International Journal of Advanced Computer Research
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
Volume-5 Issue-20 September-2015

Clustering Diabetics Data Using M-CFICA

Jerusha Shalini Vaska[1]*and A. M. Sowjanya[2]
M.tech (CST), Department of CS&SE, College of Engineering (A), Andhra University[1]
Assistant Professor, Department of CS&SE, College of Engineering (A), Andhra University[2]

Received: 12-July-2015; Revised: 25-August-2015; Accepted: 06-September-2015

Abstract

E-Health has grown popular due to a wide range of services provided. The role of a patient has also changed in today's health care as they are expected to use ICT services to gain information and knowledge to know about their well-being. In the field of data mining clustering is a widely used technique for discovering patterns in underlying data. Traditional clustering algorithms are normally limited to handling datasets that contain either numeric or categorical attributes. However, datasets with mixed types of attributes are also common in real life data mining applications. In this paper a cluster feature based incremental clustering algorithm, MCIFA (Cluster Feature-Based Incremental Clustering Approach to mixed data) is applied on the diabetes dataset to check its suitability in the medical domain. The achieved clustering accuracy in results section shows that this is indeed suitable for medical domain and can be used for 'e-prescribing'. But it needs to be fine-tuned so as to increase the clustering accuracy as the percentage of allowed error-rate in medical domain should be as small as possible.

Keywords

Data mining, Clustering, Cluster feature, Incremental clustering, mixed data, E-health.

1. Introduction

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data triggering the imminent need for turning such data into useful information and knowledge. The mined information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. Data mining can also be used to improve customer service, better target marketing campaigns, identify high-risk clients and improve production processes. Since most businesses and organizations collect data about their operations this data can then be examined for insights into their operations and transactions their businesses perform. However, manual analysis of all this data becomes difficult and tedious. An automated data mining approach, however, can be used to find patterns which may not even have suspected existed, or that would take too long to find by manual means. Automated data mining allows the development of models of considerable complexity which can take many more factors into account.

Clustering is referred to as unsupervised learning or segmentation clustering partitions which segment the data into groups that might or might not be disjointed. It is usually accomplished by determining the similarity among the data on predefined attributes. In clustering, the most similar data is grouped into clusters. Since the clusters are not predefined, a domain expert is often required to interpret the meaning of created clusters. Incremental clustering can be defined as the integration of a clustering algorithm that functions incrementally whenever new data is added to the original data already present in the database [1].

2. Literature Survey

2.1 Incremental Clustering

Incremental clustering has attracted the attention of the research community with Hartigan's Leader clustering algorithm [2] which uses a threshold to determine if an instance can be placed in an existing cluster or it should form a new cluster by itself. COBWEB [3] is an unsupervised conceptual

*Author for correspondence

clustering algorithm that produces a hierarchy of classes. Its incremental nature allows clustering of new data to be made without having to repeat the clustering already made. It has been successfully used in engineering applications [4]. CLASSIT [5] is an alternative version of COBWEB.

It handles continuous or real valued data and organizes them into a hierarchy of concepts. It assumes that the attribute values of the data records belonging to a cluster are normally distributed. As a result, its application is limited. Another such algorithm was developed by Fazil Can to cluster documents [6]. Charikar et al. defined the incremental clustering problem and proposed an incremental clustering model which preserves all the desirable properties of HAC (hierarchical agglomerative clustering) while providing an extension to the dynamic case.

[7].BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is especially suitable for large number of data items [8]. Incremental DBSCAN was presented by Ester et al., which is suitable for mining in a data warehousing environment where the databases have frequent updates [9]. The GRIN algorithm, [10] is an incremental hierarchical clustering algorithm for numerical data sets based on gravity theory in physics. Serban and Campan have presented an incremental algorithm known as Core Based Incremental Clustering (CBIC), based on the k-means clustering method which is capable of re-partitioning the object set when the attribute set changes [11]. The new demand points that arrive one at a time have been assigned either to an existing cluster newly created one by the algorithm in the incremental versions of Facility Location and k-median to maintain a good solution [12].

2.2 E-Health

E-health has grown its success and popularity from time to time, due to a wide range of services provided. It is known that e-health involves a variety of users such as patients, doctors, nurses, etc. who would access electronic medical data. These various types of users have different assigned tasks and perhaps may not be allowed to access certain data. For example, doctors have right to check and modify the records of patients' diagnosis results, but this information shall not be accessed by patients [13].

The role of patient is changed in today's healthcare. Boset al. defines Health 2.0 as "the combination of health data and health information with (patient) experience through the use of information and communication technology (ICT), enabling the citizen to become an active and responsible partner in his/her own health and care pathway" [14]. While in former days the practitioner was the one who held all the knowledge about the diagnosis and the treatment of a patient, today it is expected that patients use ICT services to gain information and knowledge about what is happening to them [15].

3. Methodology

Datasets with mixed types of attributes are very common in data mining applications like banking sector, health data and web-log data. Such domains maintain dynamically growing datasets described in terms of heterogeneous attributes and hence require incremental clustering algorithms that can handle all types of attributes. Existing incremental clustering algorithms including CFICA are limited to handle either numeric or categorical attributes but not their combinations so an incremental clustering algorithm called M-CFICA (Cluster Feature-Based Incremental Clustering Approach to Mixed Data) [16]. While designing M-CFICA, the Inverse Proximity Estimate (IPE) devised in CFICA [17] was proposed to deal with mixed distances that estimate the distance between two data points represented in terms of both numeric and categorical attributes.

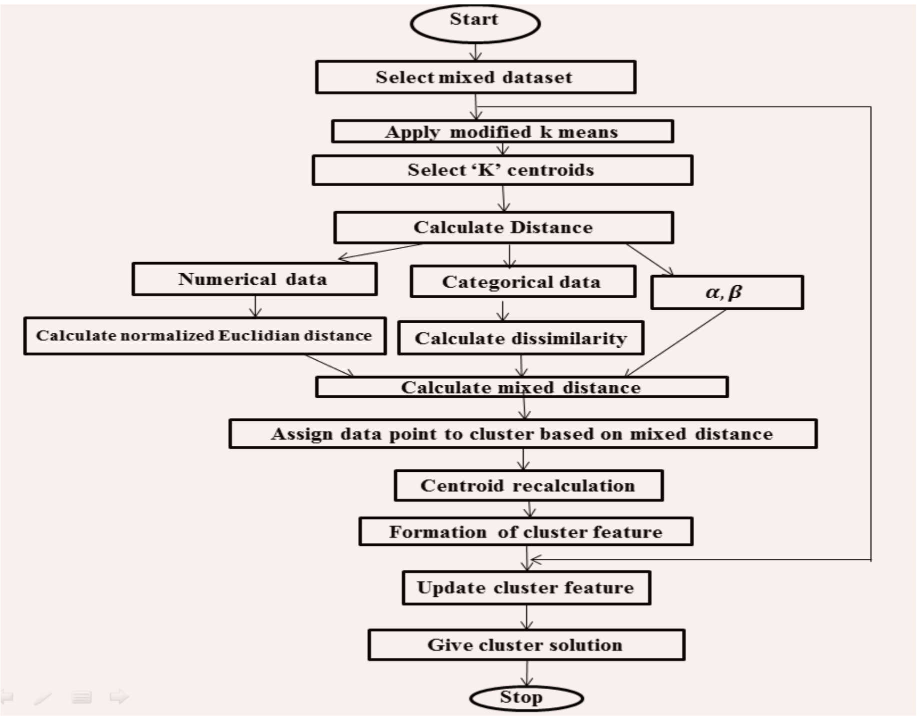

Figure 1: Flowchart

3.1 Distance Estimation for Mixed type of data

An attribute of any type can be transformed into one of the basic data types namely numeric or categorical. Hence data sets with mixed types of attributes can be represented in terms of numeric or categorical attributes either directly or after transformation. The distance between such data points is estimated separately in terms of numeric attributes and categorical attributes to arrive at Normalized Euclidean Distance (NED) and Dissimilarity (DS) respectively. The Dissimilarity (DS) between two data points is estimated based on their categorical attributes. The distinct values of categorical attributes are expressed as attribute-value pairs [18] and the prominence of various attribute-value pairs in a cluster of data points characterizes the cluster.

3.1.1 M-CFICA

M-CFICA defines mixed distance between a pair of data points described in terms of heterogeneous attributes. It is estimated as the weighted sum of Normalized Euclidean Distance and dissimilarity and is used while forming initial clusters for the static database.

$$MD(C_i, \Delta_y) = [\alpha \times NED(C_{i,}\Delta_y)] + [(1-\alpha)DS(C_i, \Delta_y)] \quad (1)$$

Where $MD(C_i, \Delta_y)$ → mixed distance between each cluster centroid C_i and data point Δ_y.

$NED(C_{i,}\Delta_y)$ → Normalized Euclidean distance between the data point Δ_y and cluster centroid C_i
$DS(C_i, \Delta_y)$ →Categorical dissimilarity is computed for the data point Δ_y and cluster centroid C_i using the dissimilarity measure!

α → Weightage calculated based on the number of numerical as well as categorical attributes in the dataset.

3.1.2 Mixed Distance

The proximity between two data points involving both numeric and categorical attributes is estimated by combining the distance estimated in terms of numeric attributes and the dissimilarity estimated in terms of categorical attributes.

1. **Numerical distance:** The commonly used distance measure for computing the distance

between the two data points with respect to its numerical attributes is Euclidean distance. A Normalized Euclidean Distance (NED) that is confined to a range [0, 1] by transforming the original Euclidean distance has been defined

$$NED(\Delta y, Ci) = \frac{ED(\Delta y, Ci)}{\sqrt{n}_n} \qquad (2)$$

2. **Categorical dissimilarity:** Normally, the described in terms of m categorical attributes is estimated as the ratio of the number of mismatches to m. This conventional approach to estimating dissimilarity ignores the specificity of an attribute-value pair giving equal importance to all attribute value pairs. Hence a data point with an essential characteristic represented by an attribute-value pair and another data point having a general characteristic found among the members of all clusters are considered equally close to a cluster. To circumvent this drawback it has been suggested that the categorical nature of a cluster is better represented in terms of specificity and prominence of attribute-value pairs. Hence, Resemblance metric [18] has been adopted while defining the dissimilarity of a data point to a cluster/data point. The dissimilarity (DS) of a data point to either a cluster prototype or another data point is confined to a range of [0, 1] based on the following definition

$$DS(\Delta\, y, Ci) = 1 - R\frac{(\Delta_y C_i)}{n_c} \qquad (3)$$

Where

n_c→ Number of categorical attributes.

$DS\ (\Delta_y, C_i)$ → Dissimilarity between centroid C_i and the incoming data point Δ_y.

$R\ (\Delta_y, C_i)$ → Signifies the resemblance of a point Δ_y to the i_{th} cluster.

$$R(\Delta_y, C_i) = \sum_{AVr\epsilon\Delta\, y} w(Ci;\ AVr) \qquad (4)$$

3. **Weightage:** Represents the relative weightage given to numerical component compared to categorical component for estimating the mixed distance. The weightage is calculated based on the number of numerical as well as categorical attributes in the dataset as in the following

$$\alpha = \frac{n_n}{n_n + n_c} \qquad (5)$$

Where n_n → Total number of numerical attributes in the dataset

n_c →Total number of categorical attributes in the dataset.

3.2 Initial Cluster formation

3.2.1 Initial Clustering of the static database

Initial cluster formation is performed on the static database first. M-CFICA employs a partitional clustering algorithm which makes use of the mixed distance estimate since the k-means algorithm cannot cluster categorical component.

3.2.2 Clustering of incremental database

A set of clusters are initially obtained using the modified k-means algorithm on the static database. The Cluster Feature is so designed to represent numeric as well as categorical nature of the members of a cluster in a nutshell, while retaining all aspects that are essential for its incremental update. The clustering solution is represented in the form of Cluster Features.

3.3 Computation of Cluster Feature (CF)

The original structure of cluster feature used in CFICA has been slightly modified to accommodate categorical data also as CFICA caters to numerical data only.

In M-CFICA the Cluster Feature is denoted as,

CF_i=$\{n_{i,}\overrightarrow{m_\imath}\ ,\ \overrightarrow{w_\imath}\ ,\ Q_{i,}\overrightarrow{ss_\imath},\overrightarrow{m_{\imath,}}\overrightarrow{AV_{\imath r}}\}$ (6)

Where CF_i → Cluster feature for cluster i (C_i)

n_i → Number of data points in cluster

$\overrightarrow{m_\imath}$ → Mean vector of numerical attributes in the cluster C_i with respect to which farthest points are identified.

$\overrightarrow{w_\imath}$ → Weights of characterizing attribute-value pairs of cluster C_i

Q_i → List of p-farthest points of cluster C_i

$\overrightarrow{ss_\imath},$ → Squared sum vector that changes

During incremental updates

$\overrightarrow{m_\imath}$ → New mean vector of the cluster C_i including newly added points to the cluster C_i

$\overrightarrow{AV_{\imath r}}$ → Updated count of all attribute value pairs occurred in i_{th} cluster.

3.4. Proximity Estimation for a new data point

The Inverse Proximity Estimate (IPE) considers the proximity of a data point to a cluster centroid as well as its proximity to a farthest point of the cluster in its vicinity to determine the membership of a data point in a cluster [19] which is formally stated below

$$IPE_{\Delta_y}^{(I)} = MD(C_i, \Delta_y) + [MD(q_i, \Delta_y) * MD(C_i, q_i)]$$
(7)

Where $IPE_{\Delta_y}^{(I)} \rightarrow$ Inverse Proximity Estimate of incoming data point Δ_y to the ith cluster C_i
$MD(C_i, \Delta_y) \rightarrow$ Mixed distance from the data point Δy to the centroid of cluster C_i

$$\Rightarrow MD(C_i, \Delta_y) = [\alpha \times NED(C_i, \Delta_y)] + [(1-\alpha) DS(C_i, \Delta_y)] \quad (8)$$

$MD(C_i, \Delta_y) \longrightarrow$ Mixed distance from farthest point qi to the data point Δ_y

$$\Rightarrow MD(q_i, \Delta_y) = [\alpha \times NED(q_i, \Delta_y)] + [(1-\alpha) DS(q_i, \Delta_y)] \quad (9)$$

$MD(q_i, \Delta_y) \longrightarrow$ Mixed distance from cluster C_i to the farthest point q_i

$$\Rightarrow MD(C_i, q_i) = [\alpha \times NED(C_i, q_i)] + [(1-\alpha) DS(C_i, q_i)] \quad (10)$$

3.5. Insertion of a new data point

After initial clustering of the static database, the clustering solution is converted into the form of Cluster Features (CFs). When there is an incoming data point y to be inserted into one of the existing k clusters decisions regarding the inclusion of new data points involve estimation of $IPE_{\Delta_y}^{(I)}$ for each cluster i and concept drift associated with the cluster. The estimation of inverse proximity estimate only considers $(\overrightarrow{m_i}, \overrightarrow{w_i})$ referred to as cluster prototype. If the incoming data point y cannot be included into any of the existing k clusters then a new singleton cluster will be formed with this data point Δ_y. In such a case the number of clusters is increased by one and the Cluster Feature is constructed for the newly added cluster.
The above algorithm is implemented on the diabetes dataset incrementally and the clustering accuracy is noted.

4. Results

M-CFICA was implemented on diabetes dataset incrementally and the clustering accuracy is noted. Diabetes dataset consist of four fields per record namely each field is separated by a tab and each record is separated by a newline.

The Code field is deciphered as follows:
(1) Date in MM-DD-YYYY format
(2) Time in XX:YY format
(3) Code
(4) Value

The dataset has been processed incrementally by dividing the 943 instances in the dataset into 5 chunks as follows:
DD1 – Consists of 300 instances
DD2– Consists of 200 instances
DD3– Consists of 200 instances
DD4– Consists of 150 instances
DD5– Consists of 97 instances

M-CFICA uses mixed distance to form initial clusters for the static database. Then cluster features are computed for those clusters. In an incremental way, the next chunk of data points i.e DD2 consisting of 200 instances are given as input. Now the inverse Proximity estimate (IPE) is computed for each data point in the second chunk. If the calculated distance is less than the preferred threshold value, λ, data points are assigned to the appropriate cluster else new cluster is formed. Then cluster feature is undated for each data point. Once all the data points in the chunk are processed, merging is done if only the mixed distance between the centroids of the pair of clusters is smaller than user defined threshold (θ). Here, λ=0.6 and θ=0.3. The set of resultant clusters are finally obtained after merging. The cluster solution is similarly updated incrementally upon receiving the remaining chunks until the whole dataset is over.

Since the dataset has class labels for each data item, the purity measure described in [20, 21] has been used for evaluating the performance of M-CFICA as given below,

$$\text{Purity} = \frac{1}{N} \sum_{i=1}^{T} X_i \qquad (11)$$

Where N is number of data points in the dataset
T is number of resultant clusters
X_i Is Number of data points of majority class in cluster i.

The performance of M-CFICA on datasets with mixed type of attributes is evaluated and compared with a well-known algorithm which handles mixed type of datasets namely, K-PROTOTYPES.

The purity of M-CFICA verses K-PROTOTYPES is tabulated in the table below. And graphically shown in figure: 2.

Table 1: Comparison Chart

S. No	**No. of Clusters (k)**	**M-CIFA**	**K prototypes**
1	10	0.928	0.532
2	20	0.847	0.873
3	30	0.762	0.755
4	40	0.743	0.732
5	50	0.648	0.652
6	60	0.535	0.582
7	70	0.648	0.732
8	80	0.793	0.823

From the above table it can be seen that M-CFICA performs almost similar to K-PROTOTYPES and that M-CFICA is suitable to handle medical data with mixed types of attributes like diabetes data.

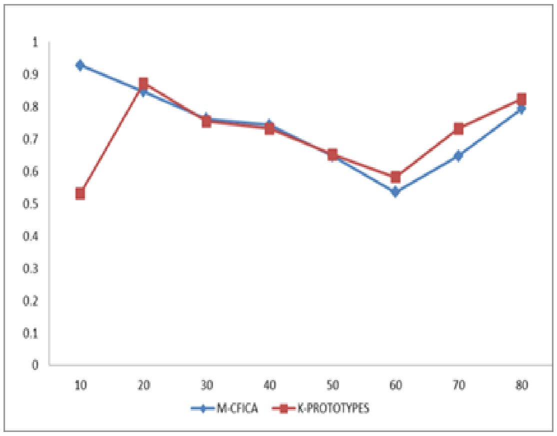

Figure 2: Purity vs. number of clusters (K) for diabetes dataset

5. Conclusion

The applicability of cluster-feature based incremental clustering algorithm for mixed data (M-CFICA) has been tested on the diabetes dataset to check this suitability of this algorithm in medical domain. The more or less similar results and accuracy convey that this algorithm used in medical domain for diagnosis and treatment of a patient. But considering it in medical domain means error rate should be as low as possible since it has the ability to affect human lives. So this algorithm though suits the purpose, it needs to be fine-tuned in the future work to increase the clustering accuracy further.

References

[1] Sowjanya A.M. and Shashi M, “New Proximity Estimate for Incremental Update of Non-Uniformly Distributed Clusters," in International Journal of Data Mining Knowledge Management Process, vol. 3,no. 5 pp. 91-109, 2013.

[2] Hartigan, J.A. Clustering Algorithms. John Wiley and Sons, Inc., New York, NY,1975.

[3] Fisher D., “Knowledge acquisition via incremental conceptual clustering,” Machine Learning, vol. 2, 1987, pp.139-172.

[4] Fisher, D., Xu, L., Carnes, R., Rich, Y., Fenves, S.J., Chen, J., Shiavi, R., Biswas, G., and Weinberg, J. Applying AI clustering to engineering tasks. IEEE Expert 8, 51–60, 1993.

[5] J. Gennary, P. Langley, and D. Fisher, “Models of Incremental Concept Formation,” Artificial Intelligence Journal, vol. 40, 1989, pp. 11-61.

[6] Fazil Can, “Incremental Clustering for Dynamic Information Processing”, ACM Transactions on Information Systems, April 1993, Vol. 11, No. 2, pp. 143-164.

[7] M. Charikar, C. Chekuri, T. Feder, and R. Motwani, “Incremental clustering and dynamic information retrieval,” 29th Symposium on Theory of Computing, 1997, pp. 626—635.

[8] Tian Zhang, Raghu Ramakrishnan, Miron Livny, “BIRCH: an efficient data clustering method for very large databases”, Proceedings of the ACM SIGMOD International. Conference on Management of Data, pp 103-114, 1996.

[9] Ester, M., Kriegel, H., Sander, J., Xu X., Wimmer, M., “Incremental Clustering for Mining in a Data Warehousing Environment”, Proceedings of the 24th International. Conference on Very Large Databases (VLDB’98), New York, USA, 1998, pp. 323-333.

[10] Chien-Yu Chen, Shien-Ching Hwang, and Yen-Jen Oyang, "An Incremental Hierarchical Data Clustering Algorithm Based on Gravity Theory", Proceedings of the 6th Pacific- Asia Conference on Advances in Knowledge Discovery and Data Mining, 2002, pp. 237 – 250.

[11] Serban G., Campan A., "Incremental Clustering Using a Core-Based Approach", Lecture Notes in Computer Science, Springer Berlin, Vol: 3733, pp. 854-863, 2005.

[12] Fotakis D., "Incremental algorithms for Facility Location and k-Median", Theoretical Computer Science, Vol: 361, No: 2-3, pp: 275-313, 2006.

[13] Boonyarattaphan, A, Yan Bai, Sam Chung, "A security framework for e-Health service authentication and e-Health data transmission" Institute of Technology, University of Washington, Tacoma 2009.

[14] L. Bos, A. Marsh, D. Carroll, S. Gupta, M. Rees, "Patient 2.0 Empowerment," Proceedings of the 2008 International Conference on Semantic Web & Web Services SWWS08, (Hamid R. Arabnia, Andy Marsh (eds)), 2008, pp.164-167.

[15] S. Nasiri, M. Dornhöfer, M. Fathi, "Improving EHR and Patient Empowerment based on Dynamic Knowledge Assets," informatik 2013: 43.Jahrestagung der Gesellschaft für Informatik, September 16-20, Koblenz, Germany, Lecture Note in Informatics(LNI) – Proceedings, Geselschaft für Informatik(GI), Köllen Druk+Verlag, Matthias Horbach (Ed.), pp. 402-413, (2013).

[16] A.M.Sowjanya, M.Shashi, "A New Distance Metric for Formation of Non-Uniformly Distributed Incremental Clusters", Research Notes in Information Science (RNIS) Volume13, May 2013.

[17] A.M. Sowjanya, M. Shashi, "Cluster Feature-based Clustering Approach (CFICA) for numerical data", International Journal of Computer Science and Network Security, 2010, Vol.10, No.9, pp.73-79.

[18] Chen H.L, Chen M.S and Chen L.S , \Catching the Trend: A Framework for Clustering Concept-Drifting Categorical Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 5, pp. 652{665, 2009.

[19] Tian Zhang , Raghu Ramakrishnan and Miron Livny, "BIRCH : A new data clustering algorithm and its applications," in Proceedings of International conference on Data Mining and Knowledge Discovery, vol. 1, no. 2, pp. 141-182, 1997.

[20] Huang Z, "Extensions to the k-Means Algorithm for Clustering Large Data Sets With Categorical Values," in Data Mining and Knowledge Discovery, vol. 2, pp. 283-304, 1998.

[21] Xiaoke Su, Yang Lan, Renxia Wan, and Yuming Qin, "A Fast Incremental Clustering Algorithm," in Proceedings of the 2009 International Symposium on Information Processing (ISIP09),pp. 175-178, 2009.

Jerusha Shalini Vaska has done her B.tech, M.tech in Computer Science. This paper is a part of the work submitted for M.tech project. Currently teaching B.tech students.
Email: shalini.vaska@gmail.com

Dr. A. Mary Sowjanya has completed her B.tech, M.tech in Computer Science and Ph.D. in Data Mining currently teaching undergraduate & postgraduate courses and pursuing research activities. Has membership in IEEE and CSI.