Comparative Analysis on the Effect of Feature Selection on Classification Performance

Sunita Beniwal, Neha Kathuria, Ashwani Kumar

Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India

Corresponding author: Sunita Beniwal, Email: sunitabeniwalcse@gmail.com

A huge amount of raw data is present in the information industry and this raw data has to be converted into useful information. Feature selection is a process of selecting the useful and relevant features from the data set. Feature selection is important as it helps in reducing the size of the data and complexity of the model and makes it simpler and easily understandable. Feature selection aims to minimize the cost and improve the performance of the model. This research work is performed using two steps, firstly feature selection is performed using Bacterial Foraging optimization algorithm on two datasets i.e., Iris and Diabetic datasets and performance of selected features is evaluated using Naïve Bayes and KNN. The results of the research are compared with accuracy of original datasets with-out feature selection. The feature selection using BFO yielded better results.

Keywords: Bacterial foraging optimization, dimensionality, feature selection, KNN, naïve bayes.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 577–582. DOI: https://doi.org/10.56155/978-81-955020-7-3-51

1 Introduction

Data mining has become increasingly valuable because of the abundance of data and the necessity to transform that data into knowledge and usable information. Data Mining is the process of solving the problems of evaluating the useful information/data already present in a large amount of database. In the information sector, there is a vast volume of unprocessed data that must be transformed into useable information for it to be helpful at all. Data mining has gained popularity due to the need to find hidden patterns and make data "information rich". Data mining is the process of extracting and identifying hidden patterns in a large volume of data. Data mining is a crucial stage in the decision-making process and expanding our knowledge base [1]. The provided data includes a significant number of characteristics that, if employed as such for developing mining models, might result in overfitting and additional modeling time. Feature selection is a process of selecting the useful and relevant features in the data set. The feature relevancy or usefulness would be based on a predictive model which is trained on the training data. Feature selection is important as it helps in reducing the size of the data and complexity of the model and makes it simpler and easily understandable. The selection of attributes would be determined based on some evaluation measure i.e., information gain, gain ratio, PCA etc [2]. Different techniques can be used for feature selection. This work focusses on the use of nature inspired algorithm, Bacterial foraging optimization for the same.

Nature is a rich source of inspiration. Nature tends to favor the animals with successful or good foraging strategies and eliminate the ones with poor strategies. The activity of foraging by animals is an optimization process. In the foraging process, animals maximize their energy by taking actions per unit time. They tend to make wise judgements and identify the optimal course of action in a dynamic setting. These algorithms were created using inspiration from nature, as their name indicates. They may be utilized to locate the overall optimum solution inside the search space and to re-solve practical optimization challenges [3].

A novel evolutionary computing method known as the Bacterial Foraging Optimization Algorithm (BFOA) was introduced by Passino [4] in 2002. E. coli bacteria's foraging behavior, or ways to find, handle, and consume food, is imitated in BFOA. E. coli bacteria go through four phases during foraging: chemotaxis, swarming, reproduction, and elimination and dispersion. The bacteria can move in two different directions, i.e., swim (unit movement in the same direction) and tumble (unit movement in a different direction). The premise behind the BFO is that animals with unsuccessful foraging tactics will be eliminated, which will help the genes of animals with good foraging methods spread [4]. The Bacterial Foraging Optimization Algorithm offers the benefit of having a lower computing time need, a lighter computational load, the ability to handle more objective functions, and global convergence [4].

2 Literature Review

Fister et al. [3] studied the various nature- inspired algorithms. They classified the existing algorithms into four main categories. These are Swarm Intelligence based, Bio-inspired, Physics and Chemistry based, and others. This classification is not unique because it largely depends on the focus, perspective, and emphasis. The emphasis or focus is about search path, the interaction of multiple agents, updating equations, and source of inspiration. Agarwal and Mehta [5] presented the review of various nature inspired algorithms and various toolboxes available and studied the efficiency of nature inspired algorithms over benchmark test problems in order to solve the “curse of dimensionality” problem.

Tang et al. [6] used the Bacterial Foraging Algorithm for the optimization in dynamic environments and named it DBFA. The capacity of dynamic settings to seek and converge is sought. The current BFO employs an artificial reproduction mechanism to increase convergence speed, but since it lacks variety, it cannot function in dynamic situations. Because of the selection strategy that DBFA uses, the bacteria in DBFA can adapt to the changing environment. When DBFA and BFA were com-pared on a variety of

fronts, DBFA performed satisfactorily. Abraham et al. [7] pro-vided a simple analysis of one step used in BFOA, i.e., reproduction. In a simple two-bacterial system operating on a one-dimensional fitness landscape, the study is focused on reproduction. The research demonstrates that the reproduction event's contribution causes bacteria to quickly converge on a nearly optimal solution.

Yan et al. [8] presented the improved BFO to overcome the shortcoming of classical BFO. The proposed algorithm is a lifecycle model in which bacteria might dynamically divide, die, and move throughout the foraging process. It offers improved performance over classical BFO and shows competitive performances compared with other algorithms on higher-dimensional problems. Jun Li et al. [9] analyzed the BFO on its various operations like elimination and dispersal to avoid the escape in local minima and chemotaxis to adjust the step length. An improved BFO was created to raise the algorithm's precision and effectiveness. According to the findings, the enhanced BFO algorithm outperforms the standard BFO method in terms of accuracy and convergence speed.

3 Results and Discussion

3.1 Methodology

The proposed research was carried out using the steps explained below:

1. Selection of dataset: Two datasets used are Iris dataset [10] and Diabetic dataset[11]. The datasets are taken from UCI data repository. Data of only two classes is used from iris dataset. The classes are Iris Versicolour and Iris Verginica.
2. Feature Selection is done using Bacterial Foraging Optimization based on the fit-ness function used i.e., Information gain. The two datasets are taken and BFO algorithm is applied on them for selecting the relevant features. If we vary the two parameters of BFO i.e., Swim length and Chemo-tactic value, the accuracy of the system may vary. Here we have taken 20 as Chemo-tactic value of bacteria and vary the Swim length as 5, 10, and 15. Best results are obtained for swim length of 15.
3. Evaluation of selected features: The relevance of selected features is evaluated using classification algorithm, Naïve Bayes and KNN. Comparison is also done with both the classifiers without using feature selection

3.2 Results

For performance analysis parameters like accuracy, precision, recall, and error rate are used.

Table 1 shows the confusion matrix after feature selection. Performance of Naïve Bayes is better than KNN. Table 2 depicts the confusion matrix of performance of both algorithms on Diabetic dataset.

Table 1. Confusion Matrix for Iris dataset

	NaïveBayes		KNN	
Classo	50	o	50	50
Class 1	o	50	o	o

Table 2. Confusion Matrix fordiabetes dataset

	NaïveBayes		KNN	
Class 0	118	3	118	32
Class1	0	29	0	0

The accuracy, precision, recall, and error rate of the model is calculated based on the confusion matrix. Table 3 given below shows the values for accuracy for both the datasets and the classifiers. Naïve bayes has better accuracy for both the datasets. The recall values are given in table 4 for both the datasets and the classifiers. Both the classifiers have recall values of 1 for both datasets. The precision is given in table 5 for both the datasets and the classifiers. Naïve bayes has better precision for both datasets. The error rate matrix is given here for both the datasets and the classifiers in table 6. The results show that the Naïve Bayes classifier gives better performance in case of Iris dataset as well as Diabetic dataset.

Table 3. Accuracy matrix

	NaïveBayes	KNN
Iris	1	0.5
Diabetic	0.975	0.7867

Table 4. Recall matrix

	NaïveBayes	KNN
Iris	1	1
Diabetic	1	1

Table 5. Precision matrix

	NaïveBayes	KNN
Iris	1	0.5
Diabetic	0.98	0.7867

Table 6. Error Rate matrix

	NaïveBayes	KNN
Iris	0	0.5
Diabetic	0.02	0.213

3.3 Comparative Analysis

The accuracy of two datasets using feature selection is compared with accuracy of same without feature selection i.e. whole dataset. Table 7 gives accuracy of both classifiers without feature selection.

Table 7. Accuracy without feature selection

	NaïveBayes	KNN
Iris	95.53	94.67
Diabetic	75.75	70.19

Comparison is made between the accuracies of both the classifiers on individual datasets with and without feature selection. The accuracies of these classifiers have been obtained using BFO for feature selection and without using feature selection. Figure 1 and 2 gives a graphical representation of comparison on iris and diabetic datasets respectively.

Figure 1 shows that better accuracy is attained using BFO selected features on Naïve Bayes classifier on Iris dataset. But it is opposite in case of KNN. It shows better re-sults without using feature selection. Comparison of performance of both classifiers using feature selection and without using feature selection on Diabetic dataset is shown in figure 2. The accuracies are com-pared on Diabetic dataset using BFO and without using BFO on both the classifiers. The above graph shows better accuracy using features selected using BFO on both Naïve Bayes and KNN classifier. When we compare the accuracies using BFO and without using BFO individually on different datasets, the comparison result goes in the favor of BFO. The BFO gives better accuracy on Iris dataset in case of Naïve Bayes classification algorithm i.e., it gives 100% accuracy but it does not perform the same for KNN. On diabetic dataset better accuracy is obtained on features selected using BFO for both KNN and Naïve Bayes. So, the feature selection using BFO in-creases the accuracy on the diabetes dataset.

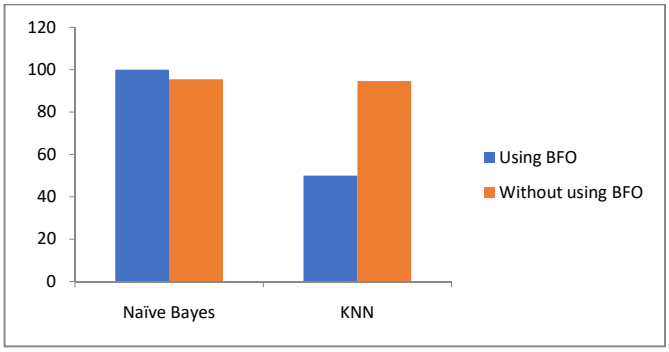

Figure 1. Comparison of accuracy on Iris dataset

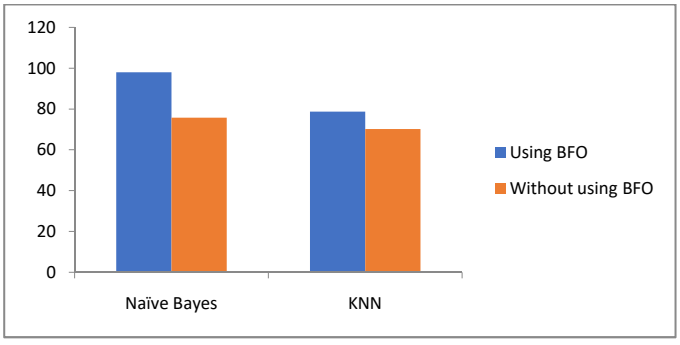

Figure 2. Comparison of accuracy on Diabetic dataset

4 Conclusion

In this research work, the Bacterial Foraging Optimization based features selection on the dataset has been studied. The result demonstrates that Naïve Bayes shows better accuracy on the datasets than KNN algorithm. The result after comparison shows better performance for BFO selected features except for Iris dataset using KNN. Ac-curacy of KNN decreased on selected features. It is observed that the Naïve Bayes demonstrate better accuracy than KNN on both the datasets.

References

[1] Han, J., Kamber, M. (2006) Data Mining: Concepts and techniques. Morgan Kaufmann,2nd edition.
[2] Rani, S., Kumar, D., Beniwal, S. (2018) Improving Medical Diagnosis using filter and Wrapper Techniques. International Journal of Innovative Science and Research Technology, 3(9), 123-129.
[3] Fister Jr, I., Yang, X. S., Fister, I., Brest, J., Fister, D. (2013) A brief review of nature-inspired algorithms for optimization. arXiv preprint arXiv:1307-.4186.
[4] Passino, K. M. (2002) Biomimicry of bacterial foraging for distributed optimization and control. IEEE control systems magazine 22(3), 52-67.
[5] Agarwal, P., Mehta, S. (2014) Nature-inspired algorithms: state-of-art, problems and prospects. International Journal of Computer Applications 100(14), 14-21.
[6] Tang, W. J., Wu, Q. H., Saunders, J. R. (2006) Bacterial foraging algorithm for dynamic environments. 2006 IEEE International conference on evolutionary computation.1324-1330.
[7] Abraham, A., Biswas, A., Dasgupta, S., Das, S. (20008) Analysis of reproduction operator in bacterial foraging optimization algorithm. 2008 IEEE Congress on Evolutionary Computation,1476-1483.
[8] Yan, X., Zhu, Y., Zhang, H., Chen, H., Niu, B. (2012) An adaptive bacterial foraging optimization algorithm with lifecycle and social learning. Discrete Dynamics in Nature and Society.
[9] Li, J., Dang, J., Bu, F., Wang, J. (2014) Analysis and improvement of the bacterial foraging optimization algorithm. Journal of Computing Science and Engineering 8(1) 1-10.
[10] Fisher, R. A. Iris. (1988) UCI Machine Learning Repository. https://doi.org/10.24432/C56C76.
[11] Kahn, M. Diabetes. UCI Machine Learning Repository. https://doi.org/10.24432/C5T59G.