2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discover

A Coarse-to- Fine Approach for Motion Pattern Discovery

Bolun Cai, Zhifeng Luo, Kerui Li
School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

***Abstract*—In this paper, we propose a coarse-to-fine approach to discovery motion patterns. There are two phases in the proposed approach. In the first phase, the proposed median-based GMM achieves coarse clustering. Moreover, the number of clusters can be heuristically found by the proposed algorithm. In the second phase, to refine coarse clustering in the first phase, a Fisher optimal division method is proposed to examine the boundary data points and to detect the change point between motion patterns. The experimental results show that the proposed approach outperforms the existing algorithms.**

***Keywords*-trajectory data clustering; GMM; motion pattern discovery;**

I. Introduction

With development of mobile computing techniques, it is easy to obtain the trajectory data of users by GPS which is integrated in smart cell phones. It is expected to discover significant motion patterns from the trajectory data, in order to learn the user behavior patterns. The motion pattern discovery may facilitate many promising applications, such as user movement prediction, logistics monitoring and urban public transport scheduling.

Many traditional classification and clustering methods are used to analyze motion patterns [1] [2] [3]. In [4] , K-means is used to cluster the preprocessed data for the discovery of motion patterns. The disadvantage of K-means is that the accuracy of clustering is sensitive to the selection on the initial cluster centers. The method proposed in [3] combines fuzzy C-means (FCM), Subtractive and GMM for clustering. But the result of FCM may be locally optimal, and the clustering results depend on the initial choice of weights. In [5], it is shown that GMM is a suitable method for motion patterns discovery. Joseph etc. model motion patterns as a mixture of Gaussian processes (GP) with a Dirichlet process (DP) prior over mixture weights. The GP provides a flexible representation for each individual motion pattern, and DP assigns the observed trajectories to particular motion patterns. But this method does not consider the case of the hybrid motion patterns in a trajectory. In this case, this method cannot detect the boundary between two different motion patterns. Suzuki etc. [4] use a hidden Markov model (HMM) to model human trajectory . In [5], HMM is used to improve expectation-maximization (EM) for learning a Bayesian model of motion patterns. The method based on HMM ignores the history location in the trajectory and only consider the current location.

As we know, the GPS trajectory data may be corrupted by the noise at the receiver. Also, the accuracy of GPS signal may be affected by multipath effect in the urban environment. In some cases, the motion pattern is too complicated to discover, such as waiting for traffic lights and driving in the congested road. It is difficult that the different motion patterns on a trajectory are correctly distinguished. In order to reduce the noise, Kalman filtering, Vondrak filtering and particle filter [6] [7] [8] are used to preprocess the GPS data. In general, there are two methods to cluster GPS trajectory data for discovering motion patterns. One is to segment the trajectory, the other is to cluster the trajectory data in temporal order. Trajectory segmentation can’t deal with the case that there are more than one repeated motion pattern in the single trajectory. For the clustering method, it is difficult to cluster correctly the data point on the boundaries between identified clusters .

In this paper, we propose a coarse-to-fine approach for discovering motion patterns in the trajectory. The proposed approach consist of two phases: phase 1. coarse clustering and phase 2. refined separation. In the phase 1, a median-based Gaussian mixture model (MGMM) is proposed to cluster the raw GPS trajectory data for the discovery of different motion patterns. Moveover, the sequential property of GPS data is utilized in MGMM to suppress the noise of GPS data. In addition, to separate the boundary of different motion patterns, we use the Fisher optimal division method (FODM) [9] for refinement in the second phase. Our contributions are as follows: (1) the performance of the traditional GMM on the trajectory data clustering is improved with a median-based method. (2) the proposed approach is feasible to detect the change points located on the boundary of two consecutive motion patterns. (3) Unlike the other clustering methods which need the input of the number of clusters, in the proposed approach, the number of clusters (motion patterns) can be heuristically discovered by the algorithm. The proposed approach is suitable for the discovery of motion patterns because the number of motion patterns in a trajectory is unknown beforehand. Moreover, the proposed approach can be used in the case that there are repeated motion patterns in a trajectory.

This paper is organized as follows: In Section II, we present the problem model. Section III presents the proposed

This work is supported by University Innovation Research and Training Program of Guangdong Province(1056111033).

978-0-7695-4810-4/12 $26.00 © 2012 IEEE
DOI 10.1109/CyberC.2012.95

coarse-to- fine approach for clustering the trajectory data. Section IV provides the evaluation of the proposed approach and experiment results. Section V is the conclusion of this paper.

II. Problem Model

We define the data format of a trajectory T in the database as $< (x_1, y_1, v_1), (x_2, y_2, v_2), ..., (x_L, y_L, v_L) >$, where (x_i, y_i, v_i) is a data item which records the GPS longitude x_i and latitude y_i at the i-th timestamp, v_i is the instantaneous velocity at the GPS data point (x_i, y_i). L is the length of T.

Definition 1: Let

$$T_{il} =< (x_i, y_i, v_i), (x_{i+1}, y_{i+1}, v_{i+1}), ..., (x_l, y_l, v_l) >$$

represent a subsequence of T. We call T_{il} as a **motion pattern** if all data points in T_{il} are assigned to a cluster which infer to an episode of user movement.

Definition 2: Assume that a trajectory contains two consecutive different motion patterns

$T_{ij} =< (x_i, y_i, v_i), (x_{i+1}, y_{i+1}, v_{i+1}), ..., (x_{c-1}, y_{c-1}$
$, v_{c-1}), (x_c, y_c, v_c), (x_{c+1}, y_{c+1}, v_{c+1}), ..., (x_j, y_j, v_j) >$,

where $< (x_i, y_i, v_i), (x_{i+1}, y_{i+1}, v_{i+1})..., (x_{c-1}, y_{c-1}, v_{c-1}) >$ and $< (x_{c+1}, y_{c+1}, v_{c+1}), ..., (x_j, y_j, v_j) >$ are two different motion patterns, respectively. (x_c, y_c, v_c) is defined as a **change point**.

In this paper, we intend to cluster a trajectory data for discovering the consecutive different motion patterns, such as $< (x_i, y_i, v_i), (x_{i+1}, y_{i+1}, v_{i+1}), ..., (x_{c-1}, y_{c-1}, v_{c-1}) >$ and $< (x_{c+1}, y_{c+1}, v_{c+1}), ..., (x_j, y_j, v_j) >$. In addition, it is expected to detect the change point (x_c, y_c, v_c) connecting different motion patterns.

III. Coarse-to-fine approach for discovering motion pattern

In this section, we will present the proposed coarse-to-fine approach in detail. We select the velocity distribution as the feature of trajectory data. In our model, there are two phases for clustering trajectory data: the first phase is the coarse clustering and the second phase is the refined separation. In the first phase, we propose a median-based GMM (MGMM) to implement the coarse clustering. The majority of data points which infer the same motion pattern may be assigned to the same cluster. It is allowed in the first phase that data points located in the boundary between two clusters cannot be correctly distinguished. These boundary data points with Fisher optimal division method (FODM) are processed in the second phase. The goal of the second phase is to detect the change point.

A. Phase 1. Coarse clustering

Based on our observations, a trajectory which contains only one motion pattern can pass Kolmogorov-Smirnov (K-S) test [12]. It is reasonable that the distribution of the velocity corresponding to the same motion pattern is assumed to follow a Gaussian distribution. The proposed Median-based GMM (MGMM) is presented as follows: Assume that the velocity of trajectory data point is modeled by a K Gaussian component mixing model with mixing proportions $\{\pi_k\}$. The probability density function of the velocity v_i can be written as:

$$P(v_i) = \sum_{k=1}^{K} p(k)p(v_i|k) = \sum_{k=1}^{K} \pi_k N(v_i|\mu_k, \Sigma_k), \qquad (1)$$

where a priori $p(k) = \pi_k$ and the conditional distribution $p(v_i|k) = N(v_i|\mu_k, \Sigma_k)$. $N(v_i|\mu_k, \Sigma_k)$ denotes a Gaussian component with the mean μ_k and the variance Σ_k.

The log-likelihood function of Equ. (1) is given by:

$$\sum_{i=1}^{L} log\{\sum_{k=1}^{K} \pi_k N(v_i|\mu_k, \Sigma_k)\} \qquad (2)$$

The maximum likelihood estimation is used to find the model parameters which maximize Equ. (2). To this aim, the expectation-maximization (EM) algorithm [10] is applied. In EM, given the i-th observation v_i, we calculate the *a posteriori* probability $\gamma(k|v_i)$ as follow:

$$\gamma(k|v_i) = \frac{\pi_k N(v_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(v_i|\mu_j, \Sigma_j)}. \qquad (3)$$

After EM reaches convergence, the normalized median of a posteriori probability for the k-th Gaussian component is given by

$$\Gamma(k|v_i) = \frac{\gamma'(k|v_i)}{\sum_{k=1}^{K} \gamma'(k|v_i)}, \qquad (4)$$

where $\gamma'(k|v_i)$ is the median of $\{\gamma(k - m|v_i), ..., \gamma(k|v_i), ..., \gamma(k + m|v_i)\}$, m is the radius of the observed window. Finally, the cluster label C_i of v_i is obtained by $C_i = \max_k \Gamma(k|v_i)$.

The value of mixing proportion π_k also can indicates which Gaussian component is a significant one. The mixing proportion π_k is calculated by: $\pi_k = \frac{\sum_{i=1}^{L} \Gamma(k|v_i)}{L}$.

Let ϵ denote a threshold on the value of mixing proportion. If $\pi_k > \epsilon$, the k-th Gaussian component represent a significant motion pattern; otherwise, it is suggested that the motion pattern represented by the k-th Gaussian component does not exist in the trajectory. In this way, the number of cluster K can be heuristically determined. ϵ can be a built-in parameter in the algorithm. In Section IV, we provide the insight to the optimal value of ϵ by experiments.

B. Phase 2. refined separation

Fisher optimal division method (FODM) is used to refine the result of the first phase. It is expected that FODM is able to detect change points between two different motion patterns. After the first phase, assume that the i'-th data point is a boundary point between two clusters identified by the proposed MGMM. Let $F =< v_{i'-n}, ...v_{i'}, ..., v_{i'+n} >$ denote a set of data points on the boundary between two

Table I
ACCURACY ON THE LDPA DATA SET

	K-means	FCM	GMM	The proposed approach
Accuracy	78.04%	75.64%	79.75%	88.15%

Table II
ACCURACY ON THE REAL GPS DATA

	K-means	FCM	GMM	The proposed approach
Accuracy	79.65%	79.73%	83.38%	93.74%

clusters. the cohesion of F with respect to the cluster corresponding to the k-th motion pattern is given by:

$$D_k(i'-n, i'+n) = \sum_{t=i'-n}^{i'+n} (v_t - \overline{v}), \tag{5}$$

where $\overline{v} = \frac{\sum_{t=i'-n}^{i'+n} v_t}{2n+1}$. n is the range of boundary points.

The index of change point c can be obtained by

$$c = \arg\min_{i'} \sum_{k=1}^{K} D_k(i'-n, i'+n) \tag{6}$$

Note that FODM only examines the boundary data points of which the amount is small for refinement.

In the proposed approach, the radius m of the observed window and the range n of boundary points are parameters in the proposed scheme. The optimal values of m and n are given by experiments in Section IV.

IV. EXPERIMENTS AND RESULTS

To validate our model, we implement an Android application so that the real GPS data can be collected when users carry their smart phone daily. The real GPS data is used to calculate the instantaneous velocity. We record a user's motion pattern including walking, biking and driving for a week.

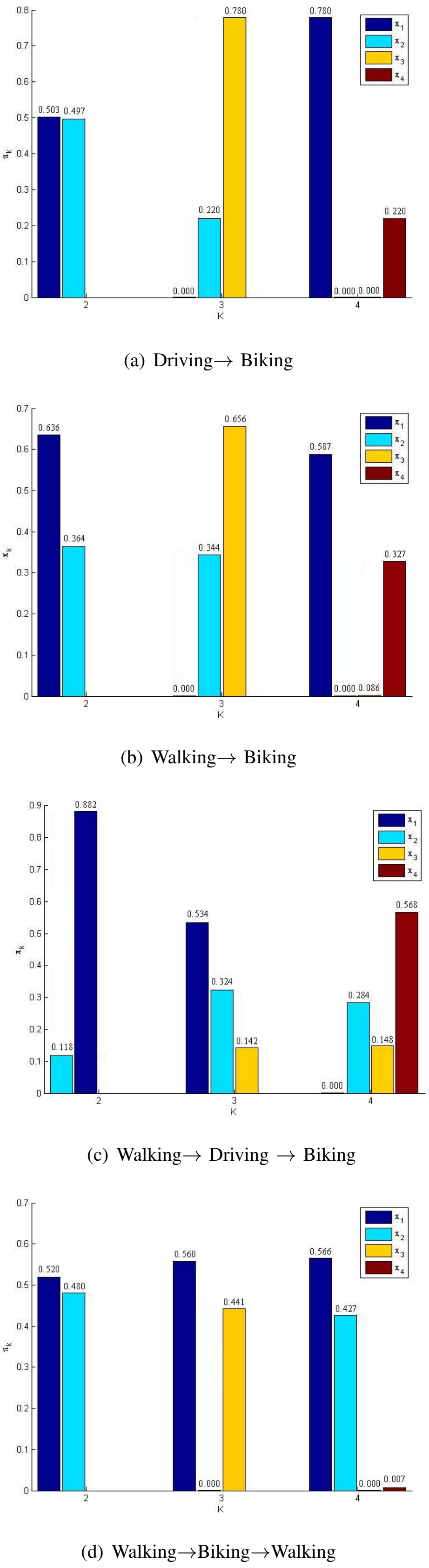
A. The evaluation on accuracy

We design an experiment to evaluate the performance of the proposed MGMM. The proposed algorithm is compare with other clustering algorithms, such as K-means, Fuzzy C-means Algorithm (FCM) and GMM. Assume that the parameter K of all algorithms in the evaluation is the actual number of clusters in the data set. For evaluation on the accuracy, we use the localization data for person activity (LDPA) data set provided in [14] and a real GPS data set collected in our experiment. The accuracy [13] is used to measure the performance.

Table I and II show the accuracy of the proposed algorithm on LDPA data set and the 10 real GPS trajectories which contain different motion patterns, respectively. By comparisons, the proposed approach can outperform the other clustering methods. The accuracy of the proposed approach can reach 93.47%.

B. The heuristic discovery of motion patterns

Assume that $\epsilon = 0.1$ in the algorithm. If $\pi_k > \epsilon$, the cluster associated to π_k is a significant motion pattern. Otherwise, the cluster represents outliers. As shown in Fig. 1, the number of π_k which is more than ϵ is the constant



(a) Driving→ Biking

(b) Walking→ Biking

(c) Walking→ Driving → Biking

(d) Walking→Biking→Walking

Figure 1. π_k for different motion patterns

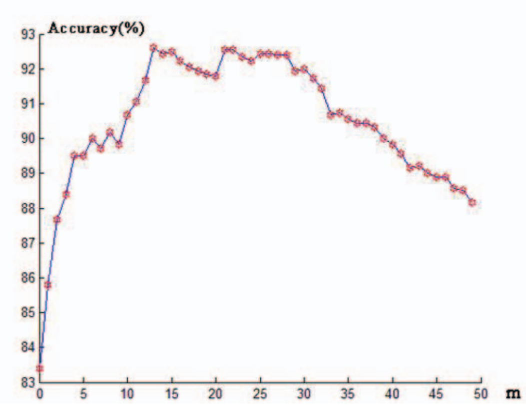


Figure 2. The effect of m on the accuracy

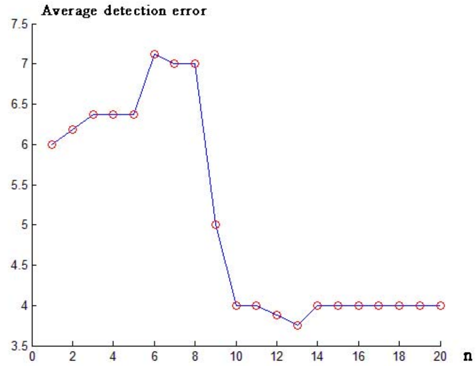


Figure 3. The effect of m on the accuracy

when K increases. Fig. 1 (a) and (b) show the case that there are two different motion patterns in the trajectory, respectively. The number of cluster detected by setting the threshold $\epsilon = 0.1$ is two when K increases from 2 to 4. Similarly, three motion patterns can be detected in Fig. 1(c). As shown in Fig. 1(d), the "Walking" motion pattern repeatedly occurs after the "Biking" motion pattern, two different motion patterns is detect by $\pi_k > \epsilon$. Our approach is still effective when there is a repeated motion pattern over time.

C. The effect of parameter selection

As shown in Fig. 2, when the the radius m of the observed window equal to 14, the accuracy of the proposed approach reach the maximum 92.6%. In the refined separation phase, the range n of FODM can be found by measurement of the average detection error which is define by the average distance between the detected positions of change points and the true position of change points. It is shown in Fig. 3 that the average detection error change with increasing n. The average detection error tends to be constant when $n \geq 10$. And there is no a significant gain when keep increasing n after $n = 10$.

V. Conclusion

In this paper, we propose an effective approach for motion pattern discovery. A median-based GMM is proposed to improve the traditional GMM. The proposed algorithm implement coarse clustering of trajectory data. And then, to refine the coarse clustering, we propose to use FODM on the boundary points of clusters. In this way, the change points between two motion patterns also can be detected. In comparison to existing approaches, the proposed approach doest not need to predefine the number of cluster which is unknown for discovering motion pattern from trajectory data. The experimental results show that the proposed approach has the advantage over the existing approaches on the accuracy.

References

[1] D. Bauer, M. Ray, N. Br, ndle, and H. Schrom-Feiertag, "On extracting commuter information from GPS motion data," in *Proc. the 5th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services.* , pp. 1-6, Dublin, Ireland, 2008.

[2] L. Haitao, K. Lingfu, and W. Peiliang, "Detecting abnormal state of elderly for service robot with H-FCM," in *ICAL '09.* , pp. 1867-1870, 2009.

[3] C. Carneiro, A. Alp, J. Macedo, and S. Spaccapietra, "Advanced Data Mining Method for Discovering Regions and Trajectories of Moving Objects: "Ciconia Ciconia" Scenario," *Lecture Notes in Geoinformation and Cartography* ,pp. 201-224, 2007.

[4] N. Suzuki, K. Hirasawa, K. Tanaka, Y. Kobayashi, Y. Sato, and Y. Fujino, "Learning motion patterns and anomaly detection by Human trajectory analysis," in *IEEE International Conference on Systems, Man and Cybernetics, 2007.*, pp. 498-503, 2007.

[5] J. Joseph, F. Doshi-Velez, A. Huang, and N. Roy, "A Bayesian nonparametric approach to modeling motion patterns," *Artificial Intelligence* , Vol. 31, No.4, pp. 383-400, 2011.

[6] A. H. Mohamed and K. P. Schwarz, "Adaptive Kalman Filtering for INS/GPS," *Journal of Geodesy* , Vol. 73, pp. 193 -203, 1999.

[7] ZHENG, D. W, ZHONG, P, DING, X. L, CHEN, and W, "Filtering GPS time-series using a vondrak filter and cross-validation," *Journal of Geodesy* , Vol. 79, No.5-6, 2005.

[8] L. Liao, D. J. Patterson, D. Fox, and H. Kautz, "Learning and inferring transportation routines," *Artificial Intelligence* , Vol. 171, pp. 311-331, 2007.

[9] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, "Fisher discriminant analysis with kernels," in *Proceedings of the 1999 IEEE Signal Processing Society Workshop* , pp. 41-48, 1999.

[10] C. M. Bishop, "Pattern Recognition and Machine Learning," Springer, USA, 2006.

[11] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing* , Vol. 10, No.6, pp. 19-41, 2000.

[12] H. W. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *J. Am. Stat. Assoc.* , Vol. 62, No. 318, pp. 399-402, 2007.

[13] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *Artificial Intelligence* , Vol. 17, No.12, pp. 1624- 1637, 2005.

[14] http://archive.ics.uci.edu/ml/datasets.html