Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 5, May 2017, pg.183 – 186

Data Mining Clustering Techniques – A Review

Shivangi Bhardwaj

CSE, Amity University Haryana, India
shivangibhardwaj28@gmail.com

Abstract— With the advent increase in health issues in our day to day life, data mining has been an essential part to fetch the knowledge and to form different patterns. Accuracy is very necessary once it involves patient care and computerizing this large quantity of knowledge enhances the standard of the complete system. Data mining works on the principal of KDD (Knowledge Discovery in Databases). Data mining tools compare symptoms, causes, treatments and negative effects so as to proceed to investigate that which action can be proved simplest for a group of patients. The data from aid organizations are voluminous and heterogeneous. It must be collected and kept in organized type and their integration allows the formation unite medical system. Data mining in health offers unlimited possibilities for analyzing models less visible or hidden to common analysis techniques. In this paper, the different techniques are compared on which research techniques have been applied.

Keywords— Data Mining, Clustering, Data sets, Database, groups.

I. INTRODUCTION

The rise in the level of knowledge due to the upcoming technology has led to a cloud of large amount of data. As people are getting more and more aware of the new methods in their respective fields so the databases have also enlarged as the need to store those outputs and inputs has come live. Data Mining is a one stage in KDD handle which contain information examination and revelation calculations [10]. Cluster Analysis is an essential strategy for database mining. It is either utilized as a remain solitary instrument to get understanding into the dissemination of an informational index, e.g. to concentrate encourage investigation and information preparing, or as a pre-processing venture for different calculations working on the identified bunches.

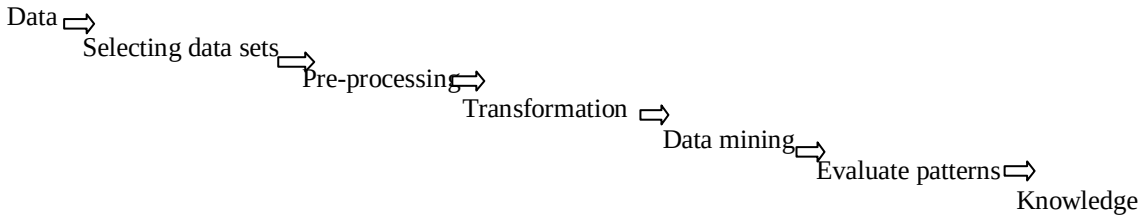
Data mining is a non-minor procedure of distinguishing legitimate, novel, potential valuable and eventually reasonable designs in information. Data mining, the extraction of the covered up prescient data from vast database, is an intense new innovation with potential to examine vital data in the information distribution centre. The term data mining alludes to the finding of pertinent and valuable data from database. Cognitive science is helpful to study human brain on which mining techniques could be applied [2].

Bigger and bigger measures of information are gathered and put away in databases expanding the requirement for productive and compelling examination strategies to make utilization of the data contained verifiably in the information. One of the essential information examination undertakings is group investigation which is planned to help a client to comprehend the common gathering or structure in an informational

collection. Hence, the development of enhanced bunching calculations has gotten a great deal of consideration over the most recent couple of years.

Data mining flow of work goes as follows [12]:

1. Data
2. Selection
3. Pre-processing
4. Transformation
5. Data mining on transformed data
6. Interpretation or Evaluation of Patterns
7. Output- Knowledge



II. DIFFERENT TECHNIQUES

Clustering is the errand of separating the populace or information focuses into various gatherings with the end goal that information focuses in similar gatherings are more like other information focuses in a similar gathering than those in different gatherings. In basic words, the point is to isolate clusters with comparative attributes and dole out them into groups. A typical spellbinding assignment in which one tries to recognize a limited arrangement of classifications or group to depict the information is known as clustering.

There are different types of clustering methods or techniques:

1. **Hierarchical Methods**: It follows two approaches which are bottom-up and top-down where all clusters are combined and are transformed into one & all observations are split into different bunches respectively.
2. **Partitioning Methods**: Assume we are given a database of "n" items and the partitioning strategy builds "k" segment of information. Each division will have a cluster and k <= n. It implies that it will arrange the information into k gatherings, which fulfil the accompanying prerequisites −
   Each gathering contains no less than one object.
   Each question must have a place with precisely one gathering.

3. **Density Based Methods**: This method is based on two function types that are connectivity and density functions. The basic idea behind it is that the thick clusters are formed and they should grow as long as they cross the threshold of the neighbouring clusters. Firstly, all data objects within the group are mutually density connected to each other. Secondly, if a data object is density connected to any other data object within the group then both the data objects must be part of the same group [11]. For high density salt-pepper noise is suggested in one of the papers [1].

4. **Grid-Based Methods**: The objects together shape a framework. The object form is quantized into limited number of cells that shape a network structure.

5. **Model Based Methods**: In this model is theorized for each group to locate the best shape of data for a given model.

There is further categorization of the above mentioned methods based on different algorithmic techniques:

1. Hierarchical methods:
   - Agglomerative Algorithms
   - Divisive Algorithms
2. Partitioning methods:
   - Relocation Algorithms
   - Probabilistic Clustering
   - K-Medoids Methods
   - K-Means Methods

3. Density-based algorithms:
   - Density-based connectivity clustering
   - Density functions clustering

4. Grid-based methods: Methods based on co-occurrence of categorical data
   - Constraint-based clustering
   - Clustering algorithms used in machine learning
   - Gradient descent and artificial neural networks
   - Evolutionary methods

5. Model Based Algorithms:
   - Algorithms for high dimensional data
   - Subspace Clustering
   - Projection Techniques
   - Co-Clustering Techniques

From the above mentioned algorithms and techniques the consolidated data has been formed which is as follows:

TABLE I

Proposed Algorithm	**Work Done**	**Data Sets**
DVBSCAN	To represent moderately uniform areas without being isolated by meagre areas [8].	Large Spatial Database
OPTICS	No limit on global parameter setting	Small and large database
Enhanced VDBSCAN	Use automatic parameter selection	2-dimensional dataset [14]

III. CONCLUSIONS

Data Mining is extracting data from the database. This is done to get some fruitful facts and figures from the huge flood and find the crux of the data sets. In this paper the different clustering techniques and their algorithms are discussed and the combination of those algorithms is jotted down in tabular form. Data mining is also done for deep cleaning of the data sets so that clustering techniques could be applied to figure out the results as needed. Moreover in this paper automatic work is also discussed. So that less time is consumed to collect the data to form clusters. It is clear from the above mentioned data that different data mining techniques can be used or applied not only on small data but on large and large amount of data sets.

REFERENCES

[1] Akansha Singh, Poonam Sharma, and Krishna Kant Singh, "A Detail Preserving Filter For High Density Salt-And-Pepper Noise," *IJEAST*, vol. 1, pp. 120-123, June 2016.
[2] Komal, "Cognitive Science: Bridging the Gap between Machine and Human Intelligence," *IJCA*, vol. 114, pp. 16-19, March 2015.
[3] Poonam Sharma, Amit Wadhwa, and Komal, "Analysis of Selection Schemes for Solving an Optimization Problem in Genetic Algorithm," *IJCA*, vol. 93, pp. 1-3, May 2014.
[4] Sonamdeep Kaur, Sarika Chaudhary, and Neha Bishnoi, "A Survey: Clustering Algorithms in Data Mining," *IJCA*, Cognition 2015, p. 12-14.
[5] Shivani Sihmar, Poonam Sharma, "Image processing techniques for offline handwritten recognition"
[6] Sonamdeep Kaur, Sarika Chaudhary, and Neha Bishnoi, "Optimization of CMAP Based Algorithms For Extracting Rare Sequence Patterns," *GE-IJER*, vol. 3, pp. 40-51, June 2015.
[7] Ashima Narang, Vijay Laxmi, "Various Load Balancing Techniques in Cloud Computing," *IJCSMC*, vol. 3, pp. 502-509, 2014.
[8] Dr.Chandra.E, Anuradha.V.P, "A Survey on Clustering Algorithms for Data in Spatial Database Management Systems," *IJCA*, vol. 24, pp. 19-26, June 2011.
[9] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, J&g Sander, "OPTICS: Ordering Points To Identify the Clustering Structure," *SIGMOD '99 Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, vol. 28, pp. 49-60, June 1999.
[10] Pragati Shrivastava, Hitesh Gupta, "A Review of Density-Based clustering in Spatial Data," *IJACR*, vol. 2, pp. 200-202, September 2012.

[11] Sourajit Behera, Rinkle Rani, "Comparative Analysis of Density based Outlier Detection techniques on Breast Cancer data Using Hadoop and Map Reduce," *International Conference on Inventive Computation Technologies (ICICT)*, January 2017.
[12] R. Tamilselvi1, S. Kalaiselvi, "An Overview of Data Mining Techniques and Applications," *IJSR*, vol. 2, pp. 506-509, February 2013.
[13] Anant Ram, Sunita Jalal, Anand S. Jalal, Manoj Kumar, "A Density based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases," *IJCA*, vol. 3, pp. 1-4, June 2010.
[14] S.Vijayalakshmi, Dr.M.Punithavalli, "Improved Varied Density Based Spatial Clustering Algorithm with Noise," *IEEE*, 2010.

