International Journal of Database Theory and Application
Vol.7, No.2 (2014), pp.99-108
http://dx.doi.org/10.14257/ijdta.2014.7.2.10

A Construction Method of Gene Expression Data Based on Information Gain and Extreme Learning Machine Classifier on Cloud Platform

Wei Sha-Sha[1], Lu Hui-Juan[2*], Jin Wei [3] and Li Chao[4]

[1,2,3,4]*College of Information Engineering, China JiLiang University*
[1]*weishasha5210@163.com,* [2]*hjlu@cjlu.edu.cn,* [3]*jinweicjlu@gmail.com,*
[4]*61322smile@sina.com*

Abstract

With the large-scale application of high dimensional gene expression data which exists lots of redundant information, it may waste a lot of time in feature selection and classification. By analyzing the process of MapReduce computing paradigms on cloud platform, it is found that the feature selection which through parallel and distributed computing in MapReduce combined with extreme learning machine is appropriate for constructing a recognition method. This paper proposed a MapReduce algorithm on high gene feature for parallel and distributed selection and classification, aiming to save time resources to make a higher accuracy in training process on large scale gene datasets. Simulation experiments on gene datasets show that the running time on cloud platform is greatly shortened by the time promising the high classification accuracy.

Keywords: *Classification; MapReduce; Cloud Platform; Feature Selection*

1. Introduction

With the development of the Internet, information is showing a trend of unlimited growth. Personal computers' ability to store and process data on becomes powerless in front of such a large data. How to find a low-cost, safe and fast way to solve the problem of unlimited storage and computing growth of Internet information becomes a proposition in front of scientists. The cloud computing finds a reasonable solution to these problems [1].

Emergence of cloud computing is the result of parallel technology, software technology and network technology development. It is a new computing model [2] which uses data, applications and IT resources through the network as a service and then provide to the users. Cloud computing can be regarded as an infrastructure management methods, which uses virtualization technology to manage resources together to form a large capacity of resource pool. Users can make request through the network to the cloud center and access to services. They can dynamically deploy, configure, reconfigure, and cancel the service or other operations on the resource pool. The definition of cloud computing from China cloud computing network: cloud computing is a development of distributed computing, parallel

* This work was supported by the National Natural Science Foundation of China (No. 61272315, No.60842009, and No. 60905034), Zhejiang Provincial Natural Science Foundation (No.Y1110342, No.Y1080950), Zhejiang Science and Technology Department of International Cooperation of special funded projects (2012C24030) and the Pao Yu-Kong and Pao Zhao-Long Scholarship for Chinese Students Studying Abroad.

Corresponding author: Hui-juan Lu, E-mail addresses: hjlu@cjlu.edu.cn.

ISSN: 2005-4270 IJDTA

computing and grid computing, or that the business achievement of those scientific concept [3]. The basic principle is that the cloud computing parallel distributed to a large distributed computer [4].

The advent of cloud computing technology makes computing become a resource. The computing capabilities of cloud computing are often likened to the power by IT professionals. As long as you can plug into the network, it will be able to "socket" up using this new "energy". We can see what a significant change "cloud" brings to computation [5].

Hadoop is an open source distributed computing framework proposed by the Apache Software Foundation open source organizations in 2005 as a part of Lucene's subproject Nutch.It is not an acronym, but a fictitious name. MapReduce and Hadoop core is a distributed file system (Hadoop Distributed File System, HDFS) and later joined the HBase. MapReduce is widely used in data mining [6], bioinformatics [7] as well as other types of data-intensive applications [8] and other fields. MapReduce is simple to use, with good error tolerance and automatic load balancing, *etc.*, which makes cloud computing platform to build effective classifiers possible.

In recent years, the application of large-scale microarray gene expression technology in cancer diagnosis data provides a new way. [9] Due to the high dimensional gene expression data, and the classification of genes that play an important role diagnosis is usually no more than a few hundred. By feature selection, select genes closely associated with the classification, one can effectively improve the classification accuracy, while reducing the cost of post-biological analysis. In gene expression data classification, Alon *et al.*, [10] in 1999 with colon cancer (Colon Cancer) data sets as experimental subjects, the data set of colon cancer classification using hierarchical clustering algorithms for classification, 2000, Rayc *et al.*, [11] using principal component analysis of yeast spore germination dataset experiments were carried out to obtain classified information in the dataset spore germination process, Khan et al [12] and Narayanan *et al.*, [13] The application of artificial neural networks to the known sample obtained classification model. Ramaswam *et al.*, use SVM studied 14 different types of tumor tissue classification problems, and achieved good classification performance [14]. Lu *et al.*, [15] use compressed sensing techniques to realize classification of gene expression data. In addition, common genetic supervised classification algorithms include KNN, artificial neural networks, decision trees, and so on.

In 2006, GuangBin Huang [16] *et al.*, according to Moore - Penrose generalized inverse matrix theory, presented a new oversight single hidden layer feed-forward neural network learning algorithm, namely Extreme Learning Machine (ELM). Compared with neural networks and support vector machines, ELM learns faster, with better generalization ability in classification applications.

In classification algorithm area, research scholars have devoted most of their time to improve the efficiency of the classification algorithm based on classification algorithm itself. Although to some extent this can be very good to improve the accuracy of classification algorithms, but the speed has not been classified a substantial upgrade. In this article, we use the environment to solve large-scale cloud platform gene expression data classifier build issues in order to improve the efficiency of building classifiers, by taking advantage of the cloud platform distributed parallel computing.

Information gain is an important concept in information theory. It has been widely applied in the field of machine learning as well as in specialty choice. For gene expression data, a classification system, the information gain is calculated for each gene, a gene of a statistical amount of information provided in the classification system to determine the classification system for the gene of importance. Information Gain method can quickly rule out a large

number of non-critical noise and irrelevant genes, refine search area of the optimal subset of genes.

Based on information gain and extreme learning machine, we proposed a filter type gene expression data classification method for cloud platforms. First, the data sets are randomly divided into several parts by a random function, based on information gain between genes and screening for genetic grouping. Second, these filtered subsets of genes utilize MapReduce parallel programming model and ELM to construct classifiers, and finally the results are returned to the client side. Our lab uses five PC to build a Hadoop cloud computing platform, with four sets of PC deployment DataNode and TaskTracker. Results show that in similar general classification accuracy, the filter-type gene expression data classification information gain and extreme learning machine on the cloud platform is faster and more efficient than the case of PC.

2. Gene Filters Based on Information Gain

Some genes are expressed only under certain experimental conditions. In order to reduce the time and space complexity of machine learning, feature selections are needed: those closely related are chosen and classified, while the classification accuracy can also be improved. Feature selection is based on the importance of the various features, characteristics after removing redundant unrelated features, picking out the classification of certain significant features to reduce the dimension of the feature space.

2.1. Information Entropy and Information Gain

Entropy is a very important concept in information theory, which represents uniformity of the distribution of any kind of energy in space. Energy distribution more uniform, more uncertainty, the greater the entropy [17]. Shannon [18] used the concept of the entropy in information processing, and proposed the concept of 'information entropy'. Entropy is a measure to quantify the information, is a measure of the degree of uncertainty of a random variable. In the information gain, the measure of the importance of the feature is to see how much information can be classified as to bring the more information, the more important features.

Order For different values random variables, corresponding to different probability, then the entropy of X is defined as:

$$H(X) = -\sum_{i} P(x_i)\log_2(P(x_i)) \tag{3-1}$$

As we can see, more changes of random variables, greater information obtained through them.

For the classification system, Class is variable, so the entropy of the classification system can be defined as

$$H(C) = -\sum_{i=1}^{l} P(c_i)\log_2(P(c_i)) \tag{3-2}$$

Here is the categories number of the classification system. In particular, for two classification problems, information entropy can be expressed as

$$H(C) = -P(c_1)\log_2(P(c_1)) - P(c_2)\log_2(P(c_2)) \tag{3-3}$$

In the classification system of gene expression data, the information gain terms for each gene, for a gene X, It may have n possible values $(x_1, x_2, \cdots x_n)$ Corresponding conditional entropy is

$$H(C \mid X) = -\sum_{j=1}^{n} P(x_j) \sum_{i=1}^{l} P(c_i \mid x_j) \log_2 (P(c_i \mid x_j)) \tag{3-4}$$

$P(c_i)$ Represents priori probability of the categorical variables C , $P(c_i \mid x_j)$ represents the conditional probability of variables C after gene X is fixed.

Thus, the information gain gene X bring to the classification system can be expressed as the difference between the original system information entropy and the conditional entropy after gene X is fixed.

$$\mathrm{IG}(\mathrm{X}) = H(C) - H(C \mid X) \tag{3-5}$$

If gene X and the Category C are not relevant $\mathrm{IG}(\mathrm{X}) = H(C) - H(C \mid X) = 0$; If relevant $H(C) > H(C \mid X)$, *i.e.*, $\mathrm{IG}(\mathrm{X}) = H(C) - H(C \mid X) > 0$ .While the larger the difference is, the stronger the correlation between X and C [100]. Therefore, the differential entropy defined information gain, represents the amount of information obtained after the elimination of uncertainty. Clearly, greater information gain value a feature item has, the larger contribution it makes, the more important for the classification. Therefore, when choosing genes, usually choose genes with great information gain to represent the original high-dimensional gene first, and use them as a basis for further gene selection.

2.2. Information Gain Process

Information gain algorithm flow can be described as follows:

Input: original gene sets S;

Output: After the information gain algorithm selection gene subset FS.

1) calculated for each category of known samples probability;

2) Calculate the entropy of the classification system according to the probability using the formula (3-2) 1) obtained;

3) For each gene, calculate the probability of all of its values Calculate conditional probabilities;

4) According to the probability 3) obtained using the formula (3-4) for each gene calculate conditional entropy;

5) Using (3-5) the information gain is calculated for all genes;

6) Sort the results obtained in 5) chooses the former K maximum gain of genetic information as a compact subset of genes FS (common value of K is 200-400).

3. Classification Model Built Based on Cloud Computing Platform

Classifiers built on the cloud computing platform, first use information gain feature selection, and then use MapReduce to feature selection model building, and finally Reduce back to Job Tracker, further implementation of the classification is executed by Job Tracker.

ELM approach is used in classifier training algorithm. Finally, send the classification results back to the client. As show in Figure 1:

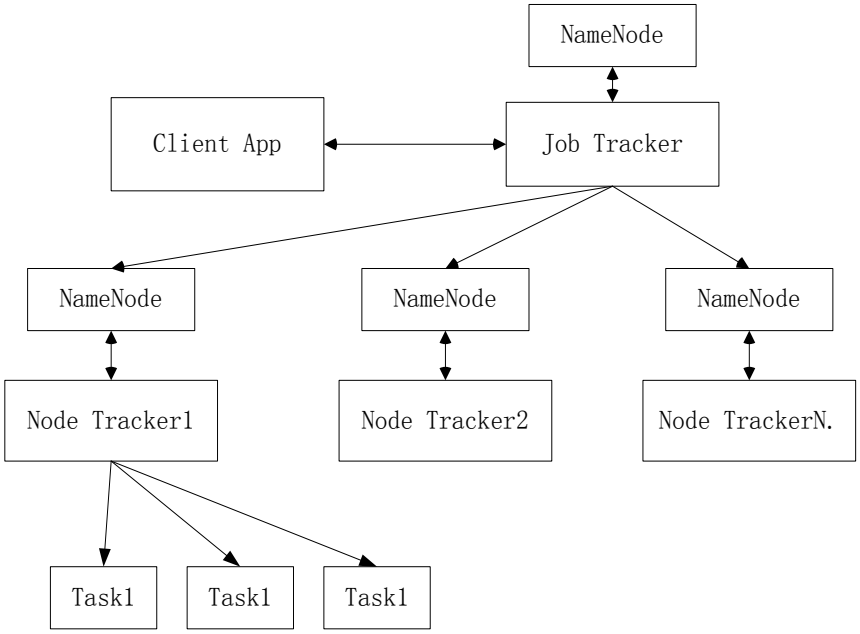


Figure 1. Cloud Computing Platform Configuration

3.1. MapReduce-based Feature Selection Model

MapReduce is composed of two verbs, namely the control task decomposition and aggregation. It is a programming model for dealing with large data sets. From technological innovation perspective, MapReduce is not an innovative technology. Distributed parallel computing programming is also simple, Streaming tool in Hadoop is convenient to use. General programming techniques can develop a distributed parallel program; the programmer can run their programs on a distributed system without knowing too many details in parallel programming and distributed programming. Wherein, Map is a highly parallel operation, usually use one parallel processor to handle the calculation of the sub-tasks. Reduce [19] is usually used to collect the results of the analysis of the final composition of each sub-task. In MapReduce environment, relationships between two tasks are typically decomposed into a plurality of tasks. One relationship between the tasks is irrelevant, these tasks can be executed in parallel; while the other relationship tasks are interdependent, their order cannot be reversed, so these tasks cannot be parallel performed. In the environment of MapReduce parallel and distributed computing model, a program can have a lot of common computer cluster automatically composed under concurrent environment.

In classification model built on cloud platforms, first do information gain characteristics filtering on gene expression data on MapReduce. Steps are as follows:

1) Genetic data sets were randomly divided into blocks by randomized functions. For simulation, experiment consists of a PC with five Hadoop cloud computing platform.

2) Map function calculates information entropy of block set features. Here Map function is defined as information gain algorithm. By setting the number of feature size, the cloud platform divide feature set reaches at time t automatically, each block of data corresponds to a

Map task, each Map task calculates its respective information entropy feature sets, among the same time different Map parallel computing, get all the information entropy at time t.

3) Execution of the Reduce task, including the selection and integration of features. Feature selection is conducted in accordance with the standard information gain algorithm.

3.2. MapReduce-based Gene Expression Data Classification Model

Classification problems are widely spread in the fields of machine learning and data mining. It has been a hot research realm at home and abroad. Its theoretical and applied research has achieved fruitful results [20]. Currently commonly used classification methods are support vector machines, Bayesian decision, artificial neural networks and their improved algorithm, *etc*. Extreme Learning Machine [21-23] (ELM) is a fast machine learning algorithm recently developed; its fast learning speed and strong generalization ability, giving it significant advantage in classification applications.

After gene expression data information gain characteristics filtering on MapReduce, characteristics for classification are feeding back to Job Tracker. Job Tracker, by executing ELM algorithm, train and test genetic characteristics obtained.

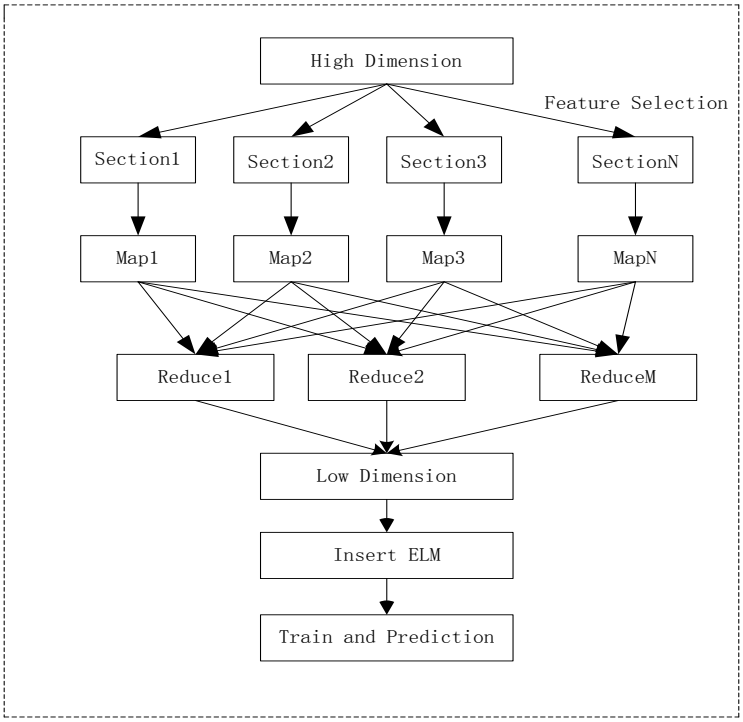


Figure 2. Classification Model on Cloud

4. Experiment

On the basis of theoretical analysis, this section selects four groups of gene expression datasets to test the performance of the Construction Method Based on Information Gain and Extreme Learning Machine Classifier Gene Expression Data on a Cloud Platform. Among them, Breast, Colon, Heart is the two types of data; Leukemia is the multi class data. The information of datasets is shown in Table 1:

Table 1. Datasets

Datasets	Sample Num	Gene Num	Distribution	
			Class	Num
Breast	97	24481	Relapese	46
			Non- Relapese	51
Leukemia	72	7129	ALL	24
			MLL	20
			AML	28
Colon	62	2000	Negative	40
			Positive	22
Heart	270	3510	Negtive	150
			Positive	120

Before feature extraction, standardize gene expression matrix elements need to be logarithmic transformed.

$$x'_{ij} = \frac{x_{ij} - \overline{x}_i}{\sqrt{\frac{1}{N-1}\sqrt{\sum_{j=1}^{N}\left(x_{ij} - \overline{x}_i\right)^2}}} \quad (4\text{-}1)$$

This article studies Breast data set as experimental example to do analysis and research. First, block data sets, using random function on the cloud platform with 4 PC deployed DataNode and TaskTracker, namely for each part of 6120. Each feature data corresponding to a Map task, every Map task information entropy of each feature set, using the mutual information maximization of mutual information method feature set start the Reduce step then. In the Reduce steps on step on to get the mutual information of sorting, filtering characteristics, from the top 1224 feature. Finally summarize, shipped to the client, on the client side with ELM to obtain the genetic traits of training and testing. This article here with Breast data set as experimental example analysis and research. First of all, random function is used on the cloud platform with 4 PC deployed DataNode and TaskTracker to block of data sets, namely for each part of 6120. Each feature data is corresponding to a Map task and every Map task information entropy of each feature set, using the method of information gain to start. The reduce step then to get the information gain of sorting, filtering characteristics, from the top 1224 feature. Finally summarize, shipped to the client, on the client side with ELM to obtain the genetic traits of training and testing. The experimental results are shown in Figures 3 and 4.

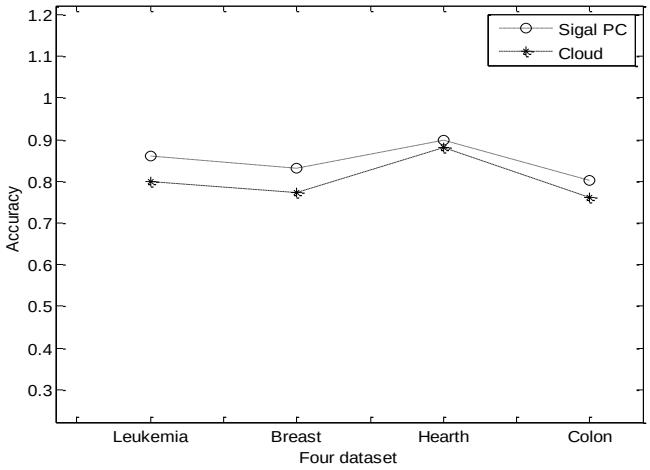


Figure 3. ClassificationAccuracy

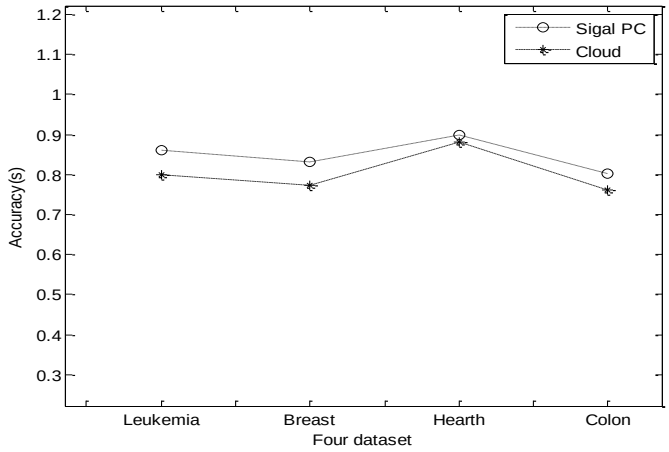


Figure 4. Classification Time

As show in the Figure 3, in a cloud platform, after filtered through the information gain of characteristics, the characteristics and class label, respectively as the input and output of ELM for training and testing, the 5 fold cross-validation, obtain the highest accuracy can reach 90%, the same as the ordinary PC environment characteristic quantity under the premise of comparison can be found that the classification accuracy is roughly same, illustrating the proposed algorithm in classification precision which has the feasibility and effectiveness. It can ensure that the extracted features are effective and have higher classification accuracy. As show in the Figure 4, due to the parallel computing performance of cloud platform, by the time it increases 4 times compared with the ordinary PC speed by the time promises the higher classification accuracy. The speed will be more obvious with the increase of number of servers, thus saving time for big data learning resources, and illustrates that the cloud platform is highly parallelized.

5. Conclusion

A recognition method is constructed based on MapReduce algorithm on high dimension feature selection in this paper. The parallel method that we build on hadoop platform distributed computing model can be used on other sample's feature selection and classification. The results of the simulation experiments demonstrate that the efficiency of construction classification method can make extraction of features in a higher classification accuracy faster, which save a lot of time resources to make a highly efficient gene feature extraction system.

Acknowledgments

The authors would like to thank the National Natural Science Foundation of China (No. 61272315, No.60842009, and No. 60905034), Zhejiang Provincial Natural Science Foundation (No.Y1110342, No.Y1080950) and the Pao Yu-Kong and Pao Zhao-Long Scholarship for Chinese Students Studying Abroad.

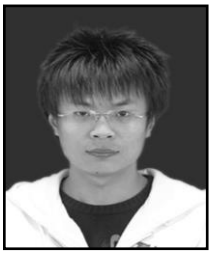
References

[1] J. Han and M. Kamber, "Data mining: concepts and techniques", San Francisco CA: Morgan Kanfnann, **(2001)**.
[2] K. Chen and W. M. Zheng, "Cloud computing: software system instance and research status", vol. 5, no. 20, **(2009)**.
[3] China cloud computing network, http://www.chinacloud.cn. **(2011)**.
[4] H. J. Lu, "A Study of Tumor Classification Algorithms Using Gene Expression Data", Xu Zhou: China Mining University, **(2012)**.
[5] G. B. Huang and Q. Y. Zhu, "Siew CK Extreme learning machine: theory and applications", Neurocomputing, vol. 70, **(2006)**.
[6] W. Shang, Z. M. Jiang, B. Adams and A. E. Hassan, "Mapreduce as a general framework to support research in mining software repositories", Proceedings of the Fourth International Workshop on Mining Software Repositories, **(2009)**.
[7] A. Matsunaga, M. Tsugawa and J. Fortes, "CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications", 4th IEEE international Conference, **(2008)**.
[8] J. Ekanayake and S. Pallickara, "MapReduce for Data Intensive Scientific Analyses", IEEE Fourth International Conference on eScience, **(2008)**.
[9] J. T. Ren and J. H. Sun, "A genetic algorithm based on information gain and feature selection algorithm", Computer Science, **(2006)**.
[10] U. Alon, N. Barkai and D. A. Notterman, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays", Proc. Natl Acad. **(1999)**.
[11] S. Raychaudhuri, J. M. Stuart and R. B. Altman, "Principal components analysis to summarize microarray experiments: application to sporulation time series", Pacific Symposium on Biocomputing, Honolulu, Hawaii, USA, **(2000)**.
[12] J. Khan, M. Bittner and Y. Chen, "DNA microarray technology: the anticipated impact on the study of human disease", Biochimica at Biophysica Acta. **(1999)**.
[13] A. Narayanan, S. S. Tatineni and J. Gamalielsson, "Reverse engineering causal networks from multiple myeloma gene expression data", **(2002)**.
[14] S. Ramaswamy, P. Tamayo and R. Rifkin, "Multiclass Cancer Diagnostic Using Tumor Gene Expression Signatures", Proceedings of the National Academy of Sciences, **(2001)**.
[15] H. J. Lu, J. J. Lu, M. Y. Wang and Y. Lu, "Based on compressed sensing of cancer gene expression data classification", China Institute of Metrology, **(2012)**.
[16] G. B. Huang and Q. Y. Zhu and S. chee-kheong, "Extreme learning machine: Theory and applications", Neurocomputing, **(2006)**.
[17] Q. H. Liu and Z. Liang, "Optimization of information gain feature selection method", Computer Engineering and Applications, **(2011)**.
[18] Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper Spectrum", IEEE Trans on Speech and Audio Processing, **(2004)**.
[19] D. Gillick, A. Faria and J. Denero, "MapReduce: Distributed Computing for Machine Learning", **(2006)**.

[20] H. J. Lu, C. L. An, X. P. Ma, E. H, Zheng and X. B. Yang, "Disagreement Measure Based Ensemble of Extreme Learning Machine for Gene Expression Data Classification", Chinese Journal of Computers, **(2013)**.
[21] Z. H. Man, K. Lee, D. H. Wang, Z. W. Cao and S. Khoo, "Robust Single-Hidden Layer Feedforward Network-Based Pattern Classifier", IEEE Transactions on Neural Networks and Learning Systems, **(2012)**.
[22] W. Ji-yi, F. Jian-qing, P. Ling-di and X. Qi, "Study on the P2P Cloud Storage System", Acta Electronica Sinica, vol. 39, no. 5, **(2011)**, pp. 1100-1107.
[23] G. B. Huang, X. Ding and H. M. Zhou, "Optimization method based Extreme Machine for Classification learning", Neurocomputing, **(2010)**.

Authors

Wei Shasha, she is currently a graduate student in College of Information Engineering, China Jiliang University. She received her B.S. from Shi Jia Zhuang Railway University in2012. Her research interests are in the field of Cloud Computing, Machine Learning and Data Mining.



Lu Huijuan, corresponding author. She received her Ph.D. and B.S. from China University of Mining & Technology, the M.S. from Zhejiang University. Now she is the Professor of China Jiliang University. She is the executive director of CCF and the member of China cloud computing Expert Committee. She is principally engaged in cloud computing， pattern recognition, bioinformatics, data mining.

Jin Wei, he is currently a graduate student in College of Information Engineering, China Jiliang University. He received his B.S. from China Jiliang University in2012. His research interests are in the field of pattern recognition, cloud computing and data mining.

Li Chao, he is currently a graduate student in College of Information Engineering, China Jiliang University. He received his B.S. from Henan Polytechnic University in 2012. His research interests are in the field of cloud computing and data mining.

