2011 IEEE International Conferences on Internet of Things, and Cyber, Physical and Social Computing

Anomaly Detection on Collective Moving Patterns

A Hidden Markov Model based Solution

Su Yang
College of Computer Science and Technology
Fudan University
Shanghai 200433, China
suyang@fudan.edu.cn

Weihua Liu
College of Computer Science and Technology
Fudan University
Shanghai 200433, China
09210240025@fudan.edu.cn

***Abstract*—Trajectories of people provide rich information about the collective behaviors, where the term collective behavior means the behavior of a large number of people as a whole. Abnormal people trajectories that are rarely observed may correspond with some unusual events, for instance, natural disasters, terrorism attacks, or traffic accidents. To detect such abnormal people trajectories, namely outliers, this paper presents a solution based on Hidden Markov Model (HMM). The motivation to use HMM for outlier detection on people trajectories is that time-varying people distribution can be modeled by using HMM and HMM can provide the probability that a sequence appears. Here, the time-varying people distributions with low probabilities to appear are regarded as outliers. Experiments with an artificial data set simulating collective behaviors and a real-world traffic data set validate the proposed solution.**

Keywords-Collective Behavior; Outlier Detection

I. INTRODUCTION

With the technological advance, tracking persons or vehicles via mobile phone, GPS, camera, and other sensors becomes practical. People's real-world digital traces were captured by the sensors distributed around and recorded by some agents like telecommunication service providers, traffic management departments, and social security departments. People's digital traces provide rich information about the patterns of people's life but how to understand the underlying patterns of people' daily life and make use of such knowledge is an open problem yet. Recently, social computing has received much attention. A new and critical topic in terms of social computing is to uncover the social patterns underlying as well as the laws governing people's daily life. In the Reality Mining project, Eagle and Pentland use mobile phone locations and proximity to model individual users' behaviors and infer social relations among users [1,2]. After analyzing a large number of mobile phone records, Gonzalez and her colleagues found that the individual travel patterns collapse into a single spatial probability distribution and are reproducible patterns [3]. Such a discovery could be utilized in emergency response, urban planning, and traffic forecasting. In [4], frequent mobility patterns are discovered from mobility paths for individual users, aiming at applications like location-based services, air pollution exposure estimation, and traffic planning. However, collective behaviors of a large group of persons viewed as a whole is not the concern of the previously mentioned researches. In the Mobile Landscapes project, Ratti and his colleagues make use of the mobile phone location data from an aggregated point of view to represent the intensity of urban activities and their evolution through space and time for urban studies and planning [5]. Candia et al found that abnormal events affect collective behaviors of people and drive their behaviors distinctly different from normal ones, which can be observed in the call patterns of people from a large scale of view [6]. However, their research does not answer how to detect automatically such abnormal patterns of large-scale collective behaviors. Recently, some researchers propose to detect abnormal collective behaviors from a data mining point of view [7][8]. The basic idea of such researches is to view the distributions of moving objects (persons or vehicles) across multiple areas as random image sequence and make use of geometrical and manifold analysis on such random image sequence to obtain features associated with every time window. Then, classical outlier detection is performed on such feature space. In this paper, we propose a new solution based on Hidden Markov Model (HMM) [9] to compute the probability for a vector sequence representing time-varying object distributions. If the obtained probability for a given pattern is low, it means that the outlier degree of such a pattern is high. In the literature of data mining, HMM is also employed for network intrusion detection [10,11]. However, network intrusions are in different application domain and with different physical characteristics other than those of the collective behaviors studied in this research. By using HMM, we achieve the outlier detection results for an artificial data set and a real-world traffic data set. The experimental results confirm the effectiveness of the proposed scheme.

The rest of this paper is organized as follows. In section 2, we describe the solution in detail. In section 3, we present the experimental results. In section 4, we make conclusions.

II. OUTLIER DETECTION ON COLLECTIVE BEHAVIORS USING HIDDEN MARKOV MODEL

Our goal is to detect abnormal time-varying people distribution. This is a new task other than the classical outlier detection problems in the context of data mining since people can be viewed as moving objects and the task is to

978-0-7695-4580-6/11 $26.00 © 2011 IEEE
DOI 10.1109/iThings/CPSCom.2011.25

make a decision on whether the trajectories of all the people of interest during a certain period is different from most cases. In the following, we outline the entire framework. First, we represent the data as follows. We divide the region of interest into M zones and count the number of people in every zone at a particular time to form an M-dimensional vector. At every sampling time, we can obtain one M-dimensional vector. As such, we can obtain an M-dimensional vector sequence representing the time-varying people distribution. Then, we divide such M-dimensional vector sequence into N pieces with partial overlap, which is referred to as subsequence. Second, we compute the probability for each subsequence. If the probability associated with a subsequence is low, it means that such a subsequence is rarely observed and the corresponding collective behavior rarely happens. So, we let the probability that a subsequence appears be the indicator of the outlier degree of such subsequence. A lower probability corresponds with higher outlier degree. Here, we apply HMM in computing the probability that each subsequence appears. Prior to online outlier detection, we use historical data to train HMM. A HMM is composed of five essential elements: Observations, hidden states, probabilities of initial states, state transition matrix, and confusion matrix. Each of the elements will be described in detail in the following.

A. Observations

In this study, we use an M-dimensional vector $O_i=[O_{i1},O_{i2},\ldots,O_{iN}]^T$ to represent the number of people located in M areas of interest at the ith sampling time. Hence, the raw data is transformed into an M-dimensional vector sequence $[O_1,O_2,\ldots,O_N]$, where N is the length of the vector sequence.

B. Hidden States

Defining hidden states is a key problem in establishing the model for this investigation. We apply the K-means clustering technique to group the N vectors $\{O_1,O_2,\ldots,O_N\}$, namely the observations, into K clusters. Hereafter, each cluster represents a hidden state. We let $C=\{C_1,C_2,\ldots,C_K\}$ represent the K hidden states. In a HMM, K is a very important parameter, which can be obtained experimentally by optimizing the performance in the training stage.

C. State Transition Probability Matrix

Once we obtain the K clusters, we can assign a class label to each vector in the sequence $[O_1,O_2,\ldots,O_N]$. We let $S(O_i)$ represent the class label of O_i, $S(O_i)\in C$. If $O_i\in C_j$, then, $S(O_i)=C_j$. By means of such, we obtain a class label based representation of the observation, that is, $[S(O_1),S(O_2),\ldots, S(O_N)]$. Then, we tie every pair of adjacent class labels in $[S(O_1),S(O_2),\ldots,S(O_N)]$. This leads to a state transition sequence $\{[S(O_1),S(O_2)], [S(O_2),S(O_3)], \ldots, [S(O_{N-1}), S(O_N)]\}$. Then, the transition probability from state C_i to state C_j can be estimated as follows:

$$a_{ij}=\frac{\#\{[S(O_t),S(O_{t+1})]\,|\,S(O_t)\in C_i \wedge S(O_{t+1})\in C_j\}}{\#\{S(O_t)\,|\,S(O_t)\in C_i\}}$$

where $\#\{.\}$ represent the number of elements contained in the subsequent given set.

D. Output Probability Matrix based on Gaussian Mixture Models

For each cluster obtained following the K-means clustering, a Gaussian Mixture Model (GMM) is used to model the distribution of the M-dimensional vectors contained in this cluster. At the training phase, we estimate the parameters of Gaussian Mixtures using the expectation-maximization (EM) algorithm. Note that since the observation vectors for each class are M-dimensional, all of the individual multivariate Gaussian distributions are correspondingly M-dimensional. Let the set of observation vectors for class j be denoted as O_j, the probability density function (PDF) of this class can be modeled to arbitrary accuracy using a mixture of Gaussians:

$$P(O_j\,|\,\theta_j)=\sum_{i=1}^{L} w_i N(O_j\,|\,\mu_i,\Sigma_i)$$

The mixtures are completely determined by the parameter set $\theta_j=\{(w_i,\mu_i,\Sigma_i)|i=1, 2,\ldots,L\}$. To initialize the EM algorithm, the GMM component means $\{\mu_i\}$ and the covariance matrices $\{\Sigma_i\}$ are derived from clustering of every Gaussian mixture and then computing the mean and size of every cluster. At the working stage, the probability that an input belongs to each cluster can be estimated using GMM. That is,

$$P(O\,|\,\theta_j)=\sum_{i=1}^{L} w_i N(O\,|\,\mu_i,\Sigma_i)$$

In the above equation, O represents a new input and the parameters are known following EM iterations.

E. Initial Probability Vector

From $[S(O_1),S(O_2),\ldots,S(O_N)]$, the probability that the sequence starts from each state can be estimated via the following equation:

$$\pi_i=\frac{\#\{S(O_j)\,|\,S(O_j)\in C_i\}}{\#\{S(O_j)\}}$$

F. Outlier Detection

We wish to detect abnormal collective activities from a dynamic view. As the time-varying object distribution can be represented in the form of a vector sequence. For each input subsequence $[O_i,O_{i+1},\ldots,O_{i+L}]$, HMM can compute the probability that such subsequence occurs. Here, we apply the forward-backward algorithm for HMM to compute the probability. Not that $[O_i,O_{i+1},\ldots,O_{i+L}]$ represent time-varying object distributions in M areas of interest within a certain duration. If the probability that such a subsequence occurs is low, then, this will be regarded as a rarely seen pattern, namely, outlier. The probability associated with each subsequence can be regarded as the indicator of the outlier degree of such subsequence. The decision for outlier detection on a given vector sequence can be made as follows:

If $Prob\{[O_i,O_{i+1},\ldots,O_{i+L}]\}<P_T$, then, it is an outlier

where P_T is a predefined threshold.

III. EXPERIMENTS

A. A Case Study with Artificial Data

We use the artificial data generation method applied in [8] to evaluate the proposed method. In this data set, people's daily motion trajectories are simulated using NetLogo. The simulated persons include employees and vagrants. Some abnormal events are set at regular intervals to affect the behaviors of employees. Our goal is to find out the abnormal collective trajectories arising from these events. In this data set, all the employees have four statuses: Staying at home, moving from home to company, staying at working place, and moving from company to home. It is supposed that the employees stay at the workplaces during the work time. Once there is an abnormal event, part of the employees move to and are gathered around the place where the event happens. The population of such participants that are subject to abnormal events is controlled by a parameter referred to as participation rate. Abnormal events have an attribute, namely lifetime, during which the abnormal event of interest lasts. After the lifetime, the corresponding event terminates and the involved employees back to their normal behaviors. For vagrants, they have two statuses: Staying at one place and moving arbitrarily. The population of the vagrants is controlled by a parameter referred to as unemployment rate. The vagrants can be regarded as noise in the experiments. They move randomly regardless of whether or not an abnormal event happens. Every day is divided into 120 time pieces. From the 110th time piece to the 20th time piece of the next day, all the employees stay at home and all the vagrants stay where they are. During the rest of the time, the vagrants move arbitrarily. The employees go to company from the 20th time piece to the 40th time piece, and go home from the 90th time piece to the 110th piece. During the rest of every day, they stay at their companies. Simulations of 500 days are generated, including 60000 time pieces in total, and an abnormal event is set every 2000 time pieces. The lifetime is a random integer between 40 and120. In this experiment, 8 communities and 4 companies are generated. Each employee has his community and company, which is not changed during the whole experiment. The number of people is 400. The parameters to control the simulation are: The unemployment rates are 0% and 20%, respectively, while the participation is 20%. The whole space is divided into M areas with the same length and width. In every time piece, we count the population in each area to figure out the distribution of people. When clustering, k is an important parameter that determines the number of clusters. We test our method with different k to examine how the parameter affects the result. We choose 40000 time pieces to train the Hidden Markov Model, and the other 20000 times pieces are used for testing. Fig. 1 shows the result when the unemployment rate is 0 and the participation rate is 20%. The x-axis represents the time piece and the y-axis represents the logarithm of the probability of the corresponding time piece. We can see that the probability changes regularly with time piece, which fits well into the people's daily motion mode. However, for every 2000 time pieces, there is a great change reflected in the probability, where the probability becomes small. It corresponds with the time at which an abnormal event takes place as we set. Intuitively, this confirms that the HMM based solution is effective for detecting outlier patterns on groups of moving objects. The recall rate and the precision rate are shown in Fig. 2. The recall rate is very high (>94%) and the threshold value does not have remarkable effect on the recall rate. The precision is also higher than 90% as the log of the threshold value varies from -300 to -400. These verify the effectiveness of the proposed solution. Fig. 3 shows the logarithm of the probability of the corresponding time piece when both unemployment rate and the participation rate is 20%. The outlier patterns can also be observed in this figure. Fig. 4 illustrates the corresponding recall rate and precision. The recall rate is relatively high, which is higher than 90% when the log of the threshold for outlier detection is greater than -335. The precision remains relatively high, higher than 90% when the log of the threshold value is less than -325. Comparing Fig. 2 with Fig. 4, it can be found that the outlier detection performance degrades when the unemployment rate is raised from 0 to 20%. The reason is that higher unemployment rate corresponds with more noisy objects. Thus, it can be concluded that noises degrade the outlier detection performance.

B. A Case Study with Real-World Traffic Data

Real-world moving objects data are also used to validate the proposed method. The data are continuously collected by the Traffic Management Center of Minnesota Department of Transportation, with a 30-second interval from over 4,000 loop detectors located around the Twin Cities Metro freeways for seven days per week. The data have two kinds of files: volume file and occupancy file. The volume files are used to record the traffic volume, and we use this kind of files only. The data that we use cover the period from 1st January to 22nd September in 2010. Due to the computational load, we only make use of the data from 2000 sensors with 10 minutes as the interval to conduct the experiment and apply principal component analysis (PCA) to reduce the data dimension to 4. The data of the first 117days, from 1st Jan to 14th May, are used for training. The data of the other 107 days, from 16th May to 22nd September, are used for testing. The training data is grouped into 5 classes via clustering. For each class, we use GMM to model the data. We use the K-means clustering method to estimate the mean and the covariance matrix for initializing GMM, where we let each GMM be composed of 3 Gaussian components. When the model is constructed, we compute the probability of every string/subsequence consisting of 7 continuous time pieces, that is, 70 minutes per sting. The logarithm of the probability of every string is illustrated in Fig. 5. If more than 6 subsequences are detected as outliers in a given day, then, this day is regarded as abnormal. Here, the threshold used for outlier detection is -170. Table 1 lists detected abnormal days and the corresponding events. It can be seen that most days detected as abnormal correspond with some events. As we are not local residents, we cannot confirm

whether all the days detected as outliers are really abnormal. Even through, the present results obtained by us are still encouraging and indicating a new effective way for outlier detection on large-scale collective behaviors.

IV. CONCLUSIONS

We propose a HMM based solution to detect abnormal collective behaviors of a large number of moving objects. It can be used in many applications such as emergency handling, traffic management, and anti-terrorism. The experiments with artificial and real-world data show that the proposed scheme is effective. The experimental results achieved from the real-world traffic data are encouraging since most of the days detected as abnormal coincide with some known events like extreme weather, festival, art show, and some other special days.

TABLE I. OUTLIER DETECTION RESULTS AGAINST THE EVENTS

28 May	Outdoor music
31 May	Memorial day
11 June	
23 June	Twin cities residents brave hot sultry weather
2 July	
9 July	Bastille day event. Road closure
16 July	Aquatennial. Road closings
23 July	
24 July	UFO spotted in twin cities
30 July	Red hot art festival
6 August	
13 August	Snow
20 August	Tornado
27 August	Record opening day attendance at minnesota state fair
3 September	
21 September	Strong storm

ACKNOWLEDGMENT

This work is supported by 973 Program (grant No.2010CB731401), NSFC (grant No. 91024011 and grant No. 61071133), Ministry of Industry and Information Technology of China (grant No. 2010ZX01042-002-003-004), and Science and Technology Commission of Shanghai Municipality (grant No. 09JC1401500).

REFERENCES

[1] N. Eagle and A. Pentland, "Reality Mining: Sensing Complex Social Systems", *Personal and Ubiquitous Computing*, vol. 10, no. 4, pp. 255-268, 2006.

[2] Nathan Eagle and Alex Sandy Pentland, "Eigenbehaviors: identifying structure in routine", *Behavioral Ecology and Sociobiology*, vol. 63, pp. 1057–1066, 2009.

[3] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns", *Nature*, vol. 453, pp. 479-482, 2008.

[4] M. A. Bayir, M. Demirbas, N. Eagle. "Discovering spatiotemporal mobility profiles of cellphone users", *In Proceedings of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks & Workshops*, 2009, pp. 1-9.

[5] C. Ratti, R. M. Pulselli, S. Williams, and D. Frenchman, "Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis", *Environment and Planning B*, vol. 33, no. 5, pp. 727-748, 2006.

[6] J. Candia, M. C. Gonzalez, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabasi, "Uncovering individual and collective human dynamics from mobile phone records", *Journal of Physics A: Mathematical and Theoretical*, vol. 41, 224015, 2008.

[7] Zhenmei Liao, Su Yang, and Jianning Liang, "Detection of Abnormal Crowd Distribution", In *Proceedings of the 3rd IEEE/ACM International Conference on Cyber, Physical and Social Computing*, 2010, pp. 600-604.

[8] Wenbin Zhou and Su Yang, "Outlier Detection on Large-Scale Collective Behaviors", *2011 International Joint Conference on Computational Sciences and Optimization*, accepted

[9] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

[10] Chun Yang, Feiqi Deng, Haidong Yang, "An Unsupervised Anomaly Detection Approach using Subtractive Clustering and Hidden Markov Model", In *Proceedings of the Second International Conference on Communications and Networking in China*, 2007, pages 313-316.

[11] Xuan Dau Hoang, Jiankun Hu, and Peter Bertok, "A program-based anomaly intrusion detection scheme using multiple detection engines and fuzzy inference", *Journal of Network and Computer Applications*, vol. 32, pp. 1219–1228, 2009.

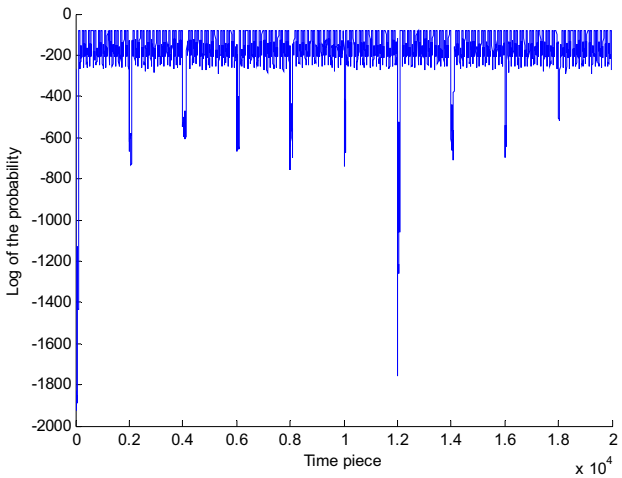

Figure 1. Probability against time piece for artificial data (Unemployment rate= 0; Participation rate=20%)

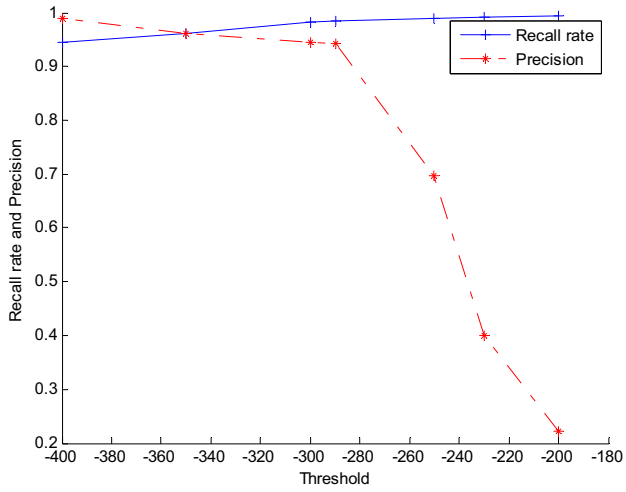

Figure 2. Recall rate and precision for artificial data (Unemployment rate= 0; Participation rate=20%)

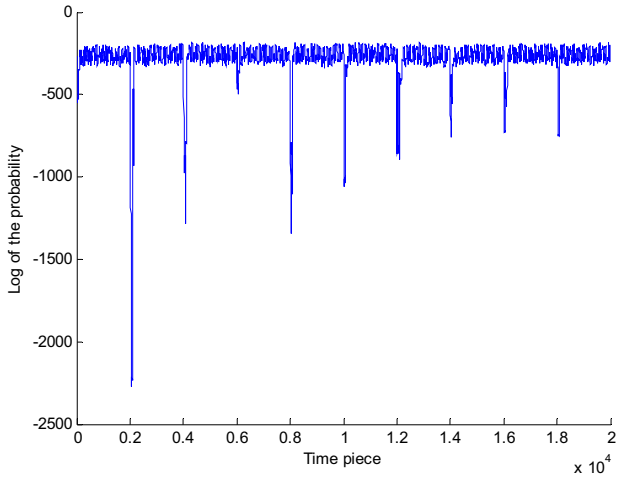

Figure 3. Probability against time piece for artificial data (Unemployment rate=20%; Participation rate=20%)

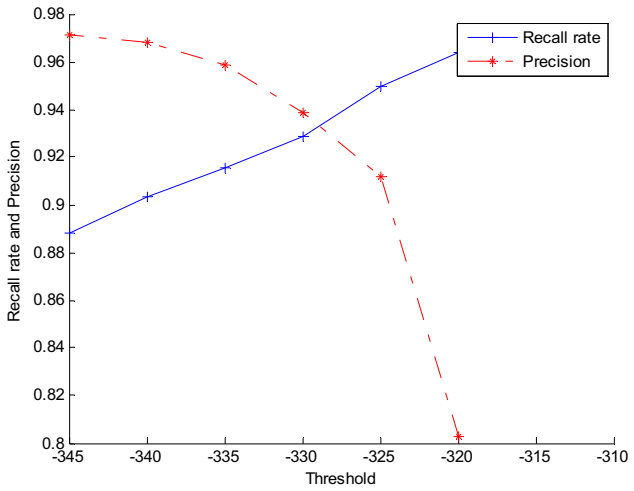

Figure 4. Recall rate for artificial data (Unemployment rate=20%; Participation rate=20%)

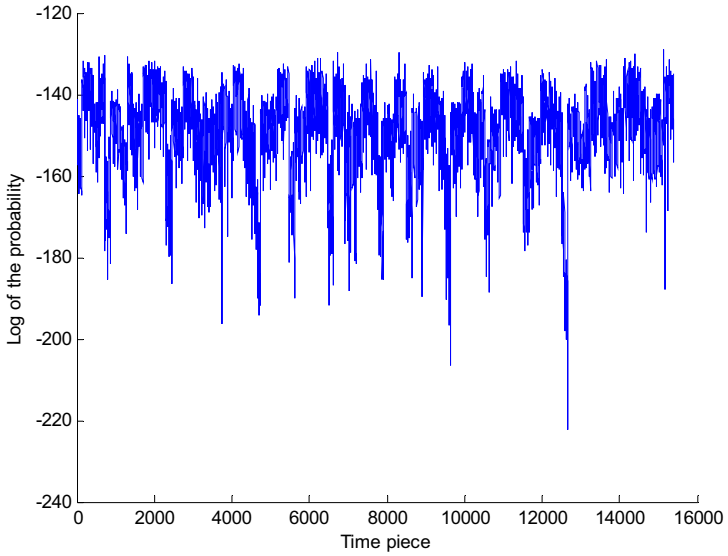

Figure 5. Probability against time piece for real-world traffic data