2014

ISSN : 0974 - 7435 Volume 10 Issue 20

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(20), 2014 [12520-12525]

Multidimensional data mining and decision tree analysis of rough set theory

Yan Li Dai
Suqian Higher Normal School, Suqian, 223800, (CHINA)

ABSTRACT

With the development and improvement of information technologies and database, data storage is used by more and more enterprises, institutes and departments, which requiring more intelligence and more accurate ways of data mining. Multidimensional data mining technology, including On Line Analytical Processing technology and data mining technology, is created on the basis of increasingly improved data warehouse technology and OLAP technology, making it more convenient for users to choose and analyze. This paper mainly researched and analyzed the multidimensional data mining of rough set, including concepts of multidimensional data and rough set, decision tree technology integrated with rough set, and the design of multidimensional data mining based on rough set.

KEYWORDS

Multidimensional data mining; Rough set; Decision tree technology.

CONCEPTS ABOUT MULTIDIMENSIONAL DATA AND ROUGH SET

Basic concept of OLAP multidimensional data set

Multidimensional database and multidimensional data set are the database presented in a multidimensional way and logistic way. Dimension is the specific way people observe the data. Levels refer to the sectors to describe the data at different degrees, for instance, the dimension of time includes such levels as day, month, year and quarter. Members are values. Data unit means the number of multidimensional arrays of which every dimension could be picked out for member, and all these members can determine a specific value. Metric, generally the index of numerical measures, is used to describe the data. Multidimensional analysis is a series of analysis (slicing, rotation and so on) on data (obtained in multidimensional ways) to observe and control the data more fully and carefully, and to get the information and substance. Data slice refers to a subset of multidimensional data set and is appointed by one or several dimensions limited by the member of dimension. Data rotation means changing the positions of dimensions for users to observe multidimensional data from other perspectives.

Concept of data mining

The function of data mining is to find information and technologies contained in the data which is hard to find. With people's increasing knowledge of the value of information, data mining is gradually developed. This technology can solve the problem of too much data with insufficient information, establish relational model and make correct prediction. It has many features and advantages over traditional processing methods, such as processing large database; random information query; useful rules and reasonably predictions on the basis of data mining technology; timely response and determination to changing data, including finding rules and managing, maintaining them. These rules are continuously updating with new data entering, and they are not suitable for all data because the database is very large, they are approved if suitable during a certain range.

Data mining process

Data mining includes four links: business object determining and theme analysis; data preparation including data selecting, cleaning, transforming and loading; data mining directly related to the selected knowledge level; results analysis and verification to get latest, usable and understandable data; knowledge assimilation to avoid contradiction. The basic process and main steps of data mining is shown as Figure 1.

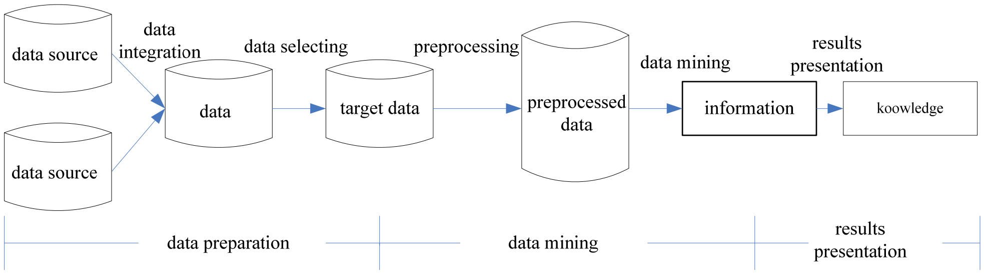

Figure 1 : Basic process and main steps of data mining

Rudimentary knowledge of rough set theory

Rough set theory has been applied in many fields (machine learning, decision making, process analysis, data mining). Rough set theory is embedding knowledge for classification into the set, and making it a part of the set. In the industrial application, some knowledge bases may be so complex and large that the redundant parts should be eliminated in order to get simplified knowledge. The simplifying process includes two fields including reduction. Knowledge dependency should also be controlled. The data in information system (decision table) of intelligent system can be presented in many ways, such as language form and number form. Inaccurate number form will make the information acquired imperfect and non-uniform, and this problem needs to be solved through the representation of knowledge. This representation is generally the information system (information table) in rough set theory. Knowing the concept of knowledge representation system makes it easy to represent the system in tabular form, namely knowledge representation system or information system attribute value table. Decision table, an important and special table, can describe some complex logic accurately and laconically. This decision table can connect independent term with several actions directly and get clear representation.

Features of rough set theory

First, it can process all kinds of data, including fragmentary data and data with many variables; second, it can process imprecise and fuzzy data, including determinacy and non-determinacy; third, it can Figure out the smallest representation of knowledge and different granulometric levels of knowledge; forth, it can reveal the manageable model with simple concept; fifth, it can generate precise rules easy to inspect and verify, especially suitable for automatic rule generation in intelligent control. In addition, the most important advantage of rough set theory is that it can provide prior information except data sets required to solve problem. Certainly, this theory is not almighty.

DECISION TREE TECHNOLOGY CONNECTED WITH ROUGH SET

Overview of decision tree algorithm

Data sorting is the most used data mining analysis method. Specifically, it is creating a classification function or model to map the data record to a pre-assumed class and make data prediction based on a well understanding of training set.

The high understandability and simple computational cost of decision tree method makes it more and more popular. Still, there are some randomness and uncertainty in this method. Decision tree is a decision analysis method used to calculate the probability that the expectation value of the Net Present Value is greater than zero through a decision tree and to value its feasibility on the premise that the probabilities of all situations are already known. It is a graphical method intuitively using probabilistic analysis. It is called decision tree because the graph of its decision branches looks like a tree. In machine learning, decision tree is a prediction model representing a mapping relation between object property and object value. Decision node is the selection of several possible schemes, namely final optimum scheme. Status node represents economic effect (expectation value) of alternative scheme. By comparing the economic effects of all status nodes, the optimum scheme can be selected in accordance with some decision standards. Result nodes represent profit and loss values of all schemes under natural condition.

Complexity and classification accuracy are the most important things in decision tree algorithm, the relationship between them are shown in Figure 2,

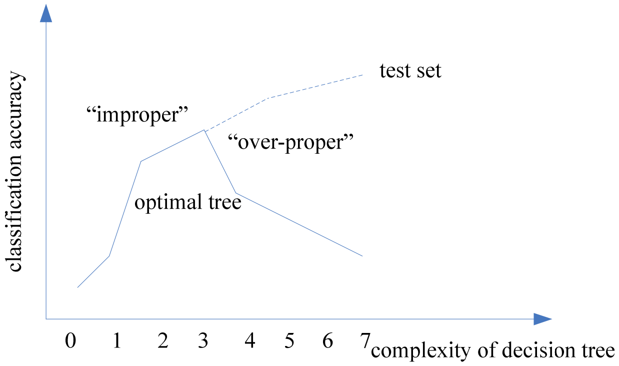

Figure 2 : Relationship between the complexity and classification accuracy of decision tree

Its assessment indicators specifically include the following five sectors: accuracy of prediction, meaning the ability of classification model to predict the type of new data; brevity of description, this assessment indicator makes decision depending on the understandability level and the way of description of the problem; complexity of calculation, mainly refers to the complexity of space and practice because they are closely related to the calculation cost; robustness of model, this assessment indicator is an complementary ability to accurately classify the data on the basis of the accuracy of prediction, especially when there is noise or the data is incomplete; scalability of treatment, means the accuracy and ability to build (classification) model when database is very large.

Algorithm design of data mining based on the integration of rough set and decision tree

Description of algorithm

Classification is predicting and assessing the type of new case by similarity through existing types in accordance with existing rules in order to master objective things. Classification mainly aims to put the elements with the same characteristics (comprised of some basic characteristics and the values of the object under this characteristic) together. Rough set cannot be set up without the support of classification mechanism. Classification represents a kind of equivalence relations, equivalence relation classification is classification of this space. Decision tree is an inductive algorithm using examples to improve the classification, prediction, processing and mining of unknown data.

Rough set theory is vital for data preprocessing and attribute reduction for its processing of mass data and elimination of redundancy are easier than other methods. However, rough set theory doesn't have cross-validation characteristic, so it may not be very accurate. Decision tree method has high-speed, simple, understandable classification rules, but it is suitable for data set with too many attributes for too many attributes may cause terrible structure in classification.

In conclusion, rough set and decision tree benefit mutually, and they are both used to process discrete data. So they can be integrated to reduce the data and eliminate the redundancy by rough set, and then find classification rules by decision tree.

Algorithm procedures

The process of data mining algorithm based on the combination of rough set and decision tree is gradually selecting the key attributes to form a new condition attributes set, and continuously repeating this process until the dependency of D to the set catch up with that of D to C. Its specific algorithm design process is as shown in Figure 3.

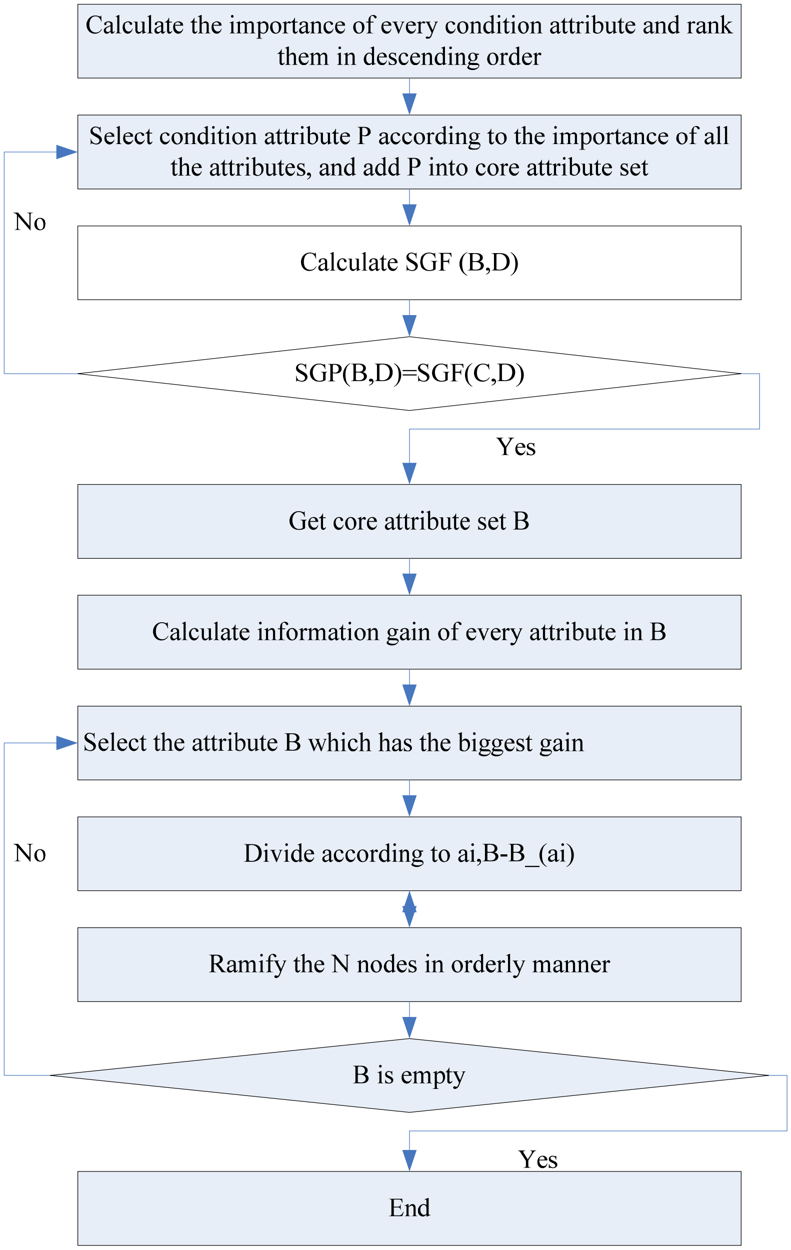
Calculate the importance of every condition attribute and rank them in descending order
Select condition attribute P according to the importance of all the attributes, and add P into core attribute set
No
Calculate SGF (B,D)
SGP(B,D)=SGF(C,D)
Yes
Get core attribute set B
Calculate information gain of every attribute in B
Select the attribute B which has the biggest gain
No
Divide according to ai,B-B_(ai)
Ramify the N nodes in orderly manner
B is empty
Yes
End

Figure 3 : Flowchart of algorithm design

Comparison of algorithm

Decision tree, which can also be called classificatory resolver, divide training set through recursion until all or most of the records in each sub-collection are the same type. The main decision tree algorithms used currently are method based on information theory and minimum GINI index method.

Usually, concluding learning system will get a decision tree which has several advantages in its application: understandable to users; less time to generate a decision tree and the ability to process large training set; simple generation algorithm and inspection; clear statement provided for sequential decision method to determine the type of a case; higher accuracy. Still, there are some disadvantages in decision tree and it is prone to be interfered by uncorrelated attributes. To solve this problem, rough set technology is required. And according to the above evaluation standards of decision tree, the combination of these two technologies can greatly reduce complexity of calculation and description.

DESIGN OF MULTIDIMENSIONAL DATA MINING BASED ON ROUGH SET

On the basis of increasingly improved data warehouse technology and OLAP technology, multidimensional technology (containing the on line analytical processing and data mining) is created. The following is the process to design and implement the data mining system through research of rough set theory and decision tree technology.

Design goal of system

This system is developed by VB.net in sql server analysis server environment. Applying this system can implement data mining on the data in data cube established in this environment. After users chose the dimension and dimension hierarchy of data, they will get different and comprehensive decision tree and implicit knowledge through this system. In addition, this system also has model validation function. According to this, users can select the most correct and appropriate mining model, and get high-accuracy information.

System design

This system has 6 data mining procedures: data preparation; data extraction; data selection; data preprocess; decision tree analysis of training data set; inspection of analysis results by inspection data set.

Data preparation needs to set up multidimensional data set for analysis in sql server analysis server environment. ADO MD technology is used in data selection. The object model of ADO MD is shown as Figure 4.

After selecting correct dimension and dimension hierarchy and cube measure, the system can generate MDX statement and get corresponding data set.

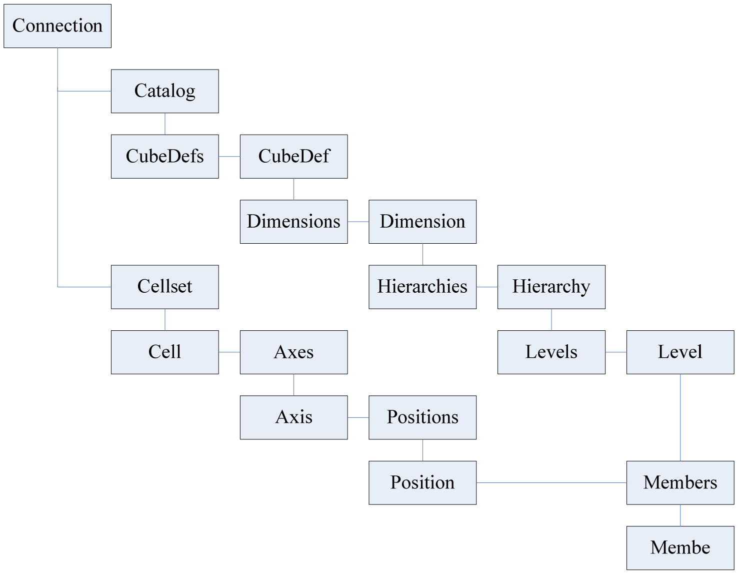

Figure 4 : Object model of ADO MD

Data extraction is to random extract data by random function to generate analysis data table, for the data generated by the previous steps are not accurate and brief enough. The data generated in this link is not preprocessed yet, so its decision attribute is sequential and it has quite a few decision attributes.

Data preprocess is to discretize and reduce the training set. Ensure all of the sequential attributes are discrete by equifrequent discretization measures. Then reduce the discretized data, and divide the data sets into training data set and inspection data set.

Carry out decision analysis on preprocessed data set, including establishing data mining model, loading selected data set into this model and analyzing these training sets by decision tree.

CONCLUSION

In conclusion, on the basis of the combination of rough set and decision tree, the system can generate more comprehensive knowledge through analysis of data in multidimensional data set by data mining algorithm, in order to raise the accuracy of decider’s decision. Currently, multidimensional data mining is not perfect and needs to be improved. For example, attentions on knowledge reduction process are not uniform; the algorithm is only suitable for discrete attribute values; there are many algorithms (sequential series analysis) can be applied into multidimensional data mining.

REFERENCES

[1] Yaojia Yi; Principle and Applications of multidimensional data analysis. Beijing: Tsinghua university press, 51-57 **(2004)**.

[2] Zhaoke Qing; Set pair analysis and its application [M]. Zhejiang Scientific Press, **(2000)**.

[3] Gaoyan; Covering rough set research[D]. Doctoral dissertation of Southwest Jiaotong University, **(2010)**.

[4] Jinxiao Fang; Application research of rough set theory in relational data base knowledge[D]. Master dissertation of

University of Electronic Science and Technology of China, **(2011)**.

[5] Xuyi, Lilong Shu; Variable Precision Rough Set Model Based on (α, λ) Connection Degree Tolerance Relation [N]. ACTA Automatica Sinica, **37(3)**, **(2011)**.

[6] Zhuxing Tong, Xubo; A XML data mining model based on rough set theory[J].Science Technology and Engineering, **20 (2011)**.