Data Analytics - Assignment 7

Jing Pang, Tian Xue, Chuqian Ma, Jiaxiang Leng

Introduction

Assignment 7 indicates that by using logic-based approaches to build classification models for two data sets, iris and congressional voting records, originating from the Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/). The dataset will be implemented with Weka software to successfully built a classification model. The Weka pre-constructs a bunch of classification algorithms to study the dataset. Therefore, we would go through a few classification algorithms in Weka and test with different training strategies.

Experiment of Iris dataset

In general, this data set contains 5 characters including sepal length in cm, sepal width in cm, petal length in cm, petal width in cm, and class. The whole dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant, such as iris setosa, iris versicolour, and iris virginica. The zeroR algorithm is used to test the dataset as a baseline performance of the learning algorithm, which in this case provides a 33.3% accuracy of classified instances on both methods of the training set and cross-validation. Therefore, in the idea of improvement of performance, the dataset will be applied into several classification algorithms, such as oneR, JRip, PART, and DecisionTable. The following table shows that the comparison of accuracy performances of different algorithms in several test options. The table below shows that the classification accuracy has influenced by different methods.

Algorithms	OneR	JRip	PART	DecisionTable
training set	96%	97.3333%	97.3333%	96%
10-fold cross-validation	92%	94%	94%	92.6667%
20-fold cross-validation	96.0784%	96%	95.3333%	93.3333%
66% split	92.6667%	96.0784%	96.0784%	96.0784%

Also, we applied the same principle of methods on tree classification models. The accuracy table reflected different running methods is shown on the below.

Algorithms	DecisionStump	HoeffdingTree	J48	LMT	RandomForest	RandomTree	REPTree
training set	66.6667%	96%	98%	98.6667%	100%	100%	96%
10-fold cross-validation	66.6667%	95.3333%	96%	94%	94.6667%	91.3333%	94%
20-fold cross-validation	66.6667%	95.3333%	96%	96.6667%	94.6667%	94%	95.3333%
66% split	62.7451%	92.1569%	96.0784%	98.0392%	96.0784%	94.1176%	92.1596%

Experiment of congressional voting records

This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition). This dataset should only conclude two options of choices between yes and no in ideally. However, the dataset contains three situations of voting records, which indicates y for yes, n for no and w for unknown. To predict whether vote or not, we extend the possibilities of results from two to three by thinking "unknown" as another parameter. And we used the result of zeroR algorithm as a baseline to compare the following algorithm research. In this case, we find out the accuracy of zeroR is 61.3793%. In the following tables, we try to different algorithms from traditional classification and tree classification on four common methods. The results are shown below.

Algorithms	**OneR**	**JRip**	**PART**	**DecisionTable**
training set	95.6322%	95.6322%	97.7011%	95.8621%
10-fold cross-validation	95.6322%	95.8621%	95.8621%	93.3333%
66% split	96.6216%	96.6216%	97.2973%	95.2703%

The result of tree classification shows at the table below.

Algorithms	**DecisionStump**	**HoeffdingTree**	**J48**	**LMT**	**RandomForest**	**RandomTree**	**REPTree**
training set	95.6322%	95.8621%	97.2414%	97.7011%	100%	100%	96.7816%
10-fold cross-validation	95.6322%	95.1724%	94.9425%	96.7816%	95.6322%	93.5632%	94.7126%
66% split	94.5946%	96.6216%	95.9459%	98.6486%	97.973%	95.9459%	96.6216%

Conclusion

The dataset of iris and congressional voting records have applied with two kinds of classifications, traditional and tree type. In details, for both classifications, there are several algorithms applied on both datasets to obtain certain classification predictions. This will also run by several methods to examine the influence of accuracy change. The algorithms of JRip and PART are better performance algorithm from all traditional classification aspect. In tree classification aspect, the algorithms of HeoffdingTree, J48, LMT, and RandomForest are the better performance algorithm from the group. However, there is no best algorithm out of all. Because algorithms are only performing better outcome in one method, and if method changes, the accuracy will drop quickly and another algorithm will become a leader.