Implementation and Evaluation of Diabetes Management System Using Clustering Technique

Snehlata Mandal
Shri Shankaracharya College of Engineering & Technology, Bhila, India, mandal.sneha@gmail.com

Vivek Dubey
SSCET Bhilai CG India, vivekdubey22@gmail.com

Implementation and Evaluation of Diabetes Management System Using Clustering Technique

Snehlata Mandal & Vivek Dubey

Shri Shankaracharya College of Engineering & Technology, Bhila, India
E-mail : mandal.sneha@gmail.com, vivekdubey22@gmail.com

Abstract - Data mining is a field of computer science which is used to discover new patterns for large data sets. Clustering is the task of discovering groups and structures in the data that are in some way or another similar without using known structures of data. Most of this data is temporal in nature. Data mining and business intelligence techniques are often used to discover patterns in such data; however, mining temporal relationships typically is a complex task. The paper proposes a data analysis and visualization technique for representing trends in temporal data using a clustering based approach by using a system that implements the cluster graph construct, which maps data to a two-dimensional directed graph that identifies trends in dominant data types over time.

In this paper, a clustering-based technique is used, to visualize temporal data to identifying trends for controlling diabetes mellitus. Given the complexity of chronic disease prevention, diabetes risk prevention and assessment may be critical area for improving clinical decision support. Information visualization utilizes high processing capabilities of the human visual system to reveal patterns in data that are not so clear in non-visual data analysis.

Key words - *Data Mining, Clustering, temporal data mining, temporal database, blood glucose, and diabetes mellitus.*

I. INTRODUCTION

Diabetes mellitus often simply referred to as diabetes—is a group of metabolic diseases in which a person has high blood sugar, either because the body does not produce enough insulin or because cells do not respond to the insulin that is produced.

This high blood sugar produces the classical symptoms of polyurea (frequent urination), polydypsia(increased thirst) and polyphagia(increased hunger).

There are three main types of diabetes:

Type 1 diabetes: results from the body's failure to produce insulin, and presently requires the person to inject insulin. (Also referred to as *insulin-dependent* diabetes mellitus, *IDDM* for short, and *juvenile* diabetes.)

Type 2 diabetes: results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency.

Gestational diabetes*:* is when pregnant women, who have never had diabetes before, have a high blood glucose level during pregnancy. It may precede development of type 2 DM. [15]

II. NEED

- Diabetes mellitus is a costly chronic disease.

- Much of the burden of preventing, diagnosing and managing diabetes falls on primary care physician who often have insufficient resources to effectively prevent and manage disease.

- At patient level monitoring and responding to changes in risk are important due to rise of pay-for-performance initiatives.

- Given the complexity of chronic disease prevention, diabetes risk prevention and assessment may be critical area for improving clinical decision support.

- Information visualization utilizes high bandwidth processing capabilities of the human visual system to reveal patterns in data that are not so clear in non-visual data analysis.[7]

III. PROBLEM DESCRIPTIONS

- A person suffering from diabetes mellitus cannot be cured.

- That is diabetes mellitus can be controlled but there are no solution to permanently cure the disease.

- Thus a person suffering from diabetes will possess a history of data related to his blood glucose level.

- The earlier age the larger the history of database.

- Thus to control diabetes an individual must always check on its glucose level, to see whether he has controlled glucose level.

- The blood glucose level is the measure of the severity of diabetes in an individual.

- This will help him to take proper medicines, diet and exercise so that he has normal glucose level.

- If required he will have to take insulin doses.

- The blood glucose value is not constant that is it changes with time.

- For a same person in a day we can have different BG values.

- Thus BG values are temporal in nature.

- Thus the changing value of BG level can be analyzed.

- The analysis can be done based on existing database on,

  1. Daily basis

  2. Weekdays

  3. Weekends

  4. Both

  5. Two weeks

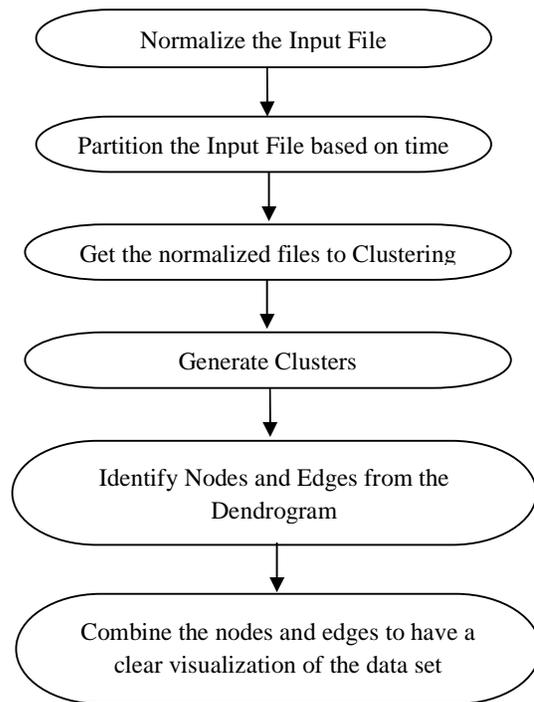 6. One month

  7. All dates

IV. IMPLEMENTATION

4.1 Overview

The Implementation step consists of two main phases: 1) Offline preprocessing of the data and 2) Online interactive analysis and graph rendering. In the preprocessing phase, the data set is partitioned based on time periods, and each partition is clustered using one of many traditional clustering techniques such as a hierarchical approach. The results of the clustering for each partition are used to generate two data structures: the node list and the edge list. Creating these lists in the preprocessing phase allows for more effective (real-time)Visualization updates of the output graphs. Based on these data structures, graph entities (nodes and edges) are generated and rendered as a temporal cluster graph in the system output window.

4.2 Process

Data Clustering - Cluster analysis or clustering is the assignment of a set of observations into subsets (called *clusters*) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.

```
┌─────────────────────────────┐
│  Normalize the Input File   │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│ Partition the Input File    │
│ based on time               │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│ Get the normalized files to │
│ Clustering                  │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│     Generate Clusters       │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│ Identify Nodes and Edges    │
│ from the Dendrogram         │
└─────────────────────────────┘
              │
              ▼
┌─────────────────────────────┐
│ Combine the nodes and edges │
│ to have a clear             │
│ visualization of the data   │
│ set                         │
└─────────────────────────────┘
```

4.3 Clustering

Hierarchical clustering creates a **hierarchy** of clusters which may be represented in a tree structure called a **dendrogram**. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations. Algorithms for hierarchical clustering are generally either agglomerative, in which one starts at the leaves and successively merges clusters together; or divisive, in which one starts at the root and recursively splits the clusters. Any valid metric may be used as a measure of similarity between pairs of observations. The choice of which clusters to merge or split is determined by a linkage criterion, which is a function of the pair wise distances between observations.

4.4 Algorithm for Clustering

Given a set of N items to be clustered, and an NxN distance (or similarity) matrix, the basic process hierarchical clustering is this:

1. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters equal the distances (similarities) between the items they contain.

2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.

3. Compute distances (similarities) between the new cluster and each of the old clusters.

4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N.

V. RESULTS

- Pattern for a month of data

For a given month the analysis shows that low BG (values<=70) rarely during morning, high BG (values>220) for some days and for most of the days BG (70<value<=220) is normal.
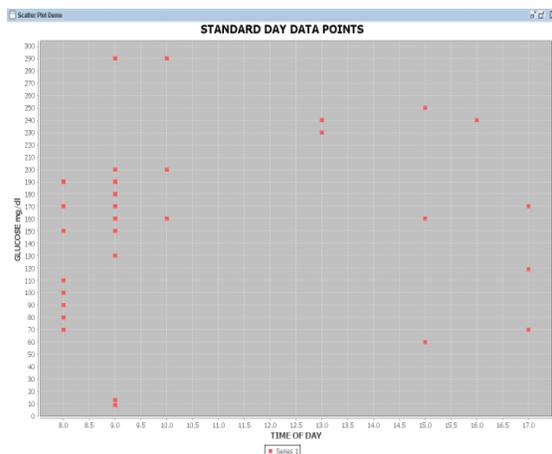
A patient BG value at different time interval is saved in the database. Weekly analysis shows that for the given week BG value is below normal (<70).
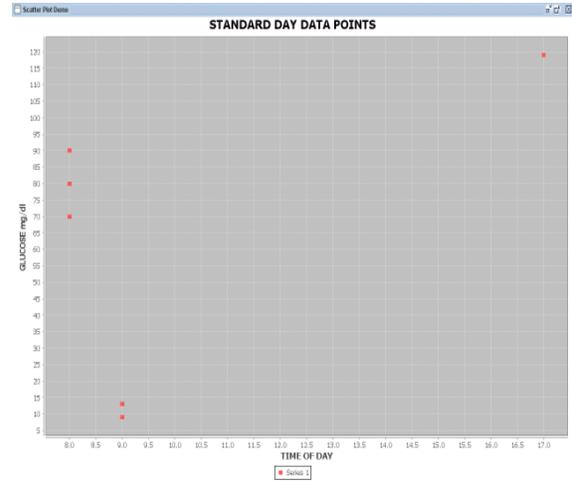
VI. CONCLUSIONS

The paper introduces cluster analysis and hierarchical algorithm which is one of the clustering methods. Hierarchical clustering algorithm is implemented on diabetes management to find tends in patient history to find out useful patterns.

6.1 Advantages

- It can produce an ordering of the objects, which may be informative for data display.

- Smaller clusters are generated, which may be helpful for discovery.

6.2 Disadvantages

- No provision can be made for a relocation of objects that may have been 'incorrectly' grouped at an early stage. The result should be examined closely to ensure it makes sense.

- Use of different distance metrics for measuring distances between clusters may generate different results. Performing multiple experiments and comparing the results is recommended to support the veracity of the original results.

REFERENCES

[1] Gediminas Adomavicius and Jesse Bockstedt, "C-TREND: Temporal cluster graph for identifying and visualizing trends in multi-attribute transactional data", IEEE Transaction on knowledge and data engineering, Vol.20, No.6. June 2008.

[2] J. Roddick and M. spiliopoulou, "A survey of temporal knowledge discovery paradigms and methods", IEEE trans knowledge and data engg., vol. 14, no. 4, pp. 750-767, July/Aug 2002.

[3] Margaret H. Dunham, "Data-mining Introductory and Advanced Topic", vol. 6, pp. 245-275, 2009.

[4] Gajendra Sharma, "Data-mining Data warehousing and OLAP", vol. 2, pp. 337-349, 2010.

[5] Paul R. Cohen and Carole R. Beal, "Temporal data mining for educational applications", Int J Software Informatics, vol. 3, no. 1, pp.31-46, March 2009.

[6] J. Abello and J. Korn, "MGV: A System of Visualizing Massive Multi-Digraphs," IEEE Trans. Visualization and Computer Graphics,vol. 8, no. 1, pp. 21-38, Jan.-Mar. 2001.

[7] Christopher A. Harle, Daniel B. Neil and Rema Padman, "An information visualization approach to classification and assessment of Diabetes risk in Primary Care," Proceedings on the 3rd informs workshop on Data Mining and Health informatics, 2008.

[8] Akash Rajak and Kanak Saxena, "Modeling Clinical database using time series based Temporal Mining," International Journal of computer Theory and Engineering, Vol. 2,No.2, April, 2010.

[9] C.M. Antunes and A.L. Oliveira, "Temporal Data Mining: An Overview," Proc. ACM SIGKDD Workshop Data Mining, pp. 1-13, Aug. 2001.

[10] C. Apte, B. Liu, E. Pednault, and P. Smyth, "Business Applications of Data Mining," Comm. ACM, vol. 45, no. 8, pp. 49-53, 2002.

[11] G.C. Battista, P. Eades, R. Tamassia, and I.G. Tollis, Graph Drawing. Prentice Hall, 1999.

[12] B. Becker, R. Kohavi, and D. Sommerfield, "Visualizing the Simple Bayesian Classifier," Proc. ACM SIGKDD Workshop Issues on the Integration of Data Mining and Data Visualization, 1997.

[13] B. Bederson, "Pad++: Advances in Multiscale Interfaces," Proc. Conf. Human Factors in Computing Systems (CHI '94), p. 315, 1994.

[14] D.J. Berndt and J. Clifford, "Finding Patterns in Time Series: A Dynamic Programming Approach," Advances in Knowledge Discoveryand Data Mining, pp. 229-248, 1995.

[15] http://en.wikipedia.org/ wiki/Diabetes_mellitus

◈ ◈ ◈