


RESEARCH

Open Access



An efficient privacy protection in mobility social network services with novel clustering-based anonymization

Zhi-Guo Chen , Ho-Seok Kang, Shang-Nan Yin and Sung-Ryul Kim*

Abstract

A popular means of social communication for online users has become a trend with rapid growth of social networks in the last few years. Facebook, Myspace, Twitter, LinkedIn, etc. have created huge amounts of data about interactions of social networks. Meanwhile, the trend is also true for offline scenarios with rapid growth of mobile devices such as smart phones, tablets, and laptops used for social interactions. These mobile devices enlarge the traditional social network services platform and lead to a greater amount of mobile social network data. These data contain more private information of individuals such as location, habit, and health condition. However, there are many analytical, sociological, and economic questions that can be answered using these data, so the mobility data managers are expected to share the data with researchers, governments, and/or companies.

Therefore, mobile social network data is badly in need of anonymization before it is shared or analyzed widely. k -anonymization is a well-known clustering-based anonymization approach. However, the implementation of this basic approach has been a challenge since many of the mobile social network data involve categorical data values. In this paper, we propose an approach for categorical data clustering using rough entropy method with DBSCAN clustering algorithm to improve the performance of k -anonymization approach. It has the ability to deal with uncertainty in the clustering process and can effectively find arbitrarily shaped clusters. We will report the proposed approach and discuss the credibility by theoretical studies and examples. And experimental results on two benchmark data sets obtained from UCI Machine Learning Repository show that our approach is second to none among the Fuzzy Centroids, MMeR, SDR and ITDR, etc. with respect to the local and global purity of clusters. Since the clustering algorithm is a key point of k -anonymization for clustering mobile social network data, our experimental results show that our proposed algorithm can be more effective to balance the utility of the mobile social network data and the performance of anonymization.

Keywords: DBSCAN algorithm, Rough entropy, Cluster purity, Categorical data clustering, k -anonymization, Mobile Social network

1 Introduction

Social network service is an evolving platform that focuses on making and maintaining social network or social relations among people who share some common activities or interests. Popular social network services, such as Facebook, Twitter, Myspace, LinkedIn, and many more are the basic carriers of multi-dimensional space which aim at shaping a virtual society for reflecting people's real life and

status in daily life. Besides, with rapid growth of mobile devices, they provide huge processing power to software applications. These devices could supply valuable user information to social network and share geographical location coordinates of the user. However, mobile social network data uploaded into the social network by these mobile devices are very sensitive. It contains lots of private information which lead to privacy leak. Meanwhile, high quality of these mobile social network data is interesting to researchers or companies with many disciplines, such as sociology, psychology, market, or habit

*Correspondence: kimsr@konkuk.ac.kr
Department of Internet and Multimedia Engineering, Konkuk University, Seoul, South Korea

research. Therefore, these data need to be more effectively anonymized so as to protect the private information before it gets published.

In recent years, a simple and practical privacy-preserving anonymization [1–7] were proposed to prevent privacy leak or against identifying individuals. However, a serious issue is that it decreases data quality a lot (information loss) after data is anonymized. k -anonymization is assigning all records into several groups so that each group contains at least k records. The observations in the same groups are similar or identical to the values of their quasi-identifier. Hence, the efficiency and accuracy of assignment affect the information loss and the performance [6] of an anonymous algorithm. Clustering is a useful technique that partitions a set of instances into subsets (called clusters) so that observations in the same cluster are similar to each other. A good clustering algorithm can present high accuracy of assignment. Hence, clustering is a key point of k -anonymization. If we improve the performance of clustering algorithm, we can obtain efficient k -anonymization which can reduce information loss of data [6]. Most of the literature adopt the k -means-based clustering algorithm for k -anonymization [4, 6, 7]. However, k -means and most of clustering algorithms are presented for clustering numerical data using some distance function. Therefore, there has been a big challenging issue for clustering mobile social network data which mostly involve categorical data values. Meanwhile, these data have often no sharp boundary between clusters. Therefore, an algorithm needs to be designed to handle uncertainty in the clustering process. Huang [8] and Kim et al. [9] have proposed some works for applying fuzzy sets in clustering categorical data to solve the uncertainty issue. Shortly afterward, Kumar et al. [10] propose an algorithm (MMeR) which use the basic rough set concepts to handle categorical attribute values as well as the uncertainty of data sets in 2009. In 2011, Panda et al. [5] uses this MMeR algorithm instead of the clustering stage algorithm of OKA [6] and show that MMeR can have a great success to improve the performance of k -anonymization for mobile social network data. In 2011, Tripathy et al. propose termed standard deviation roughness (SDR) [11] for clustering categorical data. In 2015, ITDR [12] proposed by Park et al. and the experimental results demonstrate that it has the highest performance than previous research.

DBSCAN clustering algorithm is proposed by Ester et al. in 1996 [13]. Although it employs distance function for numerical data clustering, it does not specify initial points or cluster number and it can also determine arbitrary shapes. Therefore, for improving the performance of mobile social network categorical data clustering algorithm, we refer to the basic theory of DBSCAN algorithm in our works.

Rough set is a powerful theory proposed by Pawlak in 1982 [14, 15], which is applied to data mining, machine learning, pattern recognition, and feature selection successfully [16–19]. The entropy in information theory proposed by Shannon [20] is a useful mechanism for measuring uncertainty in rough sets. Therefore, many papers are presented which combine rough set theory with Shannon's entropy theory for data labeling and outlier detection [21–23]. Especially, Reddy et al. present data labeling method based on cluster purity using relative rough entropy for categorical data clustering [23]. They apply any clustering algorithm to cluster categorical data into several clusters, then use their proposed method to cluster unlabeled data. Their experimental result demonstrates that it obtains a satisfactory performance. We employ their cluster purity theory with another way to design a novel clustering algorithm for the mobile social network categorical data.

In this paper, we propose a rough entropy method with DBSCAN algorithm for clustering mobile social network data to improve the performance of k -anonymization approach. We employ DBSCAN algorithm with rough entropy method, to calculate the purity of cluster, which can handle uncertainty and improve the performance of categorical data clustering. After adding one data point, if cluster purity is decreased to an acceptable level (threshold λ), then we add this point into the cluster. Subsequently, we use the DBSCAN algorithm to recognize next core point and generate new clusters. Finally, after merging several clusters which have common data points, we can get objective clusters. We have succeeded in showing that the proposed method is able to achieve higher local and global purity as compared to Fuzzy Centroids, MMeR, SDR, and ITDR technique [9–12]. Therefore, our clustering algorithm can improve the efficiency and accuracy of assignment for k -anonymization.

The rest of this paper is organized as follows: in Section 2, we will introduce mobile social network data, rough set theory, and rough entropy, as well as DBSCAN algorithm. Section 3 introduces our proposed method. In Section 4, a data set which is referred to paper [23] is used to illustrate our algorithm. Section 5 compares the performance of our proposed algorithm with other related algorithms by the concept of local and global purity to ensure our algorithm can be more efficiently for mobile social network categorical data clustering. Finally, in Section 6, we will conclude this paper.

2 Related Work

2.1 Mobile social network data

With the rapid development of mobile devices, users upload personal comment and share their location, habit, and emotion into social network service to reflect real life more conveniently. Therefore, these mobile devices

enlarge the traditional social network service platform and lead to huge amounts of mobile social network data. These large data are interesting to government or companies for big data analysis. The comment, location, and published time of these mobile social network data can be used to analyze the popularity of tourist attractions with the change of seasons to improve the quality of tourism environment. People’s attention or habit of these data can be used to manufacture popularity products. These data also can be used to analyze the user’s comments or emotion to adjust some policies by government, etc. However, these huge amounts of published data lead to privacy leak and can even be linked to an individual. Hence, how to balance the data quality and data privacy has become a challenging issue before mobility data managers publish these mobile social network data.

k-anonymization is assigning records into several groups so that each group contains at least *k* records. The observations in the same groups are indistinguishable in their privacy-related attributes (quasi-identifier) so that it cannot be linked to an individual, for protecting private information. Consider the data in Table 1, live location, birth year, citizenship, and nickname are regarded as quasi-identifiers. It can be easy to obtain that there is one person whose nickname is *chen*, who was born in 1988, and live in *Seoul, Gwangjindistrict* originally from *China*. If the table has more detail such as work station, educational background, religion, etc. or other published data which can be linked by quasi-identifiers, we can obtain more information about *chen* and even identify an individual. Therefore, private information of this user such as the time, location, contents of posted comment, habit, and emotion will be leaked.

The purpose of *k*-anonymization is to hide or generalize the values of quasi-identifier in the same cluster before mobility data managers publish it so as to protect personal privacy. Therefore, clustering can be treated as a key point of *k*-anonymization for these mobile social network data.

2.2 Rough set theory

In rough set, a set of data points are stored in a table, this table is referred to as an information system. This information system is defined as a quadruple $IS = (U, A, V, f)$.

U is a non-empty finite set of data points. A is a non-empty finite set of attributes. V is the union of attribute values. $f : U \times A \rightarrow V$ is an information function which associates a unique value of each attribute with every object belonging to U , such that for any $a \in A$ and $x \in U, f(x, a) \in V_a$. V_a is called the value set of attribute a [24–26].

Definition 1 For information system. With any of $P \in A$ there is an associated equivalence relation $IND(P)$ described as follows [24]:

$$IND(P) = \{(x, y) \in U^2 : \forall a \in P, f(x, a) = f(y, a)\} \quad (1)$$

The relation $IND(P)$ is called a *P*-indiscernibility relation. The partition of U is a family of all equivalence classes of $IND(P)$ and is denoted by $U/IND(P)$. If $(x, y) \in IND(P)$, then objects x and y are indiscernible from each other by attributes from P . The equivalence classes of the *P*-indiscernibility relation are denoted $[x]_P^U$.

2.3 Rough entropy

The entropy put forward by Shannon [20] as an effective measure of uncertainty has been wildly used for characterizing the information contents in all sorts of fields.

Rough entropy is an extension of entropy to measure the uncertainty in rough sets.

Definition 2 Give an information system $IS = (U, A, V, f)$. For any $P \subseteq A$, let $IND(P)$ is the equivalence relation as the form of $U/IND(P) = \{P_1, P_2, \dots, P_m\}$. The rough entropy $RE(P)$ of equivalence relation $IND(P)$ is defined by [23]:

$$RE(P) = - \sum_{i=1}^m \frac{|P_i|}{|U|} \log \frac{1}{|P_i|} \quad (2)$$

$|P_i|/|U|$ is denotes the probability of any $x \in U$ being in equivalence class P_i . $1 \leq i \leq m$ and $|M|$ denotes the cardinality of set M .

2.4 DBSCAN algorithm

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by

Table 1 Mobile social network data

User	Live location	Birth year	Citizenship	Nickname	...	Privacy information
1	Seoul, Gangnam	1980	Korean	kim	...	
2	Seoul, Dongdaemun	1985	Korean	park	...	The time, location,
3	Seoul, Gangnam	1982	Korean	lee	...	contents of posted
4	Seoul, Dongdaemun	1985	Korean	jin	...	comment, habit,
5	Seoul, Jongno	1982	Korean	jin	...	emotion, etc.
6	Seoul, Gwangjin	1988	China	chen	...	

Ester et al. in 1996 [13]. Consider a set of points in some space to be clustered. For the purpose of DBSCAN clustering, the points are classified as core point, border point, and noise [27, 28] as follows:

- A point p is a core point if it has more than a specified number of points (MinPts) within ε (Eps).
- A border point has fewer than MinPts within ε (Eps), but is in the neighborhood of a core point p .
- A noise point is any point that is not a core point or a border point.

In Fig. 1 [27], we define $\text{minPts} = 3$ and ε (Eps) = 1. Red points are core point, since at least 3 points surround it in a ε radius. Because core points are all reachable from each other, they form one cluster. Point x_7 and x_8 are border points and reachable from point x_5 and x_6 , thus, also belong to the cluster. Point x_9 is a noise point.

To summarize, if clusters have a common data point (reachable from each other), these clusters can be merged and generate a new cluster (x_9 is not reachable from any other point, it is a noise point).

3 DBASCN based on rough entropy

Rough entropy method is usually used with data labeling and outlier detection. However, researchers give a few attentions to use it in categorical data clustering. In the following paragraph, we will introduce the procedure that performs categorical data clustering using rough entropy with DBSCAN algorithm.

Definition 3 A ratio between total rough entropy (TotalRE) and maximum total rough entropy (MaxRE) is defined as cluster purity with following format [23]:

$$\text{Purity}(C_i) = (\text{TotalRE}(C_i)) / \text{MaxRE} \quad (3)$$

The sum of rough entropy on all attributes of cluster can be calculated as the total rough entropy.

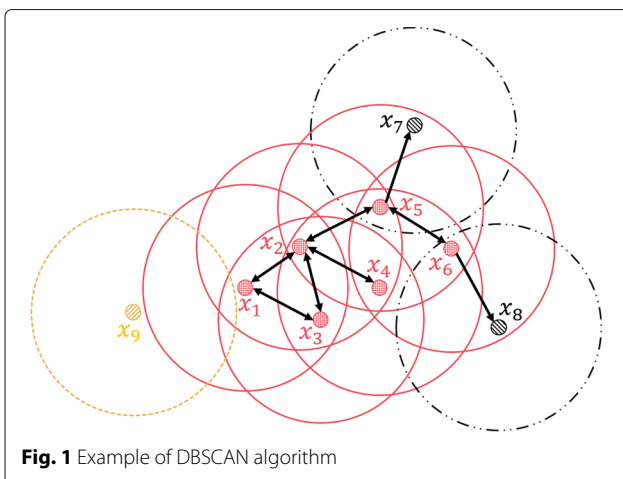


Fig. 1 Example of DBSCAN algorithm

Definition 4 The maximum total rough entropy (MaxRE) of any cluster C_i for any attribute is defined as $\text{MaxRE}(C_i) = K \times \log(m)$. Where, K is the attribute number of data points and m is the number of data points in cluster C_i [23].

This section we will introduce our algorithm which is called RE-based DBSCAN algorithm. The related definitions have been discussed in the previous section. DBSCAN algorithm can automatically find core point and generate clustering based on these points. Since the distance function is not suitable for categorical data, we employ the variation of cluster purity (calculated by rough entropy) to instead of parameter ε (Eps).

Some of social network data may contain a few numerical values. Therefore, we can adopt the method of calculating means between maximum and minimum value to group numerical values into several intervals. Hence, the numerical values can be converted to categorical values.

After DBSCAN algorithm recognizes one core point, an information system $IS = (U, A, V, f)$ can be obtained. Although, this system only has one data point in U (regarded as cluster C_i). After sorting the remaining data points according to the corresponding cluster purity, we can obtain one sorted table $(x_1, x_2, \dots, x_k, \dots, x_n)$ (the cluster purity is calculated by one core point and one of remaining point). Since clustering algorithm is the task for partitions a set of instances into groups, the observations in the same groups are most similar to each other. It does not mean that objects should be identical. Therefore, for any other data point x_k of the sorted table, add x_k into cluster C_i if cluster purity is decreased to an acceptable level (threshold λ), add data point x_k into cluster C_i . Otherwise, DBSCAN algorithm recognizes next core point and generate cluster by the above method. Finally, after merging several clusters which have common data points, we can get objective clusters. If the minimum point number of one cluster is smaller than a threshold value ν , then we can merge this cluster into another new cluster, which has the highest cluster purity.

RE-based DBSCAN algorithm

Input: Dataset D with n data points, a threshold value λ , and minimum point number values ν Output: a set of clusters

- 1 For every data point $x_i \in D(1 \leq i \leq n)$
- 2 Begin
- 3 DBSCAN algorithm recognize one core point, obtain an information system
 $IS = (U, A, V, f)$; (This system only have one data point in U regard as cluster C_i)
- 4 For every data point $x_j \in D(1 \leq j \leq n \text{ and } j \neq i)$, add x_j into cluster C_i

```

5   For every  $a \in A$ 
6   Calculate the partition  $\{U/IND(a)\}$  of two
data point  $x_i$  and  $x_j$ ;
7   Calculate the rough entropy  $RE(a)$  of two data
point  $x_i$  and  $x_j$ ;
8   Calculate Maximum rough entropy  $MaxRE$  of
data point  $x_i$  and  $x_j$ ;
9   Calculate the Cluster purity of  $x_i$  and  $x_j$ ;
10  According to cluster purity, sort data as
 $(x_1, x_2, \dots, x_k, \dots, x_n) (k \neq i)$ 
11  For every sorted data points  $x_k$ , add  $x_k$  into cluster
 $C_i$ ;
12  For every  $a \in A$ 
13  Calculate the partition  $\{U/IND(a)\}$  of cluster
 $C_i$ ;
14  Calculate the rough entropy  $RE(a)$  of cluster
 $C_i$ ;
15  Calculate Maximum rough entropy  $MaxRE$  of
cluster  $C_i$ ;
16  Calculate the Cluster purity of cluster  $C_i$ ;
17  If cluster purity( $C_i$ )  $\geq \lambda$  and cluster
purity( $C_i$ ) decrease than previous one
18  Add data point  $x_k$  into cluster  $C_i$ ;
19  End
20  If several small clusters have common data
points, merge clusters;
21  If  $minPts < \nu$ , according to cluster purity merge
this cluster with other cluster;
22  End
23  Return

```

RE-based DBSCAN algorithm treats every point as core point and estimates the cluster purity to make clusters. The evaluation criteria λ (ε for the DBSCAN algorithm) is equal for every point. Analogous to DBSCAN algorithm, if this value of λ is too large, we will suffer from the overfitting problem. However, our algorithm use the cluster purity to partition the clusters. We can let the threshold $\lambda = N \times \log(m)/K \times \log(m) (N \leq M)$. If $N = K$, λ is equal to 1, it means all values of all attributes are identical. Otherwise, all values of all attributes are different when $N = 0$. Since we do not know the actual cluster number of the unseen data set, we can guess the distribution of clusters to control the threshold value λ so as to avoid the overfitting problem and obtain the object clusters.

4 Certification by example

In this section, we make an example of mobile social network data to facilitate understanding and to ensure the reliability of our clustering algorithm. As shown in Table 2, there are some information of mobile social network data, such as a is live location, b is birth year, c is citizenship.

Table 2 A data set with 14 data points

User	Live location	Birth year	Citizenship	Privacy information
x_1	<i>Korea, Seoul</i>	1988	<i>Korean</i>	
x_2	<i>Korea, Busan</i>	1988	<i>Korean</i>	
x_3	<i>Korea, Seoul</i>	1988	<i>Chinese</i>	
x_4	<i>Korea, Seoul</i>	1988	<i>Chinese</i>	
x_5	<i>Germany, Berlin</i>	1988	<i>German</i>	<i>The time, location,</i>
x_6	<i>Germany, Munich</i>	1988	<i>German</i>	<i>contents of posted</i>
x_7	<i>Germany, Berlin</i>	1999	<i>German</i>	<i>comment, habit,</i>
x_8	<i>Korea, Seoul</i>	1988	<i>Korean</i>	<i>emotion, etc.</i>
x_9	<i>Korea, Busan</i>	2000	<i>American</i>	
x_{10}	<i>Germany, Munich</i>	1999	<i>German</i>	
x_{11}	<i>Korea, Incheon</i>	2000	<i>American</i>	
x_{12}	<i>Germany, Munich</i>	1988	<i>German</i>	

Given an information system $IS = (U, A, V, f)$, where $U = \{x_1, x_2, \dots, x_{14}\}$, $A = \{a, b, c\}$. Let threshold $\lambda = 0.67$ and $\nu = 2$.

For core data point x_1 of DBSCAN, calculate the cluster purity between x_1 and the rest points, and then sorting these points by cluster purity values.

In Table 3, for data point x_1 and x_2 , the partitions induced by all singleton subsets of $a \in A$ are calculated by formula 1;

$$U/IND(a) = \{\{x_1\}, \{x_2\}\}$$

$$U/IND(b) = \{\{x_1, x_2\}\}$$

$$U/IND(c) = \{\{x_1, x_2\}\}$$

According to formula 2, the rough entropy of each attribute of attribute set A can be calculated as follows:

$$RE(a) = - \sum_{i=1}^m \frac{|P_i|}{|U|} \log \frac{1}{|P_i|} = - \left(\frac{1}{2} \log \frac{1}{1} + \frac{1}{2} \log \frac{1}{1} \right) = 0$$

$$RE(b) = - \sum_{i=1}^m \frac{|P_i|}{|U|} \log \frac{1}{|P_i|} = - \left(\frac{2}{2} \log \frac{1}{2} \right) = 0.3010$$

$$RE(c) = - \sum_{i=1}^m \frac{|P_i|}{|U|} \log \frac{1}{|P_i|} = - \left(\frac{2}{2} \log \frac{1}{2} \right) = 0.3010$$

The maximum total rough entropy of cluster C_i is calculated by Definition 4:

$$MaxRE = K \times \log(m) = 3 \log(2) = 0.9031$$

Table 3 Data points in cluster with two unsorted data points

User	Live location	Birth year	Citizenship
x_1	<i>Korea, Seoul</i>	1988	<i>Korean</i>
x_2	<i>Korea, Busan</i>	1988	<i>Korean</i>

The total of rough entropy of two data points is calculated as follows:

$$\text{TotalRE} = \text{RE}(a) + \text{RE}(b) + \text{RE}(c) = 0.602$$

Therefore, the purity of two data points is calculated by using formula 3 as follows:

$$\text{Purity}(C_i) = (\text{TotalRE}(C_i))/\text{MaxRE} = 0.667$$

Analogously, we can find that cluster purity values between x_1 and the rest of the data points are (0.667, 0.667, 0.667, 0.333, 0.333, 0, 1.0, 0, 0, 0, and 0.333). According to these cluster purity values, we sort data points as follows: ($x_8, x_2, x_3, x_4, x_5, x_6, x_{12}, x_7, x_9, x_{10}$, and x_{11}).

Take x_1 as core point of DBSCAN, after adding data point x_8 , the data points of cluster C_i are shown in Table 4:

Analogously, we calculate the cluster purity as 1.0. It means all points in the cluster C_i is same with each other. The cluster is “clean”. Therefore, add data point x_8 into cluster C_i .

After adding data point x_2 , the cluster C_i is: $U = \{x_1, x_8, x_2\}$. Correspondingly, we can obtain that the cluster purity is 0.807 and it is bigger than threshold λ . Therefore, add data point x_2 into cluster C_i . Similarly, continue to add sorted data points until the cluster purity is smaller than threshold λ or cluster purity is increased in cluster C_i . After adding data point x_3 , the data points of cluster C_i are shown in Table 5:

The partitions induced by all singleton subsets of $a \in A$ are calculated by formula 1;

$$U/\text{IND}(a) = \{\{x_1, x_8, x_3\}, \{x_2\}\}$$

$$U/\text{IND}(b) = \{\{x_1, x_8, x_3, x_2\}\}$$

$$U/\text{IND}(c) = \{\{x_1, x_8, x_2\}, \{x_3\}\}$$

Analogously, according to formula 2, the rough entropy of each attribute of attribute set A can be calculated as follows:

$$\text{RE}(a) = -\sum_{i=1}^m \frac{|P_i|}{|U|} \log \frac{1}{|P_i|} = -\left(\frac{3}{4} \log \frac{1}{3} + \frac{1}{4} \log \frac{1}{1}\right) = 0.358$$

$$\text{RE}(b) = -\sum_{i=1}^m \frac{|P_i|}{|U|} \log \frac{1}{|P_i|} = -\left(\frac{4}{4} \log \frac{1}{4}\right) = 0.602$$

$$\text{RE}(c) = -\sum_{i=1}^m \frac{|P_i|}{|U|} \log \frac{1}{|P_i|} = -\left(\frac{3}{4} \log \frac{1}{3} + \frac{1}{4} \log \frac{1}{1}\right) = 0.358$$

Table 4 Data points in cluster with two sorted data points

User	Live location	Birth year	Citizenship
x_1	<i>Korea, Seoul</i>	1988	<i>Korean</i>
x_8	<i>Korea, Seoul</i>	1988	<i>Korean</i>

Table 5 Data points in cluster with 4 sorted data points

U	Live location	Birth year	Citizenship
x_1	<i>Korea, Seoul</i>	1988	<i>Korean</i>
x_8	<i>Korea, Seoul</i>	1988	<i>Korean</i>
x_2	<i>Korea, Busan</i>	1988	<i>Korean</i>
x_3	<i>Korea, Seoul</i>	1988	<i>Chinese</i>

The maximum total rough entropy of cluster C_i is calculated by Definition 4:

$$\text{MaxRE} = K \times \log(m) = 3 \log(4) = 1.8062$$

The total of rough entropy of two data points is calculated as follows:

$$\text{TotalRE} = \text{RE}(a) + \text{RE}(b) + \text{RE}(c) = 1.318$$

Therefore, the purity of two data points is calculated by using formula 3 as follows:

$$\text{Purity}(C_i) = (\text{TotalRE}(C_i))/\text{MaxRE} = 0.729 > \lambda$$

Therefore, add data point x_3 into cluster C_i .

After adding data point x_4 , the cluster C_i is: $U = \{x_1, x_8, x_2, x_3, x_4\}$. Correspondingly, we can obtain that the cluster purity is 0.757. Although this value is over the threshold value λ , this value is increased after adding data point x_4 into cluster C_i . It means that this point is similar to another points (like x_3) rather than core data point x_1 of DBSCAN method. Therefore, we stop this procedure and we can obtain one cluster that contains $\{x_1, x_8, x_2, x_3\}$. Similarly, all clusters of each core point can be found by repeating the above method.

Finally, by DBSCAN algorithm, merge these small clusters if they have common data points. Therefore, the result of our proposed algorithm is $C_1\{x_1, x_2, x_3, x_4, x_8\}$, $C_2\{x_5, x_6, x_7, x_{10}, x_{12}\}$ and $C_3\{x_9, x_{11}\}$.

The data set in Table 2 is given for estimating the performance of our mobile social network data clustering algorithm. We use rough entropy method to calculate the similarity between data points and employ the basic idea of DBSCAN method to merge the clusters which have common data points to generate target clusters. Our algorithm analyzes the change of cluster purity after attempting to add new data points to the cluster. If the reduction of cluster purity keeps at an acceptable level of λ after adding a data point, then add the data point into the cluster. Analogously, if the number of instances in one cluster is under the minimum threshold value ν , this cluster can be merged into other clusters by comparing the cluster purity.

5 Experimental results

In order to prove that our proposed clustering algorithm has a stronger performance so as to balance the utility of

Table 6 Local Purity of clusters in Soybean data

Clusters	C ₁	C ₂	C ₃	C ₄	Sum	Purity
1	10	0	0	0	10	1
2	0	10	0	0	10	1
3	0	0	10	0	10	1
4	0	0	0	17	17	1

data and the efficiency of anonymization, we have implemented the algorithm using JAVA language and tested on several data sets obtained from the UCI Machine Learning Repository which were used in previous clustering works [12]. The local and global purity of clusters is used to measure the quality of the clusters. The global purity of a cluster is defined as:

$$\text{Global purity} = \frac{\text{Number of data occurring in its corresponding class}}{\text{The number of data in the data set}}$$

The local purity is defined as:

$$\text{Local purity} = \frac{\sum_{i=1}^m \text{Purity}(i)}{m}$$

m is the number of clusters obtained from the proposed algorithm.

According to the above measure, a higher value of overall purity and local purity indicates a higher performance of clustering algorithm. Therefore, if these two values are 1, it means clustering algorithm is perfect for clustering data into its corresponding cluster. In this section, our proposed algorithm is compared to six algorithms based on Soybean and Zoo data sets to ensure the superiority and efficiency of our algorithm.

Experiment 1. The Soybean data set contains 47 objects. For each data point, the information of soybean diseases is described by 35 categorical attributes.

This data set totally represents four kinds of soybean diseases. The data set includes 17 objects for describing Phytophthora Rot disease, 10 objects for describing Diaporthe Stem Canker disease, 10 objects for describing Charcoal Rot disease as well as 10 objects for describing Rhizoctonia Root Rot disease [10–12]. Since there are four kinds of diseases, the results based on our algorithm generate four clusters. Table 6 summarizes the results of our algorithm on the Soybean data set.

Table 6 contains information about actual and predicted data distribution by a clustering algorithm. Each column of the Table (*C*₁, *C*₂, ..., *C*₄) represents the instances in an actual cluster, each row (1, 2, ..., 4) represents the instances in a predicted cluster by a clustering algorithm in Soybean data set.

As shown in Table 6, all of 47 objects are correctly classified into its corresponding clusters. Thus, the local and global purity of the clusters is 100%. Kim et al. [9], Kumar et al. [10], Tripathy et al. [11], and Park et al. [12] have compared this data set with different categorical data clustering algorithms. In this research, our algorithm is applied to the same data set, and therefore we can compare with these algorithms to check superiority and efficiency of our proposed algorithm.

Figure 2 shows that our algorithm outperforms Fuzzy Centroids, MMeR, and SDR method on the Soybean data set [12]. Our algorithm has the most significant local and global purity which is 1.0, which means all of the clusters is correctly classified. Since the distribution of clusters is not dense in this data set and even the threshold value *λ* is quite large, it will not suffer from the overfitting problem. Therefore, we can ensure that our algorithm has higher superiority and efficiency than MMeR, SDR algorithm.

Experiment 2. The Zoo data set contains 101 objects. For each data point, the information of an animal is

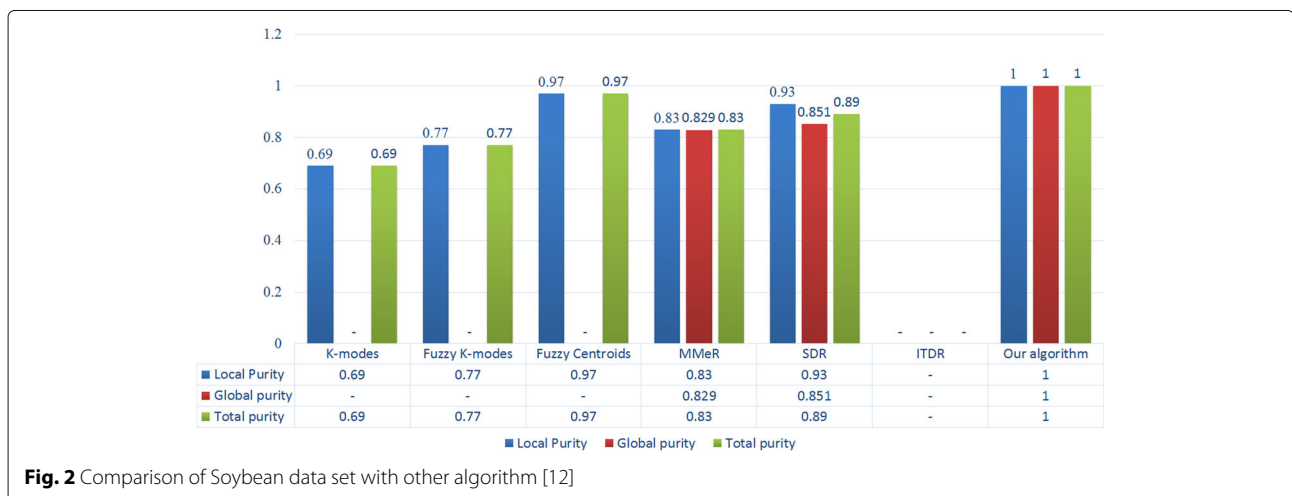


Fig. 2 Comparison of Soybean data set with other algorithm [12]

Table 7 Local purity of clusters in Zoo data

Clusters	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆	C ₇	Sum	Purity
1	27	0	0	0	0	0	0	27	1
2	14	0	0	0	0	0	0	14	1
3	0	0	2	13	0	0	0	15	0.867
4	0	0	0	0	0	0	7	7	1
5	0	20	0	0	0	0	1	21	0.952
6	0	0	0	0	0	8	2	10	0.8
7	0	0	3	0	4	0	0	7	0.571

described by 18 categorical attributes. This data set totally represents seven kinds of animals [10–12]. Therefore, the result based on our algorithm needs to generate seven clusters and compare with other algorithms. Table 7 summarizes the results of run our algorithm on the Zoo data set.

Similar to Table 6, each column of the Table 7 (C₁, C₂, ..., C₇) represents the instances in an actual cluster, each row (1, 2, ..., 7) represents the instances in a predicted cluster by a clustering algorithm in Zoo data set.

For 101 objects, 93 objects are correctly classified into its corresponding clusters. Thus, the global purity of the clusters is 92%.

The comparison of global and local purity are showed as follows:

From Fig. 3, it is clear that our algorithm has the highest purities as the local purity is 88.4%, the global purity is 92%, and total purity is 90.2%. Therefore, our algorithm performs better than Fuzzy Centroids, MMeR, SDR, and ITDR algorithm on the Zoo data set [12]. Our algorithm not only has the more significant local purity but

also has the more significant global purity. It means we decrease misclassification during running the clustering algorithm.

From above two examples, we can conclude that our algorithm is the most efficient algorithm for categorical data clustering. Meanwhile, our experiments show that it is more efficient than MMeR, SDR, and ITDR which employ the rough set theory for clustering. Therefore, our clustering algorithm is suitable for *k*-anonymization and will reduce the information loss and provide more useful mobile social network data to the government or companies.

6 Conclusions

Social networks are growing rapidly with the development of mobile devices. These devices supply valuable user information to social network including the user's geographical location coordinates. Meanwhile, mobile social network data provide interesting opportunities to researchers and/or companies with many disciplines, such as sociology, psychology, market, or habit research. Therefore, these data are badly in need of anonymization before it gets published by mobility data managers. For balancing the utility of these data and the performance of anonymization, we propose a new clustering method for anonymous mobile social network data. Our algorithm needs no initial points and cluster number from users and it can handle the impreciseness of data. The theoretical studies and experimental results show that our method is more efficient than most of the related algorithms including MMeR, SDR, and ITDR which are the algorithms based on rough set theory. We believe that our approach is useful and credible for anonymization. The further work will focus on gathering high dimension mobile social network data to estimate the feasibility of our proposed approach.

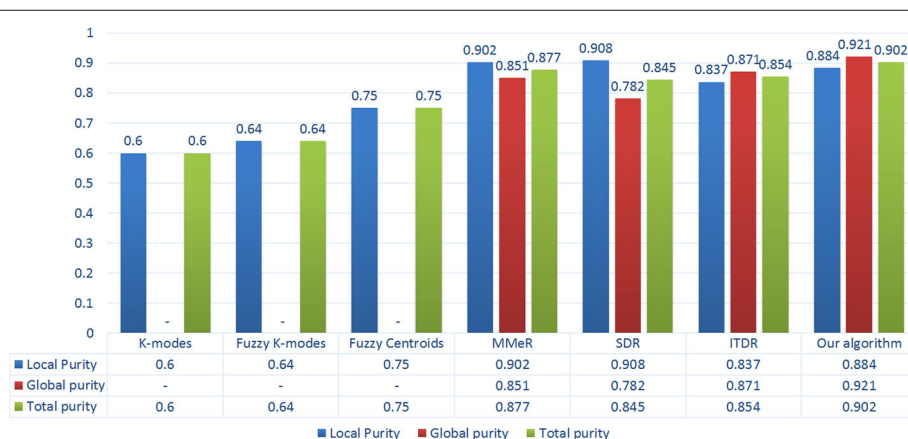


Fig. 3 Comparison of Zoo data set with other algorithm [12]

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A1B02011964).

Competing interests

The authors declare that they have no competing interests.

Received: 31 May 2016 Accepted: 2 November 2016

Published online: 29 November 2016

References

- L Sweeney, k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowledge-Based Syst.* **10**(05), 557–570 (2002)
- L Sweeney, Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowledge-Based Syst.* **10**(05), 571–588 (2002)
- G Aggarwal, T Feder, K Kenthapadi, S Khuller, R Panigrahy, D Thomas, A Zhu, in *Proceedings of the Twenty-fifth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. Achieving anonymity via clustering (ACM, New York, 2006), pp. 153–162
- X Xu, M Numao, in *2015 Third International Symposium on Computing and Networking (CANDAR)*. An efficient generalized clustering method for achieving k-anonymization (IEEE, Sapporo, 2015), pp. 499–502
- G Panda, B Tripathy, S Jha, Security aspects in mobile cloud social network services. *Int. J.* **2**(1) (2011)
- J-L Lin, M-C Wei, in *Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society*. An efficient clustering method for k-anonymization (ACM, New York, 2008), pp. 46–50
- J-W Byun, A Kamra, E Bertino, N Li, in *Advances in Databases: Concepts, Systems and Applications*. Efficient k-anonymization using clustering techniques (Springer, Berlin, 2007), pp. 188–200
- Z Huang, Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowl. Discov.* **2**(3), 283–304 (1998)
- DW Kim, KH Lee, D Lee, Fuzzy clustering of categorical data using fuzzy centroids. *Pattern Recognit. Lett.* **25**(11), 1263–1271 (2004)
- P Kumar, B Tripathy, Mmer: an algorithm for clustering heterogeneous data using rough set theory. *Int. J. Rapid Manuf.* **1**(2), 189–207 (2009)
- B Tripathy, A Ghosh, in *Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE Sdr*: An algorithm for clustering categorical data using rough set theory (IEEE, Trivandrum, 2011), pp. 867–872
- I-K Park, G-S Choi, Rough set approach for clustering categorical data using information-theoretic dependency measure. *Inf. Syst.* **48**, 289–295 (2015)
- M Ester, H-P Kriegel, J Sander, X Xu, in *Kdd*. A density-based algorithm for discovering clusters in large spatial databases with noise, vol. 96 (AAAI, Portland, 1996), pp. 226–231
- Z Pawlak, Rough sets. *Int. J. Comput. Inf. Sci.* **11**(5), 341–356 (1982)
- Z Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data vol.9*. (Springer Science & Business Media, Berlin, 2012)
- TY Lin, N Cercone. *Rough Sets and Data Mining: Analysis of Imprecise Data* (Springer Science & Business Media, Berlin, 2012)
- Y Kaya, M Uyar, A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. *Appl. Soft Comput.* **13**(8), 3429–3438 (2013)
- BB Nair, V Mohandas, N Sakthivel, A decision tree rough set hybrid system for stock market trend prediction. *Int. J. Comput. Appl.* **6**(9), 1–6 (2010)
- Y Qian, J Liang, W Pedrycz, C Dang, Positive approximation: an accelerator for attribute reduction in rough set theory. *Artif. Intell.* **174**(9), 597–618 (2010)
- CE Shannon, A mathematical theory of communication. *ACM SIGMOBILE Mobile Comput. Commun. Rev.* **5**(1), 3–55 (2001)
- F Jiang, Y Sui, C Cao, An information entropy-based approach to outlier detection in rough sets. *Expert Syst. Appl.* **37**(9), 6338–6344 (2010)
- X Li, F Rao, An rough entropy based approach to outlier detection. *J. Comput. Inf. Syst.* **8**(24), 10501–10508 (1050)
- HV Reddy, S Viswanadha Raju, P Agrawal, in *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference On*. Data labeling method based on cluster purity using relative rough entropy for categorical data clustering (IEEE, Mysore, 2013), pp. 500–506
- L Polkowski, S Tsumoto, TY Lin. *Rough set methods and applications: new developments in knowledge discovery in information systems vol.56 Physica* (Springer Science & Business Media, Berlin, 2012)
- F Jiang, Y Sui, C Cao, A rough set approach to outlier detection[J]. *Int. J. General Syst.* **37**(5), 519–536 (2008)
- F Jiang, Z Zhao, Y Ge, *A supervised and multivariate discretization algorithm for rough sets[M]//Rough Set and Knowledge Technology*. (Springer, Berlin, Heidelberg, 2010), pp. 596–603
- Visualizing DBSCAN Clustering. <http://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>. Accessed Jan 2016
- V Panchami, N Radhika, in *Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on The*. A novel approach for predicting the length of hospital stay with dbscan and supervised classification algorithms (IEEE, Bangalore, 2014), pp. 207–212

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Immediate publication on acceptance
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com