데이터마이닝 분석기법을 활용한 한방 설문 최적화

김영태*, 박상찬*, 이상철**, 유승연***, 박영재***, 박영배***, 임광혁****

요약

본 연구는 공학 분야에서 데이터 분석에 적용되던 데이터마이닝 방법론을 이용하여 한방에서 사용하는 변증진단 설문지의 중요문항을 찾아내기 위한 방법론을 개발하는 것을 목표로 한다. 체계적인 설문 데이터 분석을 통하여 설문 문항을 축소할 수 있다면, 최소한의 설문을 통하여 고객의 특성을 미리 파악할 수 있어 한방 서비스의 효율화 및 만족도 향상을 기대할 수 있다. 본 연구에서는 한방의 병인론에서 사용하는 평위산 병증 설문데이터를 이용하였으며, 군집분석을 이용하여 병증을 가지고 있는 경우와 그렇지 않은 2개의 그룹으로 분류하고, 의사결정나무 분석을 이용하여 두 개의 군집을 구분할 수 있는 중요한 규칙을 도출하여 평위산 변증에 관한 최적화 된 설문문항을 구성하였다.

Optimization of Oriental Medicine Survey Using Data Mining Techiniques

Young-Tae Kim*, Sang-Chan Park*, Sang-Chul Lee**, Seung-Yeon Yoo***, Young-Jae Park***, Young-Bae Park*** and Kwang-Hyuk Im****

ABSTRACT

The purpose of this study is to optimize the survey items for oriental medicine, appling data mining techniques used in the fields of engineering. If survey items can be minimized through systematic survey data analysis, customer attributes may be identified in advance to increase service efficiency and customer satisfaction in oriental medicine. In this study, cluster analysis was conducted to classify the groups into one with Pyungwi-san symptom and the other without the symptom. Subsequently, the rules were determined by clustering with higher scale response through decision tree analysis to construct the optimized survey for Pyungwi-san symptom.

Key Words : Oriental Medicine Survey, Clustering, Classification, Data mining

* 경희대학교 의료경영학과 (✉smarthealth@hanmail.net)
** 그리스도대학교 경영학부
* ** 경희대학교 한의과대학 진단 · 생기능의학 과학 교실
* *** 배재대학교 전자상거래학과
· 제1저자(First Author) : 김영태 · 교신저자(Correspondent Author) : 임광혁
· 접수일(2011년 9월 9일), 수정일(1차 : 2011년 10월 27일), 게재확정일(2011년 10월 31일)

Ⅰ. 서 론

한의학은 내적인 생명력을 길러 환자의 건강을 증진시킨다는 치료 이념을 가지고 있다. 그 뿐 아니라 치료 방법이 개인 의학적이고 웅변 주의적이기 때문에 의사에 따라 같은 환자에 대한 진단 결과의 차이가 클 수 있다[1]. 따라서 한방 의료 체계의 효율성을 측정하는 것이 어렵고, 환자의 생활 습관이나 병력 등과 같이 고려해야 할 변수가 많다. 그러나 다양한 한방 분야에서의 연구들은 각각의 목적에 따른 관찰과 분석만 이루어 졌을 뿐, 연구들 사이의 연관성 및 연구결과의 구조화, 시간에 따른 추이를 살피는 등의 시계열 분석이 이루어지지 못하고 있다.

위에서 언급한 대로 한의학에서 진단을 위해서는 환자의 생활 습관이나 병력 등 다양한 데이터가 필요하고, 이러한 데이터는 대부분 설문을 통하여 수집되고 있다. 그러나 설문 문항에 대한 과학적인 설계, 수집된 데이터의 체계적인 분석 및 활용은 제대로 이루어지지 못하고 있다[2].

본 연구는 공학 분야에서 데이터 분석에 적용되던 데이터마이닝 방법론을 이용하여 한방에서 사용하는 변증진단 설문지의 중요문항을 찾아내기 위한 방법론을 개발하는 것을 목표로 한다. 수집된 설문데이터의 체계적인 분석을 통하여 설문 문항을 축소할 수 있다면, 최소한의 설문을 통하여 고객의 특성을 미리 파악할 수 있고, 이를 통하여 궁극적으로는 한방 서비스의 효율화 및 고객 만족도 향상을 기대할 수 있을 것이다.

Ⅱ. 관련 연구

한방 설문지 개발과 관련된 연구로는 사상변증내용 설문조사지, 사상체질분류검사지(QSCC), 한열변증설문지 개발, 음허증 측정도구의 개발 및 신뢰도 타당도 검정, 어혈변증설문지 개발, 담음변증설문지 개발 등이 있다[3][4]. <표 1> 은 위에서 설명한 한방 설문 개발과 관련된 연구의 저자 정보, 연구 제목, 발표 연도를 정리하여 제시한 표이다.

표 1. 한방설문개발 관련 연구

Table 1. Oriental Medicine Survey development related research

Author	Title	Year
이의주 외2인	사상변증내용 설문조사지의 타당화 연구	1995
김선호 외2인	사상체질분류검사지(QSCCⅡ)의 표준화 연구	1966
이정찬 외2인	사상체질분류검사지(QSCCⅡ)의 타당화 연구	1966
김태연 외4인	사상체질분류검사지(QSCCⅡ)의 Upgrade 연구	2003
김영우	사상체질진단을 위한 사상체질분류검사지Ⅱ(QSCCⅡ)의 연구	2003
김숙경 외2인	한열변증 설문지 개발을 위한 타당성 연구	2002
백태선 외4명	한열변증 설문지와 일반적 건강검진 결과와의 상호 연관성에 관한 비교 연구	2005
권신애 외8명	어혈변증 설문지를 통한 오십견의 어혈변증 평가 및 통증, 견관절 운동범위와 어혈변증과의 관계	2011
양동훈 외3명	瘀血辨證說問紙 開發 (Development of Questionnaires for Blood Stasis Pattern)	2006
박재성 외4명	痰飮辨證 說問 開發	2006

변증진단 설문지 개발과 관련된 기존의 연구들은 대부분 한방과 관련된 기존의 서적을 이용하여 설문문항을 만들고, 요인분석을 이용하여 개발된 설문문

항의 타당도를 검증하는 방법을 이용하였다. 그러나 한방에서 사용하는 변증진단 설문지 개발에 있어서 중요한 논쟁이 있다. 첫째, 많은 설문문항을 최소화 하는 것이다. 예를 들어, 팔체질 설문문항의 경우에는 설문지가 200여개 문항으로 이루어져 있다. 따라서 어떻게 하면 설문문항을 최소화 하는 것이 문제점으로 대두되었다. 두 번째, 설문문항을 최소화하기 위해서 요인분석을 사용하거나, 아니면 평균차이검정을 통해 통계적으로 차이가 있는 설문문항을 선택해서 문항을 줄여왔다. 그러나 요인분석과 평균차이검정을 통해서 줄여진 설문문항이 변증을 진단하는데 중요한 설문문항이라고 판단할 수는 없다. 따라서 최적화된 설문문항을 추출하기 위해서는 공학적 기법들을 이용하여 고객의 특성 및 건강 상태 파악을 위한 체계적이고 효과적인 방법론을 이용한 연구가 필요하다[5][6].

III. 한방 설문 최적화

데이터마이닝 기법을 이용한 한방 설문 축소 모델은 아래 <그림 1>과 같다.

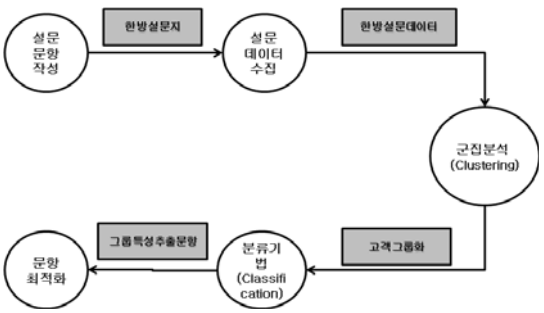

그림 1. 한방설문문항 최적화 프로세스

Fig. 1. The optimization process of oriental medicine survey

먼저 설문조사를 하기 위해서는 관련된 질병을 판단하기 위한 설문 문항을 작성하게 된다. 이렇게 작성된 설문지를 이용하여 설문을 실시하여 설문데이터를 수집한다. 수집된 설문 데이터를 데이터마이닝 기법을 이용하여 분석하기 위해서는 먼저 데이터 정제작업을 수행하고, 정제된 데이터를 분석할 수 있는 데이터 형식으로 변환하여 적재하는 작업이 이루어져야 한다. 이렇게 변환된 데이터를 이용하여 군집분석(Clustering)과 분류기법 (Classification) 을 이용하여 유용한 지식을 추출하기 위한 데이터 분석을 수행한다.

군집분석은 데이터의 특성을 분석하여 비슷한 특성을 가진 데이터들끼리 그룹을 형성해주는 데이터마이닝 기법이다. 한방 설문 데이터에 군집분석 기법을 적용하면 비슷한 특성을 가진 설문 응답자간에 그룹을 형성하게 된다. 보통 의료데이터는 설문을 수행한 해당 병증을 가지고 있는 경우와 그렇지 않은 경우 2개 그룹으로 구분하게 된다. 군집 분석의 여러 기법들 중에서 분석 속도가 빠르고 군집의 수를 미리 지정하여 분석할 수 있어서 가장 많이 사용되는 K-means 알고리즘을 사용하였다.

분류기법은 2개 이상의 그룹이 있을 때 그룹에 속한 데이터들의 특성을 추출하는 데이터마이닝 기법이다. 한방 설문 데이터에 군집분석 기법을 적용하여 2개 그룹으로 나누어진 데이터를 분류 기법의 입력으로 넣게 되면 2개 그룹을 구분하는 특성을 파악할 수 있다. 분류 기법의 여러 방법들 중 수집된 데이터의 레코드들을 분석하여 속성(Attribute)의 조합으로써 이들 사이에 존재하는 패턴에 대한 분류 규칙을 생성해 내고 이를 트리 형태로 나타내는 의사 결정 나무 (Decision Tree) 기법을 활용하였다. 의사 결정 나무 생성 알고리즘으로는 변수의 제약도 없고, 다지 분리도 가능하며, 질적 변수와 양적변수 모두를 예측변수로 사용할 수 있기 때문에 가장 널리 사용되고 있는 C 5.0을 이용하였다. 모형개발을 위한 데이터마이닝 툴은 SPSS 클레멘타인 10.1을 사용하였다.

Ⅳ. 한방 설문 최적화 실험

본 연구에서는 기존에 한방에서 사용하는 평위산 변증 설문지를 이용하였다[3][4]. 평위산 변증 설문지는 19문항으로 구성되어 있고, 리커트 형식의 7점 척도로 구성되어 있다. 설문은 성인 71명을 대상으로 실시되었으며 남성이 17명, 여성이 54명으로 여성의 비율이 상대적으로 높았다. 연령은 22~65세로 나타났다. 평위산 변증 설문지는 <표 2>와 같다[5].

표 2. 평위산 설문지

Table 2. Survey questions of Pyungwi-san

Item	Description
1	명치끝을 눌렀을 때 아프십니까?
2	식사 후에 배가 더부룩하십니까?
3	평소 식탐이 있습니까?
4	눕고만 싶고 만사가 귀찮습니까?
5	소변을 자주 보십니까?
6	몸이 잘 붓습니까?
7	팔다리에 관절통이 있습니까?
8	자주 체하지는 않습니까?
9	다른 지역에 가서 물갈이 하면 복통, 설사를 하십니까?
10	특정 음식에 두드러기가 난적이 있습니까?
11	술 마신 후 배가 더부룩하거나 설사를 하십니까?
12	식사 후 바로 배가 아프십니까?
13	식사 후 바로 대변을 보십니까?
14	트림을 자주 하십니까?
15	평소 대변이 묽은 편에 속합니까?
16	체중이 점점 늘어납니까?
17	식사 후에 피곤이 더 심해지십니까?
18	속이 메슥거려 구역감이 있습니까?
19	신물이 올라오십니까?

평위산 설문응답 결과를 군집 분석한 결과 두 개의 군집으로 분류되었다. 군집 1에는 27명, 군집 2에는 41명이 속했다. 3명은 무응답 항목이 존재하여 제외되었다. 군집 분석 결과는 <표 3> 과 같다.

표 3. 평위산 설문 군집분석 결과

Table 3. Results of clustering analysis

군집1	군집2	제외
27	41	3

<표 4> 는 군집 분석을 통해 얻은 군집별 설문문항별 평균값을 나타내고 있다.

표 4. 평위산 설문 군집별 평균값 비교

Table 4. Comparison between the average of groups

문항	군집1	군집2
1	4.74	3.51
2	4.93	3.88
3	4.48	4.71
4	4.37	4.15
5	4.48	3.24
6	4.04	3.44
7	4.11	2.63
8	5.33	2.80
9	3.37	2.15
10	2.89	1.85
11	4.33	3.27
12	3.04	1.63
13	2.93	2.02
14	4.41	3.51
15	3.81	2.95
16	4.33	3.51
17	4.63	3.73
18	4.11	1.83
19	4.26	1.98

군집별 설문문항별 평균값으로 두 군집의 특성을 살펴보면 [군집 1]에 해당하는 설문 응답자들이 평위

산 변증에 관한 질문에 보다 높은 척도로 응답한 군집임을 알 수 있다. 특히 문항 7, 8, 12, 18, 19번 문항은 비교적 두 군집 간 값의 차이가 크게 나타났음을 알 수 있다. 하지만 이러한 평균값 비교를 통하여 얻어진 결과는 해당 문항이 평위산의 증상을 예측하는 데에 있어서 중요한 문항이라고 판단하기에는 통계적으로 오류가 존재할 가능성이 있다.

그래서 설문문항 19개에 대한 설문 응답자 데이터가 군집 분석을 통해 도출된 2개의 군집으로 결정되어지는 규칙(Rule)을 추출하기 위해 의사결정나무 분석을 수행하였고, 의사결정나무 분석결과 <그림 2>의 트리가 도출되었다.

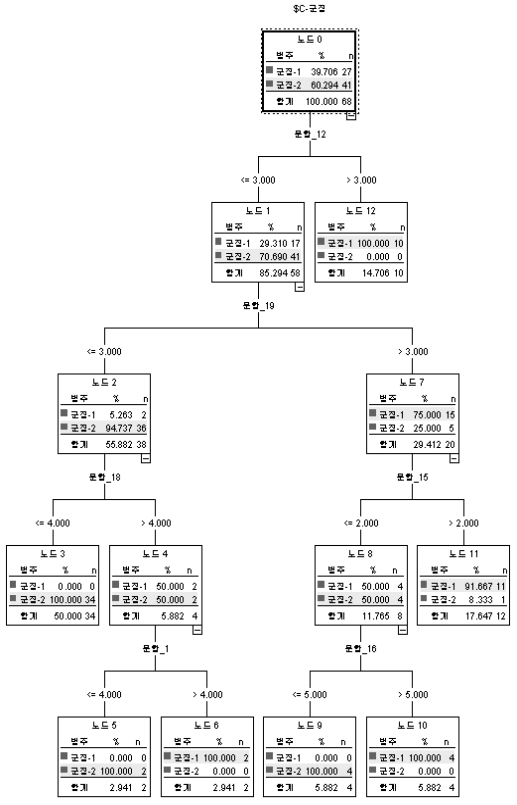

그림 2. 평위산 군집 분석 후 의사결정트리

Fig. 2. The decision tree

분석 결과를 살펴보면 설문 응답자가 군집 1 즉, 평위산 변증에 관한 질문에 보다 높은 척도로 응답한 군집으로 결정되는데 있어 가장 중요한 문항은 12번 문항이라고 할 수 있다. 12번 문항을 4점 척도 이상으로 응답한 응답자 10명은 모두 [군집 1]에 속하였다. 이는 [군집 1]에 속한 27명 중 약 37%에 해당되는 숫자이다. 이어 19번, 15번, 16번 문항을 통해 평위산 변증에 관한 질문에 보다 높은 척도로 응답한 군집으로의 결정이 완료된다.

아래 <표 5>는 의사결정나무 분석을 통하여 얻어진 평위산 변증에 관한 질문에 보다 높은 척도로 응답한 군집[군집 1]으로 결정되는 규칙(Rule)을 정리한 표이다.

표 5. 평위산 군집 결정 규칙

Table 5. Decision rules Pyungwi-san cluster

규칙 1	if 문항_12 > 3
	then 군집-1, n=10
규칙 2	if 문항_12 <= 3
	and 문항_19 > 3
	and 문항_15 > 2
	then 군집-1, n=11
규칙 3	if 문항_12 <= 3
	and 문항_19 > 3
	and 문항_15 <= 2
	and 문항_16 > 5
	then 군집-1, n=4
규칙 4	if 문항_12 <= 3
	and 문항_19 <= 3
	and 문항_18 > 4
	and 문항_1 > 4
	then 군집-1, n=2

규칙 1은 12번 문항의 응답 척도가 3을 초과하면 모두 [군집 1]에 해당함을 나타내며 그 수는 10명이다. 규칙 2는 12문항의 응답척도가 3이하, 19번문항의 응답

척도가 3을 초과 그리고 15번문항의 응답척도가 2를 초과하면 모두 [군집 1]에 해당하며 그 수는 11명이다.

규칙 3은 12문항의 응답척도가 3이하이고, 19번문항의 응답척도가 3을 초과하고, 15번문항의 응답척도가 2이하이면서 16문항의 응답척도가 5를 초과하면 모두 [군집 1]에 해당하며 그 수는 5명이다. 규칙 4는 12문항의 응답척도가 3이하이고, 19번문항의 응답척도가 3이하이고, 18번문항의 응답척도가 4를 초과하면서 1번문항의 응답척도가 4를 초과하면 모두 [군집 1]에 해당하며 그 수는 2명이다. 이처럼 위의 4개의 규칙 내에서 평위산 변증에 관한 질문에 보다 높은 척도로 응답한 군집[군집 1]으로의 결정이 이루어진다. 다만 규칙 4의 경우 의사결정나무에서 다른 규칙들과 달리 다른 노드로의 진행에 놓여져 있고 그 수가 적어 규칙의 중요도가 떨어짐을 예상 할 수 있다.

이처럼 평위산 변증 설문결과를 토대로 평위산 변증에 관한 질문에 보다 높은 척도로 응답한 군집과 그렇지 않은 군집으로 나누어 보았을 때, 평위산 변증 설문 응답자가 높은 척도로 응답한 군집으로 결정되는 규칙은 위의 1~3의 규칙 내에서 결정되면 규칙을 구성하고 있는 문항은 아래 <표 6>과 같다. 즉, 두 그룹을 구분할 수 있는 최소의 설문 문항이라 할 수 있다.

표 6. 평위산 군집 결정 주요문항

Table 6. The main questions to determine Pyungwi-san cluster

Item	Description
12	식사 후 바로 배가 아프십니까?
19	신물이 올라오십니까?
15	평소 대변이 묽은 편에 속합니까?
16	체중이 점점 늘어납니까?

군집별 평균값 비교를 통해 군집 간에 큰 차이를 보였던 문항과 차이가 있다. 12문항과 19문항은 의사결정나무 분석을 통해 도출된 군집결정 주요문항에도 포함이 되었지만 7번, 8번, 18번 문항은 통계적으로 군집간의 평균값 차이가 크긴 하였지만 군집을 결정하는 주요한 문항으로 볼 수는 없다.

결과적으로 평위산 변증에 관한 설문을 통하여 평위산 증상을 보이는 환자를 구분하고자 한다면 설문지를 구성하고 있는 1~19문항 모두 다 필요한 것은 아니라고 할 수 있다. 12번, 19번, 15번, 16번 문항만으로도 평위산 증상을 보이는 환자를 충분히 구분할 수 있으며 문항의 가중치를 고려했을 때, 위 네 가지 문항을 순서대로 질문한다면 훨씬 더 효율적으로 환자를 구분할 수 있다.

IV. 결론 및 추후 연구과제

산업공학에서 데이터 분석에 적용되던 데이터마이닝 방법론을 한의학 분야의 진료 및 진단에 참고자료로 사용하는 설문 데이터 분석에 적용한다면, 고객을 특성에 맞게 고객군으로 분류하여 각 고객군 별로 맞춤형 서비스를 제공할 수 있는 기반을 마련할 수 있다.

한방서비스 관련 설문조사 문항의 경우, 기존에 사용되는 팔체질 문항이 체질별로 200여개 문항으로 이루어져 있어 문항의 수가 환자가 응답하기에 너무 많다는 지적이 있어 왔다.

본 연구는 데이터마이닝 방법론을 적용한 한방설문 축소모형을 제안하였다. 제안 모형을 사용하여 평위산 설문데이터를 분석한 결과 19문항의 설문 중에서 중요 설문문항 4가지를 추출할 수 있었다. 이는, 고객에게 19개 모든 문항을 질문하지 않고 4가지 문항의 질문만으로 해당 고객의 평위산 관련 병증을 판단할 수 있다는 의미 있는 결과이다. 본 연구와 같이 체계적인 설문 데이터 분석을 통하여 현재 존재하는 설문지 또는 앞으로 작성할 설문지의 설문 문항을 효과적으로 최적화할 수 있다면, 최소한의 설문을 통하여 고객의 특성을 파악할 수 있어 한방 서비스의 효율화 및 만족도 향상을 기대할 수 있을 것이다.

참고문헌

[1] 김용남, "병원 양.한방 협진체제의 분석 연구", 박사학위 논문, 원광대학교, 2001.

[2] 임광혁, 박상찬, 이상철, 유승연, 박영배, "데이터마이닝 기법을 활용한 한방 설문 축소 모형", 한국 지식정보기술학회 추계학술대회 논문집, 제 5권, 제2호, pp.23-26, 2011.

[3] 하성룡, 김민용, 박영재, 박영배, "한열성향에 따른 위전도 특성 연구", 대한한의진단학회지, 제12권, 제1호, pp.131-141, 2008.

[4] 유승연, 박영재, 박영배, "심박변이도와 호흡변이도의 상관성 연구", 대한한의진단학회지, 제12권, 제2호, pp.74-83, 2008.

[5] 임준성, 박영배, 박영재, 이상철, 오환섭, "병인론적 분석에 의한 평위산 변증 설문지의 신뢰도 타당도 연구", 대한한의진단학회지, 제11권, 제2호, pp.59-67, 2007.

[6] S. Tsumoto, "Automated knowledge acquisition from clinical databases based on rough sets and attribute-oriented generalization", Proceedings of the AMIA Symposium, Vol. 1, pp. 548-552, 1998.

[7] T.C. Edwards Jr., D.R. Cutler, N.E. Zimmermann, L. Geiser and G.G. Moisen, "Effects of sample survey design on the accuracy of classification tree models in species distribution models", Ecological Modelling, Vol. 199, Issue 2, pp. 132-141, 2006.

감사의 글

이 논문은 2011년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(과제번호-20110013034).

저자소개

김영태(Young-Tae Kim)

2008년 동서대학교 상경대학 경영정보학과 졸업(학사)
2011년 경희대학교 대학원 의료경영학과 졸업(석사)
2011년~현재 경희대학교 대학원 의료경영학과 박사과정
※ 관심분야 : 데이터마이닝, 지식서비스, 스마트 헬스

박상찬(Sang-Chan Park)

1984년 서울대학교 경영학과 졸업(학사)
1985년 미국 일리노이 주립대학 졸업(경영학석사)
1991년 미국 일리노이 주립대학 졸업(경영학박사)
1989년~1995년 미국 위스컨신 주립대학 조교수
1995년~2009년 한국과학기술원 산업공학과 교수
2009년~현재 경희대학교 의료경영학과 교수
※ 관심분야 : 지식기반서비스, 데이터마이닝

이상철(Sang-Chul Lee)

1995년 아세아연합신학대학교 아세아학과졸업(문학사)
1998년 경희대학교 경영학과 졸업(경영학석사)
2004년 경희대학교 경영학과 졸업(경영학박사)
2004년 한국과학기술원 경영공학 위촉연구원
2005년~현재 그리스도대학교 경영학부 조교수
※ 관심분야 : 경영정보시스템, 데이터마이닝, 전자상거래, 품질경영

유승연(Seung-Yeon Yoo)

2007년 경희대학교 한의학과 졸업(학사)
2010년 경희대학교 한의학과 졸업(한의학석사)
현재 경희대학교 한의학과 한의학박사 과정 중
2008년~2010년 경희대학교 한방병원 전문수련의(침구과)
2011년~현재 강동경희대학교병원 건강증진클리닉 임상강사
※ 관심분야 : 한방건강평가, 미병, 건강증진

박영재(Young-Jae Park)

1989년 경희대학교 한의학과 졸업(학사)
1997년 경희대학교 한의학과 졸업(한의학석사)
2002년 경희대학교 한의학과 졸업(한의학박사)
2002년~2004년 세명대학교 한의과대학 침구과 전임강사
2010년~현재 경희대학교 한의과대학 부교수
2006년~현재 강동경희대학교 진단·생기능의학실 과장
※ 관심분야 : 한방건강평가, 미병, 건강증진

박영배(Young-Bae Park)

1977년 경희대학교 한의학과 졸업(학사)
1980년 경희대학교 한의학과 졸업(한의학석사)
1985년 경희대학교 한의학과 졸업(한의학박사)
1997년 대한한의학회 이사
2000년~2004년 대한한의진단학회 회장
1995년~현재 경희대학교 한의과대학 교수
경희대학교 진단생기능의학과장
※ 관심분야 : 한방건강평가, 미병, 건강증진

임광혁(Kwang-Hyuk Im)

1995년 한국과학기술원 전산학과 졸업(공학사)
2000년 한국과학기술원 산업공학과 (공학석사)
2006년 한국과학기술원 산업공학과 (공학박사)
2006년～2008년 : 삼성전자 반도체연구소 책임연구원
2008년～현 재 : 배재대학교 전자상거래학과 교수
※ 관심분야: 지식서비스, 경영정보시스템, 데이터마이닝, 전자상거래, 고객관계관리, 공급사슬관리