To Reduce Error Rate and Improve Performance of Classification Algorithm

S.Nagaparameshwara Chary¹, Dr.B.Rama²

¹,² Department of Computer Science, Kakatiya University, Warangal, India

Abstract: In this paper, an improved method is introduced to reduce the error rate of the standard SD in the context of a two-class classification problem. The learning procedure of the improved method consists of two stages. Initially a shorter learning periodic carried out to identify an important space where all the misclassified samples are located. And the discriminative optimal criterion is computationally intractable as it involves probabilities that are not known a priori. And we have presented an algorithmic framework for feature selection based on nonparametric Bayes error minimization and our proposed framework offers sound interpretations over the existing approaches and also provides principled building blocks for establishing new algorithms. For example when weighted features are used as the search strategy as the framework reveals that the Relief algorithm greedily attempts to minimize Bayes error estimated by classification estimator. The new interpretation of Relief insightfully explains the secret behind the heuristically margin. the lower the error rate of the classifier. Consequently, the proposed improved method by its capability of achieving higher classification accuracy.

Keywords: Classification, KNN, Random forest, Naïve byes, Decision Tree, Feature Selection, Relief

1. Introduction
In fact Data mining is the way toward separating explicit data from information and introducing important and usable data that can be utilized to take care of issues. There are various types of administrations in the process like content mining, web mining, sound and video mining, pictorial data mining and interpersonal organization data mining.

Extraction of data isn't the essential procedure we have to perform; data mining in like way consolidates different strategies, like Cleaning of the Data, Integration of data, Transformation of Data, Data Mining, Evaluation and Presentation of patterns and data respectively. Exactly when these procedures are done, we can implement them in different applications, like Detection of Fraud, Analysis In Market, Production Control etc.

The possibility of Classification Algorithms is truly basic. you are expecting the objective class by breaking down the preparation dataset. This is frequently one among the preeminent, if not the premier basic idea you concentrate after you learn Data Science.
We utilize the preparation dataset to incite better limit conditions which might be wont to decide each target class. When the limit conditions are resolved, the accompanying assignment is to foresee the objective class. The whole procedure is thought as grouping.

In grouping, the idea isn't to anticipate the objective class as in order, it's all the more ever attempting to gather the comparable very things by thinking about the principal fulfilled condition, all the things inside a similar gathering ought to be comparative and no two distinctive gathering things mustn't be comparative.

The accuracy of a classifier is the probability of correctly predicting the class of an unlabelled instance and it can be estimated in several ways (Baldiet al., 2000). Let us assume for simplicity to have a two-class problem: as an example, consider the case of a diagnostic test to discriminate between subjects affected by a disease (patients) and healthy subjects (controls).

2. Literature Review

[Shemim Begum] The “K Nearest Neighbor (KNN) technique is one of the most famous managed AI calculations in bunching and information characterization”. This calculation performs productively in the examinations on an alternate dataset. We plan to concentrate on the order issue. The calculation is experienced over a Leukemia dataset. Right off the bat, three component determination calculations i.e., “Consistency Based Feature Selection (CBFS), Fuzzy Preference-Based Rough Set (FPRS), and KFRS are applied on the dataset later which KNN is applied as a classifier onto the dataset and in doing this test as the outcomes show that the CBFS calculation by and large performs better than the other two KFRS and FPRS calculation separate”.

[MunwarAliv] AI (ML) assumes a significant job in electronic information the executives. Information the board without embracing ML or with ML utilizing metadata is over the top expensive and troublesome. So utilizing ML is important. Numerous ML calculations have been proposed to settle various information the board issues, however they generally need the forecast of secret and non-private information in information which is as yet a difficult exploration hole. A record can't be simply sorted into a solitary classification/class on the grounds that the information in one document may fall into various classifications/classes. Our fundamental goal is to anticipate the private and non-secret information of a record utilizing the K-NN calculation. We have proposed a strategy called “Training dataset Filtration Key Nearest Neighbor (TsF-KNN) classifier which arranges the information dependent on the classification level of the pattern of a (record traits)”. The proposed calculation, is proficient with regards to time and has higher precision contrasted with the customary K-NN calculations.

[Angshuman Paul] Proposed a random forest classifier which has been improved and also performs classification with minimum number of trees. This method removes some features that are unimportant. On the basis of unimportant and important features, he has released a formula which says that the upper limits on the number of trees which are to be added to the forest are important to ensure some improvement in accuracy of classification. Their algorithm is converged but consists of important features. And they also have proven that further addition of trees or any reduction does not show any change or improve the performance. Their efficiency is demonstrated through some experiments conducted on bench mark data sets. They even apply their method on industrial data set to classify different phases. And the result of their method is significant and error is less compared to some other methods.

[Marko Robnik-ˇSikonja] Proposed that “Random forests are one in every of the foremost successful ensemble methods which shows high boosting performance and SVM”. This is fast, robust and has no problem of over fitting and it also explains well about input and o/p features. They have
researched on some of the possibilities like to increase the capability and reduce the correlation between the trees which are individual and voting which is ordinary is replaced with voting which is margin weighted and gave improvement in such a way that they are highly significant in statistical manner over many sets of data. Initially they have proposed how to utilize several different attributes for the need of selecting the splits during the process of tree building. This procedure has decreased the “correlation between the trees and strength is maintained which gave a way for increase in the performance. Better improvement is got by changing voting mechanism that we propose to estimate the common margin of the trees on the instances most a bit like the new instance and then after discarding the trees with negative margin weight the trees votes with the margin”. By evaluating many data sets it is proved that the accuracy is improved significantly. And the results also have indicated that there is an area for improving in the case of using many attributes and many measures for evaluation are to be taken.

[Lior Rokach, Oded Maimon] Decision trees are viewed as the primary most mainstream approaches for illustrating information on classifiers for analyzing distinct trains for example “the insights, AI, design acknowledgment, and Data Mining have managed the issue of growing a decision tree from accessible information where the area refreshes various current strategies for building decision tree classifiers in a top-down way can be presented by the author and similarly the part recommends a bound together algorithmic system for introducing these calculations and depicts different parting rules and pruning procedures”.

[Ulrich Knoll and GholamrezaNakhaeizadeh and Birgit Tausend] The reduction of decision trees regularly depends on the grouping exactness of the decision tree. By using decision trees, we can able to show how the misclassification costs, a related measure applied if mistakes differ in their expenses, can be integrated into a few notable pruning methods. Numerous calculations for the enlistment of decision trees from characterized models dependent on ID3 have been executed in learning instruments. As riotous, inadequate, or fragmented informational collections regularly cause excessively complex decision trees, pruning techniques are applied to acquire the best tree as for standards as the grouping exactness, the intricacy of the tree, or the rules of the strategies assessed in. A related rule, the misclassification costs, applies if blunders shift in their expenses. For instance, allowing a good representative for a temperamental candidate might be more costly for a bank than rejecting it to a decent candidate.

[Wei Zhang] Proposed that, for efficiency and simplicity Naïve Bayes is mostly used in machine learning. Here we are going to analyze how Naïve Bayes is going to perform in the text classification and view results from different points. According to the views we are going to improve the way of classification with the miss classification data that is highly symmetric which is provided. The existing research is done based on the threshold adjustment in the classification which is differed by our new method. Our new method is based on feature selection we choose the tendency of features. The results have shown that if the user chooses the features which are tending to high misclassifications cost we need to improve the adjustment, this is entirely qualitative analysis. In future we want to improve it quantitatively and how to choose the future exactly.

[Jiangtao Ren] Proposed that, it is assumed that data used are precise and exact in traditional ML algorithms, but this assumption is wrong in the case of uncertainty in data like some errors in measurements and repeated values. So here for data uncertainty we use Naïve Bayes with probability distribution function. The main motto of this paper is the conditional probability estimation should be extended to handle pdf. The main problem of Naïve Bayes classifier is probability estimation of conditional class and density estimation of kernel.
3. Classification Algorithms

The initial state of this research is to identify and implement classification algorithms and the goal is to “learn how to assign class labels over the unseen data that is based on models that are built from training data sets as there exist only two class labels the classification is said to be a binary classification due to this it becomes a multiclass classification method as the study will focus on binary classification problems and further compares with different feature selection methods and identifies the impact on three commonly studied classifier algorithms: K Nearest Neighbour, Decision Tree, and Support Vector Machine (SVM)”.

Most of the tasks in performing “statistical pattern recognition for characterization over high dimensional data is deliberately processed and analyzed using statistical tools where a pattern (data sample) is a vector formed generally by many measurements or observations (features) of different physical or other quantities”. There are often tens to several hundreds or even thousands of features composing an individual pattern vector and examples of such data are “measurements arising in character, text, and face recognition from digitized images, spam email identification, diagnostics tasks in medicine and genetic engineering, recognition tasks in biology, economics, astronomy, etc”.

The recognition/classification of a given pattern is characterized by one of the two following tasks. Supervised classification is a problem of establishing decision regions between patterns and assigning an unknown input pattern into one of the predefined classes. In unsupervised classification, classes are learned based on the similarity of patterns.

A K-nearest neighbor
K-Nearest Neighbors (KNN) estimation is such an organized ML check which can be utilized for both depiction also as fall away from the faith farsighted issues. Regardless, it is generally utilized for social affair insightful issues in industry. There are two properties: “Apathetic learning calculation where KNN is a torpid learning estimation since it doesn't have a specific organizing stage and uses all the data for planning while delineation and non-parametric picking up figuring KNN is likewise a non-parametric learning estimation since it doesn't imagine anything about the essential data”.

Algorithm:
“K-Nearest Neighbors (KNN) tally uses incorporate closeness to envision the estimations of new data points which further translates that the new data point will be dispersed a value subject to how restlessly it sifts through the obsessions in the approach set and we can grasp its working with the help of following advances”:

1. For completing any figuring we need a dataset during the crucial improvement of KNN we should stack the readiness in like manner as test data
2. We need to pick the estimation of K for instance the nearest server ranches and K can be any entire number
3. For each point in the test data do the going with
   a. Calculate the division between test data and each segment of getting ready data with the help of any of the system to be unequivocal: “Euclidean, Manhattan or Hamming package and the most routinely used system to enroll package is Euclidean”
   b. Considering the section regard as sort them in climbing demand
   c. It will pick the top K lines from the composed show
   d. It will designate a class to the test point subject to most standard class of these lines
4. End
**Time Complexity**

For the savage power neighbor search of the kNN calculation, we have a period intricacy of $O(n \times m)$, where $n$ is the quantity of preparing models and $m$ is the quantity of measurements in the preparation set. For straightforwardness, expecting $n \gg m$, the intricacy of the animal power closest neighbor search is $O(n)$.

**B Naive Bayes Algorithm:**

Working of Naïve Bayes' Classifier will be comprehended with the help of the underneath model. Assume we have a dataset of climate and comparing objective variable "play" so utilizing this dataset we need to settle on a choice that whether we should consistently play or not on a particular day reliable with the environmental condition. So to determine this issue, we'd prefer to follow the beneath steps:

- Convert the given dataset into recurrence tables.
- Produce probability table by finding the conceivable outcomes of given highlights.
- Presently, use Bayes hypothesis to compute the back likelihood.

In “Bayesian request the essential interest is to find the back probabilities for instance the probability of an imprint given some watched features ($L \mid \text{features}$) and with the help of Bayes speculation we can convey this in quantitative structure” as follows and the followings are several specialists of utilizing Naïve Bayes classifiers:

- Naive Bayes gathering is certainly not difficult to finish and smart.
- It would be join speedier than discriminative models like decided fall away from the faith.
- It takes less preparing information.
- It is altogether flexible in nature, or they scale straightly with the measure of pointers and server farms.
- It can make probabilistic measures and can oversee unremitting comparably as discrete information.
- NaiveBayes gathering estimation can be utilized for twofold comparably as multi-class depiction issues both.

**Time complexity of Naïve Bayes algorithm**

The rum time complexity of naïve bayes algorithm lies in its computational efficiency which is $O(nk)$, Where $k$ is number of label classes and $n$ is number of features

4. **Proposed Method**

**Accuracy Measures in Classification**

Usually “Human and Bayes error are quite close, especially for natural perception problems and there exists little scope for improvement after surpassing human-level performance and thus learning slows down considerably due to which your algorithm is doing worse than humans, following methods can be used to improve performance:

- Get labeled data from humans
- Gain insights from manual error analysis, e.g. understand why a human got this right
- Better analysis of Bias or Variance

NavieBayes may be a fundamental and critical techniques that you just should test and using on your game plan issues.

It is certainly not difficult to know, gives astounding results and races to make a model and make needs. In this manner alone you should take a better examine the figuring.
Better dumbfounded bayes:
6 encounters to incite the head from the reasonable bayes figuring

![Fig Performance evolution of Byes](image)

1. Missing Data: Naive Bayes can supervise separately by the figuring at both model improvement at both model progress time and theory time. In that limit, if an information model unites a missing a motivation for a property, it is ignored when a probability is made due with a depiction regard.

2. Use Log Probabilities: Probabilities are dependably little numbers. To figure joint probabilities, you wish to copy probabilities together, when you increment humble number by another unnoticeable number, you get awfully unassuming number. It is possible to impel into issue with the exactness of your skimming point regards, as under-runs. To keep up a significant nice ways from this issue, meld the log probability space (take the logarithm of your probabilities). This works considering the way that to make an aching in Naive Bayes we’d like to fathom which class has the more noticeable probability (rank) rather than what the specific probability may have been.

Algorithm: Improve Accuracy of NB

Input: The list of ranked attributes \{a_1, a_2, ..., a_N\}
The dataset

Output: A set of selected attributes along with the achievable highest accuracy of NB

Process:
Step 1: Let S be a set of selected attributes for training and testing and initially S is empty a empty set.
Step 2: Set The_highest_accuracy = 0;
Step 3: if K>N goto step 10
Step 4: Add a_k into S
Step 5: Set Acc with the accuracy of NB that is tested on the attribute set S using 10 folds Cross Validation
Step 6: when Acc<=The_highest_accuracy then goto step 8
Step 7: Assign Acc to The_highest_accuracy then goto step 9
Step 8: return S along with The_highest_accuracy
Step 9: goto step 3
Step 10: return S along with The_highest_accuracy

It is clearly assumed that all the attributes and their rankings based on the attribute values attained based on the attribute evaluation method represents its effectiveness towards the algorithm implementation and due to the implementation or execution speed of the above algorithm all the list of attributes will terminate when the highest accuracy value is obtained which is merely depends on the feature combinations that are resulted from the attribute evaluation method that is implemented.
5. Conclusion
In this paper we have proposed a algorithm to implement and to reduce the error rate in implementation of classification problem and on the same hand improvising the classification algorithm using the selection and data transformation process that discriminates optimal criterion which is computationally intractable as it clearly involves probabilities which are initially not known. And we have clearly presented an algorithmic framework for feature selection that is based on nonparametric Bayes error minimization process for implementation of feature selection approaches that strive based on the key balancing factor between the ability to detect complex patterns and attain the flexibility to handle various types of data and yield computational efficiency. We propose the understandability of these advancements will leads to selecting the best key application in this approach which will guide us in creation and development of better feature selection approaches that yield negligible error rate and improve the performance drastically. In future scope we prefer to implement the above algorithm in different data sets with variance in their sizes.

References
[1] Q Chen and Y-PP Chen, “Discovery of Structural and Functional Features in RNA Pseudoknots,” IEEE Trans Knowledge Data Eng, 2009
[2] EF Combarro, E Montañés, J Diáz, J Ranilla, and R Mones, “Introducing a Family of Linear Measures for Feature Selection in Text Categorization,” IEEE Trans Knowledge Data Eng, 2005
[3] I Guyon and A Elisseeff, “An Introduction to Variable and Feature Selection,” J Machine Learning Research, vol 3 pp 1157-1182 2003
[4] S-H Yang and B-G Hu, “Feature Selection by Nonparametric Bayes Error Minimization,” Proc 12th Pacific-Asian Conf Knowl-edge Discovery and Data Mining (PAKDD ’08), pp 417-428 2008
[5] P Praveen, C J Babu and B Rama, "Big data environment for geospatial data analysis," 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2016 pp1-6doi: 10.1109/CESYS2016889816
[6] Ali Shaik, M"A survey on text classification methods through machine learning methods", International Journal of Control and Automation”, 2019 pp: 390-396
[7] B Rama, P Praveen, H Sinha and T Choudhury, "A study on causal rule discovery with PC algorithm," 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Dubai, 2017 pp 616-621 doi: 101109/ICTUS20178286083
[8] G Saon and M Padmanabhan, “Minimum Bayes Error Feature Selection for Continuous Speech Recognition”, Proc Advances in Neural Information Processing Systems 13 (NIPS ’01), pp 800–806, 2001
[9] Shikha, C (2013) Survey paper on improved methods of ID3 decision tree classification International Journal of Scientific and Research Publications 1-4
[10] Ali Shaik, M and Sampath Kumar, T and Praveen, P and Vijayaprakash, R, "Research on multi-agent experiment in clustering," International Journal of Recent Technology and Engineering", 2019 4 pp: 1126-1129
[11] Breiman, L (2001) Random forests Machine Learning 45 (1): 123 – 140
[12] Ali Shaik, M"Protecting Agents from Malicious Hosts using Trusted Platform Modules (TPM), "Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018", 2018 pp: 559-564
[13] Breiman, L, J Friedman, R Olshen, R, and Stone, C (1984) Classification and regression trees
[14] Nguyen TTS (2019) Model-based book recommender systems using Naive Bayes enhanced with optimal feature selection In: Proceedings of the 2019 8th international conference on software and computer applications, Penang, Malaysia ACM, pp 217–222
[15] Frank E, Hall MA, Witten IH (2016) The WEKA workbench Online appendix for “data mining: practical machine learning tools and techniques Morgan Kaufmann, Burlington

[16] Ali Shaik, M”Time series forecasting using vector quantization”, International Journal of Advanced Science and Technology, 2020 pp: 169-175

[17] Nguyen TTS (2019) Model-based book recommender systems using Naive Bayes enhanced with optimal feature selection In: Proceedings of the 2019 8th international conference on software and computer applications, Penang, Malaysia ACM, pp 217–222