Threshold Based Clustering Algorithm Analyzes Diabetic Mellitus

Preeti Mulay, Rahul Raghvendra Joshi, Aditya Kumar Anguria, Alisha Gonsalves, Dakshayaa Deepankar and Dipankar Ghosh

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

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.

P. Mulay (✉) · R.R. Joshi · A.K. Anguria · A. Gonsalves · D. Deepankar · D. Ghosh
Department of CS and IT, Symbiosis Institute of Technology (SIT),
Symbiosis International University (SIU), Pune, India
e-mail: preeti.mulay@sitpune.edu.in

R.R. Joshi
e-mail: rahulj@sitpune.edu.in

A.K. Anguria
e-mail: aditya.anguria@sitpune.edu.in

A. Gonsalves
e-mail: alisha.gonsalves@sitpune.edu.in

D. Deepankar
e-mail: dakshayaa.deepankar@sitpune.edu.in

D. Ghosh
e-mail: dipankar.ghosh@sitpune.edu.in

S.C. Satapathy et al. (eds.), *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, Advances in Intelligent Systems and Computing 516, DOI 10.1007/978-981-10-3156-4_3

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, 2]. The knowledge augmented through KMS is an asset for society and incremental learning triggers knowledge augmentation [3, 4]. Online interactive data mining tools are available for incremental learning [5]. The threshold acts as a key in incremental learning to investigative formed closeness factors [6]. This approach in a way may change pattern of diabetes diagnosis [6–10]. 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.

In Sect. 2, TBCA is presented. In the following sections i.e., in Sect. 3 the methodology used for its implementation, in Sect. 4 analysis of obtained results, in Sect. 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 $x_i \in I$
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 35–45 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 5 K data sets during phases of TBCA in a single and in multiple iterations.

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 Figs. 1 and 2. The graphs for some of the related attribute analysis are shown below and they depict their behaviour pattern graphically (Fig. 3).

Table 1 Impactful and non impactful attributes for diabetes mellitus

Sr. No.	Name of attribute	Range of attributes in mg/dl
1	Blood glucose fasting	115–210
2	Blood glucose PP	140–250
3	Cholesterol	140–250
4	Triglycerides	140–300
5	HDL cholesterol	40–60
6	VLDL	20–60
7	LDL cholesterol	60–115
8	Non HDL cholesterol	120–170

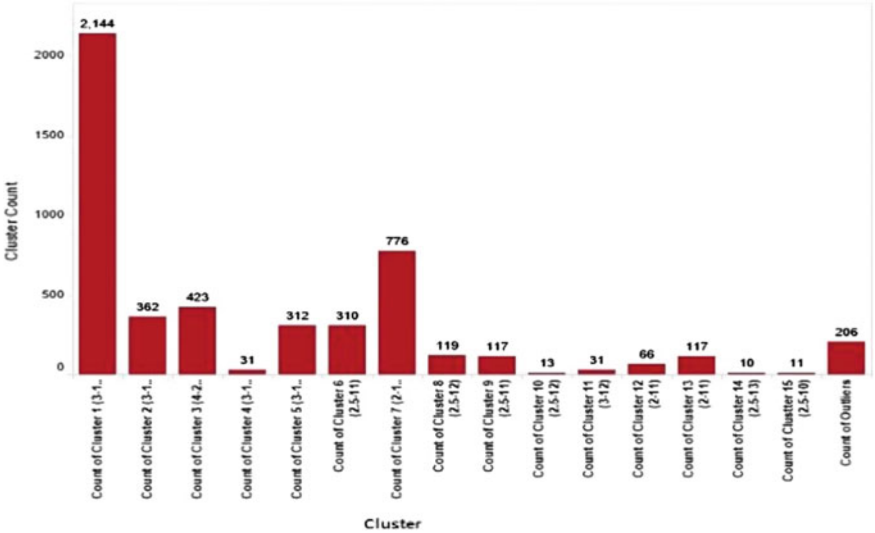

Fig. 1 Processing of 5 K data series in single iteration of TBCA

(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 Fig. 4. In Fig. 4, after analysis of deviation of each cluster against the outlier deviation, it is observed that attributes BLOOD GLUCOSE FASTING, BLOOD GLUCOSE PP, CHOLESTEROL 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 Fig. 4.

4.1 *Accuracy/Purity of TBCA*

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

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

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

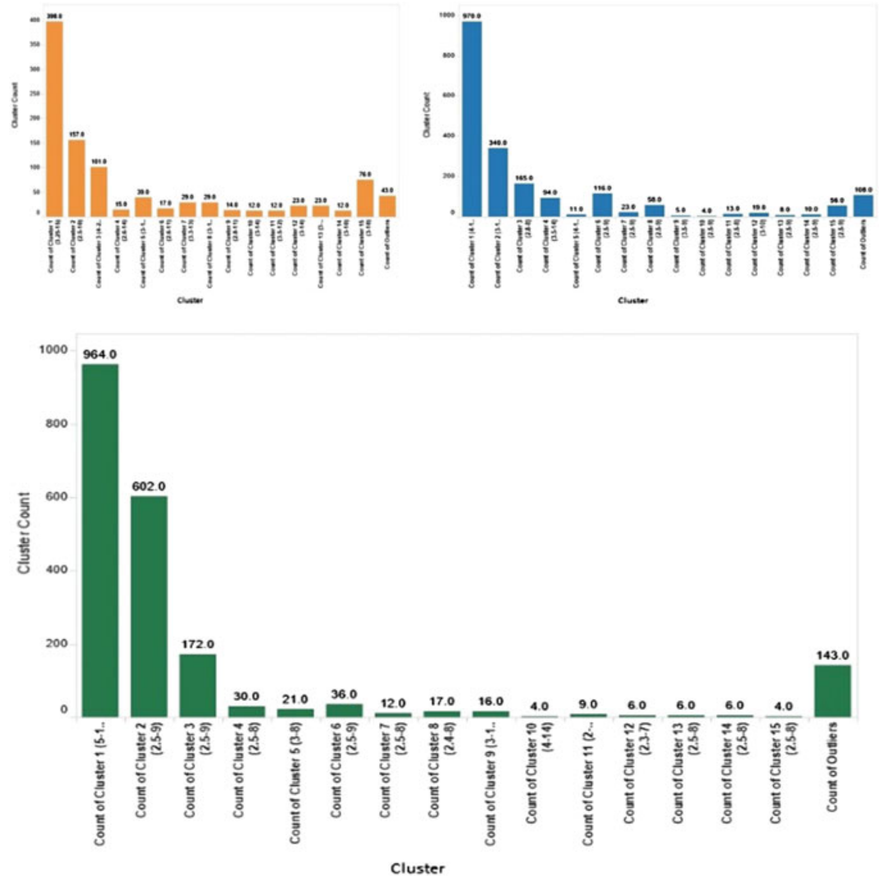

Fig. 2 Processing of 5 K data series in multiple iterations of TBCA

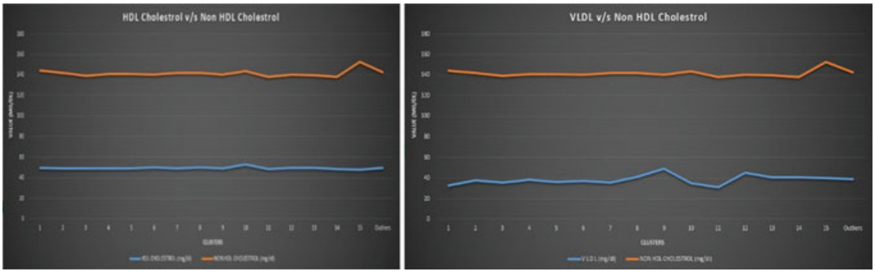

Fig. 3 HDL versus Non HDL Cholesterol, VLDL versus Non HDL Cholesterol analysis

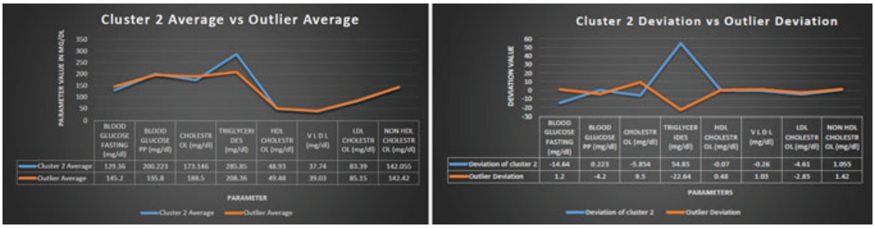

Fig. 4 Clusters, outlier average and clusters deviation, outlier deviation analysis

The accuracy/purity of TBCA is based on clustering value for single iteration and in multiple iterations on same dataset. As shown in Figs. 1 and 2, the first cluster has the maximum weight age (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 Figs. 1 and 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.

References

1. K.R. Lakshmi, S.P. Kumar, Utilization of data mining techniques for prediction of diabetes disease survivability. Int. J. Sci. Eng. Res. **4**(6), 933–940 (2013)
2. D.S. Vijayarani, M.P. Vijayarani, Detecting outliers in data streams using clustering algorithms. Int. J. Innov. Res. Comput. Commun. Eng. **1**(8), 1749–1759 (2013)
3. P. Mulay, P.A. Kulkarni, Knowledge augmentation via incremental clustering: new technology for effective knowledge management. Int. J. Bus. Inf. Syst. **12**(1), 68–87 (2013)
4. P.A. Kulkarni, P. Mulay, Evolve systems using incremental clustering approach. Evol. Syst. **4** (2), 71–85 (2013)
5. M. Borhade, P. Mulay, Online interactive data mining tool. Proc. Comput. Sci. **50**, 335–340 (2015)
6. P. Mulay, Threshold computation to discover cluster structure: a new approach. Int. J. Electr. Comput. Eng. (IJECE), **6**(1) (2016)
7. R.J. Singh, W. Singh, Data mining in healthcare for diabetes mellitus. Int. J. Sci. Res. (IJSR) **3** (7), 1993–1998 (2014)
8. S.M. Gaikwad, P. Mulay, R.R. Joshi, 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. Int. J. Appl. Eng. Res. **10** (2015)
9. M.W. Berry, J.J. Lee, G. Montana, S. Van Aelst, R.H. Zamar, Special issue on advances in data mining and robust statistics. Comput. Stat. Data Anal. **93**(C), 388–389 (2016)
10. M.S. Tejashri, N. Giri, Prof S.R. Todamal, Data mining approach for diagnosing type 2 diabetes. Int. J. Sci. Eng. Technol. **2**(8), 191–194 (2014)