International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.231-240
http://dx.doi.org/10.14257/ijdta.2016.9.5.24

Research on Parallel Computing Model and Classification Algorithm Based on Data Mining Process

Qiongshuai Lv[1,*] and Haifeng Hu[2]

[1]*School of the Software Engineering, Pingdingshan University, Pingdingshan 467000, Henan, China*
[2]*School of Computer Science and Technology, Pingdingshan University, Pingdingshan 467000, Henan, China*
E-mail: qiongshuailv@163.com

Abstract

In the big data era, with the parallel evolution of computer architecture, computing changes and modifications of industrial application mode resource expansion capability, we need to explore a new parallel computing model, to reflect the properties and large data applications form the current parallel machines, and a variety of mainstream big data processing system for unified theoretical analysis to guide large data applications tuning. Currently, despite the large data programming model study made many achievements, and is widely used in the TB level or even PB-class data processing and analysis, but the corresponding computational model study has just begun. From traditional parallel computing model, research big data programming model and large data computation model, summed up the three basic problems of large data model, in theory, need to be addressed: the three elements of the problem model, scalability and fault tolerance issues and performance optimization. Around these three questions, on the one hand and performance optimization model to calculate the theoretical study of data from a large, on the other hand these performance optimization methods in case of an actual big data.

Keywords*: Data mining; knowledge patterns; performance optimization, duplicate data, multi-core technology*

1. Introduction

With the advent of the era of big data, from a parallel machine architecture, scalable computing resources to industry application mode change significantly in the event [1-2]. Wherein between hardware and software architecture of parallel computing model is one of the guiding big data application tuning, and promote the development of the core technology of large data [3-5]. Currently industry has research and development of a variety of large data programming model, and is widely used in data processing and analysis TB level or even PB-class, and academics are trying to explore more abstract and big data computing model to reveal large data mission computing, communications and access essential characteristics of deposit behavior [6].

Faced with such a wealth of mass data, traditional data processing methods and capabilities already cannot meet the actual demand. Faced with increasingly fierce competition in the market, people can help leaders need to extract from these decisions contains a wealth of information in the data to make decisions knowledge in strong demand driven, the data mining technologies have emerged, data mining is a series of integrated application of advanced technology to extract from a large database or data warehouse of information that people are interested in, and knowledge, they are implicit, previously unknown and potentially useful concepts, rules, laws, patterns and so on [7-

ISSN: 2005-4270 IJDTA

10]. These studies include the basic process of data mining to explore the system the main feature should have their interconnections; different types of data source data mining system feature requirements; target different applications of the data mining system feature requirements; and other data mining system implementation mechanism main feature [11-12].

In this paper, data mining classification problem is especially JEP classification algorithm based on in-depth research, we propose a special JEP, it gives an efficient algorithm for mining SJEP, construct a new classification algorithm based SJEP the algorithm can be used to obtain a very small SJEP JEP-based classification algorithm similar to or even higher accuracy, in most cases, the algorithm has a higher ratio of CAB and C4.5 classification accuracy, and improve the efficiency of the algorithm, mining tasks can be completed in less time.

2. Research Status and Related Theory

2.1 Data Mining Concepts

Data mining is essentially a new business information processing, data mining technology to improve people's application data from the low-level query to the online decision support analysis and forecasting and other more advanced applications, it is through these data micro, meso and the macro statistics, analysis, synthesis and reasoning, we found correlation between the data [13]. Future trends and general overview of knowledge, knowledge of this information can be used to guide senior business activity.

According mining methods can be divided into: machine learning methods, statistical methods, cluster analysis, exploratory analysis, neural network, genetic algorithms, database methods, approximate reasoning and uncertainty reasoning method, based on evidence theory and meta-model approach modern mathematical analysis methods. The basic process of data mining can be divided into the following main stages: Cleaning and integrating data and projection data selection, data mining and visualization and evaluation of results presented in Figure 1:

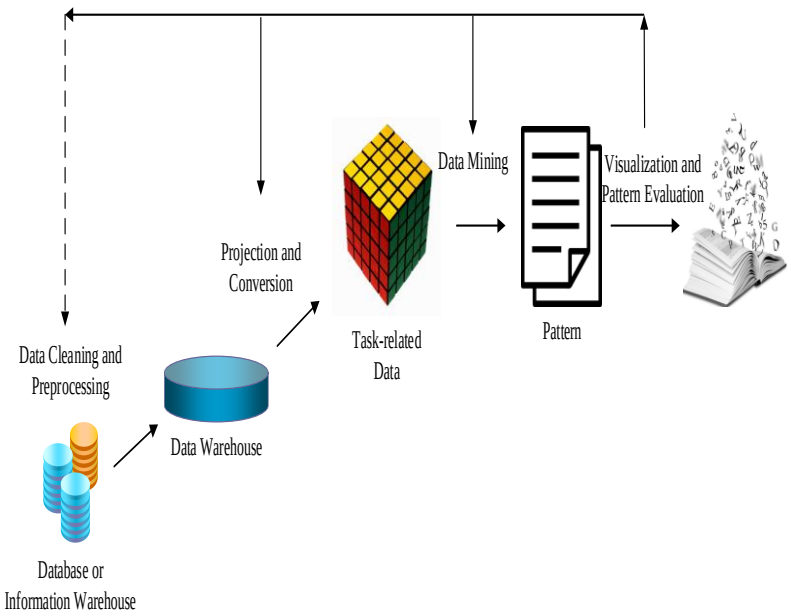

Figure 1. The Basic Process of Data Mining

In principle, the data mining should be applied to any information stored knowledge mining, but mining and technology challenges because of the different types of source

data stored in diversity. Especially in recent years studies have shown that data mining data storage types involved more and more rich, in addition to some common value model, architecture and other studies, also carried out some research techniques or algorithms for mining complex or novel data storage under.

2.2 JP Classifier Classification Algorithms

JEP these boundaries may represent a collection of very large item sets. However, when dividing a test case category, only those items set with a large degree of support has a significant impact on the final collective effect. Therefore, JEP-Classifier used only algorithm to classify the most expressive, not only greatly reduces the complexity of the algorithm, but also enhances endurance noise algorithm. Classification process JEP Classifier classification algorithm is as follows:

a. Using HORIZON MINER algorithm to find a horizontal boundary of each class;
b. True algorithm by repeatedly calling for a long time or borders based MBD_LLBORDER algorithm to dig out all of the JEP;
c. The selection of the most expressive JEP;
d. Calculate support collective effect of each class;
e. The test case is divided into collective effect of the highest scoring class.

For example, the data class contains three training data set, JEP_Classifier classification algorithms work flow shown in Figure 2.

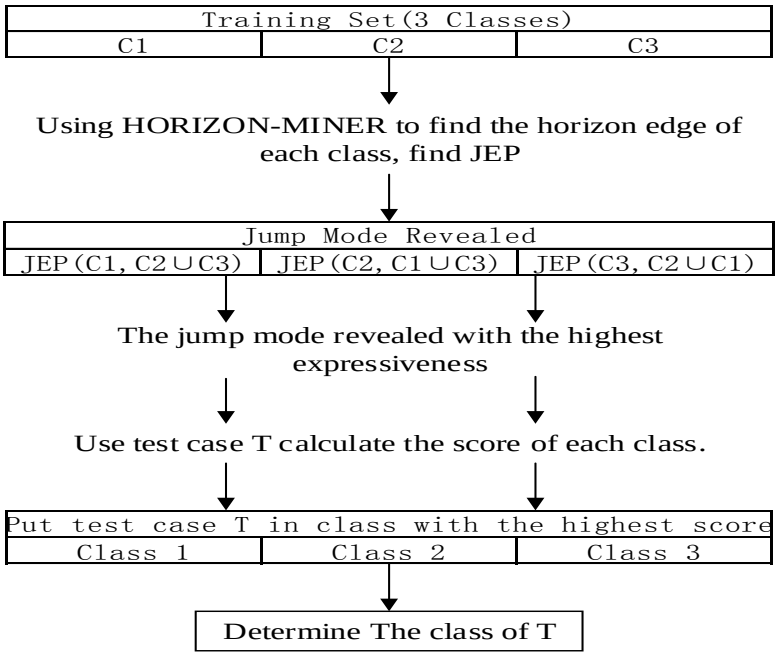

Figure 2. JEP_Classifier Classification Algorithms Work Flow

Based JEP classification algorithm JEP as the basis of the classification, with boundaries represented JEP, with algorithm-based border to dig JEP, and according to the collective effect of these JEP to classify which JEP-Classifier classification algorithm uses only the most expressive JEP classification, reducing the complexity of the algorithm, enhanced noise endurance. experimental results show that compared with some of the published classification algorithm, the class classification algorithm has higher prediction accuracy, and that the class classification algorithm can be applied to large databases.

2.3 Distributed SVM Implementations Trainer

For a statics collection containing n statics couples $\{(x_i, y_i)\}_{i=1}^{n}$, note that sample number division on m checkpoints is $\{B_1, ..., B_m\}$,then the sorting study problem can be written as the form:

$$\min_{w_1,...,w_m,z} \frac{1}{2}\|z\|_2^2 + C\sum_{j=1}^{m}\sum_{i\in B_j} \max(1-y_i x_i, 0)^2 + \sum_{i=1}^{n}\rho\|w_j - z\|^2 \tag{1}$$

Subject to $w_j - z = 0,\ j = 1, ..., m$

Which ρ is the steady step length, w_j is the weight vector of sub static collection x_{B_j} , and z is the regularization vector, $\sum_{i=1}^{n}\rho\|w_j - z\|^2$ is for strengthen the convergence.

The Lagrange transformation of problem can be written as the following form:

$$L(w, z, \lambda) = \frac{1}{2}\|z\|_2^2 + C\sum_{j=1}^{m}\sum_{i\in B_j} \max(1-y_i x_i, 0)^2 + \sum_{i=1}^{n}\left(\rho\|w_j - z\|^2 + \lambda_j\left(w_j - z\right)\right) \tag{2}$$

Which λ is a dual variable. $\lambda_j = \rho u_j$, and the (k+1) iteration can be written as the following form:

$$w_j^{k+1} = \arg\min_z C\sum_{i\in B_j} \max(1-y_i x_i, 0)^2 + \rho\|w_j - z^k + u_j^k\|^2 \tag{3}$$

$$z^{k+1} = \frac{\sum_{j=1}^{m} w_j + u_j^k}{m + 1/\rho} \tag{4}$$

$$u_j^{k+1} = u_j^k + w_j^{k+1} + z^{k+1} \tag{5}$$

A good model consists of three elements: the machine parameters (such as CPU, Memory, I/O network, node size), execution behavior (including computing behavior, communication behavior and I/O behavior), the cost function (including time cost function and space cost function, which is a function of machine parameters).

3. Big Data Model

3.1 Dryad Model

Dryad distributed programming model is a coarse-grained model for data from the Microsoft Research presented parallel applications. This model is a data flow diagram by the computing and communications vertex pipelines constituted a number of vertices in the graph is a simple, non-threaded and locking mechanism is defined by the application developer's serial program, and piping supports multiple implementations, including file transfer, TCP and Shared-memory FIFOs strategy.

In Dryad model, each task is not a table DAG, according to the dependency graph of the vertex dispatched to perform multiple or multi-core parallel processing to achieve. Figure 3 was its task structure. And ideas MapReduce model is similar, Dryad model also calculates the vertex is moved to the data storage node corresponding or close, to relieve the pressure M network transmission; also have fault tolerance mechanisms to ensure that the program can stretch run in different sizes cluster. But compared to the two-stage model of MapReduce, Dryad model can be expressed richer type of calculation; also supports multiple pipelines can avoid some unnecessary disk input and output, to

accelerate the implementation of the calculation.

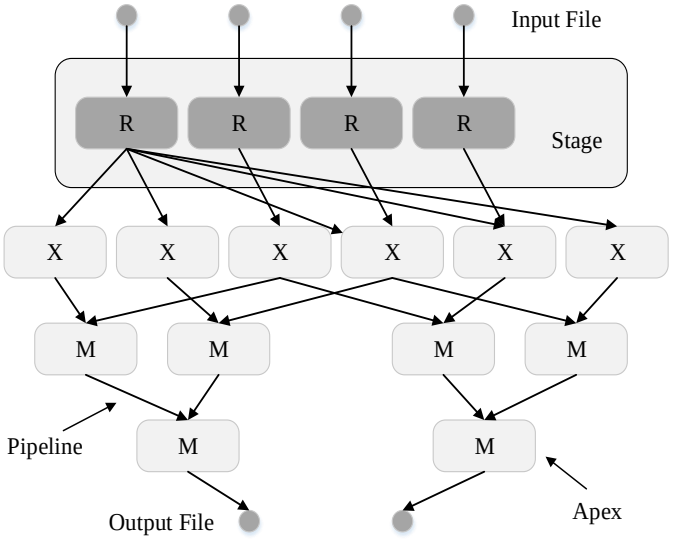

Figure 3. Dryad Task Structure Model

3.2 Big Data Calculation Model

Existing large data programming model is to improve the computing resources by supporting scale performance, and comes with fault tolerance mechanisms to cope with node failures, but they are to emphasize their scalability and fault tolerance through experiments or case, the lack of theoretical interpretation and metrics. Therefore, the model needs to calculate the angle from big data, given these two characteristics can be strictly defined and universally accepted criteria to judge them.

The current performance optimization of large data applications are mostly based on a specific programming model or framework, lack of a unified optimization theory. Requiring large data computing model based on structure optimization method is suitable for different applications and architectures. DOT model includes basic blocks, combination blocks and DOT expressions. A basic block is composed of three levels DOT, shown in Figure 4:

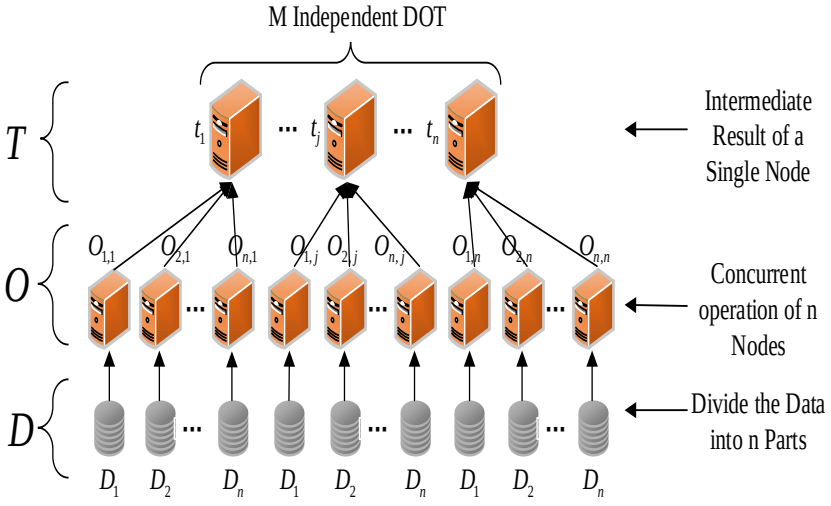

Figure 4. A Basic Block is Composed of Three Levels DOT

$$\vec{D}OT = [D_1 \quad \cdots \quad D_n] \begin{bmatrix} o_1 \\ \vdots \\ o_n \end{bmatrix} [t] = \left[\bigcup_{i=1}^{n} (o_i (D_i)) \right] [t] \tag{6}$$

Its formal description is:

$$\vec{D}OT = [D_1 \quad \cdots \quad D_n] \begin{bmatrix} o_{1,1} & o_{1,2} & \cdots & o_{1,m} \\ o_{2,1} & o_{2,2} & \cdots & o_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ o_{n,1} & o_{n,2} & \cdots & o_{n,m} \end{bmatrix} \begin{bmatrix} t_1 & 0 & \cdots & 0 \\ 0 & t_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_m \end{bmatrix}$$

$$= \left[\bigcup_{i=1}^{n} (o_{i,1} (D_i)) \quad \cdots \quad \bigcup_{i=1}^{n} (o_{i,m} (D_i)) \right] \begin{bmatrix} t_1 & 0 & \cdots & 0 \\ 0 & t_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & t_m \end{bmatrix} \tag{7}$$

3.3 Overview of the System as a Whole

H-DB draws Hadoop DB design ideas, Figure 5 shows its system architecture. As can be seen from the figure, H-DB is divided into four sections, each of which is the bottom station, a database of regional and national centers, this part of the memory of the original observation data sharing and job responsibilities function remains change; the middle layer is HDFS, not only to store metadata and query results, and increased global index information data caching layer for storing all the data dictionary tables and result buffer, and global index layer is used to store all data tables, indexes catalog and index buffer; topmost MapReduce programming model is provided only for HDFS data in parallel processing, including generating global index concurrent access to data dictionary tables and tables of information, it also provides fault tolerant protection; middleware is including database connectivity, data loader, the index generator and query engine four functional modules.

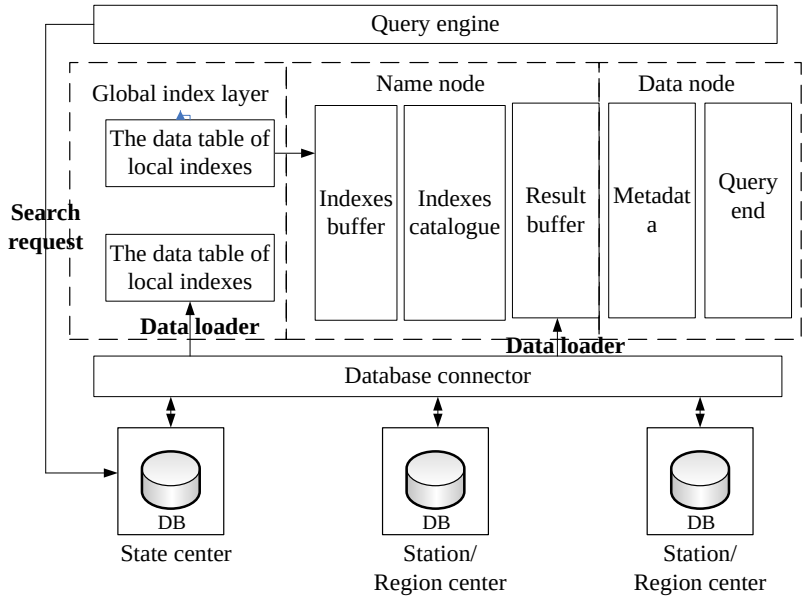

Figure 5. HDB Middleware Functions

4. Experiment and Analysis

4.1 Time Costs Related Experiments and Analysis Functions

Wordcount and terasort for two procedures, test different input data size corresponding optimal number of machines w n *, in order to verify the correctness of the time cost function P-DOT model. Note that when the former test four data sets, in order to avoid inter-process I/O memory access conflicts, only one running on each machine process; but the first five test data set, due to the limited number of machines MPI clusters, two processes run on each machine. Table 1 shows the different input data W and the corresponding optimum scale machine number n*.

Table 1. Input Data Size *W* vs. Optimal Machine Numbers η*

w / η*	0.25GB	6.25 GB	25GB	625GB	2.5TB
wordcoun(mpi)	60	164	600	2400	4800
terasort(mpi)	60	300	840	2600	5800
wordcoun(hadoop)	36	120	360	2000	/
terasort(hadoop)	36	60	280	1000	/

Figure 6 shows the use of MPI and Hadoop programming model wordcount and terasort program run time on a different data set, the number of processes in which the horizontal axis is the actual work nodes involved in the work. As can be seen from the figure, all curves are opening down and there is the lowest point, indicating a large data given task, given the scale of environmental load and input data VV, there is an optimal number of machine n *, and excessive or too little of a few machines are not effective use of computing resources.

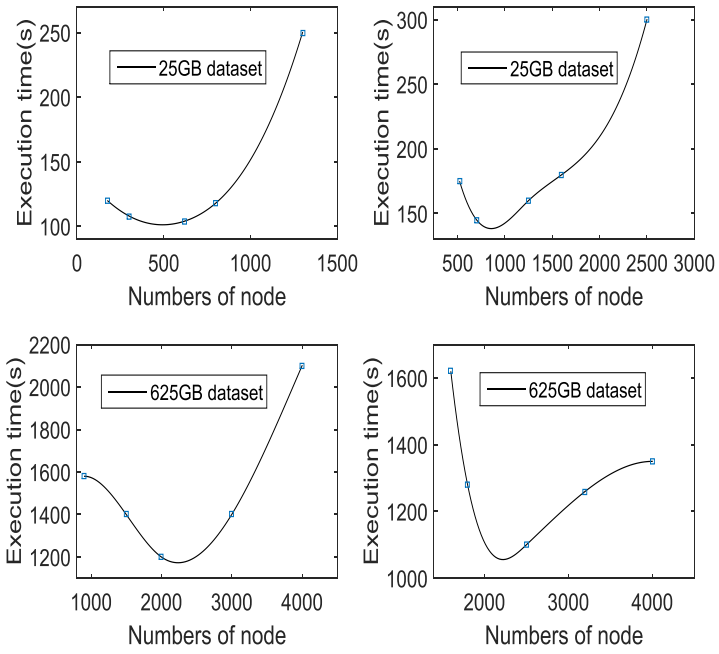

Figure 6. MPI Tests Wordcount and Terasort Run Time on Different Datasets

P-DOT model only choose the size of the input data W and machine number n as the main two parameters, so that the model is not perfect in terms of accuracy; there is interference from other tasks on the test platform, resulting in communication overhead measurement errors exist.

4.2 Accuracy Comparison of Classification Algorithm

We scale classification algorithm JPClassifier also been experimental analysis, the size of the test data set sizes affect the classification algorithm. We focus selected data from the original 25%, 50%, 75% and 90% of instances as our new data set. Algorithm is applied to the four datasets found EJP border. The results show the time to run out of linear time dependence, as Figure 7 (a) and 7 (b) below.

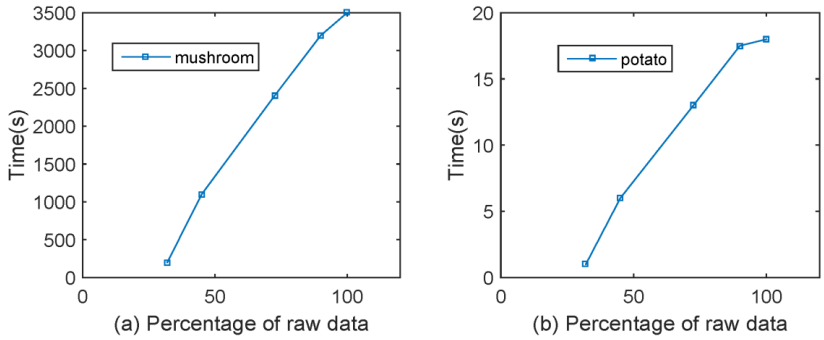

Figure 7. Scale Classification Algorithm Metrics

Data mining is a result of unpredictable work, it is difficult in advance to all the problems are designed, therefore, needs to continue to verify and modify the error, even if some knowledge is correct, it may not be of interest to us. The accuracy of mining results, not only in its credibility, and depending on whether it is useful for us, the use of restraint can help us identify the problem and make timely adjustments, so that each stage of knowledge discovery according to the right direction development. As can be seen from the experimental results, in the training phase, we can safely deleting items that support is lower than the minimum support closed set of pre-specified value, without affecting the accuracy of the classification.

5. Conclusions

Data mining core issues to be solved is how the mass of information into decision-making knowledge. At present, all kinds of enterprises, the role of data mining has attracted wide attention, as in the communications field, the field of insurance, finance and other fields. JEP-based classification algorithm is a new database for large classification algorithm proposed in recent years, and the experiment has indicated that some of the previous classification algorithm classification algorithm has higher classification accuracy. Based on the classification algorithm, inherits the advantages of the original classification algorithms, modify some of these deficiencies, the establishment of a more accurate, more efficient classification algorithm. JEP is a new knowledge model, it has a strong ability to distinguish, will be applied to classify a new direction, our job is to further improve the algorithm, so that gradually being used by the application program.

References

[1] B. P. Kelley, C. Klochko and S. Halabi, "Datafish Multiphase Data Mining Technique to Match Multiple Mutually Inclusive Independent Variables in Large PACS Databases", Journal of digital imaging, **(2015)**, pp. 1-6.
[2] M. W. King and P. A. Resick, "Data mining in psychological treatment research: A primer on classification and regression trees", Journal of consulting and clinical psychology, vol. 82, no. 5, **(2014)**, pp. 895.
[3] H. Lou, Y. Ma and F. Zhang, "Data mining for privacy preserving association rules based on improved MASK algorithm[C]//Computer Supported Cooperative Work in Design (CSCWD)", Proceedings of the 2014 IEEE 18th International Conference on. IEEE, **(2014)**, pp. 265-270.
[4] C. Jian, W. Zhang and Y. Ying, "A micro-gesture recognition on the mobile web client", Review of Computer Engineering Studies, vol. 2, no. 2, **(2015)**, pp. 19-24.
[5] T. J. Hui and Z. Quan, "Design and implementation of the crying voice detection circuit in the baby's supervision system", Review of Computer Engineering Studies, vol. 1, no. 1, **(2014)**, pp. 13-16.
[6] N. Singhal and M. Ashraf, "Performance enhancement of classification scheme in data mining using hybrid algorithm[C]//Computing, Communication & Automation (ICCCA)", 2015 International Conference on. IEEE, **(2015)**, pp. 138-141.
[7] R. Vijayakrishnan, S. R. Steinhubl, K. Ng, "Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record", Journal of cardiac failure, vol. 20, no. 7, **(2014)**, pp. 459-464.
[8] S. Bandyopadhyay, U. Maulik and C. Coello, "Guest Editorial: special issue on advances in multi-objective evolutionary algorithms for data mining", Evolutionary Computation, IEEE Transactions on, vol. 18, no. 1, **(2014)**, pp. 1-3.
[9] A. Tayyebi and B. C. Pijanowski, "Modeling multiple land use changes using ANN, CART and MARS: Comparing tradeoffs in goodness of fit and explanatory power of data mining tools", International Journal of Applied Earth Observation and Geo-information, vol. 28, pp. 102-116, **(2014)**.
[10] M. Hoogendoorn, L. M. G. Moons and M. E. Numans, "Utilizing Data Mining for Predictive Modeling of Colorectal Cancer Using Electronic Medical Records", Brain Informatics and Health. Springer International Publishing, **(2014)**, pp. 132-141.
[11] E. Roelofs, A. Dekker and E. Meldolesi, "International data-sharing for radiotherapy research: an open-source based infrastructure for multi-centric clinical data mining", Radiotherapy and Oncology, vol. 110, no. 2, **(2014)**, pp. 370-374.
[12] Z. Dauter, A. Wlodawer and W. Minor, "Avoidable errors in deposited macromolecular structures: an impediment to efficient data mining", IUCrJ, vol. 1, no. 3, **(2014)**, pp. 179-193.
[13] B. Taati, J. Snoek and D. Aleman, "Data mining in bone marrow transplant records to identify patients with high odds of survival", IEEE Journal of Biomedical and Health Information, vol. 18, no. 1, **(2014)**, pp. 21-27.

Authors

Qiongshuai Lv, He received his M.Sc. in Computer software and theory (2011) from Zhengzhou University. Now he is a lecturer at School of the Software Engineering in Pingdinshan University. He participated in the completion of a number of provincial scientific and technological projects, published a number of academic papers. His main research interests include machine learning and data mining.

Haifeng Hu, He received his M.Sc. in Engineering in computer technology (2010) from Xidian University. Now He is director of the laboratory of information engineering in Pingdingshan University. He was awarded the Second prize of Pingdingshan science and technology progress. His current research interests include data mining and network security.

