Reducing the Waiting Time of a Treatment Report by Using K-Means Algorithm in Big Data

Y.Sivamma[1] C.Yamini[2]
[1]Student [2]Assistant Professor
[1,2]Department of Computer Applications
[1,2]KMM Institute of P.G Studies, Tirupati, India

Abstract— K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed apriority. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. In this research work, proposed algorithm will perform better while handling clusters of circularly distributed data points and slightly overlapped clusters. Big data is a time period used to consult records sets which can be too large or complex for traditional facts-processing utility software program to effectively address. Big information demanding situations include taking pictures statistics, records garage, statistics evaluation, search, sharing, switch, visualization, querying, updating, records private ness and records supply.

Key words: Big Data, K-Means Algorithm, Waiting Time, Clustering

I. INTRODUCTION

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed apriority. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. Big information is a time period used to consult statistics sets which might be too massive or complicated for traditional statistics-processing software program to safely deal with. Data with many cases offer extra statistical strength, at the same time as facts with better complexity can also lead to a better fake discovery fee. Big information demanding situations include taking pictures statistics, records garage, statistics evaluation, search, sharing, switch, visualization, querying, updating, records private ness and records supply. Big data is a time period used to refer to facts units which might be too big or complex for traditional records-processing software application to properly address. Currently, most hospitals are overcrowded and they are not efficient in providing proper queue management. Providing Patient queue management and waiting time prediction is challenging and tedious job as each patient vary in scan. Some of the tasks are independent whereas some tasks are waiting to complete other dependent tasks. Most patients must have to wait in different queues for different treatments. In order to complete required treatment in a shortest duration of time waiting time of each task is predicted in real time. Data with many cases offer greater statistical strength, at the identical time as facts with higher complexity (extra attributes or columns) may moreover cause a higher false discovery fee. Current usage of the time period large statistics has a tendency to refer to the use of predictive analytics, user conduct analytics, or certain other superior records analytics techniques that extract value from statistics, and seldom to a specific size of information set. "There is little question that the portions of data now to be had are certainly massive, but that's now not the maximum relevant function of this new information ecosystem. These studies pay attention chiefly on planning patients their remedy responsibilities right now and maintain a strategic distance from packed line or queues. By using the large affordable statistics from healing facilities affected a person treatment time usage and defer time matter display is calculated. The practical affected individual data are analyzed carefully and thoroughly in slight of essential parameters, for instance, continual treatment begins time, stop time, tolerant age and element treatment content material for each precise undertaking. ID and ascertain diverse sitting tight activities for numerous patients in light of their situations and operations executed amid remedy is completed exactly. **Clustering** is one of the most common exploratory data analysis technique used to get an intuition about the structure of the data. It can be defined as the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different. In other words, we try to find homogeneous subgroups within the data such that data points in each cluster are as similar as possible according to a similarity measure such as euclidean-based distance or correlation-based distance. The decision of which similarity measure to use is application-specific.

II. RELATIVE STUDY

A. Bayesian-inference-based recommendation in online social networks

In this paper, we propose a Bayesian-inference-based recommendation system for online social networks. In our system, users share their content ratings with friends. The rating similarity between a pair of friends is measured by a set of conditional probabilities derived from their mutual rating history. A user propagates a content rating query along the social network to his direct and indirect friends. Based on the query responses, a Bayesian network is constructed to infer the rating of the querying user. We develop distributed protocols that can be easily implemented in online social networks. We further propose to use Prior distribution to cope with cold start and rating sparseness. The proposed algorithm is evaluated using two different online rating data sets of real users. We show that the proposed Bayesian-inference-based recommendation is better than the existing trust-based recommendations and is comparable to Collaborative Filtering (CF) recommendation. It allows the flexible tradeoffs between recommendation quality and recommendation quantity. We further show that informative

Prior distribution is indeed helpful to overcome cold start and rating sparseness.

B. The waiting time of a treatment report

Many substance users report that they experience multiple barriers that produce significant challenges to linking with treatment services. Being on a waiting list is frequently mentioned as a barrier, leading some people to give up on treatment and to continue using, while prompting others to view sobriety during the waiting period as proof they do not need treatment. This ethnographic study examines the views that 52 substance users have of the waiting time before treatment and the strategies they created to overcome it. Understanding how substance users react to waiting time itself and in relation to other barriers can lead to services that are effective in encouraging treatment linkage.

C. A Modified K-Means Algorithm for Big Data Clustering

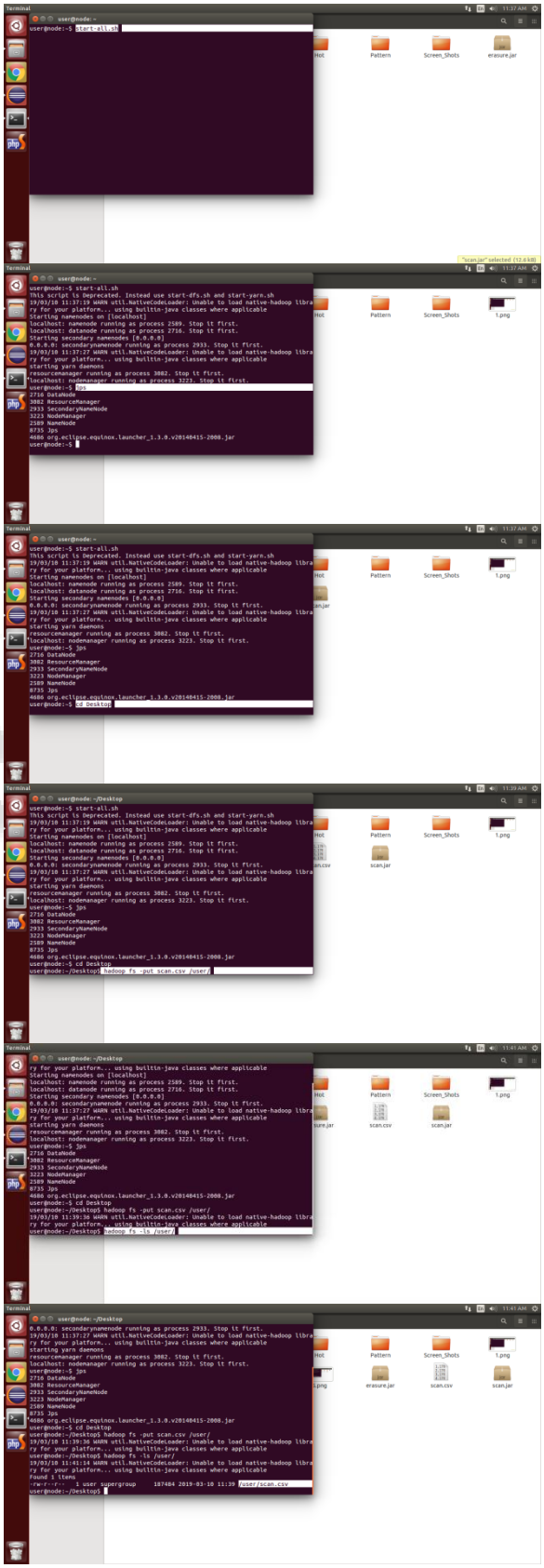
Emergence of modern techniques for scientific data collection has resulted in large scale accumulation of data pertaining to diverse fields. Conventional database querying methods are inadequate to extract useful information from these huge amounts of data. The absence of category information distinguishes data clustering (unsupervised learning) from classification or discriminate analysis (supervised learning). The aim of clustering is exploratory in nature to find structure in data. Machine learning algorithms; provide an automatic and easy way to accomplish such tasks. These algorithms are classified into supervised, unsupervised, semi-supervised algorithms. One of the most popular, simple and widely used clustering (unsupervised) algorithms is K-means. The original k-means algorithm is computationally expensive. Several methods have been proposed for improving the performance of the k-means clustering algorithm. This thesis proposes a modified k-mean clustering algorithm where modification refers to the number of cluster and running time. According to our observation, quality of the resulting clusters heavily depends on the selection of initial centroids and after a certain number of iterations, only a small part of the data elements change their cluster. So there is no need to re-distribute all data elements. Therefore, proposed method first finds the initial centroids and puts an interval between those data elements which will not change their cluster during the next iteration and those which may change the cluster to reduce the workload significantly in case of very big data sets. We evaluate our method with different sets of data and compare with others methods as well.

III. PROPOSED ALGORITHM

A. K Means Algorithm

K means algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

IV. SCREEN SHOTS



V. CONCLUSION

Proposed algorithm will perform better while handling clusters of circularly distributed data points and slightly overlapped clusters. *Big data is a time period used to consult records sets which can be too large or complex for traditional facts-processing utility software program to effectively address.* The Scikit Learn Algorithm will do game plan and backslide of enlightening lists and recommends sufferers a compelling and advantageous treatment structure with the base holding up time. Future examinations are bearing specifically to embellish a more noteworthy valuable motivation with limited way care or dataset hunting down and with low costly and substantially less time taking technique is suggested for work.

REFERENCES

[1] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, ``Robust and accurate shape model matching using random forest regression voting," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1862 1874,Sep. 2015.

[2] Apache. (Jan. 2015). Hadoop. [Online]. Available: http://hadoop. apache.org

[3] Y. Xu, K. Li, L. He, L. Zhang, and K. Li, ``A hybrid chemical reaction optimization scheme for task scheduling on heterogeneous computing systems," IEEE Trans. Parallel Distrib. Syst., vol. 26, no. 12, pp. 3208 3222, Dec. 2015.

[4] G. Adomavicius, A. Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", *IEEE Trans. Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734-749, 2005.

[5] G.Biau,"Examination of Random Forest Model", J.Mach Learn Res Vol. 13, No. 1, pp. 1063-1095, 2012.

[6] H. B. Li, W. Wang, H. W. Ding, J. Dong, Trees weighting arbitrary backwoods technique for grouping high-dimensional uproarious information" In Proc. IEEE seventh Int. Conf. e-Business Eng. (ICEBE), Nov. 2010, pp. 160-163.

[7] S. Meng, W. Dou, X. Zhang, J. Chen,"KASR: Catchphrase mindful administration proposal technique on Map Reduce for huge information applications", IEEE Trans. Parallel Distrib. Syst., Vol. 25, No. 12, pp. 3221-3231, 2014.

[8] G.Adomavicusand Y.Know,"New recommendation techniques for multicriteria systems", IEEEIntell. Syst, Vol. 22, No. 3, pp. 48-55, 2007.

