An improved weighted-feature clustering algorithm for k-anonymity

Lijian Lu and Xiaojun Ye
School of Software, Tsinghua University,
Beijing, 100084, P. R. China
lulj06@mails.tsinghua.edu.cn, ye-xj@tsinghua.edu.cn

Abstract
Chiu proposed a clustering algorithm adjusting the numeric feature weights automatically for k-anonymity implementation and this approach gave a better clustering quality over the traditional generalization and suppression methods. In this paper, we propose an improved weighted-feature clustering algorithm which takes the weight of categorical attributes and the thesis of optimal k-partition into consideration. To show the effectiveness of our method, we do some information loss experiments to compare it with greedy k-member clustering algorithm.

Keywords: k-anonymity, clustering, k-partition

1. Introduction
K-anonymity is an effective evolutionary privacy-preserving publication method \[5, 6\]. Its key idea is to ensure that each record should be identical to at least other k-1 records in the released dataset with respect to some privacy-related attributes called quasi-identifier.

Methods implementing k-anonymity based on clustering algorithm have gained much attention \[1\]. In this paper, we introduce an effective evolutionary clustering algorithm (I-WFC for short) which takes the weight of both numeric and categorical attributes into consideration and the ultimate result of this method satisfies the theory of optimal k-partition \[5\]. In Chiu’s method, the original equivalence cluster centers are selected randomly \[2\]. In our implementation, we get some gains for information loss by selecting the original equivalence cluster centers regularly. Besides all these improvements, we give an appropriate distance measurement that enables clustering of data with mixed types of numeric and categorical attributes.

The rest of this paper is organized as follows. Section 2 recalls some preliminary definitions and introduces a new distance measurement. Section 3 presents our proposed I-WFC algorithm and Section 4 shows the experiment result. Finally, Section 5 concludes this paper.

2. Preliminary definitions

Clustering algorithm for k-anonymity is the assignment of records into groups (also called equivalence clusters) so that records from the same cluster are more similar to each other than records from different clusters. An important step in any clustering is to select a distance measure, which will determine how the similarity of two elements is calculated. In this paper, we use Zhu’s method quoted as \(\text{Diss}^N\) \[8\] for distance between two numeric values, and for the distance between two categorical values, we adopt Lin’s method quoted as \(\text{Diss}^C\) \[4\] in the following sections.

**Definition 1. (Categorical records distance)** In general, records can be classified into three kinds. The first one is **numeric record** which comprises only numeric attributes and arithmetical operations can be performed on it. The second one is **categorical record** which has only categorical attributes including values over a finite set and standard arithmetical operations do not make sense. The third kind includes both numeric and categorical attributes and we name it as the **general record**.

Generally, categorical attributes have hierarchical relationships which can be regarded as taxonomy tree. If two categorical values reside in the same depth of the taxonomy tree, the value of \(\text{Diss}^C\) with Lin’s method is zero. For this reason, we propose the average of the sum of each value’s arithmetical operations can be performed on it.

**Definition 2. (Numeric record distance)** Two record distances can be classified into three kinds. The first kind includes only numeric values and we name it as the **numeric distance**. The second kind includes both numeric and categorical values and we name it as the **mixed distance**. The third kind includes only categorical values and we name it as the **categorical distance**.

In this paper, we use Zhu’s method quoted as \(\text{Diss}^N\) \[8\] for distance between two numeric values, and for the distance between two categorical values, we adopt Lin’s method quoted as \(\text{Diss}^C\) \[4\] in the following sections.

**Definition 3. (Mixed record distance)** In general, records can be classified into three kinds. The first one is **numeric record** which comprises only numeric attributes and arithmetical operations can be performed on it. The second one is **categorical record** which has only categorical attributes including values over a finite set and standard arithmetical operations do not make sense. The third kind includes both numeric and categorical attributes and we name it as the **general record**.

Generally, categorical attributes have hierarchical relationships which can be regarded as taxonomy tree. If two categorical values reside in the same depth of the taxonomy tree, the value of \(\text{Diss}^C\) with Lin’s method is zero. For this reason, we propose the average of the sum of each value’s distance between their closest common ancestors in the taxonomy tree as the distance measurement. Consider two categorical records \(t_1\) and \(t_2\) which consist of \(N\) categorical attributes and their closest common ancestor record is \(t_{12}\):

\[
\begin{align*}
&\quad t_1 = (t_{1,1}, t_{1,2}, \ldots, t_{1,N}) \\
&\quad t_2 = (t_{2,1}, t_{2,2}, \ldots, t_{2,N}) \\
&\quad t_{12} = (t_{12,1}, t_{12,2}, \ldots, t_{12,N})
\end{align*}
\]

The distance between records \(t_1\) and \(t_{12}\) is as follows.

\[
\text{Distance}(t_1, t_{12}) = \sum_{i=1}^{N} w_j \cdot \text{Diss}^C(\text{level}(t_{1,i}), \text{level}(t_{12,i}))
\]  

(1)

and the distance between \(t_2\) and \(t_{12}\) is:

\[
\text{Distance}(t_2, t_{12}) = \sum_{i=1}^{N} w_j \cdot \text{Diss}^C(\text{level}(t_{2,i}), \text{level}(t_{12,i}))
\]  

(2)

Where function \(\text{level}(i)\) means the depth of the value in the taxonomy tree and \(w_j\) means the weight of \(j\)-th attribute.

Therefore, the distance between \(t_1\) and \(t_2\) is as follows:

\[
\text{Distance}(t_1, t_2) = \left(\text{Distance}(t_1, t_{12}) + \text{Distance}(t_2, t_{12})\right) / 2
\]  

(3)
**Definition 2. (General records distance)** Consider two records $t_1$ and $t_2$ comprising $N$ attributes, where the first $p$ attributes are numeric and the rest $N-p$ are categorical:

$$t_1 = (t_{1,1}, t_{1,2}, \ldots, t_{1,N}) \quad t_2 = (t_{2,1}, t_{2,2}, \ldots, t_{2,N})$$

We define $t_{1c} = (t_{1,1:p}, t_{1,p+2}, \ldots, t_{1,N})$ as the categorical record for $t_1$ and $t_{2c} = (t_{2,1:p}, t_{2,p+2}, \ldots, t_{2,N})$ for record $t_2$. The distance between $t_{1c}$ and $t_{2c}$ can be calculated by using equation (3). As such, the distance between general records $t_1$ and $t_2$ is defined as follows:

$$\text{Diss}(t_1, t_2) = \sum_{j=1}^{N} w_j \cdot \text{DissN}(t_{1j}, t_{2j}) + \text{Distance}(t_{1c}, t_{2c})$$  \hspace{1cm} (4)

Where $w_j$ means the weight of the $i$-th attribute and $\sum_{j=1}^{N} w_j = 1$.

For a good partition, there should be appreciable intra-cluster isolation (or heterogeneity among clusters) and intra-cluster homogeneity. The homogeneity is calculated by a minimizing function – usually the sum of distances discussed above between all the pairs within or between clusters.

**Definition 3. (Inter-cluster isolation)** Assume we have one dataset which consists of $M$ records and each record comprises $N$ attributes represented as $A_j$, $1 \leq j \leq N$. Suppose during the clustering process, we get $C$ equivalence clusters. We use the minimal distance value among cluster centers to measure the inter-cluster isolation.

For numeric attribute $A_j$, by using $\text{DissN}$, the record attributes’ affect on inter-cluster isolation can be calculated:

$$\text{DistortionN}(A_j) = \min \{\text{DissN}(c_{m,j}, c_{n,j}) \mid 1 \leq m, n \leq C \text{ and } m \neq n \}$$  \hspace{1cm} (5)

Where $c_{m,j}$, $c_{n,j}$ means the value of the attribute $A_j$ for $m$-th and $n$-th equivalence cluster centers.

For categorical attribute, by using $\text{DissC}$, the record attributes’ affect can be calculated:

$$\text{DistortionC}(A_j) = \min \{\text{DissC}(\text{level}(c_{m,j}), \text{level}(c_{n,j})) \mid 1 \leq m, n \leq C \text{ and } m \neq n \}$$  \hspace{1cm} (6)

Where $c_{m,j}$, $c_{n,j}$ means the value of attribute $A_j$ for $m$-th and $n$-th equivalence cluster centers.

So the equation for the inter-cluster isolation measured by $A_j$ is as follows:

$$\text{Distortion}(A_j) = \begin{cases} 
\text{DistortionN}(A_j), 1 \leq j \leq p \\
\text{DistortionC}(A_j), p+1 \leq j \leq N
\end{cases}$$  \hspace{1cm} (7)

**Definition 4. (Intra-cluster homogeneity)** For intra-cluster homogeneity, firstly we calculate each record’s distance between their corresponding equivalence cluster center and then sum these values together as the measurement for one equivalence cluster. After this, we sum all these values of equivalence clusters as the ultimate measurement.

For numeric attribute, the record attribute’s affect on intra-cluster homogeneity is:

$$\text{DistortionNInfo}(C_i, A_j) = \sum_{c_i \in C_i} n_{c_i} \cdot \text{DissN}(c_{i,j}, c_{j}), 1 \leq j \leq p$$  \hspace{1cm} (8)

Where $t_{m,j}$ means the value of $j$-th attribute for record $m$, $c_j$ means the value of $j$-th attribute for $i$-th equivalence cluster center and $t_{m,j}$ represents the membership discussed in section 3 for $m$-th record with $i$-th equivalence cluster.

For categorical attribute, we have:

$$\text{DistortionCInfo}(C_i, A_j) = \sum_{(c_{m,j}, c_{n,j})} \text{DissC}(\text{level}(c_{m,j}), \text{level}(c_{n,j})), p+1 \leq j \leq N$$  \hspace{1cm} (9)

The whole intra-cluster homogeneity affected by $A_j$ is:

$$\text{DistortionInfo}(A_j) = \frac{\sum_{j=1}^{p} \text{DistortionNInfo}(C_i, A_j), 1 \leq j \leq p}{\sum_{j=p+1}^{N} \text{DistortionCInfo}(C_i, A_j), p+1 \leq j \leq N}$$  \hspace{1cm} (10)

By using the measurement equation from definition 3 and definition 4, we can compute the weight increment for each attribute to prepare for the next clustering iteration.

**Definition 5. (Weight increment)** We firstly sum all the attributes’ affect values to measure the whole dataset’s inter-cluster isolation and intra-cluster homogeneity respectively. After this, according to the clustering principle, we should make inter-cluster isolation smaller enough and intra-cluster homogeneity larger enough, so the target of weight adjustment should maximize the following equation:

$$\sum_{j=1}^{p} w_j \cdot \text{DistortionN}(A_j) \quad \sum_{j=p+1}^{N} w_j \cdot \text{DistortionC}(A_j)$$  \hspace{1cm} (11)

Let $w_j$ be the weight value of $j$-th attribute in the current iteration. We know that our target is to make $\text{Distortion}(A_j)$ large enough and $\text{DistortionInfo}(A_j)$ small enough. We regard this consideration as making the fraction of $\text{Distortion}(A_j)/\text{DistortionInfo}(A_j)$ large enough. In other words, this $j$-th attribute contributes “more” than other attributes for achieving the adjustment objective. Moreover, the importance of this attribute can be measured by the proportion of this fraction value to the sum of all the $N$ attribute fraction’s values. Therefore, the weight increment amount for the next iteration for weight $w_j$ is:

$$\Delta w_j = \frac{\text{Distortion}(A_j) / \text{DistortionInfo}(A_j)}{\sum_{j=1}^{N} \text{Distortion}(A_j) / \text{DistortionInfo}(A_j)}$$  \hspace{1cm} (12)

The value of the attribute weight for the next iteration is:

$$w'_j = w_j + \Delta w_j$$  \hspace{1cm} (13)

The value $w'_j$ maybe larger than one, so it needs to be normalized as the value between 0 and 1. Use the following normalization formula:

$$f(x_j) = \frac{x_j}{\sum_{j=1}^{N} x_j}$$  \hspace{1cm} (14)

Through equation (14), the value of $w'_j$ can be derived as:


\[ w_j = \frac{\sum_{j=1}^{n} w_j \cdot \text{DistortionInfo}(A)}{\sum_{j=1}^{n} \text{Distortion}(A)} \]

\[ \text{Definition 6: (Information loss)} \text{ Based on equation 4, we use the sum of each record’s distance between their corresponding equivalence clusters to measure the information loss, that is:} \]

\[ \text{IL} = \sum_{m=1}^{N} \sum_{i=1}^{m} u_{mi} \cdot \text{DissR}(t_m, c_i) \]

\[ \text{(15)} \]

\[ \text{(16)} \]

3. I-WFC algorithm

I-WFC consists of two steps. The first step is Clustering that partitions the original dataset into small equivalence clusters. The second step is Adjustment that takes the thesis of optimal k-partition \(^{[3]}\) into consideration.

3.1 Clustering

We use the following metric and equation to determine which equivalence cluster should record \( t_m \) belong to based on equation 4:

\[ u_{mi} = \begin{cases} 1, & \text{DissR}(t_m, c_i) \leq \text{DissR}(t_m, c_j) \\ 0, & \text{else} \end{cases} \]

\[ \text{(17)} \]

Where \( t_m \) means the \( m \)-th record and \( c_i, c_j \) are \( i \)-th \( j \)-th equivalence cluster centers. After this assignment, we should update each equivalence cluster centers. So, for numeric attribute, the new center for attribute \( A_j \) is:

\[ a_{ij} = \frac{\sum_{m=1}^{N} u_{mi} \cdot t_{mj}}{\sum_{m=1}^{N} u_{mi}} \]

\[ \text{(18)} \]

For categorical attribute, we first find the closest common ancestor value of each attribute and then generalize each categorical value to this ancestor:

\[ a_{ij} = \text{The value of closest common ancestor in the cluster.} \]

\[ \text{(19)} \]

The pseudo code of the Clustering step is summarized in Fig. 1: we first selecting \( C \) records as the equivalence cluster centers randomly and then initialize each quasi-identifier attribute’s weight to be equal \( 1/N \); after this, by using equation 17, we assign each record to their “should be” equivalence clusters and then update each cluster centers by using equation 18 and 19; we adjust each attribute’s weight according to equation 15 and then the procedure goes back to the “assignment” operation to reassign records to their next “should be” equivalence clusters. This iteration exists when the assignment operation ends.

3.2 Adjustment

After Clustering step, we obtained \( C \) equivalence clusters in which there maybe some clusters with size less than \( k \) or larger than \( 2k \) which are non-optimal. For these two kinds of non-optimal clusters, we treat them differently. For clusters with size less than \( k \), we merge each record in it with their closest optimal cluster. After this operation, we find clusters with size larger than \( 2k \). At this time, we treat each found cluster as a new small dataset and recursively recall Clustering and Adjustment. The pseudo-code is summarized in Fig. 2:

\[ \text{Input: A dataset } T \text{ containing } M \text{ records which comprise } N \text{ quasi-identifier attributes (Among which } p \text{ are numeric and the rest } N - p \text{ are categorical), and the value of } k \text{ for the } k\text{-anonymity requirement. For formalization, we use } T = \{ t_1, t_2, \ldots, t_M \} \text{, in which } \] \( t_i = (a_{i1}, a_{i2}, \ldots, a_{In}) \) means the \( i \)-th record of the \( M \) records. \]

\[ \text{Output: A partition of the } M \text{ records, among which there maybe clusters with size less than } k \text{ or larger than } 2k . \]

1: Compute the value: \( C = M/k \)
2: Firstly, let the weight of each quasi-identifier attribute be equally \( 1/N \).
3: Repeat
   (1): Minimize \( C \) equivalence clusters using equation 17
   (2): Compute the new cluster centers using equation 18 and 19
   (3): Adjust the weight of each quasi-identifier attribute using equation 15
   Until the membership of each equivalence cluster do not change.

\[ \text{Figure 1.Pseudo-Code of Clustering step} \]

\[ \text{Figure 2.Pseudo-Code of Adjustment step} \]

417
Lines 4-12 finds clusters with size less than k or larger than 2k; lines 13-25 treats the records in clusters with size less than k and lines 26-30 treats the clusters with size larger than 2k.

4. Experiments

We evaluate our new I-WFC algorithm compared with greedy k-member in two aspects: information loss and algorithm efficiency.

For experiment datasets, we use Adult dataset from UC Irvine Machine Learning Repository that is considered as a standard benchmark for k-anonymity. Among eight quasi-identifier attributes, age and education-num are treated as numeric attributes and other six as the categorical attributes. We remove the records with missing values represented as "?" from the original dataset firstly.

As shown in Fig. 3 and Fig. 4, our newly proposed algorithm has a good tradeoff between information loss and execution time. Although the execution time is worse than greedy k-member algorithm, we have a much better information loss result due to the consideration of weights adjustment for both numeric and categorical attributes.

5. Conclusions

In this paper, we propose an improved WFC algorithm by taking the weight of categorical and numeric attributes into consideration. Besides, we introduce an new distance measurement for the clustering iteration. Our experiments show that the I-WFC algorithm causes less information loss than greedy k-member algorithm.

Although the execution time does not looks better, this new algorithm can be more feasible in practical situations. We believe that we can improve our algorithm for better efficiency. The other working direction in the future is the other distance measurements for general records that may lead to less information loss.

Acknowledgment

This work was supported by NSFC 60673140 and NHTP (2007AA01Z156, 2008ZX01045, 2009CB320706).

References