Data mining technology based on rough set and genetic algorithm under large data environment

Liping Wang
Pingxiang University, Pingxiang Jiangxi, 337055

Keywords: Data mining; rough set; genetic algorithm; association rule mining

Abstract. New requirements drive the birth of new technologies. Data analysis is the foundation of scientific research, many scientific research is based on data collection and analysis. This paper introduces the basic theory of classical rough set, which is based on the equivalence relation, and uses an upper and lower approximation to represent an imprecise concept. Property reduction is one of the core problems of rough set theory. The heuristic algorithm find attribute reduction is an effective way to solve the problem. On the basis of rough reduction algorithm based on rough set theory, this paper makes a preliminary realization of the problem based on rough set theory, which includes: data discretization, knowledge reduction, rule extraction. The coarsening algorithm based on rough set theory for attribute reduction is essentially a range of attribute values which narrow to wide process.

Introduction

Data analysis is the foundation of scientific research, many scientific research is based on data collection and analysis in the current business activities [1-3]. The data analysis is always and some special groups of people high IQ behavior, because not every ordinary person can predict future trends or make the right decisions from past sales, but as a business or industry continues to accumulate business data, especially due to the popularity of the database, To manually sort out and understand such a large data source has been the existence of efficiency, accuracy and other issues. Therefore, to explore automated data analysis technology to provide business can bring business decision information and become inevitable [4].

In fact, data, information, and knowledge can be thought of as different forms of generalized data representation, and databases are one of the most efficient ways to organize and store data today, but database query techniques have shown limitations in the face of ever-expanding data intuitively, information or effective information refers to the data that is helpful to people [5]. The expansion of data and the advancement of technological environment, and people are demanding more and more advanced information processing, such as online decision-making and analysis, especially in telecom industry , With the monopoly of the telecommunications market to break the pattern of more intense market competition, customer service, increasing quality requirements, these factors make the domestic telecommunications companies began to build the data warehouse, and data mining system as part of the implementation of data warehousing [6, 7].

In this paper, the original data is cleaned up and the data is discretized. The attribute reduction is based on the consistency of the decision table, and the minimum attribute set is determined by judging the compatibility of the decision table after removing the attribute. Each recorded record is encoded as an individual of the initial population in the genetic algorithm with the records of the reduced information table. The individuals are selected according to the fitness of the individuals in the problem domain and the crosses and variants are combined by means of the genetic operators of the natural genetics to produce the population representing the new set. When the current solution satisfies the requirement or the evolution process reaches a certain algebra, the computation ends, and then decodes to get the final rule set.

1. Rough set theory and genetic algorithm

Rough set theory abstracts the objective world or object world into an information table system, that is, attribute-value system [8]. The information table system S can be described by a four-tuple:

$$S = <U, A, V, f> \quad (1)$$

U: is a non-empty set, is a collection of objects or instances, said domain. That all the objects in the database or tuple, set n objects, then U can be expressed as:

$$R = \{x_1, x_2, x_3, x_4, \cdots\cdots, x_n\} \quad (2)$$

Information table table is a property, a total of m columns, table rows are objects, a total of n rows, the first row i jth column of the table is f (ei, aj). A row in the table represents all the information about the object in the information table, as shown in Table 1.

Table 1. Information table

R	a1	a2	a3
e1	v11	v20	v32
e2	v10	v21	v30
e3	v12	v20	v30
e4	v11	v21	v30
e5	v11	v20	v32
e6	v12	v20	v30
e7	v10	v21	v31
e8	v11	v21	v31
e9	v11	v20	v32
e10	v10	v21	v31

Attribute dependency: there are two sets of attributes C and D, then the dependence of D on C is defined as K:

$$K = \gamma(C, D) = \sum_{X \subset U/D} \frac{|C*(X)|}{|U|} \quad (3)$$

Where C * (X) denotes the lower approximation of the set X on the attribute set C, and U / D denotes the partition class of the domain U based on the attribute set D.

Commonly used random search algorithms include genetic algorithms, simulated annealing algorithm. Because these algorithms do not depend on gradient information, their application range is wide, and their searching process is non-deterministic. The algorithm has limitation on the objective function and constraint function, so it has wide application range and the algorithm has the ability to avoid falling into the local optimum point, gradually converges to the global optimal solution. Because these algorithms are easy to be mixed with other techniques, they are widely used in the fields of optimization, machine learning, neural network and parallel processing, especially for large-scale and complicated problems that can not be solved by traditional search methods. While improving the performance of other search technology is also very good.

2. Association rules mining

Assume a set: I = {I_1, I_2, ..., I_m}, where I_1, I_2,. . . , I_m are the elements of the set, the transaction database for the D (Database), and all transactions in the database t (Transaction) are non-empty subset of I, and all transactions have TID (Transaction ID). Each transaction t in the data elements are I subset, if the item set for the k elements, known as the k item set. In D, the association rule support is the database D all the transactions also contain the proportion of X, Y, namely:

$$Support(X \geq Y) = \frac{count(X \cap Y)}{D} \quad (4)$$

Where X and Y are all non-empty subsets, and count (X ∩ Y) is the number of transactions X and Y co-occurring in database D. In the transaction database D, the association rule confidence (Confidence) is D, already contains the X case, including the proportion of Y, namely:

$$Confidence\left(X \geq Y\right)=\frac{count\left(X \cap Y\right)}{count\left(X\right)} \tag{5}$$

In all kinds of business information systems or websites, we want to find the preference relation of a user, that is, focus on the relationship between users and data items. In collaborative filtering recommendation technology, the utility matrix model is used to describe users and data items. Matrix, as in matrix (3):

$$Utility\ Matrix=\begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1n} \\ v_{12} & v_{22} & \cdots & v_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ v_{m1} & v_{m2} & \cdots & v_{mn} \end{bmatrix}_{m\times n} \tag{6}$$

3. Experiment and results

In the decision table, the condition attribute with many attribute values may be too small and the scope of the merge is relatively large. Combining these attribute values can quickly reduce the decision table and reduce the number of operations. As the decision rules are not as small as possible, for those decision-making rules are not many requirements of the feedback speed is not high production control, coarsening algorithm is not the best choice. However, for those conditions of property, and the required feedback speed is also high production control, coarsening algorithm is more effective. Attribute reduction algorithm flow chart was shown in Figure 1.

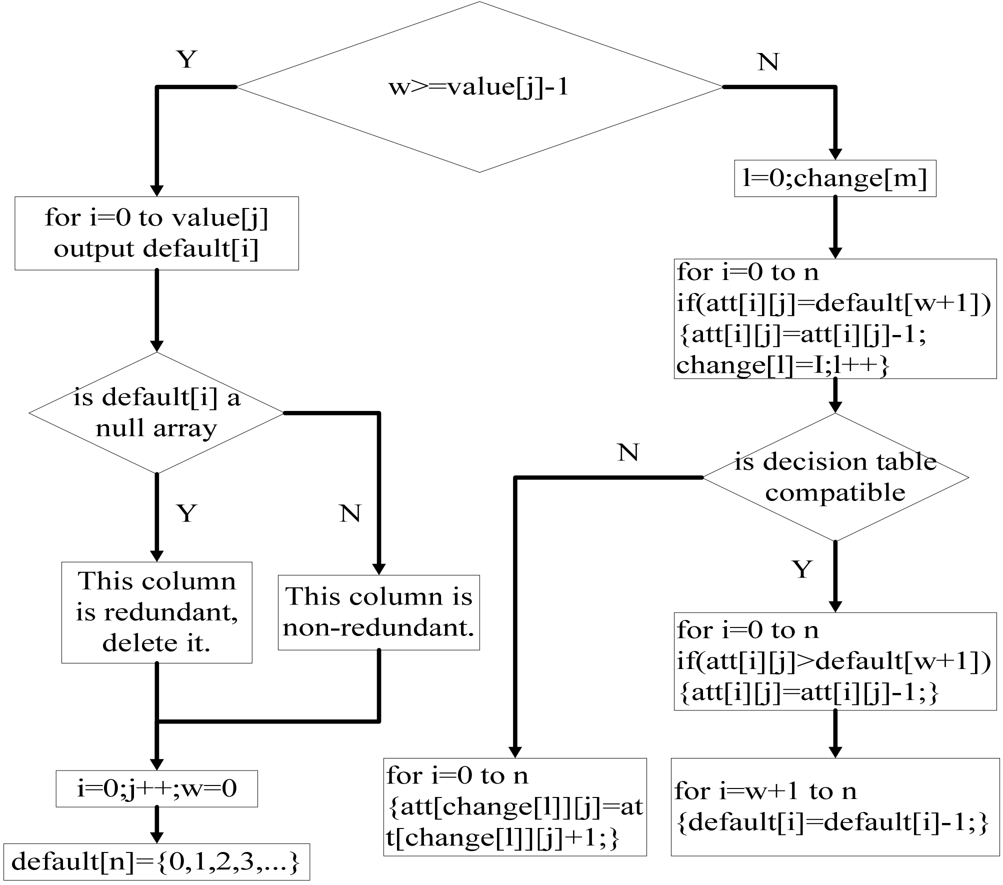
w>=value[j]-1
Y
N
for i=0 to value[j] output default[i]
l=0;change[m]
for i=0 to n
if(att[i][j]=default[w+1])
{att[i][j]=att[i][j]-1;
change[l]=I;l++}
is default[i] a null array
Y
N
is decision table compatible
N
Y
This column is redundant, delete it.
This column is non-redundant.
for i=0 to n
if(att[i][j]>default[w+1])
{att[i][j]=att[i][j]-1;}
i=0;j++;w=0
for i=0 to n
{att[change[l]][j]=att[change[l]][j]+1;}
for i=w+1 to n
{default[i]=default[i]-1;}
default[n]={0,1,2,3,...}

Figure 1. Attribute reduction algorithm flow chart

After reduction by rough set algorithm, the records of the information table are coded as the individuals of the initial population in the genetic algorithm. The individuals are selected according to the fitness of the individuals in the problem domain and the crosses and variants are combined by means of the genetic operators of the natural genetics to produce the population representing the new set. When the current solution satisfies the requirement or the evolution process reaches certain algebra, the computation ends, and then decodes to get the final rule set. The algorithm execution time was shown in Table 2.

Table 2. Algorithm performance time data

Serial No.	Data Size	Record counts	Apriori Algorithm	Parallel Algorithm
1	4003KB	1.667015*105	47.215s	121.335s
2	8005KB	3.479981*105	128.785s	160.124s
3	12009KB	4.854795*105	148.687s	224.254s
4	16012KB	7.254872*105	309.124s	377.158s
5	20147KB	9.543478*105	410.201s	490.162s
6	24018KB	10.000245*105	620.201s	533.188s

Summary

In this paper, the original data is cleaned up and the data is discretized. The attribute reduction is based on the consistency of the decision table, and the minimum attribute set is determined by judging the compatibility of the decision table after removing the attribute. Each recorded record is encoded as an individual of the initial population in the genetic algorithm with the records of the reduced information table. The individuals are selected according to the fitness of the individuals in the problem domain and the crosses and variants are combined by means of the genetic operators of the natural genetics to produce the population representing the new set.

Acknowledgement

This work was supported by Science and Technology Research Project in Department of Education, Jiangxi Province, 2015 (GJJ151274), Jiangxi Provincial Intellectual Property Soft Science Research Project (ZR201610), Humanities and Social Sciences Research Project of Jiangxi Province, 2015 (TQ1516), Art Science Planning Project of Jiangxi Province (YG2014255, YG2015189) and Pingxiang Science and Technology Support Program, 2015: Research and Development of Animation Rendering Service Platform in the Central Region Based on Cloud Computing.

References

[1] García S, Ramírez-Gallego S, Luengo J, et al. Big data preprocessing: methods and prospects[J]. Big Data Analytics, 2016, 1(1): 9.

[2] Chen F, Deng P, Wan J, et al. Data mining for the internet of things: literature review and challenges[J]. International Journal of Distributed Sensor Networks, 2015, 2015: 12.

[3] Cios K J, Pedrycz W, Swiniarski R W. Data mining methods for knowledge discovery[M]. Springer Science & Business Media, 2012.

[4] Riza L S, Janusz A, Bergmeir C, et al. Implementing algorithms of rough set theory and fuzzy rough set theory in the R package “roughsets”[J]. Information Sciences, 2014, 287: 68-89.

[5] J.van de Beek,O. Edfors, M. Sandell, S. Wilson, P. Borjesson, “On Channel Estimation in OFDM System”, in Proceedings of the IEEE Vehicular Technology Conference, pp. 815-819, 1995.

[6] Tomar D, Agarwal S. A survey on Data Mining approaches for Healthcare[J]. International Journal of Bio-Science and Bio-Technology, 2013, 5(5): 241-266.

[7] Cao J, Cui H, Shi H, et al. Big Data: A Parallel Particle Swarm Optimization-Back-Propagation Neural Network Algorithm Based on MapReduce[J]. PloS one, 2016, 11(6): e0157551.

[8] Krishnanand K N, Ghose D. Theoretical foundations for rendezvous of glowworm-inspired agent swarms at multiple locations [J]. Robotics and Autonomous Systems, 2008, 56(7): 549-569.