제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

K-Means 클러스터링 알고리즘을 이용한 사례기반 추론에 관한 연구

현우석
한국성서대학교 정보과학부
e-mail: wshyun@bible.ac.kr

A Study on Case-based Reasoning using K-Means Clustering Algorithm

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

요 약

사례 기반 추론(case-based reasoning)은 현재의 문제를 해결하기 위해서 과거에 유사하게 수행된 적이 있는 사례를 유추하여, 유추된 사례의 해를 이용하는 기법으로서 규칙 기반 추론과 함께 여러 분야에 응용되고 있다. 하지만 사례기반 추론 시 새로운 사례를 해결하기 위하여 사례베이스 안의 모든 사례를 검색해야 하기 때문에 수행시간이 증가되는 문제점을 지니고 있다. 본 연구에서는 규칙 및 K-Means 클러스터링 알고리즘에 의한 사례 기반 추론을 이용한 ADS-DAAP(Advanced Diagnosis System for Diseases associated with Acute Abdominal Pain)를 제안한다. 제안하는 시스템은 기존의 CDS-DAAP(Combined Diagnosis System for Diseases associated with Acute Abdominal Pain)와 비교해 볼 때, 수행시간을 감소시켰다.

1. 서론

지금까지 개발된 의료진단 시스템들은 규칙기반 시스템이 대부분이다. 퍼지 논리를 이용한 급성복통과 관련된 질환 진단시스템[1]은 진단에 필요한 지식을 생성규칙으로 표현한 규칙기반 시스템이다. 실제로 진단을 하는데 필요한 지식은 정형화된 규칙만으로 표현하기 어려우며, 과거의 경험을 기초로 진단하는 경우가 적지 않다. 또한 시스템의 성능 향상을 위해 규칙을 계속 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 제안되었다. 그런데 사례기반 추론 시 사례베이스로부터 유사성에 근거한 검색을 해야 하므로 사례베이스의 크기가 증가하게 되면 사례베이스 안의 모든 사례들을 검색해야 하기 때문에 검색시간이 증가되는 문제점이 발생하게 된다.

본 논문에서는 규칙 및 K-Means 클러스터링 알고리즘에 의한 사례기반 추론을 이용한 ADS-DAAP(Advanced Diagnosis System for Diseases associated with Acute Abdominal Pain)를 제안한다. 제안하는 시스템은 기존의 CDS-DAAP와 비교해 볼 때, 수행시간을 감소시켰다.

본 연구는 2002학년도 한국성서대학교 교내 연구비 지원으로 수행되었습니다.

2. K-Means 클러스터링 알고리즘

클러스터링이란 개체들을 특징이 정의되지 않은

집합으로 그룹화 하는 것을 말한다[6]. 특징이 정의되지 않았다는 의미는 그룹의 개수나 구조를 미리 고려하지 않고 분석을 수행한다는 것이다[7].

클러스터링은 두 데이터 사이의 거리나 유사성에 의하여 측정된다. 정량적 데이터의 경우 두 점 사이의 거리나 두 벡터 사이의 각으로 측정하고, 정성적 또는 범주형 데이터의 경우에는 두 데이터 사이에 일치하는 속성들의 수를 가지고 측정한다.

클러스터링 할 수 있는 모든 가능성을 가늠해 보는 것은 현실적으로 거의 불가능하므로 다양한 클러스터링 알고리즘이 나타나게 되었다. 본 논문에서 이용한 K-Means 클러스터링 알고리즘은 다음과 같은데, 숫자 K는 미리 정하거나 클러스터링 도중에 정할 수도 있고, K를 변화시켜 가면서 클러스터링 한 다음 가장 결과가 좋은 K를 사용하기도 한다.

1단계: 데이터를 K개의 초기 클러스터로 나누거나 또는 임의로 K개의 데이터를 초기치(seed)로 선택하여 K개의 클러스터를 만든다. 어느 데이터를 어느 클러스터에 할당하느냐는 사용자가 임의로 정하거나 특정 알고리즘에 의하여 정할 수 있다. 여기서 초기치는 클러스터링을 하기 위한 기준 데이터로서 처음의 K개의 데이터를 초기치로 쓸 수도 있고, 임의로 선택할 수도 있다.

2단계: 각 클러스터의 평균이나 중심을 구한다.

3단계: 임의의 한 데이터를 선택하여 각 클러스터 중심까지의 거리를 계산한다. 만일, 이 데이터와 임의의 클러스터 중심까지의 거리 중 가장 가까운 것이 자신이 속한 클러스터라면 그대로 둔다. 만일 그렇지 않다면, 거리가 가장 가까운 클러스터에 재 할당한다.

4단계: 제 3단계를 모든 데이터에 대해 수행한다.

5단계: 제 2,3,4단계를 재 할당이 없을 때까지 계속한다.

클러스터링은 자율학습(Unsupervised Learning)을 가능케 하며, 결과 값이 미리 정해져 있지 않은 상태에서 사용할 수 있다. 그러므로 데이터 베이스 내부 구조에 관한 지식이 없이도 클러스터링 기법을 사용하여 데이터의 감추어진 구조를 발견할 수 있으며, Data Mining을 수행할 때에 가장 기본적으로 사용되기도 한다[8]. 클러스터링의 단점으로는 거리측정 방법과 데이터 속성들의 가중치를 정하는 표준이 없으며, 특히 K-Means 클러스터링에서는 클러스터링을 완료하여 그 결과들을 비교해 보기 전까지는 가장 적당한 K값을 알 수가 없다는 점을 들 수 있다.

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

제안하는 급성복통 진단을 위한 규칙 및 K-MEANS 클러스터링 알고리즘에 의한 사례기반 추론을 이용한 ADS-DAAP(Advanced Diagnosis System for Diseases associated with Acute Abdominal Pain)에서는 일반적인 급성복통 진단을 위한 지식은 규칙으로 표현하고, 기존 규칙으로 처리할 수 없는 예외적인 급성 복통 진단을 위한 지식은 사례로 표현함으로써 규칙과 사례가 서로 보완적인 역할을 할 수 있도록 하였다. 또한 사례기반 추론 시 사례 사례베이스로부터 유사성에 근거한 검색을 해야 하므로 사례베이스의 크기가 증가하게 되면 수행시간이 증가되는 문제점을 해결하고자, K-MEANS 클러스터링 알고리즘을 사용하여 사례베이스를 클러스터링 함에 의해서 수행시간을 감소시키게 되었다.

제안하는 ADS-DAAP의 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 K-MEANS 클러스터링 알고리즘에 의한 예외상황 사례베이스를 기반으로 재진단을 시도하게 되어 조회시간을 감소시키게 되었다.

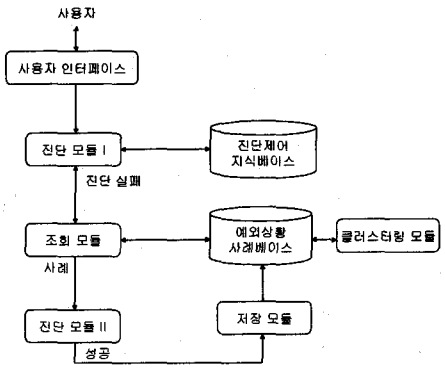

그림 1 ADS-DAAP의 구조

3.1 클러스터링 모듈

클러스터링이란 주어진 집합의 데이터들을 비슷한

성질을 가지는 그룹으로 나누는 것을 말한다[6]. 클러스터링 모듈에서는 예외상황 사례베이스에 들어있는 사례들을 클러스터링 한다. 클러스터링 모듈이 실행되기 위해서는 '수술에 관한 위험 요인 정도', '노화 정도', '탈수 정도', '장폐색 정도', '구토 정도', '상복부의 통증 정도', '하복부의 통증 정도', '만성 복통 정도', '직장 수지 검사 정도' 등의 속성이 필요하다. 모듈의 실행이 끝나면 그 결과로 K개의 클러스터와 클러스터별로 1개의 중심이 도출된다. 이 결과는 클러스터별로 인덱싱되어 예외상황 사례베이스에 저장되며 이 정보는 조회모듈에서 사례들을 조회할 때 사용된다. 본 연구에서 K값은 4로 정했다.

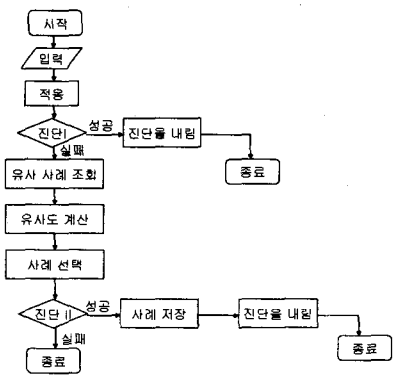

그림 2 질환 진단 과정

본 논문에서는 K-Means 클러스터링 알고리즘을 사용했으며, 다음과 같다.

1단계: 사례들을 K개의 초기 클러스터로 나눈다.

2단계: 각 클러스터의 중심 사례를 구한다. 중심 사례는 각 클러스터에 속하는 각 사례들의 속성별 평균치로 한다.

3단계: 임의의 한 사례를 선택하여 각 클러스터 중심 사례까지의 유사도를 계산한다. 만약 이 사례와 임의의 클러스터 중심까지의 유사도 중에서 가장 가까운 것이 자신이 속한 클러스터라면 그대로 두고, 그렇지 않을 경우 거리가 가장 가까운 클러스터에 재 할당한다. 여기서 임의의 한 사례와 각 클러스터 중심까지의 유사도는 퍼지 근접관계[9]를 사용하여 구했으며, 수식 (1)과 같다.

$$Similarity(case_i, case_m) = 1 - \frac{\sum_{k=1}^{n} |attr_{ik} - attr_{mk}|}{|n|} \quad (1)$$

n: 속성 수

$case_i$: 임의의 사례

$case_m$: 각 클러스터의 중심 사례

$attr_{ik}$: 임의의 사례를 구성하는 k번째 속성을 나타내는 퍼지값 $(1 \le k \le n)$

$attr_{mk}$: 각 클러스터의 중심 사례를 구성하는 k번째 속성을 나타내는 퍼지값$(1 \le k \le n)$

4단계: 제 3단계를 모든 데이터에 대해 수행한다.

5단계: 제 2,3,4 단계를 재 할당이 없을 때까지 반복한다.

3.2 예외상황 사례베이스

예외적인 경우 질환이 진단되는 각 사례들은 관계형 데이터베이스에서 하나의 테이블 형태로 저장되도록 설계하였다. 따라서 새로운 사례로 저장될 증상 데이터는 관계형 데이터 베이스에서 하나의 튜플(tuple)을 구성하며, 증상 데이터의 각 특성(property)은 튜플의 속성(attribute)이 된다.

3.3 조회모듈

진단모듈 I에서 질환 진단에 실패할 경우, K-Means 클러스터링 알고리즘을 이용한 예외상황 사례베이스에 있는 사례들을 조회하여 현재 사례와 유사한 사례를 찾아낸다. 먼저 입력된 현재 사례와 각 클러스터의 중심 사례들간의 유사도를 계산하여 가장 가까운 클러스터 안에 들어 있는 사례들에 대하여 다시 유사도를 계산하게 된다. 입력된 현재 사례와 각 클러스터의 중심 사례들간의 유사도를 계산하기 위해서 수식(1)을 사용하였으며, 가장 가까운 클러스터 안에 들어있는 사례들에 대하여 다시 유사도를 구할 때도 퍼지 근접관계[9]를 사용하였으며 식 (2)와 같다.

$$Similarity(case_i, case_j) = 1 - \frac{\sum_{k=1}^{n} |attr_{ik} - attr_{jk}|}{|n|} \quad (2)$$

n : 속성 수

$case_i$: 현재 사례

$case_j$: 과거 사례

$attr_{ik}$: 현재 사례를 구성하는 k번째 속성을 나타내는 퍼지값 $(1 \le k \le n)$

$attr_{jk}$: 과거 사례를 구성하는 k번째 속성을 나타

내는 퍼지값 ($1 \le k \le n$)

사례의 조회 과정에서 완전히 일치하는 사례를 찾는 경우는 드물다. 따라서 조회된 가장 유사한 사례에 수정 규칙을 적용하여 현재의 상황에 맞도록 적응하는 과정이 필요하다. 그러나 본 시스템에서는 일관된 수정 규칙을 발견하기가 어려우므로 이러한 수정 규칙은 포함하지 않았으며 조회과정이 끝나면 가장 유사한 과거의 사례를 유사도와 함께 제시함으로써 사용자의 의사결정을 지원하게 된다.

4. 평가

시뮬레이션 환경에서는 본 시스템의 성능을 평가하기 위해서 G 병원으로부터 획득한 300명의 실제 환자 데이터를 수집하여, 100명의 환자 데이터는 사례베이스에 저장하고 200명의 환자 데이터를 10 가지 test set으로 나누어 기존의 CDS-DAAP와 제안하는 ADS-DAAP에서 각각 진단을 하여 평균 수행시간을 비교하였는데 그림 3과 같다. 기존의 CDS-DAAP보다 제안하는 ADS-DAAP에서 수행시간이 감소되었다.

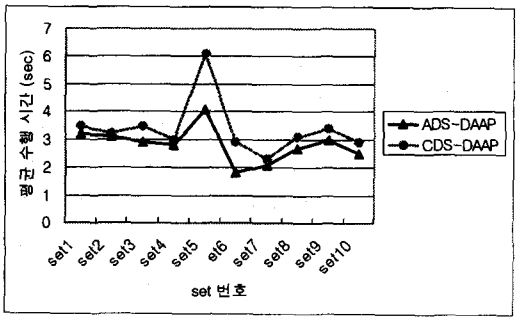

그림 3 시스템에 따른 평균 수행시간 비교

5. 결론 및 향후 과제

제안하는 ADS-DAAP는 CDS-DAAP와 비교해 볼 때 평균 수행시간이 감소되었다. 이는 예외상황 사례베이스를 검색할 때 모든 사례들에 대하여 검색을 하지 않고, K-MEANS 알고리즘에 의하여 입력된 사례와 가장 유사한 클러스터 안에 들어 있는 사례들에 대해서만 검색을 하기 때문이다. 또한 예외상황 사례베이스에 사례가 축적됨에 따라 축적된 사례를 가지고 새로운 규칙을 도출하여 지식베이스에 있는 규칙을 수정하거나 추가시킬 수 있게 함에 의해서 향후 시스템의 성능을 더욱 향상시킬 것으로 기대된다.

본 연구에 이어서 향후에 이루어져야 할 과제는 다음과 같은 항목들을 고려하여 이루어져야 한다. 첫째, 제안하는 ADS-DAAP에서 가장 좋은 결과를 내는 K 값과 중심의 개수를 구하는 차후 연구가 요구된다. 둘째, 개별 클러스터간의 구별 시 모호한 경계를 해결할 수 있는 방법에 대한 차후 연구가 요구된다. 셋째, K-Means 클러스터링시 유사도를 구하는 표준적인 방법이 존재하지 않으므로, 유사도를 구하는 표준적인 방법에 대한 차후 연구가 요구된다.

참고문헌

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

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

[3] 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.

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

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

[6] Afifi, A. A. and Clark, V., Computer-Aided Multivariate Analysis, Chapman and Hall, 1990.

[7] Johnson, R. A. and Wichern, D. W., Applied Multivariate Statistical Analysis, Prentice-Hall, Inc., 1982.

[8] Berry, M. and Linoff, G., Data Mining Techniques: For Marketing, Sales and Customer Support, John Wiley & Sons, Inc., 1997.

[9] Klir, G. and T. Folger, *Fuzzy Sets, Uncertainty, and Information*, Prentice-Hall International Editions, 1992.