K-Mean and K-Prototype Algorithms Performance Analysis

Izhar Ahmad

Abstract

Evaluation of computer and information technologies has changed the way people used to communicate and perform their tasks. Data mining algorithm is utilized for converting the data in to knowledge information that can be used in future for performing different tasks. In this research, we have presented a system design approach for the K-Mean and K-Prototype Algorithms Performance Analysis. The system architecture in this research is presents a detail discussion of the k-means and k-prototype to recommend efficient algorithm for outlier detection and other issues relating to the database clustering. The system design approach is based on the open source technologies. The verification and validation of the system is based on the simulation.

Index Terms: Database clustering, K-Mean, K-Prototype Algorithms, Performance analysis

1. Introduction

Advancement in the database storage, mining and data processing methods is considered as the centre piece of attention since years, especially in the field of computer sciences, engineering and management. The demand of the data ware housing is increasing day by day. There are number of cooperates, Small Business Companies [SME], Finical institutes and many other businesses which have been utilizing the database for different storage and automation purpose. The scientists have been using these databases for performing number of experiments.

As the information storage is increasing day by day so, it is almost becoming impossible to extract the required information from these databases.

1 Department of Computer Science, Iqra National University Peshawar.
Email: Izhahmad335@Gmail.com; Izharahmad335@yahoo.com, Mobile: 0923078353564
A large number of the databases are linked with the online real time applications. The business intelligent (BI) tools and techniques have been developed and implemented to extract the required information in the storage machines. The enterprise applications for providing the unified processes to the organizations are divided into two main logical ends, which are the front end and back end. The back end of the application is considered as the core area of these unified solutions [1]. The database is deployed in the back end of the application. In the high end, real time application the database is divided into number of clusters. These database clusters are designed and developed for performing different tasks and operations and these tasks are divided into different categories such as failover cluster, database mirroring load balancing and many others. There are number of tools and techniques which are available for performing different operations on the tasks. However, still a huge demand of the new tools and techniques exist in the software and communication industries [2].

There are number of algorithms which are designed, developed and implemented using different tools, principles and techniques. These algorithms are utilized in practical application and software key turn solutions. Due to the large number of different algorithms available for performing the different tasks interlinked with the database it has become difficult to select the best one for the required application to get the maximum output from the proposed solution. The k-means algorithm is considered as one of the most popular, reliable and effective algorithm. The k-means algorithm and k-prototype algorithm are utilized in number of practical software and key turn solutions. Many parameters are considered on to measure the performance analysis. These parameters include “Guarantee, Scalability, Exhaustiveness, Linear Separators, Mixed data type” [3]. There are several researches which have been conducted in the field of database and data mining field. Most of these researches are performed for enhancing the k-means algorithm and k-prototype algorithm.

Both the K-Means algorithm and K-prototype algorithm have been utilized for creating the database cluster. The main goal of the development of the database cluster is to separate the abnormal cluster from normal cluster [4]. The data stored in the database can be of several different categories. The information stored within the database can be similar as well as different. This storage scheme depends on the design and implementation of the application solution [5].
The information stored in the database can be taken for the cluster. The retrieved information can be utilized for performing different operations and tasks. The main purpose of clustering is to group the related data. However, getting the required data from the cluster is the real challenge. There are number of algorithms which have been utilized for the database clustering. This research study has performed the comparison between the K-Means and K-Prototype algorithm to identify the limitations, processing capability, and advantages over different layer of database. This research is helpful to identify that k-means and k-prototype algorithm have any differences. The proposed solution in the research will be presenting the detail discussion of the k-means and k-prototype to recommend efficient algorithm for outlier detection and other issues relating to the database clustering. The verification and validation (V&V) of the proposed solution is based on the conceptual framework. There are different databases which can be utilized for performing the task. However, for the verification and validation of the conceptual framework, the Microsoft tool has been utilized for performing the analysis of K-Mean and K-Prototypes.

There are number of challenges which we have faced during the design and development of the proposed solution. The main challenge is to search or design a unified and reliable simulation tool which can be modified according to our requirements. The performance analysis should be based on many parameters which are reliable and we can achieve required results.

This research is design and implemented to analyse the performance of the K-Mean and K-Prototype algorithm. However, in the next sub section, we have presented few research objectives.

The Research Objectives are as followed:

a) The research study presents a detail discussion of the k-means and k-prototype to recommend efficient algorithm for outlier detection
a. There are number of parameters on which we have analysed our proposed solution
b) To evaluate the proposed solution based on the different experiments and techniques,
c) To analyse the daily entry of a particular data, in order to monitor the daily activities of k-means and k-prototype algorithms.
The advance computing technologies have helped us to store a large amount of data. Data mining algorithm are utilized for converting the data into knowledge information that can be used in future for performing different tasks. Different organizations have utilized this information for performing different type of analysis. The information retrieved for this algorithm can be utilized for daily entry of a particular data, in order to monitor the daily activities of k-means and k-prototype algorithms.

The rest of the paper is as followed:

In the next section, we have presented the related research considering the previous research conducted in the field of database clustering. In section III, we have presented the system architecture approach designed and developed for the K-Mean and K-Prototype algorithms performance analysis. In section VI, we have performed the evaluation of the algorithms and the last section presents the conclusion.

2. Related Research

The advancement in the information technology and signal processing has changed the way the business organization, financial institute and other type of organization used to simulate their business and functional processes. In today world, a large number of business organizations are utilizing the internet based application for the simulation of their functional and automation of their environments. Most of the internet applications are hosted on the cluster architecture with multi dedicated processing power for databases. There are number of algorithms which have been designed and developed for the application and database clustering. A large number of researches have been conducted at the academic level, the software industries in the domain of internet application, database clustering, database mirroring and data mining algorithms.

The internet applications are commonly used in many organizations. There are number of advantages of the internet applications. The internet applications are rich and have high performance and computation task can be simulated on these applications as compared to the traditional language. The web application or internet application are collected with the different database. As these application can be accessed online so, these applications have high number of users as compared to the desktop application [6]. Once deployed these applications are reliable, high scalable and require less maintenance.
The Web 2.0 architecture was able to simulate the internet application. The databases are the core components of these internet applications. However, mostly each internet application consists of many databases [7].

The K-mean algorithm is used for the clustering in number of applications. Most of the advancements have been made in the architecture of the K-mean algorithm. "The basic K-mean algorithm works only on numeric values prohibit it from being used to cluster real world data containing categorical values." [8]. the research presented by [9] performs the comparison of the K-mean algorithm. The comparison is based on the Large Data Sets with Categorical Values. The simulation is performed on the bases of mixed numerical and categorical attributes. The Clustering methods partition was initially presented by [10] that "a set of objects into clusters such that objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria" [2].

The verification and validation of the proposed study was based on the commonly used approaches of "soybean disease" and "credit approval". The data sets of these two techniques were demonstrated for the clustering performance of the algorithms.

The experiments were based on the real world value presented in the data sets. Half of the millions object are considered to show that the two algorithms are efficient when clustering the large data sets, which is critical to data mining applications which are designed and developed using internet technologies [11].

The research presented by [3] was simulated for the detection and identification of the six outlier detection algorithms. There are number of limitations which are identified in this research study and yet the simulation study was not demonstrated to that much extend.

Another research was presented by [12]. The main goal of the research was to identify the CLARA algorithm. The research also concluded with the presentation of the limitations of K-Prototype and CLARA algorithms. However, few differences were identified. The main difference included was that the CLARA algorithm uses the same data. However, on the other side, the K-Prototype algorithm considers and performs the operation on the while databases [13].
The database cluster can be set according to the requirements. The cluster can be of fixed number or they can be optimal.

“Simple way to initialize cluster and that pre-processing was important for outliers” [4]

The clustering of the data is based on different groups, when there is no knowledge about the data, data is created and processed then, and the data is shifted to the specific groups. There are large numbers of groups depending upon the data and it types. However, usually the clustering is performed on the numerical data. The supervised and unsupervised learning techniques have also been integrated in number of clustering algorithm.

The Sami supervised is also another technique for the clustering. The cluster’s minimum supervision can be required for performing the operations. The operations are usually performed by the clustering itself. “The difference between classification and clustering is that the clustering is unsupervised whereas classification is supervised”[14].

The K-Mode algorithms are also used for the clustering. These algorithms categorize the data for clustering. In the K-Mode algorithms grouping same type of data and classes are predefined and can be set with a large number of the database applications. In most of the applications, the database clusters are logically linked together for a logical network between them for the database operations like searching, sorting and mapping.

In most of researches it have been identified and observed that K-Means algorithm is very popular and the best technique for the clustering especially where the databases type are numeric data. There are number of improvement and advancement which have been made in the K-Means algorithm. The improved version includes K-Modes and K- Prototype algorithms [15].

There are number of research works which are identified for the utilization of the k-Means algorithm for outlier by applying the heuristic approach [16]. The grouping of data can be used in clustering algorithms however; the grouping of the data can be performed with the help of the clustering. “The K-Means algorithm works more efficiently. It divides different clusters with respect to dissimilarity.
Outlier is the observation of suspicion that is created by using different methods” [17]

In the changing paradigm of the computing, the cloud computing which has been adopted by the number of different type of organization. However, the clustering techniques are also been simulated and a large number of the projects are still undergoing in different house developments to implement a unified techniques for the cloud computing environment.

3. System Architecture

The conceptual framework for the system architecture for the K-Mean and K-Prototype Algorithms Performance Analysis are presented in this section. The proposed solution is based on the conceptual framework on which the K-Mean and K-Prototype Algorithms are simulated. There are three main logical components of the system architecture on which the complete system architecture is designed and developed. The presented system architecture is based on the open sources technologies so that it can be utilized for different database and dataset technologies. The performance analysis and its outcomes are based on the number of parameters such as the reliability and required results can be achieved.

The concept of the system architecture for the simulation, verification and validation of the proposed solution is based on the knowledge extracted for the previous research in the field of database cluster and its algorithm. The system architecture in this research is presenting the detail discussion of the k-means and k-prototype to recommend efficient algorithm for outlier detection and other issues relating to the database clustering. There are number of challenges which we have faced during the design, development and simulation of a unified performance analysis.
The figure 1 presents the system architecture design and development of the K-Mean and K-Prototype Algorithms Performance Analysis. There are three main logical components in the presented architecture. The data grid can be considered as database centre. The data grid constrains all the datasets on which the required clustering is performed. The logical network structure is deployed on the cluster so, that these clusters can be logically linked together. Once the data grids components are inter connected with the data grid. These data grid is then connected with tracking and monitoring application. The verification and validation (V&V) of the system is based on this component. This is the core component of the performance analyser. The administrator or the research will be monitoring all the activities, both the algorithms including (K-Mean and K-Prototype). The algorithm is written in the module of K-Mean and K-Prototype algorithms. There are number of parameter which is deployed for the verification of the results. All the operations are performed in this module. Once the cycles are complete the output is send to the research outcome. Where the results module display all the information of the research simulated with the help of the K-Mean and K-Prototype module.

The system architecture is based on the open source technologies so, that proposed performance analyser can be implemented for future researches. The performance analyser is considered as a unified architecture of the comparison and performance analyser for number of databases simulations.
The research study compares the K-Means and K-Prototype algorithm to identify the limitations and advantages of these algorithms. This research is helpful to identify that k-means and k-prototype algorithm have any difference.

The tracking and monitoring application also tracks and monitors the outlier detection of the algorithm. The module will detect the outlier issue and will remove. In case if the module fails or the tracking and monitoring application is not responds the alert will be send to the administrator. The administrator or researcher will have the right to restart the cycle or cancel the operations. The Data grid, and tracking and monitoring application will be connected with web services. The web service will be using the soap protocol. This tracking and monitoring applications will help the database administrator, software or application developer to select the algorithm which is better for the database clustering. The result outcome of this research study of the proposed solution will not be based on the signal parameter. The research and performance comparison will be based on the couple of parameters. The software application will be based on the open sources architecture thus; it can easily be integrated with other line of business application working with the electric systems.

3.1 System Feature

The proposed solution consists of number of system features. The tracking and monitoring application is based on the web based architecture and can be accessed from anywhere with the help of the internet connectivity. The research doesn’t have to re program or develop a complete solution for testing of the datasets. The proposed solution doesn’t require any high performing hardware. The customized data grid can be provided from the user to the application so that the required task can be performed. The result can be saved and shared with the other researchers with the help of electronic mail. The information can be saved in the database and can be used for future tasks and activities.

3.2 Data Transmission

The data transmission between the logical components is presented. Without the proper transmission the component which are interconnected with each other will not be about to process the required result. The data grid will be connected with the web service.
The web service will upload all the information to the tracking and monitoring application. The tracking and monitoring application will have local database which will store all the information of the user and their groups including the privileges. After the analyses are performed, the information is transferred to the result outcome.

The results outcome is interlinked with the tracking and monitoring application. The alerting sub system in programmed within the tracking and monitoring application so that in case of an error the pop is generated and the data transmission is stopped and the information is transferred to the admin to perform the required task so, that the specific operation can be completed. In the presented system architecture the data flow between the three logical components. However, it depends upon the environment consisting of different protocols that can be utilized for performing communication and coordination tasks.
# Table 1: Steps of Data Transmissions

<table>
<thead>
<tr>
<th>Steps</th>
<th>Description</th>
<th>Components Involved within transmission</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data Transmission from Cluster (c) to Cluster (c)</td>
<td>Cluster</td>
</tr>
<tr>
<td>2</td>
<td>Data Transmit</td>
<td>Cluster to Cluster</td>
</tr>
<tr>
<td>3</td>
<td>Data Transmission from Cluster (c) to Data Grid (c)</td>
<td>Cluster</td>
</tr>
<tr>
<td>4</td>
<td>Data Transmit</td>
<td>Cluster to Data Grid</td>
</tr>
<tr>
<td>5</td>
<td>Data Transmission from Data Grid (DG) to Web Services (WS)</td>
<td>Data Grid</td>
</tr>
<tr>
<td>6</td>
<td>Data Transmit</td>
<td>Data Grid to Web Services</td>
</tr>
<tr>
<td>7</td>
<td>Data Transmission from Web Services (WS) to Tracking and Monitoring Application (TMA)</td>
<td>Web Services</td>
</tr>
<tr>
<td>8</td>
<td>Data Transmit</td>
<td>Web Services to Tracking and Monitoring Application</td>
</tr>
<tr>
<td>9</td>
<td>Data Transmission from Tracking and Monitoring Application (TMA) to Results outcome (RO)</td>
<td>Tracking and Monitoring Application</td>
</tr>
<tr>
<td>10</td>
<td>Data Transmit</td>
<td>Tracking and Monitoring Application to Results Outcome</td>
</tr>
</tbody>
</table>
To test the accuracy and efficiency of the proposed system architecture, the evaluations is performed. We have simulated the same dataset with different K-Mean and K-Prototype Algorithms. The main input dataset includes “E-Coli [14], Breast Cancer-Wisconsin [15] and Thyroid [16]”. The same sets of data are given as input to the original k-means algorithm and k-prototype algorithm for the verification and validation of the proposed solution.

4. Evaluation

The standard k-means algorithms and the k-prototype algorithms require the value of the centroid and K value as the input parameters. The value of the centroid and K are chosen randomly. Different experiments are conducted for the verification and validation of the proposed system architectures. The proposed system architectures is designed and developed for the Performance Analysis required to the data set and value of k. “The accuracy of clustering is determined by comparing the clusters obtained by the experiments with the already available pre-determined clusters in the UCI data set, the percentage accuracy and the time required for each experiment are calculated” [18].

![Proposed System Architecture Process](image)

**Figure 2: System Architecture Steps**
The figure 2 explains the simulation of the experiment. We have explained the simulation of the experiment on the proposed system architecture by mapping each component of the experiment on the logical layer of the proposed system architecture. The dataset is connected with three different dataset value including “E-Coli [17], Breast Cancer-Wisconsin [19] and Thyroid [20]” the set retrieved this value from these data and passed these values to the network. The secure internet connection is required to simulate the complete solution as the proposed system components works and simulates on the remote application and database clusters. Once the data is collected, the data is transferred to the web services. The services utilized the SOAP protocol and transfer this data to the tracking and monitoring application. The algorithm receives and processes the data received from the web service. The standard k-means algorithms and the k-prototype algorithms require the value of the centroid and K value as the input parameters. The parameter setting is performed by the administrator or the research, if the algorithm was executed and the result are received, the results are then passed to the performance analysis. Similarly, if the tracking and monitoring application was unable to process the result, then the error is generated and the data re-processing requests are generated through the web service. However, if the results are executed on the provided dataset, the information is send to the last logical module. This result can also be send to the email for records. The results can also be re-executed in case the research is not satisfied with the results or wants to regenerate the results for verification and validation (V&V).

![Figure 3: Performance Analysis Results](image)

We have conducted large number of experiments for the verification and validation (V&V) of the proposed solution. There are more than 50 experiments which have been conducted for the performance evaluation of the proposed solution. Number of experiments is considered on the x axis and the dataset is considered on the y axis.
It is analysed that the proposed solution presents the best output on the k-Means Algorithms and K-Prototype algorithm as the dataset increases, the performances of the proposed solution on both the algorithm increases and on the other hand as number of experiment increases the general k-Mean and k-Prototype algorithms performance decreases at many stages.

5. Conclusion

There are number of clustering algorithms which are widely used for performing the clustering on the large data. The original algorithm of the (K-Mean and K-Prototype) does not always guarantee the accuracy of the final clusters based on the selection of initial centroids. However, the proposed system architecture utilized the complete unified solution for the K-Mean and K-Prototype Algorithms Performance Analysis.

This proposed solution method ensures that the initial centroids and k values are generated depend on the distribution of the data received from the data grid. These proposed system architecture presents the better accuracy compared to the original k-means and k-prototype algorithm. The analysis have shown that the proposed system architecture produces better clusters in less computation time as compared to the standard k-means and k-prototype algorithm.

References

Tim O’Reilly, What is web 2.0, 2005.


