Data Clustering In Various Fields of Applications

Prabhu M
Assistant Professor, Department of Computer Science,
Shanmuga Industries Arts and Science College, Tiruvannamalai.
Prabhumca007@gmail.com

Abstract: Fast retrieval of the relevant information from the databases has always been a significant issue. Different techniques have been developed for this purpose, one of them is Data Clustering. In this paper Data Clustering is discussed along with its several traditional approaches and their analysis. Some applications of Data Clustering like Data Mining using clustering techniques in various fields are discussed.

Keywords: Clustering, Data Mining, Knowledge Discovery, K-means, Types of clustering.

I. INTRODUCTION

Data clustering is a method in which we make cluster of objects that are somehow similar in characteristics. The criterion for checking the similarity is implementation dependent [1].

Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre-defined classes, whereas in clustering the classes are also to be defined [2].

Precisely, Data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database systems the number of disk accesses is to be minimized. In clustering the objects of similar properties are placed in one class of objects and a single access to the disk makes the entire class available [3].

Clustering analysis has been an emerging research issue in data mining due its variety of applications. With the advent of many data clustering algorithms in the recent few years and its extensive use in wide variety of applications, including image processing, computational biology, mobile communication, medicine and economics, has led to the popularity of this algorithms [4].

II. DATA MINING

Data mining is a relatively new field of research that its objective is to acquire knowledge from large amounts of data within databases or data warehouses [5]. In medical and health care areas, due to regulations and due to the availability of computers, a large amount of data is becoming available. It encompasses classification, clustering, association rule learning, etc., whose goals are to improve decisions making and performances in organizations [6]. Data mining is an essential step in the process of knowledge discovery and the following steps involved in the kdd process in data mining are data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation[7].

Data mining is the significant to attainment a good advantage in determining the control on sales, customer satisfaction and corporate profits. Data mining brings the required insights for customer loyalty, unlocking hidden profitability and reducing the churn[8]. Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs.

Data mining techniques can be classified broadly as

1. Predictive:
   a. Classification
   b. Regression
   c. Time series Analysis
   d. Prediction

2. Descriptive:
   a. clustering
   b. Summarization
   c. Association Rules
   d. Sequence Discovery

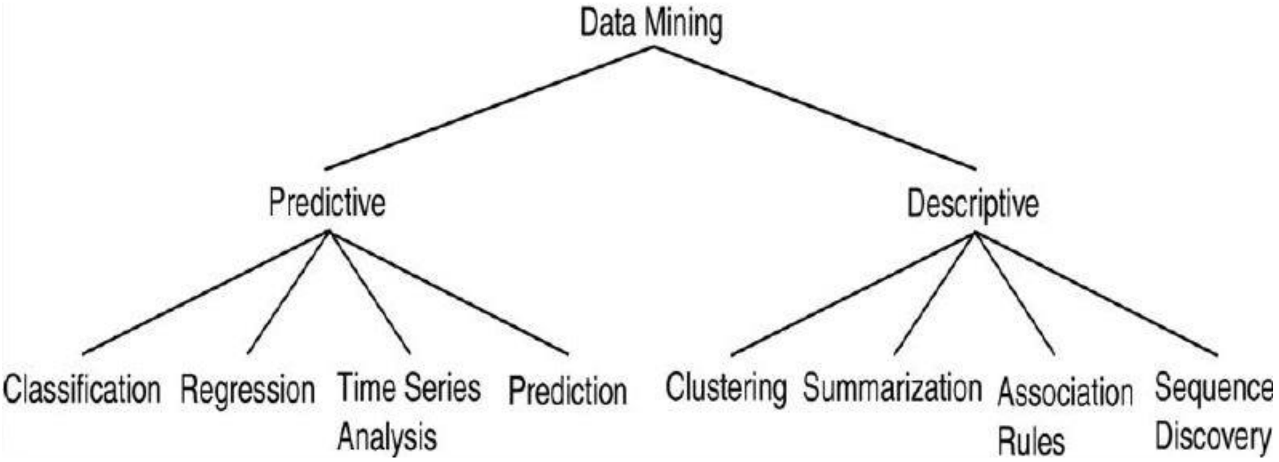

Figure1. Data mining Models

III. CLUSTERING

Clustering of data is a well-researched topic in computer sciences. Many approaches have been designed for different tasks [9]. Cluster analysis is one of the typical tasks in Data Mining, and it groups data objects based only on information found in the data that describes the objects and their relationships. Algorithm developed may give best result with one type of data set but may fail or give poor result with data set of other types [9]. Although there has been many attempts for standardizing the algorithms which can perform well in all case of scenarios but till now no major accomplishment has been achieved. Many clustering algorithms have been proposed so far. However, each algorithm has its own merits and demerits and cannot work for all real situations. Before exploring various clustering algorithms in detail let's have a brief overview about what is clustering [10].

A. Types of clustering algorithms

Every methodology follows a different set of rules for defining the 'similarity' among data points. In fact, there are more than 100 clustering algorithms known. But few of the algorithms are used popularly,

i. **Connectivity models:** As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lack scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants [11, 15, 17].

ii. **Centroid models:** These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end has to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima [12, 13, 14].

iii. **Distribution models:** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions [16].

iv. **Density Models**: These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assigns the data points within these regions in the same cluster. Popular examples of density models are DBSCAN and OPTICS [17].

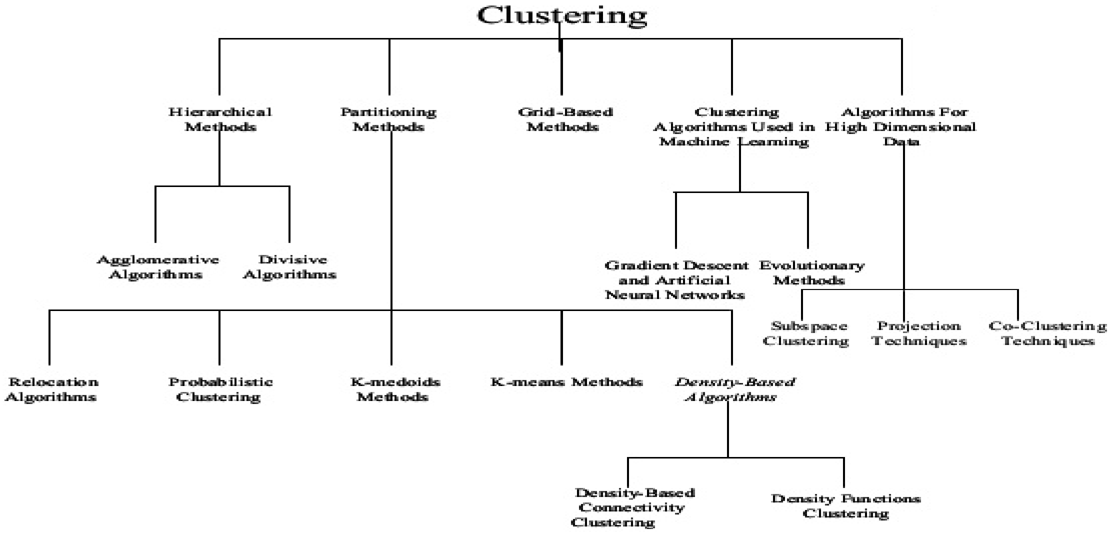

Figure2. Types of Clustering

We used K-means, seeded K-means, and constrained K-means, and clustering validity techniques to achieve our objectives. Our results indicate that: the data set contains seven classes; seeded K-means outperforms K-means and constrained K-means [18]. We used K-means, seeded K-means, and constrained K-means, and clustering validity techniques to achieve the most of the objectives.

IV. Data Clustering in various applications

Clustering analysis has been an emerging research issue in data mining due its variety of applications. With the advent of many data clustering algorithms in the recent few years and its extensive use in wide variety of applications, including image processing, computational biology, mobile communication, medicine and economics, has led to the popularity of this algorithms. Main problem with the data clustering algorithms is that it cannot be standardized.

A. Education

The use of clustering techniques for identifying learner types in Massive Open Online Courses (MOOCs) [6, 12, 20]. The different clustering techniques in the context of educational data are compared with many to identify the learners. The K means clustering for learner identification within more constrained contexts presented by a highly technical and advanced engineering MOOC [12]. Investigate on different types of learner behavior that emerge from the above-mentioned clustering and the ways in which each cluster is different from the rest. To identify appropriate labels for each user group according to their dominant behavioral characteristics and to show that the difference in learner behavior across clusters is statistically significant [19, 20].

B. Healthcare

Clustering techniques has not been widely used in large healthcare claims databases where the distribution of expenditure data is commonly severely skewed [21, 22]. The purpose is to identify cost change patterns of patients with disease by applying different clustering methods. The K-means Cluster Analysis (CA) method appeared to be the most appropriate in healthcare claims data with highly skewed cost information when taking into account both change of cost patterns and sample size in the smallest cluster [23].

C. *Multimedia*

Clustering large datasets in which the data cannot be stored in main memory in streaming model (sequential model). So the desired facilitated in order to organize the items of the collection to be clustered in a graph, where the nodes represent the items and a link between a pair of nodes exists if the model predicted that the corresponding pair of items belongs to the same cluster and the cost is given without prior knowledge of K [24]. Original K-means requires multiple passes through data reduces the number of clusters increases the cost of optimal K-means and determine better facilitate cost. A graph-based multimodal clustering approach used in order to organize the items of the collection to be clustered in a graph, where the nodes represent the items and a link between a pair of nodes exists if the model predicted that the corresponding pair of items belongs to the same cluster [25]. Using clustering techniques detecting social events and discovering problems for collections of social multimedia is processed.

V. Conclusion

In this paper, we try to give the basic concept of clustering by first providing the definition and clustering and then the definition of some related terms. We give some examples to elaborate the concept. Then we give different approaches to data clustering and types of clustering to implement that approaches. The applications of clustering are also discussed with the examples of online education, healthcare, multimedia with data mining using data clustering.

Acknowledgment

I thank our colleagues from Shanmuga Industries Arts and Science College, Department of Information systems and Management and Department of Computer science who provided insight and expertise that greatly assisted the research for this paper. I thank Mr. S.Sampath Rajan, Head of ISM Department and Dr. K. Anandaraj, Principal of our college for assistance with clustering technique, methodology and Dr. Praveen Sundar, Director of MCA Department of our college for valuable comments that greatly improved the manuscript.

REFERENCES

[1] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining.
[2] Hand, D. J. (2007). Principles of data mining. Drug safety, 30(7), 621-622.
[3] Mining, W. I. D. (2006). Data Mining: Concepts and Techniques. Morgan Kaufinann.
[4] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.
[5] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.
[6] Sundar, P. P., & Kumar, A. S. (2016). A systematic approach to identify unmotivated learners in online learning. Indian journal of science and technology, 9(14).
[7] Agrawal, D., & Aggarwal, C. C. (2001, May). On the design and quantification of privacy preserving data mining algorithms. In Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 247-255). ACM.

[8] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... & Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1), 1-37.

[9] Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008, June). Data mining algorithms to classify students. In Educational data mining 2008.

[10] Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

[10] Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. IEEE Transactions on knowledge and data engineering, 17(4), 491-502.

[11] Freitas, A. A. (2013). Data mining and knowledge discovery with evolutionary algorithms. Springer Science & Business Media.

[12] Sundar, P. P., & Kumar, A. S. (2013). Evaluation of Regional Benchmark Impact in EDM. International Journal of Computer Science Issues (IJCSI), 10(2 Part 2), 531.

[13] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.

[14] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE Transactions on Pattern Analysis & Machine Intelligence, (7), 881-892.

[15] Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.

[16] Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. IEEE Transactions on neural networks, 11(3), 586-600.

[17] Kumar, V., & Chadha, A. (2011). An empirical study of the applications of data mining techniques in higher education. International Journal of Advanced Computer Science and Applications, 2(3).

[18] Park, Y., Yu, J. H., & Jo, I. H. (2016). Clustering blended learning courses by online behavior data: A case study in a Korean higher education institute. The Internet and Higher Education, 29, 1-11.

[19] Moucary, C. E., Khair, M., & Zakhem, W. (2011). Improving student's performance using data clustering and neural networks in foreign-language based higher education. The Research Bulletin of Jordan ACM, 2(3), 27-34.

[20] Sundar, Praveen. "Quasi Framework: A new student disengagement detection in online learning." International Journal of Engineering Research & Technology (IJERT) 1, no. 10 (2012).

[21] Jeyakumar, Balajee, MA Saleem Durai, and Daphne Lopez. "Case Studies in Amalgamation of Deep Learning and Big Data." In HCI Challenges and Privacy Preservation in Big Data Security, pp. 159-174. IGI Global, 2018.

[22] Sethumadahavi R., Balajee J. (2017). "Big Data Deep Learning in Healthcare for Electronic Health Records." International Scientific Research Organization Journal, 2(2), pp.31-35.

[23] Hinneburg, A., & Keim, D. A. (1998, August). An efficient approach to clustering in large multimedia databases with noise. In KDD (Vol. 98, pp. 58-65).

[24] Faloutsos, C., & Lin, K. I. (1995). FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets (Vol. 24, No. 2, pp. 163-174). ACM.

[25] Hinneburg, A., & Keim, D. A. (2003). A general approach to clustering in large databases with noise. Knowledge and Information Systems, 5(4), 387-415.