International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-5, Issue-3 E-ISSN: 2347-2693

To Evaluate and Improve DBSCAN Algorithm with Normalization in Data Mining: A Review

P.K. Dhillon[1*], A.S. Walia[2]

[1*] Dept. of Computer Science and Engineering, Sant Baba Bhag Singh University, Jalandhar, India
[2] Dept. of Computer Science and Engineering, Sant Baba Bhag Singh University, Jalandhar, India

[*]*Corresponding Author: dparamvir8@gmail.com, Tel.: +919878895846*

Available online at: www.ijcseonline.org

Received:06/Feb/2017 Revised: 15/Feb/2017 Accepted: 09/Mar/2017 Published: 31/Mar/2017

***Abstract*-- There is ample amount of data present in the whole world. The data is generated from various sources like companies, organizations, social networking sites, image processing, world wide web, scientific and medical etc. People have less time to look at whole data. They attended towards the precious and interested information. Data mining is technique which is used to extract meaningful information from huge databases. Extracted information is visualized in the form of statics, graphs, and tables and videos etc. There are number of data mining techniques, asymmetric clustering is one of them. Asymmetric technique is type of unsupervised learning. In this, data sets which have similarity are placed in one cluster and others are in different clusters. From, number of years various asymmetric clustering techniques are introduced which work well with datasets. These techniques do not work well with the complex and strongly coupled data sets. To reduce processing time and improve accuracy neural networks are combined with asymmetric clustering algorithms.**

Keywords: Backpropagation, Data mining, DBSCAN, neural network, normalization.

I. INTRODUCTION

The sheer amount of data is stored in the whole world, which is called big data. In 2001, it was assumed that about petabytes[1] of data is stored in the world and it is expected that it will be about zettabyte[1] in 2022. Mostly, data is generated by the social websites, market analysis medical field, web mining and image processing etc. This data is stored in large databases in the forms of tables, images and videos etc. called data warehouses. The process of extracting useful patterns or knowledge from data base is called data mining. The extracted information is visualized in the form of charts, graph and tables etc. Data mining is also known by another name called KDD (knowledge discovery from the database). In data mining, frequent item set is used to find relations between numerous numbers of fields in data mining. Association rules are used to discover the frequent data item sets. The concept of association rules is used in various fields like retail stores, market strategy and stock market etc.

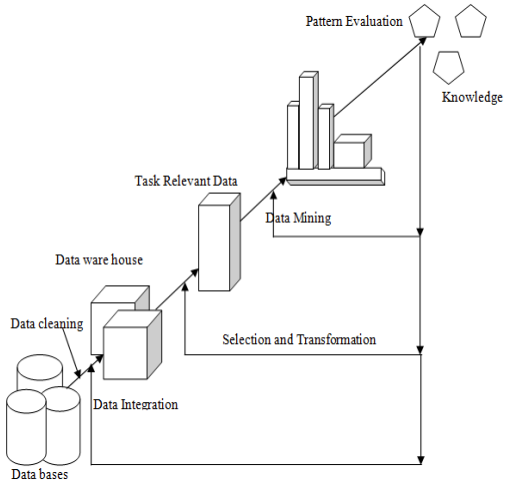

Figure 1: Data Mining Process [1-24]

These days Informational technology is mounting and databases created by organizations and companies like telecommunications, banking, marketing, transportation, manufacturing, and social networking sites etc. are becoming huge day by day. Knowledge discovery process is used to store this data in databases and efficiently access the interested or useful data from databases [2].

Knowledge discovery consist of following steps:

a) *Data Cleaning:* It is the step in which the process of detecting and removing of data which is not correct, irrelevant, containing missing values, duplicate values and noise that is dirty data from the database.

b) *Data Integration:* It is the step in which data from different sources is collected in one source to provide unified view of data.

c) *Data Selection:* It is the step in which data analysis is done in way that the selection of relevant data from databases.

d) *Data Transformation:* It is the step in which the data which is selected is reformed to correct form performing various operations like summary, aggregations, generalizations and normalized operations.

e) *Data Mining:* This is important technique in which intelligent operations are used to extract the useful pattern from the database.

f) *Pattern Evaluation:* It is the step in which the required pattern are evaluated from the given database.

g) *Knowledge Representation:* It is the step where whole process of output is visualized to user in the form of graphs, tables and graphs etc.

A. *Classification of Data Mining System:* According to following categories data mining system is classified:

   a) *According to Data source to be mined:* Data mine system can be classified according to mined techniques used like spatial data, multimedia data ,time-series data etc

   b) *According to Data models:* Data mine systems may use many models like relational model, object oriented model and transactional models**.**

   c) *According to kind of Knowledge mined:* Data mine system can be classified according to the type of knowledge is used like classification, prediction, cluster analysis and outlier analysis.

   d) *According to utilized Mining technique:* Data mine system can be classified according to techniques used for data mining techniques like decision tree, neural network etc.

   e) *According to adapted applications:* Data mine systems can be classified according to applications adapted like in finance, data mining system related to finance is used.

B. *Major issues in Data Mining:*

There are various data mining algorithms and techniques but there is enormous volume of data in world and there is continuous spike in the data, major issues that can be raised in data mining systems can be scalability and reliability of performance of data mining system [2].

Various performance issues are:

a) *Effective, Efficient and Scalable data mining:* in order to efficiently extract the useful knowledge from the large amount of databases, the technique of data mining which we are using should be effective, efficient and scalable, gives desired outputs in the desired time.

b) *Parallel, Distributed and Incremental mining algorithms:* The volume of data present in the databases is very huge and to maintain the complexity of data, data mining techniques prompt to develop the parallel and distributed data mining algorithms. Data in these algorithms is stored in different partitions and processed parallel. The output which comes from these partitions is combined to provide desired results and this is quite tough job to mine data without any scratch.

C. *Clustering in Data Mining:*

Clustering means putting objects having similar properties into one group and objects having dissimilar properties into another [2]. For example, those object who have values above threshold values can be placed in one cluster and values below into another cluster .Clustering divide the large data set into groups or clusters according to similarity in properties.

Clustering is an unsupervised learning technique as there are no classifiers and their labels .It is form of learning by observation. Cluster analysis can be used in the areas such as image processing, analysis of data, market research (buying patterns) etc. Using clustering we can do outlier [3] detection where outliers are values lying outside the cluster.

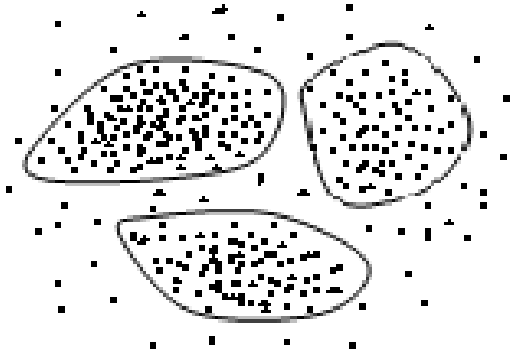

Figure 2: Clusters and Outliers [2]

In the above figure the dots which are outside the clusters represent outliers and clusters of object with similar properties.

II. PROBLEM FORMULATION

Cluster analysis is being broadly used in several applications like basket analysis, e-commerce, image processing, scientific and medical field, data analysis, and World Wide Web etc. Today in business, stock market clustering can support marketers to determine interest's vendors and customers based on their record of purchasing patterns and distinguish groups of their customers who are interested in goods. In medical science, cluster analysis can be used to derive new plant like testing new hybrid species or estimating the conditions in which they grow well and observing soil and water quality. Animal taxonomies, classify their genetic factors with similar functionality. In geology, expert can use clustering technique to recognize areas of similar interests, lands, similar, houses and infrastructure in a city or in country etc. Data clustering [4] technique is also useful in organizing data on the World Wide Web for interested knowledge or data. Clustering is an unsupervised classification technique that aims at generating collections of items, or clusters in that way that object with similar properties are grouped together in same cluster and objects with different cluster are quite distant. Mining arbitrary shaped clusters in large data sets is an open challenge in data mining. The number of solutions of these problems has been proposed with high time complexity. Computational cost can be saved by using some algorithms by shrinking a data set size to a smaller amount data examples and user defined threshold ratios can affect the clustering performances. The CLASP(clustering algorithm for arbitrary shaped clusters) algorithm is an effective and efficient algorithm for mining arbitrary shaped clusters which automatically shrinks the size of a data set while effectively preserving the shape information of clusters in the data set with representative data examples. After this it changes the locations of these data examples to improve their intrinsic relationship and make the cluster structures more clear and distinct for clustering. At last, it does agglomerative clustering to find the cluster structures with the help of pk metric called mutual k-nearest neighbor-based similarity metric. In this work, the enhancement of the asymmetric clustering algorithms to increase the quality of cluster and improve the efficiency of algorithms.

III. TECHNIQUE: BACKPROPAGATION

The backpropagation algorithm accomplishes learning on a multilayer feed-forward neural network. This algorithm iteratively learns a set of weights for prediction of the class label of tuples. This multilayer feed-forward neural network consist three layers named as input layer, one or more hidden layers, and an output layer [5].

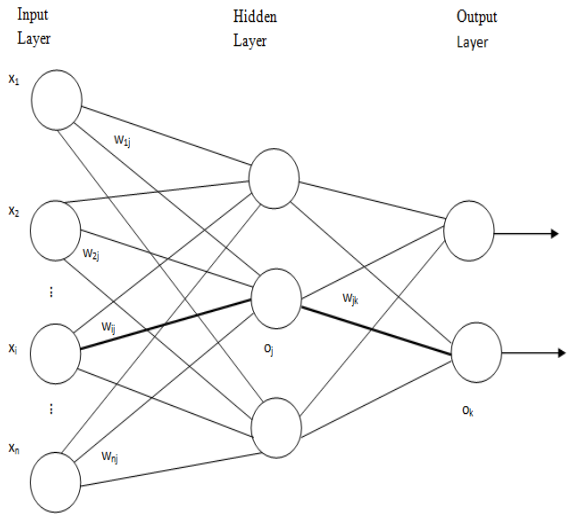

Figure 3: Layers of neural network [5]

A. *Classification of Backpropagation:*

Backpropagation is a neural network learning algorithm [6]. A neural network consist a set of input/output units which are connected to each other where each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as that correct class label of the input tuples is predicted. As Neural network learning has connection between units it is also called as connectionist learning. Neural networks involve long training times so that it is more fit for applications where this is feasible. A data set of training tuples in Backpropagation learns by iteratively processing and also compares the network's prediction for each tuple with the actual known target value. This target value may be the known class label of the training tuple or a continuous value. The weights are modified for each training tuple to minimize the mean squared error between the actual target value and the network's prediction. These modifications are made in the backpropagation algorithm which is directed from the output layer, and then through each hidden layer down to the first hidden layer that is why it is named as backpropagation algorithm [5].

B. *Process: Initialize the weights:*

The weights in the network are initialized with the small random numbers ranging from-1.0 to 1.0, or -0.5 to 0.5. Each unit has a bias associated with it and initialized with the small random numbers similarly as that of weights. Each training tuple, ***X***, is processed by following steps [6].

a) Propagate the inputs forward:

First, the input layer of the network is fed by the training tuple and the inputs pass through the input units, unchanged. That is, as there is an input unit *j*, its output is equal to *Oj*. Next, the net input and output of each unit are computed in the hidden and output layers. In the hidden or output layers the net input to a unit is computed as a linear combination of its inputs. Each such unit has number of inputs to it and the outputs of the units connected to it in the previous layer. Each connection has a weight. To compute the net input to the unit, each input connected to the unit is multiplied by its corresponding weight, and then this is summed [6].

$$I_j = \sum_i wijOi + \theta j$$

Where w_{ij} is the weight of the connection from unit in the previous layer to unit j
O_i is the output of unit i from the previous layer
Θ_j is the bias of the unit & it acts as a threshold in that it serves to vary the activity of the unit.

Each unit in the hidden and output layers takes its net input and then applies an activation function to it.

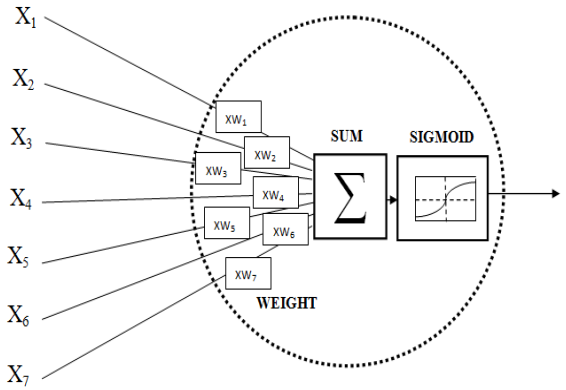

Figure 4: Internal view of Hidden Layer

b) Backpropagate the error:

To reflect the error of the network's prediction the error is propagated backward by updating the weights and biases [6]. For a unit *j* in the output layer, the error Err_j is computed by

$$Err_j = O_j(1-O_j)(T_j-O_j)$$

Where O_j is the actual output of unit j
T_j is the known target value of the given training tuple.

The error of a hidden layer unit *j* is

$$Err_j = O_j(1-O_j)\, Err_k w_{jk}$$

Where wjk is the weight of the connection from unit j to a unit k in the next higher layer.
Errk is the error of unit k.

Weights are updated by the following equations, where D*wij* is the change in weight *wij*.

$$\Delta wij = (l) Err_j O_i$$
$$wij = wij + \Delta wij$$

Biases are updated by the following equations below

$$\Delta \boldsymbol{\theta} j = (l) Errj$$
$$\boldsymbol{\theta}_j = \boldsymbol{\theta}_j + \Delta \boldsymbol{\theta}_j$$

IV. CONCLUSION

To extract useful or interested information from large set of databases data mining techniques are used. KDD (knowledge discovery from databases) is data mining method to extract information from data warehouses. Association rule is method to place the frequent item sets together to do analysis like in basket analysis, retail stores and stock market etc. Asymmetric clustering is unsupervised technique of data mining. Clustering is technique in which huge datasets are divided into small datasets in this way that objects and items with having similar properties into one group and objects having dissimilar properties into another. There are number of algorithms that work well with simple datasets in the term of accuracy and performance but, when these algorithms has to work with mixed and tightly coupled different data sets their performance in the term of accuracy is decreased. Neural networks can be combined with these existing asymmetric algorithms to improve and accuracy and reduce escape time.

References

[1] R. Buyyr, J. broberg, A. Goscinski, *"Cloud Computing Principlesand Paradigms"*, John Wiley & Sons,Inc publications , pp. 63-65, 2011.

[2] H.J. Jiawei, M. Kamber, *"Data Mining: Concepts and Techniques (3rd ed.)"*. Morgan Kaufmann publication, San Francisco, pp. 26-39, 2012.

[3] A. Nagpal,. A. Jatain, D. Gaur, *"Review based on data clustering algorithms"*, in: Proceedings of 2013 IEEE International Conference on Information and Communication Technologies (ICT 2013), pp. 171–176, 2013.

[4] W. Yu, G. Qiang, L. Xiao-Li, "*A kernel aggregate clustering approach for mixed data set and its application in customer segmentation"*, in: International Conference on Management Science and Engineering ICMSE, , pp. 121–124, 2006.

[5] Dr.R. kishore, T.Kaur, *"Backpropagation Algorithm: An Artificial Neural Network Approach for Pattern Recognition"* in:

International Journel of Scientific and Engineering Research, Vol. 3(6), pp. 1-4, 2012.

[6] K. Vora, S. Yagnik, "A Survey on Backpropagation Algorithms for Feedforward Neural Networks" in: International Journal of Engineering Development and Research(IJEDR), pp. 191-197, 2010.

[7] Z. Nafar, A. Golshani, "*Data mining methods for protein–protein interactions"*, in: Canadian Conference on Electrical and Computer Engineering, CCECE, pp. 991–994, 2006.

[8] A.M. Bakr, N.A. Yousri, M.A. Ismail, "*Efficient incremental phrase-based document clustering"*, in: International Conference on Pattern Recognition ICPR, Tsukuba (Japan), pp. 517–520, 2012.

[9] S. Nithyakalyani, S.S. Kumar, "*Data aggregation in wireless sensor network using node clustering algorithms a comparative study"*, in: Proceedings of 2013 IEEE Conference on Information and Communication Technologies (ICT), pp. 508–513, 2013.

[10] Bahm, K. Haegler, N.S Maller, C. Plant, CoCo: "*coding cost for parameter-free outlier detection"*, in: The 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 149–158, June 2009.

[11] H.P. Kriegel, M. Pfeifle, "*Effective and efficient distributed model-based clustering"*, in: Proceedings of the 5th International Conference on Data Mining (ICDM'05), pp. 285-265, 2005.

[12] K.M. Hammouda, M.S. Kamel, "*Efficient phrase-based document indexing for web document clustering"*, in: IEEE Transactions on Knowledge and Data Engineering, vol. 16(10), pp. 1279–1296, 2004.

[13] Z. Zhang, J. Zhang, H. Xue, "*Improved K-means clustering algorithm"*, in: Congress on Image and Signal Processing CISP, vol. 5, pp. 169–172, 2008.

[14] L. Li, J. You, G. Han, H. Chen, "*Double partition around medoids based cluster ensemble"*, in: International Conference on Machine Learning and Cybernetics, vol. 4, pp. 1390– 1394, 2012.

[15] D.H. Zhou, L.Y. Bin, *"An improved BIRCH clustering algorithm and application in thermal power"*, in: International Conference on Web Information Systems and Mining (WISM), vol. 1, pp. 53–56, 2010.

[16] R.T. Ng, J. Han, "*CLARANS: a method for clustering objects for spatial data mining"*, IEEE Transactions on Knowledge and Data Engineering, Vol 14 (5), pp. 1003–1016, 2002.

[17] M. Ester, H. Kriegel, J. Sander, X. Xu, "*A density-based algorithm for discovering clusters in large spatial databases with noise"*, in: Proceeding 2nd International Conference on Knowledge Discovery and Data Mining, , pp. 226–231, 1996.

[18] H. Shah, K. Napanda, L. D'mello, *" Density Based Algorithms"*, in: IJCSE International Journal of Computer Sciences nad Engineering, Vol. 3(11), pp. 54-57, 2015.

[19] Z. Wang, Y. Hao, Z. Xiong, F. Sun, "*SNN clustering kernel technique for content-based scene matching"*, in: 7th IEEE International Conference on Cybernetic Intelligent Systems, pp. 1–6, 2008.

[20] E. Achtert, C. Bohm, H-P. Kriegel, P. Kroger, I. Maller- Gorman, A. Zimek, "*Detection and visualization of subspace cluster hierarchies"*, in: Advances in Databases: Concepts, Systems and Applications, Lecture Notes in Computer Science, pp. 152–163, 2007.

[21] D.H. Widyantoro, T.R. Ioerger, J. Yen, "*An incremental approach to building a cluster hierarchy"*, ICDM Proceedings IEEE International Conference on Data Mining, pp. 705–708, 2002.

[22] S.A.L. Mary, K.R.S. Kumar, "*A density based dynamic data clustering algorithm based on incremental dataset"*, J. Computer Sci. Vol 8 (5), pp. 656–664, 2012.

[23] K.M. Hammouda, M.S. Kamel, "*Incremental document clustering using cluster similarity histograms"*, in: IEEE/WIC Proceedings International Conference on Web Intelligence, pp. 597–601, 2003.

[24] S. Young, I. Arel, "*A fast and stable incremental clustering algorithm"*, in: Seventh International Conference on Information Technology: New Generations (ITNG), pp. 204–209, 2010.

Authors Profile

Er. Paramvir Kaur Dhillon received the Bachelor degree in computer science from the Punjab Technical University, in 2015. She is presently a student of Master of Technology (Computer science) in Sant Baba Bhag Singh University. Her current research interests include data mining. She has her publication in 3rd DAV National Congress and presented a paper entitled "Expert System" in National Conference on Recent Trends in Computer Technology (RTCT-2014).

Er. Amandeep Singh Walia pursed his Bachelor degree in computer science, Master in computer science from Punjab Technical University. He is currently working as Assistant Professor in Sant Baba Bhag Singh University, Punjab. His main research work focuses on Network Routing. He has 1 years of teaching experience.

