HYBRID CLASSIFICATION MODEL IN MACHINE LEARNING: A REVIEW

Department of Computer Science and Technology, College of Technology and Engineering, MPUAT, Udaipur, India

Authors name : Suhani Jain
Department of Computer Science and Technology
College of Technology and Engineering, MPUAT
Udaipur, India
Email : suhani.jain71@gmail.com

Authors name : Naveen Choudhary
Department of Computer Science and Technology
College of Technology and Engineering, MPUAT
Udaipur, India
Email : naveenc121@yahoo.com

Authors name : Kalpana Jain
Department of Computer Science and Technology
College of Technology and Engineering, MPUAT
Udaipur, India
Email : kalpana_jain02@rediffmail.com

Abstract— Huge amounts of data are generated from various sources as a result of the digital explosion, which necessitates critical analysis in decision-making. Many academics working in the field of massive data sets have recently been interested in one of the most difficult topics and categorization. The machine learning method is used for a variety of data sets, including images, text, URLs, emoticons, and so on. However, since its discovery, it has been known that including a learning algorithm results in a greater comprehension and acceptance of a solution in terms of variability and accuracy In this paper, a comprehensive study of machine learning strategies based on different hybrid method of classification on the basis of divisions was conducted .to improve efficiency by combining multiple learning algorithms to achieve better predictable performance. In addition, towards the conclusion of the paper, some suggestions are made regarding how strong the analysis will be. Machine-based classification frequently uses algorithms, which are computer programs written in a variety of mathematical formats to speed up the automated process. To successfully execute these compatible classification jobs, such advanced, robust, fast, and reliable algorithms are required, given the dramatic increase in the size and complexity of computer data today. We evaluate and contrast a range of classification models in this paper. We'll go through the advantages of hybrid models over traditional classification algorithms. A thorough examination of all algorithms and their benefits in terms of accuracy, precision, and efficiency. We will discuss why scientists are working on hybrid models these days and putting greater emphasis on the correctness of the results. Large data sets are difficult to manage, so we need a capable system to produce them, which we are doing currently with a hybrid approach. The systematic comparison is mentioned at the end of this paper.

Keywords— machine learning, algorithms, data, classification.

I. INTRODUCTION

Machine Learning is a branch of Artificial Intelligence (AI) that interprets raw data and converts it into meaningful information. With today's exponential development in the amount of data available, it's more important than ever to create and execute appropriate data classification strategies that can handle and perform data analysis efficiently and effectively. Vector Support Machines (SVMs), Artificial Neural Networks (ANNs) networks, and other popular methodologies are employed to operate the aforementioned systems. Many advanced classification approaches are combined or combined with modern development techniques in such classification systems to improve class accuracy and reduce calculation time. Looking for the correct solution to any problem in a calculating situation is topologically equivalent to looking for a little PIN in a dark room. These searches may be extremely possible and important, but they may also be more expensive if done on a computer. If it is known that the PIN is somewhere else in the room, the cost of all calculations will be lower, and the probability of discovering an item will be higher. This condition is linked to gradient-increasing and gradient-decreasing techniques. Such tactics enable the calculating algorithm to identify suitable search regions, and searches are then conducted in those areas to ensure consistent results throughout runs. Some of the most often used gradient reduction approaches in neural networks include the BPNN, ERNN, and LM algorithms. Machine Learning Approach is a type of practical intelligence that uses existing data to train a model that can predict future behaviors, results, and trends using new test data. Supervised learning (SL) is a machine learning activity that guesses from labelled training data, whereas non-supervised learning (USL) is a type of machine learning algorithm used for drawing imaginable databases that include non-labeled data and is included in the supervised and supervised learning category. In supervised learning, there are rules set, and the outcomes are known while reading unchecked. The algorithm reads them according to particular rules, and the result is provided. Popular classification methods such as Naive Bayes (NB), Vector Support Machines (SVM), K-Nearest (KNN) Neighbors, Maximum Entropy (ME), and others are used in many sensory analysis roles, although it is unclear which one performs best. To obtain the desired accuracy in a common

domain Data classification is the process of sharing data in order to make it more efficient and effective. The classification of data is a two-step procedure: 1. Learning (Model Design) 2. Planning (Model Application) .The separator is constructed in the first stage by establishing a preset set of data classes or concepts. This is a learning step (or training phase) in which the classification algorithm reads a training set made up of copies of webpages and related class labels to develop class distinctions. The model is utilized for separation in the second stage. A test set is employed, which consists of test tuples and their corresponding class labels. These tuples are chosen at random from a large data source. They are not used to generate a separator since they are not independent of training tuples. However, a large number of studies in recent years have found that combination-based learning causes inefficiencies in one basic component in the classroom accuracy of forecast A group of dividers is a collection of dividers whose individual decisions have been brought together in a certain way to produce a consensus decision. Many machine learning programs have supervised functions. The current paper focuses on the tactics required to do this. This work, in particular, addresses issues of segregation in which exclusivity only allows for alternative, informal values.

LITERATURE REVIEW

In the k of the nearest neighbor is the constant value of k. Baoli et al. they had chosen different values of k for each class instead of constant k value, and thus had made the most critical validation (Baoli, Shiwen, & Qin, 2003). Yildiz et al. used k a nearby neighbor algorithm to specify unwanted emails once parallelism process used to reduce duration (time spent) (Yildiz, Yildirim, & Altilar, 2008).

In 2015, Junyi Xuet al. introduced the group a method of learning i.e. argument based on many things Shared Reading (AMAJL), which combines ideas from multi agent conflict, integrated learning, and integration law mines. At AMAJL, the opposition technology is introduced as an integration strategy to integrate multiple foundation dividers and produce a high performance separator. See design a controversial framework named by the Arena as communication platform for information integration. Through contradictions based on shared learning, a higher level of personality information can be extracted, thus becoming a purest world the knowledge base can be created and used independently separation. They do a lot of experiments on many public data sets use AMAJL and another benchmark methods. The results showed that their approach was possible successfully extract high quality integration information separate and improve the functioning of the partition

Xueyi Wang proposed a novel approach known as COB (core, outlier, and boundary) average measures the accuracy of most voting collections binary separation. The author says that this is a good collection methods require precise and varied individual class dividers. Data items were first divided into three sub-sets, total, external and border based on each prediction. Individual arrangements of ensembles and finally the accuracy of the compilation method was recorded In each subset and the results were grouped together. At work, the author used bags, random forest and random collection as three separate clusters and cutting trees, neighbors near k, and support equipment were used as machine learning algorithm for models.

PROPOSED METHOD

Supervised learning, in this paper we describe the strategies for classification in supervised reading. In the supervised learning, we divide the entire database into two parts one for training when the class divider learns that the data and the remaining data are used to check the accuracy of the separator. Once that is done then we may use it to test new data for future information from these supervised reading class dividers. The supervised learning categories are divided into five main groups of differentiation algorithms based on Frequency Table, Covariance matrix, parallel measurement, Vector & margin and Neural Network. From this partition group we have different partitioning algorithms.

1. Naive Bayes:

The Nave Bayes set of rules is a type of supervised learning system for solving different classification issues that is mostly based on Bayes theorem. It's most commonly used in text formats that include a high-dimensional dataset. The Nave Bayes Classifier is a unique and simple classification technique that aids in the development of rapid device studying models that can generate quick predictions. It's a type of probabilistic classifier, which means that it predicts based on the probability of an items. It's often referred to as Nave since it assumes that the presence of one characteristic is unrelated to the presence of other features. For example, if the fruit is identified by its color, shape, and flavor, purple, round, and sweet are the most likely candidates. Bayes: it is referred to as Bayes as it relies upon at the precept of Bayes' Theorem. Bayes' theorem is likewise referred to as Bayes' Rule or Bayes' regulation that is used to determine the opportunity of a speculation with prior know-how. It depends on the conditional opportunity.

Python Implementation of the Naïve Bayes algorithm:

- Statistics Pre-processing
- Associating with the schooling set as a Naive Bayes
- Forecasting the outcome of the examination
- Verify that the outcome is correct (advent of misunderstanding matrix)
- Visualizing the outcome of the lookup table.

The Naive Bayes Classifier has the following advantages:

- Naive Bayes is a fast and simple machine learning algorithm for predicting a class of datasets.
- It can be utilized for both binary and multi-magnitude classifications.
- In comparison to the opposing Algorithms, it performs well in multi-class predictions.

2. Decision Tree Classification Algorithm

The Decision Tree is another type of supervised learning technique that can be used to solve different types of classification and regression problems,

however it is most typically employed to solve classification problems. It's a tree-based classifier, with core nodes each representing dataset attributes, different branches representing choice policies, and each leaf node representing the final outcomes.
There are two nodes in a decision tree: the decision Node and the Leaf Node. Leaf nodes are the output of those selections and do not have any more branches, whereas choice nodes are utilized to make any decision and have a couple of branches. The selections or tests are carried out based on the characteristics of the given dataset. It's a graphical depiction for acquiring all viable solutions to a problem/selection depending on specified conditions. It's termed a selection tree because, like a tree, it starts with the root node and grows into a tree-like structure with additional branches.

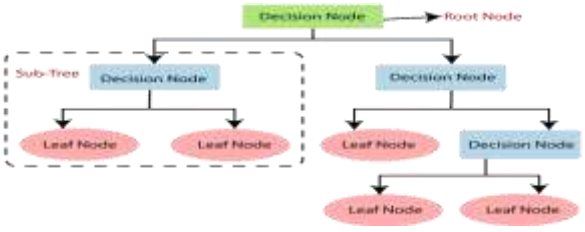

Figure 1 : Decision tree representation

Steps for Decision Tree Classification Algorithm:
Step 1: Starting the tree with the root/foundation node, which contains the complete entire dataset.
Step 2: Using the attribute choice measure, find the best attribute in the dataset (ASM).
Step 3: Subdivide the S into different subsets that comprise possible first-class attribute values.
Step four is to create the decision tree node with the high-quality attribute.
Step 5: Create new choice trees in a recursive repetitive manner.

The decision tree method is continually attempting to maximize the information's value, as well as the node / attribute that may be obtained. It can be calculated using the formula below:

Information Gain= Entropy(S)-
[(Weighted Average) *Entropy(each features)

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s) = -P (yes) log2 P (yes) - P (no) log2 P (no)
Where,
S= Total number of samples
P(yes)= probability of yes

The Decision Tree's Advantages:
1. It is simple to comprehend because it follows the identical steps that a person would take while making a decision in real life.
2. It can be extremely useful in addressing decision-making issues.
3. Thinking about all of the possible outcomes of a situation is beneficial. In comparison to previous techniques, data cleaning is not required.

3. Random Forests

"The Random Forest is a subdivision that contains a number of decision trees for the various datasets given and makes measurements to improve the database's prediction accuracy," as the name implies. Instead than depending on a single decision tree, the random forest takes a prediction from each tree and predicts the eventual conclusion based on several predictable votes. The forest's enormous number of trees ensures great accuracy while also preventing congestion. When compared to other algorithms, it takes less time to train. It has a high degree of accuracy in predicting the result, and it works well even with enormous databases. When a considerable amount of the data is missing, it can still maintain accuracy.

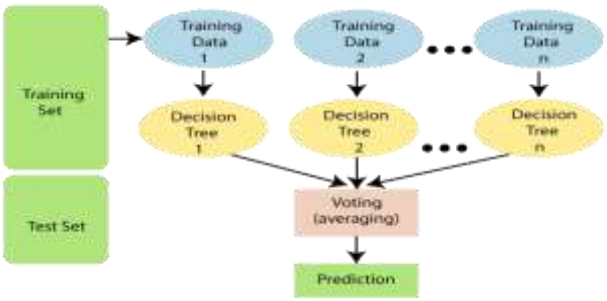

Figure 2 : Random Forest Representation

Steps for Random Forest Algorithm:
Step 1: Pick K data points at random from the training set.
Step 2: Create decision trees based on the data points you've chosen (Sub-items).
Step 3 : Decide on how many decision trees who wish to make.
Step 4: Go through Steps 1 and 2 again.
Step 5: Get a prediction for each decision tree for new data points, and allocate new data points to the category with the most votes.

Random Forest Advantages:
1. Both planning and retrieval activities are possible with Random Forest.
2. It has the ability to handle big data sets.
3. Improves model accuracy and prevents over-installation problem and Injustice of the Unplanned values.

4. KNN (K-NEAREST NEIGHBOUR)
The neighborhood near K (k-NN) approach is one of the most basic and widely used nonparametric sample separation methods (Bishop, 1995; Manocha and Girolami, 2007). Specifies the average distance between the points on the input vectors, and subsequently acts as a marker for the unmarked point in the class of its nearest neighbors. K is a key parameter in the process of generating a k-NN classifier,

and different k values will result in variable performance. If the k is set too high, the neighbors who used the prediction will spend a lot of time dividing and contributing to the forecast's accuracy. Model-based learning is a term used to describe how K-NN differs from other teaching methods (Mitchell, 1997). As a result, it lacks a model for input vectors tests and separates new conditions. As a result, k-NN On-line 'trains examples and locates a neighbor around k for a new event. A simple algorithm called Nearby Neighbors keeps track of all available examples and classifies incoming cases based on a similarity rate (e.g., grade activities). Based on their closest neighbors, KNN is employed in mathematical measurement and pattern recognition and should have an odd number. Distance characteristics such as Euclidean, Manhattan, and Minkowski are used.

Let's say we have a photo of a creature that looks like a cat or a dog, but we don't know whether it's a cat or a dog. As a result, we may utilize the KNN method in this study because it operates at the same rate. Our KNN model will extract the same features from the fresh data set in photographs of cats and dogs, and categories them as cats or dogs based on the similarities.

Figure 3 : KNN classifier example

Steps for KNN Algorithm:
Step 1: Pick a neighboring K number.
Step 2: Determine the Euclidean distance between the two K numbers.
Step 3: Using the estimated Euclidean distance, find the neighbors who are closest to K.
Step 4: Count the number of data points in each category for these k neighbors.
Step 5: In the section where the neighbor number is the limit, assign new data points.
Step 6: We've completed our model.

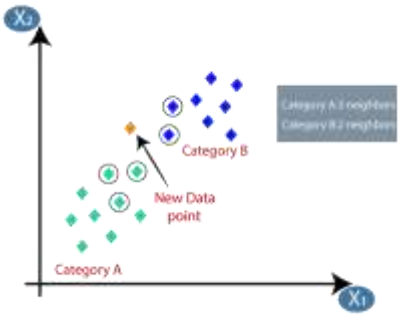

Figure 4: KNN classifier graph

KNN Algorithm Advantages:
1. Easy to operate.
2. Relying heavily on solid training data

5. SUPPORT VECTOR MACHINE

Vapnik proposes vector support equipment (SVM) (1998). SVM discovers a fine-grained aircraft for the high-resolution area after converting the vector input map into a large-size feature space. Furthermore, the decision boundary, i.e. the hyper-plane separator, is determined by the supporting vector of most training samples and so poses the greatest hazard to outsiders. The SVM separator, in particular, is designed to be separated into binary categories. To put it another way, divide a set of training vectors into two groups. The supporting vectors are training samples that are near the decision limit. A user-defined parameter called penalty factor is also available in SVM. Allows users to choose between a random number of samples and the resolution's width.

Hyperplane and Vectors Support for SVM:
Hyperplane: There are various lines / barriers for dividing classes into n-dimensional space, however we need to identify the optimum decision line for separating data points. The SVM hyperplane is a term used to describe the optimal boundary. The hyperplane's size is determined by the database's characteristics, which implies that if there are only two components (as shown in the image), the hyperplane will be a straight line. The hyperplane will be a plane with two sizes if there are three elements. We always construct a hyperplane with a large limit, implying that the distance between data points is greater. Data points or vectors that are closest to the hyperplane and have an effect on the hyperplane area are called supporting vectors.

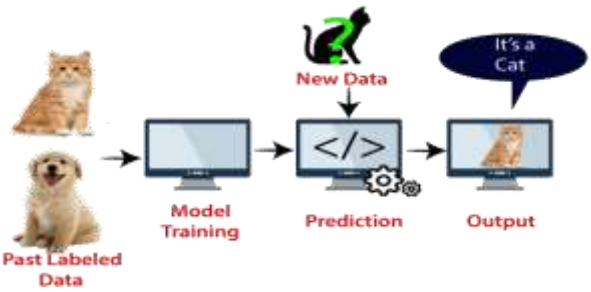

Figure 5: SVM example

Steps for SVM Algorithm:
Step 1: Create a variety of hyperplanes and then choose the best one.
Step 2: Optimize the hyperplane so that the margin between the classes is as large as possible.
Step 3: The kernel trick for non-linear hyperplanes, which is used in the SVM technique for linear hyperplane misclassifications.
Step 4: For a high-dimensional space, we reformulate the issue so that the data is implicitly mapped to it.

HYBRID CLASSIFICATION

How to Expect Progressively Internal Series Using Integrated Machine Learning Methods. With easy recording of music derived from midi files, you can train and test work on data outside of the Artificial Neural Network or the usual state power machine. A form of isolation that often entails a small number of separate procedures running in parallel. Each solution handles a distinct problem, and the choice to separate is made in the same way. The basic purpose of IDS development is to attain the highest level of accuracy in the task being done. This goal is natural, and it leads to the creation of mixed roadways to tackle the problem. A hybrid classifier is a machine learning strategy that combines different machine learning algorithms to improve system performance. In most cases, the hybrid technique is made up of two functional components. The first uses raw data as an input and generates intermediate outcomes. The second will then use the intermediate outcomes as input to generate the ultimate results (Jang, Sun, & Mizutani, 1996). Mixed divisions, in instance, may be based on the classification of many categories, such as ambiguous tactics. Mixed class dividers, on the other hand, can pre-process input samples using a composite-based strategy to eliminate bias. Training examples from each class the clustering results are then used as stage design training models. As a result, the initial level of mixed class dividers could be based on either supervised or unsupervised learning approaches. Finally, blended categories can be based on a mix of factors.

Proposed hybrid method

Bayesian methods and genetic algorithms are combined with the proposed hybrid k in the local vicinity. The fundamental concept is to use current data sets to verify them in some way. This hybrid's steps

The following is a summary of the method:

Step 1: The selected data set is first subjected to the Bayesian technique based on the predicted expansion algorithm. As test data, all of the data from the legitimate data set is used. The data was then segregated in the improper number of ways. Assume that this number is x.

Step 2: In this step, a new data set is generated at random using the old data set's maximum and minimum values for each category. The new data collection has twice as many elements as the previous one.

Step 3: This phase employs the nearest neighbor approach. For each data set, the average distances of each class are determined. The data is sorted by distance, and the section with the shortest distance is chosen. As a result, the value of k is chosen to be 50% of the total data number.

Step 4: In this step, the prior step is repeated. Test data is the same as the original data set, whereas train data is made up of produced data. A method for increasing expectations and the number of erroneously classified data has been discovered. Y is the name of this number. It has been determined that data has been categorized incorrectly. This number is known as y.

Step 5: The values of x and y are also compared in this step; if x is smaller than or equal to y, a new batch of created data will not be better than the previous one. As a result, Steps 2-4 must be repeated. This loop will continue until the value of x is smaller than the value of x. When this occurs, it signifies that the new set of data is better for training than the old one.

Step 6: After this step, a new loop is started with the loop number assigned to the user based on the data set values. The genetic algorithm is then applied to the final collection of data. The information is accurate.

Step 7: The new data set and the number of wrongly ordered data termed z are both applied to Step 1. When z is less than y, the new data set is better than the adult data set. A new data collection is added to the loop.

Step 8: This is the final step, and when the loop has completed, a set of advanced data has been detected, as well as a limited amount of inaccurate categorization data.

Results and conclusion

As previously said, ensemble based learning techniques for sentiment classification are generally used to improve a model's classification, prediction, function approximation, and performance, or to lessen the risk of an unintentional poor model selection. To solve multiclass and regression problems, an AdaBoost-based ensemble technique can be used to manipulate the weight factor of correctly and erroneously categorised instances to obtain the desired accuracy. The limitations of automated classification systems in high-dimensional classification tasks are due to concerns with machine-based hybridized classification techniques. It is critical to create classification algorithms that can greatly cut and decreases computational times and enhance classification accuracy for such datasets in order to perform and aid efficient classification for such datasets. As a result, improving the optimization capabilities of traditional optimization algorithms is necessary to minimize processing times in hybridized classification systems using optimization algorithms and enhance overall classification accuracy. The method is suggested for low-cost hardware-based clustering solutions, applications that categories noisy data sets, and data sets with few data. According to test results, the suggested technique outperforms expectation maximization classifiers in terms of classification performance. The strategy is useful for those types of data sets with a small number of data points. It generates and creates unlimited data that have similar and likewise characteristics with original data and then improve them according to the suggested algorithm.

References

[1] Abu-Nimeh, S., Nappa, D., Wang, X. and Nair, S. 2007. A comparison of machine learning techniques for phishing detection. Anti-Phishing Working Groups Ecrime Researchers Summit, pp. 60–69.

[2] Almseidin, M., Zuraiq. A.A., Al-kasassbeh, M. and Alnidami, N. 2019. Phishing detection based on machine learning and feature selection methods. International journal of interactive mobile technology, 13 : 171–183.

[3] Cao, J.C., Qiang, L., Yuede, Ji., Yukun, He. and Dong, Guo. 2014. Detection of Forwarding-Based Malicious

URLs in Online Social Networks. International Journal of Parallel Programming, pp. 1–18

[4] Chou, N., Ledesma, R., Teraguchi, Y., Boneh, D. and Mitchell, J.C. 2004. Client-side defense against web-based identity theft.Proceedings of the 11th Annual Network and Distributed System Security Symposium (NDSS), pp. 76-82.

[5] Dunlop, M., Groat, S. and Shelly, D. May 2010. Goldphish: Using images for content-based phishing analysis. Internet Monitoring and Protection (ICIMP), 5th International Conference on. IEEE, Barcelona, pp. 123–128.

[6] Fette, I., Sadeh, N. and Tomasic, A. May 2007. Learning to detect phishing emails. Proceedings of the International World Wide Web Conference (WWW), pp. 649-656.

[7] Harinahalli, G.L. and BoreGowda, G. 2020 Phishing website detection based on effective machine learning approach. Journal of Cyber Security Technology, pp. 1-14.

[8] Hassan, Y.A. and Abdelfettah, B. 2017. Using case-based reasoning for phishing detection. Procedia Computer Science. 109 : 281–288.

[9] Jain, A. K. and Gupta, B.B. 2018. PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning. Cyber Security. Advances in Intelligent Systems and Computing. 729 : 147-152.

[10] Kumar, J., Santhanavijayan, A., Janet, B., Rajendran, B. and Bindhumadhava, B.S. 2020. Phishing Website Classification and Detection Using Machine Learning. International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, pp. 1–6.

[11] Lee, L.H., Lee, K.C., Juan, Y.C., Chen, H.H, and Tseng, Y.H. 2014. Users Behavioral Prediction for Phishing Detection. Proceedings of the 23rd International Conference on World Wide Web. No. 1, pp. 337–338.

[12] Mao, J., Tian, W., Li, P., Wei, T. and Liang, Z. 2017. Phishing website detection based on effective css features of web pages. 12th International Conference on Wireless Algorithms, Systems, and Applications, pp. 804–815.

[13] Medvet, E., Kirda, E. and Kruegel, C. September 2008. Visual-similarity-based phishing detection. Proceedings of Secure Comm ACM. 129 : pp.224-230

[14] Nourian, A., Ishtiaq, S. and Maheswaran, M. 2009. Castle: A social framework for collaborative anti-phishing databases. ACM Transactions on Internet Technology, pp.

[15] Pan, Y. and Ding, X. 2006. Anomaly based web phishing page detection. Computer Security Applications Conference, pp. 381–392.

[16] Srinivasa, R.R, and Pais, A. R. 2017. Detecting phishing websites using automation of human behavior. Proceedings of the 3rd ACM workshop on cyber-physical system security, pp 33–42.

[17] Tan, C.L., Chiew, K.L, and Wong, K. 2016. Phish WHO: phishing webpage detection via identity keywords extraction and target domain name finder. Decision Support Systems. 88 : 18–27.

[18] Wu, C.Y., Kuo, C.C, and Yang, C.S. 2019. A phishing detection system based on machine learning. International Conference on Intelligent Computing and its Emerging Applications (ICEA), pp 28–32.