제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

제약된 K-means를 위한 초기 씨드 생성방법 [1]

서향숙*, 강재호**, 류광렬*
*부산대학교 정보컴퓨터공학부
**동아대학교 지능형 통합항만관리연구센터
e-mail : {hyangsuk, jhkang, krryu}@pusan.ac.kr

Initial Seed Generation for Constrained K-means

Hyangsuk Seo*, Jaeho Kang**, Kwang Ryel Ryu*
*Division of Computer Science and Engineering, Pusan National University
**Center for Intelligent & Integrated Port Management Systems, Dong-A University

요 약

군집화 시 일반적으로 개별 클래스(class) 혹은 카테고리(category) 당 하나의 군집이 형성되는 결과가 선호된다. 하지만 데이터가 비정형적인 분포를 따르는 경우에는 하나의 군집으로 개별 클래스를 온전히 표현하는 것이 불가능하거나 오히려 부자연스러운 경우가 발생할 수 있다. 본 논문에서는 예제의 클래스를 알고 있는 즉, 레이블(label)된 예제들을 그렇지 않은(unlabeled) 예제들과 함께 활용하여 군집화하는 제약된 K-means (constrained K-means) 알고리즘을 위하여 보다 자연스러운 형태의 군집이 형성될 수 있도록 초기 씨드(seed, 씨앗)를 생성하는 방안을 제안한다. 레이블된 예제들을 계층적으로 군집화하면 다양한 단계에서 제약된 K-means를 위한 씨드집합을 생성할 수 있다. 본 연구에서는 각 단계의 씨드집합을 기반으로 형성된 군집결과간의 변화정도를 측정하여 가장 적절한 것으로 추정되는 씨드집합을 선정하였다. 제안한 방안을 문서 군집화 문제에 적용하여 실험한 결과 개별 클래스마다 하나의 군집을 가정하는 경우보다 더 나은 군집을 형성할 수 있음을 확인하였다.

1. 서론

군집화 시 일반적으로 개별 클래스 혹은 카테고리 당 하나의 군집이 형성되는 결과가 선호된다. 하지만 데이터가 비정형적으로 분포하는 경우에는 하나의 클래스를 표현하기 위하여 다수의 군집을 생성하는 것이 보다 자연스러울 수 있다. 그림 1과 같은 XOR 형태로 데이터가 분포된 경우에는 클래스의 수는 2개이지만 군집은 총 4개가 생성되는 것이 보다 자연스러운 군집결과라 할 수 있다.

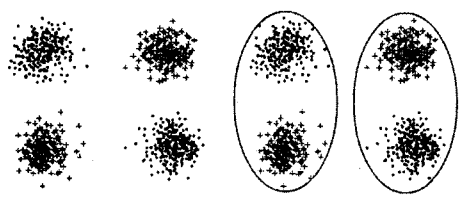

그림 1. 비정형적인 데이터예제 (XOR 문제)

본 논문에서는 예제의 클래스를 알고 있는 즉, 레이블된 예제들을 그렇지 않은 예제들과 함께 활용하여 군집화하는 준감독 군집화(semi-supervised clustering) 기법의 하나인 제약된 K-means 알고리즘을 위하여 보다 자연스러운 형태의 군집이 형성될 수 있도록 초기 씨드를 생성하는 방안을 제안하고자 한다.

준감독 군집화에서 레이블된 예제들은 군집화 수행 시 군집형태가 보다 자연스럽게 형성되도록 유도하는 역할을 수행한다. 사용자가 레이블 할 예제들을 선별할 수 있으므로 사용자의 군집화 의도를 반영할 수 있고, 상대적으로 레이블 정보를 전혀 이용하지 않는 비감독 군집화(unsupervised clustering) 기법에 비해 보다 우수한 군집 결과를 생성할 수 있다. 이러한 레이블된 예제들을 계층적으로 군집화하면 각 단계에서 제약된 K-means를 위한 씨드집합을 생성할 수 있다. 본 연구에서는 각 단계의 씨드집합을 기반한 군집결과를 보고 그 변화정도를 측정하여 가장 적절한 것으로 추정되는 씨드집합을 선정하였다. 제안한 방안을 문서 군집화 문제에 적용하여 실험한 결과 개별 클래스당 하나의 군집을 가정하는 경우보다 더 나은 군집을 형성할 수 있음을 확인하였다.

본 연구와 관련된 기존 연구로는 Dan Pelleg가 제안한 X-means[1]가 있는데 이는 비감독 군집화 방법의 하나로 우도

[1] 국가지정연구실사업(과제명: 언어 중심의 지능적 정보처리를 위한 단계적 우리말 분석기술의 개발(M10203000028-02J0000-01510))의 지원을 받아 이루어진 것임.

함수를 척도로 한 군집 분할 방법이다. 이 방법은 하나의 클래스를 여러 개의 군집으로 표현하는 것이 가능하다. 그러나 이 방법은 레이블 정보를 적용한 것이 아니다. 한편 준감독 군집화에서의 기존 연구로는 Sugato Basu가 제안한 제약된 K-means(Constrained K-means)를 위한 초기 씨드 결정 방법[2]이 있는데, 이 방법은 각 클래스마다 레이블된 데이터들의 중심점을 계산하여 초기씨드를 생성하는 방법이다. 즉 클래스 당 하나의 씨드만 생성하게 되므로 비정형적인 데이터를 적절하게 군집화하기는 힘들다. [3][4]

이어지는 구성은 다음과 같다. 2장에서는 본 논문에서 사용한 준감독 군집화 방법인 제약된 K-means방법에 대하여 설명하고, 3장에서는 씨드수를 점차적으로 늘였을 때의 군집 결과 변화에 대하여 기술하였으며, 본 논문에서 제시하는 초기 씨드 생성 알고리즘과 평가 방법에 관하여 소개하고 있다. 다음으로 4장에서는 제시한 방법을 이용하여 문서 군집화 문제를 실험하고 그 결과를 정리하였다. 5장에서는 결론 및 향후 과제에 대하여 기술하였다.

2. 제약된 K-means

알고리즘 1에 준감독 군집화 방법 중 Sugato Basu가 제안한 제약된 K-means방법[2]을 설명하였다. 알고리즘 1은 사용된 초기 씨드로 사용자가 선정해준 레이블된 데이터로부터 클래스마다 중심점을 계산하여 사용하고, 레이블된 데이터와 레이블 되지않는 데이타를 모두 포함하여 군집화를 수행한다. 군집 수행시 레이블된 데이터는 초기에 할당된 군집을 계속해서 고정시킴으로써 레이블이 변경되는 것을 막아준다.

Algorithm : Constrained K-means

Input : Set of data points $\chi = \{x_1,...,x_N\}, x_i \in \Re^d$,

number of clusters K, set $S = Y_{l=1}^{K} S_l$ of initial seeds

Output : Disjoint K partitioning $\{\chi_l\}_{l=1}^{K}$ of χ such that the KMeans objective function is optimized.

Method :

1. initialize : $\mu_h^{(0)} \leftarrow \frac{1}{|S_h|}\sum_{x \in S_h} x$, for $h = 1,...,K; t \leftarrow 0$

2. Repeat until convergence

2a. assign_cluster: For $x \in S$, if $x \in S_h$ assign x to the cluster h (i.e., set $\chi_h^{(t+1)}$). For $x \notin S$, assign x to the cluster h^* (i.e., set $\chi_{h^*}^{(t+1)}$), for $h^* = \arg\min_h \|x - \mu_h^{(t)}\|^2$

2b. estimate_means: $\mu_h^{(t+1)} \leftarrow \frac{1}{|\chi_h^{(t+1)}|}\sum_{x \in \chi_h^{(t+1)}} x$

2c. $t \leftarrow t+1$

알고리즘 1. 제약된 K-means

본 논문은 알고리즘 1의 입력(Input)으로 주어지는 클래스수/1개의 초기 씨드집합을 대신하여 군집을 더 자연스럽게 형성하는 초기 씨드를 생성하기 위한 방안을 제시한다.

3. 씨드수에 따른 군집 결과와 씨드 생성방법

3.1 씨드수에 따른 군집결과 변화

우선 씨드수가 군집화에 어떠한 영향을 미치는지 Reuter-21578문서집합[5]으로 수행한 실험에서 살펴보았다. Reuter-21578문서집합[5][2]에서 주제(TOPIC)별로 분류된 클래스 집합 중 문서수가 가장 많은 상위 2개의 earn과 acq클래스 문서를 수집하여 실험을 하였다. 문서수는 earn클래스가 3704개(65.8%), acq클래스가 1923개(34.2%)로 총 5627개의 문서를 대상으로 하였다. 전체 문서에서 클래스 비율 별로 0.3%(17개), 0.6%(34개), 0.9%(50개)만큼의 임의 데이터를 레이블링(labeling)하고 군집화를 수행하여 성능측정을 하였다. 이러한 실험을 비율 별로 각각 10번씩 실험하여 그 평균값을 그림 2에 나타내었다.

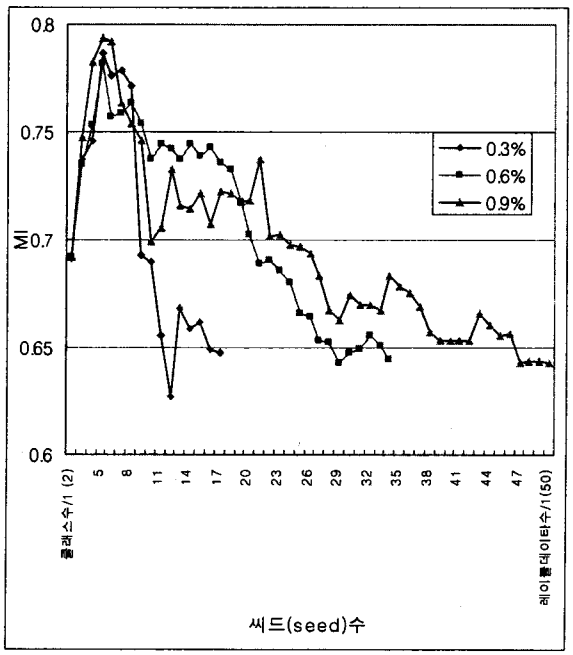

그림 2. 씨드수에 따른 군집화 결과

표 1. 레이블된 문서 집합별 실험결과.

레이블데이타비율	0.3%	0.6%	0.9%
클래스수/1	0.691	0.692	0.692
MAX	0.787	0.782	0.794

그림 2의 x축은 씨드수를 나타내는데 그 중 가장 왼쪽(클래스수/1)은 최소 씨드수인 클래스 별로 하나의 씨드를 생성한 것으로 Sugato Basu가 제안한 방법[2]을 사용한 경우이다. 가장 오른쪽(레이블 데이터수/1)은 최대 씨드수이며 레이블된 데이터 각각을 하나의 씨드로 사용한 경우이다. 중간의 20에서 60까지의 씨드는 3장에서의 씨드 병합알고리즘

[2] Reuter-21578 은 특별히 비정형적인 데이터로 직접적인 판단을 하기는 어렵다. 그러나 씨드수가 일정 수 증가 했을 때 더 좋은 결과를 나타내었다. 만약 실험 대상이 뚜렷하게 비정형적인 데이터 분포를 나타낸다면 그 결과가 더 좋게 나올 것으로 추정할 수 있다.

인 알고리즘 2를 사용하여 생성하였는데, 클래스 레이블이 같고 유사도가 가장 높은 씨드 군집간의 병합을 우선적으로 수행하는 방법으로 씨드를 레이블 데이터수/I로부터 클래스수/I 까지 씨드를 줄여주는 방법이다. 그리고 y축은 군집 평가방법중 일반적으로 사용되는 MI(Mutual Information)[6]값을 나타내었다. MI값은 0보다 크면 긍정적인 방향의 유사도가, 0보다 낮은 경우 부정적인 방향의 유사도를 나타내는 특성이 있으며, 값의 크기는 그 정도를 나타낸다. 그림 2에서 알 수 있듯이 가장 군집화가 잘된 초기 씨드수는 클래스 별로 하나씩 씨드를 생성한 경우와 레이블된 데이터 별로 하나씩 씨드를 생성한 경우 사이에 존재하는 것을 알 수 있다. 표 1은 그림 2의 결과를 MI 수치로 나타내었다. MAX는 모든 씨드집합에서의 MI 수치중에서 최대인 값을 나타낸 것이다.

군집화가 얼마나 잘 되었는지 나타내는 MI척도는 모든 데이터가 레이블을 가질 때 측정할 수 있는 방법이다. 그러나 실제로는 군집화 시에 레이블이 부여되지 않은 데이터의 레이블을 알 수 없으므로 MI척도를 이용하는 방법으로는 직접적인 평가가 어렵다. 본 논문에서는 실제 레이블 값이 없는 군집화 문제에서 휴리스틱한 방법으로 평가 가능한 기준을 제시하고자 한다. 3.2절에서는 씨드수를 줄이는 알고리즘을 기술하였고, 3.3절에서는 기존의 MI척도(알고리즘 3)와 본 논문에서 제시하는 방법을 설명하였다.

3.2 씨드수를 줄이는 알고리즘

이 절에서는 3.1장의 실험에서 사용된 씨드를 생성하는 방안에 대해 알고리즘 2에 기술하였다. 최대(레이블 데이터수/I)지점에서부터 씨드를 하나씩 줄여가며 각 단계별로 씨드집합을 생성할 수 있다. 군집방법으로 HAC(hierarchical agglomerative clustering)를 이용하였다. 이를 간단하게 설명하면, 우선 레이블된 데이터를 각각 하나의 군집으로 생성하고 전체 군집에서 레이블이 같고 유사도가 가장 높은 2개의 군집을 찾아 병합한다. 이 과정을 씨드수가 최소(클래스수/I)가 될 때까지 계속 반복하면 각 단계별로 씨드집합을 얻을 수 있다.

Algorithm : Create_Seedset
Input : Set of labeled data points $\alpha = \{a_1,..., a_j\}, a_i \in \Re^d$,
number of labeled data points j, number of categories (classes) g
Output : Set of Seedset $S_i = \{\mu_1,..., \mu_i\}, \mu_k \in \Re^d$, i = g, ..., j
Method :
1. initialize:, $i \leftarrow j, S_i \leftarrow \alpha, \mu_h \leftarrow a_h, h = 1,..., j$
2. Repeat until I < g
2a. calculate_min_seed_cluster : for $x, y \in S$,
$if\left(x \neq y, label(x) = label(y), \arg_i \min \|\mu_x - \mu_y\|^2\right)$
then assign x to min_x, y to min_y
2b. merge_ min_seed_cluster: $\mu' \leftarrow (\mu_{\min\_x} + \mu_{\min\_y})/2$
2c. $i \leftarrow i + 1$

알고리즘 2. 씨드 병합 알고리즘

3.3 MI와 ACMI

MI값이 최고인 지점(MAX)은 긍정적인 유사도가 최고인 지점이다. 즉 형성된 군집 분포와 실제 데이터의 분포가 유사함을 의미한다. 그림 2에서 MAX와 그 이전(씨드수가 하나 작을 때)과 이후(씨드수가 하나 많을 때)의 MI값 변화가 완만한 것을 볼 수 있다. 즉 군집결과가 상대적으로 유사한 것이 추정된다. 그러므로 MI값이 최고인 지점을 찾기 위해서는, 각 단계의 생성된 초기 씨드집합에 기반한 군집결과간의 MI값이 가장 높은 지점을 찾으면 된다.

Algorithm : MI
Input : g categories (classes)
$\kappa_h (h \in \{1,..., g\}, x_j \in \kappa_h \Leftrightarrow \kappa_j = h)$
"true" classification label(class label) κ
Output : $\Lambda^{(M)}(\kappa, \lambda)$
Method : $n_l^{(h)}$ denote the number of objects in cluster C_l that are classified to be h as given by κ

$$\Lambda^{(M)}(\kappa,\lambda) = \frac{1}{n}\sum_{l=1}^{\kappa}\sum_{h=1}^{g} n_l^{(h)} \frac{\log\left(\frac{n_l^{(h)} n}{\sum_{i=1}^{\kappa} n_i^{(h)} \sum_{i=1}^{g} n_l^{(i)}}\right)}{\log(\kappa \cdot g)}$$

알고리즘 3. MI

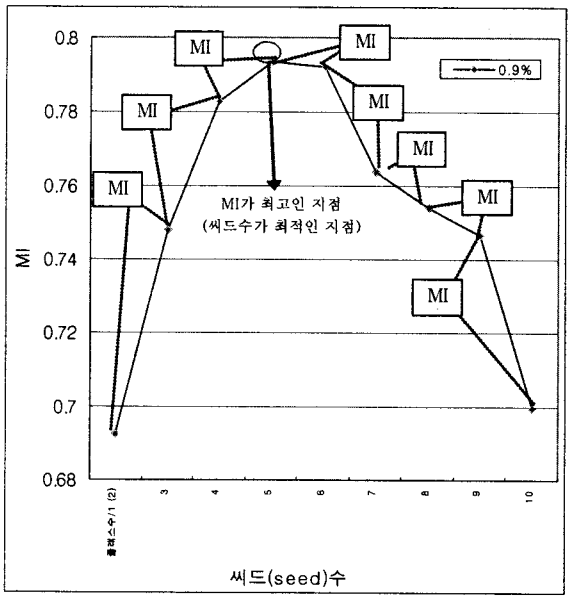

그림 3. ACMI 계산

알고리즘 3은 MI 알고리즘을 기술하였고, 그림 3은 ACMI알고리즘을 그림으로 설명한 것이다.

4. 실험 결과 및 분석

앞에서 제안한 ACMI를 이용한 초기 씨드를 생성방법이 실제 효과적인지 확인하기 위해 아래와 같은 실험을 수행하였다.

실험 데이터는 3.1 장에서 소개한 Reuter-21578 문서 데이터를 사용하여 3.1장과 동일한 형태로 실험하였고, 3.3절에서 제시한 방법으로 씨드수를 예측하여 씨드를 생성한 뒤 해당 씨드집합을 기반하여 제약된 K-means를 수행하였다. 비교실험으로는 기존 방법인 클래스수/1만큼의 씨드를 생성하여 군집화 수행후 성능을 비교하였다. 군집결과가 좋을수록 적절한 씨드가 생성되었다고 생각할 수 있으며 군집 결과 평가로는 MI를 이용하였다. 예제와 군집간의 유사도 계산방법은 문서 유사도 방법으로 널리 쓰이는 코사인 유사도를 이용하였다[7].

그림 4는 ACMI값이 최대인 지점의 씨드를 생성한 후 군집화한 결과에 MI값을 그래프로 나타낸 것이고, 표 2는 구체적인 수치를 기입하였다. 그림 4에서 문제별로 집합별 MI값이 최고인 지점을 MAX로, 비교실험인 클래스별로 하나씩 씨드를 생성하였을 때의 MI값을 클래스수/1로, 본 논문에서 제시한 ACMI값을 이용하여 씨드를 생성하였을 때의 MI값을 ACMI로 나타내고 있다. 모든 경우에 ACMI값을 이용한 경우가 클래스수/1인 경우보다 좋은 결과를 보였다. 또한 MAX값은 비율이 증가 함에 따라 조금씩 좋아지는 경향을 있으며 ACMI를 이용한 경우도 이를 따른다.

특이하게도 클래스수/1은 큰 변화가 없는데 이는 기존 방법인 클래스별 하나씩 씨드를 주는 방법이 한계가 있음을 보여주는 것이라 할 수 있겠다. 아래의 x축은 0.1%에서 0.9%까지 0.1%단위로 전체 데이터 중 임의 데이터를 레이블링한 비율을 나타낸다. 레이블된 데이터로부터 초기 씨드 집합을 생성하고, 군집화를 수행하였다. 이러한 실험을 비율별로 각각 10번씩 실험하여 그 평균값을 그림 4에 나타내었다.

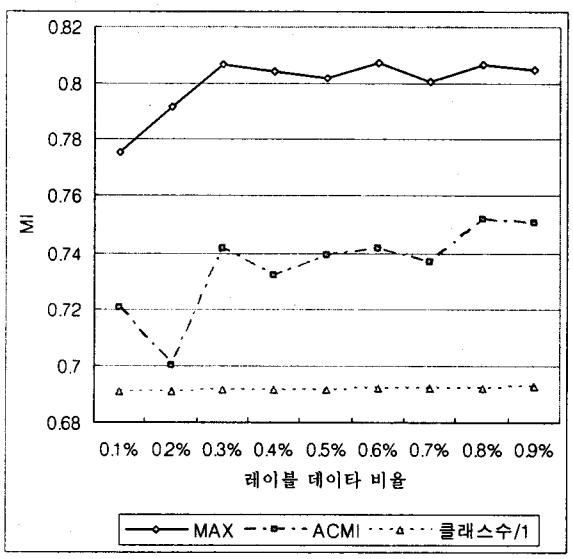

그림 4. 군집결과 MI의 그래프

표 2. 군집결과에 대한 MI값

레이블 데이타비율	0.1%	0.2%	0.3%	0.4%	0.5%
MAX	0.775	0.792	0.807	0.804	0.802
ACMI	0.721	0.700	0.742	0.732	0.739
클래스수/1	0.691	0.691	0.691	0.691	0.692
레이블 데이타비율	0.6%	0.7%	0.8%	0.9%	
MAX	0.807	0.801	0.807	0.805	
ACMI	0.741	0.737	0.752	0.750	
클래스수/1	0.692	0.692	0.692	0.692	

5. 결론 및 향후 과제

이상의 실험으로 제한된 K-means를 이용한 준감독 군집화에서 초기 씨드를 적절하게 생성해주면 클래스별 하나씩 씨드를 생성하는 방법보다 더 나은 군집을 형성 할 수 있음을 확인하였고, 제시한 ACMI를 이용한 초기 씨드 생성 방법의 유효함을 확인하였다.

향후 연구로는 보다 성능을 향상 시킬 수 있는 척도와 방법에 대한 연구가 필요하고 군집화를 수행하지 않고도 적절한 씨드를 생성하는 방법론에 관하여 연구할 필요가 있다. 또한 문서 군집화 문제가 아닌 보다 다양한 군집화 문제에서도 적용가능한지 연구해 볼 필요가 있다.

참고문헌

[1] Dau Pelleg, Andrew Moore. "X-means : Extending K-means with Efficient Estimation of the Number of Clusters". In Preceedings of the 17th International Conf. on Machine Learning, pp 727-734, 2000

[2] Sugato Basu, Arindam Banerjee, Raymond J.Mooney. "Semi-supervised clustering by seeding". In Proceedings of 19th International Conf. on Machine Learning, pp 19-26, Sydney, Australia, July 2002

[3] Kiri Wagsta, Claire Cardie, Seth Rogers, Stefan Schroedl. "Constrained K-means clustering with background knowledge". In Preceedings of 18th International Conf. on Machine Learning, pp 577-584, 2001

[4] Kiri Wagstaff, Claire Cardie. "Clustering with Instance-level Constraints". In Proceedings of 17th International Conf. on Machine Learning, pp 1103-1110, 2000

[5] Reuters-21578 Text Categorization Collection http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html 참고

[6] Strehl, A., Ghosh, J. & Mooney, R. "Impact of similarity measures on web-page clustering". In Proceedings of the 17th National Conference on Artificial Intelligence : Workshop of Artificial Intelligence for Web Search (AAAI 2000), 30-31 july 2000, Austin, Texas, USA, pp 58-64

[7] Ricardo Baeza-Yates, Berthier Rivero-Neto. Modern Information Retrieval, 1999