A Survey on Applications of Data Mining using Clustering Techniques

Neha D.
Dept. of Computer Science and Engineering
Ballari Institute of Technology & Management

B.M. Vidyavathi, PhD
Dept. of Computer Science and Engineering
Ballari Institute of Technology & Management

ABSTRACT
Data mining is a process that explores and analyses large data sets in order to discover meaningful patterns. Clustering is a main task of exploratory data analysis and data mining applications. Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). Clustering has wide applications in the field of medicine, business and marketing, World Wide Web, computer science, social science, educational data mining, climatology and many more. This paper mainly presents an overview of types of clustering techniques and some of the applications of data mining where clustering techniques can be applied. The main goal of clustering is to produce a good and high quality clusters that depends mainly on the similarity measure which has the ability to discover some or all hidden patterns and also make the analysis of data easy

Keywords
Data mining, Clustering, Clustering Techniques, Applications, K-Means Clustering

1. INTRODUCTION
The era of data has arrived. The amount of data has been increasing at a faster rate. Since the data has been increasing at a faster rate, it is of great challenge to manage the data. Data Mining is a process of discovering meaningful patterns and rules and to find the relationship among the data [1]. Data mining is a multi-step process that can be divided into six common classes of tasks: Anomaly Detection (identification of unusual data records that might require further investigation), Association Rule learning (searches for relationship between variables), Clustering (Task of grouping a set of objects that are in some way or another "similar"), Classification (generalizing known structure to apply new data), Regression (Finding a function which models the data with least error) and Summarization (Providing a more suitable representation for a data set which includes visualization and generation of report) .

Clustering is one of the most fundamental techniques in data mining. Clustering is a process of dividing the data elements into groups which are similar to each other [1]. Each group is referred to as a cluster that consists of objects that are similar to one another and dissimilar to objects of another group. It is a technique that recognizes different patterns of data [1]. A good clustering method will produce a good or a high quality cluster.

The clusters that are formed need to satisfy the following two principles:

1) Homogeneity: Elements of the same cluster are maximally close to each other.
2) Separation: Data elements in separate clusters are maximally far apart from each other.

The resulting cluster that is produced has the ability to discover hidden and meaningful patterns.

2. TYPES OF CLUSTERING TECHNIQUES

2.1 Hierarchical based clustering
Hierarchical based clustering is also known as connectivity based clustering. It is a method in which hierarchies of clusters are constructed [2]. In this type of clustering, small clusters are merged into a larger one and a large cluster are splinted into smaller clusters. The clusters are constructed by partitioning the instances into a top down or a bottom up approach which can be visualized as a tree like diagram called a "Dendogram" that records the sequence of merges or splits and also shows how the clusters are related. Once the desired numbers of clusters have been formed, the process of splitting or merging will stop. Each cluster nodes consists of child nodes and the node that belongs to the same parent are called as sibling nodes.

Hierarchical clustering is based on two types of algorithms [2]:

- Agglomerative algorithm: It is a bottom up approach. It starts by merging each object which are closer to each other or by merging a number of smaller clusters into a larger cluster or until all the objects are merged and a termination condition is met.
- Divisive algorithm: It is a top down approach. It starts by splitting a larger cluster into smaller clusters until there remain only clusters of one data object and the termination condition is met.

2.2 Partition Based Clustering
Partition based clustering is a method in which a number of objects are given and the data sets will be partitioned into a number of clusters and each cluster contains similar objects [2]. It generates a specific number of flat and dis-joint clusters and the clusters that are formed will be represented by a centroid or a cluster representative.

2.3. Density based clustering

In this type of clustering, the data objects are separated based on their connectivity, boundary or their region [2] which plays a vital role in finding non-linear shape structure based on the density. This type of clustering helps to separate low dense region (noise data) from high dense region of clusters.

2.4. Grid based clustering

Grid based clustering is a type of clustering that divides the space into a finite number of cells that are known as grids and all the operations of clustering are applied on these cells [2]. The grids are then combined together to construct a grid like format.

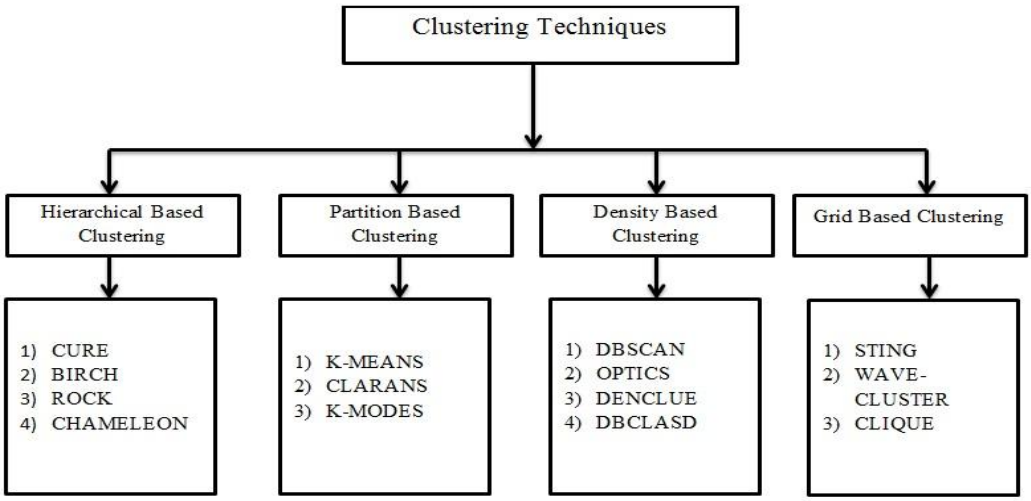

Figure 2.1: Representation of types of clustering techniques

3. APPLICATIONS OF DATA MINING USING CLUSTERING TECHNIQUES

Some of the applications of data mining where clustering techniques are implemented are [3]:

1) **Field of medicine**: In Medical imaging, cluster analysis can be used to differentiate between different types of tissues and blood (PET Scans). It is also used in the analysis of antimicrobial activity to analyse the patterns of antibiotic resistance.

2) **Business and marketing**: Partitioning the general population of consumers into market segments and to better understand the relationship between different groups of customers can be done with the help of clustering and the analysis will be used by many of the market researchers. It can also be used to group all the shopping items on the web into a set of unique products.

3) **World Wide Web**: In the study of social networks, clustering may be used to recognize communities within large groups of people. Clustering may be used to create a more relevant set of search results compared to normal search engines like Google. There are currently a number of web based clustering tools such as Clusty.

4) **Computer science**: In Image segmentation, clustering can be used to divide a digital image into distinct regions for border detection or object recognition.

5) **Social science**: In crime analysis, Cluster analysis can be used to identify areas where there are greater incidences of particular types of crime. By identifying these distinct areas or "hot spots" where a similar crime has happened over a period of time, it is possible to manage law enforcement resources more effectively.

6) **Educational data mining**: Cluster analysis is for example used to identify groups of schools or students with similar properties.

7) Climatology: To find weather regimes or preferred sea level pressure atmospheric patterns.

4. CLUSTER ANALYSIS IN SOCIAL SCIENCE (CRIME PATTERN DETECTION USING CLUSTERING)

Crime, an unlawful act is one of the most dangerous factors in the country. The rate of crime has been increasing every day that it has become impossible to find a country which has a crime free society. Crime detection is an area that is of vital importance for the police department. The police department has a large collection of information about the crime incidents that have been taking place. Earlier, the police departments were using paper based information storing systems. Due to this they had to spend a lot of time as well as man power to analyse existing crime information in order to identify the suspects. With the rapid technological advances including the increasing applications of computerized systems [5], the process of solving crimes is made easy. Data mining techniques have been employed for the crime data analysis. Clustering is one approach of data mining that is used to perform crime analysis. Clustering is a process that groups that data with similar attributes. There are different types of clustering techniques that helps in analysis of crime data. By making use of different types of clustering techniques “hot spots” can be generated by mapping the crime instances. Hot spots are the areas on the map that have high crime intensity or an area in which occurrences of crime are highly predictable. The hot spot methods predict the areas of increased crime risk based on historical crime data. . All the hot spot methods are related to clustering [4]. Thus, with the

help of clustered results based on the analysis of the crime data, identification of crime trend can be done.

For the analysis of the crime activities using clustering techniques, a given description of a crime including its location, type and physical description of the suspects have to be considered.

Based on this, the crime types are categorized into [5], [6]:

1) Theft, 2) murder, 3) kidnap, 4) fraud, 5) traffic violation, 6) drugs, 7) cyber-crime 8) Sex crime, 9) arson gang articles, 10) burglary, 11) pickpocketing, 12) gambling, 13) Forgery, 14) child labour, 15) rape etc.

Based on the location, there are six types of areas where crime incidents often take place and they are [6]:

1) Slums, 2) Residential areas, 3) Commercial areas, 4) VIP zones- high security zone areas for Very important people, 5) Travel points and 6) Markets

Primary database is created in accordance with the information collected based on the types of crimes, the location, and the physical description of the suspects (criminal behavior) and also the time (day, month, and year) at which crime has taken place including all the other available data. The available data is pre-processed, cleaned, refined, linked and extracted and then it is fed to a tool such as WEKA where clustering is done. WEKA is free software available under the GNU General Public License .It allows user to explore a wide variety of data mining techniques .The WEKA tool contains a collection of visualization tools and algorithms for data analysis and predictive modeling . They run quickly and can work with large datasets. In crime analysis, the tool can be used to find commonalties across crimes.

5. CLUSTER ANALYSIS IN BUSINESS (TRAVEL PACKAGE RECOMMENDER SYSTEM)

Tourism has become one of the fastest growing industries in the recent years. India has attracted 6.85 million international tourist arrivals and has seen a steady growth year on year from 4.45 million arrivals in 2006 to 7 million arrivals in 2013. With the improvement in living standards and advancement of time, even an ordinary family can travel comfortably on a small budget. But, the growth of the online information is imposing a great challenge to the tourists as they have to choose from a large number of travel packages [8].A travel package is a general service package provided by a travel company for the individual or a group of tourists based on their travel preferences. A package usually consists of the landscapes and some related information, such as the price, the travel period, and the transportation means. One package usually consists of many landscapes that are located in one or more areas [8]. Specifically, the travel topics are the themes designed for this package, and the landscapes are the travel places of interest and attractions, which usually locate in nearby areas [7]. There are two models that have been used to design a travel package and they are: TAST (Tourist Area Season Topic) Model and TRAST (Tourist Relation Area Season Topic) Model. The TAST model is based on the Bayesian networks. Bayesian networks in travel package system are used to measure the similarity between the travel packages and the tourists. It captures the unique characteristics of the travel data. TRAST model [7] captures the relationship among the tourists in each travel group. To compute the relationship among many tourists, a k-means cluster technique will be used to cluster the tourists into k groups and the relationship serves as the feature for clustering. The result of the clustering will then be compared with the other clustering result. The better the selected features, better clustering results will be observed. Therefore the relationship identified by TRAST model can be better used for clustering tourists and it thus it helps to find the most possible co-travel tourists for a given tourist and can be used as an assessment for travel group automatic formation.

6. CLUSTER ANALYSIS IN MEDICINE (GROWTH PATTERN IDENTIFICATION)

Obesity trends are causing serious health concerns in many countries. Children have fewer weight related health problems than adults. However, overweight children are at a higher risk of developing chronic diseases such as heart diseases and diabetes later in life. Children become overweight for a variety of reasons like lack of physical activity, unhealthy eating patterns, generic factors etc. By making use of clustering methods, the causes for obesity in children can be detected. It is done by grouping together children who share similar body measurements. The cluster that contains the same measurements of children will be plotted as a function of age which results in growth pattern curves. Growth pattern curves that will be displayed can be separated into top most, middle or bottom most clusters based on the body measurements of children belonging to different clusters. For this purpose, Electronic Medical Records (EMR) will be used to gather the data. Out of many clustering algorithms, the two most commonly used algorithms for the identification of growth patterns are: k-means using Euclidian Distance and Expectation Maximization (EM) based algorithms. A comparison of both the results of the algorithms will be done that result in clusters which helps in identification of groups of children that are assigned to the same cluster. Those children that belong to the top most growth pattern can be identified as the ones having highest weight over a period of time and hence are likely to be at risk of obesity [9].

7. CLUSTER ANALYSIS IN EDUCATION (STUDENT'S ACADEMIC PERFORMANCE DETECTION)

Educational data mining is an emerging discipline that is concerned with developing methods for the exploration of unique and large data sets that is generated from the educational settings (ex: universities , intelligent tutoring systems etc.) . Prediction of student's future learning behavior, discovering or improving the learning models, studying the effects of different strategies of teaching that learning software can provide etc. are the goals of educational data mining. Educational data mining mainly helps in studying the performance of the students based on past records [10]. Predicting the performance of students is a challenging task [11] because it involves extraction of large amounts of information from the universities containing student records [10] and analyzes the progress of each student's academic progress. Clustering in this field can be mainly used to group all the students based on their similarity measures (marks, talents, practical knowledge in a particular field, family background). By using clustering methods like k-means clustering, hierarchical clustering, statistical clustering and other clustering techniques, it helps in creation of groups with students having similar learning style which can be improved and can also be made faster. For example, k-means clustering combined with other methods can be used to cluster

those students that share similar learning patterns which helps in identification of cognitive skills in reference to reading comprehension skills for each group [12]. It can also be used to improve the web based learning environment and improve e-learning (via internet) and analyze the student's behavior in an online learning platform that might affect their performance [12]. Another example where clustering is widely used is in the annotations of a digital text. Students, in order to improve their learning process usually highlight, underline, write comments or mark in the margins of the text to make their learning and understanding more efficient and clustering can be used to group the students based on the similarity between their annotations and a quick response recommender system can be developed that provide the students with bookmarks, shared thoughts, discussions, references for further readings and important notes made by the students in the past who belong to the same cluster Examples are kindle , apple iBook etc. with self-learning and reading activities. [13]. Thus, there are different types of clustering techniques which helps in improving the performance of students and also help them to progress in their academics.

8. CLUSTER ANALYSIS IN COMPUTER SCIENCE (IMAGE SEGMENTATION)

To represent and make the analysis of an image easier, image segmentation was introduced. Image segmentation is a process of partitioning an image into multiple segments that is used to locate objects and boundaries (lines, curves etc.) in images. Image segmentation finds its applications in medical imaging, object detection, face detection, face recognition and finger print recognition, video surveillance and many more. According to [15], there has been an increase in usage of collection of data through camera technologies as well as video surveillance cameras for the purpose of recognition or detection of objects in images or for the detection of unwanted activity. Therefore clustering in image segmentation plays a significant role in the field of computer science because it aims to identify a group of similar pixels that belong together or to a specific region and is different from other regions. Clustering finds out a structure in a collection of unlabeled data [14]. The most widely used clustering technique in this field is k- means clustering algorithm as it is said be simple and easy to build an image segmenter from it . It also said to be efficient and it helps to recognize patterns in the images in a better way. There are also other clustering algorithms (Fuzzy C-means, improved Fuzzy C-means, improved k-means etc.) that have been used in image segmentation where all of the algorithms aim to group the image data sets into a number of disjoint groups of clusters [14]. All the algorithms have their advantages and disadvantages and the result that is obtained from one algorithm might not be the same as the other. Therefore image analysis and retrieval of images can be made easier by using clustering techniques in image segmentation.

9. CONCLUSION AND FUTURE SCOPE

As there are different types of clustering techniques based on their cluster model, all of the clustering types are involved in grouping the data into different groups so that the data in each group will share the similar trends and patterns. The quality of clusters produced by clustering method is measured by its ability to discover some or all of the hidden patterns. It has been observed that, the most common type of clustering technique that has been used by different applications of data mining is the k-means clustering technique. It is most widely used because it produces better cluster results as compared to the other clustering techniques and is also said to be computationally faster. But, there are also other types of clustering techniques employed by the applications. Based on the different clustering techniques, the one which provides a better result will be chosen as the best clustering technique that helps in analysis of the data and also helps in prediction. Thus with the different types of clustering techniques and its use in many of the applications, it has been adopted by many of the researchers in various fields to make the analysis of the data easier.

In the field of crime detection area using clustering techniques, there always remains a scope of improvement in terms of visual, intuitive and investigation techniques that can be developed in an effective way for the detection of crime and social link networks can be developed to link the criminals and study their interrelationships. In the field of Education, there are large numbers of factors that play an important role in prediction other than the academic which includes non-cognitive factors, to measure and monitor these factors, suitable data mining techniques are required. Since this paper mainly concentrates on k-means clustering technique in many of the applications, researching on the improvement of K-means clustering algorithms are still not solved completely and hence further attempt and explore will be needed.

The below table illustrates a comparative study of Applications of Data Mining using k-means clustering technique along with advantages and disadvantages. Alternatively there are clustering techniques other than k-means clustering technique that can be used in the applications of data mining.

Table 1 Comparative Study of Applications of Data Mining using k-means clustering technique.

References	Applications	Clustering technique	Advantages	Disadvantages	Other techniques
[16]	Social Science(Crime Pattern Detection)	k-means clustering technique.	Helps to make predictions by identifying the number of active suspects based on the graphical data. Simple, flexible, easy to understand and can be easily implemented.	Mapping real data to Data Mining attributes is not always an easy task and often requires skilled Data Miner and Data Analyst with good domain knowledge.	k-means++ clustering, Nearest Neighbor clustering, Spatio-temporal clustering, Kernel density estimation.
[17]	Business(Travel Package Recommender System)	k-Nearest Neighbor (kNN) clustering technique.	Stores all the available information and classifies new information based on similarity measure.	Scalability is one of the main problems faced when using this technique.	Fuzzy C-means clustering, k-means clustering, Modified k-means clustering (MKC).
[18]	Medicine(Medical Image Segmentation)	k-means clustering technique.	Fast, robust and easier to understand. Gives best result when data sets are distinct or well separated from each other.	Two highly overlapping data sets cannot resolve into clusters.	Fuzzy C-means clustering, Density based clustering.
[11]	Education(Student's academic performance detection)	k-means clustering technique.	Improves the student's performance and enhances the academic planners to monitor the performance and progression level of each student.	Does not handle noisy data. Need to identify the number of clusters in advance.	UCAM (Unique Clustering with Affinity Measure), hierarchical clustering, statistical clustering.

10. REFERENCES

[1] Saurabh Arora, Inderveer Chana ," A Survey of Clustering Techniques for Big Data Analysis", 5th International conference –Confluence the Next Generation International Summit (Confluence),IEEE, pp. 59-65, 2014.

[2] Nisha, Puneet Jai Kumar, "A Survey of Clustering Techniques and Algorithms" Second International Conference on Computing for sustainable Global Development (INDAICom), pp. 304-307, 2015.

[3] Applications, Cluster analysis: https://en.wikipedia.org/wiki/Cluster_analysis#Applications

[4] Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, John S. Hollywood, "Predictive Policing The Role of Crime Forecasting in Law", ©RAND Corporation, 2013.

[5] Qusay Bsoul, Juhana Salim*, Lailatul Qadri Zakaria, " An Intelligent Document Clustering Approach to Detect Crime Patterns", The 4th International Conference on Electrical Engineering and Informatics, Procedia Technology ,Elsevier Ltd. ,pp. 1181-1187, 2013

[6] Priyanka Gera, Rajan Vohra, "City Crime Profiling Using Cluster Analysis", Priyanka Gera et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5, No.4, pp. 5145-5148, 2014.

[7] Qi Liu, Enhong Chen, Senior Member, IEEE, Hui Xiong, Senior Member, IEEE, Yong Ge, Zhongmou Li, and Xiang Wu," A Cocktail Approach for Travel Package Recommendation", IEEE Transactions on Knowledge and Data Engineering, Vol. 26 ,No.2, pp. 278-293, February 2014.

[8] Q. Liu, Y. Ge, Z. Li, E. Chen, and H. Xiong,"Personalized Travel Package Recommendation", IEEE 11th International Conference on Data Mining, pp. 407–416, December 2011.

[9] Moumita Bhattacharya, Deborah Ehrenthal, MD, MPH, Hagit Shatkay," Identifying Growth-Patterns in Children by Applying Cluster analysis to Electronic Medical Record" ,IEEE International Conference on Bioinformatics and Biomedicine, pp. 348-351,2014.

[10] Ritika Saxena, "Educational Data Mining: Performance Evaluation of Decision Tree and Clustering Techniques Using WEKA Platform", International Journal of Computer Science and Business Informatics", IJCSBI.ORG, Vol. 15, No. 2. March 2015.

[11] J. James Manoharan, Dr. S. Hari Ganesh, M. Lovelin Ponn Felcia, "Discovering Student's Academic Performance Based on GPA using k-Means Clustering Algorithm" , IEEE World Congress on Computing and Communication Technologies,2013.

[12] Ashish Dutt, Saeed Aghabozrgi, Maizatul Akmal Binti Ismail, and Hamidreza Mahroeian, "Clustering Algorithms Applied in Educational Data Mining", International Journal of Information and Electronics Engineering, Vol. 5, No. 2, pp. 112-116, March 2015.

[13] Keith Ying, Maiga Chang, Andrew F. Chiarella, Kinshuk, Jia-Sheng Heh, "Clustering Students based on Their Annotations of a Digital Text", IEEE Fourth International Conference on Technology for Education, pp. 20-25,2012.

[14] B.Sathya, R.Manavalan, Image segmentation by clustering methods: Performance analysis", International Journal of Computer Applications, Volume 29, No.11, pp. 27-32, September 2011.

[15] Himanshu S. Bhatt, Richa Singh, Member, IEEE, and Mayank Vatsa, Member, IEEE, "On Recognizing Faces in Videos Using Clustering-Based Re-Ranking and Fusion", IEEE Transactions on Information Forensics and Security, VoL 9, no. 7, pp. 1556-6013, July 2014.

[16] Bashar Aubaidan, Masnizah Mohd and Mohammed Albared, "Comparative Study of k-means and k-means and ++ clustering algorithms on crime domain", Journal of Computer Science, pp.1197-1206, 2014.

[17] Parnika Patil1, V. L. Kolhe, "Survey of Travel Package Recommendation System", International Journal of Science and Research (IJSR), VoL 3, no. 12, pp. 1557-1561**,** December 2014.

[18] Juilee Anil Katkar, Trupti Baraskar, " A Review: Clustering Techniques for Medical Image Segmentation", International Journal of Advance Foundation and Research in Computer (IJAFRC) Vol 1, no.12, pp. 103-110, December 2014.