집단 건강검진을 위한 계층군집 기반 개인화 피드백 시스템

강현경*, 김준우**

요약

생활수준의 향상 등으로 건강관리에 대한 관심이 높아짐에 따라, 학교나 직장 등에서도 구성원들을 대상으로 정기적인 집단 건강검진을 실시하는 것이 일반적이다. 이러한 건강검진은 질병의 조기 발견 등에 많은 도움이 되는 것이 사실이나, 검진에 포함된 수많은 항목들은 대부분 단순히 질환의 유무 판단에만 활용되는 한계점이 있다. 본 논문은 검진에서 발견되지 않은 질환들에 대한 위험을 예측하고, 이에 대비하기 위한 일상적인 유의사항을 포함하는 개인화된 건강검진 피드백 시스템을 제안한다. 이 과정에서 검진 대상 집단의 특성을 파악하기 위하여 데이터 마이닝 기법인 계층군집 분석을 활용하였고, 제안하는 시스템은 실제 고등학생 대상 치위생 검진 결과 데이터에 적용되었다.

Hierarchical Clustering based Personalized Feedback System for Mass Health Examination

Hyun-Kyung Kang* and Jun-Woo Kim**

ABSTRACT

As the living standards have improved, the healthcare becomes a major concern of the people and organizations such as schools and companies. The mass health examination is known to be helpful for early detection of diseases, however, there is a limitation that the examination results are generally used to simply identify specific symptoms and diseases. To make full use of the mass health examination result, this paper proposes an intelligent feedback system which can evaluate the individual risk for symptoms and diseases not identified in the examination results and provide the health-related advices. To this end, the hierarchical clustering analysis, one of the common data mining techniques, is deployed to extract the characteristics of the target examinee group. Our system is applied to analyze the mass dental examination result of high school students.

Key Words : Health Examination, Medical Data Mining, Hierarchical Clustering, Personalized Service, Recommendation

* 신라대학교 치위생학과(✉icando@silla.ac.kr)
** 동아대학교 산업경영공학과
· 제1저자(First Author) : 강현경 · 교신저자(Correspondent Author) : 김준우
· 접수일(2011년 6월 24일), 수정일(1차 : 2011년 7월 22일), 게재확정일(2011년 7월 25일)

www.kci.go.kr

Ⅰ. 서 론

생활수준 향상 등으로 인하여 건강에 대한 관심이 높아지면서, 여러 가지 건강관리 서비스의 중요성이 커지고 있다[1]. 건강검진은 질병을 조기 발견하여 예방 내지 치료하는 것을 목적으로 하는 기본적인 건강관리 서비스로, 조기치료를 통한 수검자 건강 증진 및 사회적 부담 경감을 줄이는데 의의가 있다[2][3][4]. 이러한 건강검진이 효과적으로 운영되기 위해서는 검진 자체의 질을 향상시키는 것과 동시에, 실질적 생활 습관 개선 및 효과적인 사후 관리로 이어져야 함이 지적된다[3].

현재 건강검진이 활성화되어 학교, 직장 등의 조직에서 집단 검진을 정기적으로 실시하는 것이 일반화되었고, 여러 가지 성과들도 거두고 있으나, 이와 관련된 한계점들도 존재하는 것이 사실이다. 특히 1회의 건강검진을 실시할 경우, 조사하는 항목들이 상당히 많으나 이 항목들을 체계적으로 분석하기보다는 개별 항목에서 특이한 결과가 관측되는 경우, 추가적 검진이나 치료를 요망하는 정도로 검진 결과를 활용하고 있다. 또한 일반적으로 시간과 인력의 한계로 인하여 문진 절차가 형식적으로 진행되는 경우가 많아, 검진 항목 중 전문의가 관측 및 기록하는 항목이 아닌 수검자 개인이 작성하는 설문 항목에 대한 활용이 미비하기 쉽고, 당장 질환이 관측되지 않는 경우 개인 수검자에게 적절한 피드백을 제공하는데 한계가 있다[3][5].

더불어, 최근 활성화된 집단 검진의 경우, 인구사회학적, 신체적 특성이 유사한 수검자 집단에 대해 시행되기 때문에 타 집단과 차별화되는 집단 내 특성을 신속히 파악하여 이를 반영한 피드백 및 사후 관리에 대한 보완도 필요하다[4].

이러한 점을 보완하기 위하여 본 논문은 집단 검진을 실시한 이후, 개별 수검자들에게 보다 유용하고 개인화된 피드백을 제공하는 시스템을 제안하고자 한다. 제안하는 시스템은 먼저 검진 대상 집단에서 관찰되는 질환들 간의 상호 관련성을 데이터 마이닝 기법의 하나인 계층군집(Hierarchical Clustering) 분석을 통해 도출하고, 이를 이용하여 개인 수검자별 향후 위험 질환을 선별한다. 선별된 위험 질환에 대해서는 이를 예방 및 관리하기 위하여 일상생활에서 유의할 필요가 있는 사항들을 함께 제시함으로서, 보다 유용한 개인별 피드백이 가능하며, 이러한 과정에서 수검자들이 직접 작성한 설문 항목들도 이용된다.

Ⅱ. 관련 연구

데이터 마이닝은 방대한 양의 자료에 숨겨져 있는 유용한 지식이나 패턴을 탐사하는 분석 방법으로[6][7], 최근 고객 관계 관리, 교육, 의료 등 다양한 서비스 분야에서 성공적으로 활용되고 있다 [8][9][10].

데이터 마이닝의 한 분야인 군집 분석(Clustering Analysis)은 본래 데이터를 구성하는 레코드(record)들 중, 서로 유사한 것들을 같은 군집으로 묶는 작업에 해당하며, 데이터 안에 숨겨진 구조나 특성을 파악하는데 유용하다[7]. 군집 분석 중, 계층군집 방법은 두 레코드 간 상호 연관성에 기반하고 있다는 특징이 있으며, 크게 병합형과 분할형으로 구분된다. 이러한 계층군집 방법은 모든 레코드의 연관성을 한 눈에 파악할 수 있는 계통도(dendrogram)를 생성할 수 있다는 장점이 있지만, 레코드 수가 많아질 경우, 수행 속도가 느려진다는 단점도 있다[11].

반면, 최근에는 전통적인 군집 분석의 의미인 레코드를 군집화하는 것과 달리, 데이터를 구성하는 필드(field)를군집화하는 데에도 계층군집이 사용되고 있다[12]. 앞서 말한대로 계층군집은 두 항목 간의 상호 유사성을 적절히 정의할 수 있는 경우, 수행이 가능하다는 특성이 있는데, transaction 데이터의 경우에는 항목(item) 간의 유사도를 동시 발생 빈도(co-occurrence, 공기 정보) 또는 이를 변형하여 정의

www.kci.go.kr

할 수 있어, 주로 transaction 데이터의 항목을 군집하거나, transaction 데이터로 변형이 가능한 이진(binary) 데이터의 필드를 군집하는 데에도 계층군집이 널리 사용된다. 현재 동시 발생 빈도와 관련된 여러 가지 유사도 척도를 이용하는 필드 계층군집 방법 연구를 찾아볼 수 있고, 구체적으로는 단수 동시 발생 빈도를 이용하는 방법[13], 빈발 항목 집합들의 하이퍼그래프를 이용하는 방법[14], 각 항목들의 선호도를 이용하는 방법[15] 및 별도의 항목 taxonomy를 함께 이용하는 방법[16] 등이 다양하게 제안되었다.

건강검진 시 조사되는 다양한 항목들 중에는 장비나 검사자의 측정에 의해 수치적으로 기록되는 항목들도 있지만, 나머지 상당수의 항목들은 발생여부나 인지여부를 검사자가 진단하거나 수검자 스스로 기록하는 설문 항목들로 이진 문항이라는 특징을 가지고 있다. 본 논문에서 제안하는 시스템은 이러한 이진 문항들의 분석에 초점을 맞추고 있으며, 이진 데이터의 특성 상, 앞에서 설명한대로 계층군집을 통하여 필드들의 연관성 정보를 파악할 수 있다. 본 논문은 이러한 정보를 수검자 개인별로 검진 사후 일상적인 건강관리에 유용한 피드백을 제공하는데 사용한다.

III. 집단 건강검진 개인화 피드백 시스템

본 논문에서 제안하는 집단 건강검진 수검자 개인화 피드백 시스템의 전체적인 구조는 그림 1에 요약되어 있다. 시스템의 분석 절차는 크게 사전 처리에 해당하는 대상 집단 분석 단계와 여기서 추출한 지식을 토대로 개인 수검자에 대한 피드백을 생성하는 단계로 구분된다. 또한, 본 시스템의 분석 대상은 건강검진 중 특정 분야의 조사 항목들이며, 이 항목들은 크게 전문의의 진단과 관찰에 의해 기록되는 '질환 진단 항목'과, 수검자들이 작성하는 '설문 항목'으로 구성된다고 가정한다.

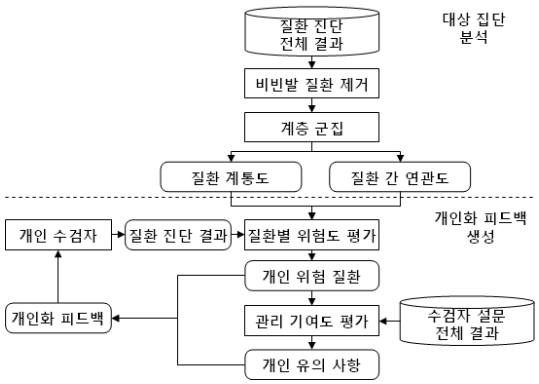

그림 1. 개인화 피드백 시스템 전체 구조

Fig. 1. The Overall Framework of The Personalized Feedback System

먼저, 대상 집단 분석 단계에서는 검진 항목 중, 질환 진단 항목에 초점을 맞추어, 각 질환들 간 연관성을 파악하기 위해 계층군집 분석을 실시한다. 이 때, 두 질환 간 유사성 척도로 동시 발생 빈도를 사용하고, 분석 결과 함께 나타나는 경우가 잦은 질환들의 관계를 파악하여, 질환 계통도 및 질환 간 연관도를 도출한다.

대상 집단 분석 단계가 완료되면, 개별 수검자들에게 개인화된 피드백을 생성하여 제공할 수 있다. 이 단계에서는 먼저 개인 수검자의 질환 진단 결과에 이전의 대상 집단 분석에서 도출한 질환 계통도 및 질환 간 연관도를 적용하여, 해당 수검자의 향후 질환별 발생 위험도를 정량적으로 산출한다. 위험도가 높은 상위 질환들의 경우, 수검자 개인 위험 질환으로 선별된다.

선별된 각 개인 위험 질환에 대해서는 설문 항목들을 관찰하여, 해당 질환을 예방하는데 유용한 개인 유의 사항을 도출한다. 최종적으로는 이렇게 도출된 개인 위험 질환과 개인 유의 사항들이 수검자에게 피드백되어 이번 건강검진에서 드러나지 않은 증상에 대해서도 일상생활 중에 주의와 관리를 해 나갈 것을 추천하게 된다.

www.kci.go.kr

3.1 대상 집단 분석

건강검진에는 여러 분야에 해당하는 다양한 검사 항목들이 포함되어 있다. 이러한 항목들은 각기 건강 관리에 있어 중요하고 의미있는 변수들이지만, 집단 건강검진의 경우, 대상 집단의 특성에 따라 발생 빈도가 극히 낮게 조사되는 항목들도 존재한다. 예를 들어, 10대 고교생을 대상으로 실시한 건강검진과 노년층을 대상으로 실시한 건강검진에서 조사된 결과들의 특성은 서로 차이점이 있는 것과 같다.

따라서, 본 논문에서 제안하는 시스템은 먼저 현재 검진 대상 집단의 특성을 도출해내는 것으로 분석을 시작한다. 이를 위해 먼저 전문의에 의해 조사된 질환 진단 자료에서 발생 빈도가 낮은 비빈발 질환 항목들을 제거한 후, 남은 T개의 질환 항목들에 전통적인 병합형 계층군집 분석을 적용하여, 항목 간 연관성을 산출한다. 이 때, 계층군집을 위해 필요한 두 질환 항목 간 유사성 측정 척도로는 전체 검진 대상자 집합에서 관찰되는 두 질환의 동시 발생 빈도를 사용한다.

병합형 계층군집을 완료할 경우, 전체 T개의 질환 항목들 간 유사성을 표현한 계통도를 얻을 수 있고, 이를 적절히 분할하면 동시 발생 위험이 큰 질환들로 구성된 군집 k개를 얻을 수 있다.

이러한 계통도는 각 질환 간의 전체적인 연관성을 한 눈에 보여줄 수 있지만, 두 질환 간의 정량적인 연관성을 파악하기 위해서는 이를 좀 더 가공할 필요가 있다. 이에 따라, 본 논문에서는 두 질환 i, j간의 연관성 $R_{i,j}$를 (1)과 같이 산출할 것을 제안한다.

$$R_{i,j} = \frac{CO_{i,j}}{N} + \triangle \quad (1)$$

단, $CO_{i,j}$와 N은 질환 i, j간 동시 발생 빈도수와 전체 수검자 수를 의미하고, $\triangle$는 (2)와 같다.

$$\triangle = \begin{cases} \frac{k}{T}, & \text{같은 군 소속인 경우} \\ 0, & otherwise \end{cases} \quad (2)$$

(1), (2)에서 질환 간 유사성 $R_{i,j}$는 동시 발생 빈도가 높을수록 값이 큼을 알 수 있다. $\triangle$는 두 질환 i, j가 동일 군집에 소속되어 있을 경우, $R_{i,j}$에 더해지는 일종의 가중치로 볼 수 있으며, 군의 개수 k가 많을수록 동일 군집 소속 가중치 $\triangle$는 커지도록 되어 있다. 이러한 $R_{i,j}$는 다음 절에서 제안하는 개인화 피드백에 이용된다.

3.2 개인화 피드백 생성

대상 집단 분석이 완료된 경우, 이제 개별 수검자에 대한 개인화 피드백을 생성하게 된다. 먼저, 한 개인 수검자의 건강검진 결과가 질환 진단 항목에 대한 결과인 행벡터 $D = [D_1\ D_2\ \ldots\ D_T]$와 설문 항목에 대한 결과 벡터 $S = [S_1\ S_2\ \ldots\ S_s]$로 표현됨을 가정하자. 단, s는 설문 항목의 개수를 의미하고, 본 논문에서는 주로 검진 항목 중, 이진 항목들에 대해 분석하고 있으므로, D, S는 모두 이진 벡터로 가정한다. 즉, D_i의 경우, 질환 i가 해당 수검자로부터 관찰되었으면 1, 그렇지 않을 경우 0의 값을 갖는 이진 변수이다.

개인화 피드백 생성을 위해서는 개인 수검자에 대해 이번 건강검진에서 발견되지는 않았으나, 향후 발견될 가능성이 큰 위험 질환을 선별한다. 이를 위해 개별 수검자 질환 진단 결과 D의 T개 질환 항목 중, 이번 검진에서 이 수검자로부터 관찰되지 않은 항목 i에 대한 위험도 A_i를 (3)과 같이 산출할 것을 제안한다.

$$A_i = \sum_j R_{i,j} D_j \quad (j = 1, 2, \ldots, T \mid j \neq i) \quad (3)$$

개별 수검자로부터 관찰되지 않은 질환들의 위험도

www.kci.go.kr

가 산출되면, 이 질환들 중 높은 위험도를 보이는 항목들은 현재 검진 대상 집단의 특성 상, 이 수검자도 향후 유의해야 할 위험 질환이라 볼 수 있다.

질환의 발생 여부는 수검자 개인이 일상생활에서 판단하기 어려운 측면이 있는 반면, 간단한 예후 증상이나 생활 습관 등은 개인이 일상생활에서 느끼거나 조절하기 용이하다. 건강검진 문항의 설문 항목들은 이러한 부분들에 대한 조사를 목적으로 하고 있으며, 따라서 개인 위험 질환이 선별된 이후에는 설문 문항 중, 개인 위험 질환의 조절에 도움이 되는 항목들을 선택하여 수검자에게 위험 질환과 함께 해당 질환 예방 또는 조기 발견을 위한 유의사항으로 제시하는 것이 바람직할 것으로 생각된다.

표 1. $(S_s,\ D_d)$의 분할표

Table 1. Contingency Table for $(S_s,\ D_d)$

S_s / D_d	1	0
1	f_{11}	f_{10}
0	f_{01}	f_{00}

이를 위해 설문 항목 S_s의 유무가 위험 질환 D_d를 관리하는데 기여하는 정도를 의미하는 관리기여도 설문 항목 $C_{s,d}$를 산출하는 것이 필요하다. $C_{s,d}$는 표 1과 같은 분할표를 이용하여 계산한다.

표 1의 f_{ab}가 의미하는 바는 전체 수검자 중, S_s 항목의 결과값이 a, D_d 항목의 결과값이 b 인 수검자의 인원수를 의미한다. 이러한 분할표를 이용하여 본 논문에서는 관리기여도를 간단히 (4)와 같이 산출한다.

$$C_{s,d} = \frac{\max(f_{11}+f_{00}, f_{10}+f_{01})}{N} \qquad (4)$$

(4)가 의미하는 바는 두 변수 S_s, D_d 사이의 상관관계가 강할수록 S_s가 D_d를 관리하는데 기여하는 바가 크다는 것을 의미한다. 구체적으로는 (4)의 분모에서 $f_{11}+f_{00} > f_{10}+f_{01}$인 경우 양의 상관 관계가 있고, 그 반대의 경우는 음의 상관 관계를 의미한다. $f_{11}+f_{00}$와 $f_{10}+f_{01}$의 값이 유사한 경우에는 두 문항 간 선형 관계가 미약함을 의미한다. 최종적으로는 앞에서 선별한 개인 위험 질환 각각에 대하여 관리기여도가 큰 설문항목들을 선정하여 제시함으로서 개별 수검자로 하여금 자신이 해당하는 검진 집단을 기준으로 보았을 때, 어떤 질환을 조심해야 하는지, 그리고 이 질환을 조기 진단 또는 예방하기 위해서는 어떤 자각 증상이나 생활 습관에 유의해야 하는지를 알려주는 것으로 개인화 피드백이 완수된다.

Ⅳ. 치위생 분야 검진 결과 분석

본 논문에서는 앞에서 설명한 집단 건강검진 개인화 피드백 방법을 실제로 적용해 보기 위하여, 부산 소재 D 고등학교에서 2011년 실시한 건강검진 결과 중, 치위생 분야 문항들을 선정하고, 1학년 학생 278명에 대한 검진 결과를 수집하였다.

4.1 분석 대상 데이터 구성

치위생 분야 검진 결과 자료는 크게 전문의에 의한 진단에 의해 기록되는 질환 진단 항목과 수검자들이 사전에 작성한 설문 항목으로 구성되어 있다. 대부분의 치위생 분야 검진 항목은 이진 문항이며, 간혹 발견되는 다진 문항은 적절히 이진화하였다.

www.kci.go.kr

표 2. 분석 대상 질환 진단 항목

Table 2. Diseases Checked by Doctor

번호	항목
1	우식 치아 유무
2	우식 발생 위험 치아 유무
3	결손치 유무
4	구내염 및 연조직 질환 유무
5	부정교합 유무
6	구강 위생 상태 불량 여부
7	기타 치아 이상 유무
8	치주 질환 유무
9	악관절 이상 여부
10	치아 마모증 유무

표 3. 분석 대상 설문 항목

Table 3. Survey Questions

번호	항목
1	치아가 깨지거나 부러짐
2	차갑고 뜨거운 식음료 섭취 시 통증
3	치아가 쑤시거나 욱신거리며 아픔
4	잇몸이 아프거나 피가 남
5	혀 또는 입 안쪽 뺨이 욱신거리며 아픔
6	불쾌한 입 냄새가 남
7	지난 1년간 치과병원 방문 여부
8	아침 식사 전 이를 잘 닦는지
9	아침 식사 후 이를 잘 닦는지
10	점심 식사 후 이를 잘 닦는지
11	저녁 식사 후 이를 잘 닦는지
12	잠자기 직전 이를 잘 닦는지
13	간식 섭취 후 이를 잘 닦는지
14	단 음식이나 청량음료를 즐겨 먹는지
15	현재 불소 함유된 치약을 사용하는지

질환 진단 항목은 표 2에 정리되어 있으며, 이들은 보통 일반인이 스스로 자가 진단하기 어려운 항목들임을 알 수 있고, 그렇기 때문에 발견 및 진단을 위해서는 건강검진이나 치과 내원 등이 필요하다.

반면 표 3에 나열된 설문 항목들의 경우, 치위생 관련 평소 자각 증상이나 생활 습관들을 묻고 있고, 전문의의 존재 없이 일반인이 평소에도 발견하거나 조절할 수 있는 항목들임을 알 수 있다.

4.2 대상 집단 분석 과정

표 2의 항목 중에서는 서로 동시에 발생하는 관련성이 있을 수 있다. 이러한 관계를 전체적으로 한 눈에 파악하기 위하여, 먼저 비빈발 질환들을 제거한 후, 병합형 계층군집 알고리즘을 적용하였다.

이 과정에서 전체 수검자 중, 5% 미만에서 관찰된 4개 질환들을 비빈발 질환으로 보아 제거하였고, 나머지 6개 질환들에 대하여 동시 발생 빈도를 유사도 척도로 사용한 병합형 계층군집 분석을 실시하였다. 표 3은 남겨진 6개의 빈발 질환들을 보여주며, 10대 고교생 집단에서 결손치, 악관절 이상, 치아 마모증 및 기타 치아 이상의 경우 흔히 발견되지 않는 질환들임을 알 수 있다.

표 3. 빈발 질환

Table 3. Frequently Identified Diseases

기호	질환
D_1	부정교합
D_2	위생 상태 불량
D_3	우식 치아
D_4	치주 질환
D_5	구내염 및 연조직 질환
D_6	우식 위험 치아

표 4. 동시 발생 행렬

Table 4. Co-occurrence Matrix

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	35	32	18	9	7	2
D_2	32	127	64	21	16	19
D_3	18	64	83	19	14	10
D_4	9	21	19	33	26	6
D_5	7	16	14	26	26	4
D_6	2	19	10	6	4	26

www.kci.go.kr

표 4는 표 3의 D_1~D_6 질환들 간의 동시 발생 빈도를 행렬로 보여주고 있으며, 이러한 정보를 근거로 병합형 계층군집 분석을 실시한 결과, 그림 2와 같은 계통도를 얻을 수 있었다.

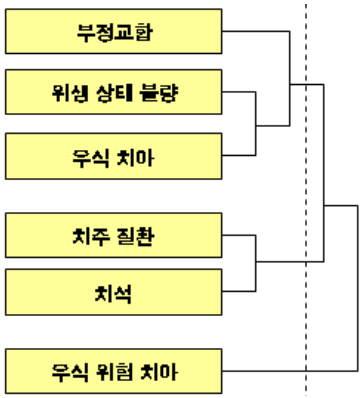

그림 2. 질환 문항 계통도

Fig. 2. The Dendrogram of Diseases

그림 2의 계통도를 수직 점선 위치에서 분할할 경우, 질환 문항들은 (부정교합, 위생 상태 불량, 우식 치아), (치주 질환, 치석), (우식 위험 치아)의 3개 3개 군으로 나뉘어지고, 이 군집들은 각각 함께 발생하는 경우가 많은 질환들로 이루어져 있다. 현재 군의 개수 k=3, 질환의 개수 T=6이므로, △=3/6=0.5이다. 이러한 정보들과 (1)을 이용하면 표 5가 보여주는 것과 같은 질환 간 연관성 $R_{i,j}$를 얻는다.

표 5. $R_{i,j}$ 행렬

Table 5. $R_{i,j}$ Matrix

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	-	0.62	0.56	0.03	0.03	0.01
D_2	0.62	-	0.73	0.08	0.06	0.07
D_3	0.56	0.73	-	0.07	0.05	0.04
D_4	0.03	0.08	0.07	-	0.59	0.02
D_5	0.03	0.06	0.05	0.59	-	0.01
D_6	0.01	0.07	0.04	0.02	0.01	-

4.3 개인화 피드백 생성

대상 집단 분석 과정이 완료되면, 이제 각 수검자에게 개인화된 피드백을 제공할 수 있다. 이 과정을 살펴보기 위해 실제 수검자 중, 5명을 표본으로 추출하여 질환 진단 항목 결과와 함께 표 6에 나타내었다. 표 6에서 값이 1인 항목은 그 질환이 관찰되었음을 나타내고, 값이 0인 항목은 그렇지 않음을 의미한다.

표 6. 표본 수검자 목록

Table 6. Sample Examinee List

번호	D_1	D_2	D_3	D_4	D_5	D_6
1	0	1	0	1	1	0
2	1	1	0	0	0	0
3	0	0	0	0	0	1
4	0	0	0	0	0	0
5	0	1	1	0	0	0

이제, 표 6의 각 수검자에 대하여 개인 위험 질환을 선별하기 위해 위험도를 계산한 결과는 표 7과 같다. 참고로 이미 발생한 질환에 대한 위험도는 계산하지 않으며, 이는 이미 발생한 질환은 건강검진 당시 문진이나 건강검진 사후에도 명시적으로 병원 내원 및 치료를 권하고 있기 때문이다.

표 7의 결과를 살펴보면, 본 논문에서 제안하는 시스템의 기능과 한계점에 대해 파악할 수 있다. 먼저, 수검자 1, 2에 대해서, 본 시스템은 가장 위험하며 조심해야 할 질환으로 D_3, 즉, 우식 치아를 선별하고 있다. 수검자 1의 경우에는 우식 치아와 동시 발생 빈도가 높았던 위생 상태 불량을 갖고 있었고, 수검자 2 역시 위생 상태 불량 및 부정 교합을 갖고 있어, 이 점은 타당하며, 수검자 2의 위험도를 더 높게 산출한다는 점 역시 합리적이다.

www.kci.go.kr

표 7. 개인별 질환 위험도

Table 7. Individual Risk Grades

번호	A_1	A_2	A_3	A_4	A_5	A_6
1	0.68	-	0.85	-	-	0.01
2	-	-	1.29	0.11	0.09	0.08
3	0.01	0.07	0.04	0.02	0.01	-
4	0.00	0.00	0.00	0.00	0.00	0.00
5	1.18	-	-	0.15	0.10	0.11

수검자 3은 우식 위험 치아만을 갖고 있었고, 질환 위험도에 따라 위생 상태 불량에 주의해야 할 가능성이 높다. 사실 우식 위험 치아가 위생 상태 불량을 초래한다기보다는 그 반대의 경우가 많겠으나, 평소 위생 상태가 불량한 수검자가 검진 당일 청결한 상태였을 가능성이 높으므로, 위생 상태 불량을 위험 질환으로 선별하는 것 역시 어느 정도의 타당성을 갖는다.

이러한 결과들을 볼 때, 본 논문에서 제안하는 피드백 방법이 건강검진 대상자 집단이 가지고 있는 건강 및 의료 관련 특성들을 적절히 찾아내어, 이를 토대로 한 개인화 피드백을 생성하는 것을 볼 수 있다.

하지만 앞의 수검자 1, 2와 수검자 3의 경우를 비교하였을 때, 질환 위험도의 값이 상대적으로 수검자 3은 작게 산출되고 있다. 이는 그림 2에서 알 수 있듯이, 수검자 3이 가진 유일한 질환인 우식 위험 치아와 같은 군집에 소속되는 질환이 없기 때문이다. 이는 같은 군집에 소속된 질환들끼리의 연관성을 계산할 때는 (1)에서와 같이 가중치 $\triangle$를 더해주게 되는데, 현재 질환의 수 T에 비해 상대적으로 k의 값이 많아서 $\triangle$ 값이 크기 때문에 일어나는 현상이다. T가 많아지는 경우는 이러한 문제점을 어느 정도 줄일 수 있겠으나, $\triangle$의 값을 적절히 정하는 방법이 향후 필요할 것으로 생각된다.

수검자 4는 아무런 치위생 질환도 관찰되지 않았다. 따라서 모든 질환에 대한 위험도가 0임을 볼 수 있다. 이러한 수검자의 경우에는 단순히 검진 대상 집단에서 빈발하는 질환을 위험 질환으로 선별할 수 있다. 현재 표 4의 대각선 셀들을 관찰했을 때, 일반적으로 가장 빈발한 질환이 D_2, 즉, 위생 상태 불량이므로, 수검자 4와 같은 고교생의 경우 일상적으로 구내 위생 상태가 불량해지지 않도록 주의할 것을 알려줄 수 있다.

수검자 5는 본 논문에서 제안하는 시스템의 한계를 보여준다. 표 7에서 이 수검자에 대하여 위험도가 가장 큰 질환은 D_1, 부정교합이다. 이는 부정교합과 같은 군집인 위생 상태 불량, 우식 치아 질환을 가지고 있기 때문이다. 하지만 상하악 맞물림이 바르지 못한 부정교합의 경우, 치열 및 골격계 형태에서 비롯되는 증상으로, 위생 상태나 우식 치아에 의해 유발된다고 보기 어렵다. 오히려 반대로 부정교합이 위생 상태 불량이나 우식 치아를 유발하는 경향이 있다. 따라서, 선별된 위험 질환을 기계적으로 제시하기 전에 의학적인 배경 지식을 통한 해석이 필요할 것으로 생각된다.

표 8. 관리 기여도가 높은 설문 항목

Table 8. Survey Questions with High Contribution Index

번호	항목	관리 기여도
3	치아가 쑤시거나 욱신거리며 아픔	0.69
1	치아가 깨지거나 부러짐	0.68
6	불쾌한 입 냄새가 남	0.66
13	간식 섭취 후 이를 잘 닦는지	0.65

끝으로 선별된 위험 질환에 대해서는 가장 관리 기여도가 높은 설문 항목들을 선택하여 위험 질환과 함께 제시한다. 예를 들어, 표 7에서 수검자 1, 2의 최고 위험 질환인 우식 치아의 경우, (4)에 의하여 관리 기여도가 큰 4개의 설문 항목은 표 8과 같다. 일상생활에서 치아 통증이나 파절, 불쾌한 입 냄새와 같은 자각 증상이 있는 경우, 우식 치아 진단을 받아볼 필요가 있으며, 우식 치아 방지를 위해서는 특히 간식 섭취 후 이를 잘 닦는 것이 좋다는 것을 추천하고 있다. 따라서 1회성의 건강 검진이 되기보다 검진 이후 평소 생활 및 습관에 대한 개인화된 피드백이 가능함을 볼 수 있다.

www.kci.go.kr

Ⅴ. 결론 및 추후 연구 과제

건강검진이 활성화되어 학교나 직장과 같은 조직에서 집단 건강검진을 정기적으로 실시하고 있어, 국민 건강관리에 상당한 성과를 내고 있는 것이 사실이나, 시간 및 비용의 한계도 존재하는 부분이 있는 것이 현실이다. 특히, 대상 집단의 특성을 반영하기 어려웠고, 설문을 통한 문진의 경우 형식적인 절차로 진행되기 쉬우며, 체계적인 사후 관리가 미비한 점이 있었다.

본 논문은 이러한 점을 보완하기 위하여, 집단 검진 실시 후, 대상 집단의 특성을 파악하고, 이에 기반하여 일상생활에서의 자각 증상이나 습관과 연계된 개인화된 건강검진 피드백 시스템을 제안하고 있다. 제안하는 시스템은 데이터 마이닝 기법인 계층군집 분석 방법을 이용하여 대상 집단의 특성을 파악하고, 이를 위험 질환 및 개인 유의 사항을 도출할 수 있었다. 이러한 시스템의 특성은 건강검진 시 문진 담당의 인력 부족 문제를 경감하고, 건강검진의 효과를 증대하는데 도움을 줄 것으로 기대된다.

한편, 앞에서도 지적하였듯이, 기술적으로 △ 값을 조절하여 질환 간 연관도 및 질환 위험도 산출 방법을 조정하는 것은 앞으로 보완되어야 하며, 산출된 정보들을 의학적인 지식을 통해 좀 더 의미적으로 가공하여 유용하게 만들 필요성도 존재한다. 그 외 부가적으로는 계층군집 분석 과정에서 유사도 척도로 단순 동시 발생 빈도 외 다른 여러 가지 척도의 개발 및 적용과 함께, 본 논문에서 주로 이진 문항만을 고려한 것에서 나아가 다진 문항들로 이루어진 검진 항목들을 다루는 문제 등이 추후 연구 과제로 남아 있다.

참고문헌

[1] 손희배, 김민수, 이영철, “RFID를 이용한 헬스케어 자가 진단 지능형 시스템 구현,” *한국지능시스템학회 논문지*, 제20권, 제1호, pp.146-152, 2010.

[2] Maciosek, M.V., Coffield, A.B., Edwards, N.M., Flottemesch, T.J., Goodman, M.J., and Solberg, L.I., "Priorities among Effective Clinical Preventive Services: Results of a Systematic Review and Analysis," *American Journal of Preventive Medicine*, Vol.31, No.1, pp.52-61, 2006.

[3] 최은진, “수요자 중심의 국가 건강검진 사업 운영 방안,” *보건복지포럼*, 제163호, pp.16-26, 2010.

[4] 최령, 황병덕, “건강보험 건강검진 대상자들의 예방적 의료서비스 이용 특성,” *한국콘텐츠학회논문지*, 제11권, 제2호, pp.331-340, 2011.

[5] 김금련, “12초만에 건강검진을 하라구요?,” *월간 인물과 사상*, 2002년 6월호, pp.194-197, 2002.

[6] Klosgen, W., and Zytkow, J., "Handbook of Data Mining and Knowledge Discovery," Oxford University Press, New York, 2002.

[7] Tan, P.-N., Steinbach, M., and Kumar, V., "Introduction to Data Mining," Addison Wesley, 2005.

[8] Ngai, E.W.T., Li, X., and Chau, D.C.K., "Application of Data Mining Techniques in Customer Relationship Management: A Literature Review and Classification," *Expert Systems with Application*, Vol.36, pp.2592-2602, 2009.

[9] Romero, C., and Ventura, S., "Educational Data Mining: A Survey from 1995 to 2005," *Expert Systems with Application*, Vol.33, pp.135-146, 2007.

[10] Ghazavi, S.N., and Liao, T.W., "Medical Data Mining by Fuzzy Modeling with Selected Features," *Artificial Intelligence in Medicine*, Vol.43, pp.195-206, 2008.

[11] Murtagh, F., "A Survey of Recent Advances in Hierarchical Clustering Algorithms," *The Computer Journal*, Vol.26, No.4, pp.354-359, 1983.

[12] 김준우, 주지영, 홍성용, 이문용, 윤완철, “효과적인 학습 전략을 위한 컨텐츠 주제어 군집 방법,” *한국정보과학회 제37회 추계학술대회 논문집*, 제37권, 제2호, pp.317-322, 2011.

[13] Tsui, C.-J., Wang, P., Fleischman, K.R., Sayeed, A.B., and Weinberg, A., "Building an IT Taxonomy with Co-occurrence Analysis: Hierarchical Clustering and Multidimensional Scaling," *Proceedings of iConference*,

www.kci.go.kr

pp.247-256, 2010.
[14] Han, E.-H., Karypis, G., Kumar, V., and Barnshad, M., "Clustering based on Association Rule Hypergraphs," *Proceedings of SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1997.
[15] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J., "Item-based Collaborative Filtering Recommendation Algorithms," Proceedings of the 10th International Conference on World Wide Web, pp.285-295, 2001.
[16] Yun, C.-H., Chuang, K.-T., and Chen, M.-S., "Clustering Item Data Sets with Association Taxonomy Similarity," Proceedings of the Third IEEE International Conference on Data Mining (ICDM '03), pp.697-700, 2003.

감사의 글

이 논문은 동아대학교 교내연구비 지원에 의하여 연구되었음.

저자소개

강현경(Hyun-Kyung Kang)

2000년 한국방송통신대학교 경영학과 졸업 (경영학사)
2004년 고신대학교 보건관리학과 졸업 (보건학석사)
2008년 고신대학교 의학과 졸업 (의학박사)

2005년~2007년 동주대학 치위생과 전임강사
2008년~2009년 동주대학 치위생과 조교수
2010년~현재 신라대학교 치위생학과 전임강사
※ 관심분야 : 구강보건, 치면세마, 치주학, 치과방사선학, 포괄치위생학

김준우(Jun-Woo Kim)

2001년 한국과학기술원 산업공학과 졸업(공학사)
2003년 한국과학기술원 산업공학과 졸업(공학석사)
2009년 한국과학기술원 산업 및 시스템공학과 졸업(공학박사)

2009년~2010년 한국기술교육대학교 산업경영학부 대우교수
2011년~현재 동아대학교 산업경영공학과 조교수
※ 관심분야 : 데이터마이닝, 지능형 시스템, 서비스관리, Operations Research, e-러닝, 융합 콘텐츠

www.kci.go.kr