But the accuracy of the clustering results heavily depends on the initial centroid and the dimension of the data. Conclusion: In this paper, they have described the problem of sample clustering on high gene dimension datasets. They have proposed new approach to improve cluster accuracy for gene data. They have achieved higher performance by our proposed method when compared with the existing methods. In future work, they will apply this method to lymphoma and SRBCT datasets.

[1] A regulation-based clustering approach which comes under a pattern matching approach for clustering gene expression data in this project this algorithm is resulted in PatternClus is being free from the initial guess about the number of cluster and non-dependency of input parameters and incremental version handles new gene data incrementally.

[2] a method called suffix tress algorithm for A clustering method for repeat analysis in DNA sequences resulted in rapid identification of all repeat in genome sequence and assignment of these repeats to similarity classes.

[4] a Heuristic search and the Mutual Reinforcing adjustment method which comes under Mining Phenotypes and Informative Genes from Gene Expression Data resulted in Effective and scalable on mining large real-world data sets. The mining results are consistently with good quality.

III. CLUSTRING TECHNOLOGY

The purpose of clustering is to analysis the similarity between the individual data and groups to infer popular structure and assign to that often correspond with self identified group, by reducing noise in the data and extracting useful information from it. Thus, it helps in predicting effective features selection

A. Partitional Clustering

Partitional clustering decomposes a data set into a set of disjoint clusters. Given a data set of *N* points, a partitioning method [13] constructs *K* partitions of the data, which representing a cluster partition. That is, it classifies the data into *K* groups by satisfying the following requirements: (1) each group contains at least one point, and (2) each point belongs to exactly one group. In fuzzy partitioning, a point can belong to more than one group [12].

Many partition clustering algorithms try to minimize an objective function. For example, in *K*-means and *K*-medoids the function (also referred to as the distortion function) is [13]

$$\sum i=1K\sum j=1|Ci|\text{Dist}\ (xj,\text{center}(i)),$$

Where $|\ C_i\ |$ is the number of points in cluster i, Dist (x_j, center (i)) is the distance between point x_j and center i. Many distance functions can be used, such as Euclidean distance and L_1 norm.

IV. PROBLEM FORMULATION

There are many data mining tools available today. There are excellent tools that are available. But they lack in some point. They are,

- Sometimes expert support is required to use these tools.
- The dataset should have a common format. If not, it will ask for the conversion to its own format. The naive user finds difficulty to do this step.
- They provide limited user interaction.
- The process output of these tools is not completely understandable.
- Manually analyzing the genetic variations is very difficult.
- Since many genetic diseases shares the same symptoms it is difficult to identify the genetic disorders accurately.
- To analyze genetic variations in medical labs it requires amino acid analyzer which is very expensive

In our proposed system we are trying to overcome it by using suitable algorithms.

V. RELATED WORK

In this section we are going to explain about how the dataset has been process and the algorithms used in it to get the result.

In this paper we are going to use our own medical genetic dataset which has been created based on consulting the medical advisers. We are mainly concentrating on Lungs genetic disease caused by Alpha-1antitrypsin deficiencies are inherited only in autosomal co-dominant mode. Studies on the records having same inheritance mode which mentioned above are sufficient. Thus, computation is reduced while prediction accuracy is increased via effective feature selection.

Sl no	Patient name	Age	Gender	Etiology	Tissue	Inheritance
1	John	30	M	Inflammatoru, Metabolic	Lungs	Autosomal-Codominant
2	Richard	35	M	Neoplastic	Liver	Autosomal-recessive
3	Alan	25	M	Metabolic	CNS	Autosomal-recessive
4	David	38	M	Degenerative	Muscles	x-chromosomal
5	Chris	48	F	Regulatory	Lungs	Autosomal-Codominant

Etiology:
This tells about cause for the diseases. For a disease, it may contain more than one etiology like inflammatory, metabolic, neoplastic, degeneration, regulatory etc.

Tissues:
This are the cells which are affected by the disease and tissues like liver, lungs, heart muscles, CNS, eyes, Brain, skin, bone, bone marrow, kidney etc are affected areas.

Inheritance:
This attribute tell us how the disease caused genetically, in other words we can say that how a person inherited the disease like autosomal dominant, autosomal recessive, autosomal codominent, X-chromosomal etc.

Pre-processing Data
The Dataset consist of hundreds of records which contains noisy data, missing data values which will affect the quality of the data in data mining and clustering the data. In order to get a good quality data and consequent results we need raw dataset to be pre-processed to improve the clustering process. In this process removal of missing data which are consider as less important will be eliminated, so that we can maintain the accuracy of the result and another main thing is, we have to convert string data (attributes like etiology, tissue) to numeric data and that data we are storing it in grid view so that result data [fig 2] can be transformed to normalization.

SINo	PatientName	Age	Gender	Etiology	Tissue	Inheritance	Etiologykey
1	john	30	Male	inflammato...	Lungs	Autosomal-...	5.25
5	chris	45	Female	neoplastic	Lungs	Autosomal-...	2
9	sirus	56	Male	degenerative	Lungs	Autosomal-...	4
12	juhah	24	Male	metabolic	Lungs	Autosomal-...	3
13	peter	23	Male	inflammato...	Lungs	Autosomal-...	5.3
14	mike	26	Male	neoplastic	Lungs	Autosomal-...	2
21	joy	29	Male	inflammatory	Lungs	Autosomal-...	5
22	jerry	55	Male	Regulatory,...	Lungs	Autosomal-...	1.25

Fig (2) Result after preprocessing dataset

Normalization approach:
In this process the cleaned dataset is transformed into specific range using normalization. Normalization is used to standardize our dataset, so that we can eliminate redundancy or noisy data and make them as reliable data which can improve our result accuracy. Before clustering we need a distance metric. In this paper we are going using Euclidian distance method is used to measure distance. Normalization helps in finding the nearest neighbor classification and clustering. Data normalization technique includes two methods min-max normalization and z-score normalization.

Min- max normalization

$$v' = \frac{v - MinA}{MaxA - MinA}$$

It involves linear transformation of raw data. MinA and MaxA are the minimum and maximum value of the attribute and v is value of attribute A [8].

Z-score normalization

$$v' = \frac{v - \bar{A}}{\sigma A}$$

σA, A are standard deviation and mean of attribute A.

K-means Clustering
The main goal of cluster analysis is to group objects that are similar in one cluster and dissimilar to other. K-means clustering algorithm is one of the most popular clustering methods. The classification of the object is based on predefined number of clusters which is given by user. Random cluster centers are chosen for each cluster. These centers are preferred to be as for as possible from each other. In this algorithm mostly

Euclidean distance is used to find distance between data points and centroids. The Euclidean distance between two multi-dimensional data points X = (x1, x2, x3... xm) and Y = (y1, y2, y3, ….ym) is described as follows

$$D(X, Y) = \sqrt{(x1-y1)2 + (x2-y2)2 + \cdots + (xm-ym)2}$$

Algorithm:

K-Means clustering

1. The algorithm selects k points arbitrarily as the initial cluster centers.
2. Each point is assigned to the closed cluster in the data-set, based upon the Euclidean distance between each cluster center and each point .
3. Each cluster center is recalculated as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters coincide.

VI. EXPERIMENTAL RESULT AND DISCUSSION

This proposed methodology is experimented with more than 100 dataset to predict the genetic diseases related to lungs. The result obtained is used to give treatment for the patients who are all affected. In general manually analyzing the genetic variations is very difficult.

Since many genetic diseases shares the same symptoms it is difficult to identify the genetic disorders accurately, and

To analyze genetic variations in medical labs it requires amino acid analyzer which is very expensive. In this proposed method it can predict how many members are affected by the Alpha-1 antitrypsin deficiency and who are not affected by that disease and what type of genetic treatment they have to take will be displayed as a result.

The below figure (3) is the raw data which contain the X and Y axis values before clustering

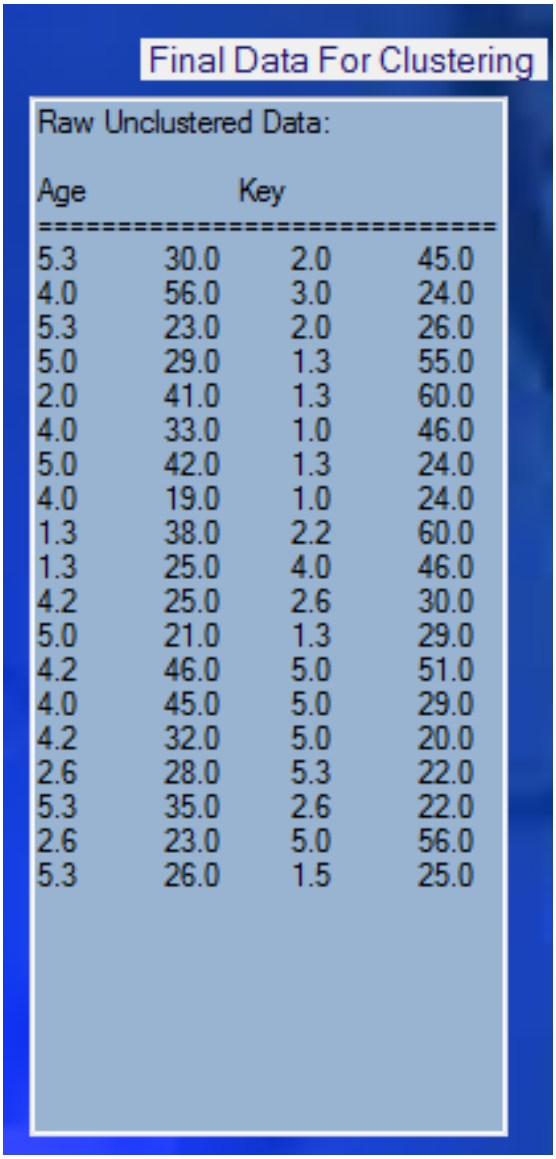

Fig (3) raw unclustered data

The key means the converted string values of attribute etiology. And the age of patient in the dataset.

Result:

The following figure graphically represent the result after applying of K-means algorithm.

The below fig [4] shows the clustering of dataset

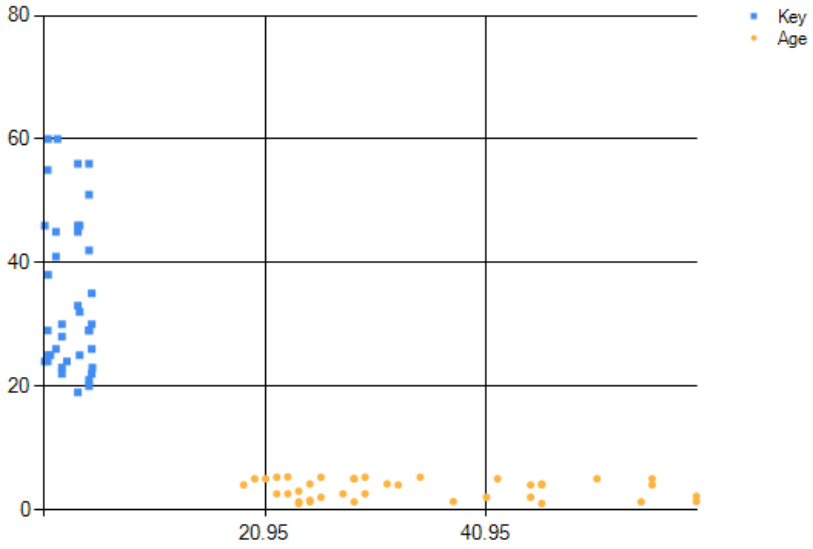

Fig (4) Fast points chart for clustering dataset

The X axis represented Age and the Y axis represented the key value of etiology.

As we can see from the fig(4), there are two clusters and which are having same features are grouper in one cluster.

The fig [5] tells us how many patients are affected by Alpha-1 antitrypsin deficiency.

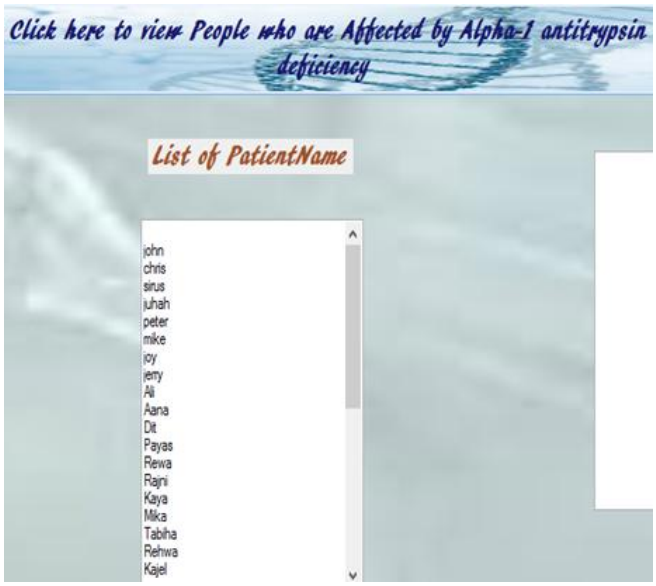

Fig (5) List of patients who are affected by genetic diseases

Conclusion

We have proposed K-mean algorithm for clustering data according to their phenotype similarities. The prediction of genetic disease is by analyzing phenotypic features. We are giving a rough look on amino acid sequence mutation occurred which cause the genetic disorder –Alpha-1antitripsin. And also giving brief information about the treatment, gene therapy a patient has to undergo. In the proposed paper we have tried to implement the K-mean algorithm for a string value by converting it into numeric value. In future work, we will go to enhance the algorithm by implementing least square method

References

[1]. International Journal. Data Mining, Modelling and Management, Vol. 3, No. 2, 2011: A pattern matching approach for clustering gene expression data

[2]. The Institute for Genomic Rsearch, 9712 Medical Center Drive Rockville, MD 20850, USA: A clustering method for repeat analysis in DNA sequences

[3]. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 3, May 2010: An Effective Technique for Clustering Incremental Gene Expression data Sauravjyoti Sarmah1 and Dhruba K. Bhattacharyya

[4]. International Journal of Advanced Science and Technology Vol. 27, February, 2011 85: A new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets

[5]. International Journal of Advanced Science and Technology Vol. 27, February, 201185: A new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets, by Tajunisha N , Saravanan V

[6]. Vol. 22 no. 22 2006, pages 2729–2734 doi:10.1093/bioinformatics/btl4237. Li,W. et al. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. Bioinformatics, 17, 282–283. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information

[7]. International Journal of Advanced Science and Technology Vol. 27, February, 2011 85: A new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets

[8]. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September 2011 ISSN (Online): 1694-0814 www.IJCSI.org 331 Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm

[9]. Jiang, D., Tang, C., and Zhang, A. "Cluster analysis for gene expression data: A survey", Available:www.cse.buffalo.edu/DBGROUP/bioinformatics/papers/survey.pdf, 2003.

[10]. Chung, S., Jun, J. andMcLeod, D. "Mining gene expression datasets using density based clustering", Technical Report, USC/IMSC, University of Southern California, No. IMSC-04-002, 2004

[11]. Goswami, M., Sarmah, R. and Bhattacharyya, D. K. "CNNC: a common nearest neighbour clustering approach for gene expression data", *International Journal of Computational Vision and Robotics* 2011, 2(2), pp. 115 - 126,2011.
[12]. Bezdek, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum, New York, USA, 1981
[13]. http://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_631