© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 08 | October – December 2016
Paper is available at www.jrps.in | Email : info@jrps.in

REVIEW PAPER ON K-MEAN CLUSTERING

[1]Preeti saini, Research Scholar, Department of CSE, JIET Jind. preetisaini06@gmail.com
[2]Sapna aggarwal, Assistant professor, Department of CSE, JIET Jind. Sapna.ruby@gmail.com

Abstract: K-Means Clustering algorithm is an idea, in which there is need to classify the given data set into K clusters; the value of K (Number of clusters) is defined by the user which is fixed. In this first the centroid of each cluster is selected for clustering and then according to the chosen centriod, the data points having minimum distance from the given cluster, is assigned to that particular cluster.

Keywords—Data mining, web mining, web intelligence, knowledge discovery, fuzzy logic, K-mean

1. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. The popular book "Data mining: Practical machine learning tools and techniques with Java" (which covers mostly machine learning material) was originally to be named just "Practical machine learning", and the term "data mining" was only added for marketing reasons. Often the more general terms "(large scale) data analysis", or "analytics" – or when referring to actual methods, artificial intelligence and machine learning – are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms *data dredging*, *data fishing*, and *data snooping* refer to the use of data mining methods to sample parts of a larger

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 08 | October – December 2016
Paper is available at www.jrps.in | Email : info@jrps.in

population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

2. SOFT COMPUTING

Soft computing is the use of inexact solutions to computationally hard tasks such as the solution of NP-complete problems, for which there is no known algorithm that can compute an exact solution in polynomial time.

The process of knowledge discovery in databases, often also called **data mining**, is the first important step in knowledge management technology. End users of these tools and systems are at all levels of management operative workers and managers. And these are their demands on the processing and analysis of data and information that affect the development of these tools.

Components of soft computing include:

- Neural networks (NN)
  - Perceptron
- Support Vector Machines (SVM)
- Fuzzy logic (FL)
- Evolutionary computation (EC), including:
  - Evolutionary algorithms
    - Genetic algorithms
    - Differential evolution
  - Metaheuristic and Swarm Intelligence
    - Ant colony optimization
    - Particle swarm optimization
    - Firefly algorithm
    - Cuckoo search
- Ideas about probability including:
  - Bayesian network
- Chaos theory

Generally speaking, soft computing techniques resemble biological processes more closely than traditional techniques, which are largely based on formal logical systems, such as sentential logic and predicate logic, or rely heavily on computer-aided numerical analysis (as in finite element analysis). Soft computing techniques are intended to complement each other.

Unlike hard computing schemes, which strive for exactness and full truth, soft computing techniques exploit the given tolerance of imprecision, partial truth, and uncertainty for a particular problem. Another common contrast comes from the observation that inductive reasoning plays a larger role in soft computing than in hard computing.

3. DATA MINING PROCESS

The **Knowledge Discovery in Databases (KDD) process** is commonly defined with the stages:

(1) Selection
(2) Pre-processing
(3) Transformation
(4) *Data Mining*
(5) Interpretation/Evaluation

It exists, however, in many variations on this theme, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) which defines six phases:

(1) Business Understanding
(2) Data Understanding
(3) Data Preparation
(4) Modeling
(5) Evaluation
(6) Deployment

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 08 | October – December 2016
Paper is available at www.jrps.in | Email : info@jrps.in

or a simplified process such as (1) pre-processing, (2) data mining, and (3) results validation.

Polls conducted in 2002, 2004, and 2007 show that the CRISP-DM methodology is the leading methodology used by data miners. The only other data mining standard named in these polls was SEMMA. However, 3-4 times as many people reported using CRISP-DM. Several teams of researchers have published reviews of data mining process models, and Azevedo and Santos conducted a comparison of CRISP-DM and SEMMA in 2008.

4. AREA OF APPLICATIONS

Bioinformatics and Biomedicine

SC has attracted close attention of researchers and has also been applied successfully to solve problems in bioinformatics and biomedicine. Nevertheless, the amount of information from biological experiments and the applications involving large-scale high-throughput technologies is rapidly increasing nowadays. Therefore, the ability of being scalable across large-scale problems becomes an essential requirement for modern SC approaches.

5. Survey of earlier work

The use of data mining techniques in manufacturing began in the 1990s and it has gradually progressed by receiving attention from the production community. These techniques are now used in many different areas in manufacturing engineering to extract knowledge for use in predictive maintenance, fault detection, design, production, quality assurance, scheduling, and decision support systems. Data can be analyzed to identify hidden patterns in the parameters that control manufacturing processes or to determine and improve the quality of products. A major advantage of data mining is that the required data for analysis can be collected during the normal operations of the manufacturing process being studied and it is therefore generally not necessary to introduce dedicated processes for data collection. Since the importance of data mining in manufacturing has clearly increased over the last 20 years, it is now appropriate to critically review its history and application.

Data mining techniques becomes the basic element of modern business. Although the idea is not new, new technologies and implemented standards make a contribution to their growing popularity. Regarding to mining model usage SQL Server 2005 stands breakthrough in this area. Thanks to the DMX language either programmers or database administrators are able to create Data Mining Systems in simple way.

Although economical and business publications are very fruitful of data mining approaches, the described problem is presented rather weak in the international publications. Nethertheless some industrial appliances of data mining technology were considered in (Duebel, C., 2003).

Industrial usage of data mining techniques opens new possibilities in decision making not only for top level management, but also for advisory or control systems. Several prediction, classification or even anomaly detection algorithms implementation may become lucrative tool for industrial process appropriate stages optimization, that combines diagnosis and control functions.

The reviewed literature shows that there is a rapid growth in the application of data mining in industry and manufacturing. However, there is still slow adoption of this technology in some industries for several reasons including both difficulties in determining the type of data mining function to be performed in any

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 08 | October – December 2016
Paper is available at www.jrps.in | Email : info@jrps.in

particular knowledge area and question ofchoice the most appropriate data mining technique regarding to many possibilities.

Waldemar Wójcik and Konrad Gromaszek (*Lublin University of Technology,Poland) introduced* "Data Mining Industrial Applications". Data mining is blend of concepts and algorithms from machine learning, statistics, artificial intelligence, and data management. With the emergence of data mining, researchers and practitioners began applying this technology on data from different areas such as banking, finance, retail, marketing, insurance, fraud detection, science, engineering, etc., to discover any hidden relationships or patterns.

Jiawei Han and Jing Gao University of Illinois at Urbana-Champaign wrote paper on "Research Challenges for Data Mining in Science and Engineering"

With the rapid development of computer and information technology in the last several decades, an enormous amount of data in science and engineering has been and will continuously be generated in massive scale, either being stored in gigantic storage devices or °owing into and out of the system in the form of data streams. Moreover, such data has been made widely available, e.g., via the Internet. Such tremendous amount of data, in the order of tera- to peta-bytes, has fundamentally changed science and engineering, transforming many disciplines from data-poor to increasingly data-rich, and calling for new, data-intensive methods to conduct research in science and engineering.

CLUSTERING SYSTEM BASED ON TEXT MINING USING THE K-MEANS ALGORITHM by Prabin Lama

"Clustering System based on Text Mining using the K means algorithm," is mainly focused on the use of text mining techniques and the K means algorithm to create the clusters of similar news articles headlines. The project study is based on text mining with primary focus on data-mining and information extraction. The news headlines and the links to the different news portal are fetched via an XML file to the clustering system. The news headlines within the XML file are then preprocessed using document preprocessing techniques and finally grouped in the clusters based on their similarities. These clusters are displayed in a sample webpage with the corresponding links to the news portal sites.

Performance Improvement Of Web Usage Mining By Using Learning Based K-Mean Clustering

Due to the increasing amount of data available online, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. Web mining technologies are the right solutions for knowledge discovery on the Web. The knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering, and Web based data warehousing. In the present work, we propose a new technique to enhance the learning capabilities and reduce the computation intensity of a competitive learning multi-layered neural network using the K-means clustering algorithm.

6. K-MEANS CLUSTERING ALGORITHM

K-means clustering is a well known partitioning method. In this objects are classified as belonging to one of K-groups. The result of partitioning method is a set of K clusters, each object of data set belonging to one cluster. In each cluster there may be a centroid or a cluster representative. In case where we consider real - valued data, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases.

Types of Clustering Algorithms are:

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 08 | October – December 2016
Paper is available at www.jrps.in | Email : info@jrps.in

1. K-means Clustering Algorithm
2. Hierarchical Clustering Algorithm
3. Density Based Clustering Algorithm
4. Self-organization maps (SOM)
5. EM clustering Algorithm

STEPS OF K-MEANS CLUSTERING ALGORITHM

K-Means Clustering algorithm is an idea, in which there is need to classify the given data set into K clusters; the value of K (Number of clusters) is defined by the user which is fixed. In this first the centroid of each cluster is selected for clustering and then according to the chosen centriod, the data points having minimum distance from the given cluster, is assigned to that particular cluster. Euclidean Distance is used for calculating the distance of data point from the particular centroid. This algorithm consists of four steps:

1. **Initialization:** In this first step data set, number of clusters and the centroid that we defined for each cluster.

2. **Classification:** The distance is calculated for each data point from the centroid and the data point having minimum distance from the centriod of a cluster is assigned to that particular cluster.

3. **Centroid Recalculation:** Clusters generated previously, the centriod is again repeatly calculated means recalculation of the centriod.

4. **Convergence Condition:** Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between the clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied.

Main advantages:

1. K-means clustering is very Fast, robust and easily understandable. If the data set is well separated from each other data set, then it gives best results.
2. The clusters do not having overlapping character and are also non-hierarchical in nature.

Main disadvantages:

1. In this algorithm, complexity is more as compared to others.
2. Need of predefined cluster centers.
3. Handling any of empty Clusters: One more problems with K-means clustering is that empty clusters are generated during execution, if in case no data points are allocated to a cluster under consideration during the assignment phase.

The experimental results demonstrated that the proposed ranking based K-means algorithm produces better results than that of the existing k-means algorithm

7. PROPOSED IMPLEMENTATION

Our Proposed implementation is to apply fuzzy modeling methods for web mining.

The main aim is to eliminate the limitations of K-mean clustering algorithm, we will customize algorithm as follow.

1. **Initialization:** In this first step data set, number of clusters and the centroid should

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 08 | October – December 2016
Paper is available at www.jrps.in | Email : info@jrps.in

be calculated automatically according to size of data.

2. **Classification:** The distance is calculated for each data point from the centroid and the data point having minimum distance from the centriod of a cluster is assigned to that particular cluster.

3. **Centroid Recalculation:** Clusters generated previously, the centriod is again repeatly calculated means recalculation of the centriod.

4. **Convergence Condition:** Some convergence conditions are given as below:

4.1 Stopping when reaching a given or defined number of iterations.

4.2 Stopping when there is no exchange of data points between the clusters.

4.3 Stopping when a threshold value is achieved.

5. If all of the above conditions are not satisfied, then go to step 2 and the whole process repeat again, until the given conditions are not satisfied.

6. ***Elimination of Empty Clusters: Clusters*** *generated previously are rechecked*

Clusters where no data points are allocated to a cluster under consideration during the assignment phase are eliminated.

Benefits of proposed Implementation over traditional

1) No need of predefined cluster center
2) There will be no Empty clusters at the end

8. Conclusion

The Internet of Things concept arises from the need to manage, automate, and explore all devices, instruments, and sensors in the world. In order to make wise decisions both for people and for the things in IoT, data mining technologies are integrated with IoT technologies for decision making support and system optimization. Data mining involves discovering novel, interesting, and potentially useful patterns from data and applying algorithms to the extraction of hidden information

Due to the increasing amount of data available online, the World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. Web mining technologies are the right solutions for knowledge discovery on the Web. The knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering, and Web based data warehousing.

3) The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure. In data mining K-means clustering algorithm is one of the efficient unsupervised learning algorithms to solve the well-known clustering problems. The disadvantage in k-means algorithm is that, the accuracy and efficiency is varied with the choice of initial clustering centers on choosing it randomly.The main aim of our research is to eliminate the limitations of K-mean clustering algorithm

References

- J. Liu, S. Zhang, Y. Ye, Agent-based characterization of web regularities, in N.

Zhong, et al. (eds.), Web Intelligence, NewYork: Springer, 2003, pp. 19–36.

- J. Liu, N. Zhong, Y. Y. Yao, Z. W. Ras, The wisdom web: new challenges for web intelligence (WI), J. Intell. Inform. Sys.,20(1): 5–9, 2003.

- Congiusta, A. Pugliese, D. Talia, and P. Trunfio, Designing GridServices for distributed knowledge discovery, Web Intell. Agent Sys, 1(2): 91–104, 2003.

- J. A. Hendler and E. A. Feigenbaum, Knowledge is power: the semantic web vision, in N. Zhong, et al. (eds.), Web Intelligence: Research and Development, LNAI 2198, Springer, 2001, 18–29.

- N. Zhong and J. Liu (eds.), Intelligent Technologies for Information Analysis, New York: Springer, 2004.

- Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project".

- Journal of Machine Learning Research **11**: 2533–2541. the original title, "Practical machine learning", was changed ... The term "data mining" was [added] primarily for marketing reasons.