The study of the removal redundant association rules based on Hypergraph in Large data environment

Li Xin-liang[1, a] Liu Li-yun[2,b]

[1,2]Loudi Vocational and Technical College, Loudi , Hunan 417000

[a]499514310@qq.com, [b] liuliyun1975@163.com

Keywords: Large data; hypergraph; redundant association rules

Abstract. This paper takes the high dimensional data for large data mining technology as the research object, using the adjacency matrix and directed hypergraph detection relationship between association rule items, explores the method, classification based on spanning tree removal algorithm for detection of large data condition redundant association rules, the algorithm can effectively improve the efficiency of association rule mining, the need to reduce the actual processing time.

1. Introduction

The arrival of the era of big data, the traditional data processing methods have been faced with new challenges, only for the development of information technology, to will bring a big data era challenge change into opportunities, in order to better use resources strategy, truly massive data for effective information; with the rise of the “intellectual economy”, collecting data, master data, using the data will become the core competitiveness of state, enterprises and units（Cui and Yang. (2010)）。

2. The definition of big data

Several theories about big data definition, Amazon data scientists John Rosser (John Rauser) think: Big data is “any more than a computer processing ability of the huge amount of data”, Informatica, chief China advisor product but bin think “big data = huge amounts of data and complex types of data”. Wikipedia, the data is defined as one large and complex, it is difficult to use the existing database management Data set processing. Large data with a large number of (Volume), diversification (Variety), fast (Velocity) and low value density (Value), etc.(Li . (2013))。

3. Hypergraph

Hypergraph is a system of subsets of a finite set, is a generalization of graph theory, the word “hypergraph” is first put forward by Berge in 1966 monograph “hyper graphs”, the hypergraph has in circuit partitioning, knowledge representation and organization methods, cellular communication system and expression of non typical molecular structure of chemical compounds and polycyclic conjugated molecules have a wide range of applications, hypergraph into directed hypergraph and undirected hypergraphs (Jiawei Han and Micheline Kamber.(2011))。

3.1 Non directed

Hypergraph is a dual of H = (V, E), which {V1,V2,V3 ,...,Vn }hypergraph on n vertices and E ={e1,e2,e3 ,...,em }, Said hypergraph m ultra edge. Super edge set E is defined on the vertex set subset of V, namely, That is ∪ej∈V，j=1,2,3, ...m，,and meet the..:

(1) ej≠∮;(2)∪ej=V

Hypergraph H can be represented by a graph, namely a collection of said elements V, connected with a line of dots represent the relationships between elements. The corresponding relationship between hypergraph and adjacency matrix..

3.2 Directed hypergraph

Directed hypergraph in undirected hyperedges hypergraph add to that direction, super edge vertices sequence. It is defined as: directed hypergraph is H= (V, E) to two yuan, V is the vertex set, E is set to super edge. To super edge e in E is defined as a sequence of (T (e)), the T (e) are subset of V, and T (e) ∩ H(e) = ∮， T（e）, T(e) said to hyper edges to the end node, H (e) known as to the hyper edges to the head node[4]. This paper attempts to use to hypergraphs to association rules are represented, and the association rules mining in the presence of redundancy and loop problem to the properties of the hypergraph solution.

4. The association rules of directed hypergraph

4.1 Association Rule Mining

In the era of big data, the data set is a large and high dimensional data set, which is the goal of data mining researchers. Association rules can be expressed as: R :X → Y， which X∈I ,Y ∈I， and X∩ Y≠ ∮， It said if the itemset X in a transaction, will inevitably lead to y will also appear in the same transaction. X said the prerequisite for the rule (the Y), known as the result of a rule (back).

The task of association rule mining is to find the strong association rules between the minimum support degree and the minimum confidence level in the database D. Strong association rules R :X → Y corresponds to the set X∪Y must be frequent item set, and frequent itemset X∪Y derived association rules R :X → Y confidence degree and by the frequent item sets X and X∪Y support degree calculation. Therefore, association rule mining can be decomposed into two steps:

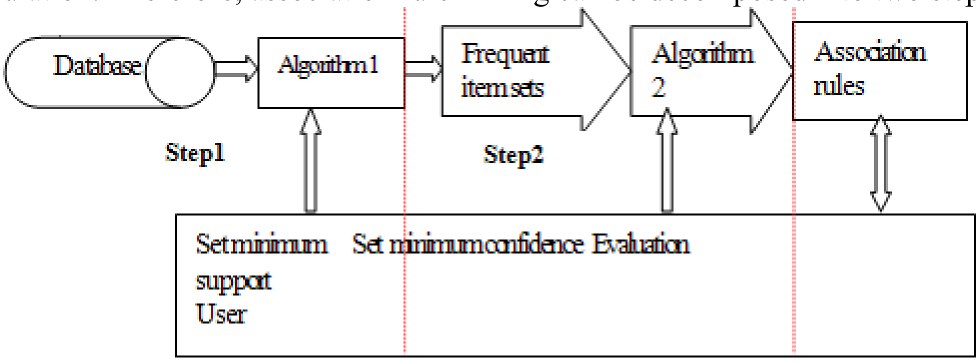

Fig.1 basic model of association rule mining

(1) a frequent item set which meets the minimum support degree given by the user from the database;

(2) an association rule generating all minimum confidence of a given minimum reliability for a user based on the frequent itemsets.

4.2 directed hypergraph on the association rules

In to the hypergraph, to the super edge e in E is defined as the pretext node H (e) and tail node of T (e) consisting of an ordered pair, and H (e), T (e) for the vertex set V subset, which can consist of a collection of vertices in. This feature is conducive to association rules expressed as a directed hypergraph. A set of M corresponding to the node corresponding to the node according to the definition of the first H (e) and association rule consequent, tail T(e) and association rules referred to in the preceding paragraph, and to ensure that each association rules are uniquely expressed as a directed hypergraph in a hyper edge(J.Zaki(2010)) association rules: $X_1,X_2,X_3,\ldots,X_m \rightarrow Y_1,Y_2,Y_3,\ldots,Y_n$, is referred to in the preceding paragraph Ante(R) is composed of a plurality of components, the consequent Cons (R) also contains more than one item, the definition of the consequent contains only the rules for simple rules, the latter contains a rule known as complex rules(J.Zaki(2010)). In this paper, we define a to super edge said a association rules, association rules in each referred to in the preceding paragraph shall be corresponding to a to the head of the hypergraph node, corresponding to the consequent of association rules are of the same to the tail node of hypergraph, each to the hyper edges to the head node and end node are multiple, such as this that the composite rules.

4.3 Association rule redundancy reconstruction based on directed hypergraph

Redundant rules has two kinds of forms: the dependency rule, namely rule Xi and Xj the same conclusion, and the premise of Xi is a sufficient condition for the Xj premise, Xj is redundant, repeat visual for the special case of dependency rule; second is repeated path rules, if in the rule base choice Xi, Xj, and between Xi and Xj exist at least two paths, to determine the existence of redundant rules。

Due to the hypergraph of the adjacency matrix is mainly used in the simple graph and directed hypergraph is composite, the composite rules can only be used to the hypergraph representation. Therefore, it is necessary to redefine the adjacency matrix. The adjacency matrix of complex and simple as a row or column, information indicative of the association rules, while ensuring that the adjacency matrix is a 0-1 matrix in order to eliminate the non 0-1 matrix algorithm to deal with the difficulty and the expression of ambiguity, and the algorithm to remove redundant after the adjacency matrix can be simply reduced to association rules.

5. Research on the algorithm of classification and removal of redundant association rules based on minimum spanning tree

5.1 Concept of minimum spanning tree

In figure H = (V, E), (u,v) represents the connection vertex u and vertex v, namely (u, v) ∈E, and W (u, v) said the weight of the edge. If T is a subset of E, namely T∈E and a graph without loops, the W(T) =∑W(u,v) W(T) minimum, said this T H minimum spanning tree. The minimum spanning tree is the abbreviation of the minimum weight spanning tree.

In figure H = (V, E), (u,v) represents the connection vertex u and vertex v, namely (u, v), E, and W(u, v) said the weight of the edge. If T is a subset of E, namely T∈E and a graph without loops, the W (T) = =∑W(u,v) of W(T) minimum, said this T is the minimum spanning tree. The minimum spanning tree is the abbreviation of the minimum weight spanning tree(Xu Zipei(2012)).

5.2. Detection redundancy process of association rules

Based on the classification of the spanning tree, the redundancy method of association rules is shown:

(1) of experimental data scanning and generate association rules with a directed hypergraph representing the association rules, re definition and its adjacency matrix.

(2) the adjacency matrix is obtained by removing the subordinate rule algorithm.

(3) a spanning tree is obtained by means of spanning tree algorithm.

(4) the resulting tree adjacency matrix is reduced to the corresponding association rule, and the final processing result is obtained.

5.3 Research on the algorithm of removing the subordinate rule by the method of defining the adjacency matrix

Redefine the adjacency matrix is becomes the every article of association rules definition to one side of a hypergraph, the adjacency matrix of the said association rules referred to in the preceding paragraph, the columns of the adjacency matrix representation of association rule consequent. The algorithm for removing the subordinate rules of the adjacency matrix is as follows:

(1) For adjacency matrix per column (association rule consequent), the query is 1, 1 corresponds to a location (association rules referred to in the preceding paragraph) for intersection operation, if there's an intersection and the corresponding (association rules referred to in the preceding paragraph) contains a bulls for 1 position is set to 0. Similarly, for each row as such a treatment.

(2)If the row and column values of a particular item are all 0, if it is, then delete the row and column.

(3) Finally, for rows and columns respectively traverse from the end, if a column / row is 0, delete the column / row, until I met 1 column / row exists.

(4) Output matrix, delete the dependency rule after the adjacency matrix, the adjacency matrix composite nodes because and other nodes have no intersection, which makes it in subsequent

processing of independence with simple nodes as, therefore, being considered for simple node processing. Delete all the slave rules in the redundant rule.

The flow chart of the algorithm is shown in Figure 3. After this algorithm, all the subordinate rules in the redundant rule can be removed, and the pre processed adjacency matrix is obtained.

5.4 the results of association rules to remove redundant algorithm and analysis of hypergraph

Based on directed hypergraph removal method of redundant association rules mainly three modules, respectively is the redefinition of the adjacency matrix module, remove the dependency rule module and a generating module tree. The first two module through VB programming realization, spanning tree module in matlab implementation, contains the specific content such as shown in Figure 2

In the process of experiment, 2 UCI data were selected, Balloons data set and Shuttle-landing-control data set, and the minimum support and minimum confidence of Balloons data set was 5%. The data set has 4 attributes, and the Aprioir algorithm is, and the 18 association rules are obtained. The minimum support degree of Shuttle-landing-control data set is 40%, and the minimum confidence is 100%. The data set has 7 attributes, and the 15 rules are obtained by Aprioir algorithm. Balloons data set as an example, according to Article 18 of association rules obtained redefine the adjacency matrix (V2=SMALL with V2 indicating the, the other and so on).

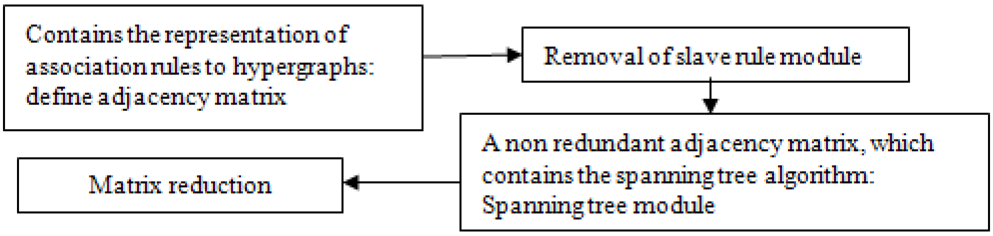
Contains the representation of association rules to hypergraphs: define adjacency matrix
Removal of slave rule module
A non redundant adjacency matrix, which contains the spanning tree algorithm: Spanning tree module
Matrix reduction

Fig 2: module composition of the method of classification and removal of redundancy

After removing the dependency rule algorithm and the spanning tree algorithm get no loop connected spanning tree, the formation of the specific process and the corresponding tree algorithm of directed hypergraph as shown in Figure 3. Figure 3 the left half is in MATLAB Nakamo Nariki algorithm of the screenshot can be seen by the pretreatment adjacency matrix (removing dependency rule algorithm results obtained have been completely removed redundancy adjacency matrix, the right part is the directed hypergraph changes in before and after the spanning tree algorithm.

```
>> w=[0 1 1 1 ;1 0 1 1 ;1 1 0 0; 1 1 0 0]
   w= 0  1  1  1
      1  0  1  1
      1  1  0  0
      1  1  0  0
   >>w1=treedgraf(w)
   w1= 0  0  0  0
       1  0  0  0
        1  0  0  0
       1  0  0  0
```

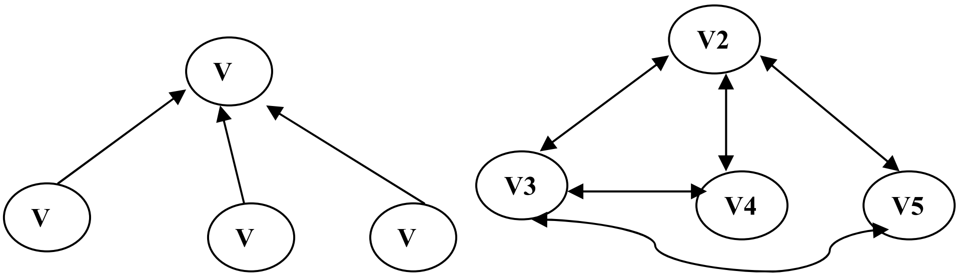

Fig3 Matlab simulation and directed hypergraph

The redundant rules can be removed by the method of minimum spanning tree classification. The results are shown in Table 1.

Table 1 the results of removing redundant rules

	Balloons	Shuttle-landing-control
Total association rules	18	15
Subordinate rule	8	9
Repeated path rules	7	2
The number of redundant rules	15	11
The number of remaining association rules	3	4

6. Summary

Under the condition of big data, the detection of redundant association rules can improve the effectiveness, the spanning tree based classification to remove the redundant association rules method, Through data analysis and verification, the algorithm can effectively improve the efficiency of association rules mining, reduce the time needed for the actual processing.

Acknowledgement

The reseach was sponsored by Hunan scientific research project (No. 15C1160)

Reference

[1] Xindong Wu, Xingquan Zhu, Gongqing Wu, Wei Ding. Data mining with big data [J]. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2014, 26(1): 97-107

[2] https://archive.ics.uci.edu/ml/datasets.html

[3] J.Zaki. Mining Non-redundant Association Rules [J]. Data Mining and Knowledge Discovery, 2010

[4]Xu Zipei. Big data is coming from the data revolution [M].2012

[5] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques,Second Edition [M]. 2011