Journal of Advances and Scholarly Researches in Allied Education
Vol. 21, Issue No. 1, January-2024, ISSN 2230-7540

Feature Selection and Dimensionality Reduction for Large Data set Clustering using Self-Organizing Maps

Gyan Chand Sharma[1]*, Dr. Mohit Gupta[2]

[1] Research Scholar, University of Technology

[2] Associate Professor, Department of Computer Science, University of Technology

Abstract - Data mining is a form of analyzing the data which relates the newline techniques from fields like statistics machine learning databases artificial newline intelligence etc Clustering is one of the most important data mining newline techniques in which cluster of objects are grouped based on their similarity newline "Clustering is the process of accumulating the data records into considerable newline subclasses clusters in a way which enhances the relationship within clusters newline and reduces the similarity among two different clusters. The activity occurs every new academic year and schools with plenty of new students registered may feel a bit overwhelmed with this grouping assignment. A decision support system which can automatically perform grouping on a list of students may be able to help the school's staffs with this repetitive task. A self-organizing map (SOM) is an example of unsupervised learning algorithm using an artificial neural network structure to produce a low dimensional representation from a given input.

***Keywords -** Data Sets, Self-Organization, Maps*

----------X----------

INTRODUCTION

The problem determines which data are utilised as well as what constitutes an effective solution. The results may be applied to fresh data thanks to modeling, which makes this feasible. On the other side, data modeling can cause issues if it is done without a solid grasp of the data and without thorough preparation of the data. In conclusion, the entire mining process is pointless if the newly acquired knowledge is not put to good use. The objective of the survey is to obtain a better understanding of the data, including both the possibilities and the issues associated with them, as well as to evaluate the adequacy of the data and choose the appropriate preprocessing and modeling methods. In most cases, it is necessary to take into consideration a number of distinct data sets as well as preprocessing methods. Because of this, having effective visualizations and summaries is really necessary. Clusters are crucial characterizations of data, which is why we will be focusing on them throughout this work. Due to the notable visualization qualities it possesses, the self-organizing map, often known as SOM, is particularly well suited for data surveying. It does a topology-preserving projection of the prototypes from the high-dimensional input space onto a low-dimensional grid, after first generating a collection of prototype vectors that represent the data set and then carrying out that projection. This ordered grid may be utilised as a practical visualisation surface for the purpose of displaying various characteristics of the SOM (and, by extension, of the data), such as the cluster structure.

Nevertheless, one can only utilise the visualisations to get qualitative information through their utilisation. In order to construct summaries, which are quantitative representations of the attributes of the data, it is necessary to choose interesting groupings of map units from the SOM. The whole map is the most evident example of such a group. In spite of the fact that its features are undeniably fascinating, it is possible to come up with summaries that are even more helpful if the SOM (and, by extension, the data) is broken up into two or more distinct sections and

each of them is investigated independently. There is also the possibility of looking at each unit of the map separately; but, if the map is very vast, doing so can provide an excessive number of summaries. Therefore, in order to properly utilise the information that is offered by the SOM, procedures are necessary that yield good candidates for map unit clusters or groups. It should be highlighted that the aim here is not to identify a clustering that is ideal for the data; rather, the goal is to acquire a decent insight into the cluster structure of the data for the purposes of data mining. As a result, the approach for clustering should be as quick as it is resilient and as aesthetically effective as possible. The clustering is accomplished by the use of a two-level strategy, in which the SOM is initially used to cluster the data set, and then the SOM itself is clustered. The computing burden is significantly reduced as a result of this strategy, which makes it feasible to cluster big data sets and to evaluate a variety of preprocessing approaches within a constrained amount of time. This is the most significant advantage of the procedure. Naturally, the method can only be considered reliable if the clusters discovered by employing the SOM are comparable to those found in the primary data. A comparison is made in the experiments between the outcomes of directly clustering the data and clustering the prototype vectors of the SOM, and it is discovered that the correlation is satisfactory.

An example of an unsupervised learning algorithm is a self-organizing map, often known as a SOM. This type of method makes use of an artificial neural network structure to generate a low-dimensional representation of a given input. This model takes use of the Kohonen Map or network which is frequently utilised for dimensionality reduction in order to make it simpler to view data. Even though SOM is a form of artificial neural network (ANN), its operation is quite distinct from that of a conventional ANN. SOM employs competitive learning, in contrast to ANN's use of error-correction learning through the use of gradient descent. In SOM, each node "competes" against other nodes to be the winning node, which would then obtain an updated weight. ANN applies error-correction learning. Applying a neighborhood function is another strategy that the SOM employs in its attempt to preserve the topological structure of the input space. The technique of artificial intelligence has been employed extensively in recent years to find solutions to issues that arise in everyday life. Learning algorithms, both supervised and unsupervised, have been created in order to handle the traditional issues of classification, grouping, and association. Eggs have been categorised in with the assistance of a Support Vector Machine. In [a Bayesian Network Model is utilised for the hepatitis diagnosis, and in Relief feature selection is integrated with the Bayesian Network Model.

An artificial neural network model was utilised in and to forecast a student's academic progress based on early semester grade inputs. This model was merged in with linear regression and support vector regression to further improve prediction accuracy. An application of SOM has been applied in unsupervised learning as a means of clustering students for the purpose of a scholarship granting system. A further use of the SOM algorithm in education is in the monitoring of e-learning activities and the visualisation of students' cognitive structural models. The use of SOM to cluster students has been examined in and in is used to study students' interest on the Maths topic. and all deal with the investigation of students' interest on the Maths subject. Using a SOM learning algorithm, the purpose of this study is to categorise pupils who have just entered in high school based on the academic grades they earned at the junior high schools they attended previously. The grades that were utilised came from their rapport books as well as the results of the national test that they had taken in their prior schooling.

OBJECTIVES

1. To improve the mixed data clustering accuracy using techniques of SOM and variants.

2. To improve training speed of SOM by introducing new fast learning model.

DATA MINING

Huge volumes of data are gathered and stored in today's contemporary society. From this data, information that is helpful and necessary may be retrieved and processed as needed. "Data mining" refers to the process of collecting relevant information from massive data sets and is a common term used in the industry. It is one of the most important steps in the datasets and consists of applying data investigation and discovery algorithms that are constrained by adequate levels of computational efficiency (Fayyad et al 1996). Additionally, it generates a specific enumeration of patterns that are found over the data. The process that is responsible for the actual finding of knowledge is known as data mining. To drive home the point that data mining techniques need to be able to handle enormous amounts of data, the necessary patterns have to be discovered within the constraints of a sufficient level of computing efficiency. E-commerce, bioinformatics, information retrieval, internet search engines, and other sectors

are among the most common and important applications for data mining.

CLUSTERING

In this real world, data are clustered, and pattern extractions are started by the clusters, which serve as a step ladder for data analysis. The process of clustering organises a collection of tangible or mental things into groups consisting of objects with similar characteristics. It is possible to do extra processing on a cluster of data items by treating the cluster as a single assembly. This simplifies the management of large volumes of different data. On the other hand, pattern removal provides metaphors for categorising the material, provided that a brief and punchy summary is provided. This is especially true with regard to class narrative. The information that is being clustered is often organised into sets in such a way that the similarity between members of the same cluster is minimised while the similarity between other clusters is maximised. The clustering approach has seen extensive use in a wide variety of sectors, including information mining, pattern recognition, client segmentation, investigation, and trend analysis, to name just a few of these applications. It has been determined that data clustering is the major data mining approach for the purpose of information discovery (Nimrat & Rajneet 2013). The work of Halkidi et al. contains a significant number of different clustering techniques (2001 & 2002). In general, significant clustering systems are able to be included into hierarchical or split categories. A given collection of data patterns can be hierarchically disintegrated through the use of a method that is hierarchical. A partition will travel toward the patterns and make other partitions of them, with each partition identifying a different cluster.

NEED FOR CLUSTERING

The purpose of using clustering is to determine the natural assemblage that exists inside a collection of unlabeled data. Despite the fact that the process of decision making is an outstanding example of clustering. It is possible to demonstrate that there is no one principle that is superior to all others and that would be capable of independently choosing the ultimate goal of the clustering. As a consequence of this, it is the responsibility of the customer to offer this concept in such a way that the outcome of the clustering will be appropriate to meet their requirements.

RELATIONSHIP BETWEEN CLUSTERING AND CLASSIFICATION

In this part, we will talk about the relationship that exists between clustering and categorization. Clustering is the process of organising the data records into major subclasses, often known as clusters, in a way that strengthens the connections between members of the same cluster while weakening the similarities between members of other clusters. Unsupervised data learning, often known as machine learning, and segmentation are both alternative names for clustering. In many cases, all that is required to get insight into the distribution of the data included inside a dataset is a set of clusters. The purpose of clustering is to provide a comprehensive perspective of a given dataset. The preparation of data for use in other data mining methods is an additional significant use of clustering algorithms.

The process of learning a function that maps data items to a subset of a given class set is what we mean when we talk about classification. As a consequence of this, a classifier is educated using a labelled collection of training items, one for each distinct class. The categorization process aims to accomplish both of these goals. Finding a more effective generic mapping that provides a more accurate prediction of the class of unidentified data items is one of the goals of this project. A simple function is used as the classifier for this purpose. In order for the classifier to accomplish this goal, it is necessary for it to determine the characteristics of the specific training instances that are typical for the whole class, as well as the characteristics that are definite for individual objects in the training set.

An additional goal of classification is to uncover a simplified and understandable model of each class that can be applied to the whole set of classes. A class model has to provide an explanation as to why specific objects belong to a particular class as well as what makes the members of a particular class unique from one another. Because a model's applicability increases in direct proportion to the degree to which it is simplified, the class model ought to be simplified to the greatest extent feasible. In addition, class models that are short and straightforward have less information that may be considered distracting, making them easier to comprehend.There is a close relationship between clustering and classification. Clustering is the process of locating a group of classes within a given dataset, whereas classification is an attempt to understand the characteristics that differentiate a given set of classes. It is not necessary to establish a set of specimen objects in order to do clustering, which is a big advantage of this method. As a consequence of this, clustering can be utilised in applications in which there is either no prior information or some prior knowledge regarding the groups or classes included inside a database. On the other hand, the efficiency of a particular clustering method is frequently open to individual interpretation and is heavily dependent on the choice of an adequate similarity metric. The use of categorization is more appropriate in situations when it has been determined in advance that a certain group of classes already exists for the application in question. In these circumstances, the assembly of a feature space in which the predefined classes are arranged into delimited clusters is often far more

complicated than the process of delivering immediate objects for each and every class.

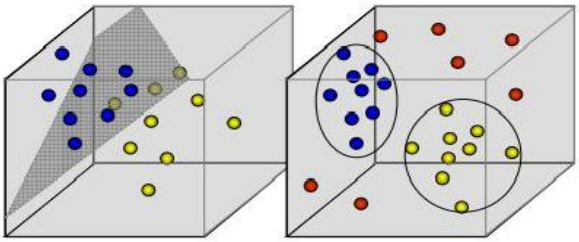

Figure 1: Classification separates the data space (left) and clustering groups data objects (right)

In addition, the performance of a classifier may be easily evaluated by counting the number of correct predictions for each category that it generates. It has been discovered that data mining tasks, such as clustering and classification, are applicable in a wide variety of contexts. On the left side of Figure 1 is a presentation of class separation achieved by a classifier, and on the right side is a show of the clustering of two clusters inside a noisy dataset.

CONCLUSION

An effective strategy for data analysis involves first clustering big datasets and then visually representing the clusters with the use of self-organizing maps (SOM). SOMs, which take their cue from neural networks, offer a powerful method for unearthing hidden patterns and structures buried behind enormous volumes of data. SOMs may organise data into clusters based on similarity. Because of this, they are particularly effective for exploratory data analysis and dimensionality reduction. This is accomplished by iteratively altering the weights of the neurons in the SOM. Complex information can be better comprehended and interpreted by humans because to the simplification afforded by the visual representation of clusters on a two-dimensional map. Each neuron that makes up the SOM represents a different cluster, and the proximity of neurons on the map shows the degree to which different clusters are similar to one another. This visualisation not only helps in understanding the underlying data distribution, but it also helps in recognising outliers and abnormalities, which can be extremely important in a variety of applications, such as customer segmentation in marketing and anomaly detection in cybersecurity.

REFERENCES

1. Arumugadevi,S& Seenivasagam,V 2014,' Color image segmentation using feedforward neural networks with FCM', International Journal of Automation and Computing, Springer.

2. Arunachalam,T 2017,' An efficient color image segmentation using edge detection and thresholding methods', International Journal of Advanced Networking & Applications, vol.8,no.5,pp.19-23.

3. Avraham,T &Lindenbaum, M 2010,' Esaliency (extended saliency): meaningful attention using stochastic image modeling', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 4, pp. 693–708.

4. Baldevbhai, PJ & Anand, RS 2012,' Color image segmentation for medical images using L*a*b* color space', IOSR Journal of Electronics and Communication Engineering, vol. 1, no. 2, pp. 24–45.

5. Bhattacharya,P, Biswas,A & Maity,SP 2014, 'Wavelets-based clustering techniques for efficient color image segmentation. In: Kumar Kundu M., Mohapatra D., Konar A., Chakraborty A. (eds) Advanced Computing, Networking and Informatics-vol.1. Smart Innovation, Systems and Technologies, Springer, Cham

6. Celik, T & Tjahjadi, T 2010, 'Unsupervised color image segmentation using dualtree complex wavelet transform', Computer Vision and Image understanding, vol.114, pp.813-826.

7. Chaobing Huang, Liu,Q & Li,X 2010,'Color image segmentation by seeded region growing and region merging ', Proceedings of the IEEE Seventh International Conference on Fuzzy Systems and Knowledge Discovery.

8. Chebbout,S & Merouani,HF 2012,' Comparative study of clustering based colour image segmentation techniques', Proceedings of the Eighth International Conference on Signal Image Technology and Internet Based Systems, IEEE, pp.839-843.

9. Chi,D 2011,' Self-organizing map-based color image segmentation with k-means clustering and saliency map', ISRN Signal Processing, vol.2011, doi:10.5402/2011/393891

10. Chitradevi,B & Srimathi, P 2014, 'An overview on image processing techniques', International Journal of Innovative Research in Computer and Communication Engineering, vol.2, no.11,pp.6466-6472.

11. Chong, RM & Tanaka,T 2010,'Motion blur identification using maxima locations for blind color image restoration', Journal of Convergence, vol.1, no.1,pp.49–56.

12. Christ, MCJ & Parvathi, MS 2011,' Fuzzy C-Means Algorithm for Medical Image

Segmentation', Proceedings of the Third International Conference on Electronics Computer Technology, IEEE, pp.33-36.

13. Dehariya, VK, Shrivastava,SK & Jain, RC 2010,'Clustering of image data set using K-Means and fuzzy K- means algorithms', Proceedings of the International Conference on CICN, pp. 386- 391.

14. Deng,YN & Manjunath,BS 2001,'Unsupervised segmentation of color-texture regions in images and video', IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.23, no.8, pp.800 – 810.

15. Destrempes,F, Angers, JF & Mignotte, M 2006,'Fusion of hidden Markov random field models and its Bayesian estimation', IEEE Transactions on Image Processing, vol. 15, no. 10,pp. 2920–2935.

16. Dharampal & Mutneja,V 2015,' Methods of image edge detection: a review', Journal of Electrical & Electronic Systems, vol.4, no.2, pp.1-5.

17. Dilpreet Kaur & Yadwinder Kaur 2014,'Various image segmentation techniques: A review', International Journal of Computer Science and Mobile Computing, vol.3, no.5, pp. 809-814.

18. Ding, Z, Sun,J &Zhang,Y 2013,'FCM Image Segmentation Algorithm Based on Color Space and Spatial Information', International Journal of Computer and Communication Engineering, vol. 2, no.1,pp.48-51.

19. Fan, JC Han,M& Wang,J 2009,' Single point iterative weighted fuzzy C-means clustering algorithm for remote sensing image segmentation," Pattern Recognition, vol. 42, no. 11, pp. 2527-2540.

20. Farooque, MY & Raeen,S 2016',Self-organizing map based improved color image segmentation', International Journal of Advanced Research in Computer Science and Software Engineering,vol.6,no.3,pp.456-462.

Corresponding Author

Gyan Chand Sharma*

Research Scholar, University of Technology