Threshold Based Clustering Algorithm Analyzes Diabetic Mellitus

Chapter *in* Advances in Intelligent Systems and Computing · March 2017

DOI: 10.1007/978-981-10-3156-4_3

Preeti Mulay
51 PUBLICATIONS **121** CITATIONS
SEE PROFILE

Rahul Joshi
Symbiosis Institute of Technology
33 PUBLICATIONS **68** CITATIONS
SEE PROFILE

Dipankar Ghosh
Symbiosis Institute of Technology
1 PUBLICATION **1** CITATION
SEE PROFILE

Threshold Based Clustering Algorithm analyzes Diabetic Mellitus

Dr. Preeti Mulay [1,1], Rahul Raghvendra Joshi[1], Aditya Kumar Anguria[1], Alisha Gonsalves, Dakshayaa Deepankar[1], Dipankar Ghosh[1],

[1] Department of CS and IT, Symbiosis Institute of Technology (SIT), affiliated to Symbiosis International University (SIU), Pune, India
{preeti.mulay, rahulj, aditya.anguria, alisha.gonsalves,
dakshayaa.deepankar, dipankar.ghosh}@sitpune.edu.in

Abstract. Diabetes Mellitus is caused due to disorders of metabolism and its one of the most common diseases in the world today, and growing. Threshold Based Clustering Algorithm (TBCA) is applied to medical data received from practitioners and presented in this paper. Medical data consist of various attributes. TBCA is formulated to effectually compute impactful attributes related to Mellitus, for further decisions. TBCAs primary focus is on computation of Threshold values, to enhance accuracy of clustering results.

Keywords:

Incremental Clustering, Knowledge Augmentation, Closeness Factor Based Algorithm (CFBA), Threshold based Clustering, Diabetes Mellitus, Data Mining and TBCA

1 Introduction

Diabetes is emerged as a major healthcare problem in India and every year it is affecting large number of people. The data science based Knowledge Management System (KMS) in health care industry is getting attention to draw effective recommendations to cure the patient in its early stages [1,10]. The knowledge augmented through KMS is an asset for society and incremental learning triggers knowledge augmentation [2, 3]. Online interactive data mining tools are available for incremental learning [4]. The threshold acts as a key in incremental learning to investigative formed closeness factors [5]. This approach in a way may change pattern of diabetes diagnosis [5, 6, 7, 8 and 9]. In this study proposed TBCA is applied on the values of attributes that are collected from patient's medical reports. TBCA implementation unleashes hidden relationships among attributes to extract impactful and non impactful attributes for diabetes mellitus.

[1] Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

In section 2, TBCA is presented. In the following sections i.e., in section 3 the methodology used for its implementation, in section 4 analysis of obtained results, in section 5 concluding remarks and at the last section, references used to carry out this study are listed.

2. TBCA

This section presents a high level pseudo code for TBCA in two parts to show TBCA is an extended version of Closeness Factor Based Algorithm (CFBA).

Input: Data series (DS), Instance (I)
Output: Clusters (K)

CFBA

```
1. Initial cluster count K = 0.
2. Calculate closeness factor (CF) for series DS(i).
3. Calculate CF for next series DS(i+1).
4. Based on CF cluster formation takes place for considered data series (DS).
5. If not (processed_Flag) then
   CF(newly added cluster) = x_i
   ins_counter(newly added cluster) = 1
   Clusters_CFBA ← Clusters ∪ newly added cluster
```

TBCA

```
6.  for all x_i ∈ I
7.  As processed_Flag = False
8.  For all clusters ∈ clusters do
9.  if || x_i - center(cluster)|| < threshold then
10. Update center(cluster)
11. ins_counter(cluster)
12. As processed_Flag = True
13. Exit loop
14. end if
15. end for
```

3. Methodology used to implement TBCA

TBCA data set considers medical reports of working adult diabetic patients having age group between 35 to 45 years for the year 2015-2016. TBCA works in three different phases as mentioned below:

1) In pre-processing input is taken as a CSV file and closeness factor value is calculated by taking into account different possibilities like sum wise, series wise, total weight and error factor for each data series set. The computed values are exported as a CSV file.

2) In clustering, clusters are formed based on closeness values that are generated through preprocessing for a particular data series and formed clusters are stored in a new CSV file in an incremental fashion.

3) Post clustering phase is used to extract values of attributes from the formed clusters for further analysis. The attributes related to diabetes mellitus are extracted on the basis of threshold where lower limit is mean of a cluster and upper limit is its higher

value. These eight attributes are mentioned in table 1 where first four are impactful and remaining are non impactful. The following figures represent processing done on 5K data sets during phases of TBCA in a single and in multiple iterations.

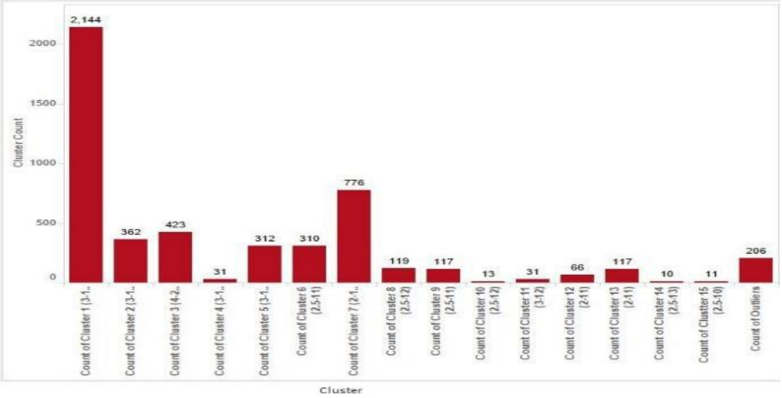


Figure 1: Processing of 5k data series in single iteration of TBCA

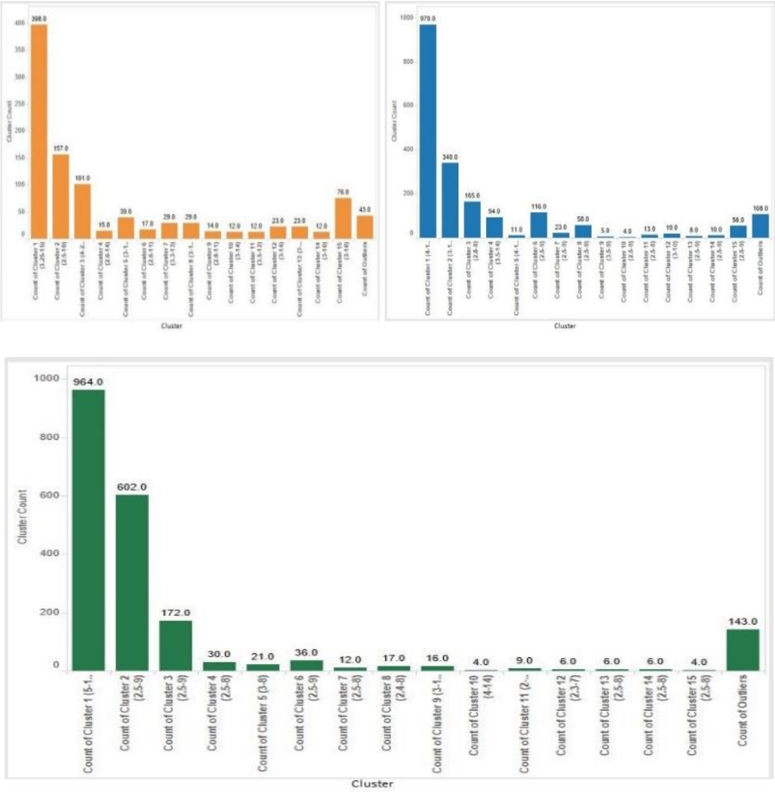


Figure 2: Processing of 5K data series in multiple iterations of TBCA

Sr. No.	Name of attribute	Range of attributes in mg/dl
1	BLOOD GLUCOSE FASTING	115-210
2	BLOOD GLUCOSE PP	140-250
3	CHOLESTROL	140-250
4	TRIGLYCERIDES	140-300
5	HDL CHOLESTROL	40-60
6	VLDL	20-60
7	LDL CHOLESTROL	60-115
8	NON HDL CHOLESTROL	120-170

Table 1: Impactful and non impactful attributes for diabetes mellitus

4. TBCA's analysis

TBCA aims to find out impactful and non impactful attributes and for the same following types of analysis are carried out.

1) Related attributes analysis: The mean value of each attribute of every cluster is taken into account to analyze related attributes in a single and multiple iterations on data sets as shown in figure 1 and 2. The graphs for some of the related attribute analysis are shown below and they depict their behaviour pattern graphically.

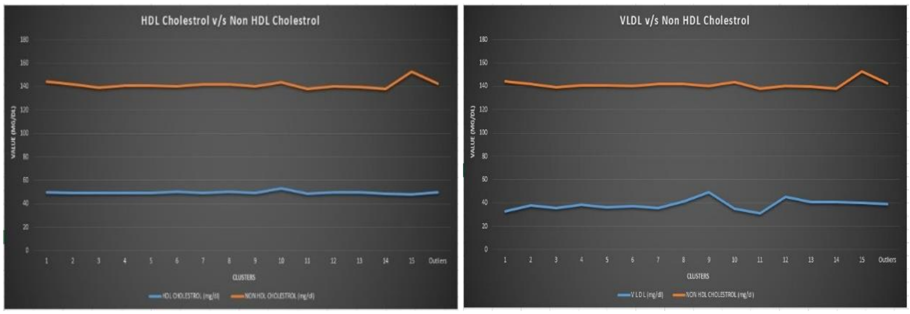
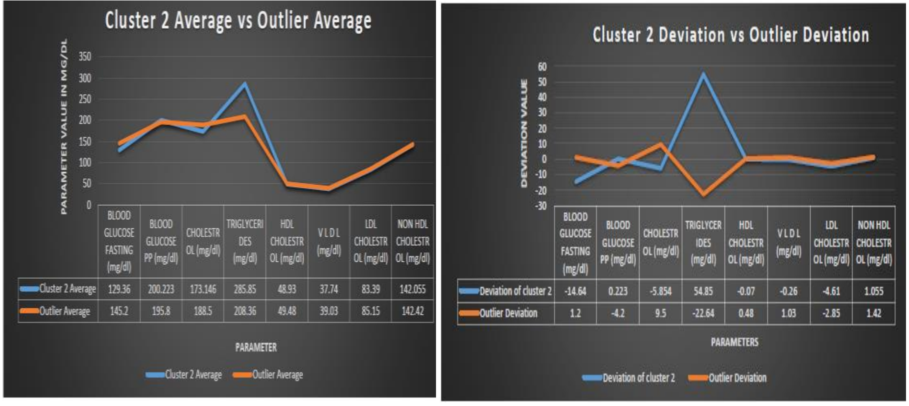


Figure 3: HDL v/s Non HDL Cholesterol, VLDL v/s Non HDL Cholesterol analysis

2) Outlier analysis to extract impactful attributes: The outlier deviation analysis of datasets with extracted eight attributes is carried out which results in depiction of the deviation of the outlier values from the cluster deviation values. The generated pattern

in shown in outlier analysis and it is observed that outlier detection in clustering plays a vital role. The patterns depicted via the statistical graph in Cluster 2 deviation versus outlier deviation for diabetes datasets in figure 4. In figure 4, after analysis of deviation of each cluster against the outlier deviation, it is observed that attributes BLOOD GLUCOSE FASTING, BLOOD GLUCOSE PP, CHOLESTROL and TRIGLYCERIDES are the main factors that are responsible for the generation of the outliers as deviation of the other cluster attributes are overlapping with the outlier deviation. This pattern is cross verified through cluster 2 averages versus outlier average graph shown in another part of figure 4.



Cluster 2 Average vs Outlier Average

	BLOOD GLUCOSE FASTING (mg/dl)	BLOOD GLUCOSE PP (mg/dl)	CHOLESTROL (mg/dl)	TRIGLYCERIDES (mg/dl)	HDL CHOLESTROL (mg/dl)	V L D L (mg/dl)	LDL CHOLESTROL (mg/dl)	NON HDL CHOLESTROL (mg/dl)
Cluster 2 Average	129.36	200.223	173.146	285.85	48.93	37.74	83.39	142.055
Outlier Average	145.2	195.8	188.5	208.36	49.48	39.03	85.15	142.42

Cluster 2 Deviation vs Outlier Deviation

	BLOOD GLUCOSE FASTING (mg/dl)	BLOOD GLUCOSE PP (mg/dl)	CHOLESTROL (mg/dl)	TRIGLYCERIDES (mg/dl)	HDL CHOLESTROL (mg/dl)	V L D L (mg/dl)	LDL CHOLESTROL (mg/dl)	NON HDL CHOLESTROL (mg/dl)
Deviation of cluster 2	-14.64	0.223	-5.854	54.85	-0.07	-0.26	-4.61	1.055
Outlier Deviation	1.2	-4.2	9.5	-22.64	0.48	1.03	-2.85	1.42

Figure 4: Clusters, outlier average and Clusters deviation, Outlier Deviation analysis

4.1 Accuracy/Purity of TBCA

The following formula is used for calculation of accuracy or purity of TBCA.

$$= (100 - \frac{(\text{Clustering value of multiple iteration} - \text{Clustering value of single iteration})}{\text{Clustering value of multiple iteration}} * 100)$$

Where clustering value = cluster count for cluster that contains maximum clustered data for a particular iteration

The accuracy/purity of TBCA is based on clustering value for single iteration and in multiple iterations on same dataset. As shown in figure 1 and 2, the first cluster has the maximum weightage (42% and 46% of the total data resides there) and hence it contains maximum clustered datasets. Therefore, the cluster count or clustering value of this cluster is used calculate the accuracy or purity of TBCA. This accuracy signifies processing of raw datasets and creation precise clusters in single as well as multiple iterations as shown in figure 1 and figure 2 over the same datasets. The multiple iterations on same dataset work in an incremental fashion and confirm cluster members independent of their order, CFBA parameters.

5. Concluding remarks and outlook

TBCA proved to be very useful in obtaining inter attribute relationship and outlier value knowledge over various iterations in an accurate manner which eventually triggered towards finding of key attributes related to diabetes mellitus. TBCA has showed 91.9% of accuracy over single or in several iterations on data set under consideration. It can be effectively used in healthcare domain for prediction of a particular disease like diabetes mellitus. It involves novel mechanism of formation of clusters based on closeness factor and then by using threshold to extract required attributes leading to crisp prediction of impactful set of attributes among them for diabetes mellitus. If a person is suffering from diabetes mellitus properly keeps track of impactful attributes then he/she can manage to cure at early stages. These extracted impactful attributes can act as a catalyst for IT industries for those that are working on medical reports of patients in order to suggest life style management recommendations to cure them from certain diseases. These impactful attributes can also bring revolution in diabetic mellitus patient’s treatment in terms of test on a patient for its diagnosis. TBCA algorithm in turn plays a vital role in augmentation of generated knowledge for diabetes mellitus and may also change current way of pathology practices for diagnosis of diabetes mellitus. So, TBCA may prove best in all other disease prediction, being applied across domain, not restricted.

6. References

1. Lakshmi, K. R., & Kumar, S. P. (2013). Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability. *International Journal of Scientific & Engineering Research*, 4(6), 933-940.

2. Mulay, P., & Kulkarni, P. A. (2013). Knowledge augmentation via incremental clustering: new technology for effective knowledge management. *International Journal of Business Information Systems*, *12*(1), 68-87.

3. Kulkarni, P. A., & Mulay, P. (2013). Evolve systems using incremental clustering approach. *Evolving Systems*, *4*(2), 71-85.

4. Borhade, M., & Mulay, P. (2015). Online Interactive Data Mining Tool. *Procedia Computer Science, 50,* 335-340.

5. Mulay, P. (2016). Threshold Computation to Discover Cluster Structure: A New Approach. *International Journal of Electrical and Computer Engineering (IJECE), 6*(1).

6. Singh, R. J., & Singh, W. (2014). Data Mining in Healthcare for Diabetes Mellitus. *International Journal of Science and Research (IJSR),* 3(7), 1993-1998.

7. Gaikwad, S. M., Mulay, P., & Joshi, R. R. (2015). Attribute Visualization and Cluster Mapping With The Help of New Proposed Algorithm and Modified Cluster Formation Algorithm To Recommend An Ice Cream To The Diabetic Patient Based on Sugar Contain In It. *International Journal of Applied Engineering Research*, *10*.

8. Berry, M. W., Lee, J. J., Montana, G., Van Aelst, S., & Zamar, R. H. (2016). Special Issue on Advances in Data Mining and Robust Statistics. *Computational Statistics & Data Analysis*, *93*(C), 388-389.

9. MS.Tejashri, N.Giri, Prof S.R.Todamal. (2014). DATA MINING APPROACH FOR DIAGNOSING TYPE 2 DIABETES. *INTERNATIONAL JOURNAL OF SCIENCE, ENGINEERING AND TECHNOLOGY*, 2(8), 191-194.

10. Dr. S. Vijayarani, Ms. P. Jothi. (2013). Detecting Outliers in Data streams using Clustering Algorithms. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(8), 1749-1759.