International Journal of Database Theory and Application
Vol.9, No.4 (2016), pp.301-308
http://dx.doi.org/10.14257/ijdta.2016.9.4.28

Multi-Data Association Rule Mining Algorithm Based on Grey Relational Analysis

Hongyan Chu

Nanjing College of Information Technology, Nanjing, Jiangsu, 210023 China
chu_hy@163.com

Abstract

Mining association rules for data is not only an essential part of data mining but also a hot issue in knowledge engineering and researches on data mining technology. Since multi-data mining is characterized as being multi-type, multi-level, multi-implicational and complicated, the efficiency of multi-data association rule mining usually cannot be high, precision and accuracy are of a relatively low degree, and the targets of mining cannot be obtained quickly. Therefore, on the basis of improving traditional association rule mining algorithm, this paper researched on multi-data association rule mining algorithm and based on grey relational analysis, proposed a multi-data association rule mining algorithm. Firstly, the associate objects most relevant to the target objects are obtained through the grey relational analysis, which helps to form single- or multi-target data associates; after that, the multi-data association rule mining model which sets data associates as the new mining objects is established. Under the conditions that the level of support and confidence are met, the frequent patterns of corresponding data associates and further, the multi-data association rule, are obtained. Simulation experiments implied that the model have the advantages of simplicity, practicality, operability, decent precision and accuracy.

1. Introduction

Along with the rapid development of science technologies in modern society, information and data grows explosively. How to obtain valuable knowledge among so much information and data is an important research issue of data mining. At present, the proceeding and practicing of big data mining technologies makes it possible to mine valuable knowledge out of information and data. Series of research fruits of big data mining also came out [1-4]. However, data mining gets to be more complicated as when the information and data increases by one order of magnitude, the computational complexity grows in an exponential order. What's more, due to the sharp increase of information and data, the implication relation between information and data turns more uncertain and vague. Consequently, it is harder to obtain the effective frequent patterns between information and data [5-6]. Hence, some scholars discussed and researched on mining association rules for data and made some achievements [7-10]. But the traditional mining association rules for data still has some shortcomings in terms of obtaining the frequent patterns of big data and information, that is, it cannot find the implication relation between the two effectively and precisely. To solve this problem, this paper introduces the grey system theory and applies grey relational analysis [11-14] to find the associate objects most relevant to the target objects. Thus single- or multi-target data associates are formed, in other words, the substrate of frequent patterns is obtained. Then, the substrate of frequent patterns is used to analyze association rules and further, achieve multi-data association mining.

ISSN: 2005-4270 IJDTA

2. The Obtainment of Data Associates Based on Grey Relational Analysis

Grey relational analysis is an intelligent approach of decision analysis carried out through analyzing and determining the degree of influence among factors or factors' scale of contribution to the main behavior. It measures the association degree of factors according to the similarity or differentiation of developing tendencies among factors, in order to obtain the information of association and implication among different factors. Grey relational analysis does not rely on the amount of data or information, so it can analyze small samples which have poor information and regularity with high efficiency. The application of grey relational analysis in multi-data mining association rules helps to determine the associates relevant to the target mining objects quickly and also the mining direction of multi-data mining association rules. Meanwhile, mining frequent patterns on the basis of associates significantly improves the efficiency and precision of data mining. Besides, the convergence of multi-data mining is guaranteed since grey relational analysis is essentially a decision analysis based on knowledge because it is carried out through comparing the indeterminacy association among factors or between systems factors and main behavior factors. The data associates gained though grey relational analysis can be classified into single-target data associates and multi-target data associates.

2.1. The Generation of Single-Target Data Associates

Single-target data associates suggest that the target mining objects have one-to-one association relations with the implication information. Considering that different analysis objects and target objects may possibly have different measure standards and dimensions, normalization of different types of analysis objects and target objects is required.

Assume the analysis object i corresponds to the numerical value $c(i)$. If the analysis object has forward characteristic, that is to say, if the bigger $c(i)$ is, the better, the normalized numerical value is $u(i)$:

$$u(i)=\frac{c(i)-min(c(i))}{max(c(i))-min(c(i))} \tag{1}$$

If the analysis object i has reverse characteristic, that is to say, if the smaller $c(i)$ is, the better, the normalized numerical value is $u(i)$:

$$u(i)=\frac{max(c(i))-c(i)}{max(c(i))-min(c(i))} \tag{2}$$

If select analysis objects i of m time nodes to carry out association analysis with target objects, the relational coefficient ζ_{ij} between the analysis object i at the j th time node and target object is:

$$\zeta_{ij}=\frac{\min\limits_{i}\min\limits_{j}\left|u_o(ij)-u(ij)\right|+\beta\max\limits_{i}\max\limits_{j}\left|u_o(ij)-u(ij)\right|}{\left|u_o(ij)-u(ij)\right|+\beta\max\limits_{i}\max\limits_{j}\left|u_o(ij)-u(ij)\right|} \tag{3}$$

Where β is the resolution ratio which is generally set at 0.5.

If the analysis object i has fuzzy value, that is, $c(i)=\left[c_1(i),c_2(i)\right]$, the formula (1) can be changed into:

$$u(i)=\left[u_1(i),u_2(i)\right]=\left[\frac{c_1(i)-min\left(c(i)\right)}{max\left(c(i)\right)-min\left(c(i)\right)},\frac{c_2(i)-min\left(c(i)\right)}{max\left(c(i)\right)-min\left(c(i)\right)}\right] \tag{4}$$

The formula (2) can be changed into:

$$u(i)=\left[u_1(i),u_2(i)\right]=\left[\frac{max\left(c(i)\right)-c_2(i)}{max\left(c(i)\right)-min\left(c(i)\right)},\frac{max\left(c(i)\right)-c_1(i)}{max\left(c(i)\right)-min\left(c(i)\right)}\right] \tag{5}$$

If select analysis objects i of m time nodes to carry out association analysis with target objects, the relational coefficient ζ_{ij} between the analysis object i at the j th time node and target object is:

$$\zeta_{ij}=\frac{\min\limits_i\min\limits_j\left|d\left(u_o(ij)-u(ij)\right)\right|+\beta\max\limits_i\max\limits_j\left|d\left(u_o(ij)-u(ij)\right)\right|}{\left|d\left(u_o(ij)-u(ij)\right)\right|+\beta\max\limits_i\max\limits_j\left|d\left(u_o(ij)-u(ij)\right)\right|} \tag{6}$$

Where $d\left(u_o(ij)-u(ij)\right)$ is the distance between the analysis object i at the j th time node and target object.

Based on this, the grey relational degree ρ_i between analysis object i and the target object is:

$$\rho_i=\frac{1}{m}\sum_{j=1}^{m}\zeta_{ij} \tag{7}$$

2.2. The Generation of Multi-Target Data Associates

Multi-target data associates suggest that the target mining objects have many-to-many association relations with the implication information. Same as the generation of single-target data associates, considering that different analysis objects and target objects may possibly have different measure standards and dimensions, normalization of different types of analysis objects and target objects is required. For the convenience of analysis, this paper determines analysis objects' numerical value from generalized value angle.

Assume that the generalized analysis object i corresponds to the generalized numerical value $c^{\Gamma}(i)=\left(\Gamma_1(i),\Gamma_2(i),\cdots,\Gamma_s(i)\right)$. If the generalized analysis object has forward characteristic, that is to say, if the bigger $c^{\Gamma}(i)$ is, the better, the normalized numerical value is $c^{\Gamma}(i)$:

$$u^{\Gamma}(i)=\frac{c^{\Gamma}(i)-min\left(c^{\Gamma}(i)\right)}{max\left(c^{\Gamma}(i)\right)-min\left(c^{\Gamma}(i)\right)} \tag{8}$$

If the generalized analysis object i has reverse characteristic, that is to say, if the smaller $c^{\Gamma}(i)$ is, the better, the normalized numerical value is $u^{\Gamma}(i)$:

$$u^{\Gamma}(i)=\frac{max\left(c^{\Gamma}(i)\right)-c^{\Gamma}(i)}{max\left(c^{\Gamma}(i)\right)-min\left(c^{\Gamma}(i)\right)} \tag{9}$$

If select generalized analysis objects i of m time nodes to carry out association analysis with target objects, the relational coefficient ζ_{ij}^{Γ} between the generalized analysis object i at the j th time node and target object is:

$$\zeta_{ij}^{\Gamma}=\frac{\min\limits_{i}\min\limits_{j}\left|u_o^{\Gamma}(ij)-u^{\Gamma}(ij)\right|+\beta\max\limits_{i}\max\limits_{j}\left|u_o^{\Gamma}(ij)-u^{\Gamma}(ij)\right|}{\left|u_o^{\Gamma}(ij)-u^{\Gamma}(ij)\right|+\beta\max\limits_{i}\max\limits_{j}\left|u_o^{\Gamma}(ij)-u^{\Gamma}(ij)\right|} \tag{10}$$

Where β is the resolution ratio which is generally set at 0.5.

Similarly, if the generalized analysis object i has fuzzy value, select generalized analysis objects i of m time nodes to carry out association analysis with target objects, and the relational coefficient ζ_{ij}^{Γ} between the generalized analysis object i at the j th time node and target object is:

$$\zeta_{ij}^{\Gamma}=\frac{\min\limits_{i}\min\limits_{j}\left|d\left(u_o^{\Gamma}(ij)-u^{\Gamma}(ij)\right)\right|+\beta\max\limits_{i}\max\limits_{j}\left|d\left(u_o^{\Gamma}(ij)-u^{\Gamma}(ij)\right)\right|}{\left|d\left(u_o^{\Gamma}(ij)-u^{\Gamma}(ij)\right)\right|+\beta\max\limits_{i}\max\limits_{j}\left|d\left(u_o^{\Gamma}(ij)-u^{\Gamma}(ij)\right)\right|} \tag{11}$$

Where $d\left(u_o^{\Gamma}(ij)-u^{\Gamma}(ij)\right)$ is the distance between the generalized analysis object i at the j th time node and target object.

Based on this, the grey relational degree ρ_i^{Γ} between generalized analysis object i and the target object is:

$$\rho_i^{\Gamma}=\frac{1}{m}\sum_{j=1}^{m}\zeta_{ij}^{\Gamma} \tag{12}$$

3. The Mining Association Rules for Data Based on the Substrate of Frequent Patterns

3.1. The Basic Definitions of Substrate of Frequent Patterns

Definition 1 Substrate of frequent patterns: The implication associates formed by target objects and the most relevant analysis objects are named substrates of frequent patterns. They are the basic units of association rules mining together with other data.

Definition 2 Assume that the transaction set $S=\{T\mid T\subseteq\Omega\}$ exists, where T is the transaction recorded by a set of data or associates, $|S|$ represents the number of transactions in the transaction set S, $\Omega=\{T\mid T=(N_T,C_T,V_T)\}$ is the record set of data or associates. The association rules mined based on the transaction set S is named implication association rules.

Definition 3 Assume that $A(T)$ and $B(T)$ are both the subsets of data's or associates' records and the condition that $(A(T)\subseteq\Omega)\cap(B(T)\subseteq\Omega)=\varnothing$ is met, then the implication association rules can be expressed as the implication formula $(A(T)\subseteq\Omega)\Rightarrow(B(T)\subseteq\Omega)$. If and only if $A(T)\subseteq T$, transaction T supports $A(T)$.

Definition 4 If $|A(T)|$ represents the number of transactions which belong to S and support $A(T)$, $|(A(T)\subseteq\Omega)\cup(B(T)\subseteq\Omega)|$ represents the number of transactions which belong to S and support $A(T)$ and $B(T)$ at the same time. The level of support κ_{sup} of $A(T)\Rightarrow B(T)$ is

$$\kappa_{sup}=|(A(T)\subseteq\Omega)\cup(B(T)\subseteq\Omega)|/|S| \tag{13}$$

The level of confidence κ_{con} is

$$\kappa_{con}=|(A(T)\subseteq\Omega)\cup(B(T)\subseteq\Omega)|/|A(T)\subseteq\Omega| \tag{14}$$

From the definition 4 it can be seen that if both the level of support κ_{sup} and the level of confidence κ_{con} meet the condition of minimum threshold value, that is,

$$(\kappa_{sup}\geq min(\kappa_{sup}))\wedge(\kappa_{con}\geq min(\kappa_{con})) \tag{15}$$

The implication rules is strong association rules; otherwise, it is weak association rules.

3.2. The Improved Multi-Data Association Rule Mining Algorithm

On the basis that the multi-data associates have been obtained, through establishing the frequent patterns tree (IFP-tree) corresponding to the transaction set S and mining the IFP-tree, the frequent patterns mined based upon collective multi-data association rules can be obtained. The improved multi-data association rule mining algorithm is described as what follows:

Step 1 Mark the data records, associates as well as corresponding data items in the transaction set S. Randomly select one record from the transaction set S as the initial tree node of IFP-tree establishment;

Step 2 In the established hierarchical IFP-tree, generate parent nodes set and child nodes set of each node and give pointers of corresponding nodes;

Step 3 Carry out frequent patterns search in parent nodes set and select the data items or associate items in parent nodes set according to the stack order. Mark combinatorials through pointers and subnodes and get the corresponding records of combinatorials;

Step 4 Determine whether the parent nodes set is empty or not. If yes, carry out frequent patterns search in child nodes set; if not, repeat step 3-step 4;

Step 5 Determine whether subnode is leaf node. If yes, do not carry out frequent patterns search in next hierarchy and stop the search; if not, repeat step3-step 5;

Step 6 Combine the node combinatorials that are searched out and have the same marks. Gain the data combinatorials and corresponding recorded values;

Step 7 Considering the condition of minimum threshold value of the level of support and the level of confidence, get the frequent patterns that meet the condition, that is to say, establish the association rules based on data items.

3.3. Realization of the Multi-Data Association Rule Mining Model Based on Grey Relational Analysis

To sum up, multi-data association rule mining model based on grey relational analysis can be built up by following steps:

Step 1 Select the target associate objects from multi-data sets which will go through association rules analysis;

Step 2 Normalize the target associate objects and analysis objects and get associated data with normalized scale;

Step 3 Based on the discussions in 2.1, carry out grey relational analysis on target associate objects and analysis objects and get single-target multi-data associates;

Step 4 Based on the discussions in 2.2, carry out grey relational analysis on target associate objects and analysis objects and get multi-target multi-data associates;

Step 5 Establish the IFP-tree of data records or associates corresponding to target associate objects and analysis objects;

Step 6 Carry out multi-data association rule mining algorithm on the established IFP-tree and get the implication association rules that meets conditions.

4. Verification and Analysis of Algorithm

This paper used sales data analysis of certain superstore's brand products as an example to explain the algorithm. The superstore launched serialized brand products and sets. In order to the association features between the brand products and By-products, this paper collected and analyzed the sales data during three months in selling season. After normalization, the initial mined data is shown in Table 1.

Table 1. The Initial Mined Data of Sales Data

Brand PA	Set T1	Set T2	Set T3	By-product S1	By-product S2	By-product S3	By-product S4	By-product S5
1.000	0.783	0.614	0.492	0.836	0.717	0.435	0.902	0.876
1.000	0.668	0.605	0.326	0.803	0.781	0.324	0.825	0.732
1.000	0.738	0.546	0.475	0.892	0.696	0.451	0.768	0.627

Based on grey relational analysis, the grey relational coefficient and grey relational degree between various set as well as By-products and brand PA were obtained. The results are shown in Table 2.

Table 2. The Grey Relational Coefficient and Grey Relational Degree Between Various Set as Well as by-Products and Brand PA

	Set T1	Set T2	Set T3	By-product S1	By-product S2	By-product S3	By-product S4	By-product S5
relational coefficient	0.786	0.602	0.515	0.869	0.702	0.483	1.000	0.943
	0.651	0.595	0.431	0.815	0.783	0.430	0.850	0.719
	0.727	0.551	0.571	0.978	0.679	0.492	0.765	0.613
relational degree	0.721	0.583	0.506	0.887	0.721	0.468	0.872	0.758

After calculation of relational degree, it can be known that By-product S1 and By-product S4 are most relevant to the brand product PA. Therefore, the initial implication association rules can be obtained: $PA \Rightarrow S1$ and $PA \Rightarrow S4$. Since By-product S1 and By-product S4 have similar relational degree with the brand product

PA, the expanded association rules can be obtained: $PA \Rightarrow (S1 \cup S4)$. This paper adopts Lenovo ThinkPad new X1 carbon, processor Intel Core i5-5200U with memory of 4G and database SQL Server 2000 to mine and analyze 1246908 sales data records of the superstore. The results are shown in Table 3.

Table 3. The Mining and Analysis of Association Rules

Association Rules	level of support	level of confidence	Time(s)		Number of Records	Data Items
			improved algorithm	classical algorithm		
$PA \Rightarrow S1$	0.020	0.668	87.4	115.8	24688	$S1$
$PA \Rightarrow S4$	0.018	0.596			22046	$S4$
$PA \Rightarrow (S1 \cup S4)$	0.013	0.445			16412	$S1 \cup S4$
					36973	PA

From the calculation results of data mining, it can be seen that the improved algorithm given in this paper can effectively improve the efficiency of data mining.

5. Conclusions

Considering the problems existing in multi-data mining association rules, this paper proposes a multi-data association rules mining algorithm based on grey relational analysis. At first, the algorithm get the associate objects most relevant to the target objects through grey relational analysis, and then uses the associate objects as the database for search of frequent patterns' substrate. Finally, based upon the improved association rules algorithm, the multi-data association rules that meets the condition of minimum threshold value of the level of support and the level of confidence is obtained. Example verified that the algorithm has simple calculation, high efficiency and relatively satisfactory precision and accuracy.

References

[1] L. Li, X. Su, Y. Wang, Y. Lin, Z. Li and Y. Li, "Robust causal dependence mining in big data network and its application to traffic flow predictions", Transportation Research Part C: Emerging Technologies, vol. 58, **(2015)**, pp. 292-307.
[2] E. J. Khatib, R. Barco, A. G. Andrades, P. Muñoz and I. Serrano, "Data mining for fuzzy diagnosis systems in LTE networks", Expert Systems with Applications, vol. 42, no. 21, **(2015)**, pp. 7549-7559.
[3] A. Weichselbraun, S. Gindl and A. Scharl, "Enriching semantic knowledge bases for opinion mining in big data applications", Knowledge-Based Systems, vol. 69, **(2014)**, pp. 78-85.
[4] P. Perner, "Mining Sparse and Big Data by Case-based Reasoning", Proceeding Computer Science, vol. 35, **(2014)**, pp. 19-33.
[5] G. Czibula, I. G. Czibula, A. M. Sîrbu and I. G. Mircea, "A novel approach to adaptive relational association rule mining", Applied Soft Computing, vol. 36, no. 11, **(2015)**, pp. 519-533.
[6] D. Nguyen, L. T. T. Nguyen, B. Vo and T. P. Hong, "A novel method for constrained class association rule mining", Information Sciences, vol. 320, no. 11, **(2015)**, pp. 107-125.
[7] J. Sahoo, A. K. Das and A. Goswami, "An efficient approach for mining association rules from high utility itemsets", Expert Systems with Applications, vol. 42, no. 13, **(2015)**, pp. 5754-5778.
[8] Ö. M. Soysal, "Association rule mining with mostly associated sequential patterns", Expert Systems with Applications, vol. 42, no. 5, **(2015)**, pp. 2582-2592.
[9] L. Vu and G. Alaghband, "Novel parallel method for association rule mining on multi-core shared memory systems", Parallel Computing, vol. 40, no. 10, **(2014)**, pp. 768-785.
[10] J. Lai, Y. Li, R. H. Deng, J. Weng, C. Guan and Q. Yan, "Towards semantically secure outsourcing of association rule mining on categorical data", Information Sciences, vol. 267, **(2014)**, pp. 267-286.

[11] X. Li, K. W. Hipel and Y. Dang, "An improved grey relational analysis approach for panel data clustering", Expert Systems with Applications, vol. 42, no. 23, **(2015)**, pp. 9105-9116.
[12] H. V. Trivedi and J. K. Singh, "Application of Grey System Theory in the Development of a Runoff Prediction Model", Biosystems Engineering, vol. 92, no. 4, **(2005)**, pp. 521-526.
[13] L. Ke, S. Xiaoliu, T. Zhongfu and G. Wenyan, "Grey Clustering Analysis Method for Overseas Energy Project Investment Risk Decision", Systems Engineering Proceeding, vol. 3, **(2012)**, pp. 55-62.
[14] T. F. Chuang and Y. H. Chang, "Comparison of physical characteristics between Rana latouchtii and Rana adenopleura using grey system theory and Artificial Neural Network", Ecological Engineering, vol. 68, **(2014)**, pp. 223-232.

