CBR을 위한 FCM 기반 퍼지 소속 함수 결정 방법

연지현*, 김은주, 이일병
연세대학교 컴퓨터과학과

Decision Method of Fuzzy Membership Function based on FCM for CBR

Ji-Hyun Yeon*, Eun-Ju Kim, Yill-Byung Lee
Dept.of Computer Science, Yonsei University

요 약

사례 기반 추론(Case-Based Reasoning)은 새로운 문제를 해결하기 위해 유사한 기존 문제를 추출하여 그 해결과정을 사용한다. 그러므로, 기존의 문제와의 유사성을 얼마만큼 잘 판별하는가가 매우 중요한 관건이다.

연구된 유사성 판단 방법으로는 퍼지 소속 함수(Fuzzy membership function)를 이용하여 사례마다 각 클래스에 대한 소속 함수 값을 주는 방법이 있다. 이 방법은 퍼지 소속 함수를 어떻게 주는가에 따라 성능이 달라진다. 본 논문에서는 적당한 퍼지 소속 함수를 주기 위하여 Fuzzy C-Means를 사용하는 방법을 제안하였다. 이 방법은 각 클래스에 대한 소속 함수 값을 결정하는데 있어서 좀 더 전체적인 데이터 분포 정보를 이용할 수 있다.

1. 서 론[1)]

사례 기반 추론(CBR)이란 새로운 문제를 해결하기 위해 기존에 해결된 문제와의 유사성을 이용하는 방법이다. 즉, 새로운 문제에 대해 이 문제와 유사한 기존의 사례를 찾아 기존 사례의 해결과정을 새로운 문제의 해결과정에 맞도록 변형하여 해결한다. 또한 이 문제를 하나의 사례로 저장함으로써 미래에 이와 유사한 문제가 발생했을 때 이를 이용하여 해결할 수 있다.

사례 기반 추론 기법은 유사성 판단 기법에 많이 좌우되므로, 이 기법의 연구가 많이 이루어지고 있다. 연구된 유사성 판단 기법으로 퍼지 소속 함수를 이용한 기법이 있다. 이 방법은 각 사례가 완전히 한 클래스에만 속한다고 전제하던 기존의 방법에서 탈피하여 각 클래스마다 속하는 소속 정도를 고려한 방법이다. 이 방법에서는 어떻게 퍼지 소속 함수를 줄 것인가가 중요한 문제이다.

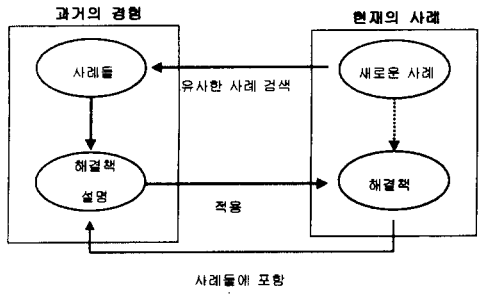
본 논문에서는 퍼지 소속 함수를 주는 방법으로 데이터의 전체적인 분포를 이용할 수 있는 Fuzzy C-Means(FCM) clustering을 제안하였다. 본 논문의 구성은 다음과 같다. 2장에서 사례 기반 추론의 개념을 설명하고, 3장에서 퍼지 소속 함수를 사용한 경우를 설명하고, 4장에서 본 논문에서 제안한 FCM을 이용한 사례 기반 추론을 설명한다. 5장에서는 제안된 방법을 평가하기 위한 실험에 관한 내용이고, 6장에서 결론을 내린다.

2. 사례 기반 추론

사례기반 추론 처리과정은 크게 사례의 표현, 사례 추출, 사례 적용, 사례 학습 등으로 구분할 수 있다. [6]

첫 번째 단계인 사례 표현은 전문가의 경험인 사례를 컴퓨터에 적절하게 표현하고자 하는 것이다. 대개의 경우 간단하게 특정한 결과를 이끌어내는 특성들의 리스트로 표현한다. 두 번째 단계인 사례 추출 단계에서는 사례 베이스로부터 새로운 사례와 유사한 사례를 효율적으로 찾아내는 단계이다. 이 때에는 매칭 규칙을 사용하게 되는데, 이 매칭 규칙을 어느 것을 사용하느냐에 따라서 성능이 많이 달라진다. 기존에 많이 사용되는 매칭 규칙으로는 유클리디안 거리, L-norm 거리, HEOM, IVDM 등이 있다.[5] 세 번째 단계는 사례 적용 단계로써, 추출된 과거의 사례로부터 추론하여 새로운 사례에 대한 해답을 구하는 단계이다. 새로운 사례와 정확히 일치하는 과거의 사례를 찾는다면 해답을 구하는 것은 간단하나, 대부분의 경우에 새로운 사례와 정확하게 일치하는 사례를 찾는다는 것은 불가능하기 때문에 과거의 사례를 적절히 조정하여 현재 주어진 새로운 사례에 대한 해답을 구하여야 한다. 사례를 적용할 때는 적응 규칙을 사용하게 되는데, 매칭 규칙과 마찬가지로 적응 규칙을 어떻게 적용하느냐에 따라서 성능이 많이 달라진다. 기존에 많이 사용되는 방법은 거리에 반비례하도록 가중치를 부여하는 것이다. 네 번째 단계는 사례 학습이다. 앞의 사례 기반 추론 절차를 거쳐 새로운 사례에 대한 해답을 구한 뒤, 그 해답에 따라 학습을 시키는 것이다. 즉, 사례 베이스를 수정(update)하고, 과거의 새로운 사례에 대한 유사도에 따라 색인에 대한 수정을 하는 것을 말한다.

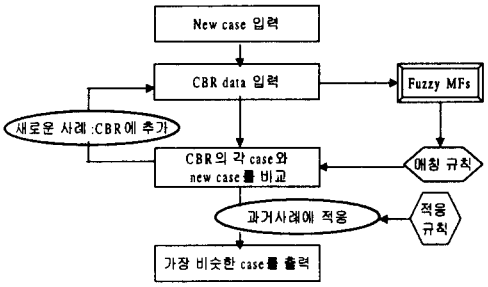
1) 본 연구는 과학기술정책관리연구소(STEPI)의 지원을 받은 것임.



< 그림 1 > 기본적인 사례 기반 추론

3. 퍼지 소속 함수를 weight로 사용한 사례 기반 추론

퍼지를 이용한 CBR의 개념이란 CBR 알고리즘의 기본 개념은 그대로 유지하면서, 결정에 중요한 역할을 하는 이웃 패턴들(K개)이 제공하는 한정된 정보에 좀 더 전반적인 클래스의 분포 정보를 담게 함으로써, 이를 근거로 내리는 판단도 어느 정도 전체 클래스 상황을 잘 반영할 수 있다는 주장이다.[1] 즉, 각각의 학습 패턴들을 자기가 속한 클래스 번호 하나로만 다루는 것이 아니라, 모든 소속 가능한 클래스에 대한 membership 값으로 다루는 것이다. 따라서, membership 값을 얼마나 잘 주느냐에 따라 전체 성능이 좌우된다.



< 그림 2 > Fuzzy Membership Function을 이용한 사례 기반 추론

퍼지 소속 함수를 이용한 CBR의 구조는 < 그림 2 >와 같다. 기존의 사례(y) 각각에 대해서 미리 초기 membership 값을 부여해 주는 과정이 선행되어야 하는데, 이 초기 membership 값을 얼마나 잘 주느냐에 따라 전체 성능이 좌우된다. 본 논문에서는 [1]에서 제안한 < 식 1 >을 사용하였다. 그리고 나서 이 '초기 membership 값'과 '이웃 사례와의 거리 정보'를 이용하여 새로운 입력 사례(x)가 각 클래스에 대해서 갖는 membership 값을 결정한다. 이 때는 < 식 2 >를 사용하여 새로운 입력 사례(x)가 어느 클래스에 속하는지를 판단한다.

이 방법은 각 사례가 완전히 한 클래스에만 속한다고 전제하여 클래스 소속정도($\tilde{\mu}$)를 0과 1로만 주었던 이전의 방법에서 각 클래스마다 속하는 소속정도를 [0, 1] 값으로 준 것이다.

$$\tilde{\mu}_i(y) = 0.51 + (\frac{n_i}{K}) * 0.49, \text{if}(j = i)$$

$$(\frac{n_i}{K}) * 0.49, \text{if}(j \neq i)$$

x:새로운 입력 사례, y:기존의 사례,
K:선택된 유사한 사례 개수, C:총 클래스 수,
j:y가 속한 클래스 번호, $1<i,j<C$, $\sum_{i=1}^{c} n_i = K$

< 식 1 > 초기 membership function

$$\tilde{\mu}_i(x) = \frac{\sum_{j=1}^{K} \tilde{\mu}_i(y_j)(1/||x-y_j||)}{\sum_{j=1}^{K}(1/||x-y_j||)}$$

K : 참조할 이웃 패턴의 수, $1<i<C$, $1<j<K$

< 식 2 > 판단 membership function

4. Fuzzy C-Means를 이용한 퍼지 소속 함수 결정

퍼지 소속 함수를 이용한 CBR에서는 초기 membership 함수를 어떻게 주는가에 따라 성능이 달라진다. 본 논문에서는 초기 membership 함수를 찾는데, Fuzzy C-Means(FCM)를 사용하려는 것이다.

4.1 Fuzzy C-Means(FCM) 알고리즘

FCM 알고리즘은 n개의 데이터를 k개의 군집으로 분할하는 방법으로, 다음과 같은 처리과정을 거치게 된다. 이 때에는 퍼지 정도 m과 cluster 개수 k, 정지 조건을 정해 주어야 한다.

단계 1 : 식 (1)를 만족하도록 [0,1]사이의 임의의 값으로 membership matrix U를 초기화

단계 2 : 식 (3)을 사용하여 퍼지 centers c_i를 계산, i = 1,...,c

단계 3 : 식 (2)에 따라서 cost function을 계산, 정지 조건에 맞으면 실행 중지 (d : 데이터와 center와의 거리)

단계 4 : 식 (4)를 사용하여 U를 수정, 단계 2로 반복

$$(1)\ \sum_{i=1}^{k} u_{ij} = 1, j = 1, 2, \ldots, n$$

$$(2)\ J(U, c_1, c_i, \ldots, c_k) = \sum_{i=1}^{k} J_i = \sum_{i=1}^{k}\sum_{j}^{n} u_{ij}^{\ m}\ d_{ij}^{\ 2}$$

$$(3)\ c_i = \frac{\sum_{j=1}^{n} u_{ij}^{\ m} X_j}{\sum_{j=1}^{n} u_{ij}^{\ m}} \qquad (4)\ u_{ij} = \frac{1}{\sum_{i=1}^{k}(\frac{d_{ij}}{d_{ij}})^{2/m-1}}$$

4.2 FCM을 이용한 CBR의 구조

본 논문에서 제안하는 방법은 FCM으로 퍼지 소속 함수를 구하는 방법이다. < 그림 3 >에서 보듯이 'FCM' 부분이 더 첨가되었다.

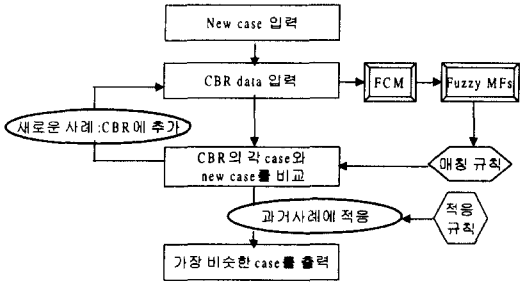
이 방법에서는 먼저 데이터를 정규화 시킨 후, 학습 데이터로 FCM을 실행시킨다. 그 다음 각 center로 모인 사례들이 어느 클래스에 속하는 것들인지 알기 위하여, 각 클래스마다 속하는 사례의 개수를 센다. 이 개수의 비율로 그 center의 각 클래스에 따른 소속 정도를 계산할 수 있다. 이렇게 해서, 각 center가 각각의 클래스에 얼

마만큼 속하는지를 알 수 있는 점수가 나오면, 이 center의 점수와 FCM의 결과로 나온 소속정도 값을 곱한다. (이 때, 한 center의 점수가 클래스별로 차이가 없을 때는 그 center 점수의 적용을 기각한다.) 그리고, 곱하여 나온 값들을 각 클래스에 속하는 값들만을 더하여 그 클래스의 소속정도 값으로 결정한다. 이 소속정도 값이 초기 membership 값이 되는 것이다. 이 방법으로 기존에 일정하게 주었던 초기 membership 값을 좀 더 전체적인 데이터의 분포에 따라 줄 수 있다. 수식으로 설명하면 < 식 3 >과 같다.

$$\widetilde{\mu}_i(x) = \sum_{i=1}^{C}\left(\sum_{j=1}^{M} u_{kj} * g_{ji} \right)$$

u_{kj} : 데이터 k가 cluster j에 속하는 정도 (FCM 결과), g_{ji} : n_j/N,
n_j : j center에 모인 데이터 수, N : 데이터 총수, C : 클래스 총수,
M : cluster 총수, $1<k<N$, $1<i<C$, $1<j<M$

< 식 3 > 제안하는 초기 membership 함수



< 그림 3 > FCM을 이용한 사례 기반 추론

5. 실험 및 결과

5.1 실험 환경

본 논문에서는 Pentium PC 233 MHz를 사용하였으며 Windows 98환경에서 Visual C++ 5.0을 사용하였다.

5.2 실험 데이터

본 시스템에 사용된 데이터는 australian, iris, glass 데이터로 < 표 1 >과 같다.

< 표 1 > 실험 데이터

실험 DATA	INSTANCES 수	ATTRIBUTE 수	CLASS 수
australian	690	15	2
iris	150	5	3
glass	214	10	7

5.3 실험 방법

데이터로는 australian, iris, glass 데이터를 사용하였으며, 본 논문에서 제안한 방법[FCM_FMS_CBR]의 성능을 확인하기 위하여 단순히 유클리디안 거리를 사용하는 기본적인 CBR 방법[CBR]과 Fuzzy membership function을 사용한 CBR 방법[FMS_CBR]도 구현하여 그 성능을 비교하였다.

australian 데이터는 10등분하여 이들 중 9개를 학습하는데 사용하고, 1개를 테스트하는데 사용하는 10-fold cross validation 방법을 사용하였다. 즉, 학습 데이터와 테스트 데이터를 달리 하여 10회 실시하고 평균을 취했다. 그리고, iris 데이터와 glass 데이터는 random 하게 9 : 1로 나누어, 90%의 데이터는 학습 데이터로 나머지 10%의 데이터는 테스트 데이터로 사용하였다.

5.4 실험결과 및 분석

다음 결과는 FCM의 m값은 2로, cluster 수는 class 개수의 3배로 하였으며, 반복회수는 20회로 실험한 결과이다.[2,4] 또, 선택할 유사 사례 개수는 5개로 하였다.

< 표 2 >에 정리된 실험 결과는 제안한 방법[FCM_FMS_CBR]을 단순히 유클리디안 거리를 사용한 기본적인 CBR 방법[CBR]과 Fuzzy membership function을 사용한 CBR 방법[FMS_CBR]의 성능과 비교한 것이다.

이 실험에서 Fuzzy C-Means를 사용하여 Fuzzy membership 값을 주었을 때, <식 1>을 사용하여 Fuzzy membership 값을 주었을 경우보다 성능이 향상된 것을 알 수 있다.

< 표 2 > 실험 결과

실험 DATA	CBR	FMS_CBR	FCM_FMS_CBR
australian	80.0	85.3	86.1
iris	87.5	100.0	100.0
glass	66.0	72.7	77.3

6. 결 론

사례 기반 추론에서 어떻게 입력 사례와 기존 사례와의 유사성을 판별하는가는 매우 중요한 문제이다. 본 논문에서는 최근에 부각되어진 한 방법인 퍼지 소속 함수를 이용하는 방법을 사용하였으며, 이 때, 퍼지 소속 함수를 어떻게 주는가에 따라서 성능이 많이 좌우되는 것을 고려하여, 퍼지 소속 함수를 주는 한 방법으로 Fuzzy C-Means를 이용하는 방법을 제안했다.

참고 문헌

[1] J.M.Keller, M.R.Gray, J.A.Givens, Jr., "A Fuzzy K-Nearest Neighbor Algorithms", IEEE Trans. Syst, Man Cybern., vol. SMC-15, no. 4, pp.580-585, July/August 1985.

[2] Bruno Bosacchi and James C. Bezdek, "Applications of Fuzzy Technology II", Proc. SPIE 2493, 1995.

[3] Mu-Song Chen, Shinn-Wen Wang, "Fuzzy clustering analysis for optimizing fuzzy membership functions", Fuzzy Sets and Systems 103, pp 239-254, 1999.

[4] Nikhil R. Pal, James C. Bezdek, "On Cluster Validity for the Fuzzy c-means Model", IEEE Transaction on Fuzzy Systems, vol 3. No 3, 1995

[5] D.Randall Wilson, Tony R. Martinez, "Improved Heterogeneous Distance Functions", Journal of Artificial Intelligence Research 6, 1997

[6] 김다윗, "신경망 분리 모형과 사례 기반 추론을 이용한 기업 신용 평가", 한국과학기술원 테크노 경영 대학원 석사 학위 논문, 1997