PCA와 SOM을 이용한 자동 군집화 에이전트

Automatic Clustering Agent using PCA and SOM

박정은, 김병진, 오경환
서강대학교 컴퓨터학과

Jung-Eun Park, Byung-Jin Kim, Kyung-Whan Oh
Dept. of Computer Science
Sogang University
E-mail : favemint@ailab.sogang.ac.kr

요 약

인터넷의 정보 홍수 속에서 원하는 정보를 정확하게 제시간에 얻기란 쉬운 일이 아니며, 따라서 이러한 작업을 대신해주는 에이전트의 역할이 점점 커지고 있다. 대부분의 이벤트들이 실시간에 발생되고 처리되어야 하는 인터넷 환경에서는 분석가가 군집화의 방법과 결과 해석에 지속적으로 관여하기 어렵기 때문에 이러한 분석가의 업무를 대신하는 지능화된 에이전트가 필요하게 된다.

본 논문에서는 특히 자율학습 군집화에 대한 자동화된 시스템으로서 자동 군집화 에이전트를 제안하며 이 시스템은 군집화 수행 에이전트와 군집화 성능 평가 에이전트로 이루어져 있다. 두 개의 에이전트가 서로 정보를 교환하면서 자동적으로 최적의 군집화를 수행한다. 군집화 과정에서는 데이터를 분석하는 분석가가 군집화의 방법과 결과 해석에 실시간으로 관여하기 어렵기 때문에 이러한 작업을 담당하는 지능화된 에이전트가 자동화된 군집화를 담당하면 효과적인 군집화 전략이 될 수 있다. 또한 UCI Machine Repository의 IRIS 데이터와 Microsoft Web Log Data를 이용한 실험을 통해 제안 시스템의 성능 평가를 수행하였다.

1. 서론

최근 인터넷을 기반으로 하는 전자상거래의 급속한 발전과 더불어, 인터넷 환경의 e-Business 시스템에서는 에이전트 기반 모델에 대한 필요성이 날로 증가하고 있다. 즉, 인터넷의 정보 홍수 속에서 원하는 정보를 정확하게 제시간에 얻기란 쉬운 일이 아니며, 따라서 이러한 작업을 대신해주는 에이전트의 역할이 점점 커지고 있다.

인터넷에서의 군집화는 사용자, 웹 문서 등 인터넷 거래 행위의 주체 및 개체들을 서로 유사한 것들끼리 묶어 주는 역할을 담당한다. 대부분의 이벤트들이 실시간에 발생되고 처리되어야 하는 인터넷 환경에서는 분석가가 군집화의 방법과 결과 해석에 계속적으로 관여하기 어렵기 때문에 이러한 분석가의 업무의 상당 부분을 담당하는 지능화된 에이전트가 필요하며 이러한 에이전트가 자동화된 군집화를 담당하게 된다.

본 논문에서 제안하는 자동 군집화 에이전트는 군집화 수행 에이전트와 군집화 성능 평가 에이전트로 구성된 멀티 에이전트로서 두 개의 에이전트가 서로 정보를 교환하면서 최적의 군집화를 수행하는 멀티 에이전트 시스템이다.

제안한 시스템에 대한 성능 평가를 위한 실험은 UCI Machine Repository의 IRIS Data와 Microsoft Web Log Data를 이용하였다.

이 논문의 2절에서는 제안 시스템에 대한 관련 연구를 알아보고, 3절에서는 제안 시스템에 대한 통합설계와 프로세스를 설명하고 있다. 4절에서는 UCI Machine Repository 데이터를 이용한 실험을 통해 제안한 시스템의 성능 평가를 수행하였고 마지막으로 5절에서 결론 및 향후 연구과제에 대하여 논의하였다.

2. 자동 군집화 에이전트(AuCA)

자동 군집화 에이전트(Automatic Clustering Agent : AuCA)는 군집화 수행 에이전트와 군집화 성능 평가 에이전트로 구성된 다중 에이전트로서 두 개의 에이전트가 서로 정보를 교환하면서 최적의 군집화를 수행한다.

2.1 군집화 수행 에이전트

자동적으로 군집화를 수행하는 군집화 수행 에이전트에서는 군집화 알고리즘으로서, 자기 조직화 지도(Self-Organizing Map ; SOM)를 이용한다. 이 기법을 사용한 이유는 자기 조직화 지도가 매우 빠른 군집화를 가능하게 하여 실시간 웹 데이터의 군집화에 적합하기 때문이다. 자기 조직화 지도의 단점은 최적의 군집 수를 결정해주는 형상 지도(feature map)의 차원을 주관적으로 결정해야 한다는 것과 형상 지도의 차원이 크면 군집수가 그에 따라 증가하며, 반면에 차원이 작으면 군집수가 감소하게 된다는 점이다. 이러한 자기 조직화 지도의 문제점을 해결하기 위해 본 논문에서는 주성분 분석(Principle Component Analysis)을 이용해 최적의 형상 지도 차원을 객관적으로 결정하였으며, 이를 가지고 초기 군집수를 결정하였다.[1][2][4][5]

2.2 군집화 성능 평가 에이전트

군집화 성능 평가 에이전트는 군집화 수행 에이전트의 결과에 대한 성능 평가를 담당한다. 만약 성능이 좋지 않은 결과를 얻게 되면 군집화 수행 에이전트에게 더 나은 군집 분석을 요구한다. 군집 결과의 성능 평가는 본 논문에서 제안하는 Variance Criterion(VC)을 통해 수행한다. 이 기준은 군집화에 사용된 군집 변수 중 연속형 변수에 대해서 분산(variance)을 적용하여 이 값들이 작은 것을 좋은 결과로 결정하였다. 이러한 척도를 적용한 근거는 군집화의 개념이 같은 군집내의 개체들은 서로 동질성이 크고 서로 다른 군집간의 개체들간에는 이질성이 크도록 하기 때문이다. 또한 동일 데이터에 대한 군집 수가 증가하면 군집 결과의 각 군집에 대한 동질성은 증가하지만 군집의 수가 너무 많아지면 분석의 의미가 없어지므로 군집수의 증가는 결과의 성능 평가에서 페널티를 포함시켜 작용하게 하였다. 즉, 본 논문에서 제안한 VC는 연속형 군집 변수의 분산과 군집 수 증가에 따른 페널티로 이루어져 있으며 식 (2-1)과 같이 정의하였다.

$$VC_M = \sum_{i=1}^{M} v_i / M + p \times M \qquad (2\text{-}1)$$

여기서 M 은 군집의 수이다. 그리고 v_i 는 i 번째 군집의 평균 분산이 된다. 두 번째 항에 있는 $p \times M$ 군집수에 따른 패널티이다. 이 패널티에서 p 는 군집수를 알고 있는 데이터로부터 휴리스틱하게 결정을 하였다. 이 결정 방안은 3.1절의 Iris Data를 이용하여 논하였다. 결론적으로 식 (2-1)의 값이 작을수록 군집 결과의 성능이 우수하게 된다.

2.3 AuCA 시스템

군집화 수행 에이전트와 군집화 성능 평가 에이전트로 구성되어 있는 자동 군집화 에이전트(AuCA)는 그림 1과 같이 서로 대화식의 구조를 띠고 있다.

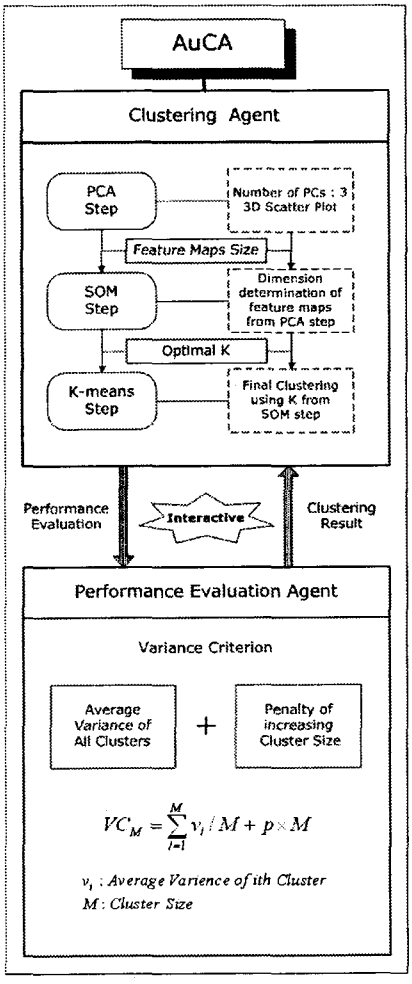

그림 1 AuCA의 시스템 구조

우선 학습 데이터에 대하여 주성분 분석을 통하여 3개의 주성분을 이용하여 전체 데이터에 대한 산점도를 그린다. 보유 주성분의 개수를 3개로 하여 3차원 산점도를 그린 이유는 시각적으로 관찰할 수 있는 산점도가 최대 3차원이기 때문이다. 이 산점도를 이용하여 전반적인 데이터의 군

집 구조를 파악한다. 이 산점도를 통해 자기 조직화 형상 지도의 차원을 결정한다. 본 논문에서는 형상지도의 차원은 산점도에 의한 (군집수*군집수)로 결정하였다. 이는 여러 차례의 실험을 통하여 휴리스틱하게 결정되었다. 주성분 분석을 통하여 형상지도의 차원이 결정된 자기 조직화 지도(SOM)를 이용하여 최적 군집수 결정을 위한 자율 학습(unsupervised learning)을 수행한다. 이 결과를 이용하여 최적의 군집수를 결정하고 이 값을 K-평균 군집 분석의 초기 군집수로 결정하여 최종적인 군집화를 수행한다. 이 과정까지를 군집화 수행 에이전트가 담당한다. 다음은 군집화 수행 에이전트가 VC 판단기준을 이용하여 군집화 성능 평가를 수행한다.[2][4]

3. 실험 및 평가

3.1 Iris 데이터를 이용한 실험 및 결과

Iris Plants Database는 150개의 학습 데이터와 4개의 입력변수로 이루어져 있으며, 이 변수들이 붓꽃의 종류를 결정해준다.[3]

우선 4개의 입력 변수를 가지고 있는 Iris 데이터에 대한 주성분 분석을 수행한 결과가 표 1에 나타나 있다. 보유 주성분 3개가 전체 데이터의 99.48%를 설명하고 있으므로 차원 축소에 따른 정보 손실은 거의 없다고 볼 수 있다.[4]

성분	고유값	누적(%)
1	2.91082	72.77
2	0.92122	95.80
3	**0.47353**	**99.48**

표 1 Iris Data의 주성분 분석 결과

3개의 주성분을 이용하여 전체 데이터에 대한 3차원 산점도를 그린 결과가 그림 2와 같이 나타났다. 전체적으로 2개의 군집수를 관찰할 수 있으므로 SOM의 형상 지도의 차원을 (2*2)로 결정하였다.

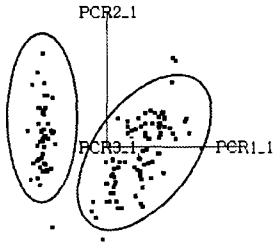

그림 2 Iris Data의 3개 주성분의 3D 산점도

다음으로 (2*2)의 형상 지도 차원을 갖는 SOM의 군집 결과가 그림 3과 같이 나타났다.

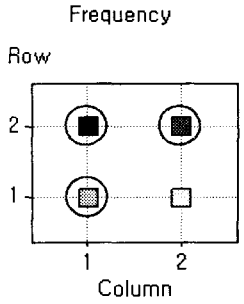

그림 3 Iris Data의 (2*2) SOM의 학습 결과

그림 3의 4개의 노드 중에서 진한 색을 나타내고 있는 노드가 군집이 형성된 노드이며, 3개의 군집이 형성됨을 알 수 있다.

클러스터	x_1	x_2	x_3	x_4	평균
1	0.426	0.879	0.100	0.141	0.387
2	0.490	0.587	0.318	0.408	0.451
3	0.593	0.605	0.362	0.434	0.499

표 2 Iris Data의 k-평균(k=3) 결과

표 2는 Iris 데이터를 이용한 K-평균 군집화 결과이며, 각 군집에 대한 각 입력 변수의 분산 값을 나타내고 있고 평균은 각 군집의 평균 분산을 나타내고 있다. 이 표의 값으로부터 3개의 군집에 대한 VC 값은 0.6605이며, 각 군집수에 대한 군집 결과에 대한 VC 값들은 표 3에 나타나 있다. 제안 시스템으로부터 결정된 3개의 군집 결과의 VC 값이 가장 작음을 알 수 있다.

군집수	VC값
2	0.6610
3	**0.6605**
4	0.6610
5	0.7145

표 3 각 군집 결과 VC값

VC값은 계산에서 앞의 식 (2-1)에서 p 값을 0.0715로 결정하였다. 이러한 결정은 군집의 수가 3개인 것을 알고 있는 Iris 데이터를 이용하여 p 값을 0.0001부터 0.2000까지 0.0001씩 증가시켜 각 군집수에 대한 VC 값을 모두 계산한 결과 p 값이 0.0715일 때 군집수 3이 다른 군집수에 비해 가장 작은 VC 값을 가지게 됨을 알 수 있었다. 따라서 본 논문에서 제안한 군집화 결과의 성능 평가 측도인 VC 식에서 p 값은 휴리스틱하게 0.0715로 결정하였다. 본 논문에서 제안하는 군집화 전략에서 모든 데이터에 대한 입력 변수

들의 척도를 모두 표준화(standardization)를 시킨 후에 군집화를 수행하기 때문에 이 값은 다른 군집화 데이터에서도 그대로 적용될 수 있다. 입력 변수들에 대한 표준화를 수행하면 변수들간의 척도에 대한 차이가 없어져서 객관적인 군집화가 가능하다. 다음 절에서는 Iris 데이터를 통해 완성시킨 VC 측도와 이 논문에서 제안하는 군집화 전략을 이용하여 실제 웹 데이터를 이용한 군집화 실험 결과를 보였다.

3.2 Microsoft web log data를 이용한 실험 및 결과

이 실험에서 사용한 데이터는 1998년 2월 중에 Microsoft 회사의 홈페이지로부터 1주일 동안 얻은 웹 로그 데이터의 표본을 추출하여 전처리한 것이다. 데이터의 속성은 총 294개의 매우 산재(sparse)되어 있는 웹 페이지를 인자분석(factor analysis)을 이용하여 10개의 입력변수로 축소시켜 주성분 분석을 수행한 결과 보유 주성분 3개가 전체 데이터의 43.19%를 설명하였다. 따라서 많은 차원을 축소했음에도 불구하고 절반 정도의 정보만이 손실되었다고 볼 수 있다.[3][6]

3개의 주성분을 이용하여 전체 데이터에 대한 3차원 산점도를 그린 결과 2개의 군집수가 관찰되었고, 이 결과로부터 (4*4)의 형상 지도 차원을 갖는 SOM의 군집 결과 5개의 군집이 형성됨을 알 수 있다. 이 값으로부터 5개의 군집에 대한 VC 값은 1.1813이다. 표 4는 각각 군집수가 다른 K-평균 군집화 결과에 대한 VC 값이다.

군집수	VC값
3	1.4427
4	1.4166
5	**1.1813**
6	1.5537
7	1.5339

표 4 각 군집 결과에 따른 VC값

표 4로부터 제안 시스템으로부터 결정된 5개의 군집 결과의 VC 값이 가장 작음을 알 수 있다. 따라서 Microsoft Web Log Data는 5개의 군집 결과를 얻게 된다.

4. 결론 및 향후 연구 과제

본 논문에서 제안된 자동 군집화 에이전트는 군집화 수행 에이전트와 군집화 성능 평가 에이전트로 구성된 멀티 에이전트로서 두 개의 에이전트가 서로 정보를 교환하면서 최적의 군집화를 수행한다. 최적의 초기 군집수를 주성분 분석과 자기 조직화 지도에 의해 결정하고, 최종적으로 K-평균 군집화 알고리즘을 사용하였다. 그리고 군집 결과의 성능 평가는 VC를 제안하여 적용하였다.

UCI Machine Repository의 IRIS 데이터와 Microsoft Web Log Data를 이용하여 실험해본 결과 본 논문에서 제안한 최적 군집화 수행 절차에 의한 결과가 다른 방법에 비해 군집의 동질성을 높이는 결과로 나타났다. 하지만 주성분분석과 자기 조직화 지도를 이용하여 초기 군집수를 결정하는 과정에서 기존의 다른 방법들에 비해서 나은 객관적인 결정을 하였으나, 여전히 주성분 분석에서의 3D 산점도에서와 같이 분석가의 주관적 판단이 요구되었다.

본 논문에서 통합 설계된 자동 군집화 에이전트는 각 에이전트간 정보교환이 이루어진다면, 웹 마이닝 프로세스에 동적으로 적용될 수 있게 될 것이며, 더불어 인터넷 상거래에서의 추천 시스템, 사용자 기호에 맞는 웹 페이지 예측, 고객 분석에 있어서 객관적이고 자동화된 활용을 할 수 있으리라 기대한다.

감사의 글

본 연구는 과학기술부 주관 뇌신경 정보학 사업에 의해 지원되었음.

5. 참고문헌

[1] Joutsensalo, J. · Miettinen, A., "Self-organizing operator map for nonlinear dimension reduction", Neural Networks, 1995. Proceedings. IEEE International Conference on, Volume: 1, vol.1 111 -114쪽

[2] T.Kohonen, "Self-Organizing Maps", Springer, 1995년

[3] http://www.ics.uci.edu/~mlearn/MLRepository.html

[4] 김기영 · 전명식, "SAS 주성분 분석", 자유아카데미, 4~7쪽 · 55~64 쪽, 1992년

[5] 박민재 · 전성해 · 오경환, "붓스트랩 기법과 유전자 알고리즘을 이용한 최적 군집 수 결정", KFIS 2002, 제12권 제 2호, 263~266쪽

[6] 전성해 · 임민택 · 오경환, "인자 점수를 이용한 이상치 데이터의 군집화", KFIS 2002, 제12권 제1호, 77~80쪽, 2002년 춘계