International Innovative Research Journal of Engineering and Technology
ISSN: 2456-1983 Vol:4 No: 1 September 2018

To Propose Prediction Analysis Algorithm based on k-means and SVM Classification

Aliza Sarlan

Software Quality and Quality Assurance (SQ2E) Research Cluster, UniversitiTeknologi PETRONAS

Abstract: The data mining is the technique which can extract useful information from the rough data. The clustering is the approach which can group similar and dissimilar type of data. The prediction analysis is the technique which can predict new values from the existing techniques. The prediction analysis contains two phases, in the first phase k-mean clustering is applied which can cluster similar and dissimilar type of data. In the second phase, the SVM classifier is applied which can classify similar and dissimilar type of data.

Keywords: Multiclass SVM, K-means, prediction analysis, linear SVM

1. INTRODUCTION

The process through which important information and patterns can be extracted from huge databases is called data mining. It is also referred as knowledge discovery process, knowledge extraction process or knowledge data mining system. Various data mining tools are used in order to analyze the data. With the help of management of data and the database, the numerous functionalities are also used which help in performing proper analysis of such huge amount of data [1]. However, due to the immense growth of data with the passage of time, new methods have been developed to store and retrieve this data in efficient manner from huge databases. These methods include query and transaction processing mechanisms within them as well. The evolution of data repository architecture is referred to as data warehouse. In order to define the unified methods, numerous heterogeneous data sources are organized. This helps in providing proper decision making with single site. Many other functionalities like summarization, consolidation and aggregation are also provided by this mechanism [2]. Ensuring the storage of data in efficient manner within the databases, data warehouses and other repositories can be done with the help of numerous interesting patterns. An unsupervised type of classification method is referred to as data clustering mechanism which majorly groups or clusters the objects on the basis of their similarities with each other. The objects which have different properties are clustered separately and the objects which are similar are clustered within one cluster. The analysis of clusters is very important within the research area of data mining. It is the initial step towards the knowledge discovery process. With the help of clustering mechanism, the data objects are grouped in the set of disjoint classes [3]. The objects which are added within similar class are put together and the objects which are highly dissimilar are put in different classes. The classification of huge data can be pre-classified with the help of most common data mining method which is known as classification. The pre-classification examples are used here from which the complete data can be classified further [4]. The decision tree or neural network based classification algorithms are utilized within this method. The learning and classification methods are performed within data classification. The classification algorithm is used for training the data is learning mechanisms. The accuracy of classification rules can be estimated with the help of test data present within the classification. The rules are applied to new data tuples if there is acceptable form of accuracy. On the basis of record-by-record method, the complete records of both fraud and valid activities are involved within the fraud detection methods. In order to determine the set of parameters which are needed to provide proper discrimination, the classifier-training algorithm is utilized [5]. The predictions of future can be determined only with the help of collected data from the statistical techniques of

predictive modeling, data mining as well as machine learning techniques. The technique which provides such predictions is known as predictive analytics technique. The various risks and opportunities can be identified within the business applications also with the help of prediction analysis. Within numerous applications such as traveling, finance, marketing and health care applications, the prediction analysis techniques are utilized. The SVM classifier is used within this method in order to provide regression, classification and general pattern recognition of the data. The SVM classifier performs better than the other classifiers as it has high generalization performance and does not need any prior knowledge to be added within it. The determination of best function can be done with the help of SVM which can be done by maximizing the margin available within two different classes due to the presence of various linear hyperplanes within these classes [6]. Hyperplane is mainly referred to as the amount of space or distance which exists between the two different classes. Margin is referred to as the shortest distance in between the closest data points to the point present on hyperplane. A recursive system is followed within the K-means clustering algorithm. The identification of user-specified number of clusters (k) can be done with a prototype based partitional clustering method which is known as k-means. In this algorithm, n numbers of observations are to be divided into k clusters. With the nearest mean, each cluster has a position within the cluster due to which is serves as a prototype of that cluster [7]. The Voronoi cells are thus formed due to partitioning of the complete data space. This method is very much similar to the k-nearest neighbor classifier which is also a famous machine learning mechanism that is utilized for organization. In k-means algorithm, each point is assigned to the nearest present centroid. A cluster is formed with the collection of points that are assigned to that centroid.

2. LITERATURE REVIEW

Vadlana Baby, et.al, (2016) proposed in this paper [8], an efficient distributed threshold privacy-preserving k-means clustering algorithm. In this algorithm, the privacy preserving method known as code based threshold secret sharing mechanism is used. In comparison to the existing protocols, the proposed protocol includes very less number of iterations. The comparisons are made in this paper in order to analyze the performance of this proposed mechanism. As per the evaluated experimental simulations, it is seen that there is no need of trust amongst the servers or users with the application of this algorithm. Thus, an efficient privacy preserving mechanism is provided here in order to preserve the data of client in proper way.

Vaibhav Kumar, et.al, (2016) presented in this paper [9], the k-means clustering based unsupervised learning method which helps in enhancing the cooperative spectrum sensing performance which occurs within the k- μ fading channels. In order to characterize the receiver operating characteristics, various system parameters within characterization are analyzed through different simulations. There is an up gradation seen in the performance of learning based methods with the application of ROC in comparison to the classical energy detection based CSS methods. On the basis of results achieved it is seen that the performance of k=2 is the best amongst the all the other values applied.

R. Kumari, Sheetanshu, et.al, (2016) presented in this paper [10], the utilization of simplistic Euclidean distance based method which was found within the Mlib library. The distance between the various features of data set can be related with the help of various distance methods applied within this technology. The k-means iterative model is further substituted with DBSCAN or Gaussian mixture model which might help in enhancing the overall performance. The system intrusion thus does not include the clustering and identification of the abnormalities which is seen amongst the results achieved as well. Thus, the various features or criteria of the data can be studied in proper manner while extending this research work in future.

Kaustubh S. Chaturbhuj, et.al (2016) presented in this paper [11] that the management of such immense amount of data growing every day is not possible with the help of traditional data processing mechanisms. In order to process such data, Hadoop has been used. The initial centroids were identified with the help of PSO. Further, the clusters were to be identified by using k-means algorithm. The utilization of this method has provided enhanced results during its implementation. The large data parallel can be handled with the help of Hadoop and MapReduce methods. The parallel processing can be provided by the multi hubs which thus might result in increasing the scalability of the mechanism.

Daniele Casagrande, et.al (2012) presented in this paper [12] that in order to provide analysis and management of huge amount of data there is a need of huge measures. These measures will only help in extracting the required information from such huge datasets. The trajectories of Hamiltonian system can be determined as the level lines.

The Hamiltonian function can be interpreted by level function due to which the relating Hamiltonian system can be integrated. Numerous enhancements are made here which help in improving the efficiency of the method. The results have shown the enhancements made and the effectiveness level achieved through application of this method.

Manish Kumar Sharma, et.al, (2015) proposed in this paper [13], an unconventional method in order to detect fatigue within the vehicular drivers. In order to distinguish cognitive fatigue of the driver, the Oximetry Pulse (OP) signal is used. This results in minimizing the accidents and causalities being faced. The fundamental and enhanced k-means are incorporated within the implementation in this paper in order to provide such mechanism. The classification accuracy of the enhanced k-means algorithm is determined to be better in comparison to the existing method. Through the analysis provided within the paper, it is seen that the methods which include selected features can be provided with efficient outcomes with the application of method proposed in this paper.

3. RESEARCH METHODOLOGY

This research work is based on the prediction analysis. The prediction analysis contains two steps, the first step contains the clustering and second step contains the classification. The clustering is approach which can group similar and dissimilar type of data. The k-mean clustering is the partitioned based clustering algorithm. The k-mean clustering algorithm has following steps :-

1. Input the Dataset for the clustering
2. Calculate Arithmetic mean of the whole dataset which will be the centriod point for the clustering
3. Calculate Euclidian distance from the centriod point to analyze data similarity. The Euclidian distance will be calculated by given formula

Euclidean Distance $=\sqrt{(X1-X)^2+(Y1-Y)\text{^}2}$

3. Calculate similarity of the Euclidean distance and cluster data according to their similarity

The clustered data will be given as input to the SVM classifier for the classification.

4. SVM CLASSIFIER

The performance is even better when the dimension of the input space is extremely high. In order to differentiate between the two classes of the training data, the SVM requires identifying the best classification function.The hyperplane f(x) is separated through the linear classification function for the linearly separable dataset. This hyperplane passes through the middle of two classes which can be said to separating them. The new data instance xn is classified by testing the sign function function f(xn); xn which belongs to the positive class if f(xn)> 0. This is done after the determination of a new function. It is the main objective of SVM to determine the best function by maximizing the margin between the two classes. This is due to the fact that there are many such linear hyperplanes. The amount of space or distance amongst two classes is known as hyperplane. The shortest between the closes data points to a point on the hyperplane is known as margin. This can further help us in defining the way to extend the margin which can help in selecting only a few hyperplanes for the solution to SVM even when so many hyperplanes are available

5. PROPOSED ALGORITHM

Input : Dataset, Number of clusters
Output : Classified Data
1. Initialize the centiods randomly C_1,C_2,C_3---------C_n
2. Repeat while data clustered {
For every point “i” in the dataset {
Set C(i)=Mean(Dataset(i))
For each point “j” in the dataset {
Set E.D(i)= √((X1-X)^2+(Y1-Y)^2)
For each point in the Euclidean distance
If (points ==similar(i))
Form cluster 1
Else
form cluster 2
end
3. Classification (){
1. Training set { xi, yi, i=1..l}
2. W_{Height} qi, i=1..l
3. Bias b
4. Training set partitioned into subsets (s) Error Set(s) and Remaining set R
5. Params::e,C, kerenal type and kernel parameters
6. R Matrix
7.Classified Data C = (xc, yc)

6. RESULTS AND DISCUSSION

The prediction analysis technique which consists of k-means and SVM classifier is implemented in MATLAB by considering parameters given in table 1

Table 1. Dataset Description

Parameters	Values
Attributes	24
Instances	634
Missing Values	No
Prediction Attributes	YES

In this research work, the multiclass SVM classifier is compared with the linear SVM and it has been analyzed that multiclass SVM performs better in terms of accuracy and execution time.

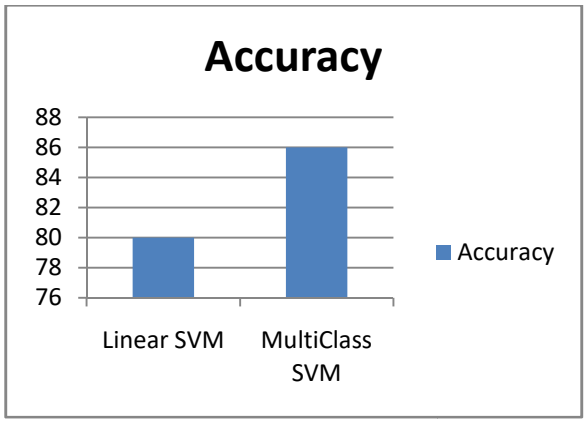

Figure 1. Accuracy Comparison

As shown in figure 1, the accuracy of linear SVM and Multiclass SVM is compared. It has been analyzed that Multiclass SVM performs better as compared to linear SVM.

7. CONCLUSION

In this paper, it has been concluded that prediction analysis is the efficient approach to predict new values from the existing approach. The prediction analysis consists of clustering and classification. In this research work, algorithm is proposed which is based on k-means and SVM classification. The multiclass SVM classifier is compared with linear SVM classifier to analysis reliabilities of the modal. In future, proposed algorithm will be implemented and results will be analyzed in terms of accuracy, execution time.

REFERENCES

[1] R. Karakis, M. Tez, Y. Kilic, Y. Kuru, and I. Guler," A genetic algorithm model based on artificial neural network for prediction of the axillary lymph node status in breast cancer," 2013, Engineering Applications of Artificial Intelligence, vol. 26, no. 3, pp. 945–950

[2] Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin," Analysis of Data Mining Techniques for Heart Disease Prediction", 2016, IEEE

[3] Kamaljit Kaur and Kuljit Kaur," Analyzing the Effect of Difficulty Level of a Course on Students Performance Prediction using Data Mining", 2015 1st International Conference on Next Generation Computing Technologies (NGCT)

[4] Monali Paul, Santosh K. Vishwakarma, Ashok Verma," Analysis of Soil Behaviour and Prediction of Crop Yield using Data Mining Approach", 2015, IEEE

[5] J. Refonaa, Dr. M. Lakshmi, V.Vivek," ANALYSIS AND PREDICTION OF NATURAL DISASTER USING SPATIAL DATA MINING TECHNIQUE", 2015, International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[6] Richa Sharma, Dr. Shailendra Narayan Singh, Dr. Sujata Khatri," Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey", 2016, IEEE, 978-1-5090-0210-8

[7] Sonali Shankar, Bishal Dey Sarkar, Sai Sabitha, Deepti Mehrotra," Performance Analysis of Student Learning Metric using K-Mean Clustering Approach", 2016, IEEE, 978-1-4673-8203-8

[8] Vadlana Baby, Dr. N. Subhash Chandra," Distributed threshold k-means clustering for privacy preserving data mining", 2016, IEEE, 978-1-5090-2029-4

[9] Vaibhav Kumar, Deep Chandra Kandpal, Monika Jain," K-mean Clustering based Cooperative Spectrum Sensing in Generalized k-μ Fading Channels", 2016, IEEE, 978-1-5090-2361-5

[10] R. Kumari, Sheetanshu, M. K. Singh, R. Jha, N.K. Singh," Anomaly Detection in Network Traffic using K-mean clustering", 2016, IEEE, 978-1-4799-8579-1

[11] Kaustubh S. Chaturbhuj, Mrs. Gauri Chaudhary," Parallel Clustering of large data set on Hadoop using Data mining techniques", 2016, IEEE, 978-1-4673-9214-3

[12] Daniele Casagrande, Mario Sassano, and Alessandro Astolfi," Hamiltonian-Based Clustering Algorithms for static and dynamic clustering in data mining and image processing", 2012, IEEE, IEEE CONTROL SYSTEMS MAGAZINE, 1066-033X

[13] Manish Kumar Sharma, Mahesh M. Bundele," Design & Analysis of K-means Algorithm for Cognitive Fatigue Detection in Vehicular Driver using Oximetry Pulse Signal", 2015, IEEE International Conference on Computer, Communication and Control IC4

