PROCEEDINGS OF SCIENCE

A Simple Methodology for Database Clustering

Hao Tang[1,2]
Guangdong University of Technology, Guangdong, 201503, China
E-mail: `122260851@qq.com`

Mei Zhang
Guangdong University of Technology, Guangdong, 201503, China
E-mail: `646054552@qq.com`

Database clustering is a preprocess technology for multi-database mining. Existing algorithms for database clustering are successful in terms of having a cluster, but the time complexity is high or excessive pursuit in respect of the non-trivial complete clustering, which may lead to a bad clustering or application-dependent. The practical application of these algorithms have high time instability and complexity. In this article, we put forward the application-independent database clustering methodology by using hierarchical clustering method to avoid instability and reduce time complexity. This methodology is called Database Hierarchical Clustering. We firstly construct a multi-objective optimization problem, and then use hierarchical clustering algorithm to find the optimal cluster structure thought. We also use the cophenetic correlation coefficient to evaluate the best cluster. Experiments on the synthetic databases and the real-world databases show that our method of clustering stability features lower time complexity than that of the BestClassification while also highlighting strong generalization ability.

CENet2015
12-13 September 2015
Shanghai, China

[1] Speaker
[2] Correspongding Author

1. Introduction

As we know, there are many large companies which have different database distributions in different branches; therefore, multi-database mining is an important branch of data mining which has become more and more important. In order to reduce the search cost, we need to determine which database is associated with our data mining. This important step we call the database selection [1]. Furthermore, let's think a need to handle the multiple large database of Multi-National Corporation. This company may need to find the non-profit association analysis project (product).The ultimate goal is to identify those without much profit nor other products to promote profitable products. The correlation analysis may find such products, thereby the company may terminate the transactions of products. The nature of the analysis may need to identify similar databases. We note that, if the two databases contain many similar transactions, the two databases are considered to be similar; if the two transactions contain many of the same goods, the two transactions are similar. In this sense, the more same frequent itemsets two databases contain, the more similar they are [2]. We could cluster multi-databases according to the similarity between two databases. Then, carry out mining in the same class in the database. It is a significant procedure in terms of the analysis of exploring patterns, grouping, decision-making and machine-learning. In this article, we mainly discuss the transaction database clustering.

2. Related Works

As to the multi database mining, the first idea (single database mining) is that a number of data in the database together constitutes a single data set[1]. Liu et al. put forward a search related database of multi database mining technology[3]. Their main researches on the identification were associated with the database application.

Zhang et al.designed a new multi-database mining process to mine the multi-databases[4]. The database clustering was the first step, and then find the local mode. The paper designed a similarity by using the items of transaction or high-frequency rules. Furthermore, they put forward algorithms called *GreedyClass* and *BestClassification*, in which, a database was chose to be one class firstly and then judg whether or not the next of database is incorporated into the known classes or considered as a new class iteself. H. Li et al. [5] designed an improved method to reduce the time complexity of *BestClassification*. The improved algorithm can obtain the best classification correctly for *m* given databases and the time complexity has been reduced from $O(n^2m^2+m^4)$ to $O(n^2m^2+m^3)$.

Animesh Adhikari et al.proposed two similarity measures by the frequent item sets of databases and designed an arithmetic to cluster the data set. In order to improve the efficiency of cluster., the multiple data clustering method is proposed based on high-class cohesion and low coupling between classes of application-independent.

There are other effective researches. Wu et al.proposed a pattern recognition method of weighted by a multi database[6]. Yin and Han put forward a new strategy for high dimensional heterogeneous database, this strategy may not apply to the aggregation transaction database[7]. Yin et al.proposed two scalable high classification algorithms: CrossMine-Rule (based on the association rules) and CrossMine-Tree (based on the decision tree)[8]. Bandyopadhyay et al.,based on the aggregation of K-means algorithm suitable for the image sensor network, proposed the isomorphism of environmental data technology [9].

3. Clustering Multiple Databases

3.1 Similarity measures

Definition 1: appoint $D=\{D_1,D_2,\cdots,D_m\}$ as a set for all database. In the case of D, this similarity matrix is $DSM(D,\alpha)$ as described by the measure of similarity *sim.* This distance matrix is $DDM(D,\alpha)$ expressed by the measure of distance *1-sim*, they are square matrix, whose (i,j) element are $DSM_{i,j}(D,\alpha)=sim(D_i,D_j,\alpha)$ and $DDM_{i,j}(D,\alpha)=1-sim(D_i,D_j,\alpha)$ respectively, while $D_i,D_j \in D$, $i,j=1,2,\cdots,m$.

As to m databases, the two in one group, there are $C_m^2=\frac{1}{2}m(m-1)$ databases. As to not a group of database, we calculated the similarity of them. In case of high similarity or distance, the two database may be assigned to the same class.

Definition2: there have *m* data pool D_1, D_2,$\cdots$, D_m. *; a*ppoint *D* is a set for m while *C*={ $c_1,c_2,...,c_n$ },($1 \le n \le m$) is a candidate cluster (partition) of D_1, D_2,$\cdots$, D_m ,if

(1) $c_i \neq \Phi$, for $1 \le i \le n$.

(2) $D=c_1 \cup c_2 \cup \cdots \cup c_n$,

(3) $c_i \cap c_j = \Phi$,for $i \neq j, 1 \le i,j \le n$.

where, c_i ($1 \le i \le n$) means a class of *C*.

Definition3: given *D* is a assembly of *m* databases, that is D_1,$\cdot\cdot$,D_m. *C*={ $c_1,c_2,...,c_n$ },($1 \le n \le m$) is a candidate cluster of *D.* The similarity between databases c_i and c_j under threshold *α* is defined as follows:

$$sim(c_i,c_j,\alpha)=\frac{\left|Itemsets\left(\bigcup_{D_s \in c_i} D_s,\alpha\right) \cap Itemsets\left(\bigcup_{D_t \in c_i} D_t,\alpha\right)\right|}{\left|Itemsets\left(\bigcup_{D_s \in c_i} D_s,\alpha\right) \cup Itemsets\left(\bigcup_{D_t \in c_i} D_t,\alpha\right)\right|} \tag{3.1}$$

Definition 4: *D* specifies a dataset for *m* databases, that D_1,D_2,$\cdots$,D_m. *C*={ $c_1,c_2,...,c_n$ },($1 \le n \le m$) is a candidate cluster of *D.* The class distance matrix $CDM(C,\alpha)$ of *C* expressed by the measure of distance *1-sim*, is a n-order matrix, The first (i, j)th element $CDM_{i,j}(C,\alpha)=1-sim(c_i,c_j,\alpha)$, $i,j=1,2,\cdots,n.$

3.2 Relevance of Databases

Definition 5: *D* is a collection of databases m and *C* is a candidate cluster of *D*, *D*={D_1,D_2,$\cdots$,D_m}. Under measure *sim,* it is defined as the sum of distances as follows:

$$dist^{\alpha}(c)=\sum_{D_i,D_j \in c; i<j}(1-sim(D_i,D_j,\alpha)) \tag{3.2}$$

from the database D_i and D_j is expressed as 1−*sim*(D_i,D_j,*α*). The shorter the distance between two databases is, the higher the sum of similarity of a class will be.

Definition 6: designate a data set D, and *C*={ $c_1,c_2,...,c_n$ },($1 \le n \le m$) is a candidate cluster of *D*, *D*={D_1, D_2,$\cdots$, D_m}. The sum of distance of *C* under threshold *α* is

$$sum\text{-}dist(C,\alpha)=\sum_{c \in C} dist^{\alpha}(c) \tag{3.3}$$

The above definition is to reveal the sum of distance of a cluster based on the distance of all classes. Lower value means higher cohesion which is an important parameter to a good cluster.

Definition 7. D is a particular data set and *C*={ $c_1,c_2,...,c_n$ } is a candidate cluster of *D*, *D*={D_1, D_2,$\cdots$, D_m}. The coupling of cluster *C* is

$$coupling(C,\alpha)=\sum_{c_i,c_j \in C; i<j} sim(c_i,c_j,\alpha) \tag{3.4}$$

The definition *coupling* represents the relevance of each pair of classes in a candidate cluster.

3.3 Finding the Best Clustering

A good cluster of multiple database must be of high cohesion, low coupling and as fewer classes as possible[1][2][4].The best cluster is selected among alternative candidate clusters. The task of our clustering is to find a clustering function $g: D \to C$ by using the database similarity matrix $DSM(D,\alpha)$ and the database distance matrix $DDM(D,\alpha)$ of D so that the expected features $sum\text{-}dist(C,\alpha)$, $coupling(C,\alpha)$ and $|C|$ are taken to a minimum. We can write:

$$\begin{cases} \min sum\text{-}dist(C,\alpha) \\ \min coupling(C,\alpha) \\ \min |C| \end{cases} \text{,subject to all feasible } g\,. \tag{3.5}$$

Such issues refer to the more objective (standard) optimization problem [10]. Here our problem is three-criteria optimization. Our linear weighted the sum method [11] and put it into a single objective optimization problem, as follows:

$$\min \lambda_1 sum\text{-}dist(C,\alpha) + \lambda_2 coupling(C,\alpha) + \lambda_3 |C| \tag{3.6}$$

subject to all feasible *g,* (3.6), where $\lambda_i \geq 0, (i=1,2,3)$, is the i criterion weight .

And then we can use the linear function as follows:

$$y = \frac{x - MinValue}{MaxValue - MinValue}, (MinValue \leq x \leq MaxValue) \tag{3.7}$$

To transform the objective functions into the range of 0~1; then we have the normalized objective functions as follows:

$$\begin{cases} sum\text{-}dist'(C,\alpha) = \dfrac{sum\text{-}dist(C,\alpha) - 0}{\frac{1}{2}m(m-1) - 0} = \dfrac{sum\text{-}dist(C,\alpha)}{\frac{1}{2}m(m-1)}, \\ coupling'(C,\alpha) = \dfrac{coupling(C,\alpha) - 0}{\frac{1}{2}m(m-1) - 0} = \dfrac{coupling(C,\alpha)}{\frac{1}{2}m(m-1)}, \\ |C|' = \dfrac{|C| - 1}{m-1}. \end{cases} \tag{3.8}$$

According to Formulas (3.6)(3.8), we give the best cluster definitions:

Definition 8. The goodness of a candidate cluster *C*=$\{c_1, c_2, ..., c_n\}$; use sim similarity that is defined such as :

$$goodness(C) = \min \lambda_1 sum\text{-}dist'(C,\alpha) + \lambda_2 coupling'(C,\alpha) + \lambda_3 |C|' \tag{3.9}$$

When $\lambda_i = 1, (i=1,2,3)$, it is known as an uniform calculation weight, which means that all objectives are equally important. As to our task, we treat the *sum-dist, coupling* and $|C|$ symmetrically in the methodology.

Let $F(C,\alpha) = sum\text{-}dist'(C,\alpha) + coupling'(C,\alpha) + |C|'$,then we have

$$goodness(C) = \min F(C,\alpha) \tag{3.10}$$

4. Algorithm of the Clustering

Since the number of the candidate clusters is much larger because there are not a few databases needed to be clustered, it is impossible to obtain all of them. As we know, the databases with high similarity should be clustered into one class, so we can cluster databases based on the similarities hierarchically. The procedure Database Hierarchical Clustering (DHC) for generating candidate clusters and identifying the best one is adopted in *Procedure 1*.

```
Procedure 1. Database Hierarchical Clustering(DHC)
begin
Input: $D_i(1\leq i\leq m)$:databases, $\alpha$: threshold value;
Output: $C_{best}$: the best cluster;
(1) find the frequent item sets of each $D_i$ under threshold $\alpha$;
(2) construct the database similarity matrix $DSM(D,\alpha)$ and the database distance matrix $DDM(D,\alpha)$;
(3) let $k$=1;
(4) construct a candidate cluster $C_k$. $C_k$=$\{c_1,c_2,\cdots,c_m\}$ where $c_i=\{D_i\}$, $1\leq i\leq m$;
(5)let $CDM(C_k,\alpha)=DDM(D,\alpha)$,calculate $F(C_k,\alpha)=sum\text{-}dist'(C_k,\alpha)+coupling'(C_k,\alpha)+|C_k|'$;
(6)let $goodness(C_{best})=F(C_k,\alpha)$ ,$C_{best}=C_k$;
(7)while ( $k<m$ ) do
begin
(7.1)find the pair of classes ( $c_p,c_q$ ), the distance value $CDM_{p,q}(C_k,\alpha)$of which is the minimum in the upper triangle of $CDM(C_k,\alpha)$;
(7.2) let $k$=$k$+1;
(7.3) let $c=c_p\cup c_q$;
7.4)    let $C_k$= $C_{k-1}-\{c_p\}-\{c_q\}+\{c\}$;
7.5)    (7.5) if $k\neq m$
begin
construct $CDM(C_k,\alpha)$ and calculate $F(C_k,\alpha)$;
else
calculate $F(C_k,\alpha)$;
end if
(7.6) if $F(C_k,\alpha)<goodness(C_{best})$
  begin
  let $goodness(C_{best})=F(C_k,\alpha)$ ,$C_{best}=C_k$;
  end if
end for
(8) output the best cluster $C_{best}$;
End procedure
```

Procedure Hierarchical Clustering generates *m* candidate clusters and obtains the best cluster according to *goodness*. When the allocation threshold *α*, step initializes Procedure (1), (2) and (3), Procedure(4) constructs a candidate cluster that each database to be one class. Step (5) and (6) initialize the *goodness*. Step (7) finds the minimum *goodness* and the best cluster by using hierarchical process. Step (8) outputs the best cluster C_{best}.5. ExperimentsWe conducted a series of experiments to verify the validity of our approach. We use a synthetic database T10I4D100K to split into ten database experiments on the effectiveness of our proposed algorithm. The T10I4D100K split into ten databases, the basic characteristics of the ten databases as shown in Table 1. As to a database, according to the threshold of change, we get different optimal clusterings, more small. The algorithm runs longer, as shown in Table 2. We still use similarity of the two-dimensional tables (Table 3). Compare our algorithm with BestClassification, the results are given in table 4. As for other values of BestClassification, which are not the most clustering, we only give two value comparison algorithm. From Table 4, our algorithm is superior to DHC BestClassification.

Database	quantity of transactions	Transaction's average length	The average number of frequent item sets	Item number set
DB1	10,000	11.06	127.66	866
DB2	10,000	11.13	128.41	867
DB3	10,000	11.07	127.65	867
DB4	10,000	11.12	128.44	866
DB5	10,000	11.14	128.75	865
DB6	10,000	11.14	128.63	866
DB7	10,000	11.11	128.56	864
DB8	10,000	11.10	128.45	864
DB9	10,000	11.08	128.56	862
DB10	10,000	11.08	128.11	865

Table 1: Enter the Database Features

Datast Set	Mim support α	Best cluster	Time
D	0.001	$\{\{DB_1,DB_3,DB_7,DB_5,DB_8,DB_{10}\},\{DB_2,DB_4,DB_5,DB_9,DB_6\},$	25min
	0.003	$\{\{DB_7,DB_3,DB_1\},\{DB_2,DB_4\},\{DB_5,DB_9,DB_6\},\{DB_8,DB_{10}\}\}$	15min
	0.004	$\{\{DB_1,DB_3\},\{DB_4,DB_2\},\{DB_6,DB_9,DB_5\},\{DB_7\},\{DB_8\},\{DB_{10}\}\}$	12min
	0.005	$\{\{DB_1,DB_3\},\{DB_2,DB_5\},\{DB_4,DB_9\},\{DB_7\},\{DB_9\},\{DB_3\},\{DB_{10}\}\}$	10min
	0.006	$\{\{DB_4,DB_1\},\{DB_2,DB_5\},\{DB_4,DB_9\},\{DB_9\},\{DB_4\},\{DB_5\},\{DB_{10}\}\}$	7min
	0.007	$\{\{DB_3\},\{DB_1,DB_5\}\{DB_2\},\{DB_3,DB_7\},\{DB_4\},\{DB_2\},\{DB_5\},\{DB_{10}\}\}$	5min
	0.008	$\{\{DB_5\},\{DB_3,DB_6\},\{DB_2\},\{DB_5,DB_4\},\{DB_8\},\{DB_9\},\{DB_8\},\{DB_{10}\}\}$	2min
	0.01	$\{\{DB_2\},\{DB_5\},\{DB_7\},\{DB_2\},\{DB_5,DB_6\},\{DB_3\},\{DB_9\},\{DB_2\},\{DB_{10}\}\}$	68s
	0.015	$\{\{DB_{10}\},\{DB_9\},\{DB_8\},\{DB_7\},\{DB_6\},\{DB_5\},\{DB_4\},\{DB_3\},\{DB_2\},\{DB_1\}$	45s

Table 2: Enter the Database Features

sim	DB_1	DB_2	DB_3	DB_4	DB_5	DB_6	DB_7	DB_8
DB_1	1	0.4	0.56	0.4	0.5	0.4	0.1	0.2
DB_2	0.4	1	0.4	0.79	0.4	0.45	0.2	0.1
DB_3	0.56	0.4	1	0.4	0.45	0.4	0.15	0.1
DB_4	0.4	0.79	0.4	1	0.4	0.45	0.2	0.1
DB_5	0.5	0.4	0.45	0.4	1	0.4	0.1	0.1
DB_6	0.4	0.45	0.4	0.45	0.4	1	0.2	0.2
DB_7	0.1	0.2	0.15	0.2	0.1	0.2	1	0.28
DB_8	0.2	0.1	0.1	0.1	0.1	0.2	0.28	1

Table 3: Similarity Telation

Algorithm	α	Best cluster	Time (ms)
BestClassification	0.6	$\{\{DB_7\},\{DB_2,DB_4\},\{DB_1\}\{DB_3\},\{DB_6\},\{DB_5\},\{DB_8\}\}$	20
DHC		$\{\{DB_6,DB_3,DB_5\},\{DB_7\},\{DB_8\},\{DB_6,DB_4,DB_2\}\}$	14
BestClassification	1	$\{\{DB_8\},\{DB_7\},\{DB_6\},\{DB_5\},\{DB_4\},\{DB_3\},\{DB_2\},\{DB_1\}\}$	15
DHC		$\{\{DB_1\},\{DB_2\},\{DB_3\},\{DB_4\},\{DB_5\},\{DB_6\},\{DB_7\},\{DB_8\}\}$	10

Table 4: Experimental Results on 3 Similarity

5. Conclusion

This paper presents a multi database classification method based on high cohesion and low coupling. The definition of distance used to measure the cohesion and the definition of coupling degree is used to measure the Multi-target optimization problem. Afterwards, the hierarchical clustering algorithm is used to find the ideological structure for optimal clustering. The we use an artificial database T10I4D100K and a real database similarity in two-dimensional tables shows effectiveness of the algorithm model; then we carry on the contrast experiment with the BestClassification algorithm. The experiments show that our method of clustering stability is strong and the time complexity is lower than that of BestClassification while featuring strong generalization ability.

References

[1] X.Wu,C.Zhang,S. Zhang. *Database classification for multi-database mining*, Information Systems. 30 (1) , 71–88(2005).

[2] A. Adhikari, P.R. Rao. *Efficient clustering of databases induced by local patterns*. Decision Support Systems. 44(4),925-945(2008).

[3] H. Liu, H. Lu, J. Yao. T*oward multidatabase mining: identifying relevant databases*, IEEE Transactions on Knowledge and Data Engineering 13(4): 541–553(2001).

[4] D.Yuan, H. Fu, Z. Li, H. Wu. *An application-independent database classification method based on high cohesion and low coupling*, Journal of Information & Computational Science 7(1),1-6(2012).

[5] H. Li, X. Hu, and Y. Zhang. *An improved database classication algorithm for multi-database min-ing.* Proc. of Frontiers of Algorithmics Workshop in LNCS. Hefei, China, 2009:187-199.

[6] X.D.Wu, S.C.Zhang. *Synthesizing high-frequencyrules from different data sources*, IEEE Trans. Knowledge Data Eng. 15 (2):353–367(2003).

[7] X. Yin, J. Han. *Efficient classification from m ultiple heterogeneous databases*. In:Proceedings of 9-th European Conf. on Principles and Practice of Knowledge Discovery in Databases,Lecture Notes in Computer Science Volume 3721,pp ,404–416(2005)

[8] X. Yin, J. Han, J. Yang, PS. Yu. *Efficient classification across multiple database relations: A crossmine approach*. IEEE Transactions on Knowledge and Data E ngineering 18(6): 770–783(2006)

[9] S .Bandyopadhyay, C. Giannella , U .Maulik , H. Kargupta , K .Liu , S. Datta. *Clustering distributed data streams in peer-to-peer environments*. Information Sciences 176(14): 1952–1985(2006)

[10] C. Hillermeier. *Nonlinear multiobjective optimization.* Birkhäuser Basel.pp,135,2001.

[11] L. Zadeh. *Optimality and non-scalar-valued performance criteria*. IEEE TAC, 8(1):59–60(1963).