Fuzzy C-Means 클러스터링을 이용한 웹 로그 분석기법

김미라[0] 곽미라 조동섭
이화여자대학교 과학기술대학원 컴퓨터학과
{memory[0], mirakwak, dscho}@ewha.ac.kr

Web Log Analysis Technique using Fuzzy C-Means Clustering

Mi ra Kim[0] Mira Kwak Dong sub Cho
Dept. of Computer Science and Engineering, Ewha Womans University

요 약

클러스터링이란 주어진 데이터 집합의 패턴들을 비슷한 성질을 가지는 그룹으로 나누어 패턴 상호간의 관계를 정립하기 위한 방법론으로, 지금까지 이를 위한 많은 알고리즘들이 개발되어 왔으며, 패턴인식, 영상 처리 등의 여러 공학 분야에 널리 적용되고 있다. FCM(Fuzzy C-Means) 알고리즘은 최소자승 기준함수(least square criterion function)에 퍼지이론을 적용한 목적함수의 반복최적화(iterative optimization)에 기반을 둔 방식으로, 하드 분할에 의한 기존의 클러스터링 방법이 승자(winner take all) 형태의 방법론을 취하는데 비하여, 각 패턴이 특정 클러스터에 속하는 소속정도를 줌으로써 보다 정확한 정보를 형성하도록 도와준다. 본 논문에서는 FCM 기법을 이용한 웹로그 분석을 하고자 한다.

1. 서 론

클러스터링이란 주어진 데이터 집합의 패턴들을 비슷한 성질을 가지는 그룹으로 나누어 패턴 상호간의 관계를 정립하기 위한 방법론으로, 지금까지 이를 위한 많은 알고리즘들이 개발되어 왔으며, 패턴인식, 영상 처리 등의 여러 공학 분야에 널리 적용되고 있다.

하드(hard) 클러스터링 기법은 주어진 데이터 상호간의 관계가 명확하다는 가정 하에 각 패턴을 하나의 클러스터에 소속시키는 하드 분할 (hard partition)에 의한 방식을 사용한다. 따라서 이 방법은 다루고자 하는 데이터의 경계가 명확하지 않을 경우 실제 데이터 상호간의 군집성을 묘사하기에 부적절할 뿐만 아니라, 주어진 데이터 분포의 성질을 손실하는 결과를 초래할 수도 있다. 이를 개선하기 위하여, Bezdek은 FCM(Fuzzy C-Means) 알고리즘이라고 불리는 퍼지 분할에 의한 방법을 고안하였다. FCM 알고리즘은 최소자승 기준함수(least square criterion function)에 퍼지이론을 적용한 목적함수의 반복최적화(iterative optimization)에 기반을 둔 방식으로, 하드 분할에 의한 기존의 클러스터링 방법이 승자(winner take all) 형태의 방법론을 취하는데 비하여, 각 패턴이 특정 클러스터에 속하는 소속 정도를 줌으로써 보다 정확한 정보를 형성하도록 도와준다.

본 논문에서는 FCM 기법을 이용한 웹로그 분석을 하고자 한다. 2장에서는 데이터 클러스터링 알고리즘에 대해 알아보고, 3장에서는 FCM기법을 웹로그 분석에 적용하여 데이터 클러스터링을 한다. 4장에서는 향후연구과제를 제시하도록 한다.

이 논문은 2002년도 두뇌한국21사업에 의하여 지원되었음.

2. 데이터 클러스터링 알고리즘

클러스터링(Clustering)이란 주어진 데이터 집합을 서로 유사성을 가지는 몇 개의 클러스터로 분할해 나가는 과정으로, 하나의 클러스터에 속하는 데이터 점들 간에는 서로 다른 클러스터 내의 점들과는 구분되는 유사성을 갖게 된다. 데이터 마이닝에서 클러스터링 방법은 기존의 통계, 기계 학습, 패턴인식에서 쓰이던 방법에 부가적으로 데이터베이스 지향적인 제약 사항들을 첨가시킨 것으로서, 최근의 멀티미디어 데이터와 같이 혼합되고 다양한 다차원 데이터를 효율적으로 사용하기 위한 방안으로 연구되고 있다.

클러스터링 방법은 크게 분할(partitioning)접근과 계층적(hierarchical)접근으로 나눌 수 있다. 분할 접근은 범주 함수를 최적화시키는 k개의 분할영역을 결정해 나가는 방법으로, 유클리드 거리 측정법에 기반한다. 클러스터의 무게중심점을 대표 값으로 분할해 나가는 k-means 방법과, 클러스터내의 중심과 가장 가까운 객체로 대표점을 찾아가는 k-medoid 방법이 있으며. 분할을 위한 초기 값과 대표 값 선정 방식에 따라, 또는 거리 대신 밀도를 기반으로 하느냐에 따라 여러 가지로 변형될 수 있다. 계층적 접근은 처음에 각각의 데이터 점을 하나의 클러스터로 설정한 후 이들 쌍간의 거리를 기반으로 하여 분할, 합병해 나가는 상향식 방식으로 모든 점들이 하나의 대형 클러스터에 속하게 될 때까지 그 히스토리 정보를 유지해 나가게 된다. 다음은 클러스터링의 대표적인 알고리즘이다.

2.1 CLARANS

k-medoid 방법을 사용하는 대표적인 알고리즘인 PAM과 CLARA를 바탕으로 개발된 알고리즘으로, 적절

한 클러스터 값을 찾아가는 각 단계마다 이웃의 표본만을 고려한다. 임의 추출 검색(randomized search)을 사용하고 있지만, 1000여 개 이상의 데이터 집합에 대해서는 적용이 불가능하다는 제약점을 가지고 있다.

2.2 BIRCH

기존의 클러스터링 알고리즘에서 입력 데이터 크기 n이 커지면 이들에 대한 다중 I/O 스캔으로 병목 현상이 일어나고, 비선형 시간 복잡도(non-linear time complexity)로 인한 처리비용이 급격히 증가된다는 제약점을 극복하기 위해 제안되었다. 알고리즘 수행 방식은 먼저 전체 데이터를 스캔해내는 사전-클러스터링(pre-clustering)단계를 수행한 후 가능한 메모리에 맞는 부클러스터에 대해 요약정보를 갖고 있는 CF-tree를 검색함으로써 방대한 데이터베이스에 대해 효율적인 클러스터링을 수행한다. I/O 비용을 최소화하면서 모든 가능한 부클러스터를 파생시키고 가능한 메모리를 최대한 상용토록 하며, 다차원 데이터들이 증가되거나 동적으로 입력되는 상황에서 좋은 클러스터를 생성할 수 있다.

2.3 DBSCAN

기존의 클러스터링 알고리즘은 방대한 데이터 집합을 효율적으로 다루는 방법에 대해서만 다루었던데 비해, DBSCAN에서는 다차원적이고 공간적인 특성을 갖는 다양한 모양과 크기의 데이터에 대한 클러스터링 방법을 제시한다. DBSCAN에서는 클러스터의 밀도(density)를 결정하기 위해 2개의 파라미터, 즉 점의 이웃의 범위를 나타내는 반경과 최소 이웃의 수를 입력받는다. 부정형의 클러스터를 찾는데 있어서 CLARANS에 비해 약 100배 정도의 효율성 증가를 보인다.

2.4 DBCLASD

방대한 공간 데이터베이스에 요구되는 클러스터링 알고리즘은 최소의 입력 파라미터와 부정형의 클러스터 발견, 그리고 대형 데이터베이스에 대해서도 좋은 성능을 나타내야 한다. DBSCAN에서는 입력 파라미터에 대해 k-dist 그래프를 사용하는 휴리스틱 방법에 제안되었지만 모든 점들에 대해 k번째 근접 이웃(k-th nearest neighbor)을 구하는 것은 대형 데이터베이스에서는 어려운 일이다. 또한 작은 데이터베이스라도 k-dist 그래프를 가시화 하는 것은 스크린 크기에 의해 제한되며, 표본추출을 통해 수행하면 정확도가 떨어지게 된다. 이러한 제한을 없애기 위해 DBCLASD는 입력 파라미터를 제거하였다. 성능면에서는 DBSCAN 보다 2-3배 수행시간이 더 걸리지만 50만개 이상의 데이터에도 좋은 규모 확장성(scalability)을 보인다.

2.5 CURE

전통적인 계층적 클러스터링 알고리즘이 가질 수 있는 연쇄효과(chaining effect) 문제를 해결하기 위한 방안으로 CURE는 클러스터 당 하나 이상의 대표점을 가지며, 이들은 클러스터의 평균값(mean)으로 수렴된다. 계층적 클러스터링 방법을 적용시킬 때 합병되는 두 클러스터에 대한 대표점은 합병되는 클러스터 내의 모든 점에 대해서가 아닌 두 클러스터로부터 미리 선택되며 특히, 무작위 표본추출(random sampling)과 분할, k-d tree와 힙(heap) 데이터 구조를 사용함으로써 기존에 하나의 중심값만으로는 찾아낼 수 없었던 비구형 클러스터, 예를 들어 긴 모양의 클러스터를 발견 가능하게 하는 특징이 있다. DBCAN과 비교할 때 입력 파라미터에 대한 영향력이 적고, 밀도 높은 선으로 연결된 두 개의 서로 다른 클러스터를 구분해 낼 수 있으며 대형 데이터베이스에 적용할 때 사전-클러스터링을 수행할 있는 장점을 갖는다.

3. FCM(Fuzzy C-means) 알고리즘

본 논문에서는 데이터 클러스터링 알고리즘의 하나인 Fuzzy C-means 알고리즘을 이용하여 가공된 웹로그 데이타를 클러스터링하고자 한다.

FCM 알고리즘은 각 데이타와 각 클러스터 중심과의 거리를 고려한 유사도 측정에 기초한 목적 함수의 최적화 방식을 사용하며, 목적 함수를 다음과 같이 정의된다.

$$Jm(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^{m}\,||x_j-v_i||^2$$

여기서 n은 데이터의 개수, c는 클러스터의 개수이고. 주어진 입력 데이터 집합 X = $\{x_1,\cdots,x_n\}$에 대한 퍼지 c분할을 $c\times n$의 행렬 U로 나타낼 때 u_{ij}는 데이터 x_j가 클러스터 i에 속하는 소속 정도를 나타낸다. 또한 $||\cdot||$은 유클리디안 노름(Euclidean norm)이고, v_i는 i번째 클러스터의 중심을 나타내며, $m\in\{1,\infty\}$은 퍼지 정도를 나타내는 매개변수이다.

FCM 알고리즘의 수행절차는 다음과 같다.

단계 1: $c(2\le c\le n)$ 값과 $m(1\le m\le\infty)$값을 결정한다.

단계 2: 다음의 조건을 만족하는 퍼지 c 분할 $U^{(0)}$을 초기화한다.

$$\sum_{i=1}^{c}u_{ij}=1,0<\sum_{j=1}^{n}u_{ij}<n;u_{ij}\in[0,1],1\le i\le c,1\le j\le n$$

단계 3: 각 클러스터에 대한 클러스터의 중심 $v_i^{(l)}$를 구한다. (l=0,1,2, $\cdots$).

$$v_i^{(l)}=\frac{\sum_{j=1}^{n}(u_{ij}^{(l)})^{m}x_j}{\sum_{j=1}^{n}(u_{ij}^{(l)})^{m}},\quad 1\le i\le c$$

단계 4: 구해진 $v_i^{(l)}$를 이용하여 $U^{(l+1)}$을 계산한다.

$x_j\neq v_i^{(l)}$인 모든 $i\in N_c$에 대하여,

$$u_{ij}^{(l+1)}=\frac{1}{\sum_{k=1}^{c}\left(\frac{||x_j-v_i^{(l)}||}{||x_j-v_k^{(l)}||}\right)^{2/(m-1)}},1\le i\le c,1\le j\le n$$

$x_j=v_i^{(l)}$인 어떤 $i\in I\subset N_c$에 대하여,

$\sum_{i\in I} U_{ij}=1, i\in N_c-I$ 인 경우에 대하여 $u_{ij}=0$.

단계 5: 만약 $||U^{(l+1)}-U^{(l)}||\leq\varepsilon$ 이면 알고리즘을 끝내고, 그렇지 않으면 단계 3으로 간다.

4. FCM을 이용한 웹로그 분석

FCM을 이용한 웹로그 데이터의 클러스터링을 위해서는 우선 웹로그 데이터의 가공이 필요하다. 본 논문의 웹로그 분석에서는 웹로그의 모든 필드 중에서 x축을 시간로 하고, y축을 IP으로 가정하였다. 이 데이터셋을 WLDF(Web Log Data for FCM)라고 부르도록 하겠다. 본 논문에서 사용한 WLDF의 일부분은 다음 그림 4-1과 같다.

time	IP
3:35:28	203.255.178.32
3:53:30	203.255.120.15
4:06:46	203.255.178.14
4:07:51	203.255.178.32
4:20:22	203.255.120.15
4:33:51	203.255.178.32
4:43:08	203.255.178.47
5:08:49	203.255.178.39
5:21:00	203.255.177.147

그림 4-1 WLDF(Web Log Data for FCM)

완성된 WLDF를 데이터셋으로 하여 FCM 알고리즘에 적용을 하면 그 결과값이 다음과 같이 나타난다. 몇 번의 이터레이션 후에 다음 그림 4-2와 같은 결과값이 나오게 된다.

```
Final V1 = <4.6850, 170.7718>
Final V2 = <4.6656, 185.3862>
Final V3 = <6.2960, 119.7389>
```

그림 4-2 FCM 알고리즘을 적용한 후 결과값

웹로그 데이터를 시간과 IP로 가공한 WLDF 데이터셋을 이용하여 FCM 알고리즘을 적용한 후에 데이터들의 클러스터링 되는 결과를 볼 수 있다. 4시대에 203.255.170.771 IP를 중심으로 데이터들이 클러스터링 되고, 4시대의 203.255.185.386 IP들이 또 하나의 클러스터를 이루며, 6시대의 203.255.119.738 IP들이 클러스터링 되는 결과 값을 볼 수 있다.

5. 결론 및 향후 연구과제

FCM(Fuzzy C-Means)알고리즘은 최소자승 기준함수(least square criterion function)에 퍼지이론을 적용한 목적함수의 반복최적화(iterative optimization)에 기반을 둔 방식으로, 하드 분할에 의한 기존의 클러스터링 방법이 승자(winner take all) 형태의 방법론을 취하는데 비하여, 각 패턴이 특정 클러스터에 속하는 소속정도를 줌으로써 보다 정확한 정보를 형성하도록 도와준다. 본 논문에서는 데이터 클러스터링 알고리즘의 하나인 Fuzzy C-means 알고리즘을 이용하여 가공된 웹 로그 데이타를 클러스터링 하였다.

향후연구과제는 FCM 알고리즘을 이용하여 웹로그의 다른 필드들을 이용하여 다차원 클러스터링을 하고자 한다. 또한 성능평가를 통해 알고리즘을 발전시켜서 최선의 방법을 제시할 것이다.

참고문헌

[1] 박승수 이상호, 용환승, 김현희, 최지영, "데이타마이닝 알고리즘의 분류 및 분석," 정보과학회논문지 : 데이타베이스 제28권 제3호, 2001년.

[2] 정창호, 임영희, 박주영, 박대희, "진화프로그램을 이용한 퍼지 클러스터링," 정보과학회논문지(B) 제26권 제1호, 1999년 1월.

[3] Raymond T. Ng, Jiawei Han, "Efficient and Effective Clustering Method for Spatial Data Mining," In Proc. of the VLDB Conference, Santiago, Chile, pp.145-155, September 1994.

[4] Tian Zhang, Raghu Ramakrishnan, and Miron Livny, "BIRCH : An Efficient Data Clustering Method for Very Large Databases," In Proc. of the ACM SIGMOND Conference on Management of Data, Montreal. Canada, pp.103-114, June 1996.

[5] Martin Easter, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," In Proc. of ACM SIGMOD 3rd International Conference on Knowledge Discovery and Data Mining, pp. 226-231, AAAI Press, 1996.

[6] Xiaowei Xu, Martin Ester, Hans-Peter Kriegel, and Jorg Sander, "A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases," In Prcc. of 14th International Conference on Data Engineering(ICDE), Orlando, Florida, USA, pp.324-331, February, 1998.