고속도로 통행료 수납자료를 이용한 주행특성 클러스터링 기법

김동근*, 박원식*, 양영규*
"경원대학교 전자계산학과"
e-mail: cocacu@naver.com, pws1981@paran.com, ykyang@kyungwon.ac.kr

Driving Characteristics Clustering use TCS Data

Dong-Keun Kim*, Won-Sik Park*, Young-Kyu Yang*
*Dept of Computer Science, Kyungwon University

요 약

고속도로의 다양한 주행특성으로는 과속하는 차량, 휴게소나 기타목적의 이용차량, 운전자의 습관이나 피로도등이 있는데 이에 따라 고속도로 주행시간에 차이가 나타난다. 하지만 현재에는 이러한 특성을 고려하지 않고 통행시간 분류가 되고 있어 정확성과 신뢰성을 보장하지 못하고 있는 실정이다.

이에 본 연구에서는 데이터 분포에 따른 해석을 통하여 TCS데이터의 특성을 고려 할 수 있는 Fuzzy c-means 알고리즘과 단순히 임의의 초기값으로 분류하는 K-means와의 비교를 통해서 주행특성을 고려한 클러스터링 기법이 경우에 따라서 더 효과적이고 신뢰성 있는 분류방법이 될 수 있음을 증명하였다.

1. 서론

PDP, 모바일 기기등 휴대용 교통정보 수단의 보급화, 교통정보에 대한 사회적 요구 증가, 교통혼잡비용의 기하급수적인 증가 등으로 인하여 그 어느때 보다 교통정보의 중요성이 커지고 있다. 동시에, 정확하고 신뢰성 있는 교통정보 제공의 필요성 역시 증가하고 있는 실정이다.

최근 연구가 활발히 진행되고 있는 고속도로 통행료수납시스템(TCS: Toll Collection System)자료는 주행차량이 경험한 통행시간을 포함하고 있어 통행시간 예측에 아주 유용하다. 하지만, 기존의 TCS데이터를 이용한 통행시간 클러스터링 방법은 최초 주어진 값에 따라 주어진 데이터가 고르지 않을 경우 오차가 커지는 문제점이 있다. 따라서, 정확하고 효율적으로 TCS데이터를 분류하기 위하여 효과적인 클러스터링 방법을 적용할 필요가 있다. 그런면에 있어서 퍼지이론은 도로상황의 정체 및 운전자의 습관 등 여러 고속도로 주행특성들의 복잡성 등을 고려할수 있는 해결책이 될 수 있다.

이에 본 연구에서는 클러스터링의 대상이 되는 데이터 분포에 따른 해석을 통하여 초기값을 설정함으로써 주행특성을 고려할수있는 Fuzzy c-means를 단순히 임의로 하나의 초기값을 지정해 주는 전통적인 군집화 알고리즘인 K-means 알고리즘과 비교/평가해 보았다.

2. 관련연구

본 연구에서 적용한 데이터 분포에 따른 해석을 통하여 초기값을 설정하는 Fuzzy c-means 알고리즘과 전통적인 군집화 알고리즘 방법인 K-means 알고리즘, 그리고 TCS 데이터와 TCS 데이터의 문제점을 해결하기위한 전진반복 전후방탐색법에 대하여 간단히 기술하며 다음과 같다.

2.1 TCS 데이터

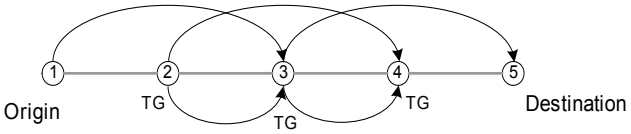
TCS 데이터는 전국 262개 영업소에서 매일 평균 310만건의 자료가 수집된다. 경로통행시간 추출을 위해 사용되는 TCS 데이터의 원시자료는 그림 1에서와 같이 출발TG번호(FROMTOLL_ID),도착TG번호(TOTOLL_ID),출발시간(START_DATE),도착시간(END_DATE),차종(CAR_TYPE)으로 구성되어있다.

FROMTOLL_ID	TOTOLL_ID	START_DATE	END_DATE	CAR_TYPE
179	174	200708132352	200708140002	1
179	174	200708132355	200708140002	1
179	174	200708132356	200708140003	1
179	174	200708132353	200708140003	1
217	174	200708132322	200708140003	1
217	174	200708132322	200708140003	1
179	174	200708132354	200708140004	6
216	174	200708132336	200708140004	1

<그림 1> TCS 데이터 구성요소

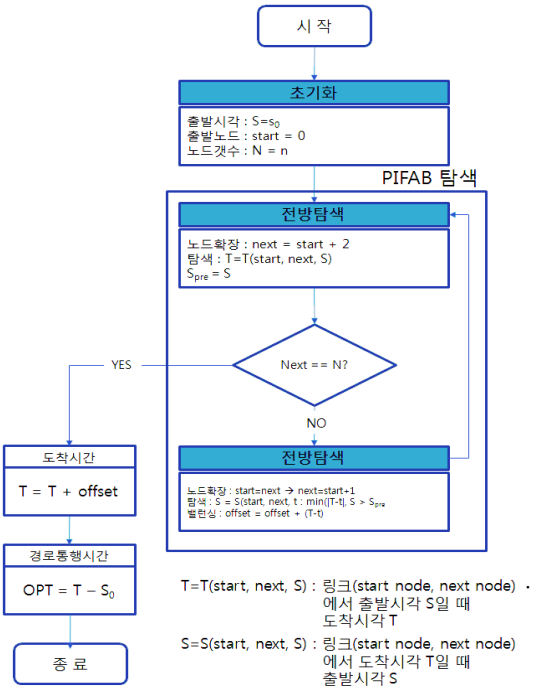
2.2 전진반복 전후방탐색법(PIFAB: Progressive Iterative Forward And Backward search)

TCS 통행시간으로부터 본선외 주행시간을 제외한 경로통행시간을 산출하는 기능을 하는 PIFAB 탐색법은 그림 2와 같이 기본 3개구간을 단위로 하여 TCS통행시간 출발/도착시간을 전진반복 전후방탐색한다. 이 방법의 장점으로 통행시간의 램프이용시간을 제거하고 본선 및 진출입 램프 정체시에도 본선 주행시간만을 추출한다는 점이 있다.



<그림 2> 전진반복 전후방 탐색법

PIFAB 탐색법의 알고리즘은 그림 3과 같다.



<그림 3> 전진반복 전후방 탐색법 알고리즘

2.2 K-means 알고리즘

전체적인 알고리즘은 초기화단계, 개체분산단계, 새로운 클러스터의 중심단계로 나누어 볼 수 있는데, 각 단계의 역할과 수렴성에 관한 사항을 간략히 알아본다.

첫째, 초기화 단계에서는 생성할 클러스터의 개수 k를 정하고, 각 클러스터에 대한 초기값을 설정하는데 특별한 조건 없이 전체 데이터 중에서 식 1과 같이 임의로 선택한다.

$$\{z_1, z_2, \ldots, z_k\} \subseteq s_i \; i = 1,2,\ldots,N \quad \text{(식 1)}$$

둘째, 개체분산 단계에서는 각 개체들과 각 클러스터의 중심과의 유클리디안 거리(J)를 식 2와 같이 구하고, 이때 개체들은 계산된 거리가 식 3과 같이 가장 최소가 되는 클러스터($C_i,\ i=1,2,\ldots k$)에 속하게 된다. 식 4에서 i와 m은 각각의 클러스터를 의미한다.

$$J_{il} = \| x_i - z_i \|^2 \text{ for } i = 1,2,\ldots,N,\ I = 1,2,\ldots,K$$

(식 2)

$$\text{if } J_{il} < J_{im} \text{ for } I,m = 1,2,\ldots,K \;,\; I \neq m \text{ then } x_i \in C_i$$

(식 3)

여기서 계산된 거리는 개체간의 유사성과 비유사성을 나타낸다. 개체들 간의 거리는 일반적으로 유클리디안 거리측정 방법을 사용한다.

$$z_i(\text{new}) = \frac{1}{N_i} \sum (x_i \in C_i) \; i = 1,\ldots,N, I = 1,\ldots,K$$

(식 4)

여기서, N_i는 각 클러스터에 새롭게 구성된 총 개체의 수를 나타내고, $x_i \in C_i$는 i번째 클러스터에 속한 개체들을 의미한다. 이러한 클러스터의 중심값 z_i(new)이 반복적으로 갱신되는데 그러한 반복에 대한 횟수와 전체 수렴성에 대한 조건이 최종 알고리즘의 결과를 좌우하게 된다.

K-means 알고리즘의 수렴여부에 관해서는 식 4와 같이 더 이상 각 클러스터의 중심에 변화가 생기지 않을 때 종료되는데, 만일 클러스터의 중심에 변화가 생겼다면 두 번째 단계로 피드백(feedback)되어 반복된다.

2.3 Fuzzy c-means 알고리즘

Fuzzy c-means 알고리즘은 먼저 초기화가 이루어지고 이로부터 클러스터의 중심을 계산하여 얻은 중심과 각 개체 사이의 유클라디안 거리를 구한 다음, 새로운 분할 행렬을 갱신한다. 이후 이와같은 갱신과정을 반복하면서, 데이터 특성에 맞는 중심점에 수렴하게 된다. Dunn in에 의해 제안된 이 알고리즘은 식 5와 같이 목적함수를 최소화하도록 하는 반복적인 알고리즘으로서 주어진 데이터로부터 유사한 클러스터를 나누며 생성된 클러스터는 시스템의 특성적인 동작을 기술하는 규칙으로서 사용되어진다.

$$J = \sum_{i-1}^{N} \sum_{j-1}^{C} u_{ij}^{m} \| x_i - c_j \|^2 \quad \text{(식 5)}$$

Fuzzy C-means 알고리즘은 다음과 같이 요약된다.

STEP 1 : 초기값 U=[u_{ij}] matrix, $U^{(0)}$

STEP 2 : 중심값 계산 $C^{(k)}=[C_j]$ with $U^{(k)}$

$$c_i = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \quad \text{(식 6)}$$

STEP 3 : $U^{(k)}$ 업데이트, $U^{(k+1)}$

$$U_{ij} = \frac{1}{\sum_{k=1}^{c} \left\{ \frac{\| x_j - c_k \|}{\| x_i - c_k \|} \right\}^{\frac{2}{m-1}}} \quad \text{(식 7)}$$

STEP 4 : if $\| U^{(k+1)} - U^{(k)} < \delta \|$ then STOP;
다른경우엔 스텝2부터 다시 반복한다.

3. 비교평가

본 연구에서는 주행특성을 반영할수 있는 Fuzzy c-means 클러스터링 방법과 K-means 클러스터링 방법에 PIFAB를 적용, 7월 11일, 13일 서울-청주구간의 통행시간을 추출하여 비교한다.

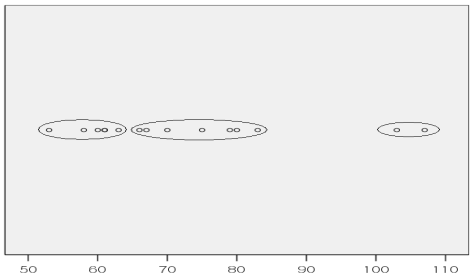
표 1은 Fuzzy c-means와 K-menas의 7월 11일 14:00시에 경우 통행시간(분)이 Fuzzy c-means는 61.33분, K-means는 61.33분으로 동일한 결과를 보였고 표준편차 역시 동일한 결과를 나타내었다.

<표 1> 7월 11일 14:00 서울-청주 14:00 Fuzzy c-means와 K-means 비교 결과

분류방법	구간	시간	그룹	댓수	통행시간	표준편차
Fuzzy c-means	서울-청주	14:00	1	8	61.33	4.45
			2	5	77.4	5.03
			3	2	105	2.83
K-means	서울-청주	14:00	1	8	61.33	4.45
			2	5	77.4	5.03
			3	2	105	2.83

그림 4는 Fuzzy c-means 알고리즘을 이용하여 7월 11일 14:00시 서울-청주 구간의 클러스터링 분류 결과이다.

X축은 통행시간을 나타내며 결과 그림에서 작은 원은 통행시간별 차량그룹이고, 큰 원은 이를 다시 주행특성에 따라 Fuzzy c-means의 분류방법으로 분류한 그룹을 나타낸다. K-means와 Fuzzy c-means의 결과가 동일하게 나타났으며, 통행시간의 분포가 큰 이상치를 포함하지 않고 균일하게 분포되는 경우 두 가지 클러스터링 방법의 결과는 같음을 보인다.



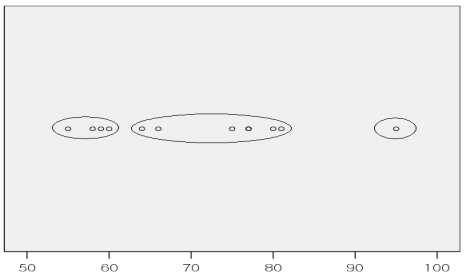
<그림 4> 7월 11일 서울-청주 14:00 Fuzzy c-means 분류결과

표 2는 7월 13일 서울-청주 14:00시 Fuzzy c-means와 K-means의 분류한 결과로써 그룹 1에서 Fuzzy c-means의 표준편차는 2.16, K-means의 표준편차는 4.03, Fuzzy c-means 그룹 2의 표준편차는 1.41, K-means 그룹 2의 표준편차는 2.44로 각 개체의 유사성과 연결성이 K-means의 방법보다 Fuzzy c-means를 이용한 분류방법이 더 정확함을 보여준다.

<표 2> 7월 13일 서울-청주 14:00 Fuzzy c-means와 K-means 비교 결과

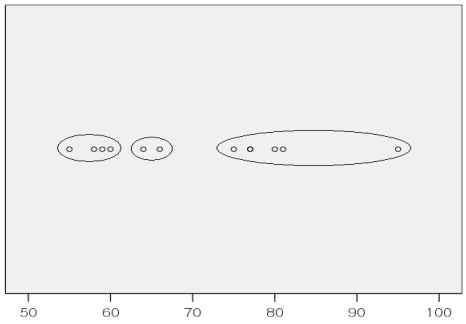
분류방법	구간	시간	그룹	댓수	통행시간	표준편차
Fuzzy c-means	서울-청주	14:00	1	4	58	2.16
			2	2	65	1.41
			3	6	80.8	7.27
K-means	서울-청주	14:00	1	6	60.3	4.03
			2	5	78	2.44
			3	1	95	0

그림 5는 K-means를 이용 7월 13일 서울-청주 구간을 분류한 결과이다. K-means의 분류 결과 특징을 살펴보면 다음과 같다. 그룹 내에 통행시간 간 높은 유사성을 보여주지만 다른 그룹과의 유사성은 떨어지고, 또한 그룹을 1개의 값으로 설정하여 효과적으로 분류하지 못하였다.



<그림 5> 7월 13일 서울-청주 14:00 K-means 분류결과

그림 6은 Fuzzy c-means를 이용 7월 13일 서울-청주 구간을 분류한 결과이다. 특징을 살펴보면 같은 클러스터링 내의 통행시간끼리는 높은 유사성을 보여주는 점은 K-means와 같지만 각 그룹과의 차이가 적고, 또한 표준편차도 작아 더 정확한 분류방법이라 할 수 있다.



<그림 6> 7월 13일 서울-청주 14:00 Fuzzy c-means 분류결과

표 3은 K-means방법과 본 연구에서 적용한 Fuzzy c-means를 이용하여 경로통행시간과 통행시간을 뺀 오차를 보여준다. 오차의 범위를 보면 K-means는 0.43~3.85분, Fuzzy c-means는 0.38~3.85분을 나타낸다. 13:00시 경우 K-means 오차는 2.72분, Fuzzy c-means 오차는 0.53분, 14:00시 K-means 오차는 2.22분, Fuzzy c-means 오차는 0.38분으로 Fuzzy c-means의 방법이 클러스터링 방법 중 하나인 K-means 보다 더 나은 결과를 도출하였고, 주행특성을 반영한 클러스터링 방법이 경우에 따라서 더 효율적일수 있음을 보였다.

<표 3> k-menas와 c-means의 오차 비교 결과

시간대	K-means오차(분)	Fuzzy c-means오차(분)
13:00	2.72	0.53
13:10	1.49	1.49
13:20	2.45	2.45
13:30	1.04	1.04
13:40	1.81	1.81
13:50	0.43	0.43
14:00	2.22	0.38
14:10	2.63	2.63
14:20	2.06	2.06
14:30	2.32	2.32
14:40	5.39	5.39
14:50	3.87	3.87
15:00	2.06	2.06
15:10	0.84	0.84
15:20	0.56	0.56
15:30	1.38	0.46
15:40	2.52	0.24
15:50	3.85	0.52

4. 결론

주행자의 피로도나 주행목적등 고속도로의 다양한 특성을 고려하지 않는 클러스터링 방법은 TCS데이터의 분포에 따라서 고속도로 주행시간에 오차가 커진다는 문제점이 발생하고 있다. 이러한 통행시간 분류방법의 문제점을 해결할수 있는 방안으로서 고속도로 주행특성을 고려할수 있는 클러스터링 방법 중 Fuzzy c-means를 K-means와 비교하여 실험하였다. Fuzzy c-means 알고리즘은 데이터의 집합을 각각의 개체들이 클러스터에 소속하는 정도를 분할 행렬로 표시하고, 초기 중심을 전체 데이터 분포의 중간 정도에 설정하는 특징이 있어 주행특성을 반영하기에 적합하다 할수 있겠다.

일반적으로 널리 사용되고 있는 단일 대표값 군집화 클러스터링 방법중 하나인 K-means 알고리즘과 Fuzzy c-means 알고리즘을 비교 평가하기 위해 경로통행시간과 통행시간을 뺀 오차의 범위를 구해본 결과 K-means는 0.43~3.85분, Fuzzy c-means는 0.38~3.85분을 나타내었다.

이는 결론적으로 고속도로의 다양한 특성을 고려하여 보다 정확한 분류결과를 추출하기 위해 클러스터링의 대상이 되는 데이터 분포에 따른 해석을 통하여 초기값을 설정하는 Fuzzy c-means 방법이 기존의 주행특성을 고려하지 않고 단순히 관측된 데이터에서 임의로 하나의 초기값을 지정해주는 클러스터링 방법보다 경우에 따라서 더 효과적이고 신뢰성있는 방법이 될수 있음을 증명하였다.

감사의 글

본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로수행되었음.
(IITA-2009-C1090-0902-0040)

본 연구는 한국도로공사 OASIS(Operations Analysis and Supportive Information System)의 자료지원을 받아 수행되었습니다.

참고문헌

[1] 남궁성, “고속도로 경로통행시간 산출을 위한 전진반복 전후방탐색법(PIFAB)의 개발,” 대한교통학회, 대한교통학회지, 제23권 제5호, pp. 147-155, 2005.

[2] 박원식, 최진우, 양영규, “고속도로 통행료수납자료를 이용한 통행시간 군집현상에 관한 연구,” 한국GIS학회, 한국GIS학회춘계학술대회연구집, pp 195-210, 2008.

[3] 강지혜, 김성주, “적응적인 초기치 설정을 이용한 Fast K-means 및 Fuzzy-c-means 알고리즘”, 한국정보과학회, 정보과학회연구지, 제 31권 제4호, pp. 516-524, 2004.

[4] 한학용, “패턴인식 개론”, 한빛미디어, 2005.

[5] 원태연, 정성원, “통계조사분석”, SPSS아카데미, 2004.

[6] 정영근, 박창호, “퍼지 추론을 이용한 최단 경로 탐색 알고리즘 개발”, 대한교통학회지, 제23권 제 8호 pp 171-179, 2005.

[7] 김인재, 오성권, "Type-2 퍼지집합 기반의 퍼지 C-Means 클러스터링의 설계“, Proceedings of KIIS Conference 2008, Vol. 18 No2, 2008.