Review on Clustering Mechanism For Web Based Data Mining

Ritu[1], Mahesh Kumar[2]

[1]M.Tech. Student ,Computer Science & Engineering
Ganga Institute of Technology and Management Kablana, Jhajjar, Haryana, India
[2]*Associate Professor*, Computer Science & Engineering
Ganga Institute of Technology and Management Kablana, Jhajjar, Haryana, India
[1] lohchabritu7@gmail.com; [2]maheshmalkani@gmail.com

Abstract— K-Means Clustering algorithm is need to classify given data set into K clusters; value of K is defined by user which is fixed. In this first centroid of each cluster is selected for clustering & then according to chosen centriod, data points having minimum distance from given cluster, is assigned to that particular cluster.

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

I.INTRODUCTION

Data mining *an interdisciplinary subfield of computer science, is computational process of discovering patterns in large data sets involving methods at intersection of artificial intelligence, machine learning, statistics, & database systems. Data mining process is to extort information of data set & transform it into an understandable structure for further use. Aside from raw analysis step, it involves database & data management aspects, data pre-processing, model & inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, & online updating.*
The term is a goal of extraction of sample & knowledge from large amount of data, not extraction of data itself. It also is a buzzword & is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, & statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, & business intelligence.

II. *SOFT COMPUTING*

Soft computing *has been useful of inaccurate solutions to computationally very solid tasks such as solution of NP-complete problems, for which there is no known algorithm that could compute an exact solution in polynomial time.*
The process of knowledge discovery in databases, often also called ***data mining****, is first important step in knowledge management technology. End users of these tools & systems are at all levels of management operative workers & managers. & these are their demands on processing & analysis of data & information that affect 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 & 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 & predicate logic, or rely heavily on computer-aided numerical analysis (as in finite element analysis). Soft computing techniques are intended to complement each other.

III. DATA MINING PROCESS

*The **Knowledge Discovery in Databases (KDD) process** is commonly defined within 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 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

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

Polls conducted in 2002, 2004, & 2007 show that CRISP-DM methodology is leading methodology used by data miners. only for data mining standard named in these polls was SEMMA. However, 3-4 times as many people reported using CRISP-DM. Several teams of researchers had been published data mining process models.

AREA OF APPLICATIONS

Bioinformatics & Biomedicine

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

IV. *SURVEY OF EARLIER WORK*

Waldemar Wójcik & Konrad Gromaszek (Lublin University of Technology,Poland) introduced "Data Mining Industrial Applications". *Data mining is blend of concepts & algorithms from machine learning, statistics, artificial intelligence, & data management. within emergence of data mining, researchers & 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 & Jing Gao University of Illinois at Urbana-Champaign wrote paper on "Research Challenges for Data Mining in Science & Engineering"

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

Text mining using k-means algorithm Clustering system based by Prabin Lama

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

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

Due to increasing amount of data available online, World Wide Web had becoming one of most valuable resources for information retrievals & knowledge discoveries..

V. *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. 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, arithmetic mean of 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:

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 given data set into K clusters; value of K (Number of clusters) is defined by user which is fixed. In this first centroid of each cluster is selected for clustering & then according to chosen centriod, data points having minimum distance from given cluster, is assigned to that particular cluster.

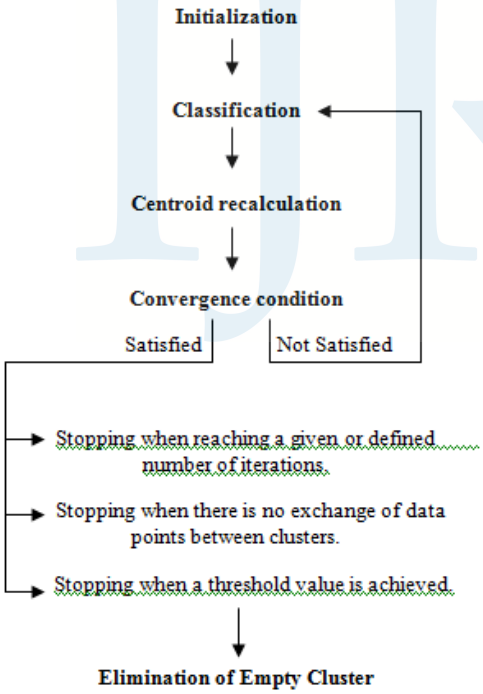

Figure 1 Proposed algorithm

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 within 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 assignment phase.

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

VI. PROPOSED IMPLEMENTATION

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

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

1. ***Initialization:*** *In this first step data set, number of clusters & centroid should be calculated automatically according to size of data.*

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

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

4.3 Stopping when a threshold value is achieved.

5. If all of above conditions are not satisfied, then go to step 2 & whole process repeat again, until 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 assignment phase are eliminated.

Benefits of proposed Implementation over traditional

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

VII. CONCLUSIONS

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

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

REFERENCES

1. 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.

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

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

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

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

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

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