Intelligent Systems and Computer Technology
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC200184

Dimensionality Reduction of Production Data Using PCA and DBSCAN Techniques

Umadevi S[a,1], NirmalaSugirthaRajini[b]
[a]*Research scholarDepartment of Computer Applications*
[b]*Professor, Department of Computer Applications*
Dr.MGR Educational and Research Institute, Chennai

Abstract. Now a day's data mining concepts are applied in various fields like medical, agriculture, production, etc. Creation of cluster is one of the major problems in data analysis process. Various clustering algorithms are used for data analysis purpose which depends upon the applications. DBSCAN is the famous method to create cluster. This article describes DBSCAN clustering concept applied on production database. The main objective of this research article is to collect and group the related data from large amount of data and remove the unwanted data. This clustering algorithm removes the unwanted attributes and groups the related data based upon density value.

Keywords. Dimensionality Reduction, DBSCAN, Production Data. PCA, Clustering

1. Introduction

In current scenario computing technologies are used in various domains. Specifically data mining algorithms are used in healthcare domain, manufacturing domain fraud detection area, customer segmentation area etc. Data mining concept is also used in production department to predict the production value. In this department large amount of data are available. To process this large amount of data is a very tedious process. Data mining PCA (Principle Component Analysis) is applied on the production data to reduce the number of attributes. This is an approach used to find the strong patterns from the original data set and remove the weaker patterns. The usage of strong pattern is easy to take a look at and envision the data. DBSCAN clustering concept is used to group the related data and remove the noise data from original dataset.

The second part of this article reviews the usage of DBSCAN clustering concept in various domains. The third part deals with the process flow diagram of proposed system. Part four deals with result and discussion part. Finally part five concludes the article.

[1]S.Umadevi, Research scholar, Dept of Computer Applications, Dr.MGR Educational and Research Institue, Chennai;
Email.umakalyani70@gmail.com

2. Literature Review

D. Hemavathi et al., says that collection of related group is the important concept in real world applications. Related group selection concept was applied in various fields like financial institutions, education system, healthcare system etc.The relevant information produces only the better result compared with original data. In this research article the authors used Terry data set for applying unsupervised concept. Here the important features are detected by using PCA concept. After those DBSCAN concepts was used to evaluate the data for better performance [1].RekhaAwasthi et al., explained about the importance of data mining tools. It is used to analyze various data.

Data mining concepts provides a facility to analyze the information from various angles, group it and make the summary of their relationships was finalized. In medical data set, mining techniques were used to help the healthcare professionals to predict the diseases in earlier manner and provides the better solution to the patients. Quantities of medical documents are very high. Various patterns are available in the medical data. Detect the diseases in earlier stage by using discover the hidden patterns. To preprocess the data normalization concept or PCA technique was used in medical data set analysis. DBSCAN concept and OPTICS technique were applied on the medical data set for improve processing speed[2].

SnehalD.Borase ET AL., says that clustering is a technique to find the homogeneous groups from the classes. Clustering techniques is used in various application areas like biotechnology, healthcare domain, data recovery CRM, sales and web content analysis. Many researchers involved to create clustering techniques. In big data contains large volume of data. In this research article the authors reviewed various existing clustering concepts and the importance of dimensionality reduction and connected with unwanted data [3].

Shuaipeng Liu ET AL., using dimensionality reduction and clustering concept to identify objects n the road. Tiny obstacles on the path are creating very dangerous hazards while driving. In this article authors designed a new algorithm to detect tiny objects on the road using radar with multi layer. The path edge points are extracted by using filtering concept. The interference point was detected by using Hough Transform. Small objects are detected by using DBSCAN clustering algorithm. This proposed algorithm was designed and developed and tested with real time date [4].

Qi Chen et al., explained about the process of mage segmentation. Using this concept the image is divided into small regions for future usage. Using DBSCAN algorithm the colored pictures are segmented. This algorithm makes easy for detecting the arbitrary shape of clusters. The main disadvantage of this algorithm is it has high complexity when the image size is larger. Self-Organizing Map (SOM) is used to reduce the dimensionality of image processing concept. Here the authors used hybrid technique called SOMDBSCAN (SOM and DBSCAN) for segment the images. Using four images the proposed method was evaluated [5].

Martin Ester, et al., was discussed about the usage of clustering approach spatial type databases. In this article the authors used DBSCAN concept to find the clusters from spatial database. The shape of the cluster is arbitrary shape. This clustering concept uses only one input argument and helps the user for determining the actual value. This new algorithm was tested with SEQUOIA 2000 benchmark real data.

The experiment result shows that DBSCAN clustering concept is more effective compared with other clustering techniques [6].

Tianfu Wang et al.,says that clustering is the important concept in spatial data mining. It categorizes the groups based on their similarities in attributes and also based on location wise. It plays a major part in distribution recognition based on density, detection of hot spot and discovers the recent trends. This article described about DBSCAN clustering concept [7].

SlavaKisilevich et al., says that analyzing large volume of data is very difficult task. Growth of communication technology leads most of the people to use cheap storage. Due to this reason large amount of users data are available on the internet like photos and human's movement etc. These users' generated data provides a new challenging task to the researchers. Here the authors presented PDBSCAN concept based on DBSCAN approach for analysis users data. In this algorithm two new concepts were introduced. The first concept is density threshold value based on the number of persons in the surroundings. The second concept is adaptive density[8].

Xiaolu Li et al., said that from the location based applications, huge location data were composed in real time. In this research paper the authors used DBSCAN with Gauss mixture model to create clusters and remove noises from location information. Here to find the density level initially by using Gaussian mixture model. Then DBSCAN approach can be used to cluster the data locally. At the same time seed values are selected for finish the cluster formation. In the final step all clusters are combined.

This new method was validated in terms of accuracy of clustering, intensity level of noise and efficiency. The experiment result showed that this new concept provided better result compared with an existing cluster method[9]

3. Proposed System

Production database contains large amount of data with its attributes. Using PCA algorithms the dimensionality is reduced. After that the reduced value will be input of DBSCAN concept. This algorithm is used to make the clusters of related data and remove the unwanted noise data. The following fig 1 shows the proposed model.

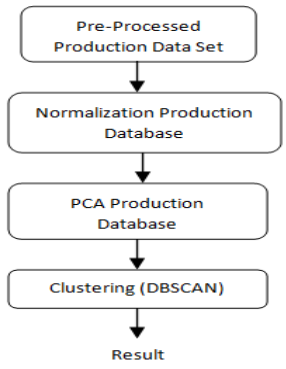

Figure 1. Proposed Model

Before going to apply PCA algorithm[10] the original production data set has been preprocessed by using various data mining concepts. After preprocessing the data should be normalized. DBSCAN(Density Based Spatial Clustering Applications with Noise) is one of the major clustering technique used to create a cluster based on density value. DBSCAN groups data points as core points, non-core points and outliers. It detects the noise data and put it separately in low density area. This algorithm uses two input arguments such as radius(Eps) and MinPtsie., minimum number of points on the data set. The following figure 2 shows the overall process of DBSCAN algorithm.

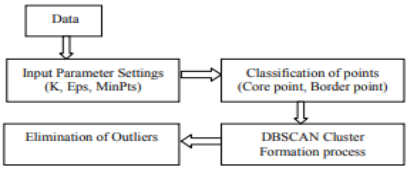

Figure 2. Overview of DBSCAN algorithm

To execute the DBSCN algorithm initially set the input parameters like K, EpS and MinPts values. The nest level of DBSCAN algorithm is classifying the data points such as core point, Border point and Noise point. DBSCAN is executed based density value of the dataset. If the point has more than a particular number of points (MinPts) with in Eps value is called core point. A border point means it has fewer than Minpts within the range of Epsvalue. A point is neither a core point nor a border point is called noise point. The noise points are considered as a outlier and removed from the clusters.

4. Result And Discussion

DBSCANclustering algorithm is used to make clusters based on the density value. It removes the noisy points from the reduced dataset. The following diagram 3 shows the output of DBSCAN concept. As we can see graph below there are four clusters that identified with DBSCAN algorithm. In each clusters some of the attributes are common. Using this common attributes the data will be classified using any machine learning algorithms.

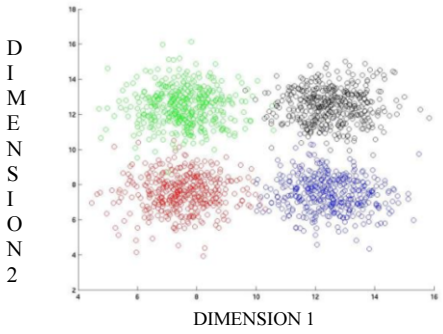

Figure 3. Output of Proposed Concept

5. Conclusion

In production department in any organization maintains a database. This database contains large amount of data. Various data mining concepts are used to analyze the production data. If the volume of the data is high, it is very difficult to analyze the data. Here PCA concept is used to reduce the data. The data is clustered by using DBSCAN algorithm. The clusters will grow based upon the density value. The input value of DBSCAN algorithm is taken from the output of PCA algorithm. DBSCAN clustering concept removes the unwanted data from reduced dataset. The output of clustering algorithm has been used to classify the various values using machine learning concepts. This proposed concept is used to predict the production value in manufacturing organizations easily.

References

[1] D. Hemavathi H. Srimathi& K. Sornalakshmi(2019), A Hybrid Technique for Unsupervised Dimensionality Reduction by Utilizing Enriched Kernel Based PCA and DBSCAN Clustering Algorithm, International Conference on Inventive Computation Technologies, Springer, pp 476-488

[2] RekhaAwasthi, Anil Kumar Tiwari&SeemaPathak, An Analysis Of Density Based Clustering Technique With Dimensionality Reduction For Diabetic Patient, International Journal of Computer Engineering and Applications, 2015,Volume IX, Issue IV,pp. 165-171.

[3] SnehalD.Borase&SatishS.Banait, Dimensionality Reduction Using Clustering Technique, International Journal of Computer Applications , Emerging Trends in Computing 2016, ISSN: 0975 – 8887, pp. 17-22.

[4] Shuaipeng Liu , KekeGeng , Guodong Yin ,Conglei Wu & Jianwei Ye, Small Objects Detection with Multi-layer Laser Radar Based on Projection Dimensionality Reduction, 2019 Chinese Control Conference (CCC).

[5] Chen, Q., Yuen, K. K. F., & Guan, C. , Towards a Hybrid Approach of Self-Organizing Map and Density-Based Spatial Clustering of Applications with Noise for Image Segmentation, 10th International Conference on Developments in eSystems Engineering (DeSE),2017, pp 238-241.

[6] Martin Ester, Hans-Peter Kriegel, Jiirg Sander &XiaoweiXu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise From: KDD, Proceedings,1996, pp 226-231.

[7] Tianufu Wang, Chang Ren, Yun Luo and Jing Tian NS-DBSCAN. A Density-Based Clustering Algorithm in Network Space, International Journal of Geo-Information, 2019,Vo. 8, No. 5.

[8] SlavaKisilevich Florian Mansmann& Daniel Keim, P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos".

[9] Xiaolu Li ,Peng Zhang &Guangyu Zhu , DBSCAN Clustering Algorithms for Non-Uniform Density Data and Its Application in Urban Rail Passenger Aggregation Distribution,2019, pp 1-22.le (WA): IASP Press; c2003. p.437-68.

[10] V.D.Ambeth Kumar and Dr.M.Ramakrishan (2012) "Footprint Recognition with COP Using Principle Component Analysis (PCA) in the month of June for the Journal of Computational Information Systems (JCIS) Journal Volume 3, 4939-4950, June 2012