Application of Open Source Data Mining Software Weka in Marketing Teaching

Zhengteng Hao(✉)

School of Economics and Management, Qinghai Nationalities University, Xining 810007, Qinghai, China

Abstract. Marketing is a professional basic course for economic management students. In the process of teaching, we should not only teach students theoretical knowledge, but also teach students to integrate into the real economic management activities. In order to stimulate students' interest in learning and improve students' practical ability, this paper discusses the teaching method of introducing open source data mining software Weka into classroom teaching, and gives practical teaching examples to improve the teaching effect. The training quality of the marketing course has been a useful attempt, and the classroom effect is good.

Keywords: Weka · Marketing teaching · Apriori correlation method · Shopping basket analysis

1 Introduction

Marketing is a professional basic course for students majoring in economic management. With the rapid development of information technology, both commodity information and customer information are massive. How to mine useful marketing information in big data and then apply it to our marketing is a problem that marketing students need to solve. Data mining technology in computer science is to solve the problem of automatically analyzing and discovering useful information in large databases. Apriori association algorithm is mainly used to discover meaningful connections hidden in large data. This paper attempts to explain the application of Apriori association algorithm in supermarket data analysis by taking open source software Weka as an example in marketing, In order to guide students to use data mining methods to solve practical problems [1].

2 Data Mining Weka

2.1 Open Source Software WEAK

The full name of Weka is Waikato environment for knowledge analysis. Its source code can be downloaded from http://www.cs.waikato.ac.nz/ml/WEKA Get it. At the 11th acmsigkdd International Conference, the Weka group of Waikato University won the

M. A. Jan and F. Khan (Eds.): BigIoT-EDU 2021, LNICST 392, pp. 356–365, 2021.
https://doi.org/10.1007/978-3-030-87903-7_44

highest service award in the field of data mining and knowledge exploration. Weka system has been widely recognized and has become one of the more complete data mining tools. As an open source data mining platform, Weka integrates a large number of machine learning algorithms that can undertake the task of data mining, including data preprocessing, classification, regression, clustering, association rules and visualization on the new interactive interface [2–5]. In modern business society, the data of enterprises are generally massive. If students can use the advanced software Weka to analyze the marketing data management and dig out the hidden relationships from the massive data, it will certainly be of great benefit to the mining and utilization of marketing data and the discovery of business opportunities.

2.2 Definition and Research Significance of Software Defect Prediction

With the increasing dependence on computer software, how to effectively improve the quality of software has become the focus and difficulty in the field of software engineering. The traditional software quality assurance methods (such as static code review or dynamic software testing) are inefficient, which need a lot of manpower and material, and need additional modification operations such as code instrumentation. Software defect prediction technology has gradually become one of the research hotspots in the field of software engineering - 4, 5, 6, 8, 9, 10.1, 12 software defect prediction, generally refers to the analysis of software source code or development process, design the degree element that has correlation with software defects, and then mine and analyze the software history warehouse to create defect prediction data set [6]. Based on the defect prediction data set, a specific modeling method (machine learning) is used to build the defect prediction model, which can predict and analyze the potential defect modules in the subsequent versions of the software. 13 software defect prediction can greatly reduce the manpower and material resources required for testing, and there is no need to modify the code. After the development of software department and module is completed, the defect prediction can be carried out in time, and the potential defect modules can be identified in advance, so that the project director can optimize the allocation of test resources and improve the test efficiency and software product quality [8–10]. Therefore, the research on software defect prediction technology, the establishment of software defect prediction model, and the early prediction and identification of software defects are not only of great research significance, but also of great application value.

2.3 Data Mining Software Weka

With the rapid development of science and technology and social economy, computer technology has been widely used in various industries. Computer software (system) is more and more closely related to people's work and life. The impact of software quality and system reliability on the efficiency and safety of production and management activities is also growing. However, with the continuous growth of people's demand for software, the scale of software is becoming larger. With the increase of complexity, software development and maintenance become more and more difficult. Software with hidden defects may lead to software failure or system collapse in the process of operation. Serious software defects will bring huge economic losses to enterprises, and may even lead to casualties. Unclear user requirements, non-standard software development process, lack of experience and ability of software developers and other reasons will lead to software defects. In order to ensure the quality of software, software testers will check software defects through certain software testing methods, and then inform developers to deal with software defects. However, affected by the actual factors such as the development schedule and cost control of software projects, software test engineers can not completely test all software modules.

3 Software Defect Prediction Method

According to the different methods of software defect prediction, it can be divided into static prediction and dynamic prediction. Static prediction mainly quantifies software code into static features, makes statistical analysis on these features and historical defect information, mines the distribution law of historical defects and constructs a prediction model, and then forecasts new program modules based on the prediction model. Dynamic prediction is to analyze the occurrence time of software defects, mining the relationship between software defects and their occurrence time [11]. At present, with the wide application of machine learning algorithm, static software defect prediction has achieved good prediction results, which has attracted more attention of researchers.

3.1 Software Defect Prediction Process

The process of software defect prediction can be divided into four stages. The software history warehouse is mined and analyzed, from which the program modules are extracted. The granularity of program module can be set as file, package, class, function, etc. After the modules are extracted, these program modules are marked as defective or non defective modules respectively. Design metrics. Based on the software static code or software development process, the corresponding metrics are designed to measure the software of program modules, and the defect prediction data set is constructed. Build defect prediction model. After the necessary data preprocessing of the defect prediction data set, the software defect prediction model is constructed with the help of specific modeling methods (such as machine learning method). Defect prediction. The software defect prediction model is used to predict the new program module, and the program module is predicted to be defect free and defect freep [12]. The prediction target can also be the defect number or defect density contained in the program module.

3.2 Metric Meta Design

The design of metrics is the core problem in the research of software defect prediction. Typical metrics include the number of lines of code, McCabe loop complexity and Halstead scientific metrics. Taking the complexity of the McCabe loop as a metric, the complexity of the control flow of the program is mainly considered. The assumption is that the higher the complexity of the control flow of the program module, the higher the possibility of containing defects. The assumption is that if there are more operators and operands in the program module, the more difficult it is to read the code, and the more likely it is to contain defects. Some researchers design metrics based on software development process. 2021 based on software development process mainly considers software project management, developer experience, code modification characteristics and so on. Researchers have conducted a comparative study on which way to design metrics is more effective in building defect prediction models. Graves 22, moser21 and others believe that the degree element designed based on the development process is more effective [13]. Menzies23 and others believe that the metric element of static code can build a high-quality defect prediction model. It can be seen that there is no obvious difference between the two. Reasonable design of metrics can achieve good prediction results.

3.3 Construction and Application of Defect Prediction Model

Before building the defect prediction model, it may be necessary to preprocess the defect prediction data set according to the quality of the data set. This is because the data set may have problems such as noise, dimension disaster and class imbalance. Preprocessing can improve the quality of the data set. 2 after preprocessing the data set, the defect prediction model is constructed with the help of certain algorithms. As the core of artificial intelligence and teaching data science, machine learning has been widely used in software defect prediction, which is selected as the algorithm to build software defect prediction model [14]. Machine learning methods commonly used in software defect prediction can be divided into classification methods and regression methods. Classification methods mainly include classification regression tree, naive Bayes, k nearest neighbor, support vector machine, ensemble learning and cluster analysis; regression methods mainly include linear regression, polynomial regression, stepwise regression and elastic regression.

4 Analysis Method of Curriculum Relevance

Pearson product distance correlation is mainly used to calculate the correlation between continuous variables. The Pearson correlation coefficient is meaningful only when the population of variables is normal distribution or can be approximately regarded as normal distribution, and the number of samples is not less than 30.

(1) The formula for calculating Pearson product distance correlation coefficient is as follows 1:

$$r_{AB} = \frac{\sum_{i=1}^{n} (a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n} (a_i - \overline{a})^2 \sum_{i=1}^{n} (b_i - \overline{b})^2}} \tag{1}$$

(2) The test statistic formula is shown in Fig. 2:

$$t = \frac{r_{AB}\sqrt{n-2}}{\sqrt{1 - r_{AB}^2}} \tag{2}$$

In terms of technical implementation, the system is developed based on spring MVC web framework and mybatis architecture [15]. The platform can be divided into two core parts, one is data transmission function, the other is data mining function, each part is the complete framework of MVC theory.

The overall architecture of the system is shown in Fig. 1:

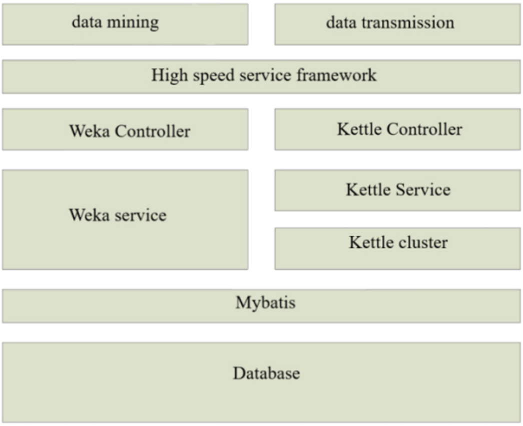

Fig. 1. Platform technical architecture

The technical architecture of the system is mainly divided into two functional applications, both of which are realized by three-tier MVC structure. The two functional applications are data transmission application and data mining application. The two applications implement the method call through the framework, which makes the application relatively independent and convenient for development.

5 Application of Open Source Software Weka in Marketing Teaching

In the process of teaching, we choose the shopping basket analysis experiment. The shopping basket analysis is to apply the correlation technology to the transaction process, especially to analyze the supermarket cashier data, and find out the commodities that appear in groups. For marketers, this is the main source of sales information for data mining. For example, after automatic analysis of the cashier data, it is found that customers who buy beer also buy potato chips. This discovery is of great significance to supermarket managers [16]. This information can be used for a variety of purposes, For example, planning the location of shelves, selling only one of the products that will be purchased at the same time at a discount, and providing product coupons that match the products sold separately. Businesses can also identify special customers from customers' purchase behaviors, not only analyzing their historical purchase patterns, but also providing a new way to identify the special customers, Moreover, it can accurately provide special purchase information that may be of great interest to potential users.

In the teaching experiment, we use a supermarket shopping basket analysis data set of Weka, the file name is supermarket.ar. This data set is collected from a real supermarket in New Zealand. There are 217 attributes and 4627 instances in the data set. It is very suitable for shopping basket analysis experiments. First, we use the preprocessing panel of Weka's Explorer interface, Load the supermarket.arf. In the current relation sub panel, we can see the basic information of the dataset. Because the dataset has many attributes and a large amount of data, students will click the Edit button on the top of the preprocessing panel to open the viewer window of the dataset and view the data file. Let students understand the properties and structure of data through proper explanation.

Among the ten association rules obtained from the calculation results, many commodities appear many times, and the total amount is very high. We can conclude that: first, customers who buy Biscuits, frozen food and other fast food will buy fruits and vegetables to supplement their body's vitamins; second, customers who buy Biscuits, frozen food, fruits and vegetables will buy bread and cakes; third, customers who buy the above food will buy a large amount at a time, and the total amount will be very high; fourth, transactions with a high total amount, They usually buy bread and cakes and so on. If the information is provided to the supermarket, we can rearrange the shelves, rearrange the supermarket, provide fast payment channels, arrange delivery and other additional services according to the knowledge, so as to enhance the market competitiveness.

For data mining, data is the core. The quality of data can directly affect the results of data mining. However, the actual data is very heavy. The so-called forest is big, there are all kinds of birds, and any problems may appear when there are more data, such as incomplete data, default attribute values, noisy data in the data set, wrong data, different coding formats or naming rules in the data set [17]. This will bring extra difficulties for data mining, even poor data will lead to the result of the error, so the data preprocessing is a step that can not be omitted.

If a data mining tool wants to have good performance, preprocessing is essential. In the process of mining, the preprocessing process will take up most of the time. There are four steps in preprocessing: data cleaning, integration and transformation, rule constraint and concept layering. Data cleaning includes: incomplete data processing, processing noise points and inconsistent data (see Fig. 2).

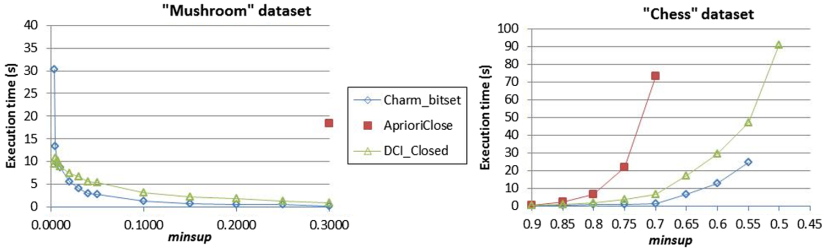

Fig. 2. Results of simulation with open source data mining (a)

K-means is a classic clustering algorithm. The basic principle of kmeans is: after determining the number of clusters K, k points are randomly selected as the center of the cluster. According to the Euclidean distance between each instance point and the center point in the data, each instance point is assigned to the cluster where the nearest center point is located. Then we get the cluster through the previous step, get its centroid point, redefine the centroid point as the center point, and then repeat the whole process [18–20]. In the continuous iteration process, the center of the cluster will continue to change, until in several successive iterations, the center of each cluster is exactly the same as the center of the previous round, which means that the cluster has been allocated, and the kmeans algorithm has been successfully implemented (see Fig. 3).

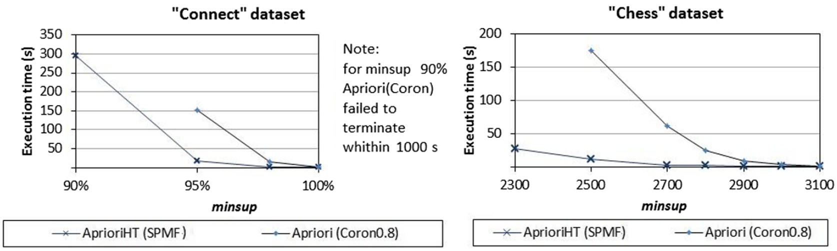

Fig. 3. Results of simulation with open source data mining (b)

6 Teaching Objectives and Contents

The purpose of this course is to enable students to understand the basic direction of machine learning; to master the basic algorithm of machine learning; to master the method of using Weka platform to realize machine learning algorithm; to understand the relevant research ideas of machine learning, from which to learn some methods of pioneers to solve problems; to further understand the usage and performance of learning algorithm through experiments, As shown in Fig. 4. In order to achieve these goals, on the basis of fully referring to the existing classic machine learning textbooks, the course team designed the following teaching content and syllabus: This course will take the classification task in data mining as an example, First, the evaluation of classification model is explained, and then a number of classical and commonly used machine learning technologies are explained. The specific chapters are arranged as follows: Chapter 1: introduction [21]. Explain the definition of machine learning, the difference and connection between machine learning and data mining, the teaching ideas and content arrangement of this course, as well as the teaching materials and reference books used in this course. Chapter 2: explain the method, index and comparative test of model evaluation. Chapter 3–9: explain the basic technology of machine learning: start with linear regression, explain linear learning: end with K-means clustering, explain unsupervised learning; the middle includes support machine learning, neural network learning, decision tree learning, Bayesian learning and nearest neighbor learning. Chapter 10–13: introduce the advanced technology of machine learning, including ensemble learning, cost sensitive learning, evolutionary learning and reinforcement learning.

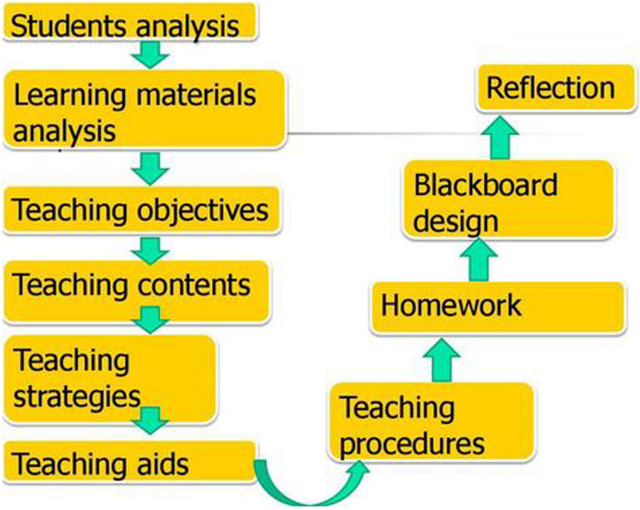

Fig. 4. Shows the teaching objectives

7 Teaching Evaluation and Assessment

7.1 Machine Learning

For machine learning, an elective course of computer major, which attaches great importance to the cultivation of hands-on ability, classroom theoretical knowledge teaching is of course important, but the more important thing is extracurricular practice, so that students' practical ability can be improved, practical application problems can be solved, and truly useful machine learning algorithms can be written [22]. In order to achieve this goal, the course team designed eight practical training and assessment problems after class, which are mainly used to consolidate and review the theory of machine learning algorithm taught in class, and master the implementation of machine learning algorithm by using the international machine learning open source experimental platform Weka. Through the experiment, we can further understand the usage and performance of machine learning algorithm, Improve the programming ability of machine learning algorithm: (1) implementation and experimental test of linear regression algorithm (10 points); (2) implementation and experimental test of logistic regression algorithm (10 points); (3) implementation and experimental test of SMO classification algorithm (15 points); (4) implementation and experimental test of BP classification algorithm (15 points); (5) implementation and experimental test of D3 classification algorithm (15 points); (4) implementation and experimental test of BP classification algorithm (15 points); (5) implementation and experimental test of D3 classification algorithm (15 points); (6) NB classification algorithm implementation and experimental test (7) KNN classification algorithm implementation and experimental test (10 points).

7.2 All Aspects of the Course

Implementation of K-means clustering algorithm and experimental test (10 points) students submit AVA code file based on Weka platform and screenshot of test results on any data set with Explorer as the basis of assessment and scoring. Machine learning is a new and very important elective course for computer and related majors. It has been widely used in many fields, such as image recognition, intelligent medical treatment, market analysis, financial investment, fraud screening, environmental protection, scientific research and so on, and has achieved considerable social effects, showing a good application prospect. How to teach this course well in University and improve students' practical ability to solve practical problems with machine learning technology is a problem that every teacher needs to seriously consider [23]. In this paper, from the teaching objectives and content, teaching methods and characteristics, to teaching evaluation and assessment, a complete description of all aspects of this course. After years of efforts, this course has been opened twice on the MOOC platform of love course China University, with 12913 and 12671 students respectively. In the third time, 2702 students have been pre selected to participate in the course, which has been widely praised by other college students and social learners, and has been selected into the first batch of excellent online open courses for undergraduates in Hubei Province [24]. The future work mainly includes to further supplement and improve the teaching content of the course, put forward new teaching methods, and improve the content and scoring standard of teaching evaluation and assessment according to the feedback of students.

8 Conclusions

In order to stimulate students' interest in learning marketing course, we try to apply an open source tool Weka in the teaching process, and with the help of its visual environment and typical algorithm, we demonstrate a practical problem-solving process for students in class. Through these teaching steps, students can gradually understand the open source software Weka and master the use of typical algorithm. We use Weka to process and analyze business data, improve data processing ability, and mine valuable information for marketing. At the same time, Weka software is open source software. For students with programming foundation, they can analyze the principle of the algorithm, and also optimize the algorithm through their own programming to further improve their ability to solve problems.

References

1. Chu, H.: Design and Implementation of University Data Exchange Platform Based on ETL. University of Electronic Science and Technology, Chengdu (2014)
2. Zhang, C.: Design and Implementation of Enterprise Data Exchange Platform Based on BTL. Harbin Institute of Technology, Heilongjiang (2016)
3. Xu, J., Pei, Y.: Review of data ETL. Comput. Sci. **38**(4), 15–20 (2011)
4. Zhang, R.: Review of ETL data extraction, software guide. **9**(10), 164–165 (2010)
5. Zhang, S.: The application of experiential teaching in the marketing teaching of secondary vocational education. Mod. Vocat. Educ. (30), 95 (2017)

6. Jiang, Y.: Application of marketing simulation training software in marketing course teaching. Contemp. Tourism (Golf Travel) (10), 193+211 (2017)
7. Weifang: The application of case teaching in marketing teaching. China Natl. Expo (10), 71–72 (2017)
8. Zhang, Y., Li, X.: A study on the teaching reform of flipped classroom mode in marketing teaching in Colleges and universities. J. High. Educ. (19), 121–123 (2017)
9. Wang, W.: Application of task driven teaching method in marketing teaching. China Market (28), 232–233+237 (2017)
10. Jin, Z., Zheng, C.: PBL teaching method in marketing teaching in Colleges and universities. Era Educ. (19), 160–161 (2017)
11. Liyaqin: The application of participatory case teaching in Marketing Teaching: based on constructivism learning theory. Contemp. Teach. Res. Cluster (10), 20 (2017)
12. Xialiping: Application of inquiry teaching method in the course of tourism marketing. Knowl. Econ. (20), 112–113 (2017)
13. Luo, X.: The application of experience teaching mode in automobile marketing teaching of Higher Vocational Colleges. Tax Payment (27), 120 (2017)
14. Liyifang: The application of mixed teaching mode in marketing course teaching. New Campus (Reading) (09), 24–25 (2017)
15. Wang, D.: The new application of case teaching method in marketing course teaching. New Campus (Early Ten Days) (09), 50 (2017)
16. Zhao, J.: Explore the application of interactive teaching mode in marketing teaching of higher vocational education. China Market (25), 227–228 (2017)
17. Xu, C.: Application analysis of project teaching method in marketing planning course teaching. Mod. Bus. Trade Ind. (25), 164–165 (2017)
18. Wang, J.: The application of project teaching method in the teaching of network marketing in secondary vocational school. Mod. Vocat. Educ. (24), 51 (2017)
19. Zousha: The application of project-based teaching method in tourism marketing teaching. Mod. Mark. (Next Issue) (07), 76 (2017)
20. This paper discusses the innovative application of SPOC in the teaching of marketing courses in Higher Vocational Colleges. China Mark. (22), 112+117 (2017)
21. Chen, H.: On the application of situational teaching method in marketing major teaching. Sci. Educ. Lit. (Next Issue) (07), 41–42 (2017)
22. Yangzhaoyun, Liyiting, Chenrainbow, Zhangdechun: Analysis of marketing teaching reform of Applied Undergraduate. Mod. Econ. Inf. (14), 421 (2017)
23. Hu, M.: The application of micro course in marketing course teaching in Higher Vocational Colleges. Educ. Mod. **4**(29), 242–243 (2017)
24. Wang, Y.: Application of case teaching method in marketing teaching. Mod. Shopping Malls (13), 251–252 (2017)