CSE 572 Data Mining

Arizona State University

Course Project Report

Fanjie Lin 1203357332
Roshan Prabakar Raj 1209565263

Introduction of project

This report attempts to understand data mining course project, utilize the method we learned in CSE 572. There are several topics in this document, data Pre-processing, data classification, data validation and attributes reduction. We were provided with the training data sets and the testing data sets.

By using training.txt and label_training.txt, we could build a classification model to predict the result of testing.txt. There are number of techniques such as data preprocessing, data classification and data validation to achieve our goal which can be used to get the desired result.

Data Preprocessing

Weka was the software which we used in our project. libsvm was the model which we used in weka to get the desired result. We had to transform the text files to exact format so that Weka or libsvm can recognize them.

Data Classification and Validation

Another issue which we faced was which classification model should we use? There are couples of classifiers to analyze the data label. Two of the classifiers , Naive Bayes classification and Support Vector Machines (SVM) classification are the two methods we learned in the class. We were given the choice of using any model and we had to use that model which returned the highest accuracy. We tried a number of models to check which returned the highest accuracy. Finally after many trials we found out that Libsvm was obviously the better classifier in given case. Then, we used Weka to generate the classification model and then use this model to predict the test data.

Project steps

- Used file format to rewrite the input file into required format.
- Used the generated file as source of libsvm tool.
- Get the most accurate coefficient.
- Use the tool to predict result.

MODEL	TESTING ACCURACY
Decision tree j48	0.698
Libsvm	0.77

Table Testing data sets result

Attributes Reduction

There were a number of attributes given to us which confused us on the classifier to be used and make efficient analysis. There is a tool in Weka, called GainRatioAtrributeEval and Ranker. After using this tool, we get the attributes weight between attributes and classification result. Besides, we choose the attributes which have weight over 0.

Conclusion

We could find the steps of classification on the analysis of data mining. Moreover, we learned how to choose suitable parameter by executing the python script in this project. To compare the differences of classifier methods helps me to realize the concept of prediction completely. It is exciting to watch the correctly classified rate becoming accurate. Besides these techniques are also really useful and popular in real world. The project and analysis provides me a whole understanding of data mining, which we have never considered before.We were glad to have this project and what we achieved.

Reference

[1] http://www.cs.waikato.ac.nz/ml/weka/index.html
[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[3] http://weka.wikispaces.com/ARFF+%28stable+version%29