Investigation of KNN and Decision Tree Induction Modelin Predicting Customer Buying Pattern

V.Umarani , M.Subathra

{*vur.mca@psgtech.ac.in, msa.mca@psgtech.ac.in}

Assistant Professor (Sr.Gr), Department of Computer Applications,
PSG College of Technology, Coimbatore, Tamilnadu, India[1,2]

Abstract- Machine Learning Techniques is playing a key role in almost all domains. More specifically its role in ecommerce industry is inevitable since these techniques are able to predict and track the buying pattern of its customers. Moreover, majority of businesses put lots of effort in order to survive and gain a competitive edge over their competitors.Today,enormous research has paved way toapply differentMachine Learning models inpredicting the buyers' behavior and thereby helping the ecommerce websitesto formulate a good marketing strategy. Hence it becomes imperative for any company to better understand the customers for increasing their digital presence and predicting how customers will respond to its marketing strategies. Classification is one of the category of supervised machine learning algorithms that has its hands on various applications domains such as in credit approval, medical diagnosis, target marketing etc. in which the input data is classified in any one of the target variable. Several studies shows the application of predictive models viz., Naïve Bayes Methods, Support Vector Machines(SVM), KNearest Neighbor (KNN), Decision Tree (DT) algorithms in various domains. The main aim of this research work is to investigate KNN and DT models for predicting customer buying pattern based on the gender, age and salary. The experimental results show that DT has a higher accuracy rate than KNN in predicting purchase rate, with an accuracy rate of 95%, while the accuracy rate of KNN was 87%.

Keywords: Machine Learning, Classification, Predicting customer behavior, KNN, DT.

1Introduction

The machine learning algorithms can predict whether a particular user likes an article based on the user's personal information. This is highly beneficial for the ecommerce websites to focus on the buying pattern of the users and frame decisions which in turn helps in building better customer relationships and take better decisions. They reduce the transaction cost of finding and selecting products in an online shopping environment, and the system can predict whether a particular user likes a product based on the user's personal information.

Tracking,Buying patterns of customers are prominent and is the way to identify, investigate, and measure the customers' behaviour. This plays a key role in aiding businesses to better comprehend and potentially expand their target buyers.

Classification is a supervised technique consisting of two phases; in the first phase, there is the process of learning model construction based on the training set and in the second phase, it is used to classify new instances. Each classifier has its own significance in terms of speed, accuracy, and other issues which in turn would be helpful in developing innovative algorithms for Research community.

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2314593

This work primarily focus on the investigation of some very well-known classification algorithms, namely K nearest neighborhood and Decision tree induction applied to analysis of prediction of purchase with social network dataset [16].

This paper is organized as follow: Section II covers related works and section III discusses K- Nearest Neighbor algorithm, while decision tree induction is discussed in section IV. Finally, section V and VI discusses the results obtained and conclusion.

2 Related works

This section presents a brief survey of the KNN and decision tree induction techniques applied in various domains.

B. Charbuty and A. Abdulazeez [1] analyzed decision tree approach in detailed manner with the help of various data set. They have achieved the better performance compared to other approaches.

C. Vaca et al., [2] analyzed market predictions with decision tree model and achieved high ROC rate with the help of Social Network Advertising Sells, Organic Purchased Indicator, and Online Shoppers databases. These results showed that decision trees are upright the fluctuation and trends from market data analysis.

H. Yang et al., [3] designed a vector homomorphic encryption scheme with distributed kNN classification algorithm for supporting large-scale data classification on distributed servers to prevent information and control flow exposure with Map/Reduce architecture.

I. Ramadhan et al.[4] compared K-Nearest Neighbor (KNN) and Decision Tree (DT) algorithms and found that DT achieved a higher accuracy than KNN in detecting Distributed Denial of Service attacks in the network domain.

K.-C. Huang et al., [5] enhanced KNN with genetic programming for finding the similarities between two instances via the transformation function. This function interpreted the relationship of two data instances into a scalar differential value which indicates the dissimilarity between two instances and achieved significant results in accuracy measures.

K. Taunk et al., [6] analyzed kNN and found its weakness in this model.They developed the variants of this model for classification.

M. F. Adak and M. Uçar [7] have used decision tree based fuzzy model for a a Book Recommendation System. Decision tree based rules of the Fuzzy model used and observed that results are outperformed well for users who want to buy books on e-commerce sites.

N. N. Qomariyah et al., [8] reviewed pairwise preference problem with Decision Tree and found J48 outperformed compared to other variants. They used 10-fold cross validation for investigating to learn pairwise preferences on a specific training split point.

P. Tamrakar et al., [9] have integrated lazy learning associative classification and kNN algorithm for improving web sources while its information can be utilized in the progress of the society. Their results showed better accuracy when compared with existing lazy learning associative classifier. It generated a worthy quality nearest neighbor class association rules based on the test query and predicted the class label.

R. L. Rosa et al., [10] have proposed an event detection system. The main task of the system was to track changesof the users' behavior in an online social network. The system was able to analyze emotion identification with the help of a tree based convolutional neural network with good accuracy

S. G. K. Patro et al., [11] developed a hybrid KNN model for finding active users or products in recommendation systems. They used user behavior data for finding similarity between specific user group and target users from a huge amount of data. The model enriched the user behavior matrix and classified the features using race classifiers from both quality and quantity aspects.

S. Pathak et al., [12] investigated decision tree models like ID3, C4.5, CHAID and CART and applied pruning techniques for improving the accuracy result. These models helped the doctors to make a critical decision for a given pathology report in the medical sector.

Saadatfar et al., [13] have used K-mean classifier which clustered the data into smaller partitions and applied the KNN with pruning. The proposed approach helped to improve classification accuracy by choosing a more appropriate cluster. They found that the selection of the appropriate cluster is a challenge for performing classification with various datasets.

Víctor Adrián et al., [14] have introduced a meta-model framework which incorporates decision to access evaluation measures. They produced c4.5 variants with 10x10- fold cross validation for candidate splits. By using a Bayesian statistical analysis, they compared and ranked the evaluation measures using different databases.

Y. -H. Shih [15] incorporated a genetic algorithm based KNN for optimizing feature weights and class weights through the distance function to improve the accuracy in imbalance of classes and noisy features. The model improved the result significantly in classification.

3 K- Nearest Neighbor Classification

KNN algorithm is the simplest of all machine learning algorithms. It is based on the principle of similar samples,usuallylocatednearby. K- Nearest Neighbor is an example-based learning method. Instance-based classifiers are also called lazystudents, as they store all training samples and do not build a classifier until new unlabeled samples need to be classified. Delayed learning algorithms require less computational time in the training phase than other learning algorithms, but require morecomputational time in the classification process.

The nearest neighbor classifier is based on similarity learning,whichconsistsof comparing a given test sample with available similar training samples. Toclassifythe data sample X, find itsclosest neighbors and then assign X to the class label to which most of its neighbors belong. The choice of k will also affect the performance of the nearest neighbor algorithm. If the value of k is too small, KNNclassifier can easily be overfitted due to noise in the training data set. On the other hand, if k is too large, the closest neighbor classifier may misclassify the test samplebecause its list of nearest neighbors may contain some data points that are located far away from its neighborhood.

KNN is working basicallyin the conviction that the data is connected in thecharacteristic space.So that, all points are considered in order and find the distance between data points. The distance from EuclideanDistance Minkowski distanceused according to the data class data type used. Thisgives a single K value used to find the latest total number of unknown sample class labels. If the value of K =1 is called classificationof the nearest neighbor. Figure 2.1 shows the mean error rate of the given K value.

The working principle of the KNN classifier is as follows:

- Initialize the value of K.
- Calculate the distance between the input sample andthe training sample.
- sorting the distance.
- Take thefirst K nearest neighbors.
- Predict class labels with more neighbors for the input sample.

However, it will be complicated to determine the K value in KNN and challenging task. Figure 2.1 shows the values of k value is varied from 1 to 40 and K value is decided based on the minimum error rate which is shown in the y axis.

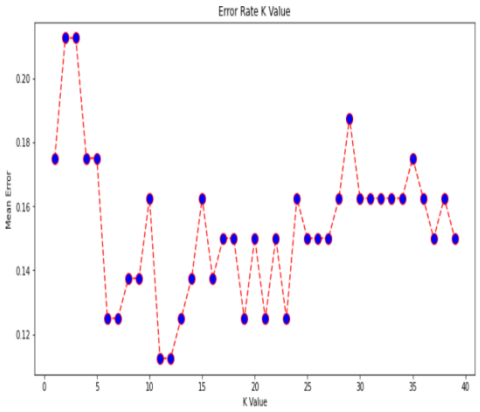

Figure 3.1

4 Decision Tree Induction

Decision tree classification technology is implemented in two stages: tree construction and tree pruning. The construction of the tree is done in a top-downmanner.At this stage, the tree is recursively partitioned until all data items belong to the same typeof label.Due totherepeated traversalof the training data set,the amount of calculation is very large.Tree pruning is done fromthe bottom up. It is used to improve the prediction and classification accuracy of the algorithm by minimizing the tree over fitting problem. The over fitting problem in the decision tree leadsfor misclassification.

Following steps are involved in implementing the decision trees:

Step 1: Select the target dataset

Step 2: Import the necessary Python Packages.

Step 3: Build a data frame

tep 4: Create the Decision Tree Model

tep 5: Predict using Test Dataset

Step 6: Predict with a New Set of Data

The basic idea behind any decision tree algorithm is as follows:

1. Select the best attribute using Attribute Selection Measures (one of the above splitting criteria) to split the records.
2. Make that attribute a decision node and break the dataset into smaller subsets.
3. Start tree building by repeating this process recursively for each child until there are no more remaining attributes.

In this work entropy is taken as a splitting criteria. Following Figure 4.1 shows the decision tree that is generated after implementing decision tree algorithm. This shows how the samples are taken based on entropy and predicted to which class it belongs to.

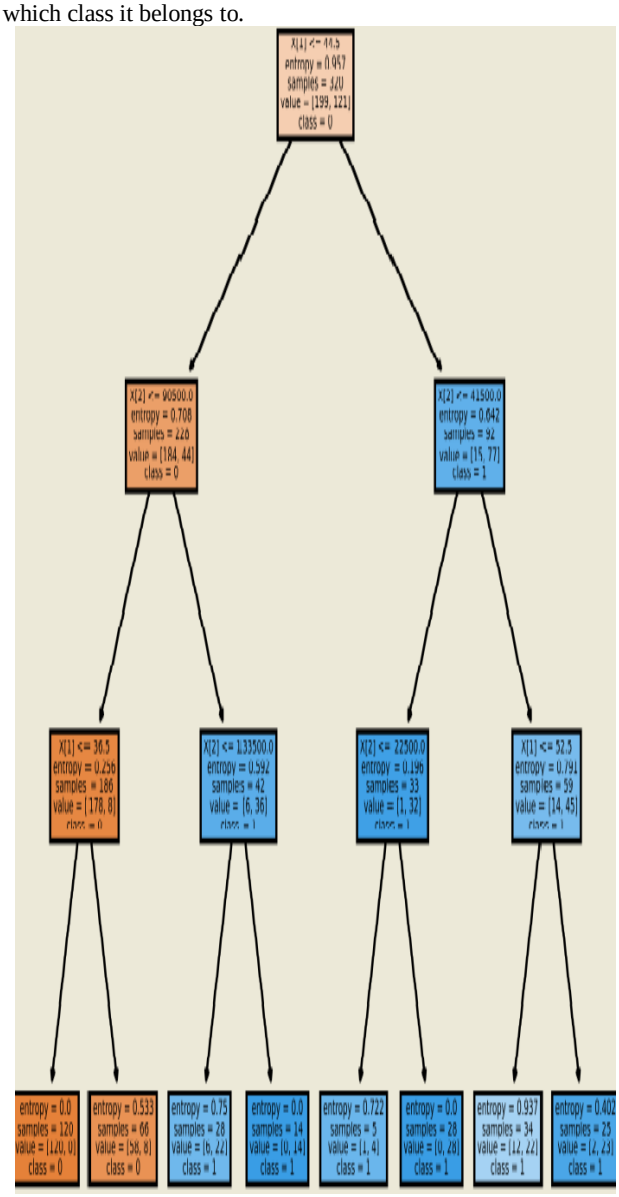

Figure 4.1

5 Results & Discussion

The simulations are done in python with Google colab with the following specifications:

i. Intel(R) Core(TM) i3 processor with4 GB RAM 3.40 GHz CPU
ii. The platform was Microsoft Windows 10.

Social Network Database [16]was considered for this study.The classification models are applied to social network dataset with 10-fold cross validation. Figure 5.1 shows the classification report generated for the KNN and DT models. Figure 5.2 shows the comparative results of the models.

KNN

Precision	recall	f1-score	support
0	0.89	0.92	0.91	52
1	0.85	0.79	0.81	28
accuracy			0.88	80
macro avg	0.87	0.85	0.86	80
weighted avg	0.87	0.88	0.87	80

Decision tree

precision	recall	f1-score	support
0	0.98	0.95	0.96	58
1	0.88	0.95	0.91	22
accuracy			0.95	80
macro avg	0.93	0.95	0.94	80
weighted avg	0.95	0.95	0.95	80

Figure 5.1

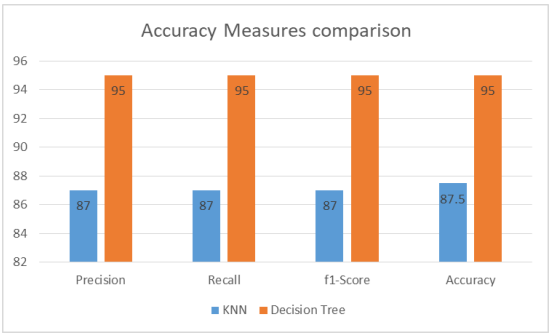

Figure 5.2

The experimental results indicate that the Decision tree has an accuracy of 95% and the KNN has an accuracy of 87%. The learning dataset is partitioned into a number of groups called “n folds”, and in this model, it is ‘10’ fold cross validation. The results may vary when we fine tune parameters and the nature of the dataset.

Conclusion

The performance of the machine learning algorithms namely KNN and DT models were tested on the social network related dataset with an intent of predicting buying behavior of the customers based on gender, age and salary. The decision tree performs well with respect to classification average accuracy and performance issues for this application. In the future study, the performance of the classifiers may be increased with the optimization of parameters of the classifiers.There is a lot of classification algorithms available now but it is not possible to conclude which one is superior to other. It depends on the application

and nature of the data. For example,the linear classifiers like Logistic regression, Fisher's linear discriminant will work better if the classes are linearly separable.

References

1. B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning", JASTT, vol. 2, no. 01, pp. 20 - 28, Mar. 2021.
2. C. Vaca, D. RiofrÍo, N. Pérez and D. BenÍtez, "Buy & Sell Trends Analysis Using Decision Trees," 2020 IEEE Colombian Conference on Applications of Computational Intelligence (IEEE ColCACI 2020), 2020, pp. 1-6, doi: 10.1109/ColCACI50549.2020.9247907.
3. H. Yang, S. Liang, J. Ni, H. Li and X. S. Shen, "Secure and Efficient k NN Classification for Industrial Internet of Things," in IEEE Internet of Things Journal, vol. 7, no. 11, pp. 10945-10954, Nov. 2020, doi: 10.1109/JIOT.2020.2992349
4. I. Ramadhan, P. Sukarno and M. A. Nugroho, "Comparative Analysis of K-Nearest Neighbor and Decision Tree in Detecting Distributed Denial of Service," *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 2020, pp. 1-4, doi: 10.1109/ICoICT49345.2020.9166380.
5. K. -C. Huang, Y. -W. Wen and C. -K. Ting, "Enhancing k-Nearest Neighbors through Learning Transformation Functions by Genetic Programming," 2019 IEEE Congress on Evolutionary Computation (CEC), 2019, pp. 1891-1897, doi: 10.1109/CEC.2019.8790163.
6. K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747
7. M. F. Adak and M. Uçar, "A Book Recommendation System Using Decision Tree-based Fuzzy Logic for E-Commerce Sites," 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2021, pp. 1-5, doi: 10.1109/HORA52670.2021.9461319.
8. N. N. Qomariyah, E. Heriyanni, A. N. Fajar and D. Kazakov, "Comparative Analysis of Decision Tree Algorithm for Learning Ordinal Data Expressed as Pairwise Comparisons," 2020 8th International Conference on Information and Communication Technology (ICoICT), 2020, pp. 1-4, doi: 10.1109/ICoICT49345.2020.9166341.
9. P. Tamrakar, S. S. Roy, B. Satapathy and S. P. S. Ibrahim, "Integration of lazy learning associative classification with kNN algorithm," 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), 2019, pp. 1-4, doi: 10.1109/ViTECoN.2019.8899415.
10. R. L. Rosa et al., "Event Detection System Based on User Behavior Changes in Online Social Networks: Case of the COVID-19 Pandemic," in IEEE Access, vol. 8, pp. 158806-158825, 2020, doi: 10.1109/ACCESS.2020.3020391.
11. S. G. K. Patro et al., "A Hybrid Action-Related K-Nearest Neighbour (HAR-KNN) Approach for Recommendation Systems," in IEEE Access, vol. 8, pp. 90978-90991, 2020, doi: 10.1109/ACCESS.2020.2994056.
12. S. Pathak, I. Mishra and A. Swetapadma, "An Assessment of Decision Tree based Classification and Regression Algorithms," 2018 3rd International Conference on Inventive Computation Technologies (ICICT), 2018, pp. 92-95, doi: 10.1109/ICICT43934.2018.9034296
13. Saadatfar, Hamid & Khosravi, Samiyeh & Hassannataj Joloudari, Javad & Mosavi, Amir & Band, Shahab. (2020). A New K-Nearest Neighbors Classifier for Big Data Based on Efficient Data Pruning. Mathematics. 8. 286. 10.3390/math8020286.
14. Víctor Adrián Sosa Hernández, Raúl Monroy, Miguel Angel Medina-Pérez, Octavio Loyola-González, and Francisco Herrera. 2021. A Practical Tutorial for Decision Tree Induction: Evaluation Measures for Candidate Splits and Opportunities. ACM Comput. Surv. 54, 1, Article 18 (April 2021), 38 pages. DOI:https://doi.org/10.1145/3429739
15. Y. -H. Shih and C. -K. Ting, "Evolutionary Optimization on k-Nearest Neighbors Classifier for Imbalanced Datasets," 2019 IEEE Congress on Evolutionary Computation (CEC), 2019, pp. 3348-3355, doi: 10.1109/CEC.2019.8789921
16 .https://www.kaggle.com/rakeshrau/social-network-ads