A HYBRID APPROACH TO DATA MINING ALGORITHMS FOR CLASSIFICATION

Avinash Pandey[1], Garima Srivastava[2], Anuradha Sharma[3]
[1] Student, [2] Assistant Professor, [3] Assistant Professor
Amity School of Engineering and Technology
Amity University, Lucknow, India

Abstract: ***These days in data investigation huge volume of data is generated each and every second. This induced data is massive to the point that it is hazardous to find the suitable data went for a positive reason on requirement. Data mining is a strategy to convey a system to selection the reason point by point data in the appropriate extend for the policymaking determination. The numerical capacities of data mining algorithms can be consolidated, changed and refined to deliver a more productive Algorithm for the order of data indexes. The data characterization is finished by the classifier additionally in view of capacity on which the entire algorithmic productivity is depended. Some top data mining algorithm require considerable amount of time to classify the data and if the data element is similar to many classes then they do not ensure the strong relation of data element with the either class. This paper gives an approach towards a hybrid data mining algorithm that combines some of the features of existing data mining algorithms. The paper gives an approach to develop hybrid data mining algorithm to classify the large data sets. This algorithm uses some of the key features of existing k-means algorithm and support vector machine algorithm.***

Key Words. Data Mining, K Means, Support vector Machine, Hybrid Algorithm

I. INTRODUCTION

Data collections like operational, non-operational and the meta Data are dissected by these numerical ways to deal with get significant actuality from the hazardous enormous measure of Data. Based on wellspring of start, the Data to be mined can be from value-based database, social database, Data distribution center, level documents, interactive media database, time arrangement database, spatial database and the electronic Data mining. This paper utilizes the best Data mining calculations on premise of productivity and attempt to investigate their scientific angles and try to describe new approach of hybrid algorithm for data classification.

The k-means and support vector machine are two most widely used data mining algorithms. The both algorithm have their own individual disadvantages. When we hybridize the support vector machine and k-mean together they can be proved very efficient. The support vector machine provides the heuristic details to k-means algorithm to work on. Due to this additional detail provided by the support vector machine algorithm we initialize the initial assignment step of k-means. The details provided by the support vector machine can be also exploited for the dimension reduction for further improving the efficiency of the k-means algorithm.

The hybrid data mining algorithms are efficient way to remove the disadvantages of the previous existing data mining algorithms. The hybrid algorithm reduces the runtime of the algorithms individually by using a heuristic approach and increases the efficiency of the algorithm in terms of runtime and precision. The hybrid algorithms are the best approach while dealing with the large data sets.

II. LITERATURE SURVEY

Dewan Md. Farid et al. expected a cross breed calculation utilizing choice tree and credulous Bayes calculation. This strategy is for the most part meant to build the precision of grouping for multi class-class order assignments. This proposed crossover calculation demonstrates the better affectability, specificity, cross approval and order exactness on genuine benchmark informational indexes from UCI [2]. The proposed framework naturally extricates the incentive in preparing informational collections. In addition, it recognizes the successful characteristics from loud complex preparing informational indexes. Information mining calculation is utilized by computational wise scientists for taking care of characterization and bunching issues. The proposed half breed calculation demonstrates the 90% exactness [1] in picture arrangement.

Lufimpu-Luviya Yannick et al. built up a half and half calculation by utilizing Support Vector Machine and Classification and Regression Tree (CART) to recognize the age band of a 2D picture confront [6]. The main strategy is utilized to discover the period of face pictures and the second technique is utilized to take care of the grouping issue. This outcome represents that second technique beats the primary strategy. The second strategy creates some dangerous perplexities. This proposed framework gives two favorable circumstances: a. It shrinkage the computational time and b. It gives similar outcomes in the whole face. It is harder to anticipate the age band and gives bring down execution in that procedure. This work is done impossible to miss the sexual orientation rendering to the promoting business imperatives. This proposed half and half calculation gives 84% precision in picture grouping.

Alireza Taravat et al. foreseen half and half calculation for to expand the advancement of sky pictures. In this test multilayer perceptron neural system and bolster vector machine calculation are half and half. This two-grouping calculation (MLPNN and SVM) have been looked at [10]. The proposed calculation is connected in dataset containing in excess of 250 pictures from different test locales. The got precision demonstrates the better identification result. The outcomes demonstrate MLPNN calculation is mechanized calculation for picture grouping. A. Kannan et al. recommended a cross breed calculation for characterizing the MRI pictures [3]. Here, the creators composed half breed strategy utilizing KNN and SVM ideas for MRI picture grouping. K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) are the most predominant information

mining technique in picture characterization. The proposed framework demonstrates the 86.7% exactness in picture characterization. The precision of the outcome is essentially expanded and the mistake rate has been continuously lessened.

Bor-Chen Kuo et al. introduced a mixture calculation utilizing Radial Basis Function (RBF) and Support Vector Machine to enhance the characterization precision in pictures. This calculation is investigated progressively datasets to gauge the grouping execution [11]. In this approach two things can be accomplished coefficients being figured: little subset of positioning and highlights. The proposed strategy shows 90% exactness in picture order [7]. From this analysis characterization exactness is expanded impressively from the subset.

M.R. Homaeinezhad et al. gifts a half and half calculation utilizing bolster vector machine and k-implies calculation. The proposed crossover calculation basically expected to build the heartiness of order precision in QRS pictures. In this proposed strategy the ECG flag distinguished and outline utilizing powerful wavelet-based calculation [4]. This technique demonstrates higher execution than other strategy. The proposed grouping strategy indicates mistake rate variety to another calculation for giving basic informational indexes [9]. To compute the execution of the proposed cross breed calculation, the succeeded comes about were related with a few related examinations. The proposed strategy demonstrates better precision in QRS pictures [12].

Utilizing innocent bayes and bolster vector machine for drusen location from fundus pictures. This proposed calculation speaks to every locale with number of highlights and afterward it's connected in this half and half classifier as an innocent bayes and SVM to group the area as drusen and non-drusen [8]. This calculation assessed by STARE database utilizing the precision, affectability and specificity. Distinctive creators have introduced diverse PC helped analytic framework for various retinal disorders. This mixture strategy gives great exactness which implies the fundus pictures are accurately ordered. The work demonstrates the precision, specificity and affectability are 0.98, 0.99 and 0.97 separately [5]. The framework decisively sectioned the pictures overwhelms than the prior proposed framework [13]. At long last, this proposed half and half calculation indicates 98% precision in picture grouping. Bolster vector machine predicts most exact outcomes in order, especially in content characterization.

III. RESEARCH METHODOLOGY

All the simulation has been done on the Weka and the analysis of the result has been done on the MatLab 2010a. The results are acquired Java as the support environment to run the algorithm.
Input of hybrid algorithm:
C(c1,c2,c3…….cn): contain the input data set on which the classification has to be done
Output of the hybrid algorithm:
A set of M clusters that are generated by the classification of
Strategy:
1. Input the data set after preprocessing.
2. Draw the outline estimations in the dataset on 2 Dimension plane and distinguish the number of bunches in light of normal outline esteem. The quantity of bunches indicating most elevated outline esteem is picked as an estimation of n for that information.
3. Segment the dataset into n a balance of.
4. Number juggling mean of respectively part is occupied as the centroid argument.
5. Register the Euclidean separation of respectively datapoint di to every one of the centroids as edist (di, oj)
6. For each di, look at the nearest centroid and appoint di to centroid.
7. Set Closest_redo[i]= nedist(ci, oj) /cj: closest centroid.
8. For each j, calculate again the centroids.
9. Rehash
10. Aimed at every numbers theme ci
10.1 Calculate its separation starting the new centroid of the current closest group.
10.2 If this separation is not exactly or equivalent to the past separation, the information point remains in that group, Else
10.2.1 Compute redo (ci, oj) from all group centroids; End for.
10.2.2 Assign the information point di to the bunch with the closest Centroid.
10.2.3 Set Closest_Redo[i] = nedist(ci, oj);
Terminate for circle.
11. Take finest normal aggregate of every Euclidean separation
furthermore, acquire the last yield.
12. Prepare the SVM classifier utilizing lessened dataset.
13. Characterize the new information utilizing SVM classifier.

IV. RESULT AND DISCUSSION

The current Algorithms are adaptable yet have a lower achievement rate in arranging huge Data collection if the Data is broadly conveyed and have almost no comparability between them. The Data mining degree is expanding step by step as the Data gathered every second is colossal. Getting the significant data from this enormous measure of Data is troublesome, so there is a need of productive Data mining Algorithm. Tis poor accuracy of the traditional data mining algorithm can be improved by using the hybrid approaches of the data mining algorithms. The algorithms will reduce the time consumption by the individual k-means algorithm by improving the convergence criteria, hence reducing the number of relocation steps which in turn reduces the number of comparison takes place during the execution.

The accuracy of the hybrid of support vector machine and k-means is better than the Sequential Minimal optimization based on support vector machine and K-means individually. The accuracy of the Hybrid of K-means and support vector machine is 93.1 % while that of simple k-means algorithm is 81.3%.

The time taken by the hybrid of support vector machine and k-means algorithm is 0.19 s while the simple k-means algorithm takes slightly more time 0.31 s to classify the clusters. The build time for the hybrid of support vector machine and k-means algorithm is 0.057 s while the build time of simple K-means algorithm is 0.040 s.

V. ACKNOWLEDGEMENT

The beginning of this examination work " A hybrid approach to data mining algorithms for classification" has been done from August 2016. My exploration control Ms. Garima Srivastava, Ms. Anuradha Sharma in branch of Computer Science and Engineering at Amity University have proposed to take every necessary step on the point of Data mining.

REFERENCES

[1] Fahim, A. M., Salem, A. M., Torkey, F. A., & Ramadan, M. A. (2006). An efficient enhanced k-means clustering algorithm. Journal of Zhejiang University SCIENCE A, 7(10), 1626-1633.

[2] Chetty, N., Vaisla, K. S., & Patil, N. (2015, May). An Improved Method for Disease Prediction Using Fuzzy Approach. In Advances in Computing and Communication Engineering (ICACCE), 2015 Second International Conference on (pp. 568-572). IEEE.

[3] Shah, S., & Singh, M. (2012, May). Comparison of a time efficient modified K-mean algorithm with K-mean and K-medoid algorithm.In Communication Systems and Network Technologies (CSNT), 2012International Conference on (pp. 435-437). IEEE.

[4] Ghumbre, S., Patil, C., & Ghatol, A. (2011, December). Heart disease diagnosis using Support Vector Machine. In International conference on computer science and information technology (ICCSIT') Pattaya.

[5] Kuramochi M, Karypis G (2005) Gene Classification using Expression Profiles: A Feasibility Study. Int J Artif Intell Tools 14(4):641–66011]

[6] Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16, 645–678.Milovic, B., & Milovic, M. (2012). Prediction and decision making in Health Care using Data Mining. Kuwait Chapter of the Arabian Journal of Business and Management Review, 1(12), 126.

[7] Huang, C. L., Liao, H. C., & Chen, M. C. (2008). Prediction model building and feature selection with Support Vector Machines in breast cancer diagnosis. Expert Systems with Applications, 34(1), 578-587.

[8] Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using Support Vector Machine. International Journal of Engineering Research and Applications, 3(2), 1797-1801.

[9] Kukar M (2006) Quality assessment of individual classifications in machine learning and data mining. Knowl Inf Syst 9(3):364–384

[10] Bharti, S., & Singh, S. N. (2015, May). Analytical study of heart disease prediction comparing with different algorithms. In Computing, Communication & Automation (ICCCA), 2015 International Conference on (pp. 78-82). IEEE.

[11] Langville AN, Meyer CD (2006) Google's PageRank and beyond: the science of search engine rankings. Princeton University Press, Princeton.

[12] Purwar, A., & Singh, S. K. (2015). Hybrid prediction model with missing value imputation for medical data. Expert Systems with Applications, 42(13), 5621-5631.

[13] Jain, A., Rajavat, A., & Bhartiya, R. (2012, November). Design, Analysis and Implementation of Modified K-mean Algorithm for Large Data-Set to Increase Scalability and Efficiency. In Computational Intelligence and Communication Networks (CICN), 2012 Fourth International Conference on (pp. 627-631). IEEE.

[14] Santhanam, T., & Padmavathi, M. S. (2015). Application of K-means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis. Procedia Computer Science, 47, 76-83.

[15] Yadav, A. K., Tomar, D., & Agarwal, S. (2013, July). Clustering of lung cancer data using Foggy K-means. In Recent Trends in Information Technology (ICRTIT), 2013 International Conference on (pp. 13-18).

[16] RKumar, G., Ramachandra, G. A., & Nagamani, K. (2014). An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets. International Journal, 4(2).