Private Trajectory Data Publication for Trajectory Classification

Huaijie Zhu[1,2](✉), Xiaochun Yang[3], Bin Wang[3], Leixia Wang[3],
and Wang-Chien Lee[4]

[1] School of Data and Computer Science, Sun Yat-Sen University, Guangzhou, China
zhuhuaijie@mail.sysu.edu.cn
[2] Guangdong Key Laboratory of Big Data Analysis and Processing,
Guangzhou, China
[3] School of Computer Science and Engineering, Northeastern University,
Shenyang, China
{yangxc,binwang,wleixia}@mail.neu.edu.cn
[4] The Pennsylvania State University, State College, USA
wlee@cse.psu.edu

Abstract. Trajectory classification (TC), i.e., predicting the class labels of moving objects based on their trajectories and other features, has many important real-world applications. Private trajectory data publication is to anonymize trajectory data, which can be released to the public or third parties. In this paper, we study *private trajectory publication for trajectory classification (PTPTC)*, which not only preserves the trajectory privacy, but also guarantees high TC accuracy. We propose a private trajectory data publishing framework for TC, which constructs an anonymous trajectory set for publication and use in data services to classify the anonymous trajectories. In order to build a "good" anonymous trajectory set (i.e., to guarantee a high TC accuracy), we propose two algorithms for constructing anonymous trajectory set, namely Anonymize-POI and Anonymize-FSP. Next, we employ *Support Vector Machine* (SVM) classifier to classify the anonymous trajectories. Finally, the experimental results show that our proposed algorithms not only preserve the trajectory privacy, but also guarantee a high TC accuracy.

Keywords: Trajectory classification · Private trajectory · Classification accuracy

1 Introduction

With the wide deployment of GPS and RFID technologies, a tremendous amount of trajectory data have been generated and used for real-world applications. *Trajectory classification (TC)*, a fundamental mining task, is essential for trajectory data analytic. TC, aiming to predict the class labels of moving objects based on their trajectories and derived features, has many important applications, such as

W. Ni et al. (Eds.): WISA 2019, LNCS 11817, pp. 347–360, 2019.
https://doi.org/10.1007/978-3-030-30952-7_35

city and transportation planning; road construction, design, and maintenance; and traffic congestion detection.

Although a lot of research works [8,9,20–22,24] on TC have been reported in the literature, to the best knowledge of the authors, there is no existing work on privacy preserving trajectory data for TC. To classify trajectory data, the privacy of users whose trajectories being collected may be jeopardized through the necessary analysis of trajectory data. If the trajectory data owner outsources the original trajectory data (without being preprocessed using any privacy protecting schemes) to some data mining vendors (or researchers) to develop predictive models for TC, the users privacy is exposed. For example, trajectory data may be explored to predict some symptoms of illness. Table 1 shows some sample trajectories collected from patients with different symptoms of illness. As shown, those records only contains record id RID, trajectory, and the illness symptoms of patients, where RID is a unique random ID, corresponding to an individual. The first record in the table indicates that patient with $RID = 1$ has visited location p_1, p_4, p_3 and p_6 at time-stamps 2, 4, 5 and 8, respectively. Assume that an adversary has the knowledge that the patient Bob has been to location p_1 and p_6 at time-stamps 2 and 8, respectively. By matching the records, the adversary may find that Bob has an HIV. This example shows that publishing user trajectory data (without any anonymization) may cause serious privacy threats. While many research studies [1,5,10,13,15] have worked on the issues of private trajectory data publication, to the best knowledge of the authors, these works have not taken the TC accuracy into consideration.

Table 1. Patients' trajectory data

RID	Trajectory	Symptoms of illness
1	(p_1, 2)→ (p_4, 4)→(p_3,5)→(p_6, 8)	HIV
2	(p_4, 1)→(p_6, 2)→(p_7, 6)→(p_2, 7)→(p_5, 8)→(p_3, 10)	Flu
3	(p_2, 2)→(p_7, 3)→(p_3, 6)→(p_6, 8)→(p_5, 9)	Flu
4	(p_6, 5)→(p_7, 7)→(p_3, 10)	HIV
5	(p_6, 3)→(p_1, 4)→(p_7, 6)→(p_8, 9)	Fever

In this paper, we study the problem of *private trajectory publication for trajectory classification* (PTPTC), which not only protects the trajectory privacy, but also guarantees a high TC accuracy. For protecting the trajectory privacy, we consider the classical Trusted Third Party framework. The data holder sends the original trajectory data to the trusted third party, and the trusted third party produces anonymous trajectories to a classifier server to train a classifier for TC. As a trade-off between user privacy and TC accuracy is expected, a major challenge faced in addressing PTPTC lies in a potential conflict of interests, i.e., the trajectory classifier requires as precise data as possible, while the users do not want to disclose their exact movements.

In order to well address the issues of private trajectory data publication for TC problem, we first analyze "what affects the accuracy of TC the most for a classifier". As we known, feature selection is crucial for improving the TC accuracy. Effective TC depends on the discriminative features. Firstly, based on the similarity of Points of Interest (POIs) (i.e., as one of features) of trajectories, we propose a trajectory similarity, called trajectory POI similarity. Based on this similarity, we propose a (k, γ)-anonymized trajectory set. Accordingly, we design an algorithm, called Anonymize-POI, to construct a (k, γ)-anonymized trajectory set. Notice that *frequent sequential patterns* are considered as the best features in most of the-state-of-the-art methods [2,8,9,17,20], as it preserves the order of visiting sequence of trajectories. Similarly, the frequent sequential pattern as the feature is also an alternative way in our classifier model, which tips us regarding how to maintain frequent sequential patterns while protecting the trajectory privacy. Thus, we propose a novel *trajectory similarity* function based on frequent sequential patterns. Accordingly, we define (k, γ)-anonymized trajectory set for trajectory privacy preservation. Meanwhile, we propose an effective *FSP-based (k, γ)-anonymized trajectory set constructing* algorithm, called Anonymize-FSP. After anomynizing all the trajectories, i.e., satisfying the k-anonymity, we employ *Support Vector Machine* (SVM) classifier to classify the anonymous trajectories. Finally, experimental results show that the proposed algorithms not only preserve trajectory privacy, but also guarantee a high TC accuracy.

In summary, the primary contributions summarized as follows made in this paper are four-fold:

- We formalize the private trajectory data publication for TC problem. To the best of our knowledge, this work is the first attempt to tackle the PTPTC problem.
- According to the *trajectory similarity* based on POIs, we propose *(k, γ)-anonymized trajectory set* for trajectory privacy preservation. For privacy preserving in TC, we design an algorithm called *Anonymize-POI*, for constructing a POI-based (k, γ)-anonymized trajectory set.
- Alternatively, we propose a novel *trajectory similarity* function based on *frequent sequential patterns* and we propose an algorithm, called Anonymize-FSP, for constructing FSP-based (k, γ)-anonymized trajectory set.
- At last, experimental result shows that our algorithms are efficient to protect trajectory privacy while guaranteeing high TC accuracy. Meanwhile, our study shows that Anonymize-FSP achieves a higher TC accuracy than Anonymize-POI.

2 Related Work

In recent years, in order to preserve user privacy in trajectory data, a lot of research studies have been reported in the literature. Generally speaking, research on trajectory privacy can be classified into trajectory privacy in location-based Service and trajectory publication.

Trajectory Privacy in Location-Based Service. In LBS, a lot of research works [11,12] study the spatial cloaking [5,25] techniques to protect trajectory privacy. Then, a distortion-based approach [12] has been proposed to aim at overcoming the drawbacks of the group-based approach. It not only requires querying users to report their locations to the location Another way to ensure k-anonymity is to use individuals' historical footprints instead of their real-time locations [15]. A footprint is defined as a user's location collected at some point of time. Similar to the previous two approaches, a fully-trusted location anonymizer is placed between users and LBS service providers to collect users' footprints. The concept mix-zones [3] have been extended to LBS for preserving trajectory privacy. Without relying on a trusted third party to perform anonymization, a mobile user can generate fake spatial trajectories, called dummies, to protect its privacy [19].

Trajectory Privacy in Trajectory Publication. In this section, we introduce privacy-preserving techniques for trajectory data publication. The anonymized trajectory data can be released to the public or third parties for answering spatio-temporal range queries and data mining. The clustering-based approach [16] utilizes the uncertainty of trajectory data to group k co-localized trajectories within the same time period to form a k-anonymized aggregate trajectory. Also, the generalization-based algorithm [10] first generalizes a trajectory data set into a set of k-anonymized trajectories. After that, generalization-based algorithm [14] with Enhanced l-Diversity [18] is proposed. Another direction is to use differential privacy techniques [7].

In summary, existing privacy-preserving techniques for spatial trajectory publication only support simple aggregate analysis, such as range queries. None of these private trajectory data publication works take TC into consideration.

3 Preliminaries

In this section, we introduce the background needed for our work, and then formulate our PTPTC problem.

3.1 Background and Definitions

In this work, we focus on TC. Give a set of trajectories $T = \{tj_1, tj_2, \ldots, tj_{num\_tra}\}$, with each trajectory associated with a class label $c_i \in C = \{c_1, \ldots, c_{num\_cat}\}$, where the trajectory, *trajectory sequences* and some other features are defined as follows:

Definition 1 *(Trajectory). A trajectory tj is a sequence of spatio-temporal points. Formally, $tj = \{ID, (x_1, y_1, t_1), \ldots, (x_n, y_n, t_n)\}$ ($t_1 < \ldots < t_n$), where ID represents a trajectory, and time t_i ($1 \leq i \leq n$) denotes that user passes by location (x_i, y_i) at time t_i.*

Definition 2 *(Point of Interests, POIs). POI is a 2-tuple (ID, loc), where ID represents the ID of trajectory which this POI belongs to, and loc = (x, y) represents the geography coordinates of this POI.*

Since the points denoting the same POI may have different position values, all the position values in a certain range are normalized to denote the same POI.

Definition 3 *(Trajectory Sequence, TS)***.** *TS means the sequence of POIs in accordance with the order of time-stamp t_i.*

Table 2. Trajectory data

RID	POI sequence	Sequential pattern	Maximum sequential subPattern
tj_1	$\langle p_1, p_2, p_5\rangle$	$\langle p_1, p_2, p_5\rangle$	$\langle p_1, p_2, p_5\rangle$
tj_2	$\langle p_1, p_2, p_3, p_5\rangle$	$\langle p_1, p_2, p_5\rangle$, $\langle p_3, p_5\rangle$	$\langle p_1, p_2, p_5\rangle$, $\langle p_3, p_5\rangle$
tj_3	$\langle p_4, p_2, p_6, p_5\rangle$	$\langle p_4, p_5\rangle$,$\langle p_6, p_5\rangle$	$\langle p_4, p_5\rangle$, $\langle p_6, p_5\rangle$, $\langle p_2, p_5\rangle$
tj_4	$\langle p_1, p_7, p_8, p_5\rangle$	*null*	$\langle p_1, p_5\rangle$
tj_5	$\langle p_6, p_4, p_3, p_5\rangle$	$\langle p_4, p_5\rangle$,$\langle p_6, p_5\rangle$,$\langle p_3, p_5\rangle$	$\langle p_4, p_5\rangle$,$\langle p_6, p_5\rangle$,$\langle p_3, p_5\rangle$

In order to define the frequency of TS, we give some necessary notations. Assume there are two TSs $\alpha = \langle a_1, a_2, \ldots; a_m\rangle$ and $\beta = \langle b_1; b_2; \ldots; b_n\rangle$. β contains α, denoted as $\alpha \sqsubseteq \beta$, iff $\exists k_1; k_2; \ldots; k_m$, $1 \leq k_1 < k_2 < \ldots < k_m \leq n$ and $a_1 = b_{k1} \wedge a_2 = b_{k2} \wedge \ldots \wedge a_m = b_{km}$. T_α denotes the set of trajectories T which contains α, i.e., $T_\alpha = \{tj | tj \in T \wedge \alpha \sqsubseteq tj\}$. In a set of TSs, a TS α is *maximal* if α is not contained in any other sequence. We now define *frequent sequential pattern* in Definition 4.

Definition 4 *(Frequent sequential pattern, FSP)***.** *Trajectory sequence ξ is a frequent sequential pattern if $\theta_\xi = |T_\alpha|/|T| \geq \theta_0$ and ξ is maximal, where θ_ξ is the frequency of ξ, and θ_0 is the minimum support threshold (0 < θ_0 <1). The set of frequent sequential patterns is denoted as $\Re$.*

FSPs are extracted form the set of trajectories T. Each trajectory may share different numbers of FSPs.

Example 1. Table 2 illustrates the FSPs. As shown, with minimum support set to 25%, i.e., a minimum support must contain at least 2 trajectories. Notice that the length of one FSP is not less than 2. Sequence $\langle p_1, p_2, p_5\rangle$, $\langle p_3, p_5\rangle$, $\langle p_6, p_5\rangle$ and $\langle p_4, p_5\rangle$ are the FSPs which satisfy the minimum support. In detail, $\langle p_1, p_2, p_5\rangle$ is supported (contained) by trajectories 1 and 2. The other FSPs contained in the other four trajectories are shown in the third column of Table 2.

To propose the private trajectory data publication methods, we define *$(k, \gamma$-anonymized trajectory set* as follow.

Definition 5 *((k, γ)-anonymized Trajectory Set)***.** *A trajectory set F is (k, γ)-anonymized trajectory set, if F consists at least k trajectories and the similarity of any two trajectories is not less than γ.*

3.2 Problem Definition

(*User Identification Problem* (Attack Model)). Given a trajectory database, trajectory tj_i exists *user identification problem* iff attacker utilizes the background knowledge to find who has visited some locations at certain time to identify the user of tj_i.

In this paper, we study the PTPTC problem. Based on this attack model, our goal is to preserve the trajectory privacy to prevent the attackers identifying the user of any trajectory in published trajectory set T. For the sake of preventing the attackers identifying the user of any trajectory, we focus on how to construct the good (k, γ)-anonymized trajectory set.

4 Private Trajectory Data Publication for Trajectory Classification Approach

In this section, we propose private trajectory data publication for TC method, based on (k, γ)-anonymized trajectory set. The proposed method consists of two main phases: (1) constructing the anonymous trajectory set; and (2) classifying the anonymous trajectories.

4.1 Constructing the Anonymous Trajectory Set

In this section, we propose the algorithms for constructing an anonymous trajectory set. In order to build a good anonymous trajectory set for guaranteeing high TC accuracy, we first analyze the essence of TC. One of the most important requirements for effective TC is to identify discriminative features. As the extensively used features, i.e., POIs and FSPs, we propose two algorithms for constructing anonymous trajectory set, namely *Anonymize-POI* and *Anonymize-FSP*.

Anonymize-POI Algorithm. According to POIs contained in a trajectory, we define the *trajectory POI similarity*.

Definition 6 *(Trajectory POI similarity). For two trajectories, the number of POIs shared in them is M, whereas the total number of POIs contained in them is N. Therefore, the trajectory POI similarity is M/N, denoted by Sim_POI.*

Based on trajectory POI similarity, we construct the (k, γ)-anonymized trajectory set, as defined in Definition 5. Thus, such a (k, γ)-anonymized trajectory set preserves the trajectory privacy according to the k-anonymity property. Accordingly, we develop an algorithm for constructing the POI-based (k, γ)-anonymized trajectory set, called Anonymize-POI. Anonymize-POI performs the following three steps: (i) obtain all the POIs; (ii) according to the similarity function, group all the trajectories; and (iii) anonymize the trajectories in each group.

In order to measure a trajectory, we first extract all the POIs, via the function getPOI. We consider the POIs in this paper, which consists of starting point,

Algorithm 1. Anonymize algorithm

```
Input: A group of trajectories $g$, and anonymous value $k$;
Output: A group of anonymous trajectory set $g'$;
if $g.size >= k$ then
    for Each trajectory $tj$ in $g$ do
        Randomly select one trajectory to replace it and attach original $ID$;
        Put the generated trajectory into $g'$;
else
    for $|g|$ to $k$ do
        Randomly choose one trajectory from $g$;
        Construct a new trajectory via copying the chosen trajectory and attach a new id;
        Put the generated trajectory into $g'$;
Return $g'$;
```

ending point, staying point and turning point. Our privacy protection of entire trajectory goal is by protecting these sensitive points. After processing all the POIs in each trajectory, we utilize the trajectory POI similarity to measure and cluster the trajectories. Since how many groups to be finally obtained and clustered is unknown, we adopt a hierarchical clustering method to cluster the trajectories (using a bottom-up strategy). Finally, we anonymize the trajectories in each group. Algorithm 1 presents the pseudo-code of Anonymize. The main idea of Anonymize algorithm is as follows. If the size of group g is not less than the anonymous value k, we randomly use one of the trajectories to replace each trajectory and keep the original trajectory id ID; Otherwise, we construct $k-|g|$ fake trajectories.

Let's show an running example for illustration of the Anonymize-POI algorithm. Recall the trajectories in Table 2, assuming that the similarity threshold γ is 0.5, and the anonymous value $k=2$. According to the POI similarity function, the groups obtained by the clustering method are {1, 2}, {3, 5}, {4}. For group {1, 2}, {3, 5}, final anonymous set {2, 2}, {5, 5} are obtained using function Anonymize. While for {4}, we construct a fake trajectory with id 6 and form the group {6, 4}. Therefore, the final 2-anonymized trajectory set is {2, 2}, {5, 5} and {4, 4}.

Anonymize-FSP Algorithm. Since POI similarity does not capture well the features of trajectory, using POI similarity does not guarantee high precision of TC. Notice that, good features play a key role in TC. As many state-of-the-art TC algorithms explore *frequent sequential pattern*, which preserves the order of visiting sequence of trajectories, we explore the frequent sequential patterns in designing a good similarity to serve the group function. Thus, we propose a novel trajectory similarity function, which takes the frequent sequential patterns into consideration, called Sim_FSP, as shown in Definition 7.

Definition 7 *(Trajectory FSP similarity).* *For two trajectories tj_i, $tj_k \in T$ and a minimum support threshold θ, the number of FSPs shared in R_{tj_i} and R_{tj_k} is M, whereas the total number of FSPs contained in tj_i and tj_k is N. Then, the trajectory FSP similarity is M/N, denoted by Sim_FSP.*

Recall the trajectories data in Table 2. $Sim\_FSP(tj_1, tj_2)$ is 0.5 according to Sim_FSP equation, $Sim\_FSP(tj_3, tj_5)$ is 0.5, and the similarity between trajectory 4 and other trajectories is 0. From this example, the FSP similarity function is not a good choice for grouping the trajectories, which leads the special case that the similarity score between two trajectories is 0. There exists two situations: (a) some trajectories only contain a subset of FSP (not the whole FSP), but the subset is not the maximal; (b) the sequential patterns shared in two trajectories are not supported by enough number of trajectories (i.e., the support degree is not high). For example, trajectory 4 contains the subset $\langle p_1, p_5\rangle$ of FSP $\langle p_1, p_2, p_5\rangle$. For the first situation, we give a new concept, *maximal FSP subset (MFS)*, which is the maximum subset of the FSP with respect to a trajectory. With regard to trajectory 4, $\langle p_1, p_5\rangle$ is the maximal FSP subset for itself. The maximal FSP subsets for other trajectories are shown in the rightmost part of Table 2. For two subsets of the same FSP, we think they are similar in some const. Thus, we utilize a function to measure the similarity between two maximal FSP subsets, r_i and r_k, which satisfy $r_i \subseteq r_k$ or $r_k \subseteq r_i$. The similarity function between r_i and r_k is as follow:

$$Sim(r_i, r_k) = \begin{cases} \frac{|r_i \bigcap r_k|}{|r_i \bigcup r_k|}, \, r_i \subseteq r_k || r_k \subseteq r_i \\ 0, \qquad\qquad\qquad\qquad else \end{cases} \tag{1}$$

According to Eq. 1, the similarity score between $\langle p_1, p_5\rangle$ and $\langle p_1, p_2, p_5\rangle$ is 2/3. Thanks to maximal FSP subset similarity function, it is not necessary for each trajectory to have FSPs and the subsets of FSP can also contribute to the similarity computation. Based on the similarity between two maximal FSP subsets, we propose another trajectory similarity function, as defined below.

Definition 8 *(Trajectory MFS similarity).* *For two trajectories tj_i, $tj_k \in T$ and a minimum support threshold θ_0, maximal FSP subset obtained in tj_i is denoted by RS_{tj_i}, and the maximal FSP subset obtained in tj_k is denoted by RS_{tj_k}. Then, the trajectory MFS similarity is defined in Eq. 2.*

$$Sim\_MFS(t_i, t_j) = \frac{\sum_{r_i \in RS_{tj_i}, r_k \in RS_{tj_k}} Sim(r_i, r_k)}{|R_{tj_i} \bigcup R_{tj_k}|} \tag{2}$$

Example 2. As shown in Table 2, according to trajectory MFS similarity, the similarity score between trajectory 1 and 2 is 1/2, the similarity score between trajectory 1 and 4 is 1/3 and the similarity score between trajectory 3 and 5 is $\frac{1+1}{1+1+2} = 1/2$.

Accordingly, based on trajectory MFS similarity, we design another (k, γ)-anonymized trajectory set. Then we construct (k, γ)-anonymized trajectory sets from original trajectories. Accordingly, we develop an effective algorithm for constructing *FSP-based* (k, γ)*-anonymized trajectory set*, called Anonymize-FSP. The Anonymize-FSP can be divided into four steps: (1) obtaining the POIs for each trajectory; (2) mining all the FSPs and find the maximal FSP subsets for each trajectory using VMSP algorithm [6]; (3) grouping all the trajectories according to the similarity function; and (4) anonymizing the trajectories in each group.

4.2 Classifying the Anonymous Trajectories

After the third trust party anonymizes the original trajectories, classifier server utilize the SVM model [4] to classify the anonymous trajectories.

5 Experiments

5.1 Experimental Setup

In this section, we evaluate the proposed algorithms, which are implemented in Java on an Intel Core 8 Duo CPU E7500 2.93 GHz PC with 6 GB RAM. The experiments are conducted using both synthetic and real datasets. Synthetic dataset (including 2012 trajectories) is generated using the software Network-based Generator of Moving Objects on website "http://iapg.jade-hs.de/personen/brinkhoff/generator/" based on Oldenburg dataset. The real dataset [23] contains 17621 GPS trajectories. Each trajectory is labelled as *driving*, *by bus*, *riding* and *walking*. All datasets are normalized in order to design the experiments.

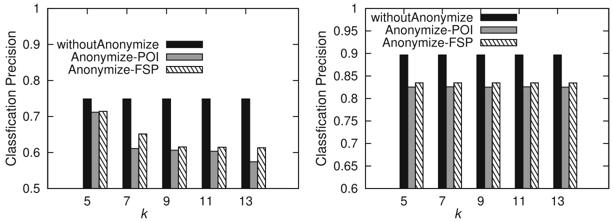
In order to investigate the efficiency and accuracy of the proposed algorithms, the experiments are carried out by varying various parameters, which are summarized in Table 3. In each experiment, we test one parameter at a time (by fixing the other parameters at their default values (i.e., in bold)). The metrics in our experimental study include *classification accuracy* on anonymous trajectories and *constructing time* of anonymous trajectories. For TC accuracy, we compared three algorithms, namely withoutAnonymize, Anonymize-POI and Anonymize-FSP.

Table 3. Parameter ranges and defaults values

Parameter	Range
k (anonymous value)	5, **7**, 9, 11, 13
Similarity threshold δ	0.4, 0.5, **0.6**, 0.7, 0.8
A minimum support threshold θ_0	0.005, **0.006**, 0.007, 0.008, 0.01

5.2 Classification Accuracy of Anonymous Trajectories

In this section, we evaluate the accuracy of the proposed anonymity algorithms, including withoutAnonymize, Anonymize-POI and Anonymize-FSP. We use SVM to classify the trajectories. We measure the TC accuracy corresponding to three different parameters: (a) anonymous value k; (b) similarity threshold γ and (c) minimum support threshold θ.

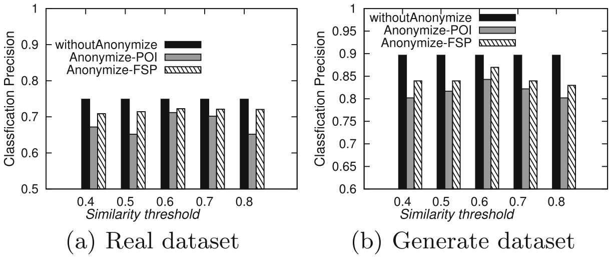


(a) Real dataset (b) Generate dataset

Fig. 1. Classification accuracy vs. k value

Effect of Anonymous Value *k*. In this experiment, we compare the accuracy of classifying the anonymous trajectories constructed by three algorithms under different anonymous k values. The results for classifying different anonymous trajectories are depicted in Fig. 1. Figure 1(a) shows the result on Real dataset. As shown, the accuracy of classifying the trajectories without anonymizing (i.e., withoutAnonymize) is more than 0.75. Obviously, it is not affected by the k values. The accuracy of classifying the anonymous trajectories generated by Anonymize-FSP algorithm is nearly 0.65. Therefore, the accuracy results of trajectories using withoutAnonymize and Anonymize-FSP are very close. This is because the Anonymize-FSP algorithm not only protects the privacy of the trajectories, but also maintains some good TC features for TC. As expected, the TC accuracy become worse when the anonymous value k increases. This shows that if we want to preserve more privacy about the anonymous data (i.e., k increases), the TC accuracy becomes lower. The results obtained on Generate dataset are shown in Fig. 1(b). Compared to the accuracy results on Real dataset, the TC accuracy is higher using different anonymization algorithms, as the generated trajectories are distributed more uniformly than Real data. The observed trend and conclusion on Generate dataset are consistent with the results on Real dataset.

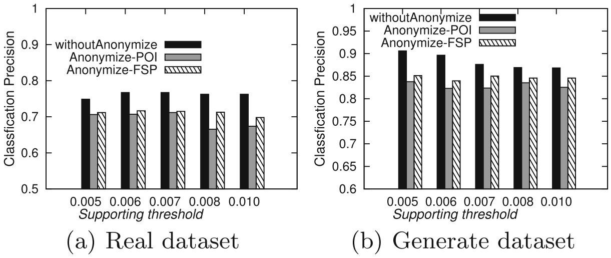
Effect of Similarity Threshold γ. Figure 2 compares the TC accuracy using different anonymization algorithms by varying the similarity threshold γ. For Real dataset, the results in Fig. 2(a) show the superiority of Anonymize-FSP algorithm over Anonymize-POI. It can be also seen that the TC accuracy using



(a) Real dataset (b) Generate dataset

Fig. 2. Classification accuracy vs. similarity threshold

without Anonymize is not affected by increasing the similarity threshold. By increasing the similarity threshold, the accuracy of classifying the anonymous trajectories generated by the proposed two algorithms first becomes higher, and then becomes lower.



(a) Real dataset (b) Generate dataset

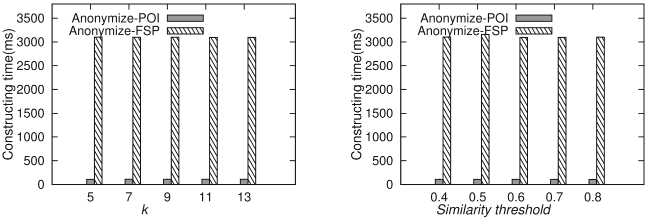
Fig. 3. Classification accuracy vs. minimum support threshold

Effect of Minimum Support Threshold θ. We then compare the accuracy of classifying different anonymous trajectories by varying minimum support threshold and show the results in Fig. 3. The accuracy results on Real dataset are shown in Fig. 3(a), where the TC accuracy of the trajectories using withoutAnonymize is more than 0.75, and it is affected when changing different minimum support thresholds. The TC accuracy using Anonymize-FSP algorithm is 0.7 under different minimum support thresholds. Moreover, it becomes higher first and then become lower when the minimum support threshold is increasing. Moreover, the Anonymize-FSP shows the superiority to Anonymize-POI for constructing the anonymous trajectories. While the TC accuracy results on Generate dataset in Fig. 3(b) is higher than on Real dataset under each minimum support threshold. This is because the FSPs can be a good feature when classifying the trajectories which are uniformly generated.

5.3 Constructing Time of Anonymous Trajectories

Finally, we evaluate the constructing time of the proposed anonymization algorithms, including withoutAnonymize, Anonymize-POI and Anonymize-FSP. Figure 4(a) shows the constructing time of the anonymous trajectories by varying different anonymous values. It can be seen that the constructing time using Anonymize-POI is less than that using Anonymize-FSP, as Anonymize-FSP needs a lot of time to mine the FSPs and MFSs when constructing the anonymous trajectories. In addition, the time of these two anonymization algorithms changes relatively stable when increasing the k values. The constructing time with respect to different similarity thresholds is depicted in Fig. 4(b). It shows similar results on these two different parameters.

In summary, by comparing the TC accuracy and constructing time, the Anonymize-FSP shows the superiority of Anonymize-POI on constructing the anonymous trajectories for TC, but it takes more time to construct anonymous trajectories.



(a) Different k

(b) Different γ

Fig. 4. Constructing time vs. k value

6 Conclusion

In this paper, we study the problem of *private trajectory data publication for trajectory classification*, which not only protects the trajectory privacy, but also guarantees high classification accuracy. We propose a private trajectory data publishing framework for trajectory classification, which first constructs the anonymous trajectory set, and then classifies the anonymous trajectory. In order to build a good anonymous trajectory set, we propose two algorithms for constructing anonymous trajectory set, namely *Anonymize-POI* and *Anonymize-FSP*. At last, comprehensive performance evaluation is conducted to validate the proposed ideas and demonstrate the accuracy and effectiveness of the proposed algorithms. The experimental results show that the proposed algorithms not only preserve the trajectory privacy, but also guarantee a high classification accuracy.

This work may lead towards several new directions for future work, e.g., using encryption techniques and differential privacy.

Acknowledgment. This work was supported by the Postdoctoral fund (2018M-643307), Key R&D Program of Guangdong Province (2018B010107005, 2019B010-120001), and the National Natural Science Foundation of China (Nos. 61532021, 61572122, U1736104).

References

1. Abul, O., Bonchi, F., Nanni, M.: Never walk alone: uncertainty for anonymity in moving objects databases. In: ICDE, pp. 376–385 (2008)
2. Agrawal, R., Srikant, R.: Mining sequential patterns. In: ICDE, pp. 1–12 (1995)
3. Beresford, A.R., Stajano, F.: Location privacy in pervasive computing. IEEE Pervasive Comput. **2**(1), 46–55 (2004)
4. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 1–27 (2011)
5. Chow, C.Y., Mokbel, M.F.: Privacy of spatial trajectories. In: Zheng, Y., Zhou, X. (eds.) Computing with Spatial Trajectories. Springer, New York (2011). https://doi.org/10.1007/978-1-4614-1629-6_4
6. Fournier-Viger, P., Wu, C.-W., Gomariz, A., Tseng, V.S.: VMSP: efficient vertical mining of maximal sequential patterns. In: Sokolova, M., van Beek, P. (eds.) AI 2014. LNCS (LNAI), vol. 8436, pp. 83–94. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-06483-3_8
7. He, X., Cormode, G., Machanavajjhala, A., Procopiuc, C.M., Srivastava, D.: DPT: differentially private trajectory synthesis using hierarchical reference systems. In: VLDB, pp. 1154–1165 (2015)
8. Lee, J.G., Han, J., Li, X.: TraClass: trajectory classification using hierarchical region-based and trajectory-based clustering. VLDB **1**, 1081–1094 (2008)
9. Lee, J.G., Han, J., Li, X., Cheng, H.: Mining discriminative patterns for classifying trajectories on road networks. TKDE **23**(5), 713–726 (2011)
10. Nergiz, M.E., Atzori, M., Saygin, Y.: Towards trajectory anonymization: a generalization-based approach. Trans. Data Privacy **2**(1), 52–61 (2009)
11. Palanisamy, B., Liu, L.: MobiMix: protecting location privacy with mix-zones over road networks. In: ICDE, pp. 494–505 (2011)
12. Pan, X., Meng, X., Xu, J.: Distortion-based anonymity for continuous queries in location-based mobile services. In: SIGSPTAIL, pp. 256–265 (2009)
13. Terrovitis, M., Mamoulis, N.: Privacy preservation in the publication of trajectories. In: MDM, pp. 65–72 (2008)
14. Wu, J., Ni, W., Zhang, S.: Generalization based privacy-preserving provenance publishing. In: Meng, X., Li, R., Wang, K., Niu, B., Wang, X., Zhao, G. (eds.) WISA 2018. LNCS, vol. 11242, pp. 287–299. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02934-0_27
15. Xu, T., Cai, Y.: Exploring historical location data for anonymity preservation in location-based services. In: INFOCOM, pp. 547–555 (2008)
16. Xin, Y., Xie, Z.Q., Yang, J.: The privacy preserving method for dynamic trajectory releasing based on adaptive clustering. Inf. Sci. **378**, 131–143 (2017)
17. Yan, X.: CloSpan: mining closed sequential patterns in large datasets. In: SAIDM, pp. 166–177 (2003)
18. Yao, L., Wang, X., Wang, X., Hu, H.: Publishing sensitive trajectory data under enhanced l-diversity model. In: MDM (2019, to appear)
19. You, T.H., Peng, W.C., Lee, W.C.: Protecting moving trajectories with dummies. In: MDM, pp. 278–282 (2008)

20. He, Z., Gu, F., Zhao, C., Liu, X., Wu, J., Wang, J.: Conditional discriminative pattern mining. Inf. Sci. **375**, 1–15 (2017)
21. Zheng, Y.: Computing with Spatial Trajectories. Springer, New York (2011). https://doi.org/10.1007/978-1-4614-1629-6
22. Zheng, Y.: Trajectory data mining: an overview. ACM Trans. Intell. Syst. Technol. **6**, 29 (2015)
23. Zheng, Y., Li, Q., Chen, Y., Xie, X., Ma, W.Y.: Understanding mobility based on GPS data. In: MOBIQUITOUS, pp. 312–321 (2008)
24. Zheng, Y., Liu, L., Wang, L., Xie, X.: Learning transportation mode from raw GPS data for geographic applications on the web. In: WWW (2008)
25. Zhu, H., Yang, X., Wang, B., Lee, W.C.: Range-based obstructed nearest neighbor queries. In: SIGMOD, pp. 2053–2068 (2016)