An Anonymous Algorithm for Hierarchical Clustering Based on K-Prototypes

YAO Yuan-jing[1] and SUN Yi*

[1] Beijing University of Posts and Telecommunications, China

[a]yaoyuanjing@bupt.edu.cn and *sunyisse@bupt.edu.cn

Keywords: k-prototypes; clustering; multi-level; information loss

Abstract. By the research on the clustering problem of multiple attributes data processing based on K-Prototypes algorithm, this paper improves distance formula, which can more accurately reflect the differences between tuples. Besides, according to the various demand of privacy preservation, the sensitive value is divided into multiple levels by (KLS, λ-clustering) – hierarchical anonymous model. The experimental results show that this algorithm is able to achieve highly accurate clustering results. It can also satisfy the requirements of multi-level privacy preservation of sensitive attributes, and effectively reduce the information loss.

Introduction

K-Prototypes [1] clustering algorithm can be used to handle data sets combined with numerical and categorical data, but its clustering result depends on the choice of the initial value. However, the improved initial value selection algorithm [2] cannot accurately measure the distance between tuples. Therefore, this article gives an improved distance formula and a cost function. In addition, a hierarchical anonymous model based on clustering is represented due to various demands for privacy preservation. Experimental results also show that the algorithm can protect the privacy of individuals and reduce the loss of information effectively.

Improved Distance Formula and Cost Function

Suppose that the quasi-identifier in the data sheet T is $QI(N_1,\ldots,N_n,M_{n+1},\ldots,M_m)$ in which $N_p(1\le p\le n)$ represents the p th numerical attribute and $M_p(n+1\le p\le m)$ represents the p th categorical attribute. During the K-anonymity, the tuple $t=(v_{N_1},v_{N_2},\ldots,v_{N_n},v_{M_1},v_{M_2},...,v_{M_m})$ is transformed into $t'=([y_{N_1},z_{N_1}],[y_{N_2},z_{N_2}],\ldots,[y_{N_n},z_{N_n}],D_{M_1},D_{M_2},...,D_{M_m})$. This paper considers the data generalization's effect on the anonymous data information loss, and the distance formula and cost function are improved as the following:

Distance Formula. By the K-Prototypes algorithm, the distance between tuples can be expressed as the sum of the distance between the numerical data and categorical data.

Definition 1 (distance formula between tuples) t_i,t_j represents tuples in the data sheet T, and the distance between t_i,t_j is:

$$d(t_i,t_j)=\sum_{p=1}^{n}x_{N_p}dist_{N_p}(t_i,t_j)+\sum_{p=n+1}^{m}x_{M_p}dist_{M_p}(t_i,t_j) \quad (1)$$

Definition 2 (distance formula between tuple and cluster) C_j represents an equivalent cluster, y_j represents the center of C_j, and the distance between t_i and C_j is defined as the distance between t_i and y_j.

$$d(t_i, C_j) = d(t_i, y_j) = \sum_{p=1}^{n} x_{N_p} dist_{N_p}(t_i, y_j) + \sum_{p=n+1}^{m} x_{M_p} dist_{M_p}(t_i, y_j) \quad (2)$$

Numerical Data. The first item of the formula represents the numerical data's contribution to the distance between tuples. The original K-Prototypes algorithm uses square Euclidean distance $|v_i - v_j|^2$ to measure. Considering the range of all values of N_p attribute, this paper uses the following improved formula:

$$dist_{N_p}(t_i, t_j) = \frac{|v_i - v_j|}{Max(N_p) - Min(N_p)} \quad (3)$$

$Max(N_p)$ represents the maximum value in N_p and $Min(N_p)$ represents the minimum value in N_p.

Categorical Data. The second item represents categorical data's contribution. The original K-Prototypes algorithm uses the Hemingway distance formula $\delta(t_i, t_j) = \begin{cases} 0, & if\ v_i = v_j, \\ 1, & if\ v_i \neq v_j \end{cases}$ to simply judge whether the tuples t_i, t_j are in the same equivalent cluster, which is ignoring the mutual influence of tuples in different clusters. This paper refers to the construction method of the generalization hierarchy tree [3] and the hierarchy distance formula based on weight. Some related definitions are given in the following:

Definition 3 (weighted hierarchical distance) h represents the level of domain generalization, layer $1, 2, ..., h-1$ represents the number of layers from the most generalized layer to the most specified layer. State $\omega_{j.j-1}, 2 \le j \le h$ represents the weight between layer j and $j-1$. When a attribute value generalizes from layer p to q, $p > q$, this weighted hierarchical distance is defined as the following formula:

$$WHD(\mathrm{p}, \mathrm{q}) = \frac{\sum_{j=q+1}^{p} \omega_{j.j-1}}{\sum_{j=2}^{h} \omega_{j.j-1}} \quad (4)$$

where $\omega_{j.j-1} = 1/(j-1)^{\beta}, 2 \le j \le h$. β is defined by the user, e.g. $\beta = 1$. The method can reflect the differences between data distortions caused by different layers' generalization

Definition 4 (distortion of the tuple's generalization) $level(v_j)$ is the domain layer of the attribute value v_j in the generalization layer, and the distortion of t and t' is defined as:

$$Distortion(t, t') = \sum_{j=1}^{m} WHD(level(x_j), level(x_j')) \quad (5)$$

Definition 5 (most recent public generalization node) All the values of any attribute form a hierarchical tree. Each node in the tree corresponds to a value and the sub-node corresponds to a more specific value. The value v_{12}^p of t_{12}, which is the most recent public generalization node of t_1 and t_2, can be defined as:

$$v_{12}^p = \begin{cases} v_1^p & , \ if\ v_1^p = v_2^p, \\ \text{value of the most recent public generalization}, & if\ v_1^p \neq v_2^p \end{cases} \quad (6)$$

in which v_1^p, v_2^p represents the p th attribute value of t_1, t_2.

Definition 6 (distance between tuples) Assume t_1,t_2 are two tuples, the distance between t_1,t_2 is defined as:

$$Distortion(t_1,t_2) = Distortion(t_1,t_{12}) + Distortion(t_2,t_{12}) \tag{7}$$

Definition 7 (distance between equivalent clusters) Assume equivalent cluster C_1 contains $n_1, t_1 \in C_1$ and C_2 contains n_2, $t_2 \in C_2$, the distance between C_1, C_2 can be defined as:

$$Dist(C_1,C_2) = n_1 \times Distortion(t_1,t_{12}) + n_2 \times Distortion(t_2,t_{12}) \tag{8}$$

According to（4）～（7）, the distance between categorical data is

$$dist_{M_p}(t_i,t_j) = \sum_{j=1}^{m} WHD(level(v_i),level(v_{ij})) + \sum_{j=1}^{m} WHD(level(v_j),level(v_{ij})) \tag{9}$$

Improved weight. In the original categorical data distance formula, γ represents the weight of categorical data in the tuple [4]. However, it only can be used to represent the different weights of numerical and categorical data instead of the weight of each attribute on the tuple. Therefore, this paper introduces the background reference matrix [5] and give the calculate method which is both suitable for numerical and categorical data.

Definition 8 (background reference matrix) q_{ij} represents the impact of the i th background data to the j th attribute of the tuple, which is defined by experts or the data owner. The background reference matrix is a $n \times p$ matrix, in which n represents the number of the tuples in the data sheet and p represents the number of the identifiers. The matrix is as the following shows:

$$J = (q_{ij})_{n\times p} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_p \\ 0 & 5 & 6 & \cdots & 3 \\ 5 & 3 & 4 & \cdots & 7 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 2 & 4 & \cdots & 7 \end{bmatrix} \tag{10}$$

where x_{N_p} represents the impact of the i th identifier to the j th tuple: $x_{N_p} = \dfrac{q_{ij}}{\sum_{i=1}^{m} q_{ij}}$.

Similarly， $x_{M_p} = \dfrac{q_{ij}}{\sum_{i=n+1}^{n+m} q_{ij}}$.

Cost Function

Definition 9 (cost function) a cost function for clustering a data set with numeric and categorical attributes can be written as:

$$\begin{cases} P(t_i,C_j) = \sum_{j=1}^{k} u_{pj} x_{N_p} \sum_{p=1}^{n} dist_{N_p}(t_i,y_j) + \sum_{j=1}^{k} u_{pj} x_{M_p} \sum_{p=n+1}^{m} dist_{M_p}(t_i,y_j) \\ \sum_{i=1}^{n+m} u_{ij} = 1; u_{ij} \in [0,1] \end{cases} \tag{11}$$

in which $u_{ij} = 1$ represents t_i is in the cluster C_j, $u_{ij} = 0$ represents it does not belong to C_j, and k represents the number of clustering data.

The Hierarchal Anonymous Model Based on Clustering

In reality, there are considerable differences in the demand of privacy preservation of different sensitive attribute values. Simply constraining different sensitive attributes to one value, instead of considering the real demand of privacy preservation, is likely to cause a lot of unnecessary information loss and low efficiency. It may also not be able to meet the requirements of high sensitive attribute values. Aiming at this problem, this paper proposes a multidimensional (KLS, λ -clustering) -hierarchical anonymous model, based on the three-dimensional hierarchical classification of sensitive attributes [6].

Definition 10 (classification of sensitive attributes) Suppose that SE is a set of values of the sensitive attribute ST in the data sheet T, that is $SE = \bigcup_{i=1}^{s} t_i.ST$. According to the demand for privacy preservation, the sensitive attribute values are divided into s classes, which are represented by $SE_1, SE_2, \ldots, SE_s$, in which the classification are usually determined by experts and data owners before the data is anonymous.

Definition 11 (KLS, λ -clustering hierarchal anonymous model) In the data sheet T, if the following two conditions are satisfied, then the data sheet T can be said to meet (KLS, λ -clustering) hierarchal anonymous model.

1) The data sheet T meets k-anonymity and l-diversity model.

2) The kinds of values on sensitive attribute ST in any equivalent cluster should be more than λ_i, in which $\lambda = (\lambda_1, \lambda_2, ..., \lambda_s)$ represents privacy preservation parameter, corresponding the controllable privacy preservation grade of $SE_1, SE_2, \ldots, SE_s$, $0 \le \lambda_i \le k (i = 1, 2, ..., s), \lambda_1 < \lambda_2 < ... < \lambda_s$. The higher the sensitivity property of is, the larger λ_i is.

(KLS，λ -clustering)-hierarchical clustering algorithm

(KLS, λ -clustering) - hierarchical clustering algorithm makes up for defect that the (KLS-clustering) algorithm clustering is not accurate and cannot classify sensitive attributes. The basic idea is to select k initial points first, then cluster tuples into k equivalent clusters. Afterwards, adjust and merge compatible equivalent cluster, and finally generalize and publish data. Specific description is shown in the algorithm.

(KLS, λ -clustering) - An anonymous algorithm for hierarchical clustering based on K-Prototypes

Input: the original data set T, QI, parameter l of l-diversity model, sensitive attribute ST, background reference matrix J, hierarchical anonymous model parameters s, λ_i.

Output: the data set meet the requirement of (KLS, λ -clustering) – hierarchical model.

Method：

/* Generation of dataset with equivalent cluster within at least l tuples */

(1) if different sensitive attribute values of T is more than l then

(2) generate equivalent cluster from T, values of quasi-identifiers of each tuple are equal;

/ * Select initial points * /

(3) calculate the center of data set y_0: for numerical data, take the average value; for categorical data, take the highest frequency value;

(4) calculate all the distance between tuples and y_0: $d(t_i, y_0)$;

(5) select $y_1 = \arg\max d(t_i, y_0)$, $Y = Y \cup \{y_1\}$,set $d(y_1, y_0) = m$;

(6) for j=2 to k do

(7) c=j-1:

(8) for $t_i \in T - Y$ do

(9) calculate $d(t_i, y_{j-1})$;

(10) end for;
(11) choose the tuple whose $d(t_i, y_{j-1})$ is most close to m/c and set it y_j, $Y = Y \cup \{y_i\}$;
(12) end for;
(13) make $Y = \{y_1, y_2, ..., y_k\}$ as the initial cluster center;
(14) calculate $d(t_i, y_i)$ and select $C_i = \arg\min d(t_i, y_i)$;
(15) if (($t_i.ST \in SE_1$) && (the number of tuples unrelated to $t_i.ST$ in the cluster $C_i \geq \lambda_1$))||
(($t_i.ST \in SE_2$) && (the number of tuples unrelated to $t_i.ST$ in the cluster $C_i \geq \lambda_2$))||…
(($t_i.ST \in SE_s$) &&(the number of tuples unrelated to $t_i.ST$ in the cluster $C_i \geq \lambda_s$))
then $C_i = C_i \cup \{t_i\}$;
(16) if i<n then to step (14);
(17) calculate the cost function $P(t_i, C_j)$;
(18) recalculate the center of equivalent cluster, to step (14) until $P(t_i, C_j)$ tends to be stable.
(19) output clustering results;
(20) end
/* Sensitive attributes anonymous algorithm * /
(21) while the existing sensitive attributes in T is more than l do begin
(22) randomly select a equivalent cluster Q whose sensitive attributes is less than l :
(23) calculate the distance between Q and other equivalent cluster $Q'[N]: d(Q, Q'[N])$;
(24) sort $Q'[N]$ from small to large according to $d(Q, Q'[N])$;
(25) for i=1 to N (N<n/l)
(26) if the sum of equivalent clusters in $Q, Q'[N] \leq l$ then
(27) merge $Q, Q'[N]$ and generalize;
(28) break;
(29) end for;
(30) end

Experimental results

In order to further verify the effectiveness of the improved algorithm, we carried out experiments. The experiment uses Adult dataset in the UCI machine learning database. This dataset is the standard dataset of anonymity protection research.

There are 45222 tuples in the dataset (after deleting the data within missing value). We use 9 attributes: Age, Gender, Race, Marital Status, Education, Native Country, Work Class, Salary Class, Occupation. Age is a numerical attribute and others are categorical data. The experiment takes the first d attributes as the quasi-identifiers and Occupation as the sensitive attribute. The number of the different sensitive values is 14.

The hardware environment of the experiment is Core i5-4200U 1.60GHz, CPU Intel, RAM 1.6GHz and the operating system is Windows Microsoft 7. All procedures are implemented with java.

This experiment compares KLS-clustering algorithm and (KLS, λ)-clustering algorithm.

Figure 1 shows, with quasi identifier dimension $|QI|$ change, the information loss of (KLS, λ)-clustering algorithm at $l = 10$, $k = n / 2l = 2260$, hierarchical anonymous model parameter $s = 3$, $\lambda_1 = 9000$, $\lambda_2 = 20000$, $\lambda_3 = 35000$; KLS-clustering algorithm at $l = 10$, $k = n / 2l = 2260$. From Figure 1, we can see that with the increase of the dimension of the quasi identifier $|QI|$, the information loss of the two algorithms is increased. This is because when $|QI|$ increase, the tuple needs to generalize more attributes, which apparently will bring more information

loss. But under the same condition, (KLS, λ)-clustering algorithm produces slightly less information loss than KLS-clustering algorithm.

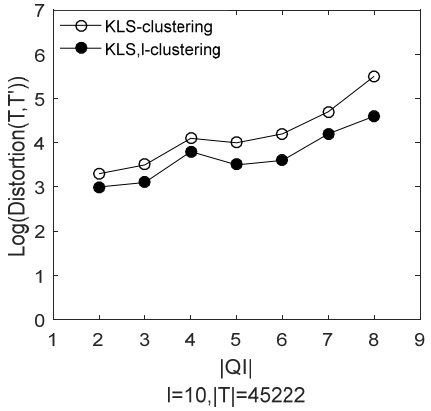

Fig. 1 Information Loss

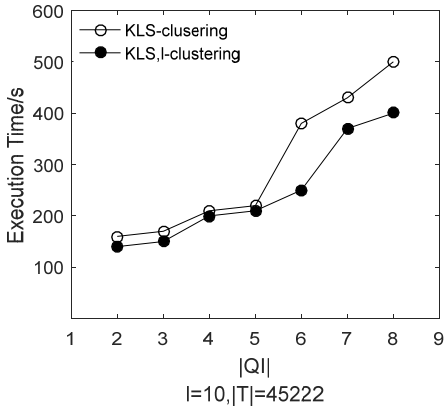

Fig. 2 Execution Time

Figure 2 shows, with quasi-identifier dimension change, the execution time of (KLS, λ)-clustering algorithm at $l=10$, $k=n/2l=2260$, hierarchical anonymous model parameter $s=3$, $\lambda_1=9000$, $\lambda_2=20000$, $\lambda_3=35000$; KLS-clustering algorithm at $l=10$, $k=n/2l=2260$. With the $|QI|$ increase, both the execution time increased. But (KLS, λ)-clustering algorithm performs better than the KLS-clustering algorithm. Combing the information loss and execution time, the (KLS, λ)- clustering algorithm is effective and feasible.

Summary

In the study of anonymous protection, it is not accurate to deal with the data distance simply. And the single constraint of sensitive attribute can result in a large amount of unnecessary information loss and low processing efficiency.

This paper presents (KLS, λ)-clustering hierarchical anonymous algorithm. This algorithm can greatly improve the accuracy of clustering and realize the hierarchical demand of different sensitive attributes so that information loss can be reduced.

Acknowledgements

This work was financially supported by the Research Innovation Fund for College Students of Beijing University of Posts and Telecommunications.

References

[1] Z.Huang. Extensions to the K-means Algorithms for Clustering Large Data Sets with Categorical Values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304.
[2] LI Shan-shan, A Research on sensitive attributes protection based on clustering [D].Jiangsu: Jiangsu University, 2012.
[3] Li Jiuyong, Wong Raymond Chi-Wing, Fu Ada Wai-Chee, etal. Achieving k-anonymity by clustering in attribute hierarchical structure [A]. LNCS 4081: Proceeding of the 8th Int Conf on Data Warehousing and Knowledge Discovery. Berlin: Springer, 2006: 405-416
[4] HUANG Z, MANG. Fuzzy K-modes algorithm for clustering categorical data [J]. IEEE Transactions on Fuzzy Systems, 1999.7(4): 446-452.
[5] SHI Li-yan, GU Bao-ping, YAO Xue-li. Application of Personal Protection Based on Improved K-anonymity Algorithm [J]. Computer Simulation, 2014. 3(31): 217-220
[6] GUI Qiong, CHENG Xiaohui, Clustering-based approach for multi-level anonymization[J], Journal of Computer Application, 2013,33(2):412-416