Report

The aim of this study is to build two classifiers of KNN (k-nearest neighbors algorithm) and NB (Naïve Bayes algorithm). With analyzing a sample dataset, hence comparing the results of MyClassifiers and the results of different classifiers within Weka, we can have a better understanding of these classifiers. Besides, with comparing results of using CFS (Correlation-based feature selection) and not using CFS, how CFS benefits our analyze is illustrated.

Our dataset is the Pima Indian Diabetes dataset which has been modified for consistency. It is originally sourced from UCI Machine Learning Repository. 768 instances described by 8 numeric attributes are in the dataset. Each entry in the dataset corresponds to a patient’s record; the attributes are personal characteristics and test measurements. There is a class containing only “yes” or “no” is to describe whether the person shows signs of diabetes or not.

The correlation feature selection (CFS) measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other". By applying Correlation-based feature selection from Weka, five attributes have been chosen, separately the second (Plasma glucose concentration a two hours in an oral glucose tolerance test), the fifth (two Hour serum insulin), the sixth (Body mass index), the seventh (Diabetes pedigree function) and the eighth (Age).

Below shows the result:

Weka:

	ZeroR	1R	1NN	5NN	NB	DT	MLP	SVM	RF
No CFS	65.1042%	70.8333%	67.8385%	74.4792%	75.1302%	71.7448%	75.3906%	76.3021%	74.8698%
CFS	65.1042%	70.8333%	69.0104%	74.4792%	76.3021%	73.3073%	75.7813%	76.6927%	75.9115%

MyClassifier:

	My1NN	My5NN	MyNB
No CFS	68.98172624076692%	71.62886994637914%	73.76987778456287%
CFS	69.24288749654937%	73.9148883665292%	74.62344093757057%

(Testing code for 10-fold cross validation is in the file “CrossValidation.java”)

From the result of Weka, 5NN, NB, MLP, SVM and RF all have accuracy nearly 75 percentage which is importantly higher than other classifiers. By comparing the results of whether CFS is applied, Weka shows that applying CFS literally makes no difference to the classifiers of ZeroR, 1R, 5NN. However, NB and DT classifiers benefit approximately 1 percent higher accuracy after CFS.

From the result of MyClassifier, what differs from Weka is that all 1NN, 5NN and NB classifiers show increased accuracy after applying CFS. Moreover, both Weka and MyClassifier reflect that naïve bayes is more sufficient than KNN classifier according to the higher accuracy.

In terms of conclusion, the accuracy of MyClassifier is generally lower than the same classifier in Weka, which means that there is still improvement can be achieved for my algorithm. And as for this data, NB, MLP, SVM and RF might be more suitable based on their high accuracy. Generally, no disadvantage of applying CFS is visible from the result, while the accuracy of all the classifiers either remains the same or slightly increased.