OCSVM(One-class SVM)과 인간의 이동을 이용한 GPS 데이터의 이상 현상 검출에 관한 연구

김 우 중, 송 하 윤
홍익대학교 정보 컴퓨터 공학부 컴퓨터공학 전공
e-mail : veritaswjkim@gmail.com

A Study on Novelty Detection of GPS Data Using Human Mobility and OCSVM(One-class SVM)

Woojoong Kim, Ha Yoon Song
Department of Computer Engineering, Hongik University

요 약

인간은 목적지를 향하여 가는 방법의 선택에 있어서 가고자 하는 목적, 목적지, 출발 시간 등에 영향을 받는다. 그러나 이러한 매개변수들과 더불어 중요하게 고려되는 것은 바로 인간의 습관이다. 다시 말해 인간이 목적지로 가는 방법을 선택하는데 습관이라는 매개변수와 밀접한 영향이 있다는 것이다.

이를 미루어 볼 때, 인간의 이동은 습관으로 인해 대부분 특정한 범주 안에서 이동을 할 것이라는 추측할 수 있다. 나아가, 사람들이 흔히 들고 다니는 GPS장치에서 측정된 데이터가 추측한 속성으로 인해 범주를 벗어나는 이상현상을 검출하는 것으로 확장을 할 수 있다. 즉, GPS장치에서 측정된 데이터는 개인별로 클래스화(Classification)가 가능함을 추론할 수 있다.

본 논문에서는 실제 사람이 이동한 좌표를 바탕으로 시간당 변화량을 계산하여 좌표에 사상시켰다. 그리고, 단일 클래스 서포트 백터 머신(OCSVM)을 가지고 클래스화 했으며, OCSVM의 커널 함수 내의 변수인 γ에 따라 클래스의 크기 혹은 클래스 내부의 밀도에 영향을 받음을 알 수 있었으며, 그 둘 사이에는 적절한 교환(Tradeoff)이 발생하였다는 결론이 나왔다.

1. 서론

인간의 이동에 관한 연구는 현재 뿐만 아니라 과거에도 지속적으로 연구를 진행 했었다. 그 중에서 인간의 이동을 심리학적인 요소를 찾아서 분석을 하였다.

연구의 예는 다음과 같다. 인간이 지도를 이용하여 경로의 계획을 잡는데 있어서 어떤 요소들이 작용 하는지에 대한 연구가 있다. 결과에 따르면, 직선의 도로가 거리가 더 멀었는데도 불구하고 짧은 곡선보다 긴 직선의 경로를 선택한 경우가 많았다[1]. 또 다른 연구로 사람이 이동 수단(Travel Mode)을 결정하는데 있어서 기존에 해 왔던 습관이 얼마나 많은 영향을 미치는지에 대한 연구도 있다[2]. 이 연구에 따르면 이동 수단에 대한 정보, 목적지까지의 거리와 더불어 습관도 무척 중요한 고려대상 중에 하나라는 결과가 나왔다. 그리고 자가용과 대중교통의 선택과 관련된 요소들을 연구한 경우도 존재한다[3, 4, 5]. 기능적인 면 보다 심리학적 요인이 많이 작용을 한다는 것인데, 그 중에서도 습관을 강제적으로 변화시켜서 통근 습관을 바꾸게 한 연구가 있다[6]. 이 연구에 따르면 인간이 이동하는데 있어서 습관이라는 매개변수가 상당한 영향을 미친다는 결과가 나왔다. 위의 연구에 따르면 기존에 자가용을 타고 출근을 하던 사람이 강제적인 외부 자극(위의 연구에서는 출근 시 사용되던 고속도로의 공사가 외부 자극)으로 인해 대중교통을 이용했는데, 외부 자극이 더 이상 주어지지 않는데도 불구하고 계속 대중교통을 이용했다는 결과가 나왔다. 이러한 것으로 미루어 보아, 인간의 이동은 지금까지 행동한 습관들에 영향을 받는다는 추론을 할 수 있다.

따라서, 본 논문에서는 인간의 행동 습관을 GPS 데이터에 적용해 이것을 클래스화(Classification)하여, GPS데이터가 과연 이상이 있는 값인지 아닌지를 판단하는 연구를 하였다.

논문의 목차는 두 번째 섹션에서는 인간의 위치 정보를 바탕으로 클래스화 하기 위한 방법인 단일 클래스 서포트 벡터 머신(One-class Support Vector Machine; OCSMVM)의 이용에 대해 간략히 설명한다. 세 번째 세션 에서는 실험 환경에 대해 설명하고 네 번째 세션 에서는 실험 결과를 분석한다. 마지막으로 앞으로 연구의 진행 방향에 대해 서술하고 마치도록 한다.

2. One-class SVM과 이용사례

OCSVM(One-class Support Vector Machine)은 2001

년에 B. Schölkopf등이 제안한 알고리즘이다[7]. 기존의 SVM(Support Vector Machine)은 -1과 +1의 두 개의 클래스로 나누어 클래스화를 한 것과는 다르게 오직 한 개의 클래스만이 존재하는 경우(Unlabelled Data)에 이용을 한다[7]. 이러한 클래스는 해당 데이터들의 분포를 추정하는데 사용이 된다. 만일 학습해야 할 데이터가

$$X_1, X_2, ..., X_n \in \aleph, n \in N$$

이라고 가정하자. 여기서 N은 관측 결과의 수이다. 그리고 이 자료를 데이터가 존재하는 차원보다 높은 차원의 특징지도(Feature Map)에 넣는다. 그리고 이 데이터를 특징공간(Feature Space)에 표시 후 초구(Hyper-sphere)를 그린다. 여기서 초구 안에 대부분의 데이터를 넣을 수 있도록 하며, 그 크기가 최소가 되도록 최적화 하는 문제(Optimal Problem)이다.

이러한 OCSVM을 이용한 연구의 사례는 다음과 같다. 우선 문서를 클래스화 한 경우이다. 문서의 클래스를 만든 다음에 이것과 관련 없는 버전의 문서를 다른 클래스로 놓는 방법을 이용하였다. 이러한 정보를 문서를 검색하는데 이용을 하였다[8]. 그리고 이미지를 검색하는 용도로 사용을 하기 위한 연구도 있다[9]. 하지만 무엇보다도 많이 이용한 분야는 이상 현상을 검출(Abnomaly Detection or Novelty Detection)하는데 이용이 된다[10, 11, 12]. 특히 윈도우즈 운영체제의 레지스터에 대한 접근의 이상현상을 검출하는 연구도 존재하였다[11]. 이처럼 2001년 발표 된 OCSVM은 기존의 SVM과는 조금 다르게 이상현상의 검출을 하는데 많이 이용이 된다. 따라서 본 논문에서도 이상현상의 검출을 위하여 OCSVM을 이용하도록 하겠다.

3. 실험

실험 환경은 다음과 같다.

· 개발 OS : Windows XP SP3
· 사용 언어 : Java 1.6(Swing)
· 사용 라이브러리 : LIBSVM ver.3.0
· 사용된 커널 함수 : RBF

LIBSVM(A Library for Support Vector Machines)은 OCSVM 뿐만 아니라 다른 SVM을 다양한 언어에서 제공하는 라이브러리이다. 사용이 편리하고 많이 쓰이는 라이브러리이기에 LIBSVM을 가지고 실험하였다[14].

RBF(Radial Basis Function)은 OCSVM에 적용한 커널 함수이다. SVM의 특성상 차원을 보다 높게 바꾸어 클래스화 하기 때문에 커널 함수가 필요한데 그 때 이용된 함수이다. 이 함수의 식은 다음과 같다.

$$\Phi(u, v) = e^{(-\gamma \times |u-v|^2)} \quad (1)$$

여기서 $\Phi(u,v)$는 커널 함수를 의미한다. u와 v는 공간의 좌표이며, γ는 사용자가 변화를 시키는 파라미터이다.

실험 진행은 우선 하루에 사람이 지나간 좌표를 불러온다. 여기서 사용된 좌표는 실제 미국의 일리노이 주에서 진행되었던 실험의 데이터를 이용하였다[13]. 그리고 실제 이 데이터를 이용하기 위하여 단위 시간을 맞추어야 한다. 따라서 위도와 경도의 초당 변화량을 구한다. 즉,

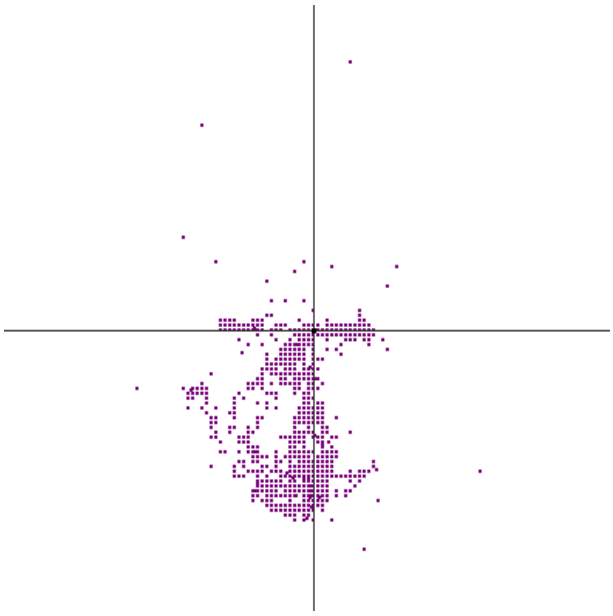
$$\Delta Latitude = \frac{Latitude_2 - Latitude_1}{t_2 - t_1} \quad (2)$$

$$\Delta Longitude = \frac{Longitude_2 - Longitude_1}{t_2 - t_1} \quad (3)$$

이다. 그리고 이 값의 좌표를

$$(\Delta Latitude, \Delta Longitude) \quad (4)$$

이라고 하자. 여기서 $\Delta Latitude$와 $\Delta Longitude$는 매우 작기 때문에 좌표 상에 잘 나타나지 않는다. 따라서 300,000배 확대하여 좌표에 그리면 다음의 <그림 1>과 같다.

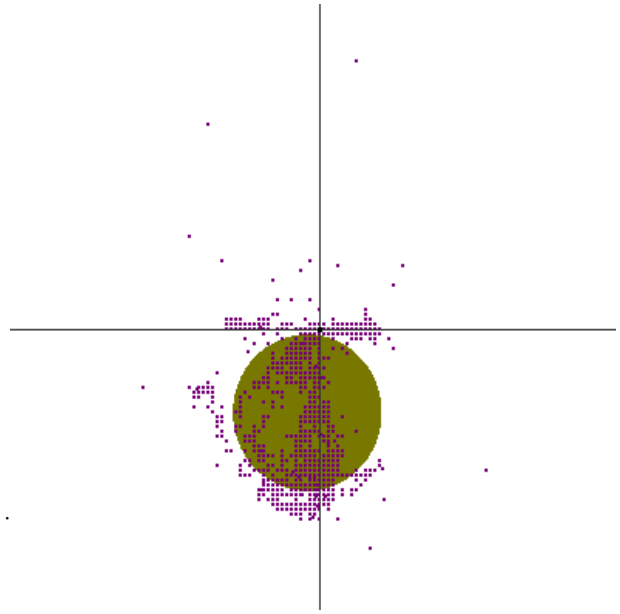


<그림 1> 좌표에 표시한 결과
(측정 된 시간 - 2006.05.30. 18:28:33 ~ 19:10:13)

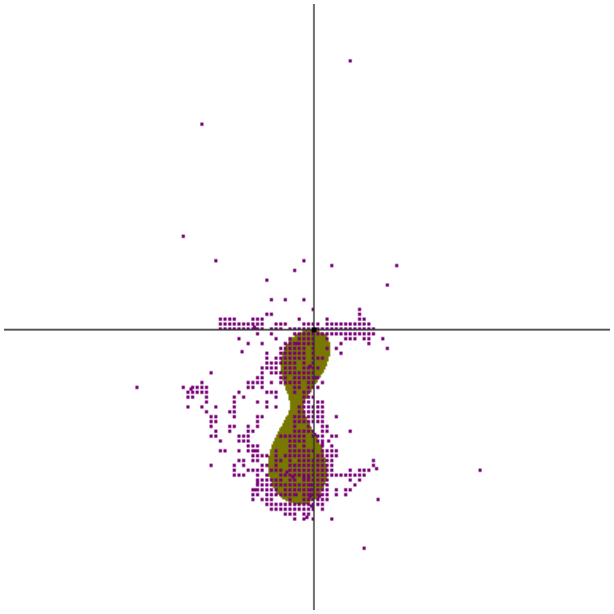
이 후 그려진 좌표들의 집합을 OCSVM을 통하여 학습시킨 다음 Class를 표기한다. 이 때, 식 (1)을 보면 γ값에 따라서 커널 함수의 결과가 달라진 것을 알 수 있다. 따라서 γ값을 변화시키면서 실험을 진행한다.

4. 결과

실제 결과는 아래의 그림과 같다.



<그림 2> 클래스화를 한 결과(γ=0.00000005)



<그림 3> 클래스화를 한 결과(γ=0.0005)

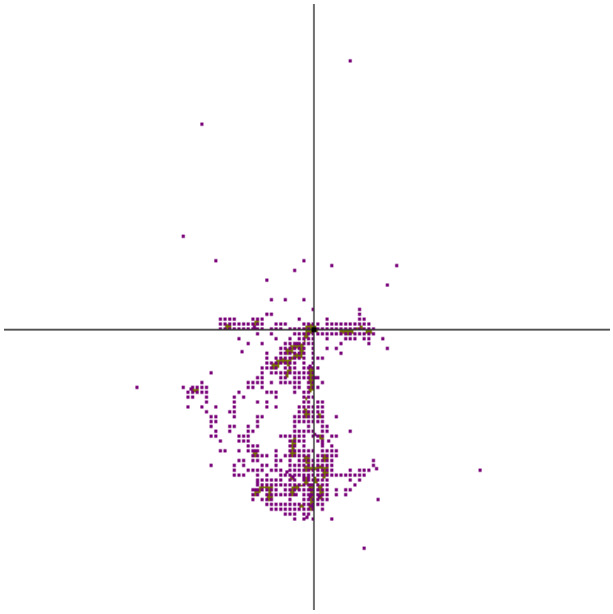
두 결과를 확인해 보면 다음과 같다. γ=0.00000005일 때가 γ=0.0005일 때 보다 훨씬 큰 클래스 영역을 나타내고 있다. 다시 말해 γ값에 따라서 클래스화 된 범위가 다름을 볼 수 있다.

그리고 이 클래스 안에 얼마나 많은 포인트가 존재하는지에 대한 계산을 하겠다. γ값에 대해서 클래스 영역 내부의 점이 얼마나 있는지 밀도를 구하는 식을 정의하면

$$Density_{point} = \frac{N(Point)}{ClassArea} \quad (5)$$

이다. 여기서 $N(Point)$는 Point의 수이고, $ClassArea$는 영역을 그릴 때 이용한 픽셀의 수이다.

이제 좀 더 γ값을 크게 해 본 결과와 종합해 보도록 하겠다. <그림 4>는 <그림 2>나 <그림 3>보다 γ값을 더 크게 하고 시행하였다.



<그림 4> 클래스화를 한 결과(γ=0.05)

위의 결과들을 분석하면 다음과 같다. 첫 번째로 γ값에 따라서 클래스의 크기가 다름을 알 수 있다. 즉 γ값이 작아지면 작아질수록, 클래스의 범위가 넓어졌다. 만일 이 클래스 내부에 존재하면 이상 현상이 발생하지 않는 허용 가능한 범위 내에서의 이동이라 할 수 있다. 즉 해당 클래스 내에 있으면 훈련 된 점들과 같은 클래스라 할 수 있기 때문이다. 클래스가 클 경우 허용 가능 범위가 넓은 것이 된다. 즉, γ값이 작아지면 작아질수록, 클래스가 커지게 되어 이상 현상이 발생하지 않는 허용 가능한 이동 범위가 넓어지게 되는 것이다.

두 번째로 클래스 내부에 있는 점의 밀도이다.

	γ = 0.00000005	γ = 0.0005	γ = 0.05
T	2299	2299	2299
P	1132	1031	873
A	7289	2698	675
D	0.1553025	0.3821349	1.293334

<표 1> γ값에 따른 밀도 계산
(T : 좌표에 사상된 점의 수, P : 영역 내부의 점의 수
A : 영역을 그리는데 사용된 점의 수, D : 밀도)

<표 1>의 결과를 보았을 때, γ값의 증가에 따라 밀도가 늘어난 것을 알 수 있다. 종합해 보면, γ값이 커지면서 클래스의 범위는 줄었는데 밀도는 증가하게 된 것이다.

<표 1>에서 보면, 클래스가 클 경우(γ값이 작은 경우) 많은 포인트를 포함 하지만, 밀도가 낮아 실제로 이상현상

이 발생 한 포인트도 이상 현상이 아니라고 판단 할 수 있다.

반면, 클래스가 작은 경우(γ값이 큰 경우) 적은 포인트를 포함하는 문제가 발생한다. 하지만 밀도가 커서 해당 클래스 내부의 포인트는 높은 신뢰성을 가지고 이상현상이 아니라고 판단 할 수 있다.

따라서, γ값에 따라 클래스의 크기와 밀도의 교환(Tradeoff)가 발생한다. 민감한 수행(Sensitive Case)에 대해서 높은 신뢰성을 갖는 검출을 원할 경우 γ값을 증가시키면 된다. 그리고 만일 민감하지 않은 수행의 경우에 대해서는 γ값을 작게 하여 클래스의 크기를 크게 해서 수행하면 될 것으로 판단된다.

5. 향후 연구 계획

본 실험에서 드러난 바와 같이, 몇 가지 정책에 대한 정의가 필요하다. 첫 번째로, γ값의 크기에 대한 문제이다. γ의 크기 밖의 좌표를 이상현상이라고 간주할 경우, 그 크기가 명확하지 않다면 이상현상에 대한 검출의 정확도 역시 명확하게 되지 않기 때문이다. 반면 너무 큰 영역은 이상현상을 유효한 좌표로 인식하게 된다. 따라서 명확한 γ값의 정의가 필요하다.

두 번째로, 클래스 외부의 좌표에 대한 오차범위이다. 본 실험 결과에서도 보다시피, 적지 않은 좌표 값들이 클래스 외부에 있다. 따라서 어느정도의 범위 까지가 유효한지 정의하는 연구가 필요하다.

세 번째로, 정지상태에 대한 정책의 설정이다. 보통 이동을 하는 경우에, 정지를 한 경우와 이동한 경우 두 가지의 경우로 나뉘게 된다. 보통 길을 걷다 보면 정지하는 경우도 상당히 많은 편인데, 이 값들까지 모두 포함해서 학습을 시키는 경우와 제외하고 시키는 경우에 대한 유효범위의 정확성을 연구해 볼 필요가 있다.

네 번째로, 더욱 더 많은 데이터의 축적이다. 이번 실험은 한명이 하루 동안 이동한 거리에 대한 실험이었는데, 과연 많은 데이터가 축적되어 학습시킨 후에 시행하였을 경우에 어떤 결과가 나오는 지 연구해 볼 필요가 있다.

마지막으로, 잘못된 값이라고 판명이 되었을 경우, 그 시점에서의 GPS 좌표 값을 예측을 할 수 있어야 한다. 예를 들어서, 유도무기체계의 경우, GPS를 통한 유도체계가 많은데 교란이 되었을 시 탐지할 뿐만 아니라 실제 값의 예측 체계가 필요하다. 따라서 향후에 잘못된 값의 정정을 할 수 있는 모델의 연구가 필요하다.

6. 참고문헌

[1] J. N. Bailenson, M. S. Shum & D. H. Uttal, Road climbing: Principles governing asymmetric route choice on maps, Environmental psychology, vol.18, issue.3, pp.251-264, 1998

[2] B. Verplanken, K. Aarts & A. V. Knippenberg, Habit, information acquisition, and the process of making travel mode choice, European journal of social psychology, vol.27, issue.5, pp.539-560, 1997

[3] T. Garling, S. Fujii & O. Boe, Empirical tests of a model of determinants of script-based driving choice, Transportation research part F: Traffic psychology and behaviour, Vol.4, issue.2, pp.89-102, 2001

[4] L. Steg, C. Vlek & G. Slotegraaf, Instrumental-reasoned and symbolic affective motives for using a motor car, Transportation research part F: Traffic psychology and behaviour, vol.4, issue.3, pp.151-169, 2001

[5] L. Steg, Car use: Lust and must. Instrumental, symbolic and affective motives for car use, Transportation research part A: Policy and practice, vol.39, issue.2-3, pp.147-162, Feb-Mar.2005

[6] S. Fujii & T. Garling, Development of script-based travel mode choice after forced change, Transportation research part F: Traffic psychology and behaviour, vol.6, issue.2, pp.117-124, 2003

[7] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, & R. C. Williamson. Estimating the support of a high-dimensional distribution. Neural computation, vol.13, 1443-1471, 2001

[8] L. Manevitz & M. Yousef, One-class SVMs for document classification, The journal of machine learning research, vol.2, pp.139-154, 2002

[9] Y. Chen, X. Zhou & T. Huang, One-class SVM for learning in image retrieval, 2001 International Conference, vol.1, pp.34-37, 2001

[10] X. Zhang, C. Gu & J. Lin, Support Vector Machines for anomaly detection, 2006 WCICA, pp.2594-2598, 2006

[11] K. Heller, K. Svore, A. Keromytis & S. Stolfo, One class support vector machines for dtecting anomalous windows registry accesses, In proceedings of the workshop on data mining for computer security, 2003

[12] J. Ma & S. Perkins, Time-series novelty detection using one-class support vector machines, Neural networks, vol.3, pp.1741-1745, 2003

[13] http://www.cs.uic.edu/~boxu/mp2p/gps_data.html

[14] C. Chang & C. Lin, LIBSVM: A library for support vector machines, available at http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001