Implementation of clustering algorithm k mean k medoid computer science essay

Technology, Computer

Data Mining is a fairly recent and contemporary topic in computing. However, Data Mining applies many older computational techniques from statistics, machine learning and pattern recognition. This paper explores two most popular clustering techniques are the k-means& k-medoids clustering algorithm. However, k-means algorithm is cluster or to group your objects based on attributes into K number of group and k-medoids is a related to the K-means algorithm. These algorithms are based on the k partition algorithms and both attempt to minimize squared error. In contrast to the K-means algorithm K-medoids chooses data points as centres. The algorithms have been developed in Java, for integration with Weka Machine Learning Software. The algorithms have been run with two dataset Facial palsy and Stemming. It is having been shown that the algorithm is generally faster and more accurate than other clustering algorithms.

Data Mining derives its name from the similarities between searching for valuable business information in a large database (for example, finding linked products in gigabytes of store scanner data) and mining a mountain for a vein of valuable ore.[1] Both process requires either sifting through an immense amount of material. Or intelligently probing it to find exactly where the value resides.

Data Mining

Data mining is also known as “ knowledge mining”. Before it was named “ DATA MINING”, it was called “ data collection”, data warehousing” or “ data access”. Data mining tools predicts the behaviours of the models that are loaded in the data mining tools (like Weka) for analysis, allowing making

predicted analysis, of the model. Data mining provides hands-on and practical information.

Data mining is the most powerful tool available now. Data mining can be used for modelling in fields such as artificial intelligence, and neural network.

What does it do?

Data mining take the data which exists in unrelated patterns and designs, and uses this data to predict information which can be compared in terms of statistical and graphical results. Data mining distil / filters the information from the data that is inputted and final model is generated.

Clustering

“ What is cluster analysis? “ Unlike classification and prediction, which analyse class-labeled data objects, clustering analyses data objects without consulting a known class label.

A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters. Each cluster “ center” is marked with a “+”.[6]

Clustering is the technique by which like objects are grouped together. The objects are clustered or grouped based on the principle of maximizing the intra class similarity and minimizing the interclass similarity. i. e. clusters of the objects are made so that the clusters have resemblance in comparison to one another, but are very divergent to objects in other clusters. Each cluster

that is made can be viewed as a class of objects, from which rules can be derived. [6]

Problem overview

The problem at hand is able to correctly cluster a facial palsy dataset which is given by our lecturer. This section will provide an overview of dataset being analysed, and description about dataset that we use in this implementation.

Data Set

1. 3. 1. 1 Facial_Palsy_svmlight_format

Facial Palsy data is for binary classification.

+1 severe facial palsy faces

-1 Non-severe or normal faces

66 Principal components generated from 50? 50 Hamming distance images

1. 3. 1. 2 A6_df2_stemming__svm:

Attributes: 100

A6_df2_stemming__svm_100. dat

+1 Open question

-1 Closed question

Section 2 – Methodology

This section will firstly discuss the methodology behind K-means & k-medoids algorithm. It is than followed by steps to implement k-means and k medoids algorithms. How many input, output and what are the steps to perform k-means and k-medoids.

2. 1 K-mean

K-means clustering starts with a single cluster in the centre, as the mean of the data. Here after the cluster is split into 2 clusters and the mean of the new cluster are iteratively trained. Again these clusters are split and the process goes on until the specified numbers of the cluster are obtained. If the specified number of cluster is not a power of two, then the nearest power of two above the number specified is selected and then the least important clusters are removed and the remaining clusters are again iteratively trained to get the final clusters. If the user specifies the random start, random cluster is generated by the algorithm, and it goes ahead by fitting the data points into these clusters. This process is repeated many times in loops, for as many random numbers the user chooses or specifies and the best value is found at the end. The output values are displayed.

The drawbacks of the clustering method are that, the measurement of the errors or the uncertainty is ignored associated with the data.

Algorithm: The k-means algorithm for partitioning, where each cluster’s centre is represented by the mean value of the objects in the cluster.

Input:

k: the number of clusters,

D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) Arbitrarily choose k objects from D as the initial cluster centers;

(2) Repeat

(3) reassign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) Update the cluster means, i. e., calculate the mean value of the objects for each cluster;

(5) Until no change;

Where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and mi is the mean of cluster Ci (both p and mi are multidimensional). In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible.[2]

Clustering of a set of objects based on the k-means method. (The mean of each cluster is marked by a “+”.)

2. 2 K- Medoids

This report recommends a new algorithm for K-medoids, which runs like the K-means algorithm. The algorithm proposed scans and calculates distance matrix, and use it for finding new medoids at every constant and repetitive step. The evaluation is based on real and artificial data and is compared with the results of the other algorithms.

Here we are discussing the approach on k- medoids clustering, using the k-medoids algorithm. The algorithm is to be implemented on the dataset which consist of uncertain data. K-medoids are implemented because they to represent the centrally located objects called medoids in a cluster. Here the k-medoids algorithm is used to find the representative objects called the medoids in the dataset.

Algorithm: k-medoids. PAM, a k-medoids algorithm for partitioning based on medoids or central objects.

Input:

k: the number of clusters,

D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) Arbitrarily choose k objects in D as the initial representative objects or seeds;

(2) Repeat

(3) Assign each remaining object to the cluster with the nearest representative object;

(4) Randomly select a no representative object, o random;

(5) Compute the total cost, S, of swapping representative object, oj, with o random;

(6) If S < 0 then swap oj with o random to form the new set of k representative objects;

(7) Until no change;

Where E is the sum of the absolute error for all objects in the data set; p is the point in space representing a given object in cluster Cj; and oj is the representative object of Cj. In general, the algorithm iterates until, eventually, each representative object is actually the medoids, or most centrally located object, of its cluster. This is the basis of the k-medoids method for grouping n objects into k clusters.[6]

2. 3 Distance Matrix

An important step in most clustering is to select a distance measure, which will determine how the similarity of two elements is calculated.

Common distance metrics:

Euclidean

Manhattan

Minkowski

Hamming etcaˆ¦

Here in our implementation we choose two distance matrix that you can see below with description.

2. 3. 1 Euclidean Distance Metric

The Euclidean distance between point’s p and q is the length of the line segment. In Cartesian coordinates, if p = (p1, p2… pn) and q = (q1, q2… qn) are two points in Euclidean n-space, then the distance from p to q is given by:

2. 3. 2 Manhattan Distance Metric

The Manhattan (or taxicab) distance, d1, between two vectors in an n-dimensional real vector space with fixed Cartesian coordinate system, is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes.

Section 3 – Discussion

In this section we are discussing about how Weka Machine learning work and how we implemented both k-means and k medoids algorithm. To implement these two algorithms we use Java and we are explaining how we implemented in java which function we use in order to implement these two algorithms.

3. 1 Weka Machine Learning

Weka is a machine learning software made using Java and many other languages. Weka has a collection of tools that are used to analyse the data that the user inputs in the form of dataset files. Weka supports more than four different input data formats. Weka uses an interactive GUI interface, which is easy for the user to use. Weka provides the functionality for testing and visual aid options that can be used by the user to compare and sort the results.

3. 2 Implementation

In this section, we discuss about implementation of 2 clustering algorithms: K-Means and K-Medoids. Here, we use Object Oriented Programming to implement these 2 algorithms. The structure of program as below:

There are 3 packages: K-Mean, K-Medoid, main.

Files in K-Mean package:

Centroid. java

Cluster. java

KMean_Algorithm. java

KMean_Test. java

KMean_UnitTest. java

Files in K-Medoid package:

KMedoid_Algorithm. java

KMedoid_UnitTest. java

Files in main package:

Attribute. java

DataPoint. java

DistanceCalculation. java

FileFilter. java

MainFrame. java

Utilities. jav

There are some main functions implemented for clustering activity as below:

3. 2. 1 read_SVMLightFile_fill_up_missing_attribute()

This function is about reading the SVM Light data file (. dat) and fill up all the missing attributes/values in data file before returning a Vector of data-points for clustering activity.

3. 2. 2 calculate_distance()

This function is providing calculation according to the distance metric input in order to calculate distance between data objects for clustering activity. Overall, this function provides calculation for 3 different distance metrics as: Euclidean, Manhattan and Minkowski.

3. 2. 3 startClustering()

This function is about running a particular clustering algorithm and returns a Vector of Clusters with their own data-points inside. All the steps of a particular clustering algorithm is implemented, here we implement K_Means and K_Medoids clustering algorithms.

3. 2. 4 calculateSumOfSquareError()

This function is about calculating the total/sum square error for all the output clusters. By calling the function “ calculateSquareError()” inside every cluster and sum up, the sum of Square Error will be calculated as long as the clustering activity finished.

3. 2. 5 calculateSumOfAbsoluteError()

This function is about calculating the total/sum absolute error for all the output clusters. By calling the function “ calculateAbsoluteError()” inside

every cluster and sum up, the sum of Absolute Error will be calculated as long as the clustering activity finished.

3. 2. 6 toString() and main()

The toString() function will return a string which represents the clustering output, including: total objects of every cluster, percent of object in every cluster, the error (such as: sum of square error or sum of absolute error), the centroid of every cluster and all the data-points clustered in the clusters.

The main() function inside MainFrame. java class will allow to execute the GUI of the program, so users can interact with system by GUI instead of console or command-line. In this GUI, users can choose type of distance metric (such as Euclidean and Manhattan), Clustering algorithm (such as K-Means and K-Medoids) and enter input parameters such as number of clusters and number of iterations for clustering activity. Besides, users also can open any data file to view or modify and save before running clustering as well as export the original data file with missing attributes/values to new processed data file with all missing values filled up by zero (0).

Section 4 – Analysis

In order to access the performance of the K-means & k-medoids clusters, two dataset of analyses was carried out. The aim of this set to tests was provide an indicator as to how well the clusters performed using the k-means and k-medoids function. The tests were involved comparing the cluster to other

cluster of various types provided within Weka cluster suite. The results are summarised throughout the remainder of this section.

4. 1 Experiment (Facial Palsy dataset) results vs. Weka

Here In this section how we did a comparison with our application algorithm vs. Weka you can see below.

In this pattern we give iterations when we run a dataset with our application and Weka.

Iterations: 10 >> 30 >> 50 >> 100 >> 200 >> 300 >> 400 >> 500

In this pattern we give a cluster when we run a dataset with our application and Weka.

Clusters: 2 >> 3 >> 4 >> 5

After we run dataset with this format than each and every run we get result we combine that result, compare with Weka, we make a total of each and every column and come with average and we are displaying in table that you can see in below table.

This Symbol is object. To see a result please click on this object it will show you result. We put as object because result is too big in size so we are not able to put in this A4 page.

4. 2 Experiment (Stemming Question dataset) results vs. Weka

Here In this section how we did a comparison with our application algorithm vs. Weka you can see below.

In this pattern we give iterations when we run a dataset with our application and Weka.

Iterations: 10 >> 30 >> 50 >> 100 >> 200 >> 300 >> 400 >> 500

In this pattern we give a cluster when we run a dataset with our application and Weka.

Clusters: 2 >> 3 >> 4 >> 5

After we run dataset with this format than each and every run we get result we combine that result, compare with Weka, we make a total of each and every column and come with average and we are displaying in table that you can see in below table.

This Symbol is object. To see a result please click on this object it will show you result. We put as object because result is too big in size so we are not able to put in this A4 page.

Section 5 – Conclusion

In evaluating the performance of data mining techniques, in addition to predicative accuracy, some researchers have been done the importance of the explanatory nature of models and the need to reveal patterns that are valid, novel, useful and may be most importantly understandable and

explainable. The K-means and k-medoids clusters achieved this by successfully clustering with facial palsy dataset.

“ Which method is more robust-k-means or k-medoids?” The k-medoids method is more robust than k-means in the presence of noise and outliers, because a medoids is less influenced by outliers or other extreme values than a mean. However, its processing is more costly than the k-means method. Both methods require the user to specify k, the number of clusters.

Aside from using the mean or the medoids as a measure of cluster center, other alternative measures are also commonly used in partitioning clustering methods. The median can be used, resulting in the k-median method, where the median or “ middle value” is taken for each ordered attribute. Alternatively, in the k-modes method, the most frequent value for each attribute is used.

5. 1 Future Work

The K-means algorithm can create some in efficiency as; it scans the dataset leaving some noise and outliners. These small flaws can be considered major to some of the users, but this doesn’t means that the implementation can be prevented. It is always possible that sometimes the dataset is more efficient to follow other algorithms more efficiently, and the result distribution can be equal or acceptable. It is always advisable to make the dataset more efficient by removing unwanted attributes and more meaning full by pre-processing the nominal values to the numeric values.

5. 2 Summery

Throughout this report the k-mean and the k-medoids algorithms are implemented, which find the best result by scanning the dataset and creating clusters. The algorithm was developed using Java API and more Java classes.