Analysis of Mass Based and Density Based Clustering Techniques on Numerical Datasets

Rekha Awasthi[1]* Anil K Tiwari[2] Seema Pathak[3]

1. Disha College Ram Nager Kota Raipur 492010 Chhattisgarh India
2. Disha College Ram Nager Kota Raipur 492010 Chhattisgarh India
3. Disha College Ram Nager Kota Raipur 492010 Chhattisgarh India

*E-mail: awasthirekha151@gmail.com

Abstract

Clustering is the techniques adopted by data mining tools across a range of application . It provides several algorithms that can assess large data set based on specific parameters & group related points . This paper gives comparative analysis of density based clustering algorithms and mass based clustering algorithms. DBSCAN [15] is a base algorithm for density based clustering techniques. One of the advantages of using these techniques is that method does not require the number of clusters to be given a prior and it can detect the clusters of different shapes and sizes from large amount of data which contains noise and outliers. OPTICS [14] on the other hand does not produce a clustering of a data set explicitly, but instead creates an augmented ordering of the database representing its density based clustering structure. Mass based clustering algorithm mass estimation technique is used (it is alternate of density based clustering) .In Mass based clustering algorithm [22] there are also core regions and noise points are used as a parameter. We analyze the algorithms in terms of the parameters essential for creating meaningful clusters. All the algorithms are tested using numerical data sets for low as well as high dimensional data sets.

Keywords: Mass Based (DEMassDBSCAN) ,DBSCAN,OPTICS.

1. Introduction

Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering . In this research paper we are working only with the clustering because it is most important process, if we have a very large database. We are using Weka tools for clustering . Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. We are using Weka data mining tools for this purpose. It provides a batter interface to the user than compare the other data mining tools.

2. Literature Survey

Clustering is an active research area in data mining with various methods reported. Xin et al. and Howard et al. made a comparative analysis of two density-based Clustering algorithms i.e. DBSCAN and DBRS which is a density-based clustering algorithm. They concluded that DBSCAN gives extremely good results and is r is efficient in many datasets. However, if a dataset has clusters of widely varying densities, than DBSCAN is not able to perform well. Also DBRS aims to reduce the running time for datasets with varying densities. It also works well on high-density clusters. In [19] Mariam et al. and Syed et al. made a comparison for two density-based clustering algorithm i.e. DBSCAN and RDBC i.e. Recursive density based clustering. RDBC is an improvement of DBSCAN. In this algorithm it calls DBSCAN with different density distance thresholds ε and density threshold MinPts. It concludes that the number of clusters formed by RDBC is more as compared to DBSCAN also we see that the runtime of RDBC is less as compared to DBSCAN. In [1] K.Santhisree et al. described a similarity measure for density-based clustering of web usage data. They developed a new similarity measure named sequence similarity measure and enhanced DBSCAN [15] and OPTICS [14] for web personalization. As an experimental result it was found that the average intra cluster distance in DBSCAN is more as compared to OPTICS and the average intra cluster distance is minimum in OPTICS. In [17] K.Mumtaz et al. and Dr. K.Duraiswamy described an analysis on Density-Based Clustering of Multi-Dimensional Spatial Data. They showed the results of analyzing the properties of density-based clustering characteristics of three clustering algorithms namely DBSCAN, k-means and SOM using synthetic two dimensional spatial data sets. It was seen that DBSCAN performs better for spatial data sets and produces the correct set of clusters compared to SOM and k-means algorithm In [11] A.Moreia, M.Santos and S.Corneiro et al. described the implementation of two density based clustering algorithms: DBSCAN [15] and SNN [12]. The no of input required by SNN is more as compared to DBSCAN. The results showed that SNN performs better than DBSCAN since it can detect clusters with different densities while the former cannot.

3. Mass Based And Density-Based Clustering Techniques

The mass estimation is another technique to find clusters in arbitrary shape data. In the clustering the mass estimation is unique because in this estimation there is no use of distance or density [20].

DEMassDBSCAN clustering mass estimation technique is used (it is alternate of density based clustering) .In DEMassDBSCAN algorithm there are also core regions and noise points are used as a parameter.

Density based clustering is to discover clusters of arbitrary shape in spatial databases with noise. It forms clusters based on maximal set of density connected points. The core part in Density-Based clustering is density-reach ability and density connectivity. Also it requires two input parameters i.e. Eps which is known as radius and the MinPts i.e. the minimum number of points required to form a cluster. It starts with an arbitrary starting point that has not visited once. Then the ε- neighborhood is retrieved, and if it contains sufficiently many points than a cluster is started. Otherwise, the point is labeled as noise. This section describes two density based clustering algorithms briefly i.e. DBSCAN (Density Based Spatial Clustering of Application with Noise) and OPTICS (Ordering Points to Identify the Clustering Structure). Here, Density=number of points within a specified radius. Density-Based clustering Algorithms mainly include two techniques:

- DBSCAN [15] which grows clusters according to a density-based connectivity analysis.
- OPTICS [14] extends DBSCAN to produce a cluster ordering obtained from a wide range of parameter setting.

Mass Based Clustering:

In the clustering approach there is various estimation techniques are used to form clusters. The mass estimation is another technique to find clusters in arbitrary shape data. In the clustering the mass estimation is unique because in this estimation there is no use of distance or density [20].

In the DBSCAN clustering algorithm hyper spheres are used to show points but in DEMassDBSCAN clustering algorithm rectangular regions are used for showing the points [21]. mass - Number of points in a given region is called mass of that region. Each region in a given data space having rectangular shape and for estimation of mass a function that called rectangular function [18].

One-Dimensional Mass Estimation: The estimation of mass in data space is depends on levels (h). If value of level h=1 the shape of function is concave. For multidimensional mass estimation the value of h is > 1.

$$mass\ (x,h) \begin{cases} \sum_{i=1}^{n-1} mass_i\ (x, h-1) p(s_i), h > 1 \\ \sum_{i=1}^{n-1} m_i(x) p(s_i), \qquad h = 1 \end{cases} \qquad \text{Eq- 1 [20]}$$

In the one dimensional mass estimation the value of $p(s_i)$ is equals to

$$p(s_i) = \frac{x_{i+1} - x_i}{x_n - x_i}$$

For one-dimensional mass estimation the value of h is equals to 1.

$$\sum_{i=1}^{n-1} m_i(x) p(s_i), \qquad h = 1$$

In the above equation the m_i(x) mass base function.

$$\overline{mass}\ (x,h) = \frac{l}{t}\sum_{k=1}^{1} mass\ (x,h/D_k)\ \ Eq-2$$

Combining these two equations-

$$\sum_{i=l}^{n-1} m_i(x)p(s_i) \approx \frac{l}{t}\sum_{i=l}^{t} m\big(T_i(x)\big)$$

$$mass\ (x,h) \approx \frac{l}{t}\sum_{i=l}^{t} m\left(T_i^h(x)\right)$$

$$\overline{mass}\ (x,h) \approx \frac{l}{t}\sum_{i=l}^{t} m\left(T_i^h(x/D_i)\right)$$

For large data bases the values of t is 1000 and ψ256. For small data sets the value of h and is ψ is varies.

In multidimensional Mass Estimation : In multidimensional mass estimation there is value of h>1. In multidimensional mass estimation there are two functions mass and random spaces generator in data space.

$\overline{gmass}\ (x) \approx \frac{l}{t}\sum_{i=l}^{t} m\left(T_i^h(x/D_i)\right)$ [20]

Mass Based Clustering:

$h\!:\!d - Tree(x) = m_j = m\left(T^h(x/D)\right)$ [20]

Multidimensional mass estimation using h:d Trees:

$\overline{gmass}\ (x) \approx \frac{l}{t}\sum_{i=l}^{t} h\!:\!d - (x)Tree_i(X)$ [20]

The multi dimensional data set the value of h (level) is greater than 1and is is similar to one dimensional mass estimation. In this equation the value of x is replaced by X.

MassTER Algorithm: In the MassTER algorithm the mass will estimate using h:d trees. In the h:d trees the h stands for no of times attribute appears and d stands for no. of dimensions. The height of tree is calculated as l=h×d.

Algorithm

Step 1 Build trees t_i

Step 2 Assign a cluster seed for every instance.

Step 3 Join two seed based clusters.

Step 4 Cluster having instances less than Ƞ are called noise instances.

Step 5: return clusters c_j,j=1,2,3,-------k. And E. (noise instances)

Data input:

D= Input Data.

T= No of trees.

Ψ= Sampling size.

h= how many times attribute come in a path.

Ƞ= minimum instances in a cluster

In this algorithm firstly h:d trees are made of instance in a region. The cluster seed is assigned for each region in data space. The cluster seed are joined and make pair with neighbor cluster seed. The instances having value less than Ƞ are defined noise [20].

Tool

In the implementation of data mining algorithms mostly Weka toolkit is used. Weka is developed by University of Waikato in New Zealand Weka . Weka is popular open source machine learning software and it provides Graphical user interface. Weka is mainly consists of four interfaces like Explorer, Experimenter, Knowledge flow and Simple CLI [23]. In the implementation of any algorithm .arff file format data is used. Weka's main user interface is the explorer, but essentially the same functionality can be accessed through the component based Knowledge Flow interface and from the command line. By using Weka we can implement all standard algorithms [23].

4. Experimental Result

In this paper there is DBSCAN and OPTICS clustering algorithms are used for benchmark. The clustering result is analyzed in terms of CPU run time (in nano seconds), clustered instances, unclustered instances. In the implementation of massTER the two data sets are used. The experimental result will described in following subsections. In the DEMassDBSCAN algorithm the values of ψ and t is taken 256 and 1000 by default. In the implementation of DEMassDBSCAN and DBSCAN and OPTICS algorithm there are only one parameter is used at one time h for DEMassDBSCAN and ε for DBSCAN and OPTICS. For implementation of DEMassDBSCAN and DBSCAN and OPTICS algorithm we take following data sets that are shown below:

Data sets :

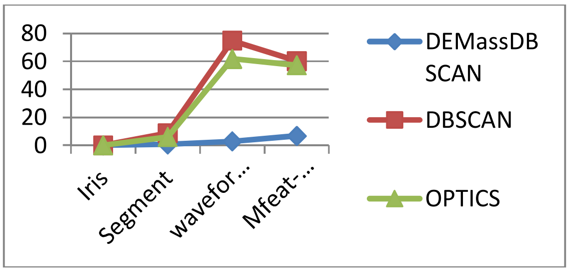
The Iris data set contains 60 instances and 16 attributes. In the clustering result the run time is better than DBSCAN algorithm , The Segment Data set contains2310 instances and 20attributes. In the clustering result the run time is better than DBSCAN algorithm , The Waveform 5000 Data set contains 5000 instances and 41 attributes. In the clustering result the run time is better than DBSCAN algorithm and The Diabetes data set contains 2000 instances and 217 attributes. In the clustering result the runtime is better than DBSCAN algorithm .

Data Set	DEMassDBSCAN			DBSCAN			OPTICS		
	Run time	Cluster	Unassigned	Run time	Cluster	Unassigned	Run time	Cluster	Unassigned
Iris Data set	0.03	1	3	0.05	1	6	0.04	1	4
Segment Data set	0.67	3	1	8.78	7	1	6	7	1
Waveform 5000 Data set	2.8	13	46	74.85	3	53	61.08	3	49
Diabetes Data set	6.69	1	36	60.22	10	19	57.22	10	19

On the basis of run time behavior the DEMassDBSCAN clustering algorithm is better than DBSCAN clustering algorithm.It gives better result when applied on large datasets. The result is shown in Table

	Run time behavior			
	Iris	**segment**	**Waveform 5000**	**Diabetes**
DEMassDBSCAN	0.03	0.67	2.8	6.69
DBSCAN	0.05	8.78	74.85	60.22
OPTICS	0.04	6	61.8	57.22

The comparison between DEMassDBSCAN And DBSCAN and OPTICS Clustering algorithm is based on their **RUN TIME** is shown below.



5. Conclusion

In this paper the runtime behavior of the DEMassDBSCAN algorithm is better than DBSCAN clustering algorithm But run time of OPTICS is better than DBSCAN .and It will gives better results on the large data sets. The un-assigned clusters are less than the DBSCAN clustering algorithm. The result of the DEMassDBSCAN is efficient and having less noise points out of whole data set. In future mass estimation is applied on other tasks. but some time runtime of the OPTICS it better than DEMassDBSCAN algorithm run time is varied according to data size .

References

[1] Ms K. Santhisree, Dr. A. Damodaram, August 2011 SSM-DBSCAN and SSM-OPTICS : Incorporating new similarity measure for Density based clustering of Web usage data, in International Journal on Computer Sciences and Engineering, August 2011

[2] , S. Chakraborty, Prof. N. K. Nagwani, Vol. 1, July 2011 Analysis and Study of Incremental DBSCAN Clustering Algorithm, International Journal of Enterprise Computing And Business Systems, Vol. 1, July 2011

[3].M. Parimala, D. Lopez, N. C. Senthilkumar, Vol. 31, June 2011 A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases, International Journal of Advanced Science and Technology, Vol. 31, June 2011.

[4] Dr. Chandra. E, Anuradha. V. P, 24, June 2011 A Survey on Clustering Algorithms for Data in Spatial Database Management System, International Journal of Computer Applications, Col. 24, June 2011

[5] J. H. Peter, A. Antonysamy, Vol. 6, September 2010 An optimized Density based Clustering Algorithm, International Journal of Computer Applications, Vol. 6, September 2010

[6] A. Ram, S. Jalal, A. S. Jalal, M. Kumar Vol. 3, June 2010, A Density based Algorithm for Discovering Density varied clusters in Large Spatial Databases, International Journal of Computer Applications, Vol. 3, June 2010

[7] Tao Pei, Ajay Jasra, David J. Hand, A. X. Zhu, C. Zhou, ,2009 DECODE: a new method for discovering clusters of different densities in spatial data, Data Min Knowl Disc, 2009

[8] Zhi-Wei SUN, 2008 A Cluster Algorithm Identifying the clustering Structure, International Conference on Computer Science and Software Engineering, 2008

[9] Marella Aditya, 2007‖DBSCAN And its Improvement‖,june 2007

[10] Stefan Brecheisen, Hans-Peter Kriegel, and Martin Pfeifle, Vol. 9, 2006 Multi-step Density Based Clustering, Knowledge and Information Systems, Vol. 9, 2006

[11] A. Moreira, M. Y. Santos and S. Carneiro, July 2005 Density-based clustering algorithms-DBSCAN and SNN, July 2005

[12] M. Rehman and S. A. Mehdi, 2005Comparision of Density-Based Clustering Algorithms, 2005

[13] Levent Ertoz, Michael Steinback, Vipin Kumar, Finding Clusters of Different Sizes, Shapes, and Density in Noisy, High Dimensional Data, Second SIAM International Conference on Data Mining, San Francisco, CA, USA, 2003

[14] M. Ankerst, M. M. Breunig, H. P. Kriegel and J. Sander, 2003 OPTICS: Ordering Points To Identify Clustering Structure, at International Conference on Management of Data, Philadelphia, ACM 1999

[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu, A Density- Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, The Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA, 1996

[16] X. Wang, H. J. Hamilton. A Comparative Study of Two Density-Based Spatial Clustering Algorithms for Very Large Datasets.

[17] K. Mumtaz, Dr. K. Duraiswamy, Vol 1 No 1 8-12 An Analysis on Density Based Clustering of Multidimensional Spatial Data in Indian Journal of Computer Science and Engineering Vol 1 No 1 8-12

[18] J.Han and M Kamber “data mining concept and technique “

[19] Mariam rehman comparison of Density based algorithm

[20]. K.M Ting, G-T Zhou, F. T Liu and J.S.C Tan, 2010 “Mass Estimation and its Applications”, Proceedings of KDD, 2010

[21]. K.M. Ting, Jonathan R Wells2010, “Multi-Dimensional Mass Estimation and Mass-based Clustering”, IEEE Proceedings 2010.

[22]. Kai Ming Ting, Takashi Washioy, Jonathan R. Wells and Fei Tony Liu, 2011 “Density Estimation based on Mass,” international conference on data mining, 2011.

[23]. Z. Markov, and I.Russell “An introduction to WEKA data mining system” tutorial. http://www.cs.ccsu.edu