Isomap을 이용한 향상된 기능의 오존 경보 예측기 구현

이태훈, 김한주, 전용권, 윤성로
고려대학교 전기전자전파공학과
e-mail : withdove@korea.ac.kr, sryoon@korea.ac.kr (corresponding)

Enhancing the Performance of an Ozone Day Predictor Using Isomap

Taehoone Lee, Hanjoo Kim, Yongkweon Jeon, Sungroh Yoon
School of Electrical Engineering
Korea University, Seoul 136-713, Korea

요 약

본 논문에서는 Isomap을 통해 기상 정보에서 특징을 추출하여, 보다 향상된 오존 경보 예측시스템의 구현을 제안한다. 큰 흐름은 전처리 과정과 특징 추출 과정 및 후처리 과정을 통해 정제한 데이터를, 기계 학습에 널리 사용되고 있는 SVM (Support Vector Machine) 등의 분류기로 오존 경보에 대한 예측을 하여 성능을 측정한다. 또한, 압축된 데이터를 분석하여 원 데이터에서의 중요한 특징들이 무엇이었는지를 분석하였다. 분류기의 실험 결과, 기후 데이터에서의 특징 추출은 제안된 Isomap 방법이 PCA 방법에 비해 성능이 우수한 것을 알 수 있었으며, 원래 데이터를 분류한 결과에 비해서는 15~35%정도가 향상되었다. 그리고 실험에 사용된 72가지의 Feature들 중, Tb, WSa, WSp 의 정보가 오존 경보 예측에 주요한 요인 인 것으로 분석되었다.

1. 서 론

오존 수준 (Ozone level)이 일정 수치를 넘어가면 인간의 건강을 포함해, 농업이나 관광 등 우리 일상의 중요한 부분에 악영향을 끼치는 것은 잘 알려져 있다. 따라서, 오존 수준이 위험수치에 이르기 전에 경보를 해줄 수 있는 시스템의 구현이 반드시 필요하다.[1] 이에 따라, 기상 관측 자료에 데이터 마이닝 기법을 적용한 다양한 선행연구가 진행되었으며, 그 중 하나로 최소 극대화 상관 분석을 이용한 경우가 있다.[2] 이는 제안한 알고리즘으로 훈련 데이터로 오존 수준이 높은 날 (Ozone day)과 평범한 날 (Normal day)의 모델을 만들고, 테스트 데이터로 각 모델에 대해 분류를 수행하여 얼마나 분류가 잘 되는지를 비교하는 연구였다. 하지만 수많은 기후 특성 중 근본적으로 어느 것이 오존 경보에 영향을 미치는지에 대한 고찰이 논외로 되어있는 상황이다. 본 논문에서는 데이터 마이닝 기법 중 특징 추출 (Feature Extraction) 에 초점을 맞추어, Isomap 알고리즘을 이용한 보다 향상된 오존 경보 예측기의 구현 방법을 제안한다. 또한, 기상 정보의 특징 추출로 Isomap이 얼마나 좋은 성능을 가지는지에 대한 비교 실험을 위해 특징 추출에 널리 사용되고 있는 주성분 분석(Principal Component Analysis)을 함께 테스트 하였다.

본 논문에서는 Isomap을 통해 72차원의 공간을 고유값의 분포를 보고 저 차원 공간으로 사영을 한다. 또한, 이렇게 축소된 공간에서의 특징들이 원래 공간의 어떠한 특징들과 매칭되는지를 분석하여 기후데이터의 어느 특징이 오존 경보와 관련이 높은지를 알아낸다. 즉, 데이터에 맞는 전처리 기법을 제안하여 분류기의 성능을 높이고, 오존 수준의 위험도는 어느 특징과 가장 관련이 있는지 분석을 하였으며, 그 개략적인 과정은 다음과 같다. 첫째로, UCI Machine Learning Repository [3] 에 있는 Ozone Level Detection 데이터를 가져왔다. 이는 2536개의 레코드가 있으며 Missing Values가 없는 1847개의 레코드로 실험을 하였다. 각 레코드는 하루의 기상 관측 데이터가 기록되어 있으며, 72가지의 Feature들로 이루어져있다. 여기에는 태양열 복사, 일출·일몰시 바람의 세기, 온도, 배기가스와 관련된 요소들이 측정되어있고 그 날이 Ozone day 였는지 Normal day 였는지 구분이 되어있다. 그리고 전처리 과정을 통해 이 데이터를 가공하게 된다. 이 안에는 정규화 및 상수배, Isomap 등이 들어가게 되며, 이는 3장 실험 방법에서 자세하게 다룰 것이다. 이렇게 정제된 데이터에 분류기를 적용하게 되는데, 실험에 사용될 분류기는 SVM (Support Vector Machine) 과 Centroid 분류기이다. SVM 은 기계 학습 분야에서 널리 사용되고 있는 분류기로써 두 그룹을 나눌 수 있는 최적의 초평면 (Optimal Hyper Plane) 을 구하여 임의의 레코드가 초평면 위에 있는지 아래 있는지에 따라 Ozone day 혹은 Normal day 로 분류하게 된다. 또한, Centroid 분류기는 두 그룹의 대푯값인 Centroid 레코드를 구하여, 임의의 레코드를 가까운 Centroid 의 그룹으로 분류한다. 이렇게 측정한 분류기의 정확도와, Isomap 과 PCA 가 추출한 특징들이 원래 데이터의 어느 요소들과 관련이 있는지 분석하게 된다

2. 예측기 구현

2.1. 분석 자료

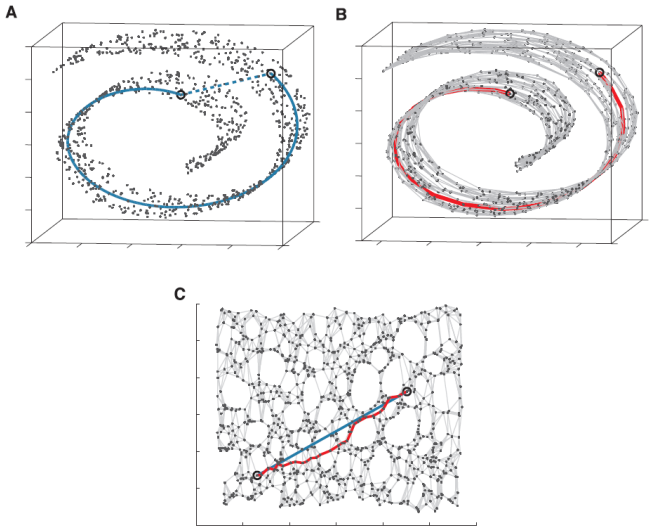
본 연구는 UCI Machine Learning Repository [3] 의 Ozone Level Detection 데이터를 사용하였으며 각 레코드는 그 날이 Ozone day 였는지 Normal day 였는지 구분이 되있고, 아래의 [표 1] 과 같은 여러가지 Feature 들로 이루어져있다.

O3	Local ozone peak prediction
Upwind	Upwind ozone background level
EmFactor	Precursor emissions related factor
Tmax	Maximum temperature in degrees F
Tb	Base temperature where net ozone production begins (50 F)
SRd	Solar radiation total for the day
WSa	Wind speed near sunrise (using 09-12 UTC forecast mode)
WSp	Wind speed mid-day (using 15-21 UTC forecast mode)

[표 1] Attribute 정보

2.2. Isomap

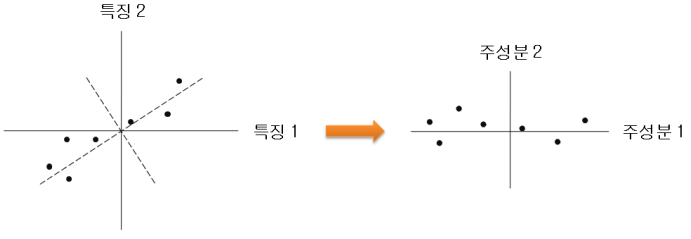
실험에 사용된 데이터는 72차원을 가지고 있다. 하지만 분명히 오존 경보 예측에 불필요한 기상 특징들도 있을 것이며, 오히려 분류기의 성능을 감소시키는 요소들도 있을 것이다. 본 논문에서는 Isomap을 통해 72차원의 공간을 저 차원 공간으로 사영을 하며 이렇게 축소된 공간에서의 특징들과 72차원의 원 공간에서의 어느 특징들이 연관되어 있는지를 분석하여 오존 경보 예측에 필요한 기후 요소들을 분석한다 Isomap 알고리즘은 먼저 각 기후 레코드의 인접 레코드를 정의해야 하는데 정의 하는 방법에 따라 ε-Isomap과 K-Isomap으로 나뉜다. [4] 전자는 거리가 ε 이내인 것들을, 후자는 가장 가까운 K 개의 레코드들을 이웃 레코드로 정의한다 그리고 새로운 공간을 구성하기 위해 모든 레코드 사이의 거리를 새로운 방식으로 계산하는데 여기에는 플루이드 알고리즘 혹은 다익스트라 알고리즘을 사용한다 이는 레코드간의 이웃관계를 정의한 그래프에서 각 레코드간의 최단거리를 구해야하기 때문인데 근거리 안에서만 유클리디안 공간이고 원거리에서는 비 유클리디안 공간 인것을 가정하기 때문이다 이렇게해서 모든 레코드 사이의 최단거리를 계산하게 되면 N개의 레코드에 대해 NxN 거리 행렬이 만들어진다. 마지막 과정은 이 거리 행렬을 고유값 분해하여 저차원 공간으로 사영 하는데 사용할 행렬을 찾아내고 이를 통해 보다 단순한 공간으로 맵핑을 시킨다.



[그림 1] Isomap

2.3. 주성분 분석 (Principal Component Analysis)

주성분 분석은 통계학에서 나온 방법으로써 여러개의 변수들이 내포된 자료를 분석할 때 유용하게 사용된다 그 핵심은 원 변수들의 몇 개의 일차결합을 통해 단순한 구조를 갖는 자료로 축소시키는 것이며 이러한 몇 개의 일차결합을 주성분이라고 부르게 된다 주성분 분석에서 주성분을 찾는 것은 자료에 대한 통계적 모형이나 어떠한 가정을 필요로 하지 않으며 공분산 행렬의 고유값과 고유 벡터에 의해 주성분이 결정되기 때문에 수학적인 절차라고 볼 수 있으며 [5], [그림 2] 에 주성분 분석의 개략적인 그림이 나와 있다.

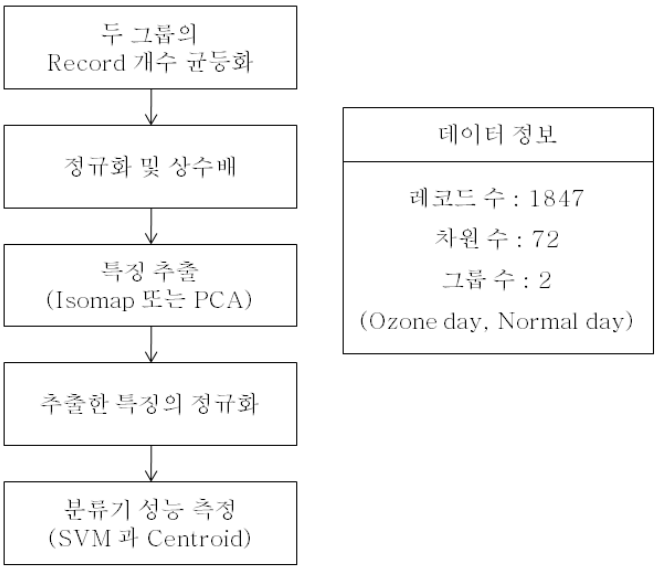


[그림 2] 주성분 분석

주성분 분석은 원래의 공간에서 분산이 가장 큰 축을 찾는 것이 목적이라고 할 수 있다. 또한, 본 연구에서는 주성분의 에너지라고 할 수 있는 고유값들의 분포를 관찰한 뒤, 원 공간의 에너지를 거의 보존하면서 주성분의 개수를 줄일 수 있도록 하였으며, 이렇게 구한 주성분은 다시 원래의 특징과 비교하여 기후 관찰 데이터의 어느 요소가 가장 분류에 영향을 끼치는지 분석하게 된다. 그리고 주성분을 추출 한다는 점에서 기후 데이터의 특징 추출 알고리즘으로 제안한 Isomap 과 같은 역할을 하고 있으므로, Isomap 의 대조군으로 PCA 를 함께 실험하였다.

3. 실험 방법 및 결과

실험은 아래 [그림 3] 과 같은 과정으로 진행하였다. 먼저, 레코드의 숫자를 볼 때 Normal day 가 Ozone day 보다 상대적으로 많기 때문에 Normal day 인 Record 를 임의로 추출하여 그 숫자를 줄였다. 이렇게 해서 실험에 사용할 두 레코드 군집의 크기를 균등하게 조절 하였다.

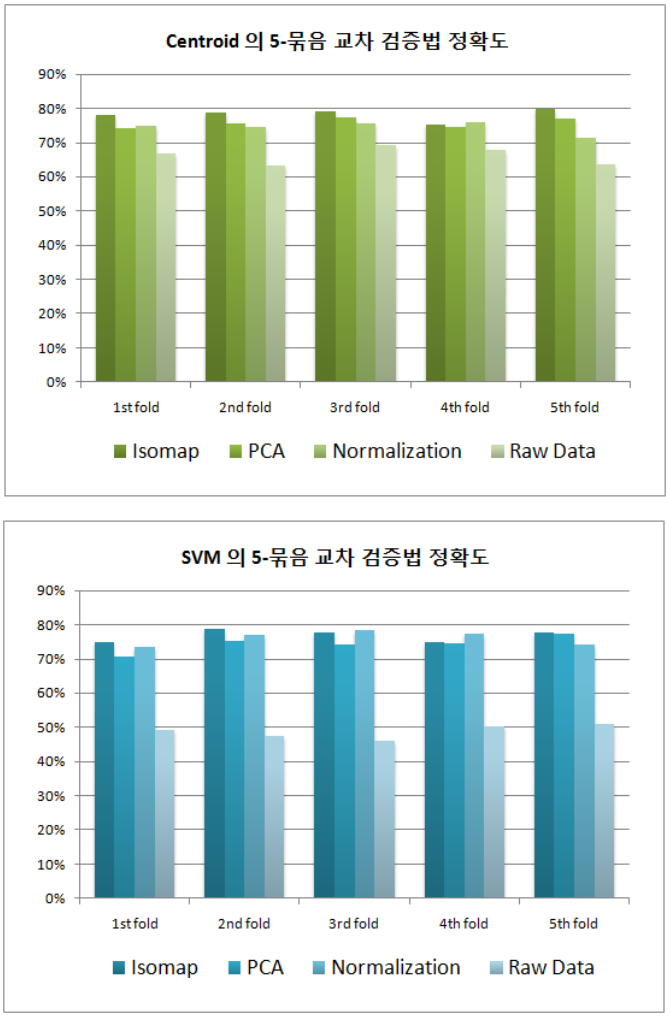


[그림 3] 실험 과정

다음으로, 각 Feature들의 숫자의 범위가 워낙 다양하다기 때문에, 전처리 과정의 향상을 위하여 정규화(Normalize)를 한 후, 제안한 Isomap 과 PCA 를 통하여 72차원의 데이터를 8차원으로 줄였다. 이 때, 두 방법 모두 고유값의 분포를 관찰하여 정보의 손실을 최소화 하면서 최대한 많이 줄일 수 있는 값을 선택하였다.

Isomap 은 72개의 고유값의 합에서 8차원으로 축소한 공간의 고유값의 합의 비율이 65.23%였으며, PCA는 85.28%였다. 이는 두 방법 모두 8차원으로 줄였을 때, 약 15~35%정도 정보가 손실되었음을 의미하지만 차원이 줄어들었을 때 결과가 좋은 것으로 보아, 원래 정보에서 15%의 필요 없는 정보를 떼어내고 새롭게 가공하는 것이 더욱 더 좋은 처리 과정임을 알 수 있다. 그리고 이렇게 압축한 데이터의 보정을 위해서 다시 한 번 정규화를 하였다.

또한 본 연구에서는 5-묶음 교차 검증법을 사용하였는데, 이는 한 개의 부분집합으로 학습을 하고, 나머지 4개의 부분집합으로 테스트를 진행하는 실험 방법이다. 그리고 SVM 과 Centroid 분류기의 정확도를 측정하여 아래 [그림 4]와 같이 나타내었다. Centroid 분류기로 실험한 결과, Isomap 이 총 5회의 실험에서 모두 우수한 결과를 보였고, 최대 79.29%의 정확도를 기록하였다. SVM 또한 최대 78.73%의 높은 정확도를 기록하였으며, 총 5회에 걸친 실험에서 4회 째 실험을 제외하고 모두 가장 좋은 성능을 내었다.



[그림 4] 분류기 정확도 측정 결과

마지막으로, 압축된 공간의 데이터들이 원 공간의 Feature 들과 얼마나 연관 되어있는지를 분석하였으며 Isomap 의 경우 첫 번째 주축이 [표 1] 의 WSa 와 상관계수 0.9183으로 가장 비슷하며, PCA 의 경우 두 번째 주축은 [표 1] 의 Tb 와 상관계수 0.9394로 가장 비슷한 것을 알 수 있었다. 두 방법에서 추출한 주축들은 공통적으로 Tb, WSa, WSb 등이 비슷한 것으로 분석 되었으며, 즉 오존 경보 예측에는 이 특징들이 주요한 요인으로 작용되는 것이라고 판단할 수 있다.

4. 결 론

본 논문은 기후 데이터의 특징 추출을 효과적으로 할 수 있는 방법을 제안하였으며, 주어진 데이터에서 오존 경보 예측에 어느 요인이 가장 우세한지를 분석하였다. Centroid 와 SVM을 통해 제안한 전처리방법이 얼마나 실효성이 있는지 테스트 해보았으며, 원 데이터에 비해 분류기의 정확도가 10~31% 정도 향상 된 것을 확인하였다. 이는 기후 데이터의 특징 추출이 효과적으로 이루어진 것이라고 할 수 있으며, 데이터의 압축 효과와 함께 오존 경보 예측에 더 적합한 새로운 데이터를 얻어낸 것을 의미한다.

5. 사 사

이 논문은 2010년 정보(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF 2010-0000631, NRF 2010-0000407)

6. 참고 문헌

[1] Kun Zhang, Wei Fan, Xiaojing Yuan, Ian Davidson, Xiangshang Li, "Forecasting Skewed Biased Stochastic Ozone Days: Analyses and Solutions," IEEE International Conference on Data Mining, pp.753-764, Dec 2006.

[2] 이태훈, 윤성로, "최소극대화 상관 분석을 이용한 향상된 기능의 오존 경보 예측기 구현, 대한전자공학회, 2009.

[3] http://archive.ics.uci.edu/ml/

[4] Joshua B. Tenenbaum, Vin de Silva, John C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction", Science, 2000.

[5] 최용석, 정광모, "SAS 를 활용한 다변량 분석 기법과 응용"