NCEP 일기도 데이터 클러스터링을 위한 특징 벡터 추출

이기범[0] 이성환 정창성 황치정
충 남 대 학 교 컴퓨터 과학과
kblee@cs.cnu.ac.kr

Feature vector extraction for NCEP weather data clustering

Ki-Bum Lee[0] Sung-Hwan Lee Chang-Sung Jung Chi-Jung Hwang
Dept. of Computer Science, Chungnam National University

요 약

방대한 양의 격자점 데이터 및 일기도 관련 데이터를 효율적으로 저장 및 검색 하기위해서는 데이터들의 유형을 찾아 서로 유형이 비슷한 데이터들을 하나의 클러스터로 연관지어 놓으면 효율적인 저장과 검색을 할 수 있다. 클러스터링에서 데이터들의 어떤 특징 벡터를 추출하는가가 클러스터링의 결과에 가장 중요한 영향을 끼친다.

본 논문에서는 격자점, 기압값 데이터로부터 일기도의 특징을 표현할 수 있는 벡터로 한반도 중심의 8방향에 대한 고/저기압의 분포와 동아시아 지역을 24영역으로 나누어 각 영역별로 고/저기압의 분포 정보를 특징벡터로 추출하여 클러스터링하였다. 클러스터링 알고리즘으로는 unsupervised mode인 SOM(Self Organizing Map) 기법을 사용하였다.

1. 서론

격자점 데이터 및 일기도 관련 데이터는 그 양이 방대하여 저장과 검색을 효율적으로 하기위해서는 데이터들의 유형을 찾아 서로 유형이 비슷한 데이터들을 하나의 클러스터로 연관지어 놓으면 효율적인 저장과 검색을 할 수 있다. 서로 유사한 데이터들을 클러스터링 기법을 통해 클러스터로 묶어 놓으면 검색시에 모든 데이터를 검색하지 않고, 찾고자 하는 target 데이터가 어느 클러스터에 해당하는 가를 먼저 알아본 후 , 해당 클러스터내에서만 유사한 일기도를 검색함으로써 성능 향상을 얻을 수 있다.

클러스터링에서 데이터들의 어떤 특징 벡터를 추출하는가가 클러스터링의 결과에 가장 중요한 영향을 끼친다. 데이터의 유형을 특징 지을 수 있는 특징 벡터를 선택해야만, 서로 유사한 데이터들이 서로 하나의 클러스터로 모아질 수 있다. 그러므로 특징벡터는 대상이 되는 데이터의 유형을 반영할 수 있는 것이어야만 한다.

본 논문에서는 격자점 기압값 데이터로부터 일기도의 특징을 표현할 수 있는 벡터로 한반도 중심의 8방향에 대한 고/저기압의 분포 정보와 동아시아 지역을 24영역으로 나누어 각 영역별 고/저기압의 분포 정보를 특징벡터로 하여 추출하였다.

본 논문은 충남대학교 BK 사업단의 지원을 받은 연구결과임

특징 벡터가 추출되면 이를 클러스터링 하기 위한 알고리즘이 필요한데 클러스터링 알고리즘은 크게 supervised mode와 unsupervised mode의 두 가지 방법으로 나뉜다. Supervised mode는 클러스터의 정보를 이미 알고 있는 상태에서 실험자의 개입에 의해 클러스터를 형성하는 방법이고, unsupervised mode는 사용자의 개입 없이 데이터들의 특징 벡터만으로 클러스터를 생성하는 방법이다. 격자점 기압값 데이터는 그 양이 방대하고 일기도의 패턴 역시 인위적으로 분류하기에는 많은 어려움이 따르므로 본 논문에서는 클러스터링 알고리즘으로 unsupervised mode인 SOM (Self Organizing Map) 기법을 사용하였다.

2. 클러스터링을 위한 특징 벡터 추출

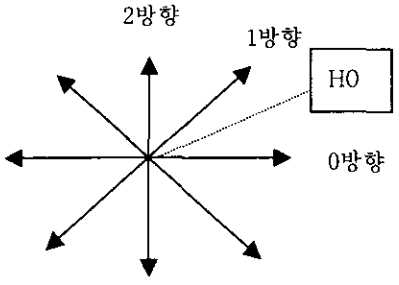
클러스터링을 하여 서로 유사한 패턴을 하나의 클러스터로 만들기 위해 각 데이터들의 특징을 표현할 수 있는 특징 벡터의 추출이 선행되어야만 한다. 특징 벡터의 추출이 적절치 못하면 클러스터링의 결과가 제대로 생성될 수 없기 때문에 반드시 데이터들의 유형을 나타낼 수 있는 특징 벡터를 생성해야 한다.

본 논문에서는 격자점 기압값의 유형을 나타낼 특징벡터로 한반도 중심의 8방향에 대한 고/저기압의 분포 정보와 동아시아 지역을 24영역으로 나누어 각 영역별 고/저기압의 분포 정보를 특징벡터로 하여 추출하였다.

2.1 한반도 중심의 8방향 기압분포 특징 벡터 추출

격자점 기압값 데이터의 유형을 분류하는데 있어 가장 먼저 고려되어야 하는 것은 한반도에 영향을 미치는 고/저기압의 분포이다. 그러므로 한반도 중심으로부터 인접한 고/저기압의 분포는 일기도의 유형에 해당한다고 볼 수 있으며 이러한 특징을 추출하기 위해서 본 논문에서는 한반도 중심으로부터 8방향에 대해 존재하는 고/저기압의 위치와 중심각과의 차이를 특징벡터로 하여 추출하였다.

[그림 1]에는 한반도 중심으로부터 8방향에 대하여 가장 인접한 고/저기압에 대해 그 거리와 기준각과의 차이각을 추출하고 이를 특징 벡터로 하는 그림이 나타나 있다.



특징 벡터 : $(H_0, H_{\theta_0}, \cdots, H_7, H_{\theta_7}, L_0, L_{\theta_0}, \cdots, L_7, L_{\theta_7})$

H_0 : 0방향에 위치한 고기압의 거리
H_{θ_0} : 0방향에 위치한 고기압의 기준각과의 차이각
L_0 : 0방향에 위치한 저기압의 거리
L_{θ_0} : 0방향에 위치한 저기압의 기준각과의 차이각

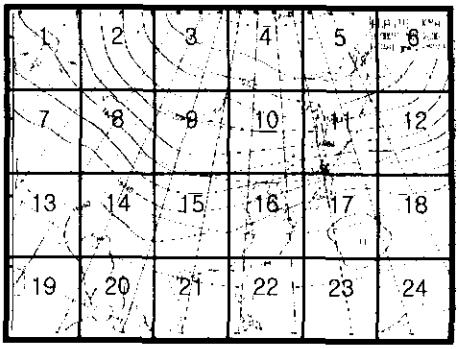
[그림 1] 8방향에 대한 고/저기압 분포를 나타내는 특징 벡터

2.2 영역별 고/저기압 분포를 이용한 특징 벡터 추출

한반도 중심의 8방향에 대한 고/저기압의 분포를 특징벡터로 추출하여 클러스터링을 하는 경우 한반도에 영향을 주는 고/저기압을 반영하는 장점이 있지만 동아시아 전반에 걸친 전체적인 고/저기압의 분포에 대한 정보는 반영하지 못할 수가 있다. 그렇기 때문에 본 논문에서는 격자점 기압값 데이터로부터 한반도 중심의 8방향에 대한 고/저기압의 분포정보를 특징벡터로 추출하는 방법과 더불어 동아시아 전체 영역에 해당하는 고/저기압의 분포를 나타내는 특징벡터를 추출하는 알고리즘을 사용하였다.

동아시아 전체 영역을 24영역으로 균등하게 나누고 각 영역내에 고/저기압의 분포를 수치로 표현하고 이를 특징벡터로 사용하였다.

[그림 2]와 같이 동아시아 지역을 24영역으로 나누어 각 영역내에 존재하는 고/저기압의 정보를 다음과 같이 표현하여 특징벡터로 하였다.



[그림 2] 동아시아 지역에 대한 24영역

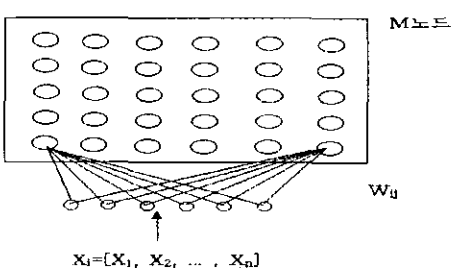
특징벡터 : $(R_1, \cdots, R_{24})$

where Ri = 0 고저기압이 없는 경우
1 고기압만 있는 경우
2 저기압만 있는 경우
3 고/저기압 둘 다 있는 경우

3. SOM을 이용한 클러스터링 기법

격자 데이터는 하루에 6시간 간격으로 4번씩 생성되고, 일년치 데이터는 365*4=1460개의 파일이 생성된다. 이들 격자 데이터는 그 양이 방대하여 유사 일기도 검색을 하기 위해 매번 모든 데이터를 대상으로 검색을 하는 것은 매우 비효율적이다. 대부분의 검색 알고리즘에서 성능 향상을 위해 대상이 되는 데이터를 자동적으로 몇 개의 클래스로 나누고, 찾고자 하는 원본의 클래스를 파악한 후에 해당 클래스의 데이터들만 검색하는 방법을 사용하고 있다. 이에 본 논문에서도 많은 양의 일기도 데이터를 보다 빠르게 검색하기 위해 격자 데이터로부터 몇몇의 유형을 분류하여 이를 클러스터링 하기위한 방법으로 SOM(Self Organizing Map)기법을 사용하였다.

SOM은 입력 정보의 통계적 특성을 추출하고 그들간의 고차원 관련성을 2차원 평면으로 투영하는 알고리즘이다. SOM은 단층 네트워크 구조를 가지고 feed-forward 방식의 노드 연결 구조로 구성되어 있으며 무교사 학습 규칙을 이용하는 신경회로망 모델로서 [그림 3]과 같이 입력 층과 경쟁 층의 간단한 2차원적인 구조로 이루어져 있으며 각 노드는 모든 입력 차원과 연결 강도 W_{ij}로 연결되어 있다.



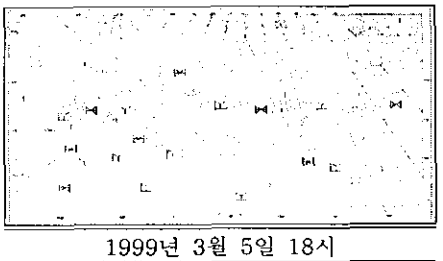
[그림 3] SOM의 구조

입력 층에서의 노드의 수는 입력 데이터 패턴의 차원 수와 일치하며, 경쟁 층의 노드들은 고밀도 연결로서 방대하게 상호 연결되어 있다. 학습 과정은 적절하게 결정된 초기 연결 강도로부터 경쟁 학습 방법을 통하여 연결 강도를 개선해 나가는 방법으로 수행되므로 SOM의 학습 규칙은 무교사 학습이다.

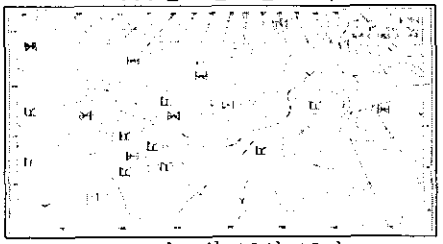
SOM은 N차원의 벡터를 2차원으로 매핑 시켜준다. 이에 우리는 격자 데이터로부터 특징 벡터를 추출하고 이를 이용하여 SOM을 구성하여 클러스터링을 수행한다. 격자 데이터로부터 추출할 특징 벡터는 한반도 중심에서 8방향으로 위치하는 고/저기압의 정보이다. 고/저기압에 대해서 8방향으로의 거리와 중심으로부터의 차이각의 쌍을 하나의 요소로 하는 32차원 벡터를 추출하고 이를 이용하여 SOM을 구성한다. 이는 한반도 중심의 고/저기압의 분포를 이용한 클러스터링으로 유사 일기도로서의 의미를 지닌다. 또한, 동아시아 지역을 24영역으로 나누어 각 영역내의 고/저기압의 분포를 특징 벡터로 추출하고 이를 클러스터링하게 된다.

4. 클러스터링 결과

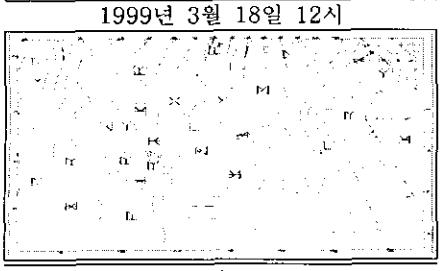
추출한 특징 벡터로 SOM을 구성하여 30 - 80개의 클러스터로 클러스터링을 시행하였다. 작게는 30여개의 클러스터로 모아져서 전체 데이터에 대한 검색을 행하는 것보다 평균적으로 30여배의 검색 성능향상을 얻을 수 있으며, 검색의 정확도도 전체 데이터에 대한 검색을 행한 경우와 거의 유사했다. [그림 4]는 30개의 클러스터로 클러스터링 된 결과 클러스터 중 하나의 일부분이며, 격자점 데이터는 1999년 1년치 365*4=1460개를 이용하였다.



1999년 3월 5일 18시



1999년 3월 18일 12시



1999년 4월 2일 12시

5. 결론 및 향후 연구

본 논문은 방대한 양의 격자점 데이터 및 일기도 관련 데이터에서 특징을 표현할 수 있는 벡터로 한반도 중심의 8방향에 대한 고/저기압의 분포 정보와 동아시아 지역을 24영역으로 나누어 각 영역별 고/저기압의 분포 정보를 특징벡터로 하여 검색에서 평균 검색 성능을 향상하였다. 향후에는 클러스터의 수를 줄이면 클러스터링이 거의 이루어지지 않아 속도 향상을 얻을 수 없고, 클러스터의 수를 늘이면 유형이 좀더 다양하게 분류되므로 정확한 클러스터를 얻을 수 있어 속도는 향상되지만 유사한 패턴이 서로 다른 클러스터에 포함될 수 있으므로 적절한 수의 클러스터를 설정하는 방법론이 필요하다.

6. 참고 문헌

[1], Christopher S. Bretherton, Catherine Smith, John M. Wallace, An Intercomparison of Methods for Finding Coupled Patterns in Climate Data, Jurnal of Climate, volume 5, JUNE, p541-560,1992

[2], Eduardo Zorita, Hans Von Storch, The Analog Method as a Simple Statistical Downscaling Technique : Comparison with More Complicated Methods, Journal of Climate, volume 12, p2474-2489, August, 1999

[3], A. Jain, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988

[4], Legras, B., T. Desponts, and B. Piguet, 1988, Cluster analysis and weather regimes, Proc. Of the ECMWF Seminar on the Nature and Prediction of Extra-tropical Weather Systems, Reading, U.K., European Centre for Medium-Range Weather Forecasts, II, 123-149

[5], H. M. Van Del Dool, Searching for analogues, how long must we wait?, TELLUS, 46A, 314-324, 1994