COMP3308_ASM2_Report

04.2022

Group Members

Yuan Feng (SID:500354182)

Yunshuo Zhang (SID:500025673)

Table of Content

1. Introduction

1.1 Aim

The aim of this study is to evaluate the performance of other classifiers on the same data set by comparing the accuracy. The aim is to write a k-nearest neighbor algorithm and a naive Bayesian algorithm, and implement them on a real data set called Pima Indian diabetes, which has 8 numeric attributes and 1 qualitative attribute. Also,we will analyze and evaluate the performance of classifiers such as zeroR, 1R, Decision tree, SVM and random forest. Furthermore, stratified cross validation methods will be used to give a better evaluation. We also compare the accuracy of our algorithm with the accuracy of the same algorithm in Weka. In this study, we also studied the impact of correlation-based feature selection on the accuracy of and classifiers.

1.2 Importance

This study is very important because if we choose the most suitable classifier, we can train data better. This is a great progress in artificial intelligence. Classifiers play an important role in machine learning and AI due to it can contribute to classify specific data points with categorical class labels by being a discrete valued function. We can find the most suitable method through different classifiers, which is also an important way for artificial intelligence to classify new data or new knowledge. By comparing the accuracy of each classifier to find the most suitable one to use, it can not only accelerate the development of artificial intelligence, but also improve the efficiency of machine learning.

2. Data resource

2.1 Data introduction & preprocessing

The owner of the database we use, *Pima Indians Diabetes Database*, is *National Institute of Diabetes and Digestive and Kidney Diseases*. There are 768 instances, each instance has 8 numerical attributes which are *Number of times pregnant, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function and Age (years)*. Each instance also has one class(yes/no) it belongs to. Class value "yes" is interpreted as tested positive for diabetes and class value "no" is interpreted as

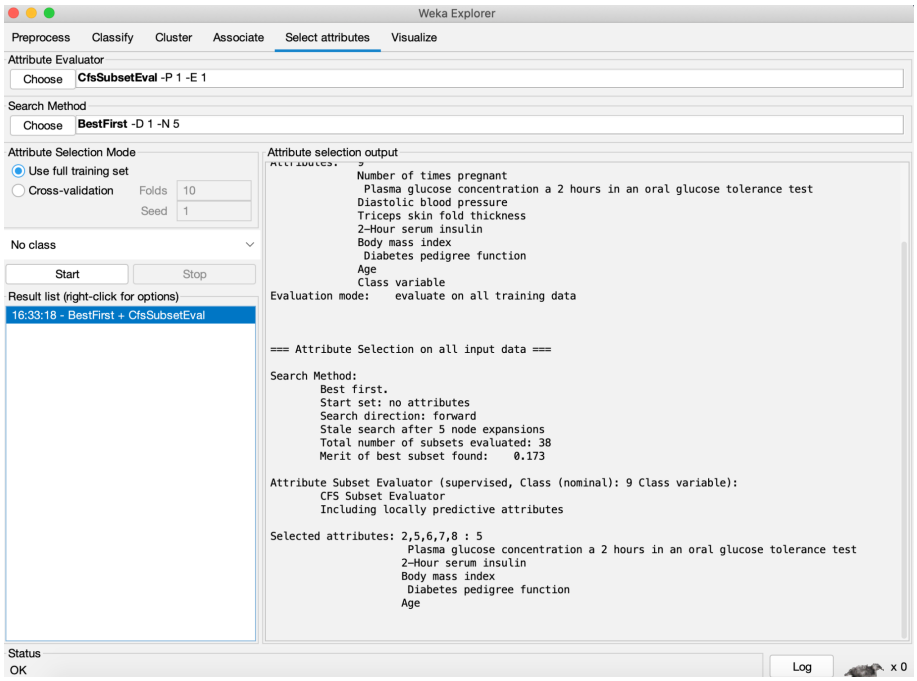
tested negative for diabete. There are 500 instances in class no and 268 instances in class yes. There are no missing attribute values. Here is a brief statistical analysis.

Attribute number:	Mean:	Standard Deviation:
1.	3.8	3.4
2.	121.7	30.4
3.	72.4	12.1
4.	29.1	8.8
5.	155.3	85.0
6.	32.5	6.9
7.	0.5	0.3
8.	33.2	11.8

From this table, we can see the data distribution of each attribute. As we can see, the distribution of attributes is very different and that will cause different influences of different attributes on the final result. To avoid these circumstances and ensure the reliability of the results, we cannot use the original data so we need to do some data preprocessing, data normalization. We did data normalization by Weka to make sure all values are in the range [0,1]. After doing the data normalization, save all data in a new file called pima.csv.

2.2 Correlation-based feature selection (CFS)

Correlation-based feature selection (CFS) is a feature selection method that can determine the number of features of the selected subset. It searches for the best subset of features, which is defined by a heuristic that considers how good a single feature is in predicting categories and how well they relate to other features. In the database, some attribute has no bearing on our class so we can remove these meaningless attributes which can also improve the accuracy of the classifier. We can do CFS in Weka.



We use “Best-First Search” as the search method. The attributes we select which have relation with our class are attribute 2, attribute 5, attribute 6, attribute 7 and attribute 8. So the selected attributes are *plasma glucose concentration a 2 hours in an oral glucose tolerance test, 2-Hour serum insulin (mu U/ml), body mass index (weight in kg/(height in m)^2), Diabetes pedigree function* and *age (years).*

3. Results and discussion

3.1 Results

The algorithm we choose are ZeroR, 1R, 1NN(k-nearest neighbors algorithm when k = 1, called IBK in Weka), 5NN(k-nearest neighbors algorithm when k = 5, called IBK in Weka), NB(Naïve Bayes), DT(Decision tree, called J48 in Weka), MLP(Multilayer Perceptron), SVM(Support vector machine, called SMO in Weka) and RF(Ransom forest). The algorithm called My1NN is the k-nearest neighbors algorithm when k equals 1 we wrote by ourselves. The algorithm called My5NN is the k-nearest neighbors algorithm when k equals 5 we wrote by ourselves. The algorithm called MyNBis is the Naïve Bayes algorithm we wrote by ourselves. For each algorithm, we calculated two accuracies, one is the accuracy after using CFS(Correlation-based feature selection) and the other one is accuracy without using CFS. The following

table contains the percentage of accuracy results calculated using the above algorithm, using 10-fold cross validation.

	ZeroR	1R	1NN	5NN	NB	DT	MLP	SVM	RF	My1NN	My5NN	MyNB
No feature Selection	65.1042 %	70.8333 %	67.8385 %	74.4792 %	75.1302 %	71.7448 %	75.3906 %	76.3021 %	74.8698 %	67.963%	74.739%	75.000%
CFS(%)	65.1042%	70.8333%	69.0104 %	74.4792 %	76.3021 %	73.3073 %	75.7813 %	76.6927 %	75.9115 %	69.000%	75.063	75.445%

3.2 Discussion of results

According to the result we get from the table above, no matter which classifier is used, the accuracy is between 65% and 78%. In these 12 classifiers with no feature selection, ZeroR has the lowest accuracy and SVM has the highest accuracy. In these 12 classifiers with CFS(Correlation-based feature selection), ZeroR has the lowest accuracy and SVM has the highest accuracy. Whether CFS is used or not, ZeroR and SVM have the lowest and highest accuracy respectively. ZeroR classifier is the simplest classifier. This method only selects a category with the highest probability as the classification result of unknown samples according to the statistical law of historical data, that is, for any unknown sample, the classification result is the same.SVM is a two class classification model. Its basic model is defined as the linear classifier with the largest interval in the feature space. In these 12 classifiers, there are 12 classifiers that have higher or equal accuracy after using CFS. The discussion of the effect of feature selection will be in 3.3. On the whole, the most suitable classifier for our database is SVM after using feature selection.

3.3 Effect of feature selection

Feature selection is to reduce the number of attributes based on *Pima Indians Diabetes Database*(original database). We hope to reduce the computational cost of modeling and improve the performance of the model by reducing attributes. There are 8 attributes in the original database and We use Weka to help us select attributes that are not related to our class. The selected attributes are the number of times pregnant, diastolic blood pressure and triceps skin fold thickness. All of them exactly have no relation with Pima Indians Diabetes. So we deleted these three attributes. Among the 12 classifiers we use, most of the classifiers that use CFS will

present higher accuracy than the same classifier but do not use CFS. Only classifiers 1R's, ZeroR's and 5NN's accuracy don't change after doing CFS. The reason for that may be the size of the database is too small, only 768 instances. And also there are only 8 attributes in this database and we only delete 3 attributes, therefore, CFS may have no effect on the final result . But in a word, for most classifiers for this database, using CFS can not only reduce the size of the database, but also improve the accuracy.

3.4 Comparison between the classifier

Compare our implementations of K-nn and NB with Weka's. The accuracy of our k-nn without using CFS when k = 1 is 67.963% and the accuracy of our k-nn with using CFS when k = 1 is 69%. The accuracy of Weka's k-nn without using CFS when k = 1 is 67.8385% and the accuracy of Weka's k-nn with CFS when k = 1 is 69.0104%. The accuracy of our k-nn when k = 1 without using CFS is higher than the Weka's one without using CFS, but the accuracy of our k-nn with using CFS is a little bit lower than the Weka's one with using CFS.

The accuracy of our k-nn without using CFS when k = 5 is 74.739% and the accuracy of our k-nn with CFS when k = 5 is 75.063%. The accuracy of Weka's k-nn without using CFS when k = 5 is 74.4792% and the accuracy of Weka's k-nn with CFS when k =5 is 74.4792%. Whether CFS is used or not, the accuracy of our k-nn when k = 5 is higher than Weka's but the difference is not very big.

The accuracy of our NB without using CFS is 75% and the accuracy of our NB with using CFS is 75.445%. The accuracy of Weka's NB without using CFS is 75.1302% and the accuracy of Weka's NB with CFS is 76.3021%. Whether CFS is used or not, the accuracy of our NB is less than Weka's but the difference is not very big.

On the whole, the classifier written by ourselves ,knn when k = 5 and NB, does not perform as well as Weka, but the difference is not big. The reason why the result calculated by our own classifier is different from that calculated by Weka is that we use different examples in 10 folds. Stratified cross-validation is a sampling technique in which samples appear in the same proportion as the population, but the specific instances in each fold are different.

4. Conclusions and future work

4.1 Conclusion

In this study, we use 12 classifiers to calculate two accuracies on the same data set. The first time we use our classifiers to calculate their accuracy on the original data set, and the second time we use the same classifiers to calculate their accuracy on the data using CFS. The results show that SVM performs best on our data set after using CFS. We also found that the accuracy of most classifiers using CFS is higher than the same classifier but without using CFS. In this study, we also compare the accuracy calculated by KNN algorithm and Nb algorithm written by ourselves with Weka's. With the same algorithm, the calculated results written by ourselves are different from those calculated by Weka.

4.2 Future work

The dataset we use, *Pima Indians Diabetes Database*, is supervised data. In the later research, we can select a group of unsupervised data and use the same steps to test the performance of each classifier on unsupervised data. This can better let us learn the performance of each classifier on different types of data sets. In addition, the capacity of this group of data is only 768, which is not enough for artificial intelligence calculus. The capacity of the data set can be increased. The more data, the more accurate the calculated classifier will be. Therefore, in future research, we can try to use larger data sets for experiments.

5. Reflection

In this study, we learned how to correctly use KNN, Nb and other classifiers and how to use Weka. By writing our own code, we have a deeper understanding of the internal algorithms of Nb and KNN classifiers and how they work. This study enables us to have a deeper understanding of machine learning. We also learned how to evaluate the performance of classifiers on the databases by comparing the accuracy and the effect of CFS on calculating the accuracy of the classifiers.