Teaching and Experimental Schema of Data Mining Technology Combined with the Cloud Computing

Hong Shi[1], Hongxia Deng[2]

[1] College of artificial intelligence, Shenzhen Polytechnic, China
[2] School of computer science and technology, Taiyuan University of Technology, China
shh51@sina.com, denghongxia@tyut.edu.cn

Abstract

The technology of data mining is abstract and difficult to understand, a teaching and experimental schema of data mining technology is designed to resolve that problem, and the proposed schema includes data classification, data clustering, data dimensionality reduction and association rules. Datasets of different types are used as the experimental datasets including sensors datasets, Internet datasets and social media datasets, at the same time, cloud computing technology is adopted to improve the efficiency of computation and storage, so that the efficiency of lessons and experiments is improved too. Graphic interface is designed to generate the graph and table results of data classification, data clustering, data dimensionality and association rules, and it make the data mining technology easy to understand. Lastly, experimental results of data mining show that the proposed schema generates accuracy plot graphs of data mining technologies, and make the data mining workflow visual.

Keywords: Cloud computing, Big data, Data mining, Curriculum revolution, Data clustering, Data dimensionality reduction, Association rules

1 Introduction

Data mining technology can find hidden internal information from massive data, including the relationship between data, the trend of data flow, data mode and other information. At present, data mining has become a very important technology in the field of electronic information [1]. Data mining technology is composed of several sub technologies, including association rule mining [2], frequent itemset mining [3], pattern mining [3], data clustering, etc. these sub technologies are closely related to mathematics and are relatively abstract, which makes it difficult for beginners to understand.

Data mining is a basic course for computer and information majors in Colleges and universities. The abstract characteristics of data mining technology make it difficult for students to understand, which brings great difficulties to teaching. Moreover, the presentation in the class proposes a higher real-time

requirement to the data mining experiments. We introduce the cloud computing to reduce the delay of the experiments by offloading the computation-intensive tasks to the cloud server. Intelligent scheduling schemes with Auto-Learning ability can be used to improve the utility of the network resource and balance the workload of the cloud server.

Data mining teaching and experimental scheme [4] should meet the following requirements: ① support different data mining sub technologies; ② quickly and accurately simulate the workflow of data mining; ③ use cloud computing technology to improve the efficiency of data analysis and storage; ④ provide a visual interface for data mining. At present, the mainstream business software of data mining mainly includes orange4ws [5] (orange for web services), knime (Konstanz information miner) analytics platform [6], rapidminer studio, etc., these commercial software are expensive, and can not meet the above four requirements at the same time, so they can not be directly used in data mining teaching and experiment in Colleges and universities.

In order to improve the comprehensibility of data mining course, a teaching and experimental scheme for data mining technology is designed. This scheme collects different types of data sets, such as sensor data, Internet data, social media data, and uses cloud computing technology to analyze and process the data sets, so as to improve the efficiency of data processing, thus improving the efficiency of teaching. In addition, the graphical interface of PC is designed, which can describe the results of data classification, data clustering and data dimension reduction in the form of image, which improves the comprehensibility of data mining technology.

The structure of this paper is as follows: the first part describes the relevant research and background knowledge; the second part describes the main framework of the data mining simulation experiment system; the third part describes the workflow of the

*Corresponding Author: HongXia Deng; E-mail: denghongxia@tyut.edu.cn

DOI: 10.3966/160792642021012201015

data mining simulation system; the fourth part describes the use method and actual effect of the data mining simulation system; the fifth part summarizes the full text and prospects for the future.

2 Data Mining Processing Based on Cloud Computing

In order to meet the four requirements of data mining teaching and experiment, this paper designs a data mining simulation experiment (DMSE) system. The system not only meets the above four requirements, but also considers user rights management, portability and scalability. Figure 1 shows the overall structure of DMSE system. The graphical user interface (GUI) in the front-end part is responsible for interacting with users, and the computing service module in the back-end part is responsible for processing the user's request.

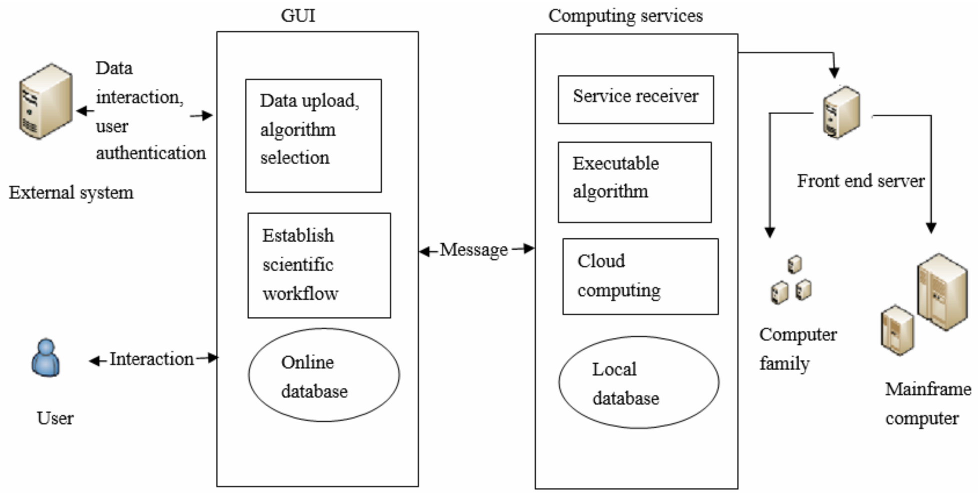


Figure 1. Overall structure of DMSE system

2.1 DMSE User Rights Management

Two kinds of user rights are designed: manager and ordinary user. Ordinary users do not need to know the internal information of DMSE, only need to operate the software to complete the simulation and experiment of data mining. Managers should be familiar with the structure of DMSE, modify the experimental data set of DMSE, and integrate new data mining technologies.

2.2 Internal Structure of DMSE

DMSE is mainly composed of GUI (graphical user interface) and computing service module. GUI and computing service module communicate through simple object access protocol [7] (soap) messages. If the GUI sends a SOAP message to the computing service module, the message contains specific data mining algorithms, and each SOAP message triggers the computing service module to run the corresponding data mining algorithm. In order to improve the processing speed of the system, cloud computing is used to implement the computing service module. The scheduler of cloud computing allocates processor resources and memory resources for data mining computing tasks.

By separating GUI from computing service module, the portability of DMSE can be effectively improved. The computing service module can be upgraded independently and the latest open source cloud computing technology can be used. GUI is responsible for the interaction between users and DMSE, provides data mining scheme for end users, and provides friendly interface for users. Users can observe the process of data mining. DMSE adopts the technology of separating cloud computing technology and computing resources, which makes DMSE have four advantages: ① using cloud storage technology to reduce the difficulty of data management; ② using cloud computing technology to improve the management efficiency of computing resources; ③ simply introducing new data mining algorithm; ④ being able to simulate the workflow of data mining algorithm.

2.3 User usage of DMSE

The process of data analysis by DMSE users using data mining algorithm is mainly divided into the following steps: firstly, DMSE users initialize and manage data mining experiments; secondly, model and simulate data analysis program; finally, DMSE saves the experimental results and data analysis results in cloud storage, so that DMSE users can access the experimental results at any time.

The initialization and management of data mining experiment is mainly realized by scientific workflow. DMSE users upload a data file, select data preprocessing, statistical features calculation, data dimension reduction, data clustering and classification and other sub technologies to observe the simulation results of DMSE. The above elements constitute the scientific workflow of data mining experiment. Figure 3 shows the GUI interface of DMSE scientific workflow. Users modify the scientific workflow of data mining by adding, deleting or modifying the above elements.

Cloud computing and Internet of things have been widely used, so data mining technology needs to deal with data sets from different devices and social media. Such data sets are generally high-dimensional data, which brings great difficulties to data mining processing. The goal of dimension reduction is to extract its low dimensional structure from high-dimensional data sets and map the data from high-dimensional space to low-dimensional space. Dimension reduction is a very important step in the field of data mining. It is very important for large data sets. After the large data sets are mapped to the low dimensional space, other data mining algorithms can be used to further analyze the low dimensional data sets. In addition, some features of the data set belong to redundant features, which can be eliminated by dimension reduction technology to improve the efficiency and accuracy of data mining.

Because the high-dimensional space cannot be represented in the form of images, and the DMSE system needs to provide users with a visual structure of the data set, so dimension reduction is an important part of DMSE. The DMSE in this paper integrates several classical dimensionality reduction algorithms, namely:

- Principal component analysis (PCA) [8]: PCA is a widely used dimensionality reduction technique.
- Multidimensional scaling (MDS) [9]: MDS is widely used for dimensionality reduction and visualization of high-dimensional data. This technology mainly uses the distance information between high-dimensional data points to transform high-dimensional data into low-dimensional data.
- Relative to MDS [10]: MDS technology retains the topological structure of high-dimensional data, but does not provide mapping of new data points. MDS provides a mapping mechanism for new data points.
- Self organizing maps multidimensional scaling (som-mds) [11]: SOM is a neural network structure, and the result is a set of neurons, which represents the high-dimensional data points, and then the MDS maps the high-dimensional data points to the two-dimensional plane. The combination of SOM and MDS can improve the dimensionality reduction efficiency of big data.

DMSE also integrates the classic algorithms of data preprocessing, data classification, data clustering and other sub technologies, such as feature selection, data partition, outlier elimination, data statistical feature calculation, random decision prediction, etc. In order to output the graphical results of data mining, the 2D coordinate graph representation of data points is designed.

3 Modeling Steps of Data Mining Workflow

DMSE supports the process of modeling data mining, and users can use DMSE to deeply observe the internal relationship of data sets. Figure 2 shows the process of data mining by DMSE. In this paper, synthetic data sets and real data sets are mixed as experimental data sets.

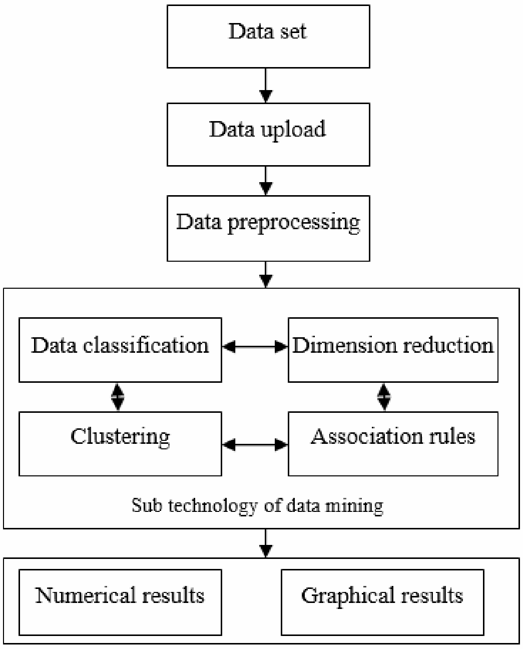


Figure 2. Flow chart of data mining by DMSE

The first data mining experiment uses the open toast cancer Wisconsin dataset [12] as the experimental data set, which includes 699 samples with 9 features. Each sample has a class label, and there are two classes: benign and malignant. The hidden information of the data set is obtained by data mining algorithm, and DMSE combines various data mining sub technologies to complete the experiment. The process of data mining experiment is divided into the following steps:

Step 1: data file upload: in order to use DMSE to analyze data, the data must be in compatible formats, such as tab, TXT, CSV, xlsx, ARFF.

Step 2: data preprocessing: this step includes data cleaning and data normalization. Data cleaning is divided into data format standardization, abnormal data clearing, error data correction, duplicate data elimination, etc. normalization processing zooms the eigenvalue to the same interval, which is to prepare for feature selection.

Step 3: data classification: DMSE integrates two classification algorithms: random decision forest (RDF) and multi-layer perceptron (MLP).

Step 4: data dimension reduction: DMSE integrates a visualization algorithm based on dimensionality reduction (smacof) [13].

Step 5: output graphical results: take two-dimensional scatter plot as the output form of data mining.

4 The Effect of Simulation Data Mining Based on DMSE

The experimental environment is PC, operating system is Ubuntu 10.04, CPU is Intel Core I5-650, main frequency is 3.2 GHz, memory capacity is 8 GB.

4.1 The Simulation Experiment of Data Dimension Reduction of DMSE

DMSE can simply realize the data mining model by establishing a scientific workflow. Figure 3 shows the operation interface of DMSE. The workflow of data mining can be established by dragging the icon with the mouse, and then the two-dimensional scatter diagram of data mining results can be output from the DMSE interface.

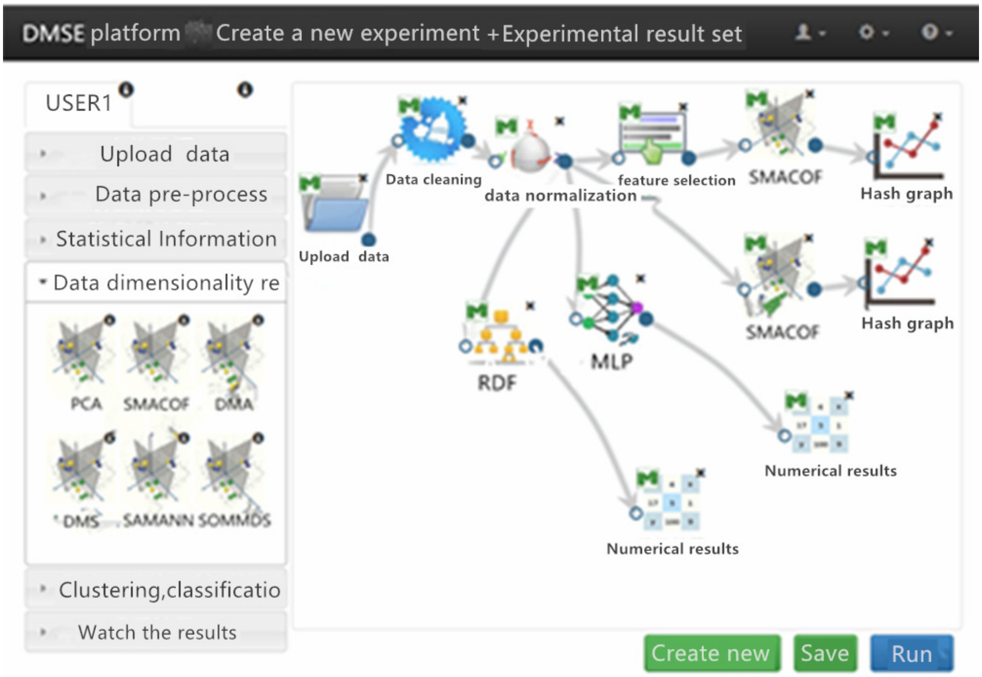
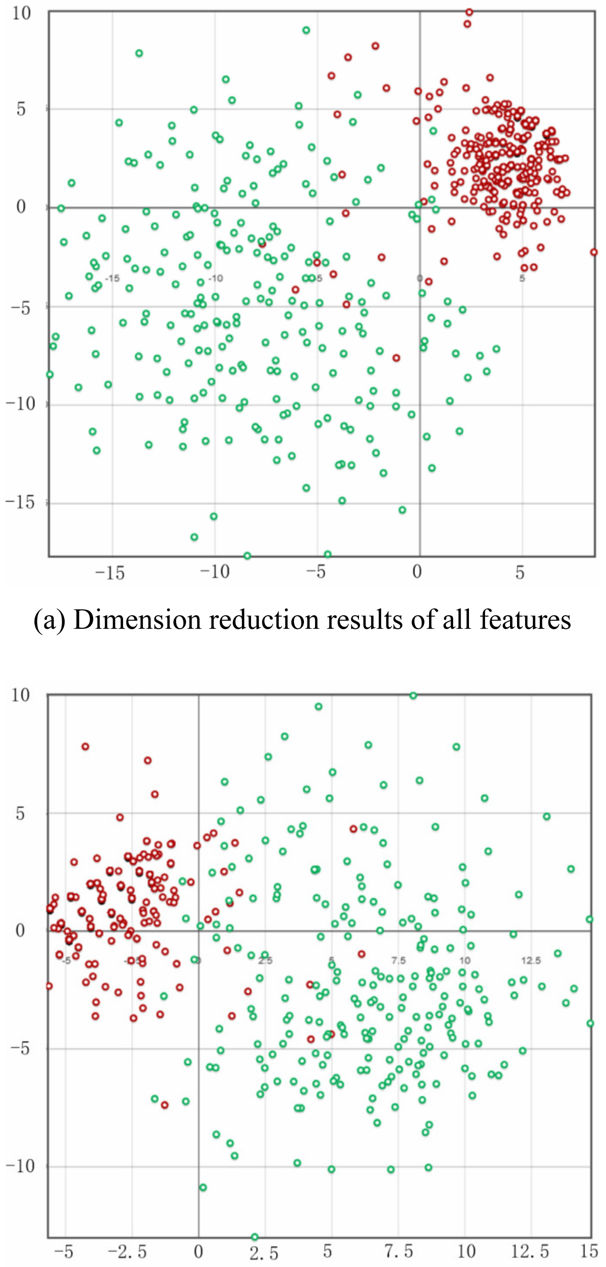


Figure 3. Operation interface of DMSEThe RDF and MLP data classification results of DMSE system are shown in Table 1, which shows that the data classification performance of DMSE is high

Table 1. classifier performance of different data mining software

Data mining software	Classifier	Average classification accuracy	recall rate	Specificity
Rrange [5]	RDF	95.65%	97.78%	91.67%
	MLP	97.10%	97.78%	95.83%
Weka [14]	RDF	98.55%	97.78%	100.00%
	MLP	95.65%	95.56%	95.83%
MS Azure ML [15]	RDF	95.59%	95.12%	96.30%
	MLP	97.01%	97.56%	96.30%
DMSE	RDF	97.10%	97.78%	95.83%
	MLP	97.10%	95.56%	100.00%

Figure 4 shows a two-dimensional hash graph after dimensionality reduction. The smacof algorithm [13] reduces the dataset to two-dimensional space. The red data points in the figure correspond to tumor data, and the red points are relatively concentrated. Other green points correspond to malignant tumor data, and the green points are extremely scattered. From Figure 4(a), we can observe the implicit relationship of the data set and improve the comprehensibility of the data set. Each breast cancer data sample is represented by nine features, but not all of them are key features. Therefore, feature selection and dimension reduction techniques are needed to filter features and retain a small number of features with strong resolution, which helps to reduce the negative impact of redundant features and accelerate the speed of data mining. Figure 4(b) shows a hash graph result of some features. Comparing the hash graphs of Figure 4(a) and Figure(b), it can be seen that feature selection processing has no obvious effect on the output hash graph, but it leads to the reduction of classification accuracy of data. Therefore, it can be concluded that all features of the data set are key feature.

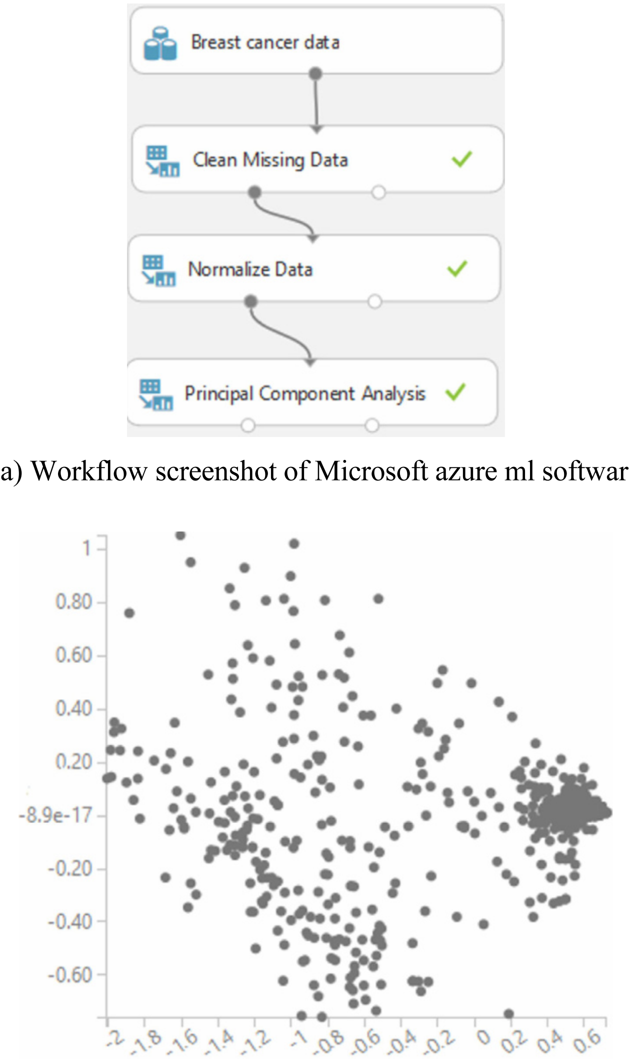


(a) Dimension reduction results of all features

(b) Dimension reduction results of some features

Figure 4. Two dimensional hash graph after dimension reduction

Microsoft azure machine learning studio [5] (Microsoft azure ml) is also a commercial software that can output data mining hash graph, which is also based on cloud computing. Comparing DMSE with Microsoft azure ml, Figure 5(a) shows the workflow of azure ml software for the toast cancer dataset, and Figure 5(b) shows the result of dimension reduction using PCA. Comparing Figure 4 with Figure 5(b), it can be seen that the output image of DMSE is color image, while the output image of azure ml software is black and white image, so the hash image of DMSE has better performance effect. In addition, DMSE integrates more dimensionality reduction algorithms than azure ml software, which proves the advantages of this visualization system.



(a) Workflow screenshot of Microsoft azure ml software

(b) Hash chart of Microsoft azure ml software output

Figure 5. Hash chart of output of Microsoft azure ml [15] software

4.2 Experiment of DMSE Cloud Computing Module

Data mining needs to consume a lot of computing and storage resources. The efficiency of processing large-scale data sets or high-dimensional data sets is low, which is not conducive to teaching. DMSE adopts cloud computing scheme to realize distributed computing and distributed storage, and adopts a high-dimensional large data set “ellipsodal data set” (http://personalpages.manchester.ac.uk/mbs/julia.handl /generators.html) For simulation, the dataset contains 10 overlapping elliptical clusters, which have 3140 50 dimensional data points. Figure 6 shows the topological map of the high-dimensional dataset. DMSE uses SMACOF to process the data set. Comparing DMSE with DMSE without cloud computing, DMSE based on cloud computing is referred to as “cloud DMSE”, and DMSE without cloud computing is referred to as “simplified DMSE”.

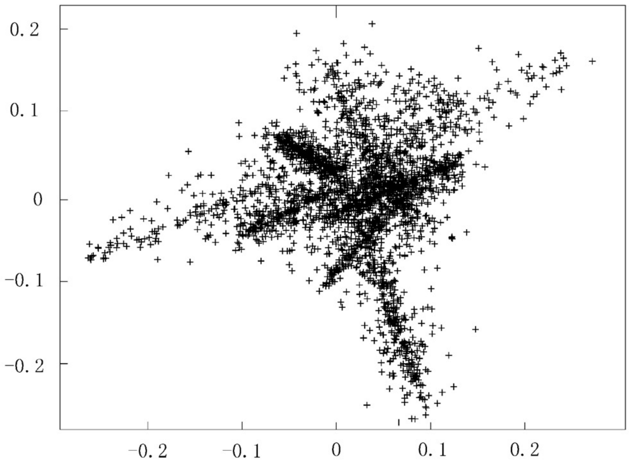


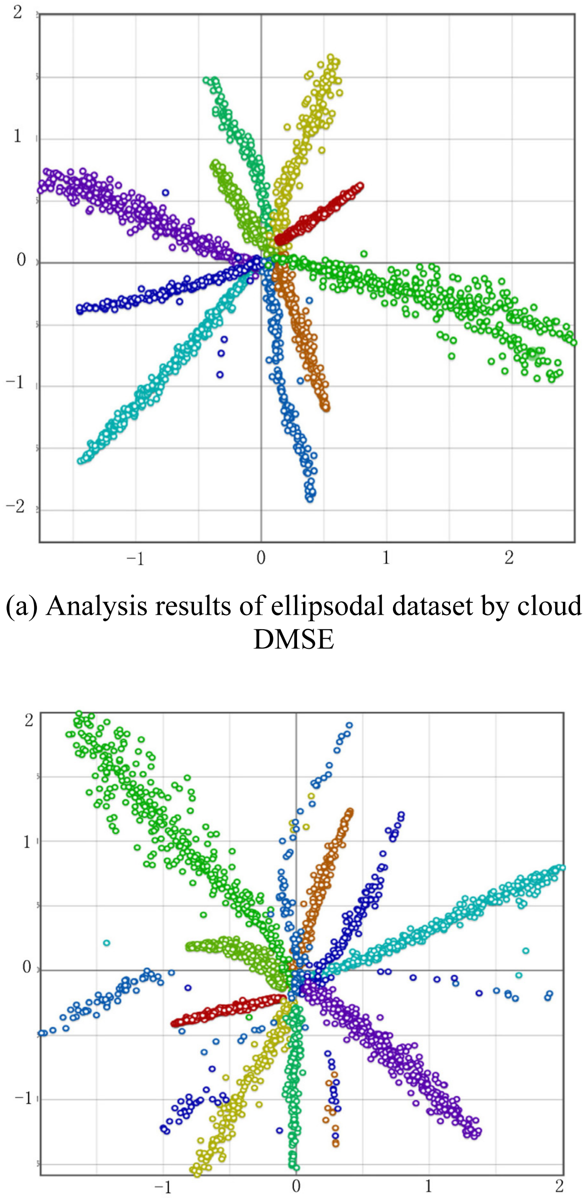
Figure 6. Topological structure of ellipsodal dataset

Figure 7 shows the analysis results of "cloud DMSE" and "simple version DMSE" on ellipsodal datasets respectively. It can be seen from the figure that the classification result of cloud DMSE is better than that of simple version DMSE. The reason is that the performance of smacof technology depends on the initialization value of low-dimensional data points. Cloud DMSE obtains a better initialization value through distributed computing, while the simplified version of DMSE fails to obtain a better initialization value. In addition, the running time of the simple version of DMSE is about 213 seconds, while that of cloud DMSE is about 9.7 seconds. Therefore, cloud computing technology not only improves the effect of data mining, but also improves the efficiency of data mining.

5 Summary

In order to improve the comprehensibility of data mining course, the experiment and teaching scheme for data mining technology is designed. This scheme includes data classification, data clustering, data dimension reduction and association rules and other data mining sub technologies. Taking sensor data, Internet data, social media data and other different types of data sets as experimental data sets, cloud computing technology is used to improve the efficiency of computing and storage, so as to improve the efficiency of teaching and experiment. The graphical interactive interface is designed, which can output the results of data classification, data clustering, data dimension reduction and association rules in the form of graph and table, so as to improve the comprehensibility of data mining technology. The final simulation results show that the cloud computing technology not only improves the visualization effect of data mining, but also improves the efficiency of data mining.

Due to the limitation of experimental conditions, the system only considers the most critical modules. With the development of big data technology, the current data security, high-dimensional small sample data, image data, video data are also the focus of teaching and scientific research. In the future, "hot plug" mechanism will be developed to integrate more data mining technology into the visual teaching system.



(a) Analysis results of ellipsodal dataset by cloud DMSE

(b) Analysis results of ellipsodal dataset by "simple version DMSE"

Figure 7. Analysis results of ellipsodal dataset by "cloud DMSE" and "simplified DMSE" respectively

References

[1] K. Jia, H. Li, Y. Yuan, Application of data mining in mobile health system based on Apriori algorithm, *Journal of Beijing University of technology*, Vol. 43, No. 3, pp. 394-401, March, 2017.

[2] Y. Cui, Z. Bao, Survey of association rule mining, *Application Research of Computers*, Vol. 33, No. 2, pp. 330-

334, February, 2016.

[3] H. Song, D. Wei, G. Tang, Y. Sun, Anomaly detection of Single User Behaviors Based on Pattern Mining, *Journal of Chinese Computer Systems*, Vol. 37, No. 2, pp. 221-226, February, 2016.

[4] H. Zhang, H. Lei, C. Gong, J. Peng, X. He, X. Ma, Quantitative analysis of virtually experimental teaching using data mining, *Research and exploration in Laboratory*, Vol. 36, No. 9, pp. 129-131, September, 2017.

[5] V. Podpečan, M. Zemenova, N. Lavrač, Orange4WS Environment for Service-Oriented Data Mining, *The Computer Journal*, Vol. 55, No. 1, pp. 82-98, January, 2012.

[6] I. Franciska, B. Swaminathan, Churn prediction analysis using various clustering algorithms in KNIME analytics platform, *International Conference on Sensing, Signal Processing and Security*, Chennai, India, 2017, pp. 166-170.

[7] Z. Li, J. Wang, Q. Song, SOAP message part information encryption mechanism based on WSE, *Computer engineering and design*, Vol. 37, No. 1, pp. 55-59, January, 2016.

[8] H. Zhang, J. F. Cai, L. Cheng, J. Zhu, Strongly Convex Programming for Exact Matrix Completion and Robust Principal Component Analysis, *Inverse Problems & Imaging*, Vol. 6, No. 2, pp. 357-372, May, 2012.

[9] L. Wei, S. Wang, F. Xu, Outliers Mining via Weighted Multidimensionality Scaling, *Computer science*, Vol. 35, No. 1, pp. 190-192, January, 2008.

[10] J. Bernataviciene, G. Dzemyda, V. Marcinkevicius, Conditions for Optimal Efficiency of Relative MDS, *Informatica*, Vol. 18, No. 2, pp. 187-202, January, 2007.

[11] M. J. Pastizzo, R. F. Erbacher, L. B. Feldman, Multidimensional Data Visualization, *Behavior Research Methods, Instruments, & Computers*, Vol. 34, No. 2, pp. 158-162, May, 2002.

[12] F. Ahmad, N. A. M. Isa, Z. Hussain, M. K. Osman, S. N. Sulaiman, A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer, *Pattern Analysis & Applications*, Vol. 18, No. 4, pp. 861-870, November, 2015.

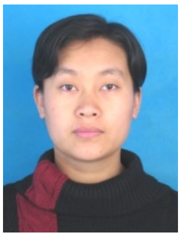
[13] J. Bernatavičienė, G. Dzemyda, V. Marcinkevičius, Diagonal majorization algorithm: Properties and efficiency, *Information Technology & Control*, Vol. 36, No. 4, pp. 353-358, December, 2007.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA data mining software: an update, *Acm Sigkdd Explorations Newsletter*, Vol.11, No.1, pp. 10-18, June, 2009.

[15] M. Andrejko, A. Zdybicka-Barabas, M. Janczarek, M. Cytrynska, Three Pseudomonas aeruginosa strains with different protease profiles, *Acta Biochimica Polonica*, Vol. 60, No. 1, pp. 83-90, 2013.

Biographies

Hong Shi obtained the master's degree from Taiyuan Institute of machinery in 2002, and she has been an information system project manager in 2015. Her research interest includes big data intelligent analysis, computer network education, image processing.



Hongxia Deng obtained the Ph.D. degree in computer application technology from Taiyuan University of Technology in 2013. And the same time she has been an associate professor. Her research interest includes computer vision, image processing, big data intelligent analysis, brain science.