ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Data Clustering and Visualization based various Machine learning techniques

Khushboo Bansal
Computer Engineering
Yadvindra College of Engineering and Technology
Talwandi Sabo
Kbansal067@gmail.com

Er.Meenakshi Bansal
(Assistant Professor)
Computer Engineering
Yadvindra College of Engineering and Technology
Talwandi Sabo
ermeenu10@gmail.com

Abstract—*Clustering is the process of grouping objects together in such a way that the objects belonging to the same group are similar and those belonging to different groups are dissimilar. Clustering Usage data is one of the important tasks of data Usage Mining, which helps to find data user clusters and data page clusters. Data user clusters establish groups of users exhibiting similar browsing patterns and data page clusters provide useful knowledge. Recent studies have supported the usage of machine learning algorithm. So, this paper has reviewed commonly used methods in the field of clustering.*

Keywords—*Clustering, Data Visualization, Support Vector Machines, Fuzzy Logic.*

1. INTRODUCTION

With the rapid growth of World Wide Web the study of modeling the user's navigational behavior in a Web site has become very important. With the large number of companies using Internet to distribute and collect information, Knowledge discovery on the Web has become an important research area [1, 2]. The purpose of Web

Usage mining is to apply statistical and data mining techniques to the preprocessed Web log data, in order to discover useful Web Usage patterns. More advanced data mining methods and algorithms, such as association rules, sequential pattern mining, classification and clustering are adapted appropriately to find suitable patterns from Web Usage data [3, 4, 5].

In addition to these methods and algorithms, the Artificial Neural Network, Genetic Algorithm and fuzzy classification methods are also used to find valuable information from the Web Usage data. Web Usage mining contains three main tasks namely:

- Data preprocessing,
- Cluster discovery and
- Cluster analysis [6, 7]

Data preprocessing consists of data cleaning, data transformation, and data reduction. Cluster discovery deals with formation of groups of users exhibiting similar browsing patterns and obtaining groups of pages that are accessed together [8,9]. Cluster analysis filters out uninteresting patterns from the user clusters and page clusters found in the Cluster discovery phase. Clustering is a data mining technique that groups together a set of items having similar characteristics [10].

2. CLUSTERING

Data mining also known as knowledge-discovery in databases (KDD) is process of extracting potentially useful information from raw data. A software engine can scan large amounts of data and automatically report interesting patterns without requiring human intervention. Other knowledge discovery technologies are Statistical Analysis [11], OLAP, Data Visualization, and Ad hoc queries. Unlike these technologies, data mining does not require a human to ask specific questions.

Here is the list of areas where data mining is widely used [12]:

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis

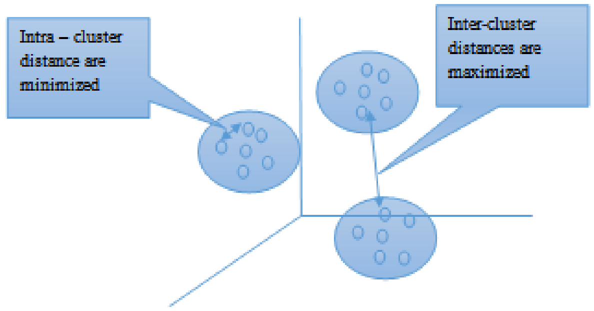

Intra – cluster distance are minimized
Inter-cluster distances are maximized

Figure 1. Clustering Principle

Clustering can be said as identification of similar classes of objects [13]. Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters.

Maximizing intra-class similarity & minimizing inter-class similarity as shown in figure 1.

3. Data Visualization

A visual can communicate more information than a table in a much smaller space. This trait of visuals makes them more

978-93-85670-72-5 © 2016 (RTCSIT)

effective than tables for presenting data. For example, notice the table below, and try to spot the month with the highest sales [14].

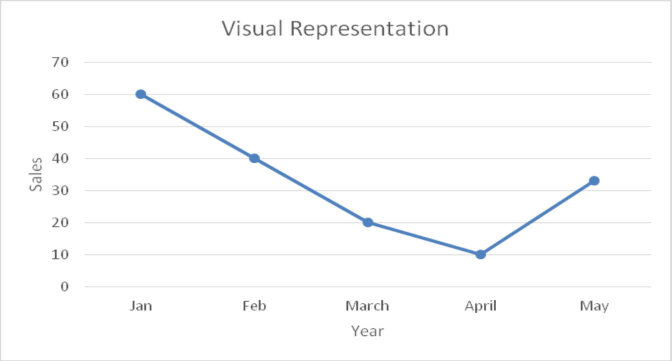


Figure 2. Sales Graph as Visual RepresentationTABLE 1 Sales table

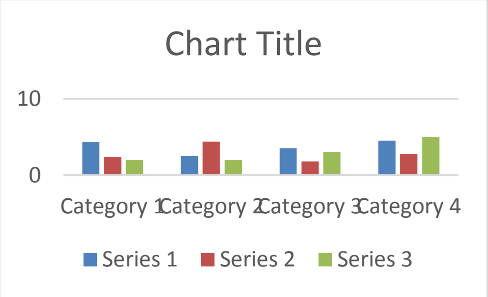
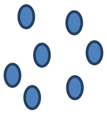
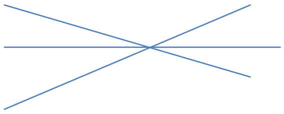
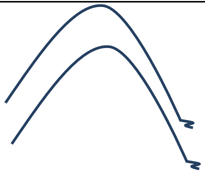
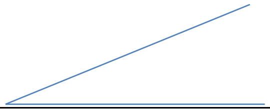
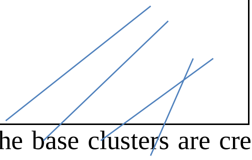
Jan	Feb	March	April	May
60	**40**	**20**	**10**	**33**

Two main goals of Visualization:

- Explanatory
- Exploratory

Exploratory visuals offer the viewer many dimensions to a data set, or compares multiple data sets with each other. They invite the viewer to explore the visual, ask questions along the way, and find answers to those questions.
Examples of Data Visualization are shown below [15]:

TABLE 2 Examples of Data Visualizations

Pattern	Example	Pattern	Example
High, Low		Clusters	
Going up and down		Wide and narrow	
Steep gradual		Intersecting or not	

4. RELATED WORK

There is difficulty in the analysis of categorical data is categorized by the fact that there is no inherent similarity between attribute values of categorical dataset. The clustering of categorical dataset is fully based on the available dataset. To cluster categorical dataset a link based cluster ensemble approach is used, in which initially the base clusters are created of the available dataset as input by applying the algorithm. From these base clusters a cluster ensemble is created. Existing clustering algorithms can be classified into two main categories: Hierarchical and Partitioning algorithm.
Arthur et al.(2007) proposed K-Means++ which is an extension of K-Means algorithm. K-Means++ find the center using probability measure which gives an optimal seed value for the existing K-Means algorithm. The author shows that K-Means++ outperforms K-Means in both speed and accuracy.

Zengyou He et al.(2008) introduced K-ANMI works similar way as K-means algorithm. It takes 'k' as input and changes class label iteratively for each object to improve the objective function value. Cluster is evaluated in each step by using mutual function based criterion-ANMI.

Zengyou He et al.(2006) proposed NabSqueezer algorithm, an improved Squeezer algorithm. NabSqueezer algorithm gives more weight to uncommon attribute value matches for finding similarity in similarity computation of Squeezer algorithm. In this algorithm weight of each attribute is precalculated using More Similar Attribute Value Set (MSFVS) method.

Li Taoying et al.(2009) proposed Fuzzy Clustering Ensemble Algorithm for Partitioning Categorical Data makes use of relationship degree of attributes for pruning a part of attributes. Descartes subset is used for finding the cluster membership. Both relationship degree and Descartes subsets are used for establishing the relationship between objet as well as minimizes the objective function.

Z.Huang and M.K. Ng (1999) presented Fuzzy K-Modes algorithm makes use of a simple matching dissimilarity measure (Generalized Hamming distance) and Mode values for clustering the categorical objects. The algorithm uses update method inorder to minimize cost function Fc (X,Z) and update z at each iteration.

Wang Jiacai and GuRuijun (2010)developed Extended Fuzzy KMeans algorithm uses expanded form of cluster centroid vector representation to keep the clustering information and update the method in the same way as in fuzzy k-means.

Desai et al. (2011) use similarity which are neighborhoodbased or incorporate the similarity computation into the learning algorithm. These measures compute the neighborhood of a data point but not suitable for calculating similarity between a pair of data instances X and Y.

Sayal et al.(2011)proposed a concept called Context Based Similarity Measure which is achieved in relational database through Functional Dependency. The Context Based similarity finds the similarity between components by checking the contexts in which they appear.

Author	Year	Algorithm	Advantage	Accuracy
Arthur et al.	2007	K-Means++	The author shows that K-Means++ outperforms K-Means in both speed and accuracy.	Good
Zengyou Ye et al.	2008	K-ANMI	Cluster is evaluated in each step using mutual function based criterion-ANMI.	Moderate
Zengyou Ye et al.	2006	NabSqueezer algorithm	Weight of each attribute is precalculated.	Medium
Li Taoying et al.	2009	Fuzzy Clustering	Minimizes the objective function.	Good
Z.Huang and M.K. Ng	1999	Fuzzy K-Modes	Keep the clustering information.	Good
Wang Jiacai and GuRuijun	2010	Extended Fuzzy K-Means	High similarity.	Medium
Desai et al.	2011	Neighborhoodbase	compute the neighborhood of a data point	Good
Sayal et al.	2011	Context Based Similarity Measure	Finds the similarity between components by checking the contexts in which they appear.	Optimum

5. VARIOUS TECHNIQUES USED IN THIS CONTEXT

A. *Cosine Similarity method*

It is the similarity measure between two vectors (or two documents on the Vector Space). It calculates the cosine angle between two documents. This measure is the evaluation for the measurement of orientation. This technique calculates the angle between documents not the magnitude of the documents. Cosine similarity can be represented as below:

$$\vec{a}.\vec{b} = \| \vec{a} \| \;\; \| \vec{b} \| \cos Q$$

B. *K-neighboring Method*

K-nearest neighbour is a classification algorithm the combines the k nearest points. It is supervised classification algorithm. It is very simple and relatively high convergence speed algorithm. However, in some applications, it may fail to produce adequate results, whilst in others its operation may render impractical. Yet, the fact that it has only one parameter, the number of neighbours used (k), makes it easy to fine-tune to a variety of situations. Its main process consists of the following steps: given a set of N points (training set), whose class labels are known, classify a set of n points (testing set) into the same set of classes by examining the k closest points around each point of the testing set and by applying the majority vote scheme.

C. *Fuzzy Logic*

Fuzzy sets were introduced by Zadeh. It was designed basically to show the uncertainty and vagueness. Fuzzy logic provides the human reasoning capabilities. The theory of fuzzy logic provides the strength to obtain the uncertainties associated with human process. The need of fuzzy logic arises in the time to describe the principle of and problem of uncertainty.

Characteristics of fuzzy logic:

- Exact reasoning is seen as the limiting case of the approximate reasoning. [16]
- Everything is the matter of a degree
- Knowledge is based on collection of variables.
- Inference is seen as the process of the elastic constraints.
- It is of imprecise data.
- It model nonlinear functions of arbitrary complexity.
- It is easy to understand.
- It can be blended with traditional methods
- Fuzzy logic is based on natural language.

D. *Support Vector Machine(SVM)*

Support vector machines (SVMs) is a binary classification algorithm developed by Vapnik. The main features of SVM are shown below, due to which its applications are quite important:

- Robust to large number of variables.
- Can be applied to & it can learn complex and simple learning models [17].
- It avoid overfitting.

Support vector machines (SVMs) have the hyperplane that classifies the various variables as shown below:

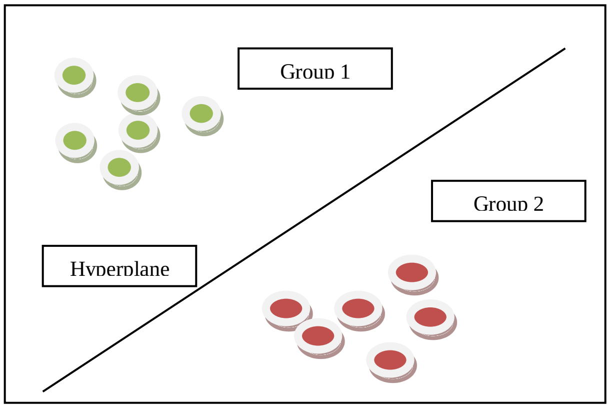


Figure 3.Support Vector Machine

6. CONCLUSION AND FUTURE SCOPE

In this paper, we dissected the data clustering and visual perception. First, we saw that all visualizations have a goal - explanatory, or exploratory. Then, in conclusion, it's obvious that we are naturally hard-wired to visualize information in a certain way. Understanding those basic principles of data visualization will help us craft outstanding visualizations, and tell compelling stories. After that basic techniques of data clustering w.r.t data visualization has been presented.

REFERENCES

[1] . Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on theWorld Wide Web", Ninth IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA, USA, 1997, Pages 558-567.

[2] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan, "Web usage mining: discovery and applications of usage patterns from Web data", ACM SIGKDD Explorations Newsletter, Volume 1, Pages 12-23, 2000.

[3] Bamshad Mobasher, Chapter: 12, "Web Usage Mining in Data Collection and Pre-Processing", ACM SIGKKD 2007 Pages 450-483.

[4] Natheer Khasawneh and Hien Chung Chan, "Active User-Based and Ontology-Based Weblog data preprocessing for Web Usage Mining", IEEE/ WIC/ACM International Conference 2006.

[5] Kobra etminani, Amin, and Noorali Rouhani, "Web usage Mining: Discovery of the user's navigational patterns using SOM", IEEE 2009.

[6] Sebastian A. Rios, and Juan D.Velasquez, “Semantic Web Usage Mining by a Concept-based approach for Off-line Web Site Enhancements”, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 2008.

[7] Murat Ali Bayir, Ismail Hakki Toroslu, Guven Fidan, and Ahmet Cosar, “Smart Miner: A New Framework for Mining Large Scale Web Usage Data”, ACM 2009.

[8] Norwati Mustapha, Manijeh Jalali , and Mehrdad Jalali, “Expectation Maximization Clustering Algorithm for User Modeling in Web Usage Mining Systems”, European Journal of Scientific ResearchISSN 1450-216X Volume 32 Number.4 (2009), Pages.467-476.

[9] Jiyang Chen, Lisheng Sun, Osmar R.Zaiane, and Ranidy Goeble, “Visualizing and Discovering WebNavigational Patterns”, Seventh International Workshop on the Web and Databases (Web DB 2004), June17-18, 2004, Paris, France.

[10] Esin Saka, and Olfa Nasraoui, “Simultaneous Clustering and Visualization of Web Usage Data using Swarm-based Intelligence”, 20th IEEE International Conference on Tools with Artificial Intelligence.

[11] Sungjune Park, Nallan C. Suresh, and Bong Keun Jeong, “Sequence based clustering for Web usage mining: A new experimental framework and ANNenhanced K-Means algorithm”, Elsevier Data and Knowledge Engineering 65 (2008) 512 – 543.

[12] Santosh K.Rangarajan, Vir V.Phoha, Kiran S.Balagani, Rastko R.Selmic and S.S. Iyengar, “Adaptive Neural Network Clustering of Web Users”, IEEE 2004 0018-9162/04.

[13] [13] Antonio S, Jose D. Martin, Emilio S, Alberto P, Rafael M and Antonio, “Web mining based on growing hierarchical Self Organizing Maps: Analysis of a real citizen Web portal”, Expert Systems with applications 34(2008)2998-2994 www.elsevier.com.

[14] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela,V. Paatero,A. and Saarela, “Self organization of a massive document collection”, IEEE Transactions on Neural Networks 11 (3)(May 2000) 574– 585.

[15] Samuel Kaski, Timo Honkela, Krista Lagus, and Teuvo Kohonen, “WEBSOM-Self organizing maps of document collections”, Neurocomputing 21(1998) 101-117 Elsevier.

[16] [Kate A. Smith, Alan Ng, “Web page clustering using a self-organizing map of user navigation atterns”,Elsevier Decision Support Systems 35 (2003) 245– 256.

[17] T. Vijaya Kumar, Dr. H. S. Guruprasad, “Clustering Web Usage Data using Concept hierarchy and Self Organizing Maps”, International Journal of Computer Applications (0975 – 8887) Volume 56– No.18, October 2012.

[18] D. Arthur and S. Vassilvitskii. 2007. K-means++: The advantages of careful seeding. In Proc. 18th Annu. ACMSIAM Symp. Discrete Algorithm.

[19] Zengyou He, Xiaofei Xu,Shenchun Deng. 2008. k-ANMI: A Mutual Induction Based Clustering Algorithm for Categorical Data. Information Fusion9 (2).

[20] Zengyou He, Xiaofei Xu,Shenchun Deng . 2006. Improving Categorical Data Clustering Algorithm by Weighting Uncommon Attribute Value Matches. ComSIS Vol.3, No.1.

[21] Taoying Li, Yan Chen.2009. Fuzzy Clustering Ensemble Algorithm for partitional Categorical Data. IEEE, International conference on Business Intelligence and Financial Engineering.

[22] Z.Haung and Michael K.Ng. 1999. A Fuzzy k-Modes Algorithm for Clustering Categorical Data. IEEE Transaction On Fuzzy systems, Vol. 7, No-4

[23] Wang Jiacai and Gu Ruijun. 2010. An Extended Fuzzy KMeans Algorithm for Clustering Categorical Valued Data. International Conference on Artificial Intelligence and Computational Intelligence.

[24] Aditya Desai, Himanshu Singh and Vikram Pudi. 2011. DISC Data-Intensive Similarity Measure for Categorical Data. Pacific-Asia Conferences on Knowledge Discovery Data Mining.

[25] Rishi Sayal and Vijay Kumar.V.2011. A novel Similarity Measure for Clustering Categorical Data Sets. International Journal of Computer Application (0975- 8887)