IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Proposed System for Data Mining Using Clustering

Pratik Sohala[1], Harshil Patel[2], Aniket Patel[3], Hezal Lopes[4]

Department of Computer, UCOE, Vasai, Kaman, Maharashtra[1-4]

Abstract: In present scenario most of the service centres manage their accounts through manual process i.e. paper work which is very tedious to create job cards, no record of parts and details of customers but this application will help in computerizing entire service centre system using software application. This proposed system is useful for any service centre for managing information of availability of spare parts, customer's information, managing accounts. This application can be useful for private service centres and franchise service centre. In service centre first job is to make job card so work can be done according to it, managing spare parts is a important task where details of spare parts is accurately managed i.e. part number, part type ,price, when it was purchased and from which retailer had purchased, even maintain records of all of customers is very important as per business perspective to grow business there is need to maintain record of each and every customer i.e. what was last service date ,what work was done in last servicing what issue customer is facing now ,what parts we used while serving their bikes ,and generating bills according to work .it also shows list of customers whose last service exceed more than 3 months as to contact them and send them reminder which helps to grow business and maintain customer relationships as details of parts changes over time there will be another panel where you can edit if details changes which makes it more flexible. There will be another panel for the Data Representation with the help of clustering algorithm for Data Mining purposes as the data will grow exponential as the proposed system will be used in the real world. Thus seeing the limitation of the current system we propose the new computerized system.

Keywords: Data Mining, Data Analysis, Clustering.

I. INTRODUCTION

Currently work is done manually i.e.by writing everything on job card or paper, which stores no data and there is no record of customers or any part of used of servicing or any kind of details. This project is to ease the work of service center manager which will store each and every details of customer and work done with storing and retrieving records which makes work easier getting everything on your fingertips. Clustering is useful in several explanatory pattern-analysis, grouping, decision-making, and machine learning situations, including data mining, document retrieval, image segmentation and pattern classification. Typical pattern clustering activity involves the following steps: Pattern representation, Clustering or grouping, Data abstraction (if needed), Assessment of output (if needed).

II. LITERATURE SURVEY

An efficient association rule based hierarchical algorithm for test clustering.The paper has been improved with the help of modifications done in hierarchical clustering based on association rules. The following paper is been used to perform clustering on text using apriority algorithm, proposed methodology works in following manner texts is been preprocessed. Preprocessing involves two techniques viz stop word removal and stemming after preprocessing feature selection method occurs then association rule miner uses apriority to find rules between text documents. After association is been found the associated documents are given as input to hierarchical algorithm it performs agglomerative clustering and results are obtained. Better results as compared to previous methods. Quality of clusters is enhanced. Speed is been increased due to related documents are only used. Can only be applied if hierarchical structure is present Scalability is less. [1]Comparative Analysis of K means Clustering Sequentially and Parallely. Clustering can be performed in 2 ways Distance based clustering and conceptual clustering. There are many ways using which clustering can be performed but here k means algorithm is been used. The paper focuses on analysis performed using sequentially and parallely the test was performed on common dataset and it was observed that parallel clustering gave a better o/p as compared to sequential because it was time consuming and no of iterations also taken where more in comparison to parallel clustering.Parallel clustering saves time. Parallel clustering may lead into more load if dataset is large. [2]An Algorithm for Spatial Data Mining using Clustering.There are many spatial data mining technique available PAM (Partitioning around Medoids), CLARA (Clustering Large Applications), CLARANS (Clustering Large Applications Based on Randomized Search), DBSCAN (Density Based Spatial Clustering of Applications with Noise), DBCLASD (Distribution Based Clustering of Large Spatial Databases), STING(Statistical Information Grid-based method), BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), Wave Cluster, DENCLUE (Density basted Clustering), CLIQUE, named for Clustering In Quest, CURE (Clustering Using Representatives) the data is been recorded from above techniques and is been compared with the modifications in existing k-means algorithm named as

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Incremental K-Means Algorithm and after observing the results it reduces the no of iterations and results into robust system. Clusters are created doesn't include outliers. Robust to any no of instances. Can only be performed on single dataset. Efficiency is less due to sequential execution. [3]Review on determining number of Cluster in K-Means Clustering. K-means is most popular technique used for clustering but the major question arises to provide the no of clusters to be formed, that can be constraint because directly providing any no won't be a feasible solution and to provide the no it requires a large study on the dataset to provide the no that produces a better result because different no of clusters changes results drastically. The paper provides 6 methods using which ideal no. of clusters can be determined that provide a better result: A. By rule of thumb, B. Elbow method, C. Information Criterion Approach, D. An Information Theoretic Approach, E. Choosing k Using the Silhouette, F. Cross-validation. [4]

III. PROPOSED SYSTEM

This project is to ease the work of employee or manager. Project contains following modules job card, billing, inventory management, and customer records etc. Proposed system stores each and every customer details including its servicing date, parts used on that date, and give reminder to manger if that customer last servicing date exceeds more than 3 months. It has inventory management of parts containing part number, price, part name and from which retailer it had been purchased. It has pre-defined details which can updated afterwards.

ALGORITHMS

The Microsoft Clustering algorithm is a segmentation or clustering algorithm that iterates over cases in a dataset to group them into clusters that contain similar characteristics. These groupings are useful for exploring data, identifying anomalies in the data, and creating predictions. The clustering algorithm differs from other data mining algorithms, such as the Microsoft Decision Trees algorithm, in that you do not have to designate a predictable column to be able to build a clustering model. The clustering algorithm trains the model strictly from the relationships that exist in the data and from the clusters that the algorithm identifies.

Eg: Consider a group of people who share similar demographic information and who buy similar products from the Adventure Works Company. This group of people represents a cluster of data. Several such clusters may exist in a database. By observing the columns that make up a cluster, you can more clearly see how records in a dataset are related to one another.

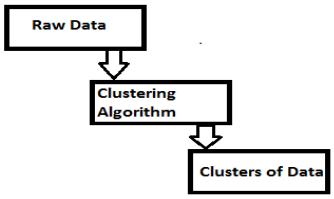

Figure 1: Clustering of Data

The figure 1 shows the data of clustering in the initial stage the raw data is been collected from the database, this raw data is used to see that how much amount of work is done in the service center and how many customers are being served in a week, month or a year. Then after collecting the data the data is sent in the clustering algorithm and the algorithm is used to divide the data in small parts of clusters that it should be easy for the user to see the data or filter the data according to the user requirements. Then these clusters of data are used to see the whole data in small parts of clusters and this data can be useful for the user to see data in easy way.

Following flowchart shows the representation of how the clustering works. In initial stage the user have to input the data in the database. Then the algorithm reads the data and finds the data according to the user defined parameters. This given parameters is then defined as cluster center. Then it will set the initial cluster center randomly. After setting the cluster center randomly the data which matches the parameters are put in the cluster center. Then it will calculate new clusters if any new data can be represented by iteration i.e.it can increase the number of clusters. When the number of clusters is fixed the data is ready to be moved to one of the clusters based on the smallest distance. Then the data is moved to the clusters. Then if the data is moved correctly then the output will be generated else the whole process will start again.

The Microsoft Clustering algorithm first identifies relationships in a dataset and generates a series of clusters based on those relationships. A scatter plot is a useful way to visually represent how the algorithm groups data, as shown in the following diagram. The scatter plot represents all the cases in the dataset, and each case is a point on the graph. The clusters group points on the graph and illustrate the relationships that the algorithm identifies.

After first defining the clusters, the algorithm calculates how well the clusters represent groupings of the points, and then tries to redefine the groupings to create clusters that better represent the data. The algorithm iterates through this process until it cannot improve the results more by redefining the clusters.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

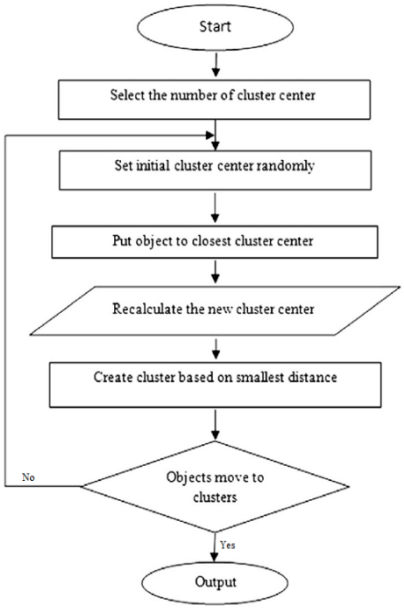

Figure 2: Flow chart of clustering

Data Required for Clustering Models

When you prepare data for use in training a clustering model, you should understand the requirements for the particular algorithm, including how much data is needed, and how the data is used. The requirements for a clustering model are as follows:

A single key column: Each model must contain one numeric or text column that uniquely identifies each record. Compound keys are not allowed.

Input columns: Each model must contain at least one input column that contains the values that are used to build the clusters. You can have as many input columns as you want, but depending on the number of values in each column, the addition of extra columns can increase the time it takes to train the model.

Optional predictable column: The algorithm does not need a predictable column to build the model, but you can add a predictable column of almost any data type. The values of the predictable column can be treated as input to the clustering model, or you can specify that it be used for prediction only. For example, if you want to predict customer income by clustering on demographics such as region or age, you would specify income as predict only and add all the other columns, such as region or age, as inputs.

Implementation of the Microsoft Clustering Algorithm

The Microsoft Clustering algorithm provides two methods for creating clusters and assigning data points to the clusters. The first, the *K-means* algorithm, is a hard clustering method. This means that a data point can belong to only one cluster, and that a single probability is calculated for the membership of each data point in that cluster.

K-Means Clustering

K-means clustering is a well-known method of assigning cluster membership by minimizing the differences among items in a cluster while maximizing the distance between clusters. The "means" in k-means refers to the centroid of the cluster, which is a data point that is chosen arbitrarily and then refined iteratively until it represents the true mean of all data points in the cluster. The "k" refers to an arbitrary number of points that are used to seed the clustering process. The k-means algorithm calculates the squared Euclidean distances between data records in a cluster and the vector that represents the cluster mean, and converges on a final set of k clusters when that sum reaches its minimum value. The k-means algorithm assigns each data point to exactly one cluster, and does not allow for uncertainty in membership. Membership in a cluster is expressed as a distance from the centroid.

IV. RESULTS AND DISCUSSIONS

The figure 3 shows the Dashboard of the project which indicates the starting of the project or it shows the main page of the project. The actual project starts from this page where the user can do all the database work and store and manage the data of the customer in the database.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

Figure 3: Dashboard of the project

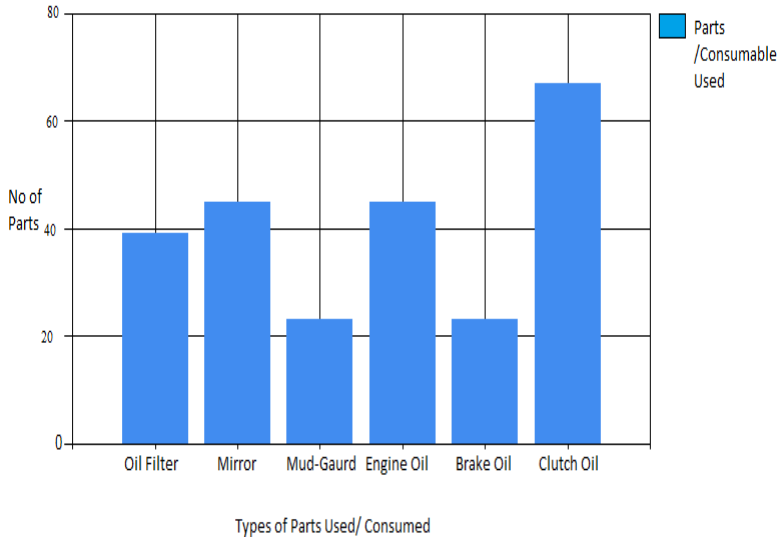

Figure 4: Graphical representation for the parts which are used or consumed

The figure 4 represents number of parts used in the month. The x-axis shows the types of parts and the y-axis shows the number of parts. The types of parts used are oil filter, mirror, mud-guard, engine oil, brake oil, clutch oil.

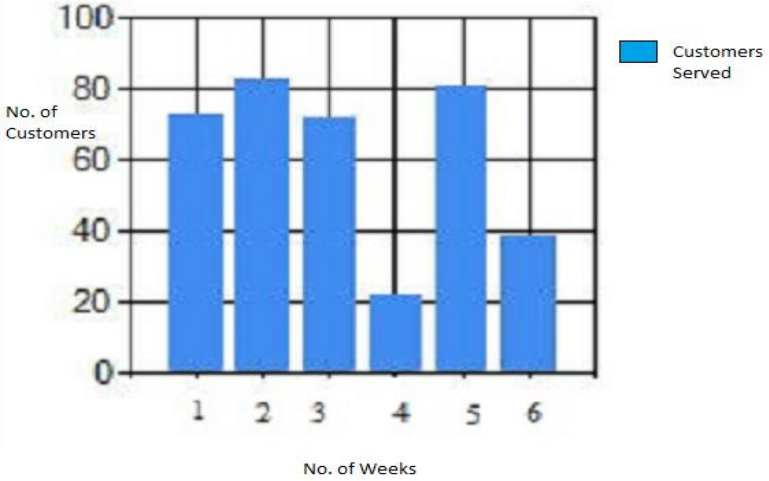

Figure 5: Graphical representation for the customers which are served in weeks

The figure 5 shows the number of customers served in the following weeks. The x-axis represents the weeks of the graph and the y-axis represents the number of customers served within the following weeks.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 3, March 2018

VI. CONCLUSION

As per the customer requirements we are developing this system because the workflow is done in an automated system instead of manually doing the record work. After some time we can generate report according to data in the system via Data mining algorithm. We can implement distributed environment and SMS alert service for customer. This project is to ease the work of employee or manager. Project contains following modules job card, billing, inventory management, customer records, etc. Proposed system stores each and every customer details including its servicing date, parts used on that date, and give reminder to manger if that customer last servicing date exceeds more than 3 months. It has inventory management of parts containing part number, price, part name, and from which retailer it had been purchased. It has pre-defined details which can be updated afterwards.

REFERENCES

[1] J. Dafni Rose, An efficient association rule based hierarchical algorithm for test clustering, in St. Joseph's Institute of Technology, Chennai, India, Jan-March 2016.
[2] Kavya D S, Chaitra D Desai, Comparative Analysis of K means Clustering Sequentially and Parallely, in REVA ITM, Bangalore, India, APR 2016.
[3] Prof. A. M. Karandikar, Karishma Vaswani, An Algorithm for Spatial Data Mining using Clustering, in Shri Ramdeobaba college of engineering and management, Nagpur, India, August 2017.
[4] Trupti M. Kodinariya, Dr. Prashant R. Makwana. Review on determining number of Cluster in K-Means Clustering, in Research Scholar in JJT University, Jhunjhunu, Rajasthan – India, GRMECT Research Center Rajkot – India. November 2013.