SOM을 적용한 선택적 샘플링에 관한 연구

김만선, 양형정, 김정식, 김선희
전남대학교 전자컴퓨터공학과
E-mail : kms9688004@nate.com, hjyang@chonnam.ac.kr,
wind1105@lycos.co.kr, wkdal749@hanmail.net

A Study on Selective Sampling using SOM

Man-Sun Kim, Hyung-Jeong Yang, Jeong-Sik Kim, Sun-Hee Kim
Dept. of Electronics and Computer Engineering,
Chon-Nam National University

요 약

데이타 마이닝을 위하여 수집된 대용량의 데이타를 여과 없이 기계학습에 적용하는 것은 많은 시간과 비용이 요구될 뿐만 아니라 저장 공간면에서도 비효율적이다. 선별적 샘플링은 이러한 상황에서 매우 효율적으로 적용할 수 있도록 원본 데이타의 특성을 가능한 반영하여 새로운 훈련 데이타를 생성하는 방법이다. 본 연구에서는 신경망의 하나인 SOM을 적용한 선별적 샘플링을 수행하는데 있어서 여러 가지 선택 문제를 효과적으로 해결하기 위한 실험을 수행한다. 실험 결과로는 두 가지 결과를 얻었다. 1) 충분한 맵 사이즈를 선택해야 학습 데이타의 함축적인 특성을 잘 반영한다, 2) 선택적 샘플링을 위한 유닛선택 방법에서는 의미없는 유닛을 제거함으로서 분류 성능향상을 얻을 수 있다.

키워드 : Selective Sampling, SOM, 양자화 에러, Clustering, data mining

1. 서론

데이타마이닝이란 대량의 데이타로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 여기서 정보는 묵시적이고 잘 알려져 있지 않지만 잠재적으로 활용가치가 있는 정보를 말한다. 데이타 마이닝을 위하여 수집한 데이타의 용량은 수십 기가바이트 이상이 될 수 있다. 이들 데이타를 모두 활용하여 기계학습을 적용하는 것은 저장 공간이나 수행시간 측면에서 매우 비효율적이다.

선별적 샘플링은 이러한 상황에서 기계학습을 보다 효율적으로 적용할 수 있도록 원본 데이타의 특성을 가능한 반영하는 부분집합을 생성하는 기법이다[1-4]. 최초 학습 예제를 임의로 선정하였을 때 선정된 데이타가 학습에 효과적이지 못한 경우 종종 발생하는데, 이것은 분류 정확도가 떨어지는 결과를 낳을 수 있다.

SOM(Self Organizing Map)은 무감독 학습 방법의 일종으로서 스스로 n 차원의 입력 데이타들을 군집화(클러스터링)하여 2차원으로 사상시켜 준다. SOM을 사용할 때 다른 신경망에서는 일반적으로 필요하지 않는 파라미터와 관련된 몇 가지 일을 수행해야 한다.

1) 층(layer)내의 뉴런의 연결강도 벡터 초기화
2) 연결강도 벡터와 입력벡터의 정규화
3) 2차원 맵 사이즈(격자의 크기)를 결정

이 중에서도 2차원 맵 사이즈 결정은 매우 중요한 이슈 중의 하나이다. 이유는 설정된 지도의 크기가 너무 작으면 자료에 존재하는 비선형성의 관계를 표현하기 어렵고, 너무 크면 분석에 많은 시간이 소요되고 관찰치가 하나도 포함되지 않은 군집이 발생하여 분석결과에 대해 그릇된 해석을 할 수 있기 때문이다.

본 연구에서는 신경망을 적용한 선별적 샘플링을 수행하는데 있어서 파라미터 결정에 대한 문제를 효과적으로 해결하기 위한 실험을 수행한다. 이를 위하여 우선적으로 맵 사이즈 선정 방법에 따른 영향과 두 가지 선별적 기준에 의한 샘플링의 영향에 대하여 자세한 실험을 통하여 알아본다.

본 논문의 구성은 다음과 같다. 2장에서 SOM에 대하여 알아본다. 3장에서 맵 사이즈의 선정 방법에 따른 영향과 두 가지 선별적 기준에 의한 샘플링의 영향에 대하여 알아본다. 4장에서는 비교 실험 결과와 실험 내용을 분석하고, 5장에서 결론 및 향후 연구 과제를 논의한다.

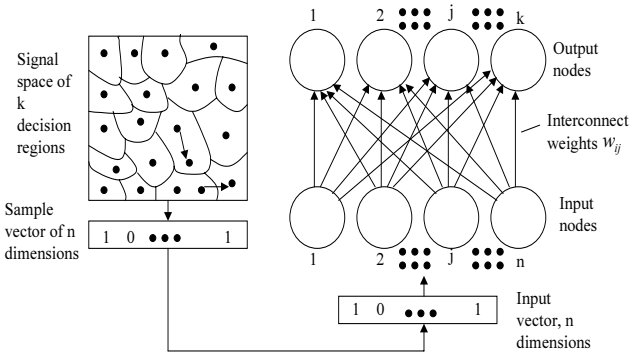
2. SOM(Self Organizing Map)

SOM[7]은 신경망 군집화 기법중의 하나로서 비슷한 속성을 갖는 군를 그룹핑하고, 동시에 다른 군에 속한 데이타와의 차이를 최대화 할 수 있도록 학습하는데, 주로 고객 세분화, 전체 데이타에 대한 이해, 특이한 세분 그룹의 발견(fraud detection) 등에 사용된다. SOM은 신경망을 이용하여 구현한 것으로 그 내부는 아래의 그림과 같다[8,9].

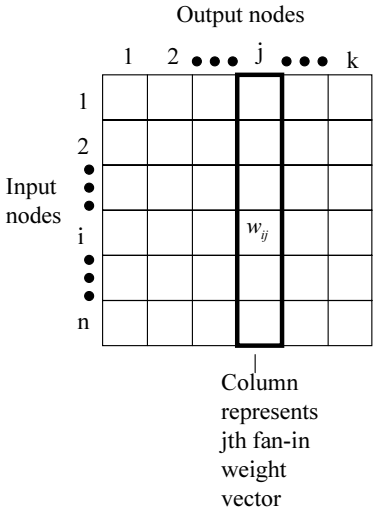
[그림 1]은 2-layer 신경망으로 n 차원의 입력 데이타

를 표현하는 n 개의 입력 노드들과 k 개의 분류 영역(decision region)을 표현하기 위한 k 개의 출력 노드로 구성되어 있다. 모든 입력 노드들은 모든 출력 노드들과 연결되어 있고 연결 가중치(weight)를 가진다[5].

[그림 2]는 입력 노드 i 와 출력 노드 j 를 연결하는 가중치 w_{ij} 들의 행렬을 보여준다. 행렬에서 i 번째 행은 입력 노드 i 로부터 각 출력 노드로 나가는 연결 가중치를 나타내며 j 번째 열은 각 입력 노드로부터 출력 노드 j 로 들어오는 연결 가중치를 나타낸다. 여기서 j 번째 열로 나타내어지는 벡터는 j 번째 Fan-in 가중치 벡터 라고 하며 입력 벡터와의 거리(유클리드 거리) 계산에 쓰인다.



[그림 1] SOM의 구조



[그림 2] 연결 가중치 행렬

초기 상태에서는 연결 가중치들을 임의로 할당 후 입력 벡터와의 유사성을 측정한다. 유사성 측정은 여러 가지 방법으로 할 수 있는 데 유클리드 거리를 이용하는 것도 하나의 방법이다. 입력 벡터와 k 개의 fan-in weight vector 사이의 유클리드 거리를 구하여 입력 벡터와 가장 유사한 (유클리드 거리가 가장 작은) j 번째 Fan-in 가중치 벡터를 찾으면 그 입력 벡터에 대해서 j 번째 출력 노드가 승자가 된다. 여기에서 승자 노드는 하나의 입력 데이터를 SOM의 맵 상에 가장 대표할 수 있는 클러스터의 위치로 표현한다. 이렇게 승자를 선택하면 승자의 Fan-in 가중치 벡터는 갱신되는 데 식으로 나타내면 아래와 같다 [5].

$$w^j(t+1) = w^j(t) + \alpha(t)[x(t) - w^j(t)] - (1)$$

$$\text{여기서}, \alpha(t) = 0.1(1 - t/10^4)$$

식(1)에서 j 번째 출력 노드가 승자가 되었으면 그 노드의 연결 가중치 벡터는 입력 벡터 쪽으로 약간 이동한다. 즉, Fan-in 가중치 벡터를 입력 데이타 벡터와 비슷하게 만들어 가는 것이다.

3. SOM의 파라미터와 관련된 두 가지 영향

본 절에서는 SOM의 파라미터와 관련된 두 가지 영향에 대하여 알아본다. 첫 번째로는 맵 사이즈의 선정 방법에 따른 분류율의 결과를 알아보고, 두 번째로는 선별적 기준 선정에 따른 샘플링의 분류 결과를 알아본다.

3. 1 맵 사이즈의 선정 방법에 따른 영향

군집화에서 양자화 에러의 값은 훈련에서 사용된 데이타와 비슷한 데이타가 입력으로 들어오면 비교적 작은 값을 가지게 될 것이고 임계치 이상이 되면 비정상 데이타로 해석할 수 있다. 그러므로 맵 사이즈는 군집화의 결과를 해석하는데 하나의 지표로 사용될 수 있다. 본 절에서는 맵 사이즈와 양자화 에러와의 관련성을 알아보기 위하여 다양한 맵 사이즈를 고려하였다.

① 유닛 의 개수를 데이터 샘플수의 제곱근과 상수를 곱하여 올림 함수로([ceil(5* $\sqrt{\text{데이터 샘플의수}}$)]) 로 결정한다. 결과 값은 속성(feature) 개수의 배수가 되도록 재선정한다. 그 결과로 격자의 행의 개수를 산출하게 되는데, 예를 들어 데이타의 수가 150이고 속성의 수가 5이면, 유닛 개수는 올림 함수를 적용한 62개가 산출되지만 속성의 수(feature=5)와 견주어 충분한 배수가 되도록 유닛의 개수가 재선정되어 65개가 된다. 최종 맵 사이즈는 [유닛 개수 속성 개수 속성 개수]인 [13 5]로 결정된다. ② 맵 사이즈를 임의로 결정한다. 본 연구에서는 임의로 [26 10]으로 설정하였는데, 첫 번째 방법에서 얻어진 맵 사이즈의 가로, 세로를 두 배로 선정한다. ③ 매우 작게, 매우 크게 [5 3], [40 20]으로 설정하였다. 생성되는 유닛의 수는 15, 800개가 된다. 150개의 데이타를 갖는 실험에서는 작은 맵은 하나의 유닛에 여러 개의 데이타가 할당될 수 있으며, 큰 맵은 하나의 클러스터에 관측 데이타가 하나도 포함되지 않는 경우가 발생할 수 있다.

충분히 훈련이 된 SOM은 임의의 입력 패턴에 대해 분류작업을 수행할 수 있다. 임의의 입력 데이타에 대해서 그 데이타와 가장 유사한 값을 가지는 노드(클래스)로 분류를 하게 된다. 이 때 훈련 단계에서 사용된 데이타와 비슷한 데이타가 입력으로 들어왔다면 양자화 에러(입력 데이타와 노드의 대표값 사이의 거리)가 비교적 작은 값

을 가지게 될 것이고 기존의 데이타들과는 다른 새로운 패턴의 데이타가 입력으로 들어오면 양자화 에러값이 매우 큰 값을 가지게 될 것이다.

<표 1>맵 사이즈에 따른 유닛 수와 양자화 에러

맵 사이즈	유닛 수	양자화 에러
[5 3]	15	0.759
[13 5]	65	0.417
[26 10]	260	0.221
[40 20]	800	0.076

실험 결과는 위의 <표 1>과 같다. 유닛의 수는 맵 사이즈의 행과 열을 곱한 값이 된다. 전체 데이타의 개수가 150개라면 유닛의 개수 중 어느 하나의 유닛에 0~n개의 데이타가 할당된다. 유닛의 수가 데이타의 수보다 많은 경우는 각기 다른 유닛에 흩어져 분포하게 되므로 함축적으로 의미 있는 유닛을 발견할 수가 없게 된다.

양자화 에러는 유닛의 수가 증가할수록 작은 값을 갖는 양상을 보였으나, 오히려 효율적인 맵 사이즈를 결정하는 문제와는 연관성이 없음을 보였다.

3. 2 두 가지 선별적 기준에 따른 샘플링의 영향

본 절에서는 두 가지 선별적 기준에 의하여 샘플링을 수행하고 그 결과를 분류 알고리즘에 이용하여 훈련하고 검증을 수행한다. 첫 번째 선별 기준은 하나의 유닛에 할당된 데이타의 개수를 count라 정하고 count가 5 이상인 유닛만 추출하여 새로운 훈련 데이타로 선정하는 것이다. 두 번째 선별 기준은 count가 3 이하인 유닛만 제거하여 새로운 훈련 데이타로 선정하는 것이다. 3.1의 실험 결과에서 맵 사이즈가 [5 3], [13 5]인 경우만 선별기준에 적합한 데이타를 샘플링 할 수 있었다.

본 연구에서는 Support Vector Machine(SVM) 분류기를 사용하였다. SVM은 Vapnik에 의해 제안된 구조적 위험 최소화를 목적으로 하는 패턴 인식 알고리즘으로써 그 성능과 안정성이 다른 알고리즘에 비하여 우수함이 입증되었고 여러 응용 분야에서도 활발히 사용되고 있다.

<표 2> 선별적 기준에 따른 SVM을 적용한 분류 결과

맵 사이즈	[13 5]	[5 3]
총 유닛 수	65	15
유닛에 5개 이상 할당된 경우만 추출	에러율 0.180 데이타 수 43개 **(a1 방법)**	에러율 0.033 데이타 수 139개 **(b1 방법)**
유닛에 3개 이하 할당된 경우만 제거	에러율 0.033 데이타 수 91개 **(a2 방법)**	에러율 0.026 데이타 수 147개 **(b2 방법)**

・a1방법과 a2방법의 비교, b1방법과 b2방법의 비교
두 가지 선별 기준에 의한 방법에서는 전자의 방법보다 후자의 방법이 더 우수함을 보인다. 소수의 데이타가 할당된 유닛을 제거하는 방법에서 에러율이 감소하였다.

・a2방법과 b1방법의 비교
두 실험에서 에러율은 0.033으로 같지만, 사용된 데이타의 개수에서 현격히 차이가 있음을 보였다. a1방법의 데이타의 개수는 전체 데이타의 60.6%, b1방법은 92.6%를 샘플링하였다. 이 결과는 유닛의 밀도가 적은 것을 제거하는 방법이 더 우수함을 입증하였다.

・맵 사이즈와 양자화 에러율의 비교
양자화 에러율을 기준으로 비교한다면 맵 사이즈가 [5 3]의 크기를 갖는 실험이 더 우수한 결과를 보였다. 이 결과는 3.1에서 보인 양자화 에러값과는 반비례적인 결과를 보였다.

4. 비교 실험 결과

실험 환경은 펜티엄4 3.39Ghz, 1GB RAM, 윈도우 XP 환경에서 MATLAB 7.01을 사용하였다.

실험 데이타로는 Iris 데이타[6]를 사용하였다. 피셔의 붗꽃 자료는 주어진 바와 같이 세 가지 품종 setosa versicolor, virginica 의 붗꽃 으로부터 각각 50개, 총 150개의 객체들을 추출한 다음 아래 4개의 변수(sepal length 꽃받침조각의 길이, sepal width 꽃받침조각의 폭, petal length 꽃잎의 길이, petal width 꽃잎의 폭)를 센티미터 단위로 측정하였다.

비교 실험을 위하여 기계학습의 대표적인 7가지(NN, 배깅, SVM, ADABOOST, 의사결정트리, 나이브 베이지안, kNN) 분류알고리즘을 사용하였다.

<표 3>의 결과에서는 NN, **b2** 방법의 분류 결과가 가장 우수했다. 큰 차이는 없으나 나이브 베이지안을 기본 분류기로 사용한 배깅의 결과와 **a2**, **b1** 방법의 분류 결과도 우수함을 보였다. SVM을 적용한 실험으로는 3가지 커널방법을 사용하였다. POLY 커널을 적용한 SVM은 매우 큰 에러율을 보였다.

<표 3> 분류 결과

SVM 적용 알고리즘	에러율
a1 방법	0.180
a2 방법	0.033
b1 방법	0.033
b2 방법	0.026
SVM(커널:LINEAR)	0.040
SVM(커널:RBF)	0.047
SVM(커널:POLY)	0.467

분류 알고리즘	에러율
NN(Neural Networks)	0.027
배깅+Naive	0.033
ADABOOST+Naive	0.047
의사결정트리	0.047
나이브 베이지안	0.053
kNN(k=5)	0.060

<표 2>의 4가지 방법은 원본 데이타를 모두 사용하지 않고 일부분만 선별적으로 샘플링하여 재구성된 데이타를 사용하였다. 그럼에도 불구하고 <표 3>의 결과에서는 뛰어난 분류 성능을 보였다. 선별적 샘플링은 기계학습을 보다 효율적으로 적용할 수 있도록 원본 데이타의 특성을 가능한 반영하는 부분집합을 생성하였다. 최초 학습예제를 임의로 선정하였을 때 선정된 데이타가 학습에 효과적이지 못한 경우 종종 발생하여 분류 정확도가 떨어지는데 이러한 문제를 해결할 수 있는 방법임을 입증하였다. 또한 유닛의 밀도가 적은 것을 제거하는 방법이 분류 성능에 더 우수함을 실험적으로 보였다.

5. 결론 및 향후 연구 과제

본 연구에서는 신경망의 하나인 SOM을 적용한 선별적 샘플링을 수행하는데 있어서 여러 가지 선택 문제를 효과적으로 해결하기 위한 실험을 수행하였다. 실험 결과로는 다음과 같이 두 가지 결과를 얻었다.

1) 맵 사이즈를 선택할 경우에는 충분한 맵 사이즈를 고려해야 학습 데이타의 함축적인 특성을 잘 반영한다. 군집화를 위하여 SOM을 사용했으나 데이타의 개수보다 유닛의 개수가 많으면 군집화의 장점을 얻을 수 없었다. 또한 양자화 에러는 유닛의 수가 증가할수록 작은 값을 갖는 양상을 보였으나, 오히려 효율적인 맵 사이즈를 결정하는 문제와는 연관성이 없음을 보였다.

2) 선택적 샘플링을 위한 유닛선택 방법에서는 의미없는 유닛을 제거하는 접근 방법이 분류 성능향상을 얻을 수 있다. 선택적 샘플링을 수행한 결과로 얻은 훈련 데이타를 SOM을 제외한 기계학습의 대표적인 7가지 방법에 적용한 결과 만족할만한 우수한 성능을 보였다.

향후 연구 과제로는 실제 대용량 데이타를 적용하여 성능을 비교해야 할 것이며, 분류 성능 향상을 위하여 군집화를 선별적 샘플링 방법에 적용하는 연구를 찾아보고 성능을 비교하는 연구가 필요하다.

참고문헌

[1] Jennifer N.M. Ballard, Gilles A. Lajoie and Ken K.-C. Yeung, “Selective sampling of multiply phosphorylated peptides by capillary electrophoresis for electrospray ionization mass spectrometry analysis,” Journal of Chromatography A, Volume 1156, Issues 1-2, 13 July 2007, pp. 101-110.

[2] Ion Muslea, Steven Minton, and Craig A. Knoblock, “Selective Sampling With Redundant Views,” American Association for Articial Intelligence, 2000.

[3] Yoav Freund, Florham Park, H. Sebastian Seung, "Selective Sampling Using the Query by Committee Algorithm," Machine Learning archive Volume 28 , Issue 2-3 Aug./Sept. 1997.

[4] Piotr Juszczak, Robert P.W. Duin, "Selective sampling methods in one-class classification problems," 13th International Conference on Artificial Neural Networks", pp. 140-148, 2003.

[5] 김인영, 장병탁, “SOM을이용한유닉스시스템사용자의 비정상행위탐지,” 한국정보과학회 학술발표논문집 한국정보과학회 춘계학술발표논문집 제28권 제1호(A), 2001.

[6] http://mlearn.ics.uci.edu/databases/

[7] T. Kohonen, Self-organizing maps, Springer Verlag, Berlin, Germany (1995).

[8] Michael Chester, Neural Networks : A Tutorial, Prentice Hall. pp.42-49, 1993.

[9] Simon Haykin, Neural Networks : A Comprehensive Foundation, Prentice Hall. pp.43-483,1993.