퍼지 클러스터링에 의한 사례기반 추론에 관한 연구

현우석°
한국성서대학교 정보과학부
wshyun°@bible.ac.kr

A Study on Case-based Reasoning using Fuzzy Clustering

Woo-Seok Hyun°
Dept. of Information and Science, Korean Bible University

요 약

주어진 현재의 문제를 해결하기 위해서 과거에 유사하게 수행된 사례를 유추하여 유추된 사례의 해를 이용하는 사례 기반 추론(case-base reasoning)은 여러 분야에 응용되고 있지만, 사례기반 추론 시 새로운 사례를 해결하기 위하여 사례베이스 내의 모든 사례를 검색해야 하기 때문에 수행시간이 증가되는 단점을 지니고 있다. 본 연구에서는 하드 클러스터링 방법으로 완전하게 분류하는 것이 불가능할 수도 있다는 문제점을 개선시키기 위하여 퍼지 클러스터링을 방법을 이용하여 사례베이스를 분류함에 의하여 시스템의 수행시간을 감소시키면서 정확성을 높이게 되었다.

1. 서 론

의사들이 환자들로부터의 다양한 정보들을 가지고 진단을 내리는 과정은 여러 가지 논리적 단계를 거치게 되며 상당히 복잡한 의사결정을 요구하게 된다. 퍼지 논리를 이용한 급성복통과 관련된 질환 진단시스템[1]은 퍼지논리를 이용한 규칙기반 시스템으로서 진단에 필요한 지식을 정형화된 규칙만으로 표현하는데 어려움이 있으며, 시스템의 성능 향상을 위해 규칙을 계속적으로 수정하고 추가해야 하며, 예외적인 상황에서 진단 시 문제점을 지니고 있다. 이런 문제점을 해결하고자 사례기반 추론[2-4]에 의해 확장된 CDS-DAAP(Combined Diagnosis System for Diseases associated with Acute Abdominal Pain)[5]가 제안되었으나, 사례기반 추론시 사례베이스의 크기가 증가하게 되면 사례베이스 안의 모든 사례들을 검색해야 하기 때문에 검색시간이 증가하게 되는 문제점이 발생하게 된다. 이것을 개선하고자 K-Means 클러스터링 알고리즘에 의한 사례기반 추론을 이용한 ADS-DAAP(Advanced Diagnosis System for Diseases associated with Acute Abdominal Pain)[6]이 제안되었다. K-Means 클러스터링과 같은 하드 클러스터링 방법은 분명한 경계선을 가지지만 경계 부근에 있는 환자의 증상의 사례를 어느 하나의 클러스터로 완전하게 분류하는 것이 불가능할 수도 있다는 문제점이 있다.

본 연구에서는 퍼지 클러스터링 방법 중의 하나인 Fuzzy C-Means 알고리즘을 이용하여 클러스터에 속하는 정도를 [0,1]로 확장하여 경계 부근의 사례들을 한 개 이상의 클러스터에 속할 수 있게 허용하는 방법을 사용한 ADS-DAAP-FC(Advanced Diagnosis System for Diseases associated with Acute Abdominal Pain using Fuzzy Clustering)를 제안한다. 제안하는 시스템은 기존의 ADS-DAAP와 비교해 보았을 때 시스템의 수행시간을 감소시키면서도 정확성을 높이게 되었다.

본 연구는 2002학년도 한국성서대학교 교내 연구비 지원으로 수행되었습니다.

2. 퍼지 클러스터링

본 논문에서는 예외상황 사례베이스에 들어있는 사례들을 클러스터링하기 위하여 Bezdek의 FCM(Fuzzy C-Means)[7-9] 알고리즘을 사용하였으며, 다음과 같다.

사례베이스를 c개의 클러스터로 분류할 때 각 클러스터의 중심 벡터 $v_i(i=1,2, \ldots ,c)$와 데이터 x_k와의 비유사도(dissimilarity) d_{ik}는 유클리디언(Euclidean) 거리로 구하며 식 (1)과 같다.

$$d_{ik} = d(x_k, v_i) = \|x_k - v_i\| = \left\{\sum_{j=1}^{p}(x_{kj} - v_i)^2\right\}^{(1/2)} \quad (1)$$

FCM 알고리즘은 식 (2)의 목적함수를 최소화하는데 목적이 있다.

$$J(U,v) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m \|x_k - v_i\|^2 \quad (2)$$

이 때, 클러스터 중심(v_i)과 소속도 함수 값(u_{ik})은 식 (3)과 (4)를 이용하여 구하며 다음과 같다.

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m}, 1 \le i \le c \quad (3)$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}} \quad (4)$$

FCM 알고리즘은 다음과 같이 4 단계로 표현할 수 있다.

1단계: 클러스터의 수 c와 지수가중치 m 값을 선정하고, 퍼지 c분할행렬 ($U^{(l)}$)을 초기화한다.

$2 \le c \le n$

$1 < m < \infty$

2단계: 1단계에서 구한 ($U^{(l)}$)과 식(3)을 이용하여 새로운 클러스터의 중심 $v_i^{(l)}$(i=1,2,...,c)을 구한다.

3단계: $x_k \neq v_i^{(l)}$ 일 때 $U^{(l)}$를 식 (4)에 의해$U^{(l+1)}$로 업데이트하고, 그 외에는

$u_{ik}^{(l+1)} = \begin{cases} 1 & i=j \\ 0 & i \neq j \end{cases}$ 로 둔다.

4단계:

$\|U^{(l+1)} - U^{(l)}\|_G \le \varepsilon$ 의 조건이 만족되면 클러스터링을 종결하고 그렇지 않으면 l을 1 증가시킨 후 2단계로 되돌아가서 반복 수행한다.

3. ADS-DAAP-FC(Advanced Diagnosis System for Diseases associated with Acute Abdominal Pain using Fuzzy Clustering)

제안하는 ADS-DAAP-FC에서는 예외상황 사례베이스를 클러스터링 함에 있어서 Fuzzy C-Means 알고리즘을 사용하였다. K-Means 클러스터링과 같은 하드 클러스터링 방법은 분명한 경계선을 가지지만 경계 부근에 있는 환자의 증상의 사례를 어느 하나의 클러스터로 완전하게 분류하는 것이 불가능할 수도 있다는 문제점을 개선시키기 위하여 Fuzzy C-Means 알고리즘을 이용하여 클러스터에 속하는 정도를 [0,1]로 확장하여 경계 부근의 사례들을 한 개 이상의 클러스터에 속할 수 있게 허용하는 방법을 사용하여 수행시간을 감소시키면서 정확성을 높이게 되었다.

제안하는 ADS-DAAP-FC의 구조는 그림 1과 같으며, 본 시스템의 진단 과정은 그림 2와 같이 먼저 환자의 데이터가 입력되어 규칙으로 표현된 진단제어 지식베이스를 기반으로 진단을 수행하고, 진단에 실패한 경우 Fuzzy C-Means 클러스터링 알고리즘에 의한 예외상황 사례베이스를 기반으로 재진단을 시도하게 되어 조회시간을 감소시키면서 정확성을 높이게 되었다.

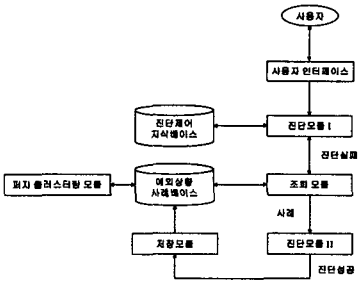

그림 1 ADS-DAAP-FS의 구조

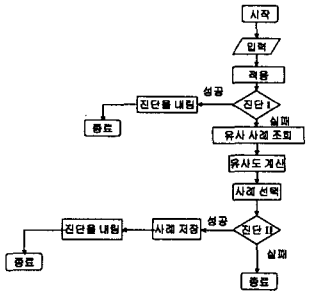

그림 2 질환 진단 과정

3.1 퍼지 클러스터링 모듈

예외상황 사례베이스에 들어 있는 사례들을 2절의 FCM[7-9] 알고리즘을 이용하여 퍼지 클러스터링한다. 퍼지 클러스터링 모듈이 실행되기 위해서는 '수술에 관한 위험 정도', '노화 정도', '탈수 정도', '장폐색 정도', '구토 정도', '통증의 완화 정도', '만성 복통 정도', '복벽의 통증 정도', '상복부의 통증 정도', '하복부의 통증 정도', '직장 검사 수지 정도' 등의 속성이 필요하다. 모듈의 실행이 끝나게 되면 그 결과로 c개의 클러스터와 클러스터별로 v_i개의 중심이 도출된다. 이 결과는 클러스터별로 인덱싱 되어 예외상황 사례베이스에 저장되며 이 정보는 조회모듈에서 사례들을 조회할 때 사용된다. 본 연구에서 c 값은 4, v_i 값은 1로 정했다.

3.2 예외상황 사례베이스 모듈

진단제어 지식베이스에 들어 있는 기존 규칙으로 처리할 수 없는 예외적인 급성 복통 진단을 위한 지식은 사례베이스에 사례로서 저장된다. 표 1은 초기 사례베이스의 일부분을 보여 준다.

표 1 ADS-DAAP-FS 의 사례베이스

수술에 관한 위험 정도	노화 정도	탈수 정도	장 폐 색 정도	구토 정도	통 증 의 완화 정도	유사도	참조 횟수
0.61	0.57	0.49	0.72	0.59	0.67		0
0.59	0.41	0.53	0.76	0.65	0.47		0
0.54	0.69	0.39	0.78	0.53	0.34		0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3.3 조회 모듈

진단 모듈 I에서 진단에 실패한 경우, 예외상황 사례베이스에 있는 사례들을 조회하여 현 상황과 유사한 사례를 찾아낸다. 유사한 사례를 찾기 위해서는 유사도를 사용하게 되는데, 식 (5)와 같다.

$$Similarity(case_i, case_j) = 1 - \frac{\sum_{l=1}^{n} |attr_{il} - attr_{jl}|}{|n|} \quad (5)$$

l : 속성 수

$case_i$: 현재 사례

$case_j$: 과거 사례

$attr_{il}$: 현재 사례를 구성하는 l번째 속성을 나타내는 퍼지값 $(1 \le l \le n)$

$attr_{jl}$: 과거 사례를 구성하는 l번째 속성을 나타내는 퍼지값 $(1 \le l \le n)$

4. ADS-DAAP-FS의 성능 평가

시뮬레이션 환경에서 제안하는 시스템의 성능을 평가하기 위해서 G 병원으로부터 획득한 200명의 실제 환자 데이터를 기존의 ADS-DAAP와 제안하는 ADS-DAAP-FS에서 각각 진단하여 평균 수행시간과 평균 진단 실패율을 비교하였는데 그 결과는 그림 3, 4와 같다.

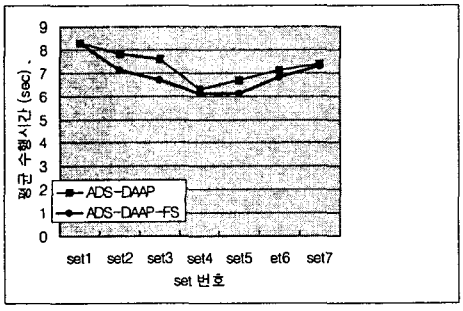

그림 4 시스템에 따른 평균 수행시간

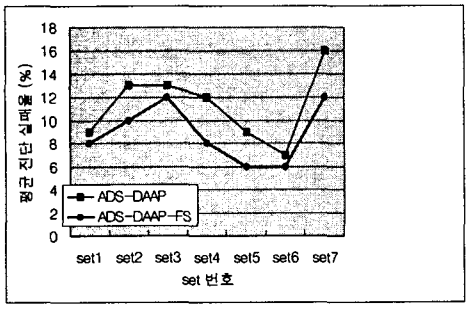

그림 5 시스템에 따른 평균 진단 실패율

5. 결론 및 향후 과제

제안하는 퍼지 클러스터링에 의한 ADS-DAAP-FS는 기존의 하드 클러스터링의 하나인 K-Means 클러스터링을 사용한 ADS-DAAP와 비교해 볼 때 평균 수행시간을 감소시켰으며, 평균 진단 실패율을 감소시켜 정확성을 높이게 되었다. 이것은 Fuzzy C-Means 알고리즘이 K-Means 클러스터링과 같은 하드 클러스터링 방법에서 경계 부근에 있는 환자의 증상의 사례를 어느 하나의 클러스터로 완전하게 분류하는 것이 불가능할 수도 있다는 문제점을 해결하였기 때문이다.

향후 연구과제로는 퍼지 클러스터링에서 클러스터 수나 중심의 개수 선정에 따라 결과가 다르게 나타날 수 있으므로 최적의 클러스터나 중심의 개수를 선정하는 연구가 남아있다.

참고문헌

[1] 현우석, "퍼지논리를 이용한 급성복통과 관련된 질환 진단시스템의 설계," 한국퍼지및지능시스템학회 2002 춘계학술발표논문집, 제 12권, 제 1호, pp.68-71, May, 2002.

[2] M. P. Feret and J. I. Glasgow, "Hybrid Case-Based Reasoning for the Diagnois of Complex Devices", *Proc, of the National Conf. on Artificial Intelligence(AAAI-93)*, pp.168-175, 1993.

[3] J. L. Kolodner, "Improving human decision making through Case-base decision aiding", *AI Magazine*, Vol.12, No.2, pp.52-68, 1991.

[4] R. Barletta, "Case-based reasoning and information retrival: Opportunities for technology sharing", *IEEE Expert*, Vol.8, No.6, pp.2-3, 1993.

[5] 현우석, "급성복통 진단을 위한 규칙 및 사례기반 추론의 통합," 한국퍼지및지능시스템학회 2002 추계학술발표논문집, 제 12권, 제 2호, pp.459-462, Dec., 2002.

[6] 현우석, "K-Means 클러스터링 알고리즘을 이용한 사례기반 추론에 관한 연구," 한국정보처리학회 2003 춘계학술발표논문집, 제 10권, 제 1호, pp.341-344, May, 2003.

[7] R. L. Cannon, J. V. Dave and J. C. Bezdek, Efficient Implementation of the Fuzzy C-Means Clustering Algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.8, No.2, 1986, pp.248-255.

[8] N. Pal and J. Bezdek, "On Cluster Validity for the Fuzzy C-Means Model," IEEE Trans. on Fuzzy Systems, Vol. 3, No. 3, pp.370-379, 1995.

[9] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.