범주형 속성 기반 군집화를 위한 새로운 유사 측도
(A New Similarity Measure for Categorical Attribute-Based Clustering)

김 민 [†] 전 주 혁 [††] 우 경 구 [†††] 김 명 호 [††††]
(Min Kim) (Joo Hyuk Jeon) (Kyung Gu Woo) (Myoung Ho Kim)

요 약 데이터의 군집을 찾아내는 문제는 패턴 인식, 이미지 처리, 시장 조사 등 많은 응용 분야에서 널리 사용되고 있다. 군집의 질을 결정하는 핵심 요소로는 유사 측도, 차원의 개수 등이 있다. 유사 측도는 데이터의 특성을 반영하여 다르게 정의되어야 하는데, 대부분 기존의 연구들은 데이터를 특징 지어주는 속성이 수치형으로 주어진 경우에 국한되어 있었다. 속성이 범주형으로 주어진 경우도 실생활에 많이 존재하지만, 범주형 변수에 대한 속성값의 유사성은 값의 순서가 고유하게 정해지지 않아서 정의하기 어렵다. 이에 더하여, 고차원 데이터에 대해서는 데이터 점들이 희박하게 위치하여 가까운 점과 먼 점간의 차이가 거의 없고, 군집화 결과가 좋지 않을 수 있다. 이 문제를 해결하기 위해 부분 차원 군집화 방법이 제안되어 왔다. 부분 차원 군집화 방법은 각 군집을 발견하기에 적합한 부분 차원을 선택하면서 군집화를 수행하는 방법이다. 본 논문에서는 범주형 속성으로 특징지어진 고차원 데이터를 부분 차원 군집화하기 위한 새로운 유사 측도를 제안한다. 유사 측도는 각 군집은 다른 군집과 구별되는 특정 정보를 잘 표현할 수 있어야 한다는 기본적인 가정 하에 속성들 사이의 상관성을 반영하여 정의되었다. 이들 모두를 반영한 유사측도는 기존에 존재하지 않았다는 점에서 본 연구는 의미가 있다. 실제 데이터 집합을 군집화하는 실험을 통해 제안하는 방법이 다른 군집화 방법보다 저차원 데이터와 고차원 데이터 모두에 대해 좀 더 정확한 군집 결과를 얻을 수 있음을 보였다.

키워드 : 군집화, 유사 측도, k-평균 군집화

Abstract The problem of finding clusters is widely used in numerous applications, such as pattern recognition, image analysis, market analysis. The important factors that decide cluster quality are the similarity measure and the number of attributes. Similarity measures should be defined with respect to the data types. Existing similarity measures are well applicable to numerical attribute values. However, those measures do not work well when the data is described by categorical attributes, that is, when no inherent similarity measure between values. In high dimensional spaces, conventional clustering algorithms tend to break down because of sparsity of data points. To overcome this difficulty, a subspace clustering approach has been proposed. It is based on the observation that different clusters may exist in different subspaces. In this paper, we propose a new similarity measure for clustering of high dimensional categorical data. The measure is defined based on the fact that a good clustering is one where each cluster should have certain information that can distinguish it with other clusters. We also try to capture on the attribute dependencies. This study is meaningful because there has been no method to use both of them. Experimental results on real datasets show clusters obtained by our proposed similarity measure are good enough with respect to clustering accuracy.

Key words : clustering, similarity measure, k-means clustering

· 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아수행된 연구임(No. 2009-0083055)

† 학생회원 : 한국과학기술연구원 인지로봇센터
minkim@kist.re.kr
†† 학생회원 : 한국과학기술원 전산학과
jhjeon@dbserver.kaist.ac.kr
††† 정 회 원 : 삼성전자 종합기술원 SW 선행연구소 전문연구원
epigramwoo@gmail.com
†††† 종신회원 : 한국과학기술원 전산학과 교수
mhkim@dbserver.kaist.ac.kr

논문접수 : 2009년 5월 4일
심사완료 : 2010년 1월 20일

정보과학회논문지: 데이타베이스 제37권 제2호(2010.4)

1. 서 론

1.1 군집화

군집화(clustering)는 각 객체의 클래스 레이블이 알려지지 않은 데이터 객체의 집합을 유사성에 기초하여 동질적인 집단으로 데이터를 그룹화시키는 과정을 말한다. 즉, 군집 안의 객체끼리는 높은 유사성을 지니고, 다른 군집들의 객체와는 매우 다르도록 분류하는 과정이라고 할 수 있다[1]. 군집화는 패턴 인식, 이미지 처리, 시장조사를 포함한 많은 응용 분야에서 넓게 사용된다는 점에서 매우 흥미로운 연구 분야이다.

우선, 주어진 데이터 객체 각각을 특징지어 줄 수 있는 속성들이 주어졌다고 가정한다. 속성 변수의 특징에 따라 데이터 유형은 크게 수치형(numeric), 범주형(categorical), 혼합형(mixed) 데이터로 나뉜다. 데이터를 특징지어 주는 각 속성이 중량이나 강도, 수명 등과 같은 수치형 변수로 정의될 경우는 수치형 데이터, 각 속성이 색상, 평가 등의 범주형 변수로 정의될 경우는 범주형 데이터로 분류한다. 수치형 변수와 범주형 변수 모두로 정의될 경우는 혼합형 데이터라고 한다[2]. 본 논문에서는 혼합형 데이터에 대하여 사용자가 수치형 변수 또는 범주형 변수로 정의된 속성만을 선택하여 군집화를 수행하고자 할 경우도, 각각을 수치형 데이터와 범주형 데이터로 부르기로 한다.

군집화에서 핵심적인 문제는 유효한 유사 측도의 정의이다. 유사 측도는 데이터의 특성을 반영하여 다르게 정의되어야 한다. 수치형 데이터에 대한 유사 측도는 기하학적 개념으로 쉽게 정의할 수 있다. 이는 실제 속성의 값에 의존하여 결정된다(예를 들어, 1000원과 100원보다 1000원과 1100원이 더 유사하다). 이와 달리, 범주형 변수에 대한 속성값의 유사성은 값의 순서가 고유하게 정해지지 않아서 정의하기가 어렵다[3]. 범주형 변수의 두 속성값에 대한 유사도로 속성값이 같으면 1, 다르면 0을 할당하고, 두 객체의 유사도는 속성값들의 일치 비율로 계산하는 것이 보편적으로 사용되어 왔다[3]. 속성값이 같으면 1을 할당하는 것은 합당해 보이나 다르면 0을 할당하는 것은 적합하지 않다. 표 1에서 속성 '평가'에서의 속성값 '좋음'과 '매우 좋음' 사이의 유사도는 '매우 좋음'과 '나쁨' 사이의 유사도보다 커야 한다는 것을 직관적으로 생각할 수 있다. 즉, 속성값이 불일치 한다고 해서 속성값 사이의 유사도를 전혀 같지 않다는 의미의 0으로 정하는 것은 알맞지 않다. 또 다른 방법으로 각 속성값에 대해 실수 값을 대응시키고, 이에 대해 수치형 변수에 대해 정의되었던 유사 측도를 사용하는 방법이 제안되어 왔다[4]. 그러나 각 속성값에 대응하는 정확한 실수 값을 찾을 수 없다는 점에서 이는 적합한 측도가 아니다.

표 1 범주형 데이터

속성 \ 객체	색	평가
O_1	빨강	좋음
O_2	노랑	매우 좋음
O_3	파랑	나쁨
O_4	노랑	좋음

1.2 부분 차원 군집화

차원을 형성하는 속성의 개수 역시 군집을 정확히 식별하는 데 영향을 준다. 10차원 이하의 저 차원 데이터에 대해서는 차원이 많아질수록 군집화 성능이 좋아진다. 그러나 그 이상의 차원에서는 차원이 증가할수록 정확한 군집을 찾기 어렵다. 그 이유는 차원이 증가하면서 적은 수의 차원만 특정 군집에 관련되어 무관한 차원의 데이터들이 실제 군집의 발견을 방해하기 때문이다. 또한, 차원이 증가하면 데이터 점들이 공간에 희박하게 위치하여 점 사이의 거리가 비슷하게 되므로 군집화 결과가 좋지 않을 수 있다.

이러한 문제를 해결하기 위해 고차원 벡터로 표현된 객체를 저차원 벡터로 표현하는 차원 축소(dimension reduction) 방법이 제안되어 왔다. 그 기법으로 속성 변환(attribute transformation method)[5]이나 속성 부분 집합 선택(attribute subset selection)[6]이 있다. 속성 변환은 원래 객체들간의 상대적 거리를 유지하면서 객체가 놓여 있는 공간을 더 작은 공간으로 변환한다. 이는 데이터가 놓인 공간을 변환을 통해 새로운 공간으로 변형시키는 것을 의미한다. 속성 부분 집합 선택은 무관하거나 불필요한 차원을 제거하면서 군집화 작업에 가장 관련도가 큰 속성의 부분 집합을 찾는 방법이다.

차원 축소 방법은 하나의 부분 공간을 찾아서 그 부분 공간 내에서 모든 군집을 탐색한다. 그러나 일반적으로 각 군집은 서로 다른 부분 공간 내에서 발견할 수 있기 때문에 차원 축소 방법은 이러한 응용에서 정확한 군집을 발견할 수 없는 한계가 있다. 부분 차원 군집화(subspace clustering)는 속성 부분 집합 선택을 확장한 것으로 다른 부분 공간은 서로 다르고 의미 있는 군집과 연관된다는 사실에 기초하고 있다[7]. 부분 차원 군집화의 어려움은 각 군집에 대한 정확한 차원의 부분 집합을 찾아야 함과 동시에 각 차원의 부분 집합에서 정확한 군집을 찾아야 한다는 것이다. 이러한 상호 의존적인 문제에 대한 해답을 얻으려고 부분 차원을 찾아가며 군집화하는 과정이 제안되어 왔다. 그러나 기존의 기법들은 고차원 수치형 데이터에 대한 부분 차원 군집화 알고리즘에 국한되어 있었다.

본 논문에서는 고차원 범주형 데이터의 부분 차원 군

집화에 적합한 새로운 유사 측도를 정의한다. 제안된 유사 측도는 좋은 군집은 특정 정보를 잘 표현할 수 있어야 한다는 기본적인 가정하에 속성들 사이의 상관성을 반영하여 정의한다. k-평균 군집화(k-Means clustering) 방법에 제안된 유사 측도를 사용하여 효과적인 군집화가 가능함을 보인다. 제안하는 유사 측도의 유효성은 범주형 데이터 군집화 알고리즘의 성능을 측정하기 위해 널리 사용되는 데이터를 군집화한 결과로부터 측정하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로써 군집화 알고리즘의 다양한 방법에 대해서 살펴보고, 3장에서는 제안하는 유사 측도를 정의한다. 4장에서는 제안하는 유사 측도를 적용한 k-평균 군집화 알고리즘에 대해 기술한다. 5장에서는 제안한 방법과 기존의 방법을 비교한 실험 결과를 보여주며, 6장에서는 본 논문을 정리하고 요약한다.

2. 관련 연구

이 장에서는 일반적인 군집화 방법과 범주형 데이터의 군집화 알고리즘에 대해서 살펴보도록 한다. 먼저 본 논문에서 사용할 표기법에 대해서 정의한다. N개의 객체로 이루어진 데이터 집합 **U**에 대해 객체 각각을 특징지어 줄 수 있는 m개의 속성 $A_1, A_2, ..., A_m$을 선택하고, m차원 공간에 표시되어 있다고 가정한다. A_t ($1 \le t \le m$)에는 유한한 이산형 값이 존재한다고 가정한다. X_t는 각 속성에서의 속성값을 가지는 이산형 랜덤 변수를 표현하고, $U = (X_1, X_2, \cdots, X_m)$는 **U**에 속하는 값을 가질 수 있는 이산형 랜덤 변수를 나타내는 것으로 한다.

2.1 군집화 방법

군집화 알고리즘은 일반적으로 분할기법(partitioning method)과 계층적 기법(hierarchical method)으로 나뉜다.

2.1.1 분할 기법

분할 기법(partitioning method)은 주어진 데이터 집합을 군집을 나타내는 k개의 분할로 만들어주는 알고리즘이다. 분할 기법을 사용하는 대표적인 알고리즘으로 k-평균 군집화(k-Means clustering)[8]가 있다. k-평균 군집화는 먼저 k개의 초기 분할을 생성하고, 반복적인 재배정 기법(iterative relocation technique)을 사용하여 지역적으로 최적의 분할을 찾으므로 수행 속도를 빠르게 한다. k-평균 군집화 알고리즘은 그림 1과 같다. 우선 데이터와 군집의 수 k를 입력 인자로 받는다. 초기화 과정에서는 k개의 객체를 임의로 선택하고, 각각을 군집으로 생각한다. 남아있는 N-k개의 객체를 가장 가까운 군집에 각각 할당한다. 이 때 어떤 군집의 중심(center)도 다시 계산되지 않는다. 정제 과정에서는 할당된 데이터를 가지고 군집의 중심을 계산한다. 군집의 중심은 각

- N 개의 객체와 군집의 개수 k 가 주어졌다고 가정하자.
- 초기화 (initialization)
  1) k 개의 객체를 임의로 선택하고 각각 군집이라 생각한다. 각 군집의 중심(center)을 찾는다.
  2) N-k 객체 할당
     - 각 객체를 가장 가까운 군집 C_q 에 각각 할당한다.
- 정제 (refinement)
  1) 새로 할당된 객체에 대한 군집의 중심을 다시 계산한다.
  2) 객체를 가장 가까운 군집에 할당한다.

  1), 2)를 군집의 정보가 바뀌지 않을 때까지 계속한다.

그림 1 k-평균 군집화

군집에 할당된 객체의 속성값에 대한 평균으로 구해진다. 각 객체에서 가장 가까이 있는 군집의 중심을 가진 것으로 확인된 군집에 객체를 위치시킨다. 정제 과정을 군집 내의 객체 정보가 바뀌지 않을 때까지 계속한다.

분할 기법을 사용하는 군집화 알고리즘은 기본적으로 초기 분할 생성 과정과 반복적인 재배정 과정을 통해 군집을 발견한다. 다만, 군집의 중심을 구하는 방법과 객체와 군집의 중심 사이의 유사도를 측정하는 방법에 따라 여러 변형들이 존재한다. k-평균 군집화에서는 군집의 중심이 각 군집에 할당된 객체의 속성값에 대한 평균으로 구해진다[8]. 범주형 데이터에 대해서는 보통 모드(mode)과 중앙 객체(medoid)로 군집의 중심이 정해져 왔다[9,10]. 수치형 변수에 대해 속성값의 평균이 군집에 할당된 객체들의 속성값 분포를 잘 반영한다는 것이 밝혀져 있으므로 수치형 데이터의 속성 값의 평균으로 군집의 중심을 정하는 것은 매우 적절해 보인다. 그러나 모드(mode)는 가장 많이 존재하는 속성값으로 군집의 중심을 결정하는 것으로 군집의 분포를 모두 반영할 수 없는 한계를 지닌다.

2.1.2 계층적 기법

계층적 군집화 알고리즘(hierarchical method)은 제일 꼭대기에 모든 객체가 포함된 하나의 군집과 바닥에 각각 하나의 객체를 가진 군집을 형성한다. 계층적 기법은 어떻게 분할이 형성되는지에 기초하여 병합법(agglomerative method)과 분리법(divisive method)으로 구분할 수 있다. 병합법은 각 객체가 하나의 군집이라는 가정하에 각 군집의 유사도를 측정하여 가장 가까운 군집을 병합하는 과정을 거치므로 상향식(bottom-up) 방법이라 볼 수 있다. 각 군집이 합쳐질 때마다 군집의 개수가 하나씩 줄어들며, 원하는 군집의 개수가 만들어질 때까지 이 과정을 계속한다. 분리법은 모든 객체가 하나의 군집에 속한다는 가정하에 시작하는 하향식(top-down) 방법이다. 군집이 나뉘어질 때마다 군집의 개수가 하나씩 늘며, 원하는 군집의 개수가 만들어질 때까지 이 과정을 반복한다.

계층적 기법의 가장 큰 문제점은 한번 병합되거나 분리되면 다시 되돌릴 수 없다는 데에 있다. 즉, 종종 병합과 분리의 선택과 관련하여 어려움을 겪고, 이러한 결정은 다음 단계의 결정에 영향을 주기 때문에 중요하다. 따라서 병합과 분리에 관한 결정이 어떤 단계에서라도 제대로 이루어지지 않으면 군집의 질이 매우 낮다. 또한, 병합과 분리의 결정 시 객체나 군집의 적절한 개수를 평가하고 점검해야 할 필요가 있기 때문에 확장성이 좋지 않다.

2.2 범주형 데이터의 군집화 알고리즘

범주형 데이터 군집화 알고리즘에 대한 연구가 최근 많이 이루어져 왔다. k-모드(mode) 알고리즘[10]은 k-평균 군집화의 범주형 데이터에 대한 확장 알고리즘이다. 군집화 과정에서 모드를 갱신하기 위해서 두 객체의 속성값 불일치 수로 새로운 거리 측도를 사용하였다. Squeezer[11]는 각 과정마다 하나의 객체를 보고 존재하는 군집에 할당할 것인지, 아니면 새로운 군집을 생성할 것인지 결정한다. ROCK[12]은 범주형 데이터에 대한 계층적 군집화 방법으로 객체의 거리를 계산하기 위해 자카드 계수(Jaccard coefficient)를 사용한다. 객체 사이의 유사도가 주어진 임계(threshold) 값을 넘으면 두 객체는 이웃(neighbor)이라고 생각한다. 두 객체 사이의 연결(link)은 공통된 이웃의 개수로 계산된다. 그 후, 병합 방법으로 계층(hierarchy)를 형성해 나아간다.

2.2.1 개념적 군집화

개념적 군집화(conceptual clustering)는 계층적 기법에 기반을 둔 범주형 데이터의 군집화 방법으로 [13]에서 제안되었다. 개념적 군집화는 비슷한 객체의 그룹을 인식하는 보통의 군집화와는 달리, 개념이나 클래스를 나타내는 각 그룹에 대해 특징적 설명을 찾는다. 개념적 군집화 방법에서는 그룹이나 군집의 관계를 정의하려고 조건부 확률을 사용한다. 범주 효용(categorical utility)은 [14]에서 제안된 측도로써 군집의 질을 판단하기 적합하다고 알려져 있다. 이는 COBWEB[15]과 그의 변형, 예를 들어 COBWEB/3[16], ECOBWEB[17], ITERATE[18] 등에서 사용되어 왔다. V_t를 A_t가 될 수 있는 모든 속성값으로 이루어진 집합이라면, 범주 효용 CU는 다음과 같이 정의될 수 있다.

$$CU = \frac{\sum_q [p(C_q)\sum_{t=1}^{m}\sum_{v_t \in V_t}\{p(A_t=v_t|C_q)^2 - p(A_t=v_t)^2\}]}{k} \quad (1)$$

$p(A_t = v_t|C_q)$는 군집C_q에 속하는 객체의 속성 값들을 고려하였을 때 속성 A_t가 속성값 v_t를 가질 확률을 의미한다. 이에 따라 범주 효용은 군집의 정보가 주어졌을 경우 그렇지 않을 경우보다 정확히 예측되는 속성값들의 개수의 증가량의 추정을 의미한다. 범주 효용은 클래스 내(intraclass) 유사성과 클래스 간(interclass) 상이성을 평가한다. 이에 대한 더 자세한 설명은 본 논문에서 생략한다. 더 자세한 설명은 [1]을 참조하기 바란다.

2.4.2 엔트로피 기반 군집화

정보량에 수학적인 정의를 부여하고 이를 연구하는 정보 이론(information theory)의 관점에서는 정보 병목 기법(information bottleneck method)을 통해 군집화를 수행하려는 시도가 있었다. 이는 다른 확률 변수에 비하여 군집의 특정 정보를 표현하는 확률 변수를 압축함으로 고차원 데이터의 잡음을 줄이고, 더 조밀하고 엄정하게 내재된 구조를 잘 반영하는 자료를 산출하도록 한다. 여기서 군집의 특정 정보란 군집에 속하는 객체의 속성값으로 특징지어질 수 있다. 정보량은 엔트로피, 상대적 엔트로피, 상호 정보량 등을 가지고 수치화할 수 있는데 이는 LIMBO[19], COOLCAT[20] 등에서 사용되어 왔다. 엔트로피는 랜덤 변수의 불확실성에 대한 척도로 이산 랜덤 변수 X의 엔트로피 H(X)는 $H(X) = -\sum_{x\in} xp(x)\log p(x)$ 이다. U의 엔트로피는 다음과 같이 정의될 수 있다[21].

$$H(U) = \sum_{t=1}^{m} H(X_t) = -\sum_{t=1}^{m}\sum_{v_t \in V_t} p(A_t = v_t)\log p(A_t = v_t) \quad (2)$$

군집 C_k의 엔트로피 $H(U|C_k)$는

$$H(U|C_k) = -\sum_{t=1}^{m}\sum_{v_t \in V_t} p(A_t = v_t|C_k)\log p(A_t = v_t|C_k) \quad (3)$$

이며, 정보 병목 기법을 이용한 군집화 방법은 이를 최소화 하는 방향으로 군집을 만들게 된다.

개념적 군집화와 엔트로피 기반의 군집화(entropy-based clustering)는 범주형 데이터를 군집화하는 데 적절하다고 밝혀져 있지만 속성들의 확률 분포가 서로 독립이라고 가정한다는 점에서 한계가 있다.

3. 새로운 유사 측도 제안

본 장에서는 좋은 군집은 특정 정보를 잘 표현할 수 있어야 한다는 기본적인 가정 하에, 속성들 사이의 상관성을 반영한 객체와 군집 사이의 새로운 유사 측도를 제안한다. 두 객체의 유사도는 모든 속성에 대한 속성값 유사도의 합으로 정의될 수 있다. 특정 속성에서의 유사도는 속성 연관 유사도(co-occurrence-based similarity)와 군집 내 속성값 편향성(skewness)을 이용하여 정의된다.

C_{kt}를 C_k에 속하는 객체들의 t번째 속성값들을 중복을 허용하여 나타낸 집합이라고 하자. 객체 $\mathbf{x} = (x_1, x_2, \cdots, x_m)$와 군집 C_k의 A_t에 대한 유사도(similarity)는 군집 C_k

내에서 x_t의 편향성(skewness) $w_t(x_t,C_{kt})$과, x_t와 군집 C_k 간 속성 연관 유사도(co-occurrence-based similarity) $Co_S_t(x_t, C_{kt})$를 반영하여 식 (4)와 같이 정의한다. 편향성 w_t 과 속성 연관 유사도 Co_S_t에 대해서는 이어지는 절에서 보다 자세히 설명하도록 한다.

정의 1. 객체 $\mathbf{x} = (x_1,x_2, \cdots,x_m)$와 군집 C_k의 A_t에 대한 유사도

$$\lambda(\mathbf{x},C_k) = \sum_{t=1}^{m}\left(w_t(x_t,C_{kt})\cdot Co\_S_t(x_t,C_{kt})\right) \quad (4)$$

3.1 속성 연관 유사 측도

속성 A_t에서의 속성 연관 유사도(co-occurrence-based similarity)는 A_t와는 다른 속성에 대한 속성값 분포 정보를 이용하여 정의한다. 아래의 예제 1은 속성 연관 유사 측도가 어떻게 정의되었는지 설명한다.

표 2 영화 데이터 베이스의 개체[19]

	director	actor	genre
$\mathbf{o}_1$	Scorsese	De Niro	crime
$\mathbf{o}_2$	Coppola	De Niro	crime
$\mathbf{o}_3$	Hitchcock	Stewart	thriller
$\mathbf{o}_4$	Hitchcock	Grant	thriller
$\mathbf{o}_5$	Koster	Grant	comedy
$\mathbf{o}_6$	Koster	Stewart	comedy

예제 1. 위의 표 2에서 객체 $\mathbf{o}_1$과 $\mathbf{o}_2$의 director속성에 대한 유사도를 계산하는 과정을 살펴 보면, $\mathbf{o}_1$과 $\mathbf{o}_2$의 director에서의 속성값이 'Scorsese'와 'Coppola'인데 이 두 값만을 보고 $\mathbf{o}_1$과 $\mathbf{o}_2$가 director에서 얼마나 유사한지 판단하기 어렵다. 이 경우 다른 속성 값에 대한 정보를 통해 특정 속성에 대한 객체들의 유사성을 가늠해 볼 수 있다. 표 2에서 객체 $\mathbf{o}_1$과 $\mathbf{o}_2$의 actor와 genre에서의 속성값을 살펴보자. $\mathbf{o}_1$과 $\mathbf{o}_2$의 속성값이 둘 다 actor에서의 속성값이 'De Niro'이고 genre에서의 속성값이 'crime'이므로 객체 $\mathbf{o}_1$과 $\mathbf{o}_2$가 director속성에 대해 유사한 것으로 볼 수 있다. 그러나 객체 $\mathbf{o}_1$과 $\mathbf{o}_3$는 director속성에 대해서는 actor와 genre에서의 속성값이 모두 다르므로 유사하지 않은 것으로 생각된다.

객체들의 특정 속성에 대한 유사도를 다른 속성에 대한 속성값이 같은 정도로 정의하는 것은 한 속성만의 속성값을 보지 않고, 더 많은 정보를 확인한다는 점에서 매우 타당해 보인다. 즉, 하나의 속성에 대한 객체 간 유사성은 다른 속성의 속성값 분포를 확인하고 비교함으로써 계산할 수 있다.

객체 $\mathbf{x} = (x_1,x_2,\cdots,x_m)$와 군집 C_k에 대한 A_t에서의 속성 연관 유사도(co-occurrence-based similarity) $Co_S_t(x_t,C_{kt})$를 정의하기에 앞서, 각각의 속성 A_s ($1\le s\le m$)에서 군집 내 속성값 x_s의 출현 확률(attribute value occurrence probability)을 정의하도록 한다. 이는 군집 C_k 에 속하는 객체가 A_s에서의 속성값으로 x_s를 가질 확률에 해당하며, 다음과 같이 표현된다.

정의 2. 속성 A_s ($1\le s\le m$)에서 군집 내 속성값 x_s의 출현 확률

$$OP_s(x_s, C_{ks}) := p(A_s = x_s \mid C_k) \quad (5)$$

A_t에 대한 객체 $\mathbf{x}$와 군집 C_k간 속성 연관 유사도 $Co_S_t(x_t,C_{kt})$는 A_t가 아닌 속성들에 대한 속성값 출현 확률의 평균으로써 다음과 같이 계산된다.

정의 3. A_t에 대한 객체 $\mathbf{x}$와 군집 C_k간 속성 연관 유사도

$$Co\_S_t(x_t, C_{kt}) := \frac{1}{m-1}\left(\sum_{s=1,s\neq t}^{m} OP_s(x_s, C_{ks})\right) \quad (6)$$

객체와 군집의 A_t에 대한 속성 연관 유사도는 A_s($1\le s\le m$ 이고 $s\neq t$)에서의 속성값 출현 확률에 대한 평균으로 정의하여 A_t와는 다른 모든 속성에서의 객체와 군집의 속성값을 고려할 수 있게 한다.

3.2 군집 내 속성값의 편향성

속성들 간의 연관성 외에, 군집 내 속성값 분포 정보도 객체와 군집 간 유사도에 영향을 줄 수 있다. 군집 내 속성값 편향성(skewness)은 1장에서 설명한 부분 차원 군집화 방법에 대한 이해를 바탕으로 정의할 수 있다. 부분 차원 군집화 방법은 좋은 군집이라면 최소한 몇몇의 속성에 대해 특정 속성값을 가진 집합으로 정의되어야 한다는 가정에 기초한다. 본 논문에서는 군집이 어떤 속성값과 연관되어 있는지를 반영하여 군집 내 속성값에 대한 편향성을 정의한다. 다시 말해, 속성의 편향성은 동일한 속성 A_t와 속성값 x_t에 대해서 각 군집별로 다를 수 있고, 동일한 군집에 대해서도 속성 및 속성값 별로 다를 수 있다.

가령, 표 3에서 actor 속성만을 고려할 때, C_1의 편향된 actor 속성값 'De Niro'를 갖는 객체 $\mathbf{o}_9$가 그렇지 않은 $\mathbf{o}_6$~$\mathbf{o}_8$보다 actor에 대해 C_1과 더 가깝다고 할 수 있다. 주어진 군집과 객체에 대해서 각 속성별 편향성을 계산하였을 때 객체의 속성값이 A_t에 대해 많이 편향되어 있을수록 A_t는 군집을 발견하는데 주요한 역할을 하는 속성으로 판단하고, 편향성 $w_t(x_t,C_{kt})$을 다음과 같이 정의한다.

정의 4. A_t의 속성값 x_t의 C_{kt}에서 편향성

$$w_t(x_t, C_{kt}) := \frac{p(A_t = x_t \mid C_k)}{p(A_t = x_t)} \quad (7)$$

식 (7)은 t번째 속성값으로 x_t가 선택될 확률에 대한, 군집 C_k에 속하는 객체의 t번째 속성값으로 x_t가 선택될 확률이며 x_t가 군집 내 얼마나 편향되어 있는지를 나타

표 3 영화 데이터 베이스의 개체

	director	actor	genre	군집
o_1	Scorsese	De Niro	crime	C_1
o_2	Scorsese	De Niro	thriller	C_1
o_3	Hitchcock	De Niro	thriller	C_1
o_4	Hitchcock	Grant	crime	C_2
o_5	Hitchcock	Stewart	comedy	C_2
o_6	Coppola	Stewart	comedy	-
o_7	Coppola	Stewart	comedy	-
o_8	Coppola	Stewart	comedy	-
o_9	Hitchcock	De Niro	comedy	-

낸다. 이러한 정의에 따라 속성의 편향성은 동일한 속성 A_t와 속성값 x_t에 대해 각 군집 별로 다를 수 있다.

예제 2에서 객체와 군집간의 유사도를 계산하는 과정을 통해 군집화 과정에서 편향성이 어떠한 영향을 주는지 살펴보도록 한다.

예제 2. 위의 표 3은 영화 데이터 베이스 개체로, 이상적 군집의 개수는 2이고 객체 o_1~o_5는 각각 군집 C_1 또는 C_2에 이미 할당되었다고 가정한다. o_9와 군집 C_1, o_9와 군집 C_2와의 속성 director에 대한 유사도를 각각 구하는 방법에 대해서 생각해 본다. 먼저, 객체 o_9와 군집 C_1의 director에 대한 유사도는 o_9와 C_1에 속한 객체들의 actor와 genre에서의 속성값을 보고 계산할 수 있다. 식 (5)에 따라 군집 C_1에 속하는 객체의 actor에서의 속성값에 대한 o_9의 actor에서의 속성값 출현 확률은 $OP_{actor}((o_9)_{actor}, (C_1)_{actor}) = p(actor = De\ Niro|C_1) = 1$ 이고, 군집 C_1에 속하는 객체의 genre에서의 속성값에 대한 o_9의 actor에서의 속성값 출현 확률은 $OP_{genre}((o_9)_{genre}, (C_1)_{genre}) = p(genre=comedy|C_1) = 0$ 이다. 그러므로 director에 대한 객체 o_9와 군집 C_1간 속성 연관 유사도는

$$Co\_S_{director}((o_9)_{director}, (C_1)_{director}) = \frac{OP_{actor}((o_9)_{actor},(C_1)_{actor}) + OP_{genre}((o_9)_{genre},(C_1)_{genre})}{2} = \frac{1}{2}.$$

마찬가지로 o_9와 군집 C_2와의 속성 director에 대한 속성 연관 유사도는 다음과 같다.

$$Co\_S_{director}((o_9)_{director}, (C_2)_{director}) = \frac{OP_{actor}((o_9)_{actor},(C_2)_{actor}) + OP_{genre}((o_9)_{genre},(C_2)_{genre})}{2} = \frac{p(actor=De\ Niro|C_2) + p(genre=comedy|C_2)}{2} = \frac{0+1/2}{2} = \frac{1}{4}$$

C_2군집에 속하는 모든 객체는 director에서 'Hitchcock'만을 속성값으로 가지므로, 'Hitchcock'은 director에서 C_2군집을 특징지어 줄 수 있는 속성값이라고 판단할 수 있다. director의 속성값 'Hitchcock'의 출현 확률에 대한 C_2군집에서 'Hitchcock'의 출현 확률을 계산하여 이 값이 크다면 'Hitchcock'은 C_2에서 편향되어 자주 나타나는 것으로 간주한다. 주어진 속성의 속성값에 대한 각 군집별 편향성을 다르게 정의함으로써 속성값과 군집의 관련성을 반영할 수 있다. 다음은 식 (7)에 따른 군집 C_1과 C_2에서 속성값 $(o_9)_{director}$의 편향성이다.

$$w_{director}((o_9)_{director}, (C_1)_{director}) = w_{director}(Hitchcock, (C_1)_{director}) = \frac{p(director=Hitchcock|C_1)}{p(director=Hitchcock)} = \frac{1/3}{4/9} = \frac{3}{4},$$

$$w_{director}((o_9)_{director}, (C_2)_{director}) = w_{director}(Hitchcock, (C_2)_{director}) = \frac{p(director=Hitchcock|C_2)}{p(director=Hitchcock)} = \frac{1}{4/9} = \frac{9}{4}$$

$w((o_9)_{director}, C_1) < w((o_9)_{director}, C_2)$ 임을 확인할 수 있다. 이는 $(o_9)_{director}$가 C_1 보다 C_2내에서 더 편향되어 있는 값이라는 것을 의미한다.

4. 군집화 방법

본 논문에서 사용하는 군집화 방법은 그림 1에 있는 k-평균 군집화 방법을 기초로 하지만, 다음과 같은 점에서 다르다. 첫째, 먼 점 초기 선택 알고리즘(furthest first algorithm)[22]을 사용하여 초기 군집의 정보가 군집화를 수행하는 데 적절한 정보를 제공하도록 한다. 먼 점 초기화 알고리즘은 좋은 성능을 낸다는 것이 이미 밝혀진 바 있다. 둘째, 초기화 단계에서 N-k개의 객체를 분류할 때, 각 객체를 군집에 할당할 때마다 군집의 정보를 갱신하도록 한다. 본 논문에서 사용하는 군집화 알고리즘은 그림 2와 같다.

먼 점 초기 선택 알고리즘(furthest first algorithm)을 사용하기 위해서는 객체 사이의 유사도를 계산해야 한다. 객체 사이의 유사도는 각각의 객체를 하나의 군집으로 간주하여 앞서 정의한 유사도 λ를 사용함으로써 계산할 수 있다. 예를 들어 객체 o_i 와 o_j 로 이루어져 있는 군집의 유사도를 계산하는 과정을 살펴보자. $\{o_i\}$를 하나의 군집으로 보고 o_j를 객체로 생각하면 유사도는 다음과 같이 계산된다.

$$\lambda(o_j, \{o_i\}) = \sum_{t=1}^{m} \left(w_t(o_{jt}, \{o_i\}_t) \cdot Co\_S_t(o_{jt}, \{o_j\}_t) \right) \quad (8)$$

여기서 o_{it} 과 o_{jt}는 각각 o_i 와 o_j 의 t 번째 속성값이다.

이와 반대로 o_i를 객체로 $\{o_j\}$를 군집으로 볼 경우에 유사도는 다음과 같다.

$$\lambda(o_i, \{o_j\}) = \sum_{t=1}^{m} \left(w_t(o_{it}, \{o_j\}_t) \cdot Co\_S_t(o_{it}, \{o_j\}_t) \right) \quad (9)$$

군집 $\{o_i\}$와 $\{o_j\}$의 유사도 $\bar{\lambda}$는 (8)과 (9)의 평균으로 다음과 같이 정의된다.

$$\bar{\lambda}(\{o_i\}, \{o_j\}) = \frac{1}{2}\left(\lambda(o_j, \{o_i\}) + \lambda(o_i, \{o_j\}) \right) \quad (10)$$

- N 개의 객체와 군집의 개수 k 가 주어졌다고 가정하자.
- 초기화 (initialization)
  1) 먼 점 초기 선택 알고리즘(Furthest first algorithm) 을 사용하여 k 개의 객체를 선택하고 각각 군집이라 생각한다.
     ① 하나의 객체를 임으로 선택한다.
     ② 두 번째는 처음에 뽑은 군집으로부터 가장 작은 유사도를 가지는 객체를 선택한다.
     ③ 세 번째는 기존에 선택하였던 두 개의 군집에서 가장 작은 유사도를 가지는 객체를 선택한다.
     ④ 일반적으로, 다음의 객체 $\mathbf{x}$ 는 $\arg\{\min_{\mathbf{x}} \max_k \bar{\lambda}(\{\mathbf{x}\},C_k)\}$, 여기서 C_k 는 위에서 이미 선택된 객체를 각각 가지는 군집을 나타낸다.
  2) N-k 객체 할당
     - 객체를 하나씩 가장 가까운 군집에 할당하고, 새로 할당된 객체의 정보를 반영하여 군집의 정보를 갱신한다.
       모든 객체가 하나의 군집에 속하게 될 때까지 계속한다.
- 정제 (refinement)
  1) 새로 할당된 객체에 대한 군집의 정보를 다시 계산한다.
  2) 객체를 가장 가까운 군집에 할당한다.
  1), 2)를 군집의 정보가 바뀌지 않을 때까지 계속한다.

그림 2 수정된 k-평균 군집화 알고리즘

식 (10)과 같이 정의함으로 유사도 $\bar{\lambda}$는 대칭성을 만족한다.

5. 실험 및 평가

본 장에서는 제안하는 유사 측도를 평가하기 위해 사용된 실험 데이터 집합 및 실험 결과에 대해 기술한다. UCI machine Learning Repository [23]에서 제공하는 Breast Cancer, Congressional voting, Soybean 데이터를 이용하여 제안하는 측도의 효용성을 평가하고자 하였다. 이 데이터들은 범주형 데이터 군집화 알고리즘의 성능을 측정하기 위해 널리 사용되어 왔다. 데이터 집합의 특징들은 표 4에 요약되어 있다.

군집화의 결과를 측정하기 위해, 우리는 이미 이상적으로 분류된 데이터를 알고 있다고 가정한다. 그리고 실험으로 얻어진 군집과 실제 분할에 모두 속하는 객체의 개수로 군집화의 질을 측정한다.

5.1 실험 데이터 및 결과 측정 방법

본 논문에서는 제안된 유사 측도의 유효성을 검증하기 위해 Breast Cancer 데이터, Congressional voting 데이터, soybean 데이터를 군집화 하였다. 우선, Breast cancer 데이터는 699개의 객체로 이루어져 있고, 각 객체는 9개의 속성으로 특징지어진다. 객체는 benign 과 malignant 중의 한 군집으로 할당되어 있다. 군집들은 458개와 241개의 객체로 이루어진다. 각 속성은 2~10개의 중복되지 않는 속성값을 갖는다. Congressional voting 데이터는 1998년에 미국의 국회의원들이 투표한 정보이다. 16개의 논쟁이 각 속성으로 나타내어지고, 각각에 대해 yes 또는 no의 속성값이 존재한다. republican과 democrat의 두 개 분할로 이루어지고, 각각 435개와 267개의 객체들이 이상적으로 속해있다. 각 속성은 2~7개의 중복되지 않는 속성값을 갖는다. 마지막으로, 47개의 객체가 35개의 속성으로 구분돼 있는 Soybean 데이터는 이상적으로 4개의 군집으로 나뉘어진다. 즉, 각 레코드는 diaporthe stem rot, charcoal rot, rhizoctonia root rot, phytophthora rot 중의 하나의 군집에 속한다. phytophthora rot에 속하는 객체는 17개이며

표 4 데이터 집합의 특징

Data set	객체 수	속성 수	군집 수	각 군집에 들어 있는 객체 수
Breast cancer	699	9	2	458 ; 241
Congressional voting	435	16	2	167 ; 168
Soybean	47	35	4	10 ; 10 ; 10 ; 17

나머지는 각각 10개의 객체로 이루어져 있다.

군집화 결과를 측정하기 위한 군집의 정확도 r[24]은 다음과 같이 정의된다. 여기서 k는 군집의 개수, N은 데이터 집합 안의 객체의 수, 그리고 a_i는 i번째 군집과 그에 대응하는 분할에 모두 나타나는 객체의 수를 가리킨다.

$$r = \frac{\sum_{i=1}^{k} a_i}{N} \quad (11)$$

5.2 실험 결과

아래 표 5, 6, 7는 제안된 유사 측도를 이용해서 군집화한 실험 결과를 오분류 행렬(misclassification matrix)로 나타낸 것이다. Breast cancer 데이터를 군집화한 결과에 대해 군집의 정확도 r_a는 (11)에서 정의한 식에 의해 다음과 같이 계산된다.

$$r_a = \frac{443 + 187}{699} = 0.90$$

마찬가지로, Congressional voting 데이터와 Soybean 데이터에 대해서도 군집화 결과에 대한 군집의 정확도 r_b, r_c를 다음과 같이 구할 수 있다.

$$r_b = \frac{225 + 159}{435} = 0.88$$

$$r_c = \frac{10 + 10 + 10 + 17}{47} = 1$$

표 8, 10, 11은 각 데이터를 서로 다른 알고리즘을 사용하여 군집화한 결과를 비교한 것이다. 표 8은 Breast cancer 데이터를 군집화한 결과를 나열한 것이다. COOLCAT과 제안된 방법을 사용하여 군집화를 수행한 결과 모두가 0.9 이상의 정확도로 정확한 군집을 찾는데 효과적인 것을 확인할 수 있다. Breast cancer 데이터의 적절한 군집은 엔트로피를 이용한 군집화 과정으로 찾을 수 있고, 제안된 유사 측도는 엔트로피 개념을 반영하여 적절하게 정의되었음을 알 수 있다.

표 5 제안된 유사 측도를 이용한 Breast cancer 데이터의 군집화 결과

	Benign	Malignant
실험으로 얻어진 군집1	443	54
실험으로 얻어진 군집2	15	187

표 6 제안된 유사 측도를 이용한 Congressional voting 데이터의 군집화 결과

	Democrats	Republicans
실험으로 얻어진 군집1	225	9
실험으로 얻어진 군집2	42	159

표 7 제안된 유사 측도를 이용한 Soybean 데이터의 군집화 결과

	D1	D2	D3	D4
실험으로 얻어진 군집1	10	0	0	0
실험으로 얻어진 군집2	0	10	0	0
실험으로 얻어진 군집3	0	0	10	0
실험으로 얻어진 군집4	0	0	0	17

표 8 서로 다른 유사 측도를 사용한 Breast cancer 데이터의 군집화 결과

알고리즘	r
제안된 측도를 이용한 군집화	0.90
COOLCAT	0.95

아래의 표 9는 Breast cancer 데이터를 ROCK을 이용하여 군집화한 결과이다. Breast Cancer 데이터를 2개의 군집으로 나누려 시도하여도 실제 두 객체가 유사하더라도 공통된 이웃 객체를 많이 가지지 않는 경우가 많아 ROCK으로는 두개의 군집을 형성시키는 데에 어려움이 발생한다. 2장에서 설명한 바와 같이 ROCK은 두 객체의 유사도를 공통된 이웃의 개수로 정의한다. 두 객체 뿐만 아니라 이웃을 확인하므로 ROCK은 좀 더 많은 정보를 사용할 수 있도록 한다. 이로써, ROCK은 대체적으로 좋은 군집을 찾아주나, 공통된 이웃 객체를 많이 가지지 않으면서 두 객체의 유사도가 높은 경우에는 적용 가능하지 않다고 알려져 있다[27].

COOLCAT은 범주형 데이터 군집화에 적절하나, 고차원 벡터로 나타내어진 데이터를 군집화 하기는 어렵다고 알려져 있다[26]. 표 10과 11은 각각 16개, 35개의 속성으로 각각 특징지어진 Congressional voting 데이터와 Soybean 데이터를 군집화한 결과이다. 이들은 본 논문에서 제안한 방법이 속성의 개수에 따라 군집화 결과가 영향을 받는 COOLCAT보다 좋은 군집화 결과를 얻을 수 있음을 보여준다. 제안한 방법이 고차원 데이터에 대해서도 좋은 결과를 가져오는 이유는 부분 차원 군집화 방법의 개념으로부터 정의된 편향성이 군집화 과정에서 잘 반영되었기 때문이다.

표 10과 표 11에서 Congressional voting 데이터에 대해서 ROCK을 사용한 군집화 결과가 제안한 방법을 사용한 군집화 결과보다 좋고, Soybean 데이터에 대해서는 그렇지 않은 것을 알 수 있다. 표 11의 결과로부터 Soybean 데이터는 ROCK으로 좋은 군집을 찾을 수 없는 데이터에 속한다고 판단할 수 있다. 이에 더하여, 제안된 방법은 k-평균 군집화에 기반하여 군집화 과정을 진행하였고, COOLCAT과 ROCK은 계층적 방법에 기반한 방법이라는 점에서 제안된 방법이 좀 더 빠르게 군집을 찾을 수 있다. 참고로, k-평균 군집화의 시간 복잡도는 O(kN)[28] 이고, 계층적 방법에 시간 복잡도는 $O(N^2 \log N)$[29] 이다.

표 9 ROCK을 이용한 Breast Cancer 데이터의 군집화 결과

실험으로 얻어진 군집	Benign	Malignant	실험으로 얻어진 군집	Benign	Malignant	실험으로 얻어진 군집	Benign	Malignant
1	442	15	71	0	1	141	0	1
2	1	0	72	0	1	142	0	1
3	1	0	73	0	1	143	0	1
4	0	14	74	0	1	144	1	0
5	0	2	75	1	0	145	0	3
6	0	1	76	0	1	146	0	1
7	0	1	77	0	1	147	0	1
8	0	1	78	0	1	148	0	1
9	0	1	79	0	1	149	0	1
10	0	1	80	0	1	150	1	0
11	0	1	81	0	1	151	0	1
12	0	1	82	0	1	152	0	1
13	0	1	83	0	1	153	0	1
14	1	0	84	0	1	154	0	1
15	0	1	85	0	1	155	0	1
16	0	30	86	0	1	156	0	1
17	0	1	87	0	1	157	0	1
18	0	1	88	1	0	158	0	1
19	0	1	89	0	1	159	0	1
20	0	1	90	0	1	160	0	1
21	0	1	91	0	1	161	0	1
22	0	1	92	0	2	162	0	1
23	0	1	93	1	0	163	0	1
24	0	1	94	0	1	164	0	1
25	0	2	95	0	1	165	0	1
26	0	1	96	1	0	166	0	1
27	0	1	97	0	1	167	0	1
28	0	1	98	0	1	168	0	1
29	0	1	99	0	1	169	0	1
30	0	1	100	0	1	170	0	1
31	0	1	101	0	1	171	0	1
32	0	1	102	0	1	172	0	1
33	0	1	103	0	1	173	0	1
34	0	1	104	0	1	174	0	1
35	0	1	105	0	1	175	0	1
36	0	1	106	0	2	176	0	1
37	0	1	107	0	1	177	0	1
38	0	1	108	0	1	178	0	1
39	0	1	109	0	1	179	0	1
40	0	1	110	0	1	180	0	1
41	0	1	111	0	1	181	1	0
42	0	1	112	1	0	182	0	1
43	0	1	113	0	1	183	0	1
44	0	1	114	0	1	184	0	1
45	0	1	115	0	1	185	1	0
46	0	1	116	0	1	186	0	1
47	0	1	117	0	1	187	0	1
48	0	1	118	0	1	188	0	1
49	0	1	119	1	0	189	0	1
50	0	1	120	0	1	190	0	1
51	1	0	121	0	1	191	0	1
52	0	1	122	1	0	192	0	1
53	0	2	123	0	1	193	0	1
54	0	1	124	0	1			
55	0	1	125	0	1			
56	0	1	126	0	2			
57	0	1	127	0	1			
58	0	1	128	0	1			
59	0	1	129	0	1			
60	0	1	130	0	1			
61	0	1	131	0	1			
62	0	1	132	0	1			
63	0	1	133	0	1			
64	0	1	134	0	1			
65	0	1	135	1	0			
66	0	1	136	0	1			
67	0	1	137	0	1			
68	0	1	138	0	1			
69	0	1	139	0	1			
70	0	1	140	0	1			

표 11은 Soybean 데이터에 대해 서로 다른 알고리즘을 사용하여 군집화한 결과이다. 특히, 본 논문에서 제안한 군집화 방법을 사용하여 정확한 군집을 찾을 수 있는데, 이는 이상적 군집의 개수가 많아서 군집 각각이 특정한 정보를 포함할 가능성이 높게 되고, 식 (7)에서 부분 차원 군집화 방법의 개념을 이용해 정의한 편향성의 영향이 극대화 된 것이라 생각된다. 결과적으로, 실제 데이터 집합을 군집화하는 실험을 통해서, 제안하는

표 10 서로 다른 유사 측도를 사용한 Congressional voting 데이터의 군집화 결과

알고리즘	r
제안된 측도를 이용한 군집화	0.88
COOLCAT	0.85
ROCK	0.93

표 11 서로 다른 유사 측도를 사용한 Soybean 데이터의 군집화 결과

알고리즘	r
제안된 측도를 이용한 군집화	1
COOLCAT	0.88
ROCK	0.57

방법이 저차원 범주형 데이터 뿐만 아니라 고차원 범주형 데이터에 대해서도 정확한 군집을 찾는데 효과적임을 확인하였다.

6. 결 론

본 논문에서는 좋은 군집은 특정 정보를 잘 표현할 수 있어야 한다는 기본적인 가정 하에 속성들 사이의 상관성을 반영한 새로운 유사 측도를 제안하였다. 개념적 군집화와 엔트로피 기반의 군집화는 범주형 데이터에 대해 대체로 좋은 결과를 낸다고 밝혀져 있다. 그러나 각 속성에서의 정보량과 엔트로피만을 이용하여 군집화를 수행하면 속성 간의 의존성을 무시하여 정확하지 못한 결과를 낼 가능성이 있다. 본 논문에서는 이러한 문제점들을 보완할 수 있는 새로운 유사 측도를 제안하였다. 그리고 개선된 k-평균 군집화 알고리즘에 제안하는 유사 측도를 사용하여 다양한 실험을 하였다. 실제 데이터 집합을 군집화하는 실험을 통하여 제안하는 방법이 고차원 데이터는 물론 저차원 범주형 데이터에 대해서도 정확한 군집 결과를 찾을 수 있음을 보였다.

참 고 문 헌

[1] H. Jiawei and K. Micheline, Data Mining: Concepts and Techniques, 2rd ed., pp.383-444, Morgan Kaufmann, 2006.

[2] A. Ahmad and L. Dey, A k-mean clustering algorithm for mixed numeric and categorical data, Data & Knowledge Engineering, vol.63, Issue 2, pp.503-527, 2007.

[3] C. Stanfill and D. Waltz, Toward memory-based reasoning, Communications of the ACM, vol.29, no.12, pp.1213-1228, 1986.

[4] P. H. A. Sneath and R. R. Sokal, Numerical Taxonomy: The Principles and Practice of Numerical Classication, W. H. Freeman and Company, 1973.

[5] C. Ding, X. He, H. Zha, and H. D. Simon, Adaptive dimension reduction for clustering high dimensional data. Proceedings of Second IEEE International Conference on Data Mining, pp. 147-154, 2002.

[6] L. Yu and H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution. Proceedings of the twentieth International Conference on Machine Learning, pp.856-863, 2003.

[7] S. Raychaudhuri, P. D. Sutphin, J. T. Chang, and R. B. Altman, Basic microarray analysis: grouping and feature reduction. Trends in Biotechnology, vol.19, no.5, pp.189-193, 2001.

[8] J. MacQueen, Some methods for classification and analysis of multivariate observation. Proceedings of the fifth Berkeley Symp. on Math. Statist. and Prob., vol.1, pp.281-297, 1966.

[9] Z. Huang, Extensions to the k-means algorithm for clustering large data sets with categorical data. Data Mining and Knowledge Discovery, vol.2, no.3, pp.283-304, 1998.

[10] L. Kaufman and P. Rousseeuw, Clustering by means of medoids. In Dodge, Y. (Ed.) Statistical Data Analysis based on the L1 Norm. pp.405-416, 1987.

[11] Z. He, X. Xu and S. Deng, Squeezer: an efficient algorithm for clustering categorical data. Journal of Computer Science and Technology, vol.17, no.5, pp.611-624, 2002.

[12] S. Guha, R. Rastogi and K. Shim, ROCK: a robust clustering algorithm for categorical attributes. Proceedings of the 15th International Conference on Data Engineering, pp.512-521, 1999.

[13] D. H. Fisher, Knowledge acquisition via incremental conceptual clustering. Machine Learning, vol.2, no.2, pp.139-172, 1987.

[14] M. Gluck and J. Corter, Information, Uncertainty, and the Utility of Categories. Proceedings of Seventh Annual Conference of Cognitive Science Society, pp.283-287, 1985.

[15] Z. Huang and M.K. Ng, A fuzzy k-modes algorithm for clustering categorical data. IEEE Transactions on Fuzzy Systems, vol.7, no.4, pp.446-452, 1999.

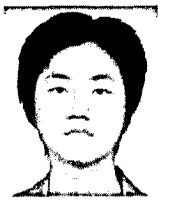
[16] K.B. McKusick and K. Thompson, COBWEB/3: A portable implementation, Report FIA-90-6-18-2, NASA, Ames Research Center, 1990.

[17] Y. Reich and S.J. Fenves, The formation and use of abstract concepts in design. Concept Formation: Knowledge and Experience in Unsupervised Learning, Morgan Kaufmann, 1991.

[18] G. Biswas, J. Weinberg, and C. Li, ITERATE: A conceptual clustering scheme for knowledge discovery in databases. Artificial Intelligence in the Petroleum Industry, B. Braunschweig and R. Day eds., pp.111-139, 1995.

[19] P. Andritsos, P. Tsaparas, R.J. Miller and K.C. Sevcik, LIMBO: Scalable clustering of categorical data. Proceedings of the 9th International Conference on Extending DataBase Technology (EDBT), 2004.
[20] D. Barbará, Y. Li and J. Couto, COOLCAT: an entropy-based algorithm for categorical clustering. Proceedings of ACM Conf. on Information and Knowledge Mgt. (CIKM), pp.582-589, 2002.
[21] T. Cover, J. Thomas, Elements of information theory, Wiley InterScience, 1991.
[22] D. Hochbaum and D. Shmoys, A best possible heuristic for the k-center problem. Mathematics of Operations Research, vol.10, no.2, pp.180-184, 1985.
[23] C. J. Merz and P. Merphy, UCI Repository of Machine Learning Databases, 1996. Available from: <http://www.ics.uci.edu/~mlearn/MLRRepository.html>.
[24] Z. Huang, A fast clustering algorithm to cluster very large categorical data sets in data mining. Proceedings of SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, pp.1-8, 1997.
[25] F. Cao, J. Liang and L. Bai, A new initialization method for categorical data clustering, Expert Systems With Applications: An International Journal archive, vol.36, Issue 7, pp.10223-102228, 2009.
[26] M. Al-Razgan, C. Domeniconi and D. Barbara, Random Subspace Ensembles for Clustering Categorical Data. Studies in Computational Intelligence, Springer, 2008.
[27] B. Broda and M. Piasecki, Experiments in Clustering Documents for Automatic Acquisition of Lexical Semantic Networks for Polish, Proceedings of the 16th International Conference Intelligent Information Systems, 2008, pp.203-202, 2008.
[28] A. M. Fahim, G. Saake, A. M. Salem, F. A. Torkey, and M. A. Ramadan, k-Means for Spherical Clusters with Large Variance in Sizes, Proceedings of World Academy of Science, Engineering and Technology, vol.35, pp.177-182, 2008.
[29] K. Qin, M. Xu, Y. Du, and S. Yue, Cloud Model and Hierarchical Clustering Based Spatial Data Mining Method and Application, Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial information Sciences, vol.37, pp.241-246, 2008.

김 민
2006년 경희대학교 수학과 학사. 2009년 KAIST 전산학과 석사. 2009년 6월~현재 한국과학기술연구원 인지로봇센터 연구원. 관심분야는 Database system, Data mining, Image processing



전 주 혁
2005년 KAIST 전자전산학과 전산학전공 학사. 2007년 KAIST 전자전산학과 전산학전공 석사. 2007년 3월~현재 KAIST 전자전산학과 전산학전공 박사과정. 관심분야는 데이터베이스 시스템, 센서 네트워크, 스트림 데이터 처리, 시멘틱 웹 등

우 경 구
1991년~1996년 서울대학교 컴퓨터공학과 학사. 1997년~1998년 한국과학기술원 전산학과 석사. 1998년~2004년 한국과학기술원 전산학과 전산학전공 박사. 2003년~현재 삼성전자 종합기술원 Future IT연구소 전문연구원(구 SW연구소). 관심분야는 Data Mining, Database, Search, Information Visualization 등

김 명 호
정보과학회논문지 : 데이타베이스
제 37 권 제 1 호 참조