Indian Journal of Science and Technology, *Vol 10(20), DOI: 10.17485/ijst/2017/v10i20/101599, May 2017*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Study on Impact of Dimensionality Reduction on Naïve Bayes Classifier

Priya Mohan[1]* and Ilango Paramasivam[2]

[1]Department of Computer Science, Bharathiar University, Coimbatore – 641046, Tamilnadu, India;
priya.vinoth13@gmail.com
[2]School of Computing Science & Engineering, VIT University, Vellore – 632014, Tamilnadu, India; pilango@vit.ac.in

Abstract

Objectives: The time complexity of the machine learning algorithm is directly proportionate to the dimension of the dataset. In this paper, he impacts of dimensionality of the dataset on the machine learning algorithm, Naïve-Bayes Classifier is evaluated with all feature subsets to analyze whether there is any variations in the performance. **Methods/Statistical Analysis**: Naïve Bayes Classifier is taken for the study to evaluate its variations in terms of its performance in correctly classified instances and incorrectly classified instances. Pima Indian Type II diabetes dataset is taken for the experimental study. Confusion matrix will be formulated for the performance of Naïve-Bayes Classifier using 10-fold cross validation for each run. The study exhibits the impact of the dimensionality on the performance of Naïve-Bayes Classifier. **Findings:** The Naïve Bayes classifier classifies the patient records either as diabetes or as non-diabetes using the values of the feature set. It is a probabilistic approach of classifying the patient records into the binary class. It is found that there is an impact on the performance of Naïve Bayes Classifier due to the dimensionality of the feature set it terms of Classification accuracy, number of true positives, true negatives, false positives and false negatives. The incorrect classification is certainly dangerous. Whereas the valid classification facilitates the healthcare systems in terms of planning effective course of treatment which will save the life of the patient. The invalid classification will lead to a wrong diagnosis while formulating the treatment plan and it will lead to loss of life. Hence, the invalid classification in terms of false negative rate is to be viewed very seriously. In this paper, the study shows that there is an impact on the performance of Naïve Bayes Classifier due to the higher dimensionality of the dataset. **Application/Improvements:** They will be used in medical Informatics for the quality diagnosis and effective treatment planning. The focus on the false positive rate in the classification accuracy of Naïve Bayes Classifier will notably help the healthcare systems to diagnose the patients accurately to save life.

Keywords: Classification Accuracy, Dimensionality Reduction, Machine Learning, Naïve-Bayes Classifier

1. Introduction

Data mining is basically the process of Knowledge Discovery in Databases (KDD) that refers to the overall acquisition of knowledge and patters from the dataset. KDD follows the estimation and possible evaluation of the patterns for the decision making activities that is considered as knowledge. It comprises of the sub processes of selection, preprocessing, transformation, data mining, interpretation or evaluation, which finally yield knowledge. Identifying patterns and Extraction knowledge has been challenge on the large volume of data collected over the period of time. In the field of knowledge extraction, researchers have worked for more than a decade, producing important results; still, this remains an open issue[1]. Data mining deals with discovery of patterns in large datasets. Classification and clustering, both are the two major functionalities of data mining employed to achieve this.

Machine learning explores the construction and study of algorithms that can learn from and make predictions on data. This is further divided into supervised learning, unsupervised learning and reinforcement learning, depending on how the learning process takes place. Some of the most commonly applied machine learning algorithms are Artificial Neural Networks (ANN), Support

**Author for correspondence*

Vector Machines (SVM),and Bayesian Networks etc[2]. Machine learning has found its use in several applications spanning various domains. Some important areas where it is being utilized are- search engines, medical diagnosis, Internet fraud detection, bioinformatics, human handwriting and speech recognition, fraud detection of banks credit and debit cards, etc. It has gained immense significance in recent times owing to the fact that machine learning is used to train computers to perform actions which cannot be programmed in advance[3]. In the machine learning context, a dataset is a collection of instances, characterized by features, where each instance has a different value for a particular feature. Depending on what the dataset is used for, the features have different purposes.

In machine learning as well as data analytics, the features coupled with a good classification algorithm are predictive in nature. In machine learning, high dimensionality of datasets poses a problem to the overall coherence and effectiveness of the machine learning algorithms[4]. One of the reasons is that, when these algorithms are applied to traditional datasets, the classifiers used assume that the features are independent. But in a high dimensionality dataset, it is highly likely that this assumption will be ignored, as high correlation is often observed between the features. Another reason is that any increase in complexity of the pattern to be learned will make the training more time consuming and will negatively affect the accuracy.

In this paper, it is proposed to experiment the Naïve Bayes Classifier with the all feature subsets generated on the machine learning dataset and the performance is evaluated to study how the dimensionality affects the performance of the classifier. Section II briefs the related work presently available on machine learning, data mining, and their applications and pitfalls in terms of performance, methods available to solve the curse of high dimensionality. Section III expresses the considered problem of experimenting with the machine learning algorithm with the different feature subsets. Section IV is about the experimental set up for the study. In the Section V, we have discussed about the performance of the machine learning algorithm. Section VI concludes with the outcomes.

Knowledge Discovery in Databases is the step by step process of identifying useful information or knowledge from a large volume of data collected over the period[1]. Data Mining is one of the phases in KDD process[5]. Data mining utilizes the features of Artificial Intelligence, Machine Learning, and Statistics to formulate the algorithms to achieve the important knowledge. Machine Learning and Data Mining are widely used in various real life applications like business, science and medicine to new areas like sports and e-commerce[2]. Machine learning offers various methods for efficient construction of descriptive models from data. At its best, the modern modeling technique should offer both: accuracy of the developed model and transparency so that the decision maker knows why and how the model derives the decision. Learning models are widely implemented for prediction of system behavior and forecasting future trends. Machine Learning and data mining, both are used in fraud detection, credit card default forecasting, customer purchase patterns, bioinformatics, stock market analysis, molecular biology etc[4]. Data mining suffers with challenges like difficulty in accessing data, preprocessing of noisy data, and inconsistency in data type, dimensions and quality, reliance on personnel trained to handle data mining tools[5].

Machine learning, which is incorporated in the data mining algorithms, has issues of its own like overfitting and high dimensionality[6]. Machine learning methods have been applied to a variety of domains to improve decision-making. One major problem of this area is dealing with datasets with a high number of features[6,7]. Researchers have found out that data preprocessing is essential in order to use data mining tools on the database[4]. High dimensionality affects the analysis and organization of data, a problem that is not faced when the data has lesser number of dimensions. The curse of dimensionality concerns various domains like numerical analysis, combinatory, data mining, and machine learning databases[7]. In data mining, the machine learning algorithms used assume independence of features in the datasets. But if there are too many features, chances are that there will be correlation between some of them[8]. Hence, we need feature selection or feature extraction, by means of which variables will be discarded, while satisfying the better classification accuracy. Feature extraction facilitates to get new features from the set of the original features, whereas feature selection provides a subset of the features[2]. When the underlying important features are known and irrelevant/redundant features are removed, learning problems can be greatly simplified resulting in improved generalization capabilities[7].

Dimensionality reduction reduces the dimension of datasets and enhances the computational efficiency of classification algorithms[4]. The processing time reduces with the decreasing number of attributes[2]. The size of the produced decision tree also reduces with the reduction of attributes, in the decision tree classifier[7]. The so-called "curse of dimensionality" pertinent to many learning algorithms, denotes the drastic raise of computational complexity and classification error with data having high number of dimensions[7].The process of Feature selection attenuates the unimportant variables, removes data redundancy or noisy data, and produces the immediate effects for applications: motivating data mining algorithms to improve the mining such as accuracy prediction and comprehensible results[8]. Most data mining methods depend on features set that define the behavior of the learning algorithm and directly or indirectly influence the complexity of the resulting models. Research[2] shows that the inclusion of useless, or irrelevant, features cause the predictive performance to decline. Reducing the data dimensionality reduces the size of the hypothesis space and allows algorithms to operate faster and more effectively[2].

In this paper, it is proposed to assess the impact of high dimensionality of the dataset and the dimensionality reduction techniques on the performance of the machine learning algorithm, Naïve Bayes Classifier using wrapper models

2. Study of Impact on Naïve Bayes Classifier

The Naïve Bayes Classifier has been applied to study the impact of dimensionality reduction. It follows a basic probabilistic formula to classify models and then produces the important class/labels in the reference to feature selection[2,4]. The reduced dimensionality will be as the smaller feature subsets such as $f = \{\{f_1\}, \{f_1, f_2\}, \ldots\ldots\ldots, \{f_1, f_2, \ldots\ldots, f_n\}\}$ and the Naïve Bayes Classifier will be evaluated for every reduced feature subset.

The Naïve Bayes Theorem goes as follows:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)}$$

Where $p(C_k|x)$ is the posterior
$p(C_k)$ is the prior
$p(x|C_k)$ is the likelihood
$p(x)$ is the evidence.

The Naïve Bayes Classifier will exhibit the stable performance if there is a strong independence assumption which states that there is no relation between any of the features used. The Classifier requires a small amount of training data to estimate the parameters necessary for classification.

3. Experimental Set Up

The Classification algorithms can be applied to compare the effects of applying dimensionality reduction and also without applying dimensionality reduction[9]. The probabilistic learner, Naive Bayes classifier is used to assess the effects of high dimensionality of the dataset on its performance. Wrapper model is taken for the study to use the Naïve Bayes classifier as an evaluation algorithm. Pima Indian Type II Diabetes dataset with 8 features and 768 instances, one class label[2].

4. Performance Evaluation

The Naïve Bayes Classifier produces every feature subset derived from the attribute set of the data. The significance is calculated with the parameters such as False Negatives and False Positives, True Negatives and True Positives, and the classification accuracy. The calculated performance of the accuracy classifier for all the features is 76.3021%, Values of True Positive, False Positive, False Negative and True Negative are 422, 78, 104, 164 respectively. If the Naïve Bayes classifier produces same results with the less number of features, then it is acceptable. If the performance is enhanced, then it is appreciable. The dimensionality reduction is applied in the study to analyze the change in the performance in any form. The features subsets were generated as $f = \{\{f_1\}, \{f_1, f_2\}, \ldots\ldots\ldots, \{f_1, f_2, \ldots\ldots, f_n\}\}$ over the dataset taken for the study. Hence, the Naïve Bayes classifier generates totally eight feature subsets and the result of is observed for each feature subset f_i. The total results of Naïve Bayes classifier for all the feature subsets are shown in Table 1.

It is observed from Table that there is a variation in performance of Naïve Bayes Classifier for the different feature subsets. The Naïve Bayes classifier shows better performance for some of the feature subsets namely $f_3 = \{f_1,f_2,f_3,f_4,f_5,f_6\}$ and $f_4 =\{f_1,f_2,f_3,f_4,f_5\}$. However, the performance of Naïve Bayes classifier deteriorates for the other 7 feature subsets namely f_n, f_1, f_2, f_5, f_6, f_7, and f_8. The impact

Table 1. Performance of Naïve Bayes Classifier for the various feature subsets

S. No.	Feature subsets	Classification accuracy (%)	True Positive	False Positive	False Negative	True Negative
1	All features	76.3021	422	78	104	164
2	$f_1,f_2,f_3,f_4,f_5,f_6, f_7$	75.78	418	82	110	158
3	f_1,f_2,f_3,f_4,f_5,f_6	75.99	420	81	105	162
4	f_1,f_2,f_3,f_4,f_5	77.01	427	73	99	169
5	f_1,f_2,f_3,f_4	78.08	427	76	96	169
6	f_1,f_2,f_3	76.9	424	76	100	168
7	f_1,f_2	74.21	418	82	108	160
8	f_1	70.13	402	93	119	154

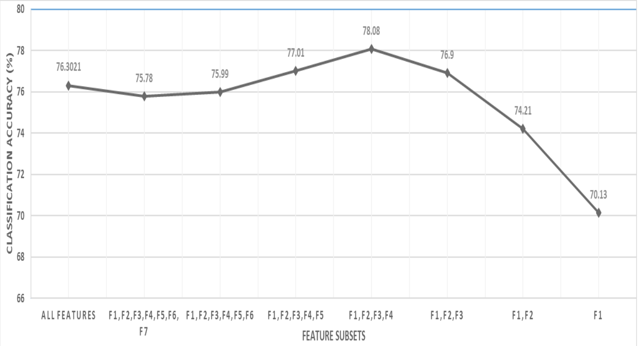

Figure 1. Impact of dimensionality of the dataset on Naïve – Bayes Classifier.

on the performance of Naïve Bayes classifier is shown in Figure 1.

5. Conclusion

The impact of the dimensionality reduction on the performance of Naïve Bayes Classifier is analyzed in terms of classification accuracy, Number of False Negatives and False Positives, True Negatives and True Positives. The study conducts experiments using Pima Indian Type II Diabetes dataset. The Naïve Bayes classifier classifies the patient records either as diabetes or as non-diabetes by reading the values of the feature subset. The incorrect classification is certainly dangerous. But, the correct classification will help to alert the patient at an early stage, so that the life can be saved. The study shows that there is an impact on the performance for every feature subset. Some feature subsets negatively impact whereas few feature subsets positively impact the performance of Naïve Bayes classifier. The study concludes that the performance of Naïve Bayes classifier can be enhanced by finding the optimum feature subset which will be very much helpful in predicting any unknown object into an appropriate class.

6. References

1. Hirota K, Pedrycz W. Fuzzy computing for data mining. Proceedings of the IEEE. 1999; 87(9):1575–600. Crossref
2. Sarojini B. An Integrated Approach of Feature Selection and Parameter optimisation of kernel to enhance the performance of Support Vector Machine. International Journal of Communication Networks and Distributed Systems. 2015; 15(2-3):265–78. Crossref
3. Freeman C, Kuli D, Basir O. An evaluation of classifier-specific filter measure performance for feature selection. Pattern Recognition. 2015; 48(5):1812–26. Crossref
4. Balakrishnan S, Narayanasamy R, Savarimuthu N. Enhancing the performance of LibSVM Classifier by Kernel F-score Feature Selection. 2009; 40:533–43.
5. Kumar A, Tyagi AK, Tyagi SK. Data Mining Various Issues and Challenges for Future - A Short discussion on Data Mining issues for future work. International Journal of Emerging Technology and Advanced Engineering. 2014; 4(1):1–8.
6. Balakrishnan S, Narayanasamy R, Savarimuthu N, Samikkannu R. SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases. Proc. IEEE International Conference on Systems Man and Cybernetics SMC, 2008. p. 2628–33. Crossref
7. Balakrishnan S, Narayanaswamy R. Performance of LibSVM Classifier by Kernel F-score Feature Selection. Springer-Verlag Berlin Heidelberg. 2009; 40:533–43.
8. Sánchez L, Suárez MR, Villar JR, Couso I. Mutual information-based feature selection and partition design in fuzzy rule-based classifiers from vague data. 2008; 49(3):607–22.
9. Imani MB, Pourhabibi T, Keyvanpour MR, Azmi R. A New Feature Selection Method Based on Ant Colony and Genetic Algorithm on Persian Font Recognition. International Journal of Machine Learning and Computing. 2012; 2(3):1–5. Crossref