International Journal of Science, Technology and Management (IJSTM) ISSN (online): 2321-774X

Volume 9, Issue 1, 2022

Survey on Analysis of Various Classification Algorithms of Data Mining

Salma

salmakhanbehal@gmail.com

M.Tech. Scholar, Department of CSE, BRCM CET, Behal, Bhiwani, Haryana (India)

Dr. Dinesh Kumar

dinesh.muwal@gmail.com

Professor, Department of CSE, BRCM CET, MDU, *Rohtak*, Haryana (India)

ABSTRACT

Data Mining is a field of search and researches of data. Mining the data means fetching out a piece of data from a huge data block. The basic work in the data mining can be categorized in two subsequent ways. One is called classification and the other is called clustering. Although both refers to some kind of same region but still there are differences in both the terms. The classification of the data is only possible if you have modified and identified the clusters. In the presented research paper, our aim is to analyze the different classification algorithm like C4.5, CART, SVM, RF, ID3, KNN etc. and find out which algorithm work better to classify different data classes or for classification technique.

Keywords: Data mining, classification, C4.5, CART, SVM, RF, ID3, KNN.

INTRODUCTION

Data mining or KDD is the non-trivial elicitation of implicative already unexplored and likely proper information from the data. Data mining may be visible because the exploration and analysis of huge amount of information on the way to find out significant styles and regulations this studies focuses on a selected branch of groups individually the buying and income department consequently the intention of facts mining is to allow groups to enhance its relevant operations via a higher understanding of the enterprise tactics the records mining strategies and tools defined here are however similarly applied in fields ranging from regulation enforcement to radio astronomy medication and purchaser relation control in fact almost not one of the facts mining technique have been first evoked with business applications in thoughts all techniques used are borrowed from records laptop science and device learning research the selection of a specific mixture of techniques to use in a specific state of affairs relies upon on the nature of the statistics mining undertaking the nature of the to be had records and the skills and choices of the records miner berry and linoff 1999 provider and povel 2003 grover and mehra 2008 records mining obligations may be directed and undirected as is said via berry and layoff 2004 and bendoly 2003 in different literature lusting et al 2010 there may be made difference between prescriptive and descriptive data mining duties in which predictive is just like directed and descriptive to undirected directed information mining tries to give an explanation for or categorize some specific target discipline such as income or reaction directed facts mining responsibilities are class forecasting regression and description profiling undirected facts mining tries to find patterns or similarities amongst agencies of data without using a specific goal discipline or series of predefined training undirected facts mining duties are affiliation clustering and outline profiling records mining maximum of the time encompass constructing fashions which can be algorithms or units of rules that connects a set of inputs to a selected target underneath the proper situations a model can result in perception with the aid of presenting an evidence of how results of unique interest including setting an order or failing to pay a bill are associated with and predicted by means of the available statistics. it uses several techniques like clustering, classification, find dependency networks data summary analyze changes and detect anomalies.

CLASSIFICATION IN DATA MINING

Classification is utilized to figure out in which collection every information occurrence is connected inside a given dataset. It is utilized for characterizing information into various classes as per some constrains. Classification methods in DM are fit for handling a lot of information. It tends to be utilized to anticipate straight out class names and arranges information in view of preparing set and class marks and it very well may be utilized for classification recently accessible data. The term could cover any setting wherein some choice or estimate is made based on as of now accessible data. Classification procedure is perceived strategy for over and again going with such choices in new circumstances. Here on the off chance that we expect that issue is a worry with the development of a strategy that will be applied to a proceeding with succession of cases where each new case should be relegated to one of a group of pre characterized classes based on noticed elements of information. Production of a classification strategy from a group of information for which the specific classes are referred to progress of time is named as example acknowledgment or managed learning. Settings in which a classification task is crucial incorporate, for instance, relegating people to credit status based on monetary and other individual data, and the underlying conclusion of a patient's sickness to choose quick therapy while anticipating wonderful experimental outcomes. Probably the most basic issues emerging in science, industry and trade can be called as classification or decision issues. [14]

LITERATURE SURVEY

In 2022, Wang, Jing, et al. [1] proposes a cloud-random forest (C-RF) model joining cloud model and RF to evaluate the gamble of CHD. In this model, in light of the CART, a weight deciding algorithm in view of the cloud model and decision making trial and assessment lab is applied to get the loads of the assessment credits. The attribute weight and the gain value of smallest Gini coefficient relating to a similar characteristic are weighted and added. The weighted aggregate is then used to supplant the original gain value. This worth rule is utilized as another CART hub split measure to build another decision tree, hence framing another RF, by name, the C-RF. The Framingham dataset of the Kaggle stage is the examination test for the experimental investigation. Contrasting the C-RF model and CART, SVM, convolutional neural network (CNN), and (RF) utilizing standard execution assessment lists, for example, precision, mistake rates, ROC curve and AUC value. The outcome shows that the classification exactness of the C-RF model is 85%, which is worked on by 8, 9, 4 and 3% separately contrasted and CART, SVM, CNN and RF. The error rate of the principal type is 13.99%, which is 6.99, 7.44, 4.47 and 3.02% lower than CART, SVM, CNN and RF separately. The AUC esteem is 0.85, which is likewise higher than other examination models. Accordingly, the C-RF model is more predominant on classification performance and classification effect in the gamble evaluation of CHD.

In 2022, Reynara, Febian Joshua et al. [2] implemented the C4.5 and CART algorithm to classify occupations based on alumni data and produce analysis and job classification models. The classification was carried out on three types of experiments, wherein the experiment eight categories of work, the highest accuracy was 42% for C4.5 algorithms and 43% for CART algorithm. In the experiment of three categories of work, the highest accuracy was 58% for C4.5 and 61% for CART. While in the experiment of two categories of work, the highest accuracy was 75% for c4.5 and 77% for CART. The best algorithm for classifying fields of work based on alumni data from the two algorithms used is the CART algorithm because all the highest accuracy of the model produced in the three experiments is better than the highest accuracy of the model generated by the C4.5 algorithms, although the difference in accuracy is not too significant. there are still several methods that can be used to overcome imbalanced classification problems in addition to the two methods used in this study, such as Cost-Sensitive Learning, Gradient Boosting with XGBoost, Ensemble Algorithm, etc. Further research can apply these methods to overcome the imbalance of classification problems encountered.

In 2021 Li, Zhuqing.[3] analyze the utilization of inertial sensors in basketball pose analysis. The information of 20 players in various postures were gathered by Micro-electromechanical frameworks inertial sensors. The mean, fluctuation, and skewness were taken as highlights to think about the performance of C4.5, RF, k-NN, and SVM algorithms in analyze pose information. It was found that the classification accuracy of the k-NN algorithm was around 90%, and the precision of C4.5, RF, and SVM algorithm was all above 90%. The accuracy of the RF technique was the most elevated (98.72%), which was fundamentally higher than C4.5 and SVM algorithms. The outcomes checked the benefits of the RF algorithm in basketball pose analysis. The examination results affirm the unwavering quality of the inertial sensor in the field of movement act investigation and make a few commitments to its application in sport preparing. This paper offers help for the analyze the motion pose.

In 2021, Amrin, Amrin, and Omar Pahlevi [4] was apply and think about a few DM and improvement classification algorithms with particle swarm optimization (pso), including the C4.5 algorithms, K-NN, C4.5 with PSO, and K-NN with PSO to analyze provocative sicknesses. Cautiously, then look at which of the few of these techniques is the most accurate. In view of the consequences of estimating the exhibition of the three models utilizing the Cross Validation, Confusion Matrix and ROC Curve techniques. In view of the outcomes, it is realized that the C4.5 techniques with PSO is the best strategy with an accuracy of 79.51% and an under the curve (AUC) worth of 0.950, then, at that point, the k-NN technique with PSO has a precision of 75.59% and an AUC worth of 0.909, then the C4.5 techniques with an exactness pace of 70.99% and an AUC worth of 0.950, then the k-NN strategy with an accuracy pace of 67.19%, and an AUC worth of 0.873. This demonstrates that PSO can work on the exhibition of the classification Algorithms utilized.

In 2020, Bunkar, Kamal, et. al. [5] give a description to apply the cycles of DM, especially classification, to help work on the nature of the advanced education platform by analyzing student data that impact student performance in courses. The proposed algorithm comprises of two significant parts first, the student learning philosophy analysis, and second is performance or accomplishment prediction. For student learning conduct examination, the paper incorporates the outcomes and algorithm choice; further, it contains the model of student performance prediction model. In this unique situation, two well-known DM algorithms named as C4.5 and CART decision trees are applied. Those methods acknowledge the accomplishment information for preparing and by utilizing student's current performance, the future performance is anticipated. The correlation of the presentation of the two algorithms is read up for prediction precision, error rate, time, and memory utilizations. The outcomes outline the C4.5 decisions tree-based performance prediction uncovers higher accuracy also, lesser memory and shorter time utilization. In this manner in the future, the C4.5 decision tree calculation is taken on.

In 2020, Yontar, Meltem et al [6] propose a model to predict regardless of whether credit card clients will pay their debts or not. Utilizing the proposed model, potential neglected dangers can be anticipated and fundamental moves can be made in time. For the expectation of clients' instalment status of the next months, we utilize Artificial Neural Network (ANN), Support Vector Machine (SVM), Classification and Regression Tree (CART) and C4.5, which are broadly utilized man-made consciousness and decision tree algorithms. Our dataset incorporates 10713 client's records got from a notable bank in Taiwan. These records comprise of client data, for example, how much credit, orientation, training level, conjugal status, mature, past instalment records, receipt endlessly measure of Visa installments. We apply cross approval and hold-out techniques to partition our dataset into two sections as training and test sets. Then, at that point, we assess the algorithms with the proposed exhibition measurements. We likewise advance the boundaries of the algorithms to work on the presentation of expectation. The outcomes show that the model worked with the CART algorithms, one of the decision tree algorithms, gives high precision (around 86%) to anticipate the clients' instalment status for the next month. At the point when the algorithms parameters are optimized, accuracy and performance are expanded.

In 2019, Asri, Hiba, et.al [9] presents performance comparison between various DM algorithms: (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbours (k-NN) applied to the Wisconsin Breast Cancer (WBC unique) datasets. We use classification exactness and disarray lattice in view of 10-overlay cross approval technique. We likewise present a combination at classification level between those classifiers to get the most reliable multi-classifiers approach. Exploratory outcomes show that the classification utilizing combination of SVM, NB and C4.5 arrived at the most noteworthy precision (97.31%) while exactness of utilizing a classifier SVM is (97.13%). All experiments are executed inside a simulation environment and led in WEKA DM tool.

In 2018, Zacharis, Nick Z. [10] test the capacity of CART analysis to anticipate progress in web-based blended learning environments, by utilizing on the web associations put away in the system log file. CART is nonparametric and in this way, is appropriate for information having a place in different disseminations. The impact of exceptions in the info factors is unimportant, furthermore, the pruning strategy it applies guarantees that no significant foundation is neglected. In this review, the CART procedure accomplished extremely high exactness (99.1 %) in classifying students into the individuals who effectively passed the class and the individuals who failed to do as such. The quantity of instant messages that a student shipped off colleagues and educators was the most significant indicator obviously success. The second most significant figure foreseeing whether a student will pass or bomb the course, was the number of commitments to bunch wiki-based assignments. Test endeavors and the quantity of documents saw had a particular however somewhat unobtrusive impact on anticipating achievement. Without a doubt more examination should be finished involving this scientific apparatus in more crowded and different learning settings, yet all at once this research major areas of strength for gives that the CART technique of analysis can actually utilize the proposed indicators furthermore, forecast student course accomplishment. Educators may depend on the prognostic force of CART investigation to plan opportune intercessions and assist students with succeeding.

In 2018, Saheed, Y. K., et al. [11] presents a technique to foresee student performance utilizing Iterative dichotomiser 3 (ID3), C4.5 and (CART). The examination was performed on Waikato Environment for Knowledge Analysis (WEKA). The exploratory outcomes showed that an ID3 exactness of 95.9%, particularity of 95.9%, accuracy of 95.9%, review of 95.9%, f-proportion of 95.9% and mistakenly ordered occurrence of 3.83. The C4.5 gave an exactness of 98.3%, particularity of 98.3%, accuracy of 98.4%, review of 98.3%, f-proportion of 98.3% and mistakenly ordered occurrence of 1.70. The CART results showed an exactness of 98.3%, explicitness of 98.3%, accuracy of 98.4%, review of 98.3%, f-proportion of 98.3% and erroneously arranged case of 1.70. The time taken to assemble the model of ID3 is 0.05 seconds, C4.5 is 0.03 seconds and CART of 0.58 seconds. Trial outcomes uncovered that C4.5 outflanks different classifiers and calls for sensible measure of investment to construct the model.

In 2017 Gupta, Bhumika, et al. [11] investigates different decision tree algorithm that are utilized in DM. We found that every Algorithm has its own benefits and burdens according to our review. The productivity of different Decision tree algorithm can be investigated in view of their precision and the property determination measure utilized. The effectiveness of algorithms also relies upon the time taken data of the decision tree by the algorithm. We found that both C4.5 and CART are superior to ID3 when missing qualities are to be dealt with through ID3 can't deal with absent or boisterous information. Yet, we likewise dissected that ID3 delivers quicker results. The paper also gives a thought of the quality determination measure utilized by different decision tree algorithms like ID3 calculation utilizes data gain, the C4.5 algorithm utilizes gain proportion and CART calculation utilizes GINI index as the quality determination measure. The paper li gives the techniques for estimation of these quality choice measures. On the whole, we track down that these calculations for decision tree induction are to be utilized at various times as indicated by the circumstance.

In 2016, Geng, Huantong, et al. [12] looks at the landfalling tropical cyclones(TCs) over China utilizing cutting edge DM techniques (for example finite Mixture Model (FMM) based cluster technique Classification and Regression Tree (CART)). Utilizing the 1951-2012 TC best track dataset delivered by the Shanghai Typhoon Institute of the Chinese Meteorological Administration, the tracks of TCs landfalling over the Chinese coast were arranged into three bunches through a FMM. A few environment records were investigated involving the CART calculation for the three groups. The expectation model worked via CART for summer track recurrence depended on an irregular examining of the information for a considerable length of time (around 75% of the complete years) as the preparation set with a preparation precision of 100 percent (Cluster-1), 89.96% (Cluster-2) and 100 percent (Cluster-3). Information for the excess 16 years (around 25%) was utilized for testing with an expectation exactness of 87.5% (Cluster-1), 62.5% (Cluster-2) and 68.75% (Cluster-3). This study centers around Cluster-1 of summer TCs landfalling over China for its high, serious areas of strength for recurrence, extreme effects and long life expectancy. Moreover, it proposes that the FMM calculation is successful for track grouping of TCs arriving over China. What's more, the CART calculation, which was utilized to assemble the expectation model of Cluster-1 for the arrangement of track recurrence, showed high exactness and its outcomes can be made sense of and saw without any problem. It gives a novel structure to determining the recurrence of TCs landfilling over China.

In 2013, Neelamegam, S., and E. Ramaraj [16] provide a review of different classification techniques in data mining & also this paper presents the fundamental arrangement methods a few significant sorts of arrangement strategy including decision tree, Bayesian networks, k-nearest neighbor classifier, neural network, support vector machine. data mining offers promising ways of revealing secret examples inside a lot of information these secret examples might possibly be utilized to foresee future way of behaving the accessibility of new data mining calculations in any case ought to be met with alert as a matter of some importance these methods are just as great as the information that has been gathered great information is the principal prerequisite for good information investigation expecting great information is accessible the subsequent stage is to pick the most fitting strategy to mine the information.

CONCLUSION

From all the above calculations we come to the conclusion that the C4.5 algorithms is an excellent algorithm when we are dealing with classifying a small or medium sized data. It simply provides good performance accuracy every time. A direct algorithm of C-means method requires time proportional to the product of number of patterns and number of clusters identified. CART algorithm is a decision based algorithm which is used here with C mean algorithm in order to improve the efficiency of classifying data in terms of accuracy, no of clusters formed to classify the data and the time taken to classify the data from an available vast amount of data.

REFERENCES

[1] Wang, Jing, et al. "Risk assessment of coronary heart disease based on cloud-random forest." Artificial Intelligence Review (2022)

[2] Reynara, Febian Joshua, Sepriana Carolina, and Iustisia Natalia Simbolon. "The Comparison of C4. 5 and CART (Classification and Regression Tree) Algorithm in Classification of Occupation for Fresh Graduate." (2022).

[3] Li, Zhuqing. "Feature Extraction and Data Analysis of Basketball Motion Postures: Acquisition with an Inertial Sensor." Journal of Engineering and Science in Medical Diagnostics and Therapy 4.4 (2021).

[4] Amrin, Amrin, and Omar Pahlevi. "Data Mining Optimization Based on Particle Swarm Optimization for Diagnosis of Inflammatory Liver Disease." JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING 5.1 (2021): 152-159.

[5] Bunkar, Kamal, and Sanjay Tanwani. "Student performance prediction using C4. 5 decision tree and CART algorithm." Parishodh Journal 9.II (2020): 1702-1716.

[6] Yontar, Meltem, Özge Hüsniye Namli, and Seda Yanik. "Using machine learning techniques to develop prediction models for detecting unpaid credit card customers." Journal of Intelligent & Fuzzy Systems 39.5 (2020): 6073-6087.

[7] Pah, Clarissa Elfira Amos, and Ditdit Nugeraha Utama. "Decision support model for employee recruitment using data mining classification." International Journal 8.5 (2020).

[8] Saputra, Dedi, et al. "Performance Comparison and Optimized Algorithm Classification." Journal of Physics: Conference Series. Vol. 1641. No. 1. IOP Publishing, 2020.

[9] Asri, Hiba, Hajar Mousannif, and Hassan Al Moatassim. "A hybrid data mining classifier for breast cancer prediction." International Conference on Advanced Intelligent Systems for Sustainable Development. Springer, Cham, 2019.

[10] Zacharis, Nick Z. "Classification and regression trees (CART) for predictive modeling in blended learning." IJ Intelligent Systems and Applications 3 (2018): 1-9.

[11] Saheed, Y. K., et al. "Student performance prediction based on data mining classification techniques." Nigerian Journal of Technology 37.4 (2018): 1087-1091.

[12]Gupta, Bhumika, et al. “Analysis of various decision tree algorithms for classification in data mining." International Journal of Computer Applications 163.8 (2017): 15-19.

[13] Geng, Huantong, et al. "A prediction scheme for the frequency of summer tropical cyclone landfalling over China based on data mining methods." Meteorological Applications 23.4 (2016): 587-593.

[14] Nikam, Sagar S. "A comparative study of classification techniques in data mining algorithms." Oriental Journal of Computer Science and Technology 8.1 (2015): 13-19.

[15] Satyanarayana, N., C. H. Ramalingaswamy, and Y. Ramadevi. "Survey of classification techniques in data mining." International Journal of Innovative Science, Engineering & Technology 1.9 (2014).

[16] Neelamegam, S., and E. Ramaraj. "Classification algorithm in data mining: An overview." International Journal of P2P Network Trends and Technology (IJPTT) 4.8 (2013): 369-374.

[17] Chrysos, Grigorios, et al. "HC-CART: A parallel system implementation of data mining classification and regression tree (CART) algorithm on a multi-FPGA system." ACM Transactions on Architecture and Code Optimization (TACO) 9.4 (2013): 1-25

[18] Kumar, Raj, and Rajesh Verma. "Classification algorithms for data mining: A survey." International Journal of Innovations in Engineering and Technology (IJIET) 1.2 (2012).

[19]Agarwal, Sonali, G. N. Pandey, and M. D. Tiwari. "Data mining

in education: data classification and decision tree approach." International Journal of e-Education, e-Business, e-Management and e-Learning 2.2 (2012): 140.

[20] Keramati, Abbas, and Niloofar Yousefi. "A proposed classification of data mining techniques in credit scoring." Proc. 2011 Int. Conf. on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia.2011.