클러스터 앙상블을 활용한 K-modes 알고리즘의 성능 향상 기법

An Ensemble Approach for performance stability of the K-modes algorithm

저자
(Authors)

조진혁, 고송, 김대원
Jin-Hyuk Jo, Song Ko, Dae-Won Kim

출처
(Source)

한국지능시스템학회 학술발표 논문집 19(1), 2009.4, 145-148(4 pages)

발행처
(Publisher)

한국지능시스템학회
Korean Institute of Intelligent Systems

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE01372408

APA Style

조진혁, 고송, 김대원 (2009). 클러스터 앙상블을 활용한 K-modes 알고리즘의 성능 향상 기법. 한국지능시스템학회 학술발표 논문집, 19(1), 145-148

클러스터 앙상블을 활용한 K-modes 알고리즘의 성능 향상 기법

An Ensemble Approach for performance stability of the K-modes algorithm

조진혁[1] · 고송[1] · 김대원[1]
Jin-Hyuk Jo, Song Ko and Dae-Won Kim

[1]중앙대학교 컴퓨터공학과
E-mail: jinhyukjo@gmail.com sko22.cau@gmail.com dwkim@cau.ac.kr

요 약

본 논문은 K-modes의 초기 값에 의한 성능 편차 문제를 극복하기 위해, 다양한 초기 값에 의한 K-modes 결과를 앙상블 하는 방법을 다룬다. 적용한 앙상블 기법은 K-modes의 초기 값 변화에 의한 결과에서 높은 신뢰도를 보이는 패턴들을 기반으로 해당 그룹의 모델을 형성 한 후 나머지 패턴을 대상으로 군집화하여 모델과의 유사성 비교를 통해 해당 그룹에 할당하는 방법에 대한 것이다. 실험 결과를 통해 제안한 앙상블 알고리즘이 안정성과 성능에서 개선됨을 볼 수 있었다.

키워드 : K-modes 클러스터링, 앙상블 알고리즘

1. 서 론

데이터의 분포 형태를 적합한 모델로 표현하기 위해서 활용될 수 있는 클러스터링 방법은 수치형, 범주형, 혼합형 등 데이터타입에 따라 다양하게 소개되고 있다. 그 중 본 논문에서는 범주형 데이터에 대한 클러스터링 기법 성능의 개선을 위한 방법을 소개한다.

범주형 데이터 측정을 위한 빈도수 기반의 분할 클러스터링 기법인 K-modes가 대표적으로 사용되고 있다. K-modes는 절차가 간단하고 수렴속도가 빠른 장점이 있지만, 단점으로 초기 중심에 의한 클러스터링 결과의 불안정성이 크다[1, 2]. 본 논문은 K-modes의 단점을 극복하기 위해 안정성 있는 클러스터링 결과를 도출하는 것에 주안점을 두고 연구를 진행하였다. 문제를 해결하기 위해서, 다양한 초기 중심에 의한 K-modes 결과를 분석하여, 지속적으로 같은 그룹에 속하는 패턴들을 앙상블하였다. 앙상블 된 패턴들을 하나의 그룹으로 모델링하고, 신뢰성 있는 패턴들로 그룹화 된 그룹들의 중심이 K-modes의 초기 중심 보다 성능의 향상 및 안정성 증가에 기여 한다는 가정 하에 연구를 진행 하였다[3].

2. 제안하는 기법

2.1 기본 정의

본 절에서는 제안하는 기법에서 활용하는 기본 정의를 다룬다. 집합 $X=\{X_1,X_2,X,\ldots X_n,\}$의 데이터는 n개($1\le i \le n$)이고, Xi($1\le i \le n$), 범주형 특징 j개를 가질 때, $Xi=\{X_{i1},X_{i2},X_{i3},\ldots X_{ij}\}$ 로 정의 된다. 두 데이터 집한 간의 비유사성을 식(2.1)과 같이 정의하였다.

$$d(X,Y)=\sum_{l=0}^{n}\delta(x_l,y_l) \quad (2.1)$$

$$\delta(x,y)=\begin{cases}0, & x=y\\ 1, & x\neq y\end{cases} \quad (2.2)$$

식 (2.2)의 δ(X,Y)는 두 값이 일치 할 때 0, 그렇지 않을 때 1을 갖는 함수이다. 그룹 수 k개가 있을 때 클러스터 중심 벡터 $Q=\{Q_1,\ldots,Q_k\}$로 정의 한다면 집합 X와 클러스터중심 벡터사이의 거리 척도 식(2.3)과 같이 정의 된다.

집합 X에서 변수 A_j가 c_{ij}를 갖는 빈도수를 n_{ij}라고 하면 X_j가 c_{ij}를 가질 상대빈도는 식(2.4)이 된다.

$$D(Q,X)=\sum_{l=0}^{n}d(X_l,Q_l) \quad (2.3)$$

$$f(A_j=c_{ij}|X)=\frac{n_{ij}}{n} \quad (2.4)$$

$f(A_i=q_i|X)\ge f(Aj=c_{ij}|X)$을 만족하는 클러스터중심 벡터 $Q=\{Q_1,\ldots,Q_k\}$는 비유사성의 합 D(Q.X)를 최소화 함으로, 결과적으로 집합 X의 모드가 된다. 즉 변수별로 빈도가 가장 큰 범주 값들의 조합이 그 집합의 모드가 된다.

2.2 기존 연구의 단점

표1은 K-modes의 단점으로 초기 결과에 의한 불안정한 성능을 증명하기 위한 실험 데이터이다. 실험 데이터는 10개의 패턴과 (p1~p10), 5개의 속성(f1~f5)으로 이루어진 데이터이고, 2개의 클래스로 분류하고 c1,c2로 지칭한다. 클러스터링 실행에 사용되는 2개 클래스 c1과

《표1 K-modes 예제 데이터》

	f1	f2	f3	f4	f5
p1	1	0	0	0	0
p2	1	1	1	1	0
p3	1	0	1	0	1
p4	0	0	1	0	0
p5	0	1	1	1	0
p6	1	1	1	0	0
p7	1	0	1	1	0
p8	0	1	1	0	1
p9	1	0	0	1	0
p10	0	1	1	0	1

《표2 K-modes 초기중심과 결과》

	초기 중심	결과
1회	m1{1,1,1,1,0}	c1{p1,p2,p5,p6,p7,p9}
	m2{1,1,1,0,0}	c2{p3,p4,p8,p10}
2회	m1{1,0,0,1,0}	c1{p1,p2,p7,p9}
	m2{0,1,1,0,1}	c2{p3,p4,p5,p6,p8,p10}

c2에 대한 초기 중심은 m1,m2로 지칭하고 유사도가 같을 경우 c1에 수렴하여 실험 하였다.

표2의 K-modes결과 1회 실행과 2회 실행의 종료 된 클러스터링의 결과가 다름을 보이고 있다. 이와 같은 K-modes의 초기중심에 의한 결과의 불안전성을 해결하기 위한 앙상블을 활용한 문제 해결을 다음 2.3절에서 제안하고 한다.

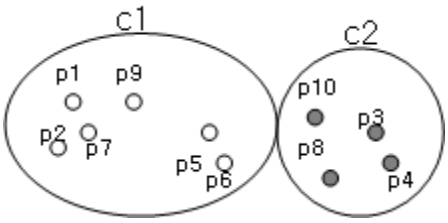
2.3 제안하는 앙상블 K-modes

앙상블의 적용 기법은 클러스터링을 반복 실행하여, 지속적으로 연관되어 같은 그룹에 속하는 패턴들을 유사도가 높은 패턴들로 보고, 패턴들을 그룹화 한다. 유사도가 높은 패턴들로 형성된 그룹의 중심은 K-modes의 초기중심 보다 신뢰성 있는 중심으로 보고, 이후 잔여 패턴들과 형성된 중심과의 유사도를 측정하여 유사도가 높은 그룹에 배속하는 것이 본 논문에서 제안하는 알고리즘의 기본 개념이다. 지속적으로 연관되어 같은 그룹에 속하는 신뢰성 정도가 α이상인 패턴들을 유사도 높은 패턴으로 본다.

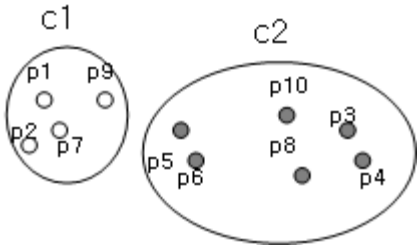
그림1은 제안하는 기법의 동작 과정을 보이는 예제이다. 사용된 데이터는 2.2절에 사용된 실험 데이터와 동일하고, 유사도가 높은 패턴들을 g1과 g2 두 개의 그룹으로 형성하였다.

데이터를 2개의 그룹으로 클러스터 할 때 최초 클러스터링 실행에서 실행 시 그림1의 (a)에서 첫 번째 클러

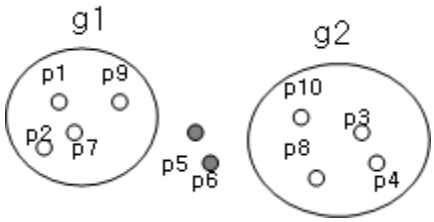
스터링 실행 시 c1은 c1={p1, p2, p3, p4, p5, p6} 소속되고 c2는 c2={p7, p8, p9}에 소속되는 결과를 알 수 있다. 그림1의 (b)에서 두 번째 실행의 c1은 c1={p1, p2, p3}, c2는 c2={p4, p5, p6, p7, p8, p9}의 결과를 알 수 있다. 그림1의 (a)(b)의 총 2회 클러스터링 결과 중 p1, p2, p3 와 p7, p8, p9는 항상 같은 그룹에 클러스터링 되는 결과를 확인 할 수 있다. 이 패턴들을 신뢰성 높은 패턴들로 보고, 그림1의 (c)에서 p1, p2, p3를 하나의 클러스터링 모델로 p7,p8,p9를 클러스터링 모델로 g1,g2를 생성하고, 두 모델의 중심을 찾아 클러스터링이 종료 되지 않은 p4, p5, p6의 유사도를 측정하여 유사도가 높은 모델로 할당되게 유도한다.



(a) k-modes 1회 클러스터 실행 결과



(b) k-modes 2회 클러스터 실행 결과



(c) 신뢰성 높은 패턴들을 초기 그룹으로 구축

그림1 제안한 알고리즘 실행 동작 예제

2.4 알고리즘 프로시저

알고리즘1 은 제안하는 앙상블 K-modes 알고리즘의 프로시저다. 알고리즘에 사용되는 변수는 다음과 같다.

X는 n개의 패턴으로 구성 된 데이터, k는 클러스터 수, e는 K-modes를 반복실행 하는 회 수, C[k]는 K-modes 실행 결과를 각 클러스터에 속하는 패턴들의 집합, A[e][k]는 K-modes 반복 실행 별 실행 결과 저장 α는 지속적으로 연관되어 같은 그룹에 속하는 신뢰성 정도, G[k]는 신뢰성 있는 패턴들로 구성된 그룹, M[k]는 G[k]의 중심, T[k]는 Xi가 A[e][k]에 나타나는 빈도 수, d(X,Y)는 2.1절의 식 (2.1) 이다.

3. 실험

3.1 실험 디자인

본 논문은 K-modes 와의 비교 실험을 통해 제안하는 알고리즘의 우수성을 보이려고 한다. 사용한 데이터는 총 3가지로, UCI Machine Learning Repository의 데이터를 사용하였다[4]. 실험 데이터는 표 3의 데이터 셋을 사용하여 실험하였다.

실험 방법은 K-modes알고리즘과 제안한 알고리즘을 대상으로 1000번의 랜덤 초기중심에 의한 실행 결과를 앙상블 하여 α 이상인 데이터를 그룹화 하여 실험하였다.(α=80) 비교 방법은 K-modes와 제안한 알고리즘의 평균 성능과 평균 편차를 실험을 통해 비교하였다.

알고리즘 1 제안하는 앙상블 K-modes

step 1. K-modes 반복 실행 결과 저장

K-modes를 e 회 실행 하며, 실행 결과를 A에 저장 :
FOR i = 1 TO e
K-modes RUN
A[e][k] = C[k]

step 2. 지속적으로 타나는 패턴들의 빈도수 검사

Xj를 A[e][k]에 속하는 패턴들과 비교 후, 나타나는 빈도수 검사 :
IF(Xi==A[e][k]) THEN T[k]=T[k]+1

step 3. 지속적으로 나타나는 패턴 그룹에 배속

Xj가 A에 속하는 빈도수 인 T[k]가 α 보다 크다면 신뢰성 있는 패턴으로 보고, 신뢰성 있는 패턴을 G[k]에 배속 :
IF(T[k]>α) THEN G[k]=Xj

step 4. 패턴 수만큼 반복 실행하여 그룹 형성

step 2~4를 패턴의 수만큼 반복 :
FOR j = 1 TO n
step 2~4

step 5. 형성된 그룹의 중심과 잔여 패턴과의 유사도 측정

G[k]의 중심을 M[k]에 저장 하고 G[k]에 속하지 않은 잔여 패턴과의 유사도 측정 :
IF (!Xj==G[k]) THEN d(Gk,Xj)

step 6. 알고리즘 종료

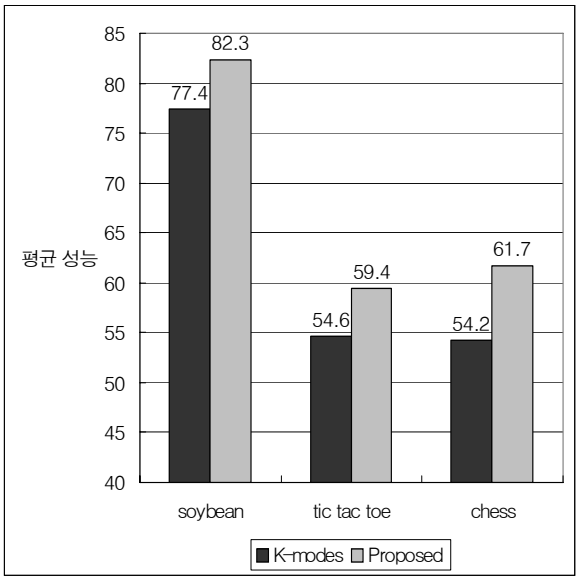
측정된 유사도가 높은 그룹으로 잔여 패턴 배속 후 종료

표3 실험에 사용된 데이터 셋 정보

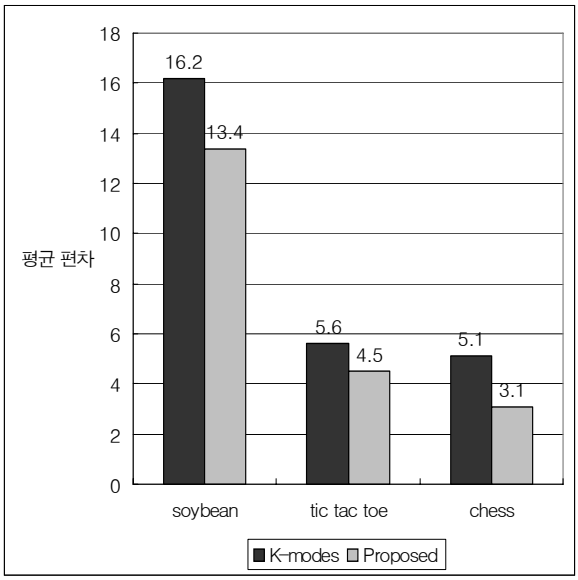
데이터 셋	패 턴	속 성	그 룹
soybean	47	35	4
tic tac toe	958	9	2
chess	3,196	36	2

3.2 결과 및 분석

그림 2를 통해 각 데이터에 대한 정확도와 안정성 실험 결과를 보이고 있다. 그림2의 (a)와(b) soybeans 데이터를 통해 K-modes는 77.4% 성능과 편차 16.2를 보였고, 제안한 알고리즘은 82.3%성능과 편차 13.4를 보여 4.9%의 성능 향상 과 2.8의 편차 감소를 확인 할 수 있었다. soybean데이터에 비해 tic tac toe 데이터와 chess 데이터는 K-modes의 편차가 5.6과 5.1로 낮았



(a)데이터 별 평균 성능



(b)데이터 별 평균 편차
그림2 K-modes와 제안한 알고리즘 1000회 실험 결과

음에도 4.5와 3.1로 편차가 감소함을 보여 안정화가 되었고, 성능 향상은 tic tac toe 데이터 K-modes 54.6%에서 제안한 알고리즘은 59.4%로 4.8%의 성능 향상되었고, chess 데이터 K-modes 54.2%에서 제안한 알고리즘은 61.7%로 7.5%의 성능 향상을 가져 왔다.

K-modes의 soybean 데이터 실험 결과 성능의 최고 100%와 최저 55%를 보였는데, 성능이 낮았던 실행의 초기 중심들과 최고 성능 실행의 초기 중심과 분석하였다. 분석 결과 초기 중심들의 유사도가 높을 경우 패턴들의 중심과의 거리가 비슷해서, 클러스터링 결과의 평균과의 편차가 높은 것을 확인 할 수 있었다. 제안한 알고리즘은 신뢰성 높은 패턴들을 모델로 구축함으로써 중심들의 유사도가, K-modes의 랜덤 한 초기 중심들의 유사도 보다 신뢰성 있다는 가정이 검증 되어 K-modes 보다 성능의 향상과 안정화가 가능하다는 분석을 하였다. 또 속성이 많은 데이터인 soybean과 chess에서 더 높은 성능 향상을 됨을 실험을 통해서 알 수 있다.

4. 결론 및 향후 연구 방향

본 논문은 K-modes의 초기 중심에 의한 성능 편차 문제를 앙상블 알고리즘의 적용을 통하여 개선하였다. 제안한 알고리즘은 α를 80으로 설정하여, 같은 그룹에 속하는 신뢰성 높은 패턴들을 α 이상의 패턴들에 대하여 수렴하여 실험하였다. 하지만 데이터 마다 최적의 수렴치가 같지 않는 문제점을 알 수 있었고, 또 클래스 분류가 작은 데이터에 대한 성능 편차의 안정화 역시 개선할 단점으로 확인 되었다. 향후 유사도가 낮은 초기 중심의 수행 결과를 앙상블하여 수렴한다면 더 높은 성능을 유도할 수 있을 것이라고 예측한다.

참 고 문 헌

[1] J. B. MacQueen, "Some Methods for classificatio n and Analysis of Multivariate Observations, Proc eedings of 5-th Berkeley Symposium on Mathem atical Statistics and Probability", Berkeley, Univer sity of California Press, 1:281-297
[2] Z. Huang, "Extensions to the K-modes algorithm for clustering large data sets with categorical values", Data Mining Knowledge
[3] E. Dimitriadou , “Voting-merging: An ensemble method for clustering”, Lecture Notes in Comput er Science
[4] http://archive.ics.uci.edu/ml/index.html