Study on simulation data analysis of complex system based on data mining method

Jinhai Zhang

Marine college of Shandong jiaotong university, Weihai , Shandong,China

Keywords: Data analysis, data mining, support vector machine ,parameter optimization

Abstract. Analysis of simulation data is purposefully collected data and analyze data, make this information or knowledge of the process is conducted in order to better understand and improve the system, is the focus of simulation problems. Due to existing simulation and data analysis methods, lacking to deal effectively with large scale and high dimension and interaction of complex method for simulation of complex data, so data mining techniques applied to the analysis of simulation data, using a prediction model based on support vector machine method for wide-range programme optimization and data trend forecast provides a solution to the problem.

Introduction

Analysis of simulation data is purposefully collected data and analyze the data and turn it into information or knowledge. Its purpose was amid a large number of random information in the simulation data set, extraction and refining, to find research on simulation system of internal rules to better understand and improve the system to provide decision support.

As gradually to the direction of complex system of military simulation, simulation input/output variables increasing effect associated with each other and become more complex, increasing uncertainty, have a huge challenge on simulation data analysis and processing. Traditional data analysis methods based on statistical theory for small size, low dimensions, relatively simple data relationships with very good handling ability. Simulation system for complex data, namely mass, higher dimensions, the relationship between complex data, there is a lot of uncertainty factors, traditional data analysis capacity will be greatly reduced, such as slow calculation, calculating costs, requires a lot of computer resources, analysis results reliability.

Data mining from large, incomplete, there is noise, fuzzy and random data to extract valuable information and knowledge effectively. Therefore, it is necessary to apply data mining techniques in the field of simulation and data analysis, simulation and data analysis for researchers to understand and improve the simulation provide powerful decision support systems, simulation of complex system based on data mining data analysis conducted in-depth research. analysis of the results of this research for the actual simulation data provides methodological guidance.

Characteristic analysis and simulation data preprocessing method

Need for data analysis of simulation data from simulation experiment, and simulation experiment different Yu traditional of experiment method: simulation experiment factors far than traditional experiment, traditional experiment General up to consider more than 10 a factors, and simulation experiment may has hundreds of a experiment factors, and each factors may has many level; traditional experiment General only a quantitative of performance index, and simulation experiment in the often will involved multiple index, and index may is qualitative of, also may is quantitative of.

Among them, the simulation of a simulation in order to achieve a certain goal, a simulation may contain more than one test scenario, each scenario has multiple simulation runs required, each want to set run times are not necessarily the same.

Precisely because of the structural characteristics of simulation and experimental purposes, and simulation data needs are as follows: individual scenario analysis, comparison of two or more scenario analysis and optimization of wide range of programmes. For single wants to set of analysis, due to simulation has randomness of features, need used Statistics analysis of method, while to on experiment results data of statistics features for analysis, while also to to out corresponding of estimated value; for multiple wants to set of compared analysis, is to compared alternative programme of differences, on both of poor of expectations building confidence interval, and through assumed test to judge two a alternative programme of expectations Zhijian whether has significantly of differences.

Mass, dimensions and other characteristics of simulation data, if these data directly for simulation data analysis, even if effective algorithm for high dimensional data, it is difficult to achieve optimal results. For this type of data, the properties should be reduced in order to achieve the purpose of dimension reduction. In addition, some data from hardware in the loop simulation of measuring element, due to the influence of the measurement error, there is often a lot of abnormal data in the data warehouse. Data cleaning and the data in the data warehouse system will not be washed away in the process, hence the need to use an algorithm to find it.

Attribute reduction, also known as dimensional reduction, is a kind of high-dimensional data dimension reduction method. Attribute reduction has many benefits, key is, if the dimension (the number of properties) is low, many data mining algorithm will produce better results. This is because attributes can remove irrelevant properties and reduce the noise reduction, and can be part of the solution "dimensionality". Moreover, the attribute reduction may make it easier for simulation model to understand, because it may involve less properties for the model. In addition, the attribute reduction can also make it easier for data visualization. Attribute reduction even if no data reduction to two or three dimensions, data through visualization of two or three properties, and this combination will greatly reduce the number of. Lastly, attribute reduction can reduce time and memory requirements of the data mining algorithms.

Outlier detection methods are commonly used residual graph, scatter graph. But because of the abnormal value is not known in advance, data regression models with outliers can make regression curve to outliers, abnormal values of residuals instead of smaller. Therefore, the residual method is no longer available. Scatter method is only valid for a regression, but in multiple regression results are poor.

Simulation data mining based on support vector machine

Typically, data mining task is divided into two main categories:

(1) the prediction task. The mission objectives are based on the values of other properties, predict the value of a specific property;

(2) describe the task. These potential links in the task's goal is to export summary data model.

Decision tree learning inductive learning algorithm is based on the instance, it focuses on a group of no order, no rule instances derived rules of classification decision tree representation, usually used to form classification and forecast model to classify unknown data or forecasts, data mining, and so on. Its advantage is that simple decision tree construction, and applies to both small samples of data and huge amounts of data. Disadvantage will produce pieces of data (data fragmentation), that records in the leaf nodes may be too little, for leaf nodes represent the class, can't make a statistically significant judgments and subtree may be repeated several times in the decision tree, decision trees are too complex.

Support vector machine is by the United States N · Professor Vapnik and others put forward between 1992 and 1995, originally from the handling of two-class classification problem. Method is to find a hyperplane, so the training set belongs to different categories of points exactly at the different side of the hyperplane, and make those points as far away as possible from the hyperplane.

Dependencies if data is non-linear, that is, linearly inseparable, you can map the data from a low-dimensional input space into a high-dimensional feature space h to obtain a linear relationship, the SVM with linear work ability, and complete the conversion method is the core of mathematics, as shown in Figure 1.

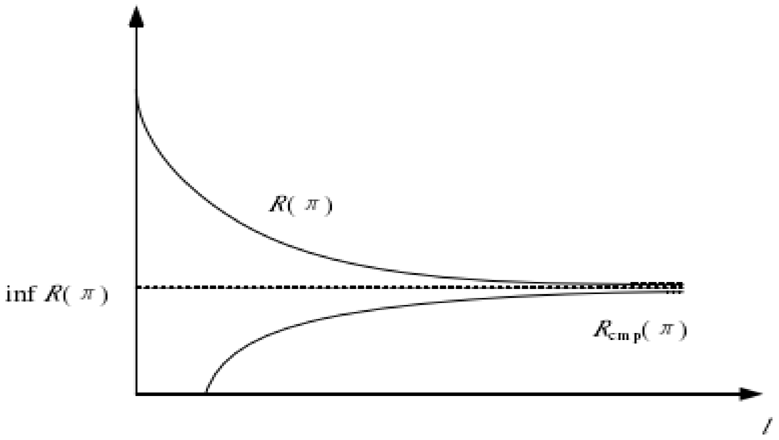

Fig.1. SVM classification

Some linearly inseparable sets of data to construct a hyperplane that does not have a category error, need to be mapped to a significant dimension of the feature space, it increases the complexity. This soft interval methods can be used, construct a hyperplane on minimum classification error probability of the entire data set, made compromise between complexity and classification errors.

Support vector machine regression model was eventually transformed into a classification model with the same problem. For non-linear regression, classification using kernel methods to map data to a higher-dimensional space, then the linear regression. Given a maximal margin classification algorithm based on contraction closed convex hull: first panning up and down through the set variable Epsilon will return problem into a binary classification problem, and the collection of two types of contraction closed convex minimum distances between packages, and finally obtained the maximum interval according to the minimum distance separating hyperplane regression function. This method establishes a relationship between a binary classification and regression, indicating regression problems can be solved by classification.

For all simulation data and not data, the difference lies in the simulation of initial conditions of the system is different, thus in the different initial conditions in the simulation data, more or less, there are properties that are not simulated data. But by analyzing the data parameters on simulation of simulation results to guess not simulation data is difficult to achieve. Support vector machine based classification algorithm can be used, to classify the simulation data, find a hyperplane; then use the category flat face for simulation data to predict, guess the likely outcome. as shown in Figure 2.

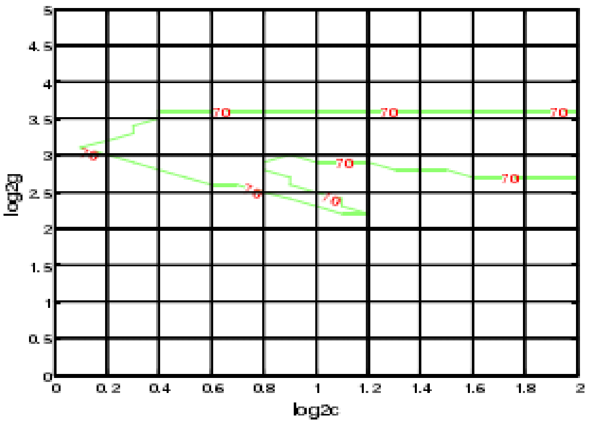

Fig.2. Three dimensional simulation map

Optimization method of model parameters

Cross validation (Cross Validation) is used to validation classification device performance of a statistics analysis method, the method of basic thought is will original data set for group, part as training set, and another part as test set, first with training set data training classification device, again using test set to test classification model of performance, to get of classification accurate rate as evaluation classification device pros and cons of performance index.

In the practical application of fitness functions according to the specific requirements and to solve the problem. Ideal fitness function is smooth, so to avoid local Optima, but in practice this is difficult to achieve. Thus in designing fitness function, we should try to avoid too much local optimal solutions, but also to avoid the global optimal solution too isolated.

Described above is the operator of genetic algorithms, in addition to this there are other genetic algorithm inversion operation, variable length chromosome genetic algorithm, genetic algorithms, and so on.

Also select instances defensive positions, will train under cross-validation method set accuracy as the fitness function in genetic algorithm value, then use GA to optimize the parameters of SVM model.

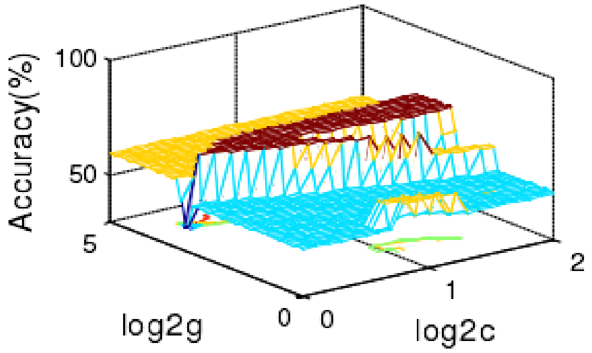

Fig.3. Parameter optimization results of grid search

References

[1] J. Han, M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2007: 113~124

[2] A Feelders, H Daniels, M Holsheimer. Methodological and Practical Aspects of Data Mining. Information & Management, 2000, 37: 271~281

[3] Y.S. Kim. Comparison of the Decision Tree, Artificial Neural Network, and Linear Regression Methods Based on the Number and Types of Independent Variables and Sample Size. Expert System with Applications, 2008, 34: 1227~1234

[4] J. Alcala-Fdez, L. Sanchez, S. Garcia, et al. KEEL: A Software Toll to Assess Evolutionary Algorithms for Data Mining Problems. Soft Computing, 2008: 17~31

[5] J.P.C. Kleijnen, S.M. Sanchez, T.W. Lucas, T.M. Cioppa. A User's Guide to the Brave New World of Designing Simulation Experiments. Tilburg University, No. 2003–01, 2003: 38~44

[6] J.C. Platt, N. Cristianini, J. Shawe-Taylor. Large Margin DAGs for Multiclass Classification. In Advances in Neural Information Processing Systems, 2000, 12

[7] B. Scholkopf and A.J. Smola. Support Vector Machines and Kernel Algorithms. In Encyclopedia of Biostatistics, John Wiley and Sons, 2003: 187~201