COMP3308 Assignment 2

Investigation of Effects Correlation-based Feature Selection Method on A Small Dataset Using K-Nearest Neighbour and Naive Bayes Algorithm

SID 440 590 457

July 15, 2019

1 Introduction

1.1 Aim

Using K-Nearest Neighbour and Naive Bayes algorithm on a small data set and stratified cross validation method, the effects of on a real data set has been evaluated. This report aims to look at these effects and also the effects of correlation based feature selection to evaluate the performance of the algorithm and compare the results with the results obtain from Weka. One of the most important aim was to figure out if removing the least important attributes i.e applying CFS on the data set has any effect on the accuracy. It turned out, it doesn't. CFS didn't impact the accuracy of the classifiers adversely.

1.2 Importance

It is of huge importance in machine learning to be able to compare different classifiers with the ability to implement K-Nearest Neighbour and Naive Bayes algorithm. It can have significant effect on decision making and it is important to know and understand whether and how classifiers are successful, as artificial intelligence is very much dependant on accurate decision making ability. Machine learning techniques have been giving freedom to experimenters in order to analyze data in large quantities. Feature selection is fundamental to machine learning. One thing about feature selection is the most common "more are the features used more will be the information or better classification power". But this is not always the case, just because the features do not mean it has more information or better classification performance. Some features can be irrelevant which create noise and affect the lining out-performance. Besides, redundant features also cause degradation in the information. Such things cause degradation is called the curse of dimensionality. Too many features and dimensions lead to degradation and more computation time. The curse of dimensionality can be overcome by feature reduction. The feature reduction can be taken place in terms of feature selection and feature extraction. Here, feature selection is used for the removal of unwanted or irrelevant data. In order to optimize, the accuracy of the classification needs to be improved or maintained. Two ways are possible to study in this case,

- With increasing features the classification accuracy initially increases and then declines.
- This could also be the case when classification accuracy follows the uptrend and then remains constant or remains the same against features.

Feature selection was applied due to all these technicalities.

2 Data

The data set used in this assignment, the Pima Indian Diabetes dataset was originally obtained from UCI Machine Learning Repository but subsequently modified for using in this assignment. The modified data has 768 instances and 8 numeric attributes with two classes as described below,

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure ($mmHg$)
4. Triceps skin fold thickness

5. 2-Hour serum insulin (muU/ml)
6. Body mass index (kg/m^2)
7. Diabetes pedigree function
8. Age
9. Class variable ("yes" or "no")

Weka was used to identify the most important attributes of the dataset, performing Correlation based Feature Selection (CFS) method and to get rid of the unimportant ones. CFS can identify the attributes which have very little impact using generated heuristic which can be ignored. The most important attributes were 2, 5, 6, 7, 8 and 9 from the above list.

3 Classification Algorithms

3.1 K-Nearest Neighbour (kNN)

The kNN is an example of lazy learners and is actually the non-parametric method used for the classification. The K-Nearest neighbor algorithm uses straight line distance (Euclidean distance) between two data points. Keeping in view, the data is continuous and not a normalized one that is why this is considered to be the best choice to use the Euclidean distance method for a kNN algorithm.

3.2 Naïve Bayes (NB)

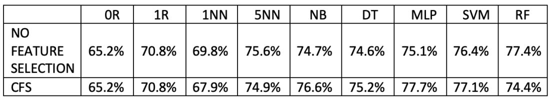
Naïve Bayes is based on Bayes theorem which assumes that the presence of a feature in a class is not related to that of any other. It is a statistical type classifier with an assumption and probability to determine a given data point with attributes. For all the attributes, a normal distribution was assumed following the Bayes theorem.

Examples were divided into training and testing sets. The training set was used to train the learner while the testing set was used to test/evaluate the learner. If the test set is small then there is variance. Coincidentally, if a small set is taken then the accuracy may be very low or high on that set. In the case of a larger test set, variance is small.

Results and Discussion

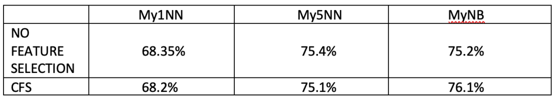
When validation fails to use all the available data then cross-validation is used. In K-fold cross-validation, the entire available data was taken and split into many k-subsets. In this study, 10-fold cross validation was used which is actually the splitting of data into 10-subsets. The accuracy of the classifiers in each subset showed different percent values. An average percent was calculated using the percent values of all the 10-subsets and finally, this average value showed accuracy. With all this, limited data were used so that the training and testing can be done.

The table shows that the NB value resulted slightly better than the kNN algorithm. The resulted algorithms (kNN and NB) are almost the same or identical. The performance of all the algorithms is almost the same except for 1NN as well as zeroR, with accuracy ranging from 70% to 75% without any feature selection. Further results showed an increase of 0.3% mean-accuracy of my algorithm. It can be seen on the other side, an increased value with 0.7% mean accuracy of the Weka's algorithm. Fold thickness of skin triceps, number of times pregnant and DBP are the attributes removed by Weka using CFS. These attributes make sense but are considered to be least likely related to diabetes. It is of great importance and advantageous to analyze and classify data containing fewer attributes. It is super simple to classify data points by using less time and computational power. However, the dataset used in this case is relatively small. Therefore, different results can be obtained from much complex dataset containing numerous attributes.



	0R	1R	1NN	5NN	NB	DT	MLP	SVM	RF
NO FEATURE SELECTION	65.2%	70.8%	69.8%	75.6%	74.7%	74.6%	75.1%	76.4%	77.4%
CFS	65.2%	70.8%	67.9%	74.9%	76.6%	75.2%	77.7%	77.1%	74.4%

Table 1: Accuracy of Weka's algorithm



	My1NN	My5NN	MyNB
NO FEATURE SELECTION	68.35%	75.4%	75.2%
CFS	68.2%	75.1%	76.1%

Table 2: Accuracy of my algorithm

4 Conclusion and Future Work

In conclusion, the data used was real and many facts could be analyzed about the success of different classifiers. Almost all the classifiers were highly accurate to the Weka responses in both the situations (after and before applying CFS on the dataset). There was a non-zero net increase recorded in the accuracy of the classifiers of Weka and my classifiers by using CFS. After applying CFS, the highest increase in the accuracy was for the multilayer perceptron classifier. Though the increase is small, the effect on this classifier as compared to other classifiers is greater. Further work on the effect of CFS on a large dataset could be interesting. Particularly, the effect of CFS on runtime and the complexity of classifiers.

5 Reflection

One can be enriched with the understanding of correlation-based feature selection (CFS) and its importance in many circumstances. The way CFS screen and detect the noisy, irrelevant or unnecessary features and end up with the removal of these redundant and useless data enabled the process to be short and timeless. The important takeaway of this assignment showed the incredible opportunity of AI algorithms. The algorithm's decision-making success is mostly affected by the classifiers, details in small implementation, attribute collection and features selection. Based on these factors, the life-changing and decision-making tools can be discovered. The study of feature selection is of great importance in mining, bioinformatics, from sequence through micro-array and spectral analysis. All these uses have given rise to feature analysis involving techniques and methods. In early ages, it was difficult to get hands-on experiments and analyzing data but the current age is informative. Accumulating data is inexpensive and easy. However, there is more work to do, like exploring AI furthermore in most important areas such as the field of medicine.