Stochastic Modeling and Applications
Vol. 25 No. 1 (January-June, 2021)

Received: 21st January 2021 *Revised: 13th March 2021* *Selected: 20th March 2021*

HYBRID MODEL FOR IMPROVING CLASSIFICATION ACCURACY

SATYANARAYANA AND ISMAIL BEARY

Abstract. K-Nearest Neighbors (KNN) and Decision Tree (DT) techniques are powerful and efficient toolbox for researchers to deal with challenges processed by rapid improvement in technologies in terms of time complexity and computational complexity. The main disadvantage of DT is that it assigns the same class for all the tuples which satisfies the same corresponding splitting criterion. Since all the features are used in computing similarities, KNN is sensitive to irrelevant feature. In this paper an efficient hybrid classification model based on DT and KNN is proposed which overcomes the above two problems. The prediction performance of the proposed method is also compared with DT, Support Vector Machine (SVM) and KNN model through a simulation study. The simulation results show that proposed hybrid classification model performs better than KNN, SVM and DT irrespective of sample size when the observations are from normal distribution. A simulation study to check the Robustness of proposed model about distributions is also carried out. The proposed model is also compared with above models based on 4 types of real-life datasets.

1. Introduction

Machine learning methods are extensively applied to various datasets to identify the hidden pattern and a predictive classifier cab be constructed for future decision making. K-Nearest Neighbors (KNN) and Decision Tree (DT) techniques are powerful and efficient toolbox for researchers to deal with challenges processed by rapid improvement in technologies in terms of time complexity and computational complexity. ID3, C4.5, C5.0, and CART are the most powerful and most commonly used decision tree algorithms (Anuradha and Velmurugan : 2014). ID3 is further improved by Ross and named it as C4. The Decision tree is uprooted tree like structure in which topmost node is root node and its branches are the outcomes of the test.

CART (Classification and Regression Tree) is the most popular, an efficient and widely used method for constructing decision trees introduced by Breiman et al (1984). CART considers binary split ($X_i \leq$ splitting point and $Xi >$ splitting point) for each variable based on the *splitting point*, which minimizes the error sum of squares resulting from the splitting point. For continuous variables all consecutive midpoints are considered to select the final best *splitting point*. Using minimum Gini Index criteria, the best variable at each node is selected. Gini index is biased as it searches all possible splitting and it can be overcome by

Key words and phrases. KNN, DT, SVM, Hybrid Model.

proper normalisation. Transparent in nature is one of the best advantages of DT and its structure is easy to classify and interpret. It searches all possible splits in each node with each possible splitting point and results only the best and simpler structure.

K-Nearest-Neighbour (KNN) algorithm has been identified as one of the top ten data mining algorithms because it is simple and accurate. KNN has the ability to produce simple but powerful classifiers. Sadegh et al (2013) explain that all features are used in computing similarities and hence KNN is sensitive to irrelevant features. This problem can be resolved by proper feature selection. The proposed model overcomes this problem by applying KNN to each group after selecting the significant variables from decision tree. Even though KNN takes longer time for computations, this method is extensively used in many fields because of its simplicity and reasonable accuracy. Mehmet et al (2009) proposed a hybrid classification model based on KNN, Bayesian and genetic algorithm and compared its performance using five UCI machine learning datasets. Gulnaz et al (2017) proposed a new hybrid model based on support vector machine (SVM) and KNN and showed that it has good classification accuracy compared to SVM and KNN.

The paper is organized as follows. In section 2, Methodology, flow chart of the model is presented. The simulation design to check the performance of the hybrid model and its results are presented in Section 3. In Section 4, numerical analysis with real life application is considered. Section 5 concludes the paper.

2. Methodology

The main disadvantage of Decision tree classification is that it assigns the same predicted class for all the tuples in a branch. Along with the classification accuracy, predicted class of the response variables plays vital role in decision making process. In this paper an efficient hybrid DT-KNN model is proposed to overcome the above problem. The proposed model overcomes problem associated with KNN by applying KNN to each group after selecting the significant variables from decision tree. The proposed model identifies significant variables and best splitting point based on the Gini Index. Instead of searching for K-nearest neighbours directly in the entire training data, first grouping of the elements is done based on the Gini Index criteria and then KNN is applied to each group separately.

2.1. Algorithm- Procedure.

1. Split the data set into train and test set
2. Built a decision tree using CART algorithm and identify the significant variables for train set.
3. For first leaf node note down the position of the tuples satisfying the corresponding splitting criterion, re-arrange the both response variable and independent variables according to position noted.
4. Apply the KNN classed to the arranged dataset to get the predicted class of the response variable based on appropriate value of K.
5. Repeat step 2 and 3 for all the leaf node

6. Print the predicted class s along with its original class of the response variable for the train set.

Algorithm to generate hybrid tree

Input:

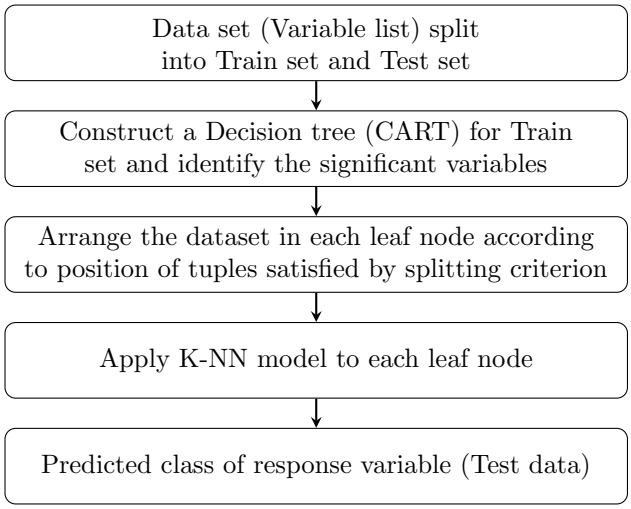
- *Data set*, Which consist of training tuples
- *Variable list*, the set of variables related to study variables
- *Variable Selection Method*, A method to identify the splitting variable and splitting point that best partition the dataset

Output: RT-KNN tree which holds predicted classes of response variable

Algorithm

(1) Create node N
(2) If tuples in dataset T, are all same class C then return N as leaf node and apply KNN to obtain the predicted class of the corresponding response variable
(3) If the list is empty then apply "*Variable Selection Method*" to determine the best splitting variable and splitting point
(4) Lable N with splitting criterion
(5) For each outcome z of splitting criterion // partition the dataset T and grow subtree for each partition
(6) Let T_z, be the set of datasets tuples in T satisfying outcome of z // a partition
(7) If T_z is empty then attach a leaf labled with the majority class in T to node N and Re-arrange the tuples in T_z and Apply KNN regression to T_z and obtained the predicted class of the corresponding response variable
(8) Else attach the node returned by 'Generate DT-KNN tree' to node N end for
(9) Return to Node N

2.2. Flow diagram of the proposed model.



2.3. Evaluating Classifier Performance. Estimating Classifier accuracy plays important role to evaluate how accurate a given classifier predict the class labels of the tuples. Most commonly used accuracy measures are total accuracy, sensitivity and specificity.

3. Simulation Study

In this section, a simulation study is carried out to highlight distinction between proposed model (Hybrid DT-KNN), DT, Gaussian kernel based SVR and KNN model. The predictive performances of these models are compared based on classification Accuracy using R. We consider the logistic model as

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3,$$

where the parametric vector is given by $\beta = (-2, 1.5, 0.5, -1)$. The decision tree was constructed using rpart package. The covariates in the simulation model X_1, X_2, X_3 are generated from normal distribution with mean vector $(13, 14, 15)$ and variance vector $(4.5, 5.5, 6.5)$ respectively. The samples size used are 30, 50, 80, 100, 300, 500, 1000. The tree was grown to consist of three leaf nodes. The threshold value for stopping parameter in DT is 0.01. For each case $5,000$ repetitions were performed and, in each simulation, tree was constructed using training data and tree performance was evaluated using independently generated test data. The appropriate value of K is taken as $K = \sqrt{number\ of\ training\ tuples}$. A simulation study to check the Robustness of the model is also carried by generating observations from multivariate t-distribution.

TABLE 1. Classification Accuracy table when observations are from Normal distribution

Sample size	**Decision Tree**	**KNN**	**Proposed model**	**SVM**
30	0.868	0.925	0.854	0.934
50	0.880	0.930	0.924	0.945
80	0.896	0.936	**0.940**	0.934
100	0.905	0.921	**0.943**	0.927
300	0.925	0.944	**0.950**	0.940
500	0.933	0.946	**0.951**	0.947
1000	0.932	0.946	**0.951**	0.942

Table 1 summarize performance of all four classification when the observations are from normal distribution. As excepted hybrid DT-KNN model has better classification accuracy than all other methods when sample size is ≥ 80. Thus, proposed model overcomes the disadvantage of RT, KNN and performance better than all other models under Total accuracy criterion.

TABLE 2. Classification Accuracy table when observations are from t-distribution

Sample size	Decision Tree	KNN	Proposed model	SVM
30	0.600	0.540	**0.600**	0.505
50	0.653	0.625	**0.720**	0.717
80	0.680	0.752	**0.845**	0.812
100	0.521	0.850	**0.905**	0.870
300	0.630	0.612	**0.720**	0.705
500	0.630	0.833	**0.930**	0.833
1000	0.856	0.889	**0.956**	0.912

Table 2 summarize performance of all classification models when the observations are from t-distribution. The proposed method has better classification accuracy than all other methods irrespective of the sample size. This shows that Hybrid method is robustness to distributions assumptions.

4. Real Life Application

The working of the proposed method is illustrated for four different real-life Datasets collected from Kaggle website. The datasets considered are

a. **Diabetes dataset:** This dataset consists of 8 medical predictor variables on 768 female patients namely number of pregnancies the patient has had, BMI, blood pressure, Skin thickness, insulin level, glucose level, diabetes pedigree function, age and a outcome variable (diabetes 1:yes, 0:no) collected from Kaggle website.
b. **Heart Disease dataset:** This Dataset consisting of 9 attributes measured on 270 patients collected for the purpose of heart disease classification of a given patient. The variables included in the study are Age, Exercise induced Angina, Gender, Serum Cholesterol level, chest pain type, fasting blood sugar, Resting Blood Pressure, maximum heart rate and a outcome variable (Heart disease 1:yes, 0:no).
c. **Glass Type dataset:** This is a Glass Identification Datasets consist of 10 attributes. The response is glass type (discrete 7 values-7 types). We restrict ourselves in this paper to two glass type classes (1 and 2). The predictors included in the study are Sodium, Silicon , Magnesium, Calcium, Aluminum, Potassium, Barium and Iron.
d. **Indian Liver Patients dataset:** This dataset contains 416 liver patients and 167 non liver patient records collected from North East of Andhra Pradesh, India. The variables included are Age, Alamine Aminotransferase, Aspartate Aminotransferase, Gender, Total Bilirubin, Alkaline Phosphotase, Total Proteins, Direct Bilirubin Albumin, Globulin Ratio, Albumin and class attribute.

All the four dataset is divided into train set and test set in the ratio 80 : 20. Each experiment is repeated 5 times with randomly assigned test and train sets and we will report average performance over 5-fold validation. $e1071$ package is

used to fit Gaussian kernel based SVM. We have used most of the default argument present in the packages.

Among the four datasets under consideration

1. Diabetes and Indian Liver patient's datasets are imbalance datasets. From the above table it is clear that DT, SVM, KNN models are not enough good to identify the second class (Specificity). But the proposed model shows high specificity as well as total accuracy and sensitivity.
2. Glass identification and Heart disease datasets are balanced datasets. The accuracy measures show that the proposed model performs better than all other model.

Data set	**Model**	**Accuracy** %	**Sensitivity** %	**Specificity** %
Diabetes	DT	62.20	85.53	15.00
	KNN	57.77	75.00	25.00
	SVM	62.29	78.45	15.79
	Proposed	**64.65**	**82.89**	**42.00**
Indian Liver patients	DT	64.36	83.07	09.09
	KNN	70.11	89.23	14.23
	SVM	68.52	87.52	21.36
	Proposed	**75.86**	**90.76**	**31.81**
Heart Disease	DT	58.53	44.44	69.56
	KNN	53.65	55.55	52.17
	SVM	65.34	50.00	72.00
	Proposed	**65.85**	**55.55**	**73.91**
Glass	DT	63.63	55.55	69.23
	KNN	72.72	88.88	61.53
	SVM	75.52	80.67	66.66
	Proposed	**77.27**	**77.77**	**76.92**

The results show that the proposed Hybrid DT-KNN classifier gives more accurate classification results compared to DT, KNN and SVM for all the four datasets under investigation.

5. Conclusion

The main focus of the study was on overcoming the disadvantages of DT and KNN method. KNN is sensitive to irrelevant or redundant variables and the proposed model overcomes this problem by applying KNN to each group after selecting the significant variables from Decision tree. The study also focusses on performance of proposed model to describe the relationship between small numbers of covariates with outcome variable in case of various sample sizes. The classification performance of the proposed method using Euclidean distance is compared with the Decision tree, SVM model and KNN in terms of the Accuracy through a simulation study. The simulation results show proposed model performs better than KNN, SVM and DT irrespective of sample size when the observations are from normal distribution. A simulation study to check the Robustness of the model is also carried by generating observations from t-distribution. The simulation results show that the proposed method has better classification accuracy than

all other methods. This shows that proposed Hybrid classification model is robust to distribution assumptions. The working of the proposed method is illustrated for four different real-life Datasets. The result shows that for both balanced and imbalance datasets under consideration accuracy, sensitivity and specificity is high for the proposed DT-KNN Hybrid model. Therefore, proposed method along with overcoming the disadvantages performs better than SVM, KNN and DT.

References

1. Berk, R. A.: *Statistical learning from a regression perspective*, Springer Science & Business, New York, 2008.
2. Bhatia, N. and Vandana.: Survey of Nearest Neighbor Techniques, *International Journal of Computer Science and Information Security*, **8(2)** (2010) 302–305.
3. Breiman L., Friedman J., Olshen R. and Stone C.: *Classification and regression trees*, CRC Press, Boca Raton, 1984.
4. Chakraborty T., Chakraborty A. and Mansoor Z.: A hybrid regression model for water quality prediction, *OPSEARCH*, **56(4)** (2019) 1167–1178.
5. Chomboon, K., Pasapichi, C., Pongsakorn, T., Kerdprasop, K. and Kerdprasop, N.: An empirical study of distance metrics for k-nearest neighbor algorithm, in: *The 3rd International Conference on Industrial Application Engineering* (2015) 280–285.
6. Cover, T. and Hart, P.: Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, **13(1)** (1967) 21–27.
7. Esposito F., Malerba D., Semeraro G. and Kay J.: A comparative analysis of methods for pruning decision trees, *IEEE Trans Pattern Anal Mach Intell.*, **19(5)** (1997) 476–491.
8. Gulnaz Alimjan., Tieli Sun., Hurxida Jumahun. and Yu Guan., Wanting Zhou. and Hongguang Sun.: A Hybrid Classification Approach Based on Support Vector Machine and K-Nearest Neighbor for Remote Sensing Data, *International Journal of Pattern Recognition and Artificial Intelligence*, **31(10)** (2017) 1750034.
9. Hothorn T., Hornik K. and Zeileis A.: Unbiased recursive partitioning: a conditional inference framework, *J Comput Graph Stat.*, **15(3)** (2006) 651–674.
10. Hothorn T. and Zeileis Partykit A.: a modular toolkit for recursive partitioning in R. *J Mach Learn Res.*, **16** (2015) 3905–3909.
11. Hothorn, T., Hornik, K. and Zeileis, A.: Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, **15** (2006) 651–674.
12. James, G., Witten, D., Hastie, T. and Tibshirani, R.: *An introduction to statistical learning with applications in R*, Springer, New York, 2013.
13. Loh W and Shih Y.: Split selection methods for classification trees, *Stat Sin.*, **7(4)** (1997) 815–840.
14. Mehmet Aci., Cigdem Inan., and Mutlu Avci.: A hybrid classification method of k-nearest neighbor, Bayesian methods and genetic algorithm, *Expert Systems with Applications*, **37** (2010) 5061–5067.
15. Shih Y.: A note on split selection bias in classification trees.*Comput Stat Data Anal.*, **45(3)** (2004) 457–466.
16. Vale, C. D., and Maurelli, V. A.: Simulating multivariate non-normal distributions. *Psychometrika*, **48** (1983) 465–471.

Satyanarayana: Department of Statistics, Mangalore University, Mangalagangothri, Mangaluru, Karnataka - 574199, India

Email address: sathya1301@gmail.com

Ismail Beary: Department of Statistics, Yenepoya (Deemed to be university), Deralakatte, Mangaluru, Karnataka - 575018, India

Email address: ismailb@yenepoya.edu.in, prof.ismailb@gmail.com