Journal of Applied Science and Engineering Innovation, Vol.2 No.4, 2015, pp.130-133
ISSN (Print): 2331-9062
ISSN (Online): 2331-9070

Analysis of Employment Data Mining for University Student based on Weka Platform

Lina Gao

Northeast Petroleum University, Qinhuangdao, Hebei066004,China

Abstract*:* This paper took the historical data of university graduates employment and the employment guidance as the study object and purposes, tried to find out the useful information hidden in the historical data of employment through the data mining of the historical data of university graduates employment. Therefore, this paper firstly did the preliminary study on the historical data of employment and established the subject of data mining, then built the data mining model to do the mining analysis by using data mining tools-Weka.

Keywords Data mining; Weka; Decision tree; ID3 algorithm; Decision making; Multi-valued and multi-fabled decision tree

INTRODUCTION

With the expansion and development of information technology in all aspects of the society, a lot of colleges and universities have accumulated and stored graduate employment information for many years in the process of their development. Facing to the huge amount of data and the complex information, the data mining technology not only can simply query and count the historical data, but also can find out the potential relationship between historical data and then extract the valuable knowledge from them; at last it can provide better decision-making advice for decision maker. [1]

In today's world, due to the rapid development of technology, the amount of data stored has been constantly increasing in every field. It is intended to obtain meaningful, valuable information that is not previously known from these data by applying data mining techniques. The data is for their sizes covers much space in pages but the merit of their usage is little. However, if we sum it up by putting numbers in an order, if we convert into meaningful sentences by arranging letters, and if we produce a melody by putting notes in a raw and if we produce a graphic or a picture of a tree by combining data on computer screen it is only at this point that we convert these data into information. Info covers less space as for its size in contrast to data but is more powerful in terms of usage worth. [2]

Because of this reason, data mining comes in the first line in the process of information exploration on databases. From these data in hand, data mining is extracting potentially handy information which is not so clear, unknown before and up closed. Data mining is not a solution itself at this point [3], instead it is a tool which supports decision making process and which tries to ensure required knowledge in order to reach the solution of the problem

ORGANIZATION OF THE TEXT AND THEORY

According to the characteristics of the employment data used by this paper, it selected the categorizing methods of decision tree on the basis of analysis of numerous data mining methods. Because the ID3 algorithms of decision tree tended to choose the attribute which had more values and its information gain computed complexly and its decision tree was too cumbersome, this paper improved the original ID3 algorithms and then compared it with the original ID3 algorithm. By comparison, the improved ID3 algorithm partly enhanced the classification accuracy and the lean degree of decision tree. Finally, this paper chose the improved ID3 algorithm to construct the data mining model. After studied and analyzed the result of the model, it obtained the following three aspects conclusions: (1) nearly half of the graduates who was in the normal logo did not choose the educational institution; (2) the native place of the students existed the relationship with the subordinate of employment units and the employment unit area; (3) different gender had a different emphasis on the employment.

Decision tree (Fig. 1) classifier takes much importance in the technology of Data mining. As the wide applications of classifier of Data mining, decision tree classifier has achieved much research achievement. Based on deep research in the main algorithms of decision tree, this dissertation finishes the designing and carrying of decision tree classifiers on the platform of WEKA, which not only realizes in effect the utilization of the data miner in existence, but also takes bravery researches and innovations in the rare concerned field of decision tree classifier, and also performs the realizing of new algorithm of decision tree. [4]

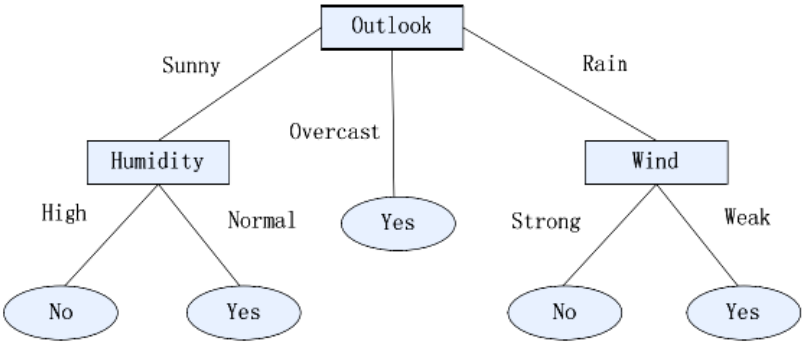

Figure 1Decision tree

Firstly, after descriptions in detail of the function and structure of data miner WEKA, based on the research of criterions for evaluating decision tree classifiers' performance, this dissertation takes analytic experiments of classic decision tree classifiers on the platform of WEKA, and does some comparisons and analyses according to different criterions for evaluating.

Secondly, this dissertation takes deep researches in the performing theory of classic decision tree classifiers. Based on serious study in the system structure of WEKA, we take the algorithm of SPRINT out, and also take test experiments for the algorithms' performance, which utilizes data miner in effect to carry out individual algorithms.

In order to improve the ability of dealing with multi-valued and mufti-Tabled data, this dissertation discusses a new mufti-valued and mufti-Tabled data decision tree classifier; on the basis of decision tree classifier in exist. In the new mufti-valued and mufti-Tabled decision tree, a new approach of measuring similarity considering both same and consistent features of label-sets is proposed. The new classifier is realized and tested in WEKA, the result of experiment shows that it has better classification efficiency for mufti-valued and mufti-Tabled data.

Model and Analysis

After applying the model, I achieved an accuracy of 88.68 % (for IE.arff), meaning that 47 instances out of 53 were correctly classified in or model. For the CIG.arff data file I reached 71.74% accuracy, meaning that 33 instances out of 45 were correctly classified. I also obtained the values of several performance measures for numeric prediction, presented in Table 1.

From the confusion matrix result for the IE.arff data I established that two instances of the Disagree class were assigned to the Agree class, four instances of the Neutral class were assigned to Agree class. For the CIG.arff three instances of the Disagree class were assigned to the Agree class, one instance of the Disagree class were assigned to the Neutral class, seven instances of the Neutral class were assigned to Agree class and two instances of the Agree class were assigned to the Neutral class. The decision tree resulted from the first data set of the IE.arff file (Figure 2) has as a central root joint the high_school (graduated high school) attribute, the main mean to differentiate the IE students' choice in continuing their education. For the second level ramification, the IE students who graduated mostly a high school oriented on Information systems are influenced in their decision by the books, course materials, case studies of the highest quality they received.

Table 1. Performance measures' results

Performance measures	Result IE.arff	Result CIG.ARFF
Kappa statistic	0	-0.01
MAE, mean absolute error	0.12	0.25
RMSE, root mean square error	0.26	0.41
RAE, relative absolute error (%)	67.18	86.03
RRSE, root relative squared error (%)	97.47	112.07

Our research guiding directions in data mining studies of students' behavior are presently continued with concentration on the following factors: clustering techniques; decisional trees for each specialization based on several algorithms; parallel analysis with the data extracted from the master degree students to exemplify detailed behavioral models.

This section is devoted to describing in detail how to implement or to import an algorithm into the Weka software tool. The Weka philosophy tries to include the fewest possible constraints for the developer, in order to ease the inclusion of new algorithms within this tool. Thus, it is not necessary to follow the guidelines of any design pattern or framework in the development of a new method. In fact, each algorithm has its source code in a single folder and does not depend on a structure of classes, making the integration of new methods straightforward.

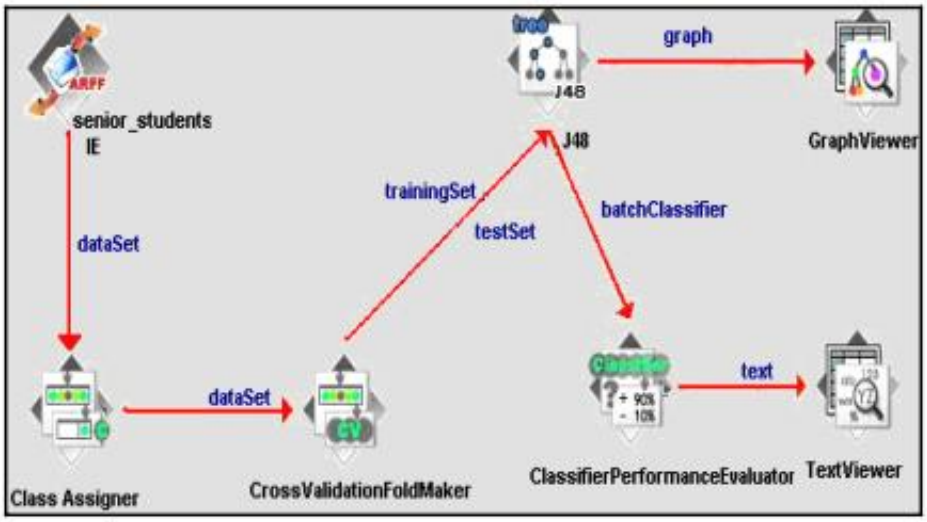

Figure 2: The graph model built after [5]

We enumerate the list of details to take into account before codifying a method for the Weka software, which is also detailed at the Weka Reference Manual (Figure 3).

Nowadays, the use of statistical tests to improve the evaluation process of the performance of a new method has become a widespread technique in the field of Data Mining. Usually, they are employed inside the framework of any experimental analysis to decide when an algorithm is better than other one. This task, which may not be trivial, has become necessary when a new proposed method offers a improvement over the existing methods for a given problem. There exist two kinds of test: parametric and non-parametric, depending of the concrete type of data employed. As a general rule, a non-parametric test is less restrictive than a parametric one, although it is less robust than a parametric when data are well conditioned.

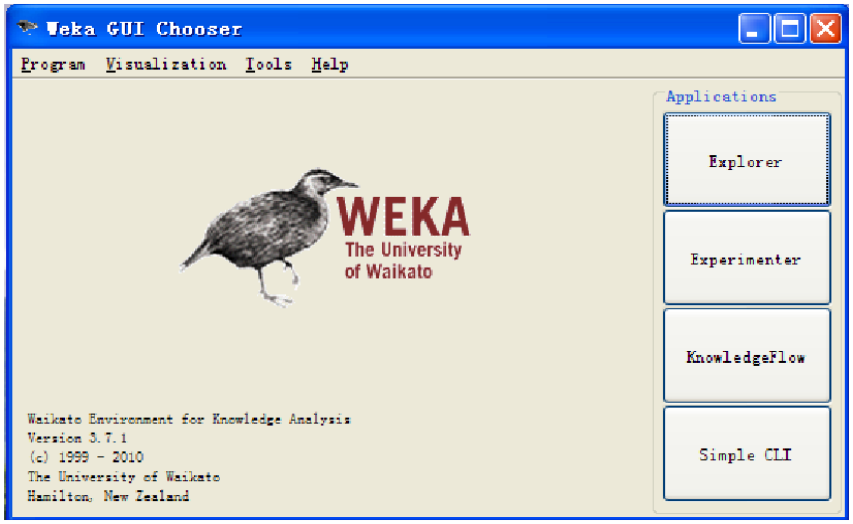

Figure 3: Weka Interface

Parametric tests have been commonly used in the analysis of experiments in DM. For example, a common way to test whether the difference between the results of two algorithms is non-random is to compute a paired t-test, which checks whether the average difference in their performance over the data sets is different from zero. When comparing a set of multiple algorithms, the common statistical method for testing the differences between more than two related sample means is the repeated-measures ANOVA (or within-subjects ANOVA). Unfortunately, parametric tests are based on assumptions which are most probably violated when analyzing the performance of computational intelligence and data mining algorithms. These assumptions are known as independence, normality.

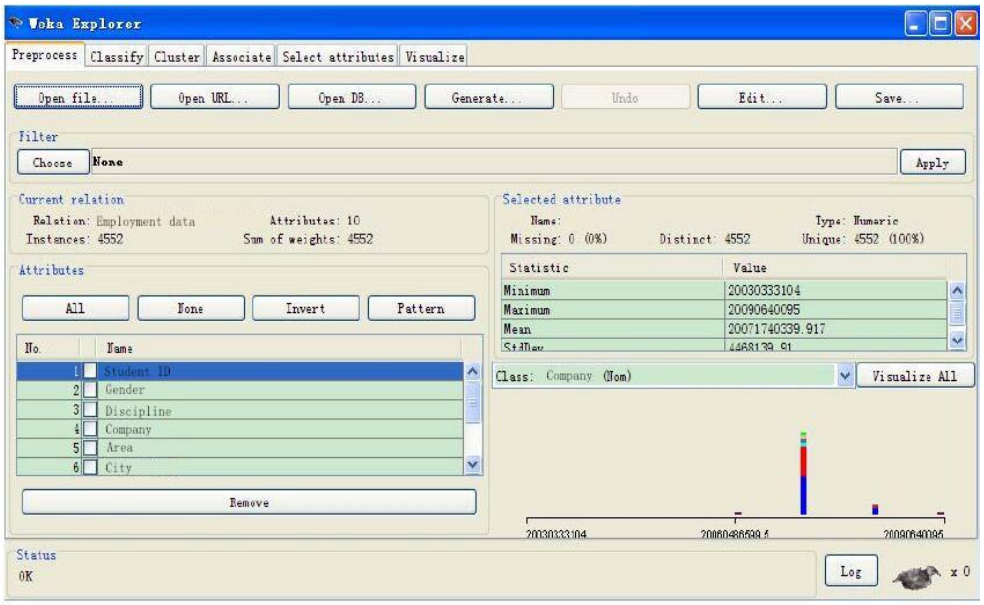

Figure 4: Weka Preprocess Interface

Weka is one of the fewest Data Mining software tools that provide to the researcher a complete set of statistical procedures and multiple comparisons. Inside the Weka environment, several parametric and nonparametric procedures have been coded, which should help to contrast the results obtained in any experiment performed with the software tool. These tests follow the same methodology that the rest of elements of Weka, making easy both its employment and its integration inside a complete experimental study. Weka preprocess interface is shown in Figure 4.

CONCLUSION

Since the data used by this paper only included the employment information data it did not involve the personal general information and performance information data and its data attributes were nominal attributes. Therefore, under the condition of the corresponding comprehensive factors consideration of the multi-dimension data set, the methods building models analysis influencing graduate employment and the mining would be left behind for further research.

REFERENCES

A. Ghosh and L.C. Jain. Evolutionary Computation in Data Mining, SpringerVerlag, 2005.

J.S. Aguilar-Ruiz, J.C. Riquelme, and M. Toro. Evolutionary learning of hierarchical decision rules, IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, 33(2) (2003) 324-331.

Witten I. Frank E. WEKA Machine Learning Algorithms in Java, Morgan Kaufmann Publishers, 2000.

Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufman, 2003.