View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by GSSRR.ORG: International Journals: Publishing Research Papers in all Fields

International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

A New Model for Iris Classification Based on Naïve Bayes Grid Parameters Optimization

Saba Abdul-baqi Salman[a], Al-Hakam Ayad Salih[b], Ahmed Hussein Ali[c]*, Mohammad Khamees Khaleel[d], Mostafa Abdulghfoor Mohammed[e]

[a]*Computer Science Department \College of Education\ Al-Iraqia University\Baghdad\Iraq*

[b]*Collage of Art\ Tikrit University\ Salah ad Din Governorate \Iraq*

[c,d]*AL Salam University College \ Computer Science Dep.\Baghdad, Iraq*

[e]*Imam Aadam University College\ Baghdad, Iraq and Ph.D Candidate Faculty of Automatic Control and Computers\University Politehnica of Bucharest 313 Splaiul Independenței, 060042\ România*

[a]*Email: saba.a.salman1@gmail.com*

[b]*Email: alhakam.ayad1987@gmail.com*

[c]*Email: msc.ahmed.h.ali@gmail.com*

[d]*Email: mohammad.cs88@gmail.com*

[e]*Email: alqaisy86@gmail.com*

Abstract

Data mining classification plays an important role in the prediction of outcomes. One of the outstanding classifications methods in data mining is Naive Bayes Classification (NBC). It is capable of envisaging results and mostly effective than other classification methods. Many Naive Bayes classification method provide low performance in classification and regression problems Ones of the facts behinds the performances of the NBC is dues to the assumptions of contingent on independence amidst predictors and the initials hyper parameters. However, this strong assumption leads to loss of accuracy. In this study, a new method for boosting the accuracy of NBC was proposed. The proposed new technique uses a grid search to give better accuracy Naïve Bayes classification.

Keywords: Data mining; Classification; Naïve Bayes Classifier; Grid optimization; Accuracy.

---

* Corresponding author.

1. Introduction

Classification is a major and efficient method for predicting results from a set of data in data mining. The Naïve Bayes Classifier is a famous classification method used for predicting the outcome of datasets [1, 2]. Generally, the NBC operates effectively compared to the other classifiers because of its excellent prediction accuracy, simplicity, reduced memory requirement, and lesser computational complexity. The assumption of self-reliance amidst the predictors is the main rationale for a superior capability of the NBC. However, this assumption and inadequate initialized factor might result in accuracy loss in the NBC [3]. There can be a higher loss of accuracy if the set of considered datasets have properties with strong interaction among themselves. Therefore, enhancing the accuracy of NBC with parameter optimization is a challenging task [4]. In this study, a strong technique for reducing the accuracy loss in the NBC was proposed owing to the poor initialization of the Naïve Bayes factor. The results obtained from the experimental trials indicate that the suggested technique effectively enhanced the validity of the proposed NBC variant compared to the conventional NBC. The suggested algorithm was evaluated on an IRIS dataset acquired from the UCI Machine Learning database.

This paper is organized in sections; section II elaborates the NBC while sections III and IV discussed the dataset and the implementation, respectively. The results of the evaluations are presented in section V while section VI presents the study conclusions.

2. Related Work

Despite of the simplicity of naïve bayes classifier , its implemented in many practical classification and recognition problems. So there are several classification models based on naïve bayes proposed by the research community. The literature survey explain the articles about naïve bayes with different optimization algorithm.M.Borrotti and his colleagues [5] proposed NACO method that combine naïve bayes classifier with ant colony optimization algorithm in order to increase the accuracy of classification for high dimensional data. Shuangshuang Cui and his colleagues [6] used naïve bayes classifier to predict osteonecrosis of the femoral head. Sujana and his colleagues [7] proposed new feature selection method using naïve bayes and cuckoo search optimization algorithm. All the above discussed method have the problem of local optima when optimize hyper parameter of naïve bayes classifier. In this study, a new method for boosting the accuracy of NBC using grid search optimization was proposed.

3. Classifiers

Classification is imperative for data mining. The learning algorithm [8] establishes a classifier in a given set of measurement, for instance, a set of characteristic data (x1,x2,…. ,xn), where xi denotes feature data Xi. Let c stand for the classification feature and $c \in C \subseteq \mathbb{R}m$ be an instance of C. The purpose of classification is to initiate the actuality of groups when given a set of observation (unsupervised learning) or where various categories prevail and the target is classified into one of the previous categories (supervised learning) [9]. Supervised learning has been employed in this study as the classification method.

3.1 Naïve Bayes

Classifier effectively anticipates the class of a data in relation to an individual case in a given dataset. Depending on the Bayes' Theorem, the NBC is a supervised classification used for predicting the class from the characteristics of a dataset.

3.2 Grid search optimization

Several studies have been conducted on the selection of kernel parameters, These studies have proposed methods such as the grid search method and its variant, as well as the bilinear method for kernel parameters selection [3]. The PSO and grid search are the two methods that often give the best values of the kernel parameters (C,γ) as earlier reported. The grid search method takes the combined values of M and N for C and γ (M*N for C,γ) and trains them respectively before estimating the promote recognition rate. Grid search technique has an increased learning precision; however, the superior quantity of the input consumes much time. The grid search is more preferable because it can parallelly participate in the training of every SVM since they do not depend on each other. Hence, it takes much time to execute a complete grid search. Thus, accomplishing the total grid search consumes much time. Hsu and his colleagues [10] suggested an improved grid search method with the basic idea of first getting the optimal values of C and γ using a large step combination of C and γ before conducting a detailed grid search within a range close to the C and γ values. More so, Li and his colleagues [11] suggested bilinear grid search technique with the main goal of obtaining the combination of parameters that can give an improved bilinear grid search before searching for the optimum values in a certain range close to the optimal parameter combination.

4. Dataset

Row No.	id	label	a1	a2	a3	a4
1	id_1	Iris-setosa	5.100	3.500	1.400	0.200
2	id_2	Iris-setosa	4.900	3	1.400	0.200
3	id_3	Iris-setosa	4.700	3.200	1.300	0.200
4	id_4	Iris-setosa	4.600	3.100	1.500	0.200
5	id_5	Iris-setosa	5	3.600	1.400	0.200
6	id_6	Iris-setosa	5.400	3.900	1.700	0.400
7	id_7	Iris-setosa	4.600	3.400	1.400	0.300
8	id_8	Iris-setosa	5	3.400	1.500	0.200
9	id_9	Iris-setosa	4.400	2.900	1.400	0.200
10	id_10	Iris-setosa	4.900	3.100	1.500	0.100

Figure 1: IRIS Dataset

The presented technique in this study used the IRIS dataset acquired from the UCI Machine Learning Repository. The dataset is in a multivariate group as it provides the statistic on the Iris plant type based on four characteristics which include width, width and petal - length, sepal - length, and values as presented in Fig-1. The dataset is composed of three groups with 50 cases each and a total of 50 cases. The type of Iris plant is the

forecasted characteristic in this dataset [5].

5. Implementation

- Step-1: The Iris dataset in CSV is computed as the input.
- Step-2: Divide the data into test and training datasets. In this study, the dataset was divided into 70% training and 30% testing.
- Step-3: Distinguish the training dataset based on the class values, that is, 1, 2 and 3.
- Step-4: Determine the standard deviation and mean values for the individual data case based on the class values.
- Step-5: Choose the naïve Bayes laplace_correction parameter as input to grid search optimization algorithm.
- Step-6: Apply the optimal value of the laplace_correction parameter as an initial value to the process of classification using naïve Bayes.
- Step-7: Utilize the model and generate predictions.
- Step-8: Determine the prediction accuracy through the comparison of the class data of test dataset. This accuracy is evaluated based on the ratio between 0 to 100%.

6. Results and Correlations

```
PerformanceVector:
accuracy: 95.33% +/- 4.27% (mikro: 95.33%)
ConfusionMatrix:
True:   Iris-setosa     Iris-versicolor Iris-virginica
Iris-setosa:    50      0       0
Iris-versicolor:        0       47      4
Iris-virginica: 0       3       46
```

Figure 2: The accuracy of 95.33% without Step-5

```
PerformanceVector:
accuracy: 97.78%
ConfusionMatrix:
True:   Iris-setosa     Iris-versicolor Iris-virginica
Iris-setosa:    11      0       0
Iris-versicolor:        0       21      0
Iris-virginica: 0       1       12
```

Figure 3: The accuracy of 95.33% without Step-5

The suggested model presented in Section IV was performed on the Iris dataset with and without Step-5. In each

run, the obtained results were evaluated based on the accuracy of the NBC. The obtained results showed that the accuracy of the NBC increased to 97.78 using Step-5 and about 95.33% without Step-5. All the results, with the optimization, are presented in Figs 2 and 3, respectively. The inbuilt NBC of Matlab achieved 79.09% accuracy using the same dataset.

Table 1 presents the combined analysis of the suggested technique in relation to the accuracy of other techniques. The results showed that the suggested technique reduced the accuracy loss in the NBC through the assumption of conditional independence. Thus, the presented model in this study can enhance the conduct of the NBC. The result of proposed method show the powerful of using grid search optimization to select the best hyper parameter of naïve bayes classifier. The grid search method divide all the parameters into a certain range, and calculate the parameters of all points on the grid, and choose the best parameters according to the accuracy.

Table 1: Comparative Analysis of Performance

Classifier	**Performance on 150 Data Instances**		
	Training (67%)	**Test (33%)**	**Accuracy %**
NBC with conditional independence and execution step – 5	100	50	97.78
NBC with conditional independence without execution step - 5	100	50	95.33
Matlab's inbuilt NBC Algorithm	100	50	79.09

7. Conclusions and Recommendation

The starting value is the main factor for accuracy loss in the NBC. However, this presumption causes the estimation of probabilities to be easier. The separation technique employed in this study improved the accuracy of the classifier using grid search optimization technique. The obtained results indicate that the employed technique achieved a significant prediction accuracy compared to the conventional inbuilt Matlab NBC. Thus, the accuracy of the NBC can be enhanced with the assumption of conditional independence.

References

[1] K. Netti and Y. Radhika, "A novel method for minimizing loss of accuracy in Naive Bayes classifier," in Computational Intelligence and Computing Research (ICCIC), 2015 IEEE International Conference on, 2015, pp. 1-4.

[2] K. Netti and Y. Radhika, "An efficient Naïve Bayes classifier with negation handling for seismic

hazard prediction," in Intelligent Systems and Control (ISCO), 2016 10th International Conference on, 2016, pp. 1-4.

[3] T. Xiao, D. Ren, S. Lei, J. Zhang, and X. Liu, "Based on grid-search and PSO parameter optimization for Support Vector Machine," in Intelligent Control and Automation (WCICA), 2014 11th World Congress on, 2014, pp. 1529-1533.

[4] L. Moore and C. Kambhampati, "The effect of features using Feature Selection for Bayesian Classifier," in Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, 2013, pp. 4641-4646.

[5] M. Borrotti, G. Minervini, D. De Lucrezia, and I. Poli, "Naïve Bayes ant colony optimization for designing high dimensional experiments," Applied Soft Computing, vol. 49, pp. 259-268, 2016.

[6] S. Cui, L. Zhao, Y. Wang, Q. Dong, J. Ma, Y. Wang, et al., "Using Naive Bayes Classifier to predict osteonecrosis of the femoral head with cannulated screw fixation," Injury, 2018.

[7] T. S. Sujana, N. M. S. Rao, and R. S. Reddy, "An efficient feature selection using parallel cuckoo search and naïve Bayes classifier," in Networks & Advances in Computational Technologies (NetACT), 2017 International Conference on, 2017, pp. 167-172.

[8] M. Duan, K. Li, X. Liao, and K. Li, "A parallel multiclassification algorithm for big data using an extreme learning machine," IEEE transactions on neural networks and learning systems, 2017.

[9] M. Khalilinezhad, B. Minaei, G. Vernazza, and S. Dellepiane, "Prediction of healthy blood with data mining classification by using Decision Tree, Naive Baysian and SVM approaches," in Sixth International Conference on Graphic and Image Processing (ICGIP 2014), 2015, p. 94432G.

[10] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," 2003.

[11] L. Li and X.-L. Zhang, "Optimization of SVM with RBF kernel," Jisuanji Gongcheng yu Yingyong(Computer Engineering and Applications), vol. 42, pp. 190-192, 2006.