A Bibliographic Survey of Sentiment Classification using Hybrid Ensemble-based Machine Learning Approaches

Rajni Bhalla [a], Amit Sharma [b], Geetha Ganesan [c]

[a] *Lovely Professional University, Jalandhar, India*
[b] *Lovely Professional University, Jalandhar, India*
[c] *Advanced Computing Research Society, Chennai, Tamilnadu, India*

Abstract

The rapid number of reviews on different fields have contributed to the rising field of data analysis. Several methods are existing for data analysis but there is a need to find the right methodology that can provide better accuracy. The objective of the paper is to find an accurate method depending upon the type of dataset. Previous researches have primarily relied on using the KNN approach and issues for deciding the K-value. For the research work, the data from the Statistics Department of the University of Wisconsin-Madison has been taken to evaluate the teacher performance. The hybrid approach uses three different machine learning models for prediction. The prediction model was tested effectively using the teacher assistant evaluation dataset. The hybrid approach has been developed to improve the identification of teacher performance. Our findings indicate that on combining KNN, decision tree, and naïve Bayes, there is a considerable increase in the performance of the prediction analysis. The results have shown that the hybrid approach called KDN (KNN, Decision Tree, Naïve Bayes) obtained better results with 53.04% accuracy as compared to the baseline system performance.

Keywords 1

Hybrid approach, KNN, Classification, machine learning

1. Introduction

Nowadays, most academic institutes face a low-quality problem in the educational field. One of these factors is educational student achievement and teacher assistant teaching quality. Some studies had been done to engorge the students to improve their academic achievement, but still, the problem of the teaching quality needs to be improved especially in the practical parts that are normally performed by the Teacher Assistants.

In this paper, the Hybrid approach is applied for checking the performance of the teacher. Naïve Bayes, KNN, and decision trees are the best examples of supervised learning where data is already labeled. A decision tree might your a good starting point. A decision tree is generated using a decision tree classifier that gives a clear visual.

K-nearest neighbor (K-NN) classification is a labor-intensive algorithm best adopted in the situation of the large training dataset. The algorithm is found to conform to the Euclidean distance measure in terms of the distance matrix.

One of supervised learning algorithm is Naive Bayes. Naive Bayes is also known as linear classification method. On the contrary, K-NN is not a linear classifier. When we process data using KNN, there are lot of calculations need to perform on each step. This is the main reason K-NN is unable

WAI-2022: Workshop on Artificial Intelligence, January 27 – 28, 2022, Chennai, India.
EMAIL: geetha@advancedcomputingresearchsociety.org (Geetha Ganesan)
ORCID: 0000-0001-7338-973X (Geetha Ganesan)

to process large amount of data. Both Naive bayes and KNN are powerful techniques. Naive Bayes is preferred over KNN when we need to process data considering speed. If you can't pick between the three, your best strategy is to mix them all and run a test on your data to determine which one delivers the greatest results.

The suggested method's technique is described in Section 2. A quick summary of the datasets is explained in Section 3. The collected results and consequences of the study are presented and compared with other methods in Section 4. This research comes to an end in Section 5.

2. Literature Review

The detecting methods used in earlier models are introduced in this section. Then we compare and contrast these strategies with those utilized in the proposed model. k Nearest Neighbors (KNN) is a common and extensively used technique for classification [1] [2], clustering[3], and regression [4] in a variety of research areas, including economic modelling [5], image interpolation (Smith et al., 1988), and visual category recognition (Smith et al., 1988). (Zhang et al., 2006). A hybrid and layered Intrusion Detection System (IDS) is suggested, which employs a mix of machine learning and feature selection approaches to deliver high-performance intrusion detection in a variety of assault types [6]. Designing a hybrid analysis is designed to increase the capacity to maintain significant findings and well-supported outcomes by combining traditional statistical analysis and artificial intelligence technologies[7]. We believe that a hybrid strategy that incorporates both machine and human-centered features can achieve greater efficacy, competence, and social significance than either method alone[8].

3. Methodology

3.1. Dataset Description

The dataset has been taken from the UCI repository. The statistics come from assessments of 151 teaching assistant (TA) assignments. By splitting the scores into three groups of about similar size, the class variable was produced ("low," "mid," and "high").

4. Experiment and Results

The analysis design is a combination of several stages and each stage contains a different number of steps as shown in figure1. Firstly, the teacher assistant dataset is retrieved and the rename operator is used to rename the English Speaker attribute. In the second phase, the Spilt Validation operator is used to divide the dataset into two groups; one potion for training data and the other for testing data, and in the third phase, the KNN operator, Decision tree, Naïve Bayes, and hybrid approach is used to train the data and then apply model operator is used for testing the data. In the fourth phase, the different-different models (KNN, decision tree, naïve Bayes, and hybrid) are applied that represent a sample, and a data accuracy algorithm is used to get the performance. The fifth phase represents the results in graphical shape.

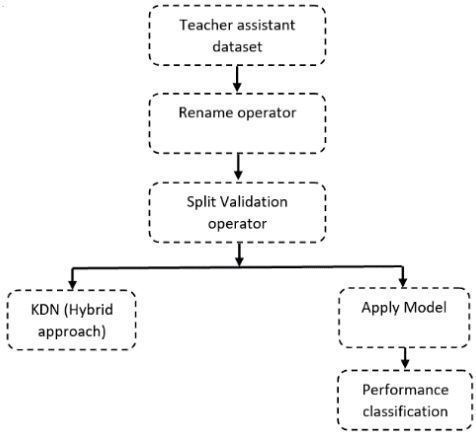

Figure 1: Pictorial representation of the methodology

4.1. KNN

K-nearest neighbours (KNN) is a simple, easy-to-implement supervised machine learning approach that may be used to solve both classification and regression problems. The KNN algorithm believes that objects that are similar are near together. To put it another way, related items are close together. The KNN algorithm relies on this assumption being correct in order for it to work. KNN combines the concept of similarity (also known as distance, proximity, or closeness) with some basic mathematics, such as computing the distance between points on a graph.

4.2. Naive Bayes

The Bayes' Theorem is used to produce the Naive Bayes classifiers, which are a set of classification algorithms based on the Bayes' Theorem. It's a group of algorithms that all work on the same principle: each pair of categorizing features is independent of the others.

4.3. Decision Tree

Decision tree is one of the powerful techniques that has been used for prediction. A decision tree always presented the result in the form of decision tree. The results of all three algorithms will be compared using ensemble approaches.

4.4. Results

The analysis of the proposed model achieved different shapes of results in the training and testing stages. By using these results, the performance of the Teacher Assistant can be analyzed and controlled. The performance output is analyzed based on accuracy, and prediction error.

Table 1: Data Performance using KNN
Accuracy: 47.83%

	True3	True2	True1	Class precision
pred3	9	4	4	52.94%
pred2	4	8	6	44.44%
pred1	3	3	5	45.45%
Class recall	56.25%	53.33%	33.33%	

We used the KNN approach to evaluate teachers and obtained a 47.83 percent accuracy, as shown in Table1. When we use naïve Bayes, we got 42.38% accuracy as shown in Table2. At the time of the decision tree, we got 37.04% accuracy as shown in Table3. We need to work on the performance of the model.

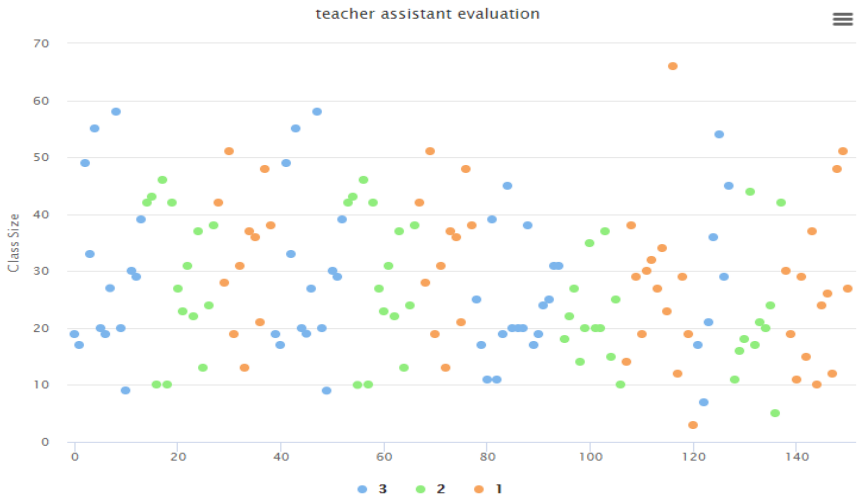

Figure 2: Scatter Plot showing Category

Table 2: Data Performance using Naïve Bayes
Accuracy: 42.38% +/-11.77% (micro average: 42.38%)

	True3	True2	True1	Class precision
pred3	41	34	31	36.68%
pred2	8	10	5	43.48%
pred1	3	6	13	59.09%
Class recall	78.85%	20.00%	26.53%	

Table 3: Data Performance using Decision Tree
Accuracy: 37.04% +/-6.79% (micro average: 37.09%)

	True3	True2	True1	Class precision
pred3	47	41	47	34.81%
pred2	4	8	1	61.54%
pred1	1	1	1	33.33%
Class recall	90.38%	16.00%	2.04%	

Finally, a vote operator has been used to combine KNN, Naive Bayes and decision tree and performance has been compared with individual models as shown in Table 4**Error! Reference source not found.**.

Table 4: Data Performance using Hybrid Approach
Accuracy: 53.04% +/-8.62% (micro average: 52.98%)

	True3	True2	True1	Class precision
pred3	39	20	19	50.00%
pred2	10	23	12	51.11%
pred1	3	7	18	64.29%
Class recall	75.00%	46.00%	36.73%	

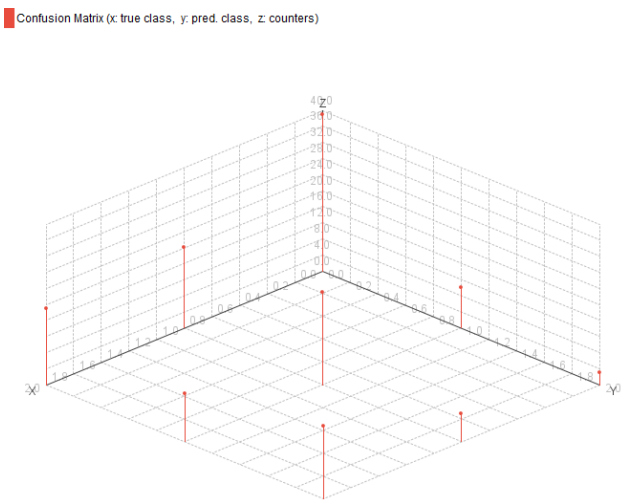

Figure 3: Plot view using hybrid approach

It is clear from Table4 and Figure 3 that hybrid produces better results as compared to the model.

5. Conclusion

This study was conducted to check the performance of different machine learning models after performing data analysis on teaching assistant evaluation. The purpose of this research is to identify effective strategies that can find an accurate model from several prediction models. As per previous studies, there can be no doubt that existing methodologies like KNN, decision tree, and naïve Bayes have proven great methodologies. As per result, KDN proved better in terms to find the accuracy of the model. A hybrid classification approach that incorporates the KNN algorithm, Decision tree, and Naive Bayes is presented here. This analysis adopts the prediction process based on the data size, time process, accuracy, estimated error factor tried to investigate and evaluate the teacher assistant. The results of the evaluation were obtained using the different sizes in the training and testing phases. The deep examinations highlighted that the group of 53.04% achieved better results in the prediction accuracy, estimated time, and error factor. In the future, we'll look at different distance and similarity options that might help us to get a more precise distance or similarity measurement. To suggest a measurement with a reduced computational cost a method of categorization that is more effective and efficient.

6. References

[1] C. H. Wan, L. H. Lee, R. Rajkumar, and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine," *Expert Syst. Appl.*, vol. 39, no. 15, pp. 11880–11888, 2012, doi: 10.1016/j.eswa.2012.02.068.

[2] Z.-H. Z. Min-Ling Zhang, "A k-nearest neighbor based algorithm for multi-label classification," *IEEE Int. Conf. Granul. Comput.*, vol. 2, no. 2, pp. 718–721, 2005.

[3] Q. B. Liu, S. Deng, C. H. Lu, B. Wang, and Y. F. Zhou, "Relative density based K-nearest neighbors clustering algorithm," *Int. Conf. Mach. Learn. Cybern.*, vol. 1, no. November, pp. 133–137, 2003, doi: 10.1109/icmlc.2003.1264457.

[4] J. K. Solano Meza, D. Orjuela Yepes, J. Rodrigo-Ilarri, and E. Cassiraga, "Predictive analysis of urban waste generation for the city of Bogotá, Colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks," *Heliyon*, vol. 5, no. 11, p. e02810, 2019, doi: 10.1016/j.heliyon.2019.e02810.

[5] Xiao-Gao Yu and Xiao-Peng Yu, “New K-nearest neighbor searching algorithm based on angular similarity,” in *2008 International Conference on Machine Learning and Cybernetics*, Jul. 2008, pp. 1779–1784, doi: 10.1109/ICMLC.2008.4620693.

[6] Ü. Çavuşoğlu, “A new hybrid approach for intrusion detection using machine learning methods,” *Appl. Intell. 49*, vol. 7, no. 49, pp. 2735–2761, 2019.

[7] F. Costa-Mendes, Ricardo and Oliveira, Tiago and Castelli, Mauro and Cruz-Jesus, “A machine learning approximation of the 2015 Portuguese high school student grades: A hybrid approach,” *Educ. Inf. Technol.*, vol. 26, no. 2, pp. 1527-1547 (Springer), 2021.

[8] A. Sartas, Murat and Cummings, Sarah and Garbero, Alessandra and Akramkhanov, *A human machine hybrid approach for systematic reviews and maps in international development and social impact sectors*, vol. 12, no. 8. Multidisciplinary Digital Publishing Institute, 2021.