Path Normalcy Analysis Using Nearest Neighbor Outlier Detection

David Luper, Muthukumaran Chandrasekaran, Khaled Rasheed, Hamid R. Arabnia
Department of Computer Science, University of Georgia, Athens, GA, USA

Abstract - *We present a machine learning technique that recognizes patterns of normal movement, using GPS data and time stamps, to gain the ability to detect regions of time containing abnormal movement. We argue people move throughout regions of time in established patterns, and a person's normal movement can be learned by machines. We use intelligent features extracted from raw GPS data with time stamps, to describe a person's movement over discrete regions of time. Then we use a nearest neighbor approach to determine outliers in a distribution of time regions. We consider outliers as time regions where patterns of established normal movement have been violated. Ultimately, we produce a distance range value for a distribution in conjunction with normalized scores depicting the degree to which each time region contained movement consistent with the other time regions being analyzed. We also produce a classification of each day as normal or abnormal.*

Key Words: Global Positioning System, Outlier Detection, Machine Learning, Unsupervised Learning, Nearest Neighbor, Data Mining

1. Introduction

GPS data is packed full of information. In the past few years the acquisition and use of GPS data has seen a rise in popularity. With this onset of information there needs to be some kind of machine processing to accommodate the plethora of information that has become available. Recent trends in data mining on GPS data consist of efforts such as research by Rogers et al. [1] which used GPS data to indentify traffic lanes and augment road models, research by Luper et al. [2] which used data mining and semantic web technology to discover levels of correlation in human association networks, and research by Yavas et al. [3] which used movement tendencies of mobile phone subscribers to predict location in order to manage wireless communication more efficiently. In addition to these, research by Horvitz et al. [4] mentions a project that has gone online within the last six months from Microsoft, Inc., which is an artificial intelligence platform accessible over the internet, titled Clearflow. It uses machine learning algorithms on collected GPS data to perform intelligent routing of traffic including back roads and main thoroughfares. These papers reflect tendencies of learning from GPS data, and it is our position, that humans have general patterns of movement. They have jobs, social commitments, and appointments, etc., and either due to obligation or preference, people need to maintain their responsibilities which generally entails physically being somewhere at a scheduled time. Outside of general commitments and responsibilities, people have places they prefer to go (points of interest to them). These places may change over time, and they may be a very large, broad collection of locations, but in their movement through these locations on a day to day basis, general patterns develop. It can be beneficial to gain insight into the patterns of movement, specifically when deviations from this established pattern develop. The work by Ashbrook et al. [5] and Hightower et al. [6] propose behavior modeling systems that could benefit greatly from this information. In any behavior modeling system, there needs to be some way of recognizing when a person deviates severely from established trends of activity. Just as it is reasonable to assume humans move in a learnable pattern, it even is more reasonable to assume that during certain time regions they will deviate from that pattern for some unknown reason. To more accurately model behavior, these systems would need to be made aware when someone is violating their normal patterns of movement.

In this paper we propose using outlier detection to isolate abnormal patterns of movement from a distribution of discrete time regions, and we use the degree to which an instance is an outlier to produce a normalized score reflecting how abnormal the pattern of movement is relative to other neighbor instances. This normalized score, in conjunction with the range of the score, provide a metric for outlier detection. Outliers are defined in Hodge et al. [7] as outlying observations that appear to deviate markedly from other members of the sample in which they occur. In this paper outliers are the search target over a distribution of time regions for the purpose of flagging days containing degrees of abnormal movement. We have chosen two present our results in two forms. First, we implement methodology that presents normalized scores and their range value, to depict a given day's degree of normality. Second,

using these normalized scores, we provide a classification of abnormal or not abnormal over the different discrete regions of time in our distribution.

In our approach we extend our analysis across several domains. As such, we make some basic yet important assumptions. First, we assume that technologies exist for GPS data collection and monitoring (such as mobile phones or personal navigation devices). For this research we implement a simulation of a GPS data collection framework which will be discussed later, as well as real data gathered from a test subject. At this point in our work, we are not taking into account privacy constraints that could potentially restrict GPS data gathering, but we focus on demonstrating the benefits and practicality of our methodology. With the recent upswing in applications utilizing the GPS system, i.e., automobiles, mobile phones and other communications devices, the usefulness of GPS data has become apparent and will only continue to grow. This will, without doubt, shadow privacy concerns and a legal framework will have to be developed for handling this increasingly useful data.

2. Data Acquisition

To collect data for our experiments two different techniques were used. They were, simulating data, and gathering real data on a test subject.

2.1 Simulated Data

In order to test our outlier detection scheme on numerous models, we needed some way of intelligently simulating data. We desired to show the robustness of our methodology over different types of movement patterns such as someone who has a nine to five job, someone who has no job, someone who has a route based job, etc. To do this we built a data simulator. The simulator follows *n* number of people through a variable length time region and logs coordinates for these people on 20 second intervals. It assigns people a home, points of interest, regular appointments, jobs, regular bed and wake times, etc. Points of interest for these people were generated in groups of 2 - 5 points for each person per time region for each day. For example, on any given day of the week, between one o'clock and 4 o'clock, a person had a pool of 2 to 5 points of interest they could go to if they were free (meaning they were not at work or in an appointment). The jobs people were assigned varied in type. They included jobs where people worked at a single place, where they worked at a set of 2 to 3 places, where they drove a route all day, and where they mimicked movement like a university student. People were forced to go to their jobs with a 15 minute window each day and they stayed a determined length of time. If someone had a reoccurring appointment on a given day, they were also made to go to that appointment within a 15 minute window of time. During free time (time not at work, appointments, or time when the person would be asleep) the people were allowed to move with randomness based on probabilities they would follow their normal patterns of movement. Throughout each simulation run, we inserted a certain number of abnormal days in which a person would deviate markedly from their normal patterns of movement. This deviation could be that they missed or arrived late for work or an appointment, or they went to a place they had never visited before, or they stayed home instead of going somewhere, etc. The abnormal days were created within reasonable bounds that provided a good way to test our methodology in many different scenarios.

2.2 Real Data

In order to show the effectiveness of our outlier detection scheme, it was necessary to test our methodology on real data. For our experiment we gathered four weeks of GPS data on a test subject, where the subject took around a Bluetooth GPS receiver that was used in conjunction with a mobile phone which logged the coordinates supplied by the GPS receiver every 20 seconds. The GPS used for this experiment was accurate to less than 15 meters under optimal conditions, however, due to lack of satellite verification, cold starting, satellite acquisition time, terrain topology, etc., the data obtained was noisier than an optimal system would have provided. There is discussion in Ashbrook et al. [5] that deals with lapses in data acquisition and there are numerous papers, of which Schmid et al. [8] presents a particularly eloquent methodology, dealing with location extraction (location extraction acts as a filtering mechanism that extracts meaningful locations from noisy data). To extract locations from the raw data we collected, a distance and time based Thresholding scheme was implemented. If logged coordinates were within the threshold distance to previous logged coordinates, over a variable length of time (the time threshold), the collective group of coordinates was viewed as a location. Due to noise, locations the test subject visited were displaced by varying distances on different, subsequent visits to the location (this displacement was usually less than 500 feet). To recognize these displaced locations as one distinct location, a clustering method was used where a distance threshold was defined, then, for each location L_1, any location L_2, within the distance threshold to L_1 was considered connected to L_1. This partitioned the location into groups of connected locations, and any location, within a group of connected locations, was recognized as being the same. There also existed segments of time where holes existed in the data, for instance, while the phone and GPS

receiver recharged, or the subject entered a building. These holes were fixed by assigning the missing data points in set X{} in between two logged points A and B (where A and B are successive, logged points) with the coordinates successfully logged for point B.

3. Feature Computation

A day is considered an outlier if it is in some way far-off or distant from its cluster or prominently different from its neighbors. We have tried to extract important features from a person's longitudinal and latitudinal coordinates sampled every 20 seconds over a variable length of days to classify a particular day clearly as an abnormal day for that person based on their patterns of movement. In extracting features, the raw GPS coordinates were typically made discrete using Discrete Mapping methodology described by work in Luper et al. [4]. This allows locations to be binned and mined accordingly. Out of an initial set of 14 features, the most successful features that we extracted are described first, the rest are briefly described after that.

1. *Coordinate Dispersion Factor* (*CDF_Score*) is a measure of the coordinate dispersion over the day being analyzed.

$$= \frac{\sum_{i=1}^{n} \sqrt{\left(\left(lat(i) - \frac{\sum_{j=1}^{n} lat(j)}{n}\right)^2 + \left(lng(i) - \frac{\sum_{j=1}^{n} lng(j)}{n}\right)^2\right)}}{n} \quad (1)$$

2. *Point Frequency Score* (*PF_Score*) can be calculated using the following formula:

$$\text{PF\_Score} = \sum_{i=1}^{n}(1 - \text{PF}) \quad (2)$$

Where *PF_Score* is the Point frequency score, *n* is the number of points visited on the day being analyzed, and *PF* is the ratio of the frequency that a particular point is visited on a day being analyzed to the average frequency that point was visited on all similar days of the week, over the entire distribution (say all Mondays, all Tuesdays and so on).

3. *Normal Percentage of Time Score* (*NPT_Score*) is the ratio of the sum of percentages of time spent at each visited point on the day being analyzed to the total number of points visited on that particular day.

4. *Anomaly Count* (*ANMLY_Score)* is the summation of points visited on a given day where the specified point was only visited one time in the entire distribution.

5. *Time Region Violation Score* (*TRV_Score)* is a summation of the percentages over the distribution of points a person was not at during a specified region of time. We keep 24 discrete time regions bins, any location a person visits gets placed into any discrete time region bin the visit overlaps. A discrete time region bin can have multiple entries for the same point (see location x in Figure 1). Every location within each discrete time region carries weight *W* where *W* is the percentage, out of the total number of points in the respective bin, that this location constitutes (i.e. in Figure 1 for the 1 time region x would have weight 0.67). Each day in the distribution is compared to these discrete time region bins. For any time region of the current day where there is a point in the previously constructed time region bins which a person is not at, the weights for these violations are summed, producing the *TRV_Score.* In Figure 1, if the day in question visited location x and did not visit location y for the 1 time region, the sum of the *TRV_Score* would increase by 0.33.

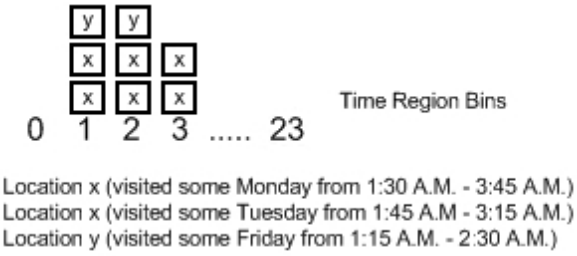

Figure 1 – This shows the time region bins built from the entire distribution. Location x is visited twice during the 1, 2, and 3 hour time regions. Location y is visited once during the 1 and 2 hour time regions.

The rest of the features are briefly described below.

6. *Distance Traveled Factor* (*DTF_Score*): The total Euclidian distance between the longitudinal and latitudinal coordinate values over the whole day.

Hourly Deviation from Mean (HDev_Score) And
Hourly Distance from Mean (HDis_Score):

For these two features each day of the week is divided into 24 time regions. The average latitude

and longitude coordinates are calculated over the entire distribution of data for each time region bin. The local average for the day being analyzed is found for each of its time regions.

7. *Hourly Deviation from* Mean (*HDev_Score*): The average of the Euclidian distances between each of these local averages and the averages over the entire distribution for the given day of the week.

8. *Hourly Distance from Mean* (*HDis_Score*): The sum of the Euclidian distances between each of the local averages and the averages over the entire distribution.

9. *Two Point Score* (*TP_Score*): For each day of the week over the whole distribution, the two points where the person spent the most time at is found. And for the day being analyzed, a score of 0 is assigned if they made it to those points; else a score of 1 is assigned.

10. *Percentage Day Moving* (*PDM_Score)*: The percentage of time a person was not at a visited location.

11. *Percentage Stationary Weight Score* (PSW_Score): (1-*PDM_Score)* / Total points visited on the day being analyzed.

12. *Lateness* Score *(Late_Score*): The average time of arrival was calculated for each location the person visited. The differences between the time of arrival on a given day and the average time of arrival for the point in question are summed over the locations visited for the day being analyzed.

13. *Normal Regions Visited* (*NRV_Score*): All locations the person visited were found and separated into the days of the week they were visited on. Each day was analyzed against any points ever visited on the respective day of the week. Any point not visited within a certain variable time threshold was scored as a violation, and all violations were summed. The weight each violation held was the number of times over the distribution, on the day of the week in question, the point was visited during the respective time frame, divided by the total number of instances of the particular day of the week in question.

14. *Region Weight Score (RW_Score)*: A summation of the total percentage of time the person spent at a location over the distribution, for each location visited on the day being analyzed.

4. Outlier Detection

We used a nearest neighbor implementation over our feature space to find the closest neighboring time regions for each discrete region of time in our distribution. Our nearest neighbor approach used Euclidian distance to compare each time region with every other time region while keeping track of the closest k neighbors for each instance. In our experiment we set $k = 2$ for 36 day distributions of data and $k = 5$ for 119 day distributions. The number of nearest neighbors found is an important variable in our methodology because higher numbers of nearest neighbors found for each instance will affect the average distance to nearest neighbors in later steps. In effect, k loosely establishes the number of instances that constitute an established trend of movement. It is important to note that, for our experiment, we chose to use discrete time regions that represented days of the week. This being the case, any nearest neighbor in our algorithm had to be of the same day of the week as its complement. In other words, we compared Mondays to Mondays, Tuesdays to Tuesday, and so on. It made intuitive sense that we would want to compare homogenous days of the week due to the fact that a person's normal schedule typically fluctuates depending on the day of the week. After calculating the k nearest neighbors for each time region in the distribution, we averaged the distance to the k nearest neighbors for each time region. This presents an average distance in the feature space each time region is from its closest neighbors. We normalized the score for each time region (the formula used to normalize the scores is given in Formula 3) over the entire set of time regions. Along with the range value, this presents us with an intuitive measurement of normality.

$$\text{normalized\_score} = (x - \min) / (\max - \min) \quad (3)$$

Where x is the property being normalized, *min* is the minimum value for the specified property over the entire distribution, *max* is the maximum value for the specified property over the entire distribution.

After computing normalized scores for a distribution, the average normalized score, along the standard deviation of the normalized scores, are used to classify days as abnormal or normal based upon the number of standard deviations a given score lies away from the average.

5. Results

To decide which of our initial 14 features best described patterns of movement throughout a discrete time region, we ran tests over simulated and real datasets. The simulated datasets we used contained 36 days of data. Each

of the different sets had 5 different types of people in it. The 5 different types of people were people who worked at a job in the same place every day, people who worked a job that consisted of a set of 2 or 3 places, people who had a route based job, people who exhibited movement similar to a college student, and people who did not have a job. Each set of data consisted of 50 simulated people. We ran each of the possible combinations of feature sets (2^{14}) over each set of data. The criteria used to judge the best feature set was the one that identified the single, generated abnormal day as the farthest outlier the most times. The other measures of success we kept track of were the distance between the farthest outlier (the generated abnormal day) and the next farthest computed outlier, the average outlier magnitude over entire distribution of days (the outlier magnitude was the average distance to the *k* nearest neighbors), and the standard deviation from this average outlier magnitude. Intuitively, these other measures of success show how far the calculated outlier was from the rest of the instances, the average closeness of the clustered instances, and the deviation from the average closeness. It is desirable for the outlier to be measured very far from the rest of the instances, and the normal days to be measured close to each other, which is why these measures of success were chosen. The 100 feature sets that performed the best were run against 10 simulated datasets (5 containing 36 days of data and 5 containing 119 days of data) and 8 real datasets. The real datasets consisted of 4 weeks of data from the test subject, however, in order to judge how well we were detecting outliers, the test subject spent a day going to places that were abnormal and spending certain amounts of time at each place. The real data was made into 8 sets of data where this abnormal day was treated as a single different day of the week in 7 of the 8 sets (i.e. in one set called "Monday", the abnormal day was inserted over an existing Monday in the distribution, and so on). We kept the same success measurements as before, but we calculated these measurements for each of the 10 simulated and 8 real datasets separately. Once the tests were done, we ranked each feature set's performance over the 10 simulated datasets and the 8 real datasets then averaged their rankings. The top feature set for the simulated data consisted of the *PF_Score*, *ANMLY_Score*, and *TRV_Score*. The top feature set for the real data consisted of *PF_Score*, *ANMLY_Score*, *TRV_Score*, *NPT_Score*, and *CDF_Score*. Using the optimal feature sets over the simulated data, we failed to detect the generated abnormal day as the farthest outlier 5 times out of 500 hundred trials over the different datasets. Of the test we performed on the real data, the abnormal day we collected on the test subject was the farthest calculated outlier over each of the 7 runs which included it. We ran another set of tests over the simulated and real data where an abnormal day was not simulated (in the case of the simulated data) or in the case of the real data, the abnormal day was removed from the distribution. This allowed us to see how our methodology would react in the absence of a true outlier. Table 1 shows the results from the simulated data and table 2 shows results from the real data. In Table 1, the Misses column represents the number of times where the embedded outlier was not calculated as the farthest outlier in the dataset. Distance I is the normalized distance between the farthest and second farthest outlier. Distance II is the average normalized distance score over the distribution. Variance is calculated as the standard deviation from Distance II. Range is the maximum non normalized distance score in the distribution minus the minimum non normalized distance score. In Table 2 these columns hold the same meaning and the DOW column displays the day of the week the abnormal day was inserted over. Formula 3 was used to normalize the scores for these tables.

After an analysis of these results, we ran one other test over the simulated data. We wanted to test the robustness of our detection methodology so we generated more simulated data for each of the 5 types of simulated people in our tests, however, for this test we generated from one to four abnormal days in the sets of data. There was a complete set of data where one abnormal day was inserted, a complete set with two, and so on. Each set of data for this test contained 119 days. We ran our outlier detection scheme on these four sets and then classified each of the days in the distributions as normal or abnormal using the days distance from the average outlier magnitude over the entire distribution. For our classification we made anything farther than four standard deviations from average outlier magnitude an abnormal day and anything closer a normal day. Figure 2 shows the confusion matrix for this test.

6. Analysis

The results obtained from the first set of tests to find the best feature sets yielded *PF_Score*, *ANMLY_Score*, and *TRV_Score* as the best features for distinguishing the simulated data and it yielded *PF_Score*, *ANMLY_Score*, *TRV_Score*, *NPT_Score*, and *CDF_Score* as the top feature set for the real data. Interestingly, the best features for the simulated data were a subset of the best features for the real data. This strengthens the results of the simulated data because the overlap in important features can be attributed to the similarity between simulated and real data. The two features that were in the real data feature set, but not in the simulated data feature set, can be attributed to the fact that simulating human behavior is difficult. There is a level of randomness in real human behavior that is hard to capture in a simulation. One of these two features used on the real dataset and not on the simulated was the *CDF_Score*. This could be due to the fact that in the simulation people were confined to an area roughly the size of the Athens, GA city limits. They were able to stray slightly beyond these

bounds, however, the test subject for the real data worked 45 minutes outside of Athens, GA every Friday, which would have made the deviation score important for distinguishing Fridays. The other feature that was used on the real data and not on the simulated was the *NPT_Score*. This can be attributed to the test subject moving in a more random pattern than the simulator could replicate. This would have increased the disparity between the abnormal days and the normal days in the test subject. Tables 1 and 2 show the results from the outlier detection we implemented. In the simulated data, we correctly computed the generated abnormal day as the outlier with 99% confidence. Our methodology performed better on the real data.

The results obtained on the tests over the simulated data with abnormal days when compared to the results on the distributions with no abnormal days, showed interesting tendencies. The scores on the sets of data are linear separable on the Distance I, Variance, and the Range value. The Distance II value is nearly linearly separable and would only misclassify 2 instances. With these different measurements combined, a Thresholding rule would be able to classify a distribution as not having an outlier or having an outlier with 100% confidence over our dataset. The real data shared these tendencies and if the Distance I, Distance II, and Variance values were given a threshold, only Sunday would have been misclassified as not having an abnormal day. The reason for the disparity among these different sets is that with no distinguished outlier, the distribution of score becomes affected. Without a distinct outlier having a large magnitude, the Distance I value decreases because the second farthest outlier is closer to the farthest outlier. The Distance II value increases when there is no distinct outlier because every score over the entire distribution is closer to the calculated farthest outlier. The Variance increases because of this even distribution of normalized scores as well, and finally, the Range decreases as the farthest calculated outlier becomes closer to the rest of the instances in the feature space. While the real data values for the set that did not include the embedded outlier were different from the sets that did include the embedded outlier, the difference was not as great as the simulated data and in fact weekend days presented scores similar to the distribution with no embedded outlier. This can be attributed to the fact that even without the embedded abnormal day the real data contained some days that were distinctly different from the rest. One of these days included the Martin Luther King Jr. holiday in January where the test subject did not go to school. This Monday was the farthest calculated outlier in the real set that did not include an abnormal day, and after analyzing the Mondays (done be plotting them on a map), this was calculated correctly.

The tests performed over the simulated data containing varying numbers of abnormal days produce the confusion matrix seen in Figure 2. The normal days are considered positive instances and the abnormal days are considered negative instances. Of the instances evaluated that were actually normal, 99.6% were classified correctly while only 0.4% were classified incorrectly. Conversely, of the instances evaluated that were actually abnormal, 90.8% were classified correctly while only 9.2% were classified incorrectly. This test shows the robustness of the detection methodology over different patterns of abnormal movement (i.e. these datasets contained variable numbers of abnormal days). The proposed methodology classifies both normal and abnormal days correctly at a rate that yields significant information gain.

One of the limitations of the experiments was the scarcity of real data. The data gathered from the test subject performed well, but the fact that there was only one distinct distribution of real data limits the conclusions that can be drawn. Also, the one distribution obtained from the test subject consisted of only 4 weeks of data. This is enough duration for our nearest neighbor approach to produce effective results, but the more instances in the dataset, the more established certain patterns of movement become. This enables the detection of outlying days to be more successful. This being said, one of the strong points of this approach is that if someone's patterns of movement change (for instance think of a student at the end of a school semester), the algorithm can adapt quickly, potentially in 2 to 3 weeks (depending on the number of nearest neighbors the algorithm is set to look for). As a final analysis, although the simulated data statistically follows some of the same trends as the real data, in further work it would be beneficial to obtain more real data from a wide distribution of test subjects.

7. Future Research

Current topics being explored further are processing the results of the outlier magnitude calculation to be able to distinguish when there is not an outlier. The normalized score in combination with the Range and other measurement shown in this paper provide a good framework for building on in order to accomplish this. Another potential area of interest for this work is to find a centroid in the feature space and treat the instances as vectors rather than just individual points. This methodology could incorporate Kohonen self organizing maps, and outliers could be determined this way. It would be interesting to compare the methodology described in this paper with this alternate way to see which performs better. Finally, future work will incorporate the algorithm described in this paper to strengthen related work in positional forecasting. When forecasting human locations, it becomes imperative to know when a person's behavior is abnormal so that this can be reflected in the forecasting of the potential future locations.

8. Conclusion

In conclusion, we have shown an interesting methodology for discovering time regions in a distribution that contain movement patterns that are abnormal from the rest. We adopt an approach that is both flexible in nature and powerful, and one that appears reasonable from the results. We describe movement through discrete time regions using various features, and then compute Euclidian distance for each instance to find its relative position in the feature space. In this process we detect and score the normality of the given days as compared to the rest of the distribution.

9. References

[1] Rogers, S., Langley, P., & Wilson, C. (1999). Mining GPS data to augment road models. Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (pp. 104{113). San Diego, CA: ACM Press.

[2] David Luper, Delroy Cameron, John A. Miller, and Hamid R. Arabnia (2007), Spatial and Temporal Target Association through Semantic Analysis and GPS Data Mining, IKE'08 - The 2008 International Conference on Information and Knowledge Engineering

[3] G. Yavaş, D. Katsaros, Ö. Ulusoy, Y. Manolopoulos. A Data Mining Approach for Location Prediction in Mobile Environment. *Data and Knowledge Engineering*, 54, 2, (2005) 121-146.

[4] E. Horvitz, Machine Learning, Reasoning, and Intelligence in Daily Life: Directions and Challenges, Proceedings of ICML, 1999 - research.microsoft.com

[5] Ashbrook, D., Starner, T.: Using GPS to learn significant locations and predict movement across multiple users. Personal and Ubiquitous Computing **7** (2003) 275–286

[6] Hightower, J., Consolvo, S., LaMarca, A., Smith, I., and Hughes, J. "*Learning and Recognizing the Places We Go*", in Proceedings of the Seventh International Conference on Ubiquitous Computing (Ubicomp 2005), pp. 159-176, Sep. 2005

[7] Hodge, V.J. and Austin, J. (2004) *A survey of outlier detection methodologies.* Artificial Intelligence Review, 22 (2). pp. 85-126.

[8] F. Schmid, K. F. Richter. Extracting Places from Location Data Streams. In A. Zipf (Eds.), Workshop Proceedings (UbiGIS), MÄunster, Germany.

[9] Cui ZHU, Hiroyuki KITAGAWA, Spiros PAPADIMITRIOU, Christos FALOUTSOS, Example Based Outlier Detection with Relevance Feedback DBSJ Letters, in DBSJ Journal, Vol.3, No.2, September 2004, pp.5-8

Results on Simulated Data

Job Type	Run Length	Outlier Embedded	Misses	Distance I	Distance II	Variance	Range
Route Based	36 Days	YES	0	0.651685	0.099027	0.079178	1.09451
None	36 Days	YES	2	0.732311	0.078975	0.060297	1.247945
Student	36 Days	YES	0	0.690527	0.069679	0.064373	1.298912
Single Place	36 Days	YES	0	0.735491	0.070966	0.056351	1.294178
Set of Places	36 Days	YES	0	0.685896	0.090925	0.070501	1.23278
Route Based	119 Days	YES	0	0.691363	0.047015	0.054092	1.270604
None	119 Days	YES	0	0.735481	0.037386	0.041517	1.381661
Student	119 Days	YES	1	0.613785	0.034397	0.050755	1.373909
Single Place	119 Days	YES	1	0.6957	0.031647	0.042454	1.409393
Set of Places	119 Days	YES	1	0.670651	0.036424	0.045195	1.380998
Route Based	36 Days	NO	N/A	0.242408	0.193354	0.175182	0.912614
None	36 Days	NO	N/A	0.252645	0.230111	0.175393	0.757411
Student	36 Days	NO	N/A	0.293269	0.177678	0.154794	0.827709
Single Place	36 Days	NO	N/A	0.289758	0.209804	0.162209	0.731772
Set of Places	36 Days	NO	N/A	0.208055	0.231486	0.182005	0.796987
Route Based	119 Days	NO	N/A	0.152807	0.122097	0.168307	0.969638
None	119 Days	NO	N/A	0.245889	0.101207	0.119415	1.016793
Student	119 Days	NO	N/A	0.230524	0.088545	0.118736	1.036878
Single Place	119 Days	NO	N/A	0.190848	0.098171	0.129749	0.928613
Set of Places	119 Days	NO	N/A	0.222621	0.10998	0.128067	0.996013

Table 1 - The averaged results from outlier detection on each of the 50 people in their respective simulated datasets.

Results on Real Data

DOW	Outlier Embedded	Misses	Distance I	Distance II	Variance	Range
Sunday	YES	0	0.1695993	0.368487	0.2541217	0.7563016
Monday	YES	0	0.5357498	0.2200804	0.1374091	1.3622149
Tuesday	YES	0	0.5242837	0.2089656	0.1366711	1.4082673
Wednesday	YES	0	0.5141501	0.2270943	0.1461674	1.336075
Thursday	YES	0	0.4690308	0.22873	0.1480276	1.1712802
Friday	YES	0	0.5437422	0.2124483	0.1444024	1.3222431
Saturday	YES	0	0.3534024	0.2968458	0.1957365	0.9880988
N/A	NO	N/A	0.2305651	0.3224752	0.2131091	1.0745948

Table 2 –The averaged results from the outlier detection on each of real datasets

Confusion Matrix on Simulated Data

	normal	abnormal	← actual values
classified as ↓			
normal	116007	229	
abnormal	493	2271	

Figure 2 – Confusion matrix for experiments using simulated data. The threshold for deciding abnormal days was set to 4 times the standard deviation from the average normalized distance value over the entire distribution.