IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 24, Issue 1, Ser. I (Jan. –Feb. 2022), PP 12-14
www.iosrjournals.org

Research on Naive Bayes by using Rapid Miner Device

Widiharto[1], M. Arief Soeleman[2]
[1](Magister Teknik Informatika, Universitas Dian Nuswantoro, Indonesia)
[2](Pascasarjana Teknik Informatika, Universitas Dian Nuswantoro, Indonesia)

Abstract:
Data mining is a procedure used to evaluate and find the hidden knowledge of a database. It is applicable in various sectors such as weather forecast, hospitals, business industries and many more. Currently, education data mining is the most useful application used to analyze data related to student performance, course outlines, and faculty performance. This paper describes different classification techniques by using large and small datasets. These two datasets are dataset examples used through repository sites. Several instances depend upon these sites. These data sets are applied on Naive Bayes type to show that it is the first-class classifier from small and big statistics sets. This paper offers the observe and evaluation of numerous methodologies used for prediction. Based on observation, Naïve Bayes is more appropriate for small datasets based at the assessment executed on this paper the use of numerous methodologies pushed through the RapidMiner device whilst equating precision, consider and accuracy.
Key Word*: Naïve Bayes; Rapid Miner.*

Date of Submission: 01-01-2022 Date of Acceptance: 12-01-2022

I. Introduction

The paper's predominant goal is to observe the effect of various type algorithms in the prediction of unidentified attributes labels. The techniques for identifying the algorithms are accuracy, recall and precision. These are useful while training statistics is used in place of checking out statistics, i.e. locating out the price of regarded values and evaluating them to realize the accuracy, recall and precision of the precise algorithm [1].

We examine the effect of the distribution entropy at the type error, displaying that surely deterministic, or low-entropy, dependencies yield suitable overall performance of naive Bayes. We additionally display that, surprisingly, the accuracy of naive Bayes isn't always correlated with the degree of function dependencies measured the class-conditional mutual facts among the function [2].

II. Naïve Bayes Classifier

Naive bayes is considered a subset of Bayesian decision theory. It also responsible to make new assumption by using its unique formulation. Naive bayes is also considered as a simplest probabilistic classifier. Moreover it perform amazingly well in real world application and its all features and assumptions are provisionally independent. Its values and variables solved many optimizing ploblems equations. The end result found that the given model display the improvement of naive Bayes classifier [3]. The Naïve Bayes classification algorithm has been used by various researchers, one of which was carried out by Hamzah in 2012 [4]. The Naïve Bayes algorithm has several advantages, namely fast in calculation, simple algorithm and high accuracy.

III. Factor Considered for Calculating Performance of Naïve Bayes Classifier

Rapid Miner is a statics analysis software design by a company with the exact name that gives a big environment for machine learning, text mining, deep learning and predictive analysis. It is used in business and commercial enterprises. Moreover, it is widely used in research, training, education, rapid prototyping and support or development of all types of machine learning processing, including result visualization, data preparation, validation and optimization [5] accuracy, precision, recall, true positive and false positive are the factors for identifying the activity of naive Bayes Classifier. Let's describe below [6]:

1. Accuracy

Accuracy is identified as the quantity of times expectedly divided through the Total quantity of times. This way, accuracy is the proportion of the appropriately expected training the various general training. In the test, the accuracy values published into desk withinside the foundation of zero to one hundred, instead of zero to one.

Accuracy = ((True Positive + True Negative) / (P + N))*100%

2. Precision

Precision is the exactness of really categorized magnificence, consequently referred to as advantageous predictive analysis. It is the share of times which without a doubt have magnificence x / Total categorized as magnificence x. So essentially, excessive precision said the correct outcomes, and it takes all applicable statistics however returns the handiest topmost outcomes. In short, it's far the number of selected objects which had been associated.

Precision = (True Positive / (True Positive + False Positive))*100%

3. Recall

Recall offers a hassle of sensitivity and its procedure values, product amount, or completeness. It lowers the back of more applicable and a part of the files that might be applicable due to the query. In different words, modules that might be virtually understood as tough to hold from the overall quantity of research. moreover, it's far the quantity of associated items that had been selected.

Recall = (True Positive / (False Negative + True Positive))*100%

4. True Positive (TP)

True advantageous are the tuples that had been efficaciously labelled through the classifier. The share is labelled as magnificence x / real general in magnificence x. True advantageous projected through the modules, which might be expected definitely because of the outcomes specific on end.

True Positive rate = (True Positive / (True Positive + False Negative))*100%

5. False Positive (FP)

False advantageous, percentage incorrectly labelled as x magnificence / Actual general of all training, besides x. It is incorrectly expected as compared to unique outcomes.

False Positive rate = (False Positive / (False Positive + True Negative))*100%

6. F-Measure

F-Measure labeled as (2*Precision*Recall / (Precision+ Recall))*100%. It is a mixed value for precision and recall.

IV. Classification using Naïve Bayes Method

To evaluate research experiments and degree activity with Naïve Bayes classifier, RapidMiner software is extensively used. The key distinction of the Naïve Bayes technique is that it requires text-based labels for numeric labels. The class procedure begins with the practice of education statistics. Fundamental necessities of education statistics for Naïve Bayes consist of 4 factors: label, index, function, and weight. In this case, TF- IDF values is for weight, which might be already identified for examine with LIBSVM. Text-based labels are usual, and multi-label class is supported as well. In this stage, education statistics must be imported into the software. At the same time, they are uploading the document; the software lets in custom function undertaking to the category in imported CSV document. The class version might be skilled based on the assigned function along with label, function, ID and weight. It indicates pattern education statistics for Naïve Bayes class by using RapidMiner. RapidMiner procedure modelling consists of numerous sorts of factors along with parameters operators and repositories. The procedure layout begins with choosing the proper document entered by different operators. Each detail has more than one port in each left facet and proper facet. Usually, ports at the left facet are for entering from previous factors and ports at the proper facet are for output that may enter to subsequent detail. The following steps explained how the procedure layout is completed in Rapid Miner steps which is [7]:

1. Import education statistics into version space design
2. From the list of operators, chosen to normalize the scale education statistics. The use of port Link is set up among document and normalizer
3. Cross-validation technique is chosen for measuring the classifier's overall performance over the statistics of education
4. Cross-validation operator include the sub-processes inclusion. In the cross-validation process, the Naïve Bayes classifier version is introduced as classifier operator
5. Apply Model operator is chosen to use the skilled classifier version for validation
6. To measure overall performance and generate an overall vector performance, the class operator is selected
7. The Cross-validation operator is connected to the foremost output ports
8. Once the procedure is designed with proper factors, then the procedure is carried out, and anticipated output is generated in the shape of an overall vector performance.

V. Conclusion

So as we see that this paper define an experimental technique, which helps studies to discover the concern region of a record based on co-phrases analysis. This study specializes in studying models that is Naïve Bayes. One of the important goals of this paper was to lay out such experimental approaches, which display that with big sufficient schooling data, it's far viable to educate naive Bayes classifiers with a greater label or even large dataset. The work could be improved to large datasets and observe different class strategies to examine overall performance and efficiencies in destiny work. The experimentation design additionally suggests that a practical machine to categorize concern regions based on consumer appearance is probably viable.

References

[1] A. Singh and R. Sathyaraj, "A Comparison Between Classification Algorithms on Different Datasets Methodologies using Rapidminer," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 5, no. 5, pp. 560–563, 2016, doi: 10.17148/IJARCCE.2016.55140.

[2] A. Kori, "Comparative Study of Data Classifiers Using Rapidminer," *Internatio nal J. Eng. Dev. Res.*, vol. 5, no. 2, pp. 2321–9939, 2017.

[3] A. S.-M. Al-Ghamdi and F. Saleem, "in Building and Implementation of Naive Bayes Classification Model: A Domain of Educational Data Mining BUILDING AND IMPLEMENTATION OF NAIVE BAYES CLASSIFICATION MODEL: A DOMAIN OF EDUCATIONAL DATA MINING," *Int. J. Adv. Electron. Comput. Sci.*, no. 6, pp. 2394–2835, 2019, [Online]. Available: http://iraj.

[4] A. Hamzah, "Klasifikasi Teks Dengan Naïve Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita Dan Abstract Akademis," *Pros. Semin. Nas. Apl. Sains Teknol. Periode III*, no. 2011, pp. 269–277, 2012, doi: 1979-911X.

[5] R. Hossain, R. Ibrahim, R. Binti, D. Zain, and A. M. Khaidzir, "Experimental Study of Support Vector Machines and Naïve Bayes Classifier on Automated Subject Area Classification," *J. Inf. Syst. Res. Innov.*, vol. 11, no. December, pp. 7–13, 2017.

[6] D. Xhemali, C. J. Hinde, and R. G. Stone, "Naive Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages," *Int. J. Comput. Sci.*, vol. 4, no. 1, pp. 16–23, 2009, [Online]. Available: http://cogprints.org/6708/.

[7] P. Kaviani and S. Dhotre, "International Journal of Advance Engineering and Research Short Survey on Naive Bayes Algorithm," *Int. J. Adv. Eng. Res. Dev.*, vol. 4, no. 11, pp. 607–611, 2017.

Widiharto, et. al. "Research on Naive Bayes by using Rapid Miner Device." *IOSR Journal of Computer Engineering (IOSR-JCE),* 24(1), 2022, pp. 12-14.