Research on Online Identification of Error Based on Multi Source Data Mining

Shihai Yang[1], Taiwen Dai[2], Shuangshuang Zhao[1], Tao Li[*2]

State Grid Jiangsu Electric Power Research Institute
Nanjing, Jiangsu, China
Fujian Yirong Information Technology Co., Ltd.
Fuzhou, Fujian, China

***Abstract* — In this paper, the author studies the online identification of error and anomaly based on multi source data mining. A new cluster algorithm in the data mining field---CMS (Classification from Multiple Sources) is introduced emphatically, which is to be used in this research. Combining with the process of knowledge discovery in multiple source data mining, the paper proposes a new segmentation model based on multiple source data mining for online identification of error and anomaly. The result shows that after using Classification from Multiple Sources (CMS) in multi-source data mining, the efficiency of online error identification can be improved.**

Keywords - online identification of error; error and anomaly; classification from multiple sources

I. INTRODUCTION

With the development of information technology, large amounts of data is produced in different applications and accumulated at different locations in distributed way. How the useful and hidden knowledge/patterns can be extracted from the accumulated data is one of the most challenging issues[1]. Grid technology enables the collaboration and sharing among distributed and heterogeneous resources. Applying data mining in Grid provides an effective solution to extract knowledge from large-scale geographically distributed data. Since data mining is a non-trivial process which is composed of many operations executed on large amounts of data, the combination of data mining and Grid will inevitably increase the complexity of data mining processes. In the previous research, data mining process is always treated as independent black-box algorithms in applications in which the functionality and intermediate steps are hidden. During this process, the execution processes of data mining are invisible to users and environments, and data mining algorithms used in central environments cannot be automatically trans-formed to the processed that can be executed in distributed environments according to the distributed resources, and users cannot control data mining execution; moreover, the independence between the interfaces for data mining services and Grid services are in-convenient for users to access data mining services in Grid. As a result, data mining cannot work efficiently as we expect in Grid environments. As the problem is encountered in railway freight application system, based on Railway Freight Grid, how to distribute computational resources can be efficiently used to extract knowledge from the freight data distributed at railway bureaus in order to support decision making. In our approach, data mining algorithms are decomposed as execution process models which are composed of finer-grained data mining operators, and then the models are optimized according to the distribution of data and computational resources in Grid to get the distributed data mining execution process models; the execution engines schedule the models and assign the tasks to different nodes in the Grid, and users can get the data mining results via unified and Grid-compliant interfaces. In the thesis, based on Grid, the approach is used to process the following data mining algorithms: association rules mining, sequential patterns mining, CART classifier and naive Bayesian classifier. The major contributions of Zhang's thesis include: Data mining execution process model composed of finer-grained data mining operators enabling to describe the execution process of data mining algorithms. Users, applications and execution environments can have a clue about the intermediate steps and intermediate results via the execution process models[2]. The data mining operators are evaluated based on simulation data by the experiments which are executed in central environment and the result shows that data mining execution process model can show the execution of every step of data mining algorithms. The optimization algorithm proposing how to transform data mining execution process models to distributed ones which can execute in Grid, the optimization algorithm is divided into three sub-processes: data localization, global optimization and local optimization, and in every sub-process, data mining operators are optimized according to the type of operators and the distribution of data. Distributed data mining execution process models are evaluated based on simulation data in Grid, the results prove that distributed models can execute in shorter response time and use computational resource in more balanced way than centralized processing. DMEP engine providing a runtime environment for data mining execution process models in Grid, in the engine.

II. THE FRAMEWORK OF ONLINE IDENTIFICATION OF ERROR

As one of the main branches of the automatic control science, system identification has been applied in many fields. In the past, system identification was mostly applied in linear system modeling, and the flawless theory for linear system had come into being for many years. However, with the development of society and science, nonlinear system is more and more important. The conflict between control and model is getting more and more evident. This fact results in the development of nonlinear system identification theory.

Testing and measure are indispensable means for human beings to understand the objective things in the nature and to quantify the phenomena of these things in order to see through the essence. And measuring instruments are essential tools to realize testing and measure. As the advanced delegate of measuring instruments virtual instruments which are the outcome of tight combination of computer software, hardware and bus technology preserve good characteristic of measure. The core of virtual instruments is to realize measuring function by means of software. Errors are inevitable in all the stationary and dynamic measuring process which prevents us to get the real value of the measured quantity directly. Nowadays technologies develop fast and people demand a higher precision for products. Measuring technology is also expected to achieve a higher accuracy. Hence, to research into the errors of measure and measuring system in order to diminish and even eliminate errors, enhance the precision of measure is necessary. Virtual instruments are inevitable to cause error to damage the precision of measuring results. However, a whole system to evaluate errors for virtual instruments is not available currently. Therefore, to research into the error of virtual instrument system and educe the error analysis and modification methods is of profound theoretical and practical value. Liu's paper researches deep into mechanical measuring theory, systematic error analysis and synthesis method, constitution of virtual instrument system, error of each subsystem, error description and modification of virtual instrument system and so on[3]. Vedachalam discusses the principle of virtual instrument and its error, carries through innovative systematic researches on error analysis and modification method for virtual instrument system[4]. According to the mechanical measuring systematic error constitution theory and based on the analysis of virtual instrument systematic construction, it researches into the errors of each virtual instrument component, and gets the error analysis and modification method of sensors, data acquisition system and virtual instruments' software. Fakhrahmad researches into the virtual instrument systematic error transfer and synthesis using mechanical measuring theory and theories and methods of error transfer and synthesis in error theory. Based on the analysis of errors of each virtual instrument component, it generalizes the error constitution formula of virtual instrument systems and gets the uniform error description and analysis method for virtual instruments[5]. It takes the advantages of the idea that the core of virtual instrument technology is to replace or partly replace hardware with software and proposes the error modification method: (1) use software to implement the function of hardware to those hardware which easily cause error or whose errors are too high; (2) design software to modify the errors of hardware which can't be replaced by software. 3) It provides example for error analysis and modification of virtual instrument system. It discusses the error analysis and modification of virtual noise analyzer and virtual ECG analyzer.

III. THE ALGORITHM

In online error identification, the object we wish to acquire is a vector, while in many application of practical interest, we often wish to reconstruct an object in the form of a matrix, which is more convenient for data acquisition, modelling, processing and analyzing. However, these data often suffer from the problem of deficiency, loss, or corrupted with noises. The goal of matrix recovery is to acquire the exact data under these situations.

The basic model for online error identification as follows:

$$TSP(t)=\begin{cases} TSP_1 & 0\le t\le \Delta t \\ TSP_2 & \Delta t\le t\le 2\Delta t \\ \cdots & \\ TSP_n & (n-1)\Delta t\le t\le n\Delta t \end{cases} \tag{1}$$

$$\min d(T)=\sum_{k=1}^{n}\sum_{i=1}^{N}\sum_{j=1}^{N}c_{i,j}(k\Delta t) \tag{2}$$

$$s.t \qquad \Delta t=\frac{T}{n},\frac{\Delta c_{i,j}}{\Delta t}=0 \tag{3}$$

In the problem of low-rank matrix recovery, traditional algorithms obtain low-rank solutions by minimizing the nuclear norm of the matrix. However, the algorithms are unstable when the data matrix is highly correlated. Therefore, we consider the matrix elastic-net regularization, which can lead to stable solutions by minimizing the empirical risk penalized with the combination of the nuclear norm and the Fresenius norm. Moreover, the solution of the model can be characterized as the fixed point of a contractive map by the proximity operator. Then we construct a fixed point iterative scheme for solving the model. The theoretical results show that the sequence of iterates converges to the solution of the matrix elastic-net regularization. The error bounds of the matrix elastic-net regularization model under the assumption of matrix restricted isometric property (RIP) are analysed. We give the definition of the RIP for matrix recovery, and show that certain classes of random measurements satisfy the matrix RIP. The analysis indicates that the error is proportional to the number of degrees of freedom times the noise level under the matrix RIP. Error bound is also established for full-rank matrices.

In the framework of statistical learning theory, we give an all-round analysis on the convergence and the generalization performance of the matrix recovery algorithms under the operator assumptions. We describe the matrix recovery problem as a learning problem, and define some Hilbert-Schmidt operators. The generalization error

bounds for matrix recovery are then obtained by estimates of Hilbert-Schmidt operators. It is worth mentioning that we also give an adaptive scheme to select the regularization parameter. The stability of the matrix recovery algorithms is studied. We consider the uniform β-stability of matrix recovery algorithms and characterize the conditions on the penalty function that lead to uniformβ-stability. In particular, we apply our results to show that the matrix elastic-net regularization is uniformly β-stable.

We may get the calculating method for the main index in the following equation (4)-(5):

$$M_{ij} = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{\sigma^2}\right) \tag{4}$$

$$L = \begin{bmatrix} L_1 & & \\ & \ddots & \\ & & L_k \end{bmatrix} \tag{5}$$

Their matching eigenvectors matrix is shown in the following equation (6):

$$H = \left[h_1, h_2, ..., h_k\right] = A^{1/2}E \tag{6}$$

So, we can get:

$$U_{ij} = \frac{H_{ij}}{\sqrt{\sum_{t=1}^{k} H_{it}^2}}, i = 1, ..., n,\ j = 1, ..., k \tag{7}$$

$$P = I - A^{-1/2}MA^{-1/2} \tag{8}$$

According to the equation (6), the calculating formula can be obtained in equation (7)-(10).

$$\boldsymbol{g}(x,\omega) = \frac{1}{(2\pi)^3}\int \boldsymbol{g}(k,\omega)\exp(-i\mathrm{k}\square\mathrm{x})d\mathrm{k} \tag{9}$$

$$\boldsymbol{g}(k,\omega) = \begin{Vmatrix} G_{ik}(k,\omega) & \gamma_i(k,\omega) \\ \gamma_k^T(k,\omega) & g(k,\omega) \end{Vmatrix} \tag{10}$$

$$G_{ik} = (\Lambda_{ik} + \frac{1}{\lambda}h_i h_k^T)^{-1},\ g = -(\lambda + h_i^T\Lambda_{ij}^{-1}h_j)^{-1},$$

$$\gamma_i = \frac{1}{\lambda}h_k^T G_{ki} \tag{11}$$

$$\Lambda_{ik}(k,\omega) = k_j C_{ijkl}^0 k_k - \rho_0\omega^2\delta_{il},\ h_i(k) = e_{kil}^0 k_k k_l,$$

$$h_l^T = e_{ikl}^{0T}k_i k_k,\ \lambda(k) = \eta_{ik}^0 k_i k_k \tag{12}$$

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ik_3 x_3'}dx_3' = \delta(k_3) \tag{13}$$

IV. EXPERIMENT AND DATA ANALYSIS

Data mining is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. It has achieved very significant results in multiple applications such as banking and communications in recent year. Both clustering and classification are key technologies of data mining. There are dozens of algorithms available for a specific data mining task. However, in practice, data mining techniques are severely limited by two main challenges. One is lack of rules or knowledge to help people to choose proper algorithms for their ongoing data mining project, followed by difficult to test the reliability and the operating efficiency of the selected model. For the above mentioned problems, we focus on the model selection of data mining, specifically, on the clustering and classification. The tasks space contains the description of the ongoing data mining project. The algorithm space defines the available algorithms for the task. Evaluating criteria space contains the metrics used to evaluate the performance of the models. Evaluating strategy describes which and how multiple criteria decision making (MCDM) methods are used in the model ranking and final selection. Some representative clustering algorithms are selected from the division based algorithm, hierarchical algorithm, density based algorithm and model based clustering algorithm. We construct the evaluating space with11performance metrics which are from external, internal and relative clustering evaluation criteria. After comprehensive comparative analysis by using the MCDM methods, a mechanism to automatic algorithm selection for a specific data mining task are built. We examine the model selection problem at the direction of binary classification and ensemble learning for the software defect prediction. It is acquired two main achievements based on the ranking results. On the one hand, empirical knowledge was acquired for the software defect prediction task. On the other hand, the ranking results are used to guide the design of feature transform, new algorithms for specific tasks. Guided by the classification models selection results, a density based over sampling algorithm and an ensemble feature selection method for imbalanced learning are proposed. The experimental results indicated that the proposed algorithms are effective for the imbalanced learning problems. Figure 1 shows the performance index for each node after using multi source data mining.

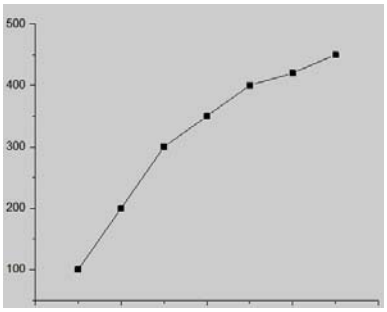

Figure 1. The performance index for each node after using multi source data mining

A zone partition method is presented according to the order and map theory based on decision table and fuzzy monotone models; two fuzzy monotone membership functions are presented based on minimum and average value of zone separately; two fuzzy monotone relationships are constructed and their parameters properties are discussed after some propositions are presented and proved; then two data mining algorithms based on minimum and average value of zone separately are presented; the UCI international wastewater data are used to validate the effect of two data

mining algorithms after the decision rules being analyzed and designed; the rough and fine degree of these two data mining algorithms is compared and analyzed; the merits and different properties are shown thorough the comparison of other attributes reduction algorithms, and the time complexity is given too. Figure 2 shows comparison of performance improvement before and after using multi source data mining. Figure 3 shows the trend rate of performance improvement before and after using multi source data mining.

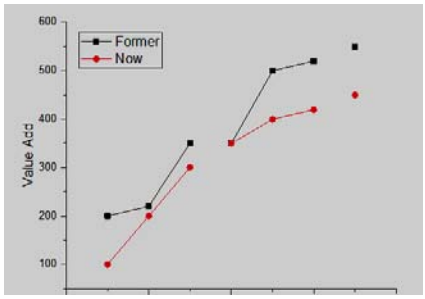

Figure 2. Comparison of performance improvement before and after using multi source data mining

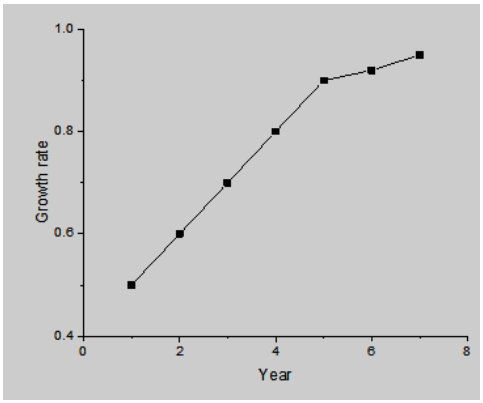

Figure 3. Trend rate of performance improvement before and after using multi source data mining

The optimization algorithm proposing how to transform data mining execution process models to distributed ones which can execute in Grid, the optimization algorithm is divided into three sub-processes: data localization, global optimization and local optimization, and in every sub-process, data mining operators are optimized according to the type of operators and the distribution of data. Distributed data mining execution process models are evaluated based on simulation data in Grid, the results prove that distributed models can execute in shorter response time and use computational resource in more balanced way than centralized processing. Figure 4 shows the impact of performance improvement before and after using multi source data mining and Figure 5 shows the influence of performance improvement before and after using multi source data mining.

According to the results, system delay is determined. Applying predictive error identification method, by comparing the different order model, model structure and parameters of the steering gear is determined. Aiming at satellite-satellite pointing/tracking system, an experiment based on xPC real-time simulation platform is designed. White noise acting as inspirit signal, the experiment data is collected. Utilizing these data and error back propagating identification method, different neuron and input-output delay are selected. By comparing approximation ability and generalization ability, the neural networks model in position mode and velocity mode is identified. The algorithms are unstable when the data matrix is highly correlated. Therefore, we consider the matrix elastic-net regularization, which can lead to stable solutions by minimizing the empirical risk penalized with the combination of the nuclear norm and the Fresenius norm. Moreover, the solution of the model can be characterized as the fixed point of a contractive map by the proximity operator. Then we construct a fixed point iterative scheme for solving the model.

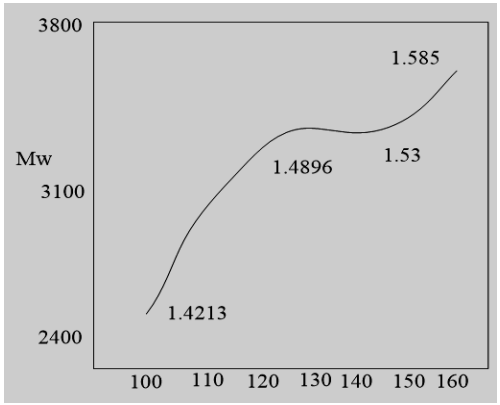

Figure 4. The impact of performance improvement before and after using multi source data mining

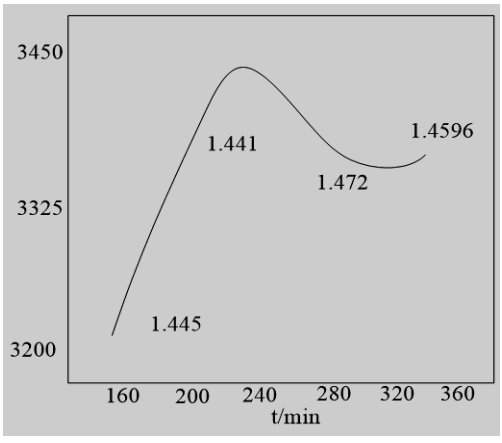

Figure 5. The influence of performance improvement before and after using multi source data mining

V. CONCLUSION

In this paper, the author studies the online identification of error and anomaly based on multi source data mining. We construct a fixed point iterative scheme for solving the model. The theoretical results show that the sequence of iterates converges to the solution of the matrix elastic-net regularization. The error bounds of the matrix elastic-net regularization model under the assumption of matrix restricted isometric property are analyzed. We give the definition of the RIP for matrix recovery, and show that certain classes of random measurements satisfy the matrix RIP. The analysis indicates that the error is proportional to the number of degrees of freedom times the noise level under the matrix RIP. Error bound is also established for full-rank matrices.

Guided by the classification models selection results, a density based over sampling algorithm and an ensemble feature selection method for imbalanced learning are proposed. The experimental results indicated that the proposed algorithms are effective for the imbalanced learning problems. A new cluster algorithm in the data mining field---CMS (Classification from Multiple Sources) is introduced emphatically, which is to be used in this research. Combining with the process of knowledge discovery in multiple source data mining, the paper proposes a new segmentation model based on multiple source data mining for online identification of error and anomaly. The result shows that after using Classification from Multiple Sources (CMS) in multi-source data mining, the efficiency of online error identification can be improved.

REFERENCES

[1] Bréant Claudine, Thurler Gérald, Borst François, et al. “Design of a Multi-Dimensional Database for the Archimed Data Warehouse”, Studies in health technology and informatics, vol. 2, No. 07, pp. 116-121, 2005.

[2] Shichao Zhang, Qingfeng Chen, Qiang Yang. “Acquiring knowledge from inconsistent data sources through weighting” Data Knowledge Engineering, vol. 12, No. 05, pp. 698-705, 2010.

[3] Zhenguo Liu, Zhengfu Bian, Fuxiang Lü et al. “Monitoring on subsidence due to repeated excavation with DInSAR technology”, International Journal of Mining Science and Technology, vol. 13, No. 07, pp. 232-242, 2013.

[4] N. Vedachalam, S. Muthukrishna Babu, G.A. Ramadass, et al. “Review of maturing multi-megawatt power electronic converter technologies and reliability modeling in the light of subsea applications”, Applied Ocean Research, Applied Ocean Research, vol. 2, No. 04, pp.46-55. 2014.

[5] S. M. Fakhrahmad, M. H. Sadreddini, M. Zolghadri Jahromi. “A proposed expert system for word sense disambiguation: deductive ambiguity resolution based on data mining and forward chaining”, Expert Systems, Applied Ocean Research, vol. 33, No. 06, pp.322-333.2015.