Volume 3, Issue 7, July 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Study on Performing Clusters in Transactional Data with Different Sizes and Shapes

Nithya. P*, T. Menaka
Department of Computer Science, NGM College
India

Abstract*— Data mining uses various clustering algorithm for grouping similar objects. One of the most popular algorithm for clustering is density based clustering algorithm, which clusters are of widely differing sizes, densities and shapes when the data contains large amounts of noise and outliers. Many of these issues become even more important when the data is of very high dimensionality, such as text, time series and sequence data. In this paper we present a novel clustering technique, which can solve mentioned issues significantly. We show that our algorithm is intuitive, easy to state and analyse than traditional methods on data set: transactional data. As a result, the algorithm can effectively find the behaviours in transactional data i.e. in banking system. However, we discuss a number of optimizations that allow the algorithm to handle large number of data sets efficiently. This paper obtained loan data sets to cluster the related data according to their behaviour.

Keywords*— Data mining, cluster analysis, novel clustering, shared nearest neighbour, density based spatial clustering.

I. INTRODUCTION

Cluster analysis [7] partition the set of data into groups based on data similarity. For example cluster analysis is used to group the related documents in file, to find the genes that have similar functionality. In business, clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. In clustering, there are five different techniques i.e. Partitioning, Density based, Hierarchal, Grid based, Model based [5]. However most of the clustering challenges, particularly those related to "quality", rather than computational resources for this reason clustering research have also been focused on these issues [9]. Crucial density based algorithm can find clusters with different sizes and shapes but failure with variant densities. Our algorithm first finds the nearest neighbour of each data point and then redefines the similarity between pairs of points in terms how many nearest neighbours the two points the share.

II. RELATED WORK

Finding clusters of different shapes sizes; especially in the presence of noise is a problem that many recent clustering algorithms have addressed. DBSCAN, CURE, and Chameleon have shown good results. In DBSCAN the neighbourhood within a radius ε of a given object is called the ε-neighbourhood of the object [2]. Then the ε-neighbourhood of an object contains at least a minimum number, minpts, of the object, and then the object called core objects. An object p is density-reachable from object q with respect to ε and minpts in a set of objects, D, if there is a chain of objects $p_1,\ldots,p_n$, where $p_1=q$ and $p_n=p$ such that p_{i+1} is directly density-reachable from pi with respect to ε and minpts, for $1\leq i\leq n$, $p_i \in D$ [10]. In Fig 1 DBSCAN algorithm fails in case of changeable densities and neck type of dataset that is shown below in the pictorial representation

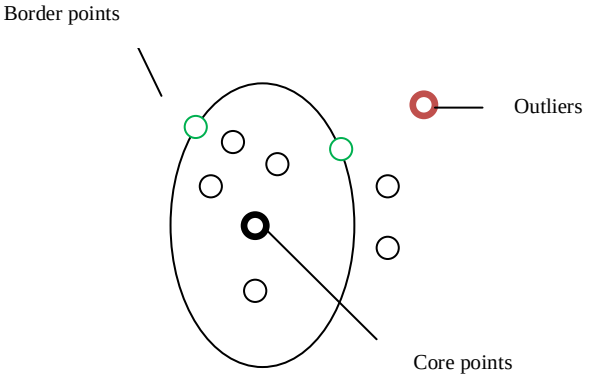

Fig 1: DBSCAN with points

CURE (Clustering using Representatives) identify clusters having non sphere-shaped and wide variances in shapes [7]. CURE employs a novel hierarchal clustering algorithm that adopts middle ground between the centric based and all point extremes. In Fig 2, a constant number c of well scattered points of a cluster are chosen and they are shrunk towards the centric of the cluster by a fraction α. The speckled points after shrinking are used as representatives of the cluster. The clusters with the nearby pair of representatives are the clusters that are combined at each step of CURE's hierarchical clustering algorithm. But in the CURE cluster proximity yet ignore cluster interconnectivity. In Chameleon (Hierarchical clustering algorithm) that uses dynamic modelling to determine the similarity between pairs of objects. Chameleon first finds the similarity between each pair of clusters C_i and C_j

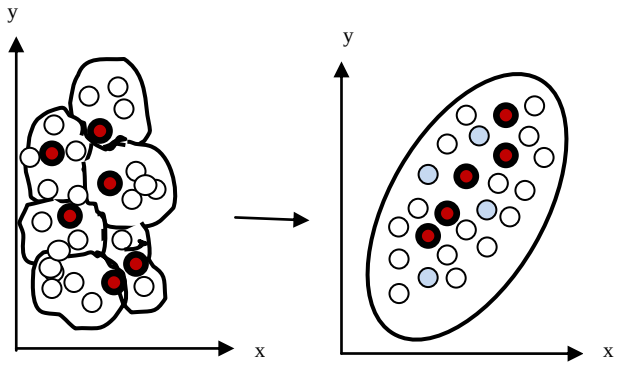

Fig 2: Shrinking representative objects

according to their interconnectivity, $RI(C_i,C_j)$ as follows

$$RI(C_I,C_J) = \frac{|EC_{\{ci,ci\}}|}{½(|EC_{ci}|+|EC_{ci}|)} \quad \text{.....(i)}$$

After finding the connectivity the closeness can be determined using the parameter Ci and Cj. However processing time for high-dimensional data may require more even in worst cases. However, we believe that using representative points to deal with differing sizes and shapes, the difficulty of dealing with clusters of conflicting densities, the significance of eliminating outliers and noise, and the problem with similarity that can arise particularly in higher dimensions.

III. ENHANCED DEFINITION OF SIMILARITY

In our role first we present a clustering approach that can simultaneously address different clustering issues for a variety of data sets and then alter the similarity between pairs of points in terms how many nearest neighbours share the two points. The noise and outliers can be eliminated using the similarity function. While our algorithm has many features [1]. First the algorithm does not cluster all the points because the noise in data is removed and then clustering is desired so that unclustered data is found as core points. Finally the points added to clusters containing the closest representative objects. Much of strength of our approach came from Chameleon, CURE and DBSCAN, although our basic inspiration derives from Jarvis-Patrick clustering technique. Clustering technique combine the notion of representative points, creating the clustering algorithm which incorporates varied ideas. Clustering allows for unsupervised learning. That is, the machine / software will learn on its own, using the learning set data, and will classify the objects into a particular class. Consider the pair of six-dimensional data points, 1 and 2, shown below, which have binary attributes. Here we have taken six attributes with two points as the binary attributes for calculating the similarity between the points.

TABLE I
BINARY ATTRIBUTES

pts	U1	U2	U3	U4	U5	U6
1	0	0	0	0	0	1
2	1	0	0	0	0	0

If we calculate the Euclidean distance between these two points, we get √2. Now consider the next pair of six-dimensional points, 3 and 4.

TABLE II
BINARY ATTRIBUTES

pts	U1	U2	U3	U4	U5	U6
3	0	0	0	0	0	1
4	1	0	0	0	0	0

If we calculate the distance between points 3 and 4, we again find out that it is √2. Clearly Euclidean distance does not capture the similarity of points with binary attributes. And also formulae will not for high dimensional data. For example cluster the following eight points with (x, y) representing locations into three clusters A1(2.5, 10) A2(2.5, 5) A3(8, 4.4) A4(5.3, 8) A5(7.3, 5) A6(6.7, 4) A7(1.1, 2) A8(4.1, 9). Initial cluster canters are: A1(2.1, 10),A4(5.1, 8) and A7(1, 2.1). The distance function between a and b are a=(x1, y1) b=(x2, y2) is defined as: $\rho(a, b) = |x2 – x1| + |y2 – y1|$ Use k-means algorithm to find the three cluster canters after the second iteration. If we calculate the above problem we will get the values that shown below in graph. For this problem the solution shown in Fig 3.However, the traditional

Euclidean distance notion of density is meaningless in number of points. When the dimension increases, volume automatically increases and the density reaches 0. Different measures, such as cosine measure and jaccard measure have been suggested to address this problem. But the triangular inequality doesn't hold. DBSCAN is the basic algorithm of density clustering. DBSCAN requires two global points' minimum objects and radius but it does not specify the upper limit of the core. An alternative to direct similarity is to define the similarity between the pairs of points in terms of their shared nearest neighbour.

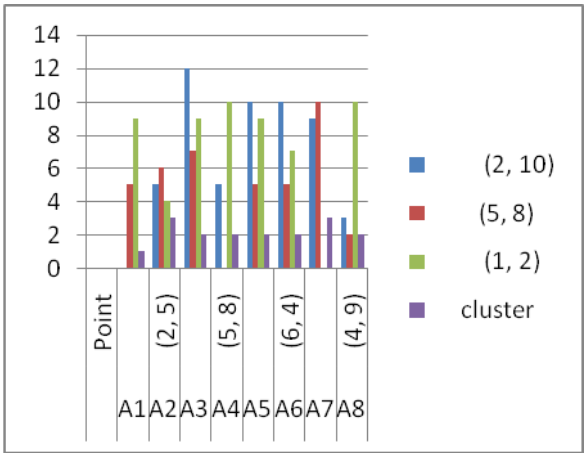

Fig 3: Clustering graph

For example if point A is close to point B and if they are both close to set of points C then we can say that A and B are close with greater confidence. The weight between the two points is calculated in the SNN graph. Let i and j be two points. The strength of the link between i and j is calculated using the equation,

Strength $(i,j)=\sum(k+l-n)$, where $i_m=j_m$ **..... (ii)**

In the shared nearest neighbours the links between the two points is constructed only if they are only in nearest neighbour lists.

A. SNN clustering Algorithm

The steps of the SNN clustering algorithm are as follows [6]:

1) Generate the similarity matrix.
2) Sparsify the similarity matrix using k-nn sparsification.
3) Generate the shared nearest neighbor graph from k-nn sparsified similarity matrix.
4) For every point in the graph, evaluate the total strength of links coming out of the point.
5) Determine representative points by choosing the points that have high total link strength.
6) Determine noise points by choosing the points that have low total link strength and remove them.
7) Remove all links that have weight smaller than a threshold.
8) Connected components points were taken to form clusters, where each point in a cluster is either a representative point or is linked to representative point.

In this section, we consider an application of our SNN clustering technique to transactional data, i.e. banking system. In particular our data consist of yearly measurement of customer level management in loan such as home, educational, vehicle. Briefly, here we interested in to show the behaviour of correlates of all transactions in banking. The number of cluster is not given to the algorithm as a parameter. In order to envisage how the SNN clustering algorithm described in this paper compare with other clustering algorithm, we formed a transactional dataset as well as CURE. The dataset consist of 4 globular data clusters of differing densities. All the links in the SNN graph are counted to get the number of

TABLE III
TRANSACTIONAL DATA

Cust_id	Loan type	Year
8785913	Home	2011-2012
8785813	Educational	
7845123	Home	
7854637	Car	
6784563	Educational	
7845123	Car	
7645324	Business	
7895623	Business	

links that a point has, regardless of the strength of the links. Clusters obtained by Jarvis-Patrick method using the smallest possible threshold. Representative points and noise points can be found by looking the sum of link strengths for every point in the SNN graph. The point that have high total link strength then become the candidate for representative points, while the low total link strength become candidate for noise points. Depending on the nature of the data, the algorithm finds clusters for the given data set. Here we have given the data in table III. For the above table we have presented the SNN clustering according to their behaviour as well as with CURE. DBSCAN is a pioneer density based algorithm to find out the clusters of different shapes and sizes from large amount of data. DBSCAN was able to separate the three clusters, but classified everything else as noise and also it fails in the case of different densities and neck points.

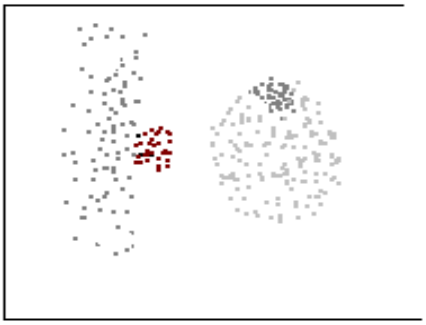

Fig 4. Clusters produced by CURE

In Fig 4, we see that CURE failed to detect the clusters correctly, it almost found one clusters correctly.In CURE the clusters produced is tedious to find accurately. Mostly there are four clusters but here CURE clearly shown only one cluster[7]. But in Fig 5 our SNN clustering algorithm was able to identify each of these clusters in the dataset, although 15% of the points were not assigned to any clusters.

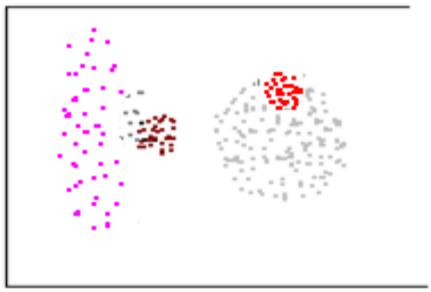

Fig 5 SNN clustering

IV. CONCLUSIONS

In this paper we described shared nearest neighbor algorithm to find the clusters even in the presence of noise and outliers and the algorithm automatically finds the clusters according to their behaviours. The SNN clustering algorithm can find clusters with respect to their surrondings and behaviours. Novel clustering technique also presented to solve the time series problem. Here clusters are found according to their behaviour in transaactional data that is loan type is clustered according to the behaviour.

REFERENCES

[1] Ahmad Ali Abin, Hamid Beigy, "An algorithm for discovering clusters of different densities or shapes in noisy data sets,"SAC'13 March 18-22,2013.

[2] Anant Ram, Sunitaj Jalal, Anand S.Jalal, Monij Kumar,"A density based algorithm for discovvering density varied clusters in large spatial databases,"International journal of computer applicationals(0975-8887), vol 3,No.6,june,2010.

[3] Mohammed T.H.Elbatta and Wesam M.Ashour,"A dynamic method for discovering density varied cluster,"International journal of signal processing, image processing and pattern recognition , vol.6, No.1,Feb,2013.

[4] J.Hencil peter, A.Antonysamy,"Heterogeneous density based spatial clustering of applications with noise," International journal of computer science and network security, vol.10,No.8,Aug,2010.

[5] R.Bharathi, S.Vijayalakshmi,"Improved varied density based spatial clustering of applications with noise by probability based k selection," International journal of computer science and management research, vol.1,issue 4,Nov, 2012, ISSN 2278-733X.

[6] (2013) The IEEE website. [Online]. Available: http://www.cs.umn.edu/rertoz/snn/

[7] Levent Ertoz, Michael steinbach, Vipin Kumar,"Finding clusters of different sizes, shapes and densities in noisy, high dimensional data,"cs.umu.edu,Feb 20,2003.

[8] Margaret H. Dunham, 'Data Mining: Introductory and Advanced Topics' Published by Dorling Kindersley (India) pvt.Ltd,2006,ISBN 978-81-7758-785-2.

[9] J.Kan, M.Kamber, "Data mining concepts and techniques", Morgan Kaufmann Publishers, San Fransisco,USA,2001,ISBN 1558604898.

[10] A.M.Khattak, A.M.Khan, Sungyoung Lee, Young-Koo Lee, "Analysing Association Rule Mining and Clustering on sales day data with XLMiner and Weka," International journal of databases theory and applications, Vol.3, No.1, March, 2010.

[11] Alexandros Nanopoulos, Apostolos N.Papadopoulos, Yannis Manolopoulos,"Information system 32 649-669, 2007.

