데이터 마이닝 기법을 이용한 직무교육 성취집단 예측모형 개발

곽기효, 서용무[a]

[a]고려대학교 경영학과

136-701, 서울시 성북구 안암동 5가 1

Tel:+82-2-3290-2521, Fax:+82-2-922-7220, E-mail: {khkwak, ymsuh}@korea.ac.kr

Abstract

국방부에서 발표한 '국방개혁에 관한 법률' 에 따라 2014년까지 현역병들에 대한 복무기간이 단계적으로 단축될 예정이다. 이에 따라 좀 더 효율적인 직무교육 방안이 필요하게 되어, '차등제 교육'을 시행하고 있다. 이 교육의 효과를 향상시키기 위해서는 훈련병들의 예상 학업 성취도를 미리 정확하게 예측하는 것이 필수적이다. 따라서, 본 연구에서는 입교 초기에 얻을 수 있는 신병들의 제한된 자료들을 이용하여 교육 성취도 예측 모형을 개발하였다.

본 모형의 목적 변수는 '일반관리 인원', '집중관리 인원' 의 값을 갖는 이진형 성취집단 변수이며, 사용된 기법은 *k*-means 군집기법과 Decision Tree 기법을 혼합한 모형, *k*-means 군집기법과 Neural Network 기법을 혼합한 모형, Decision Tree 모형, Neural Network 모형, Bayesian 모형, 그리고 Logistic 모형 등을 사용하였다. 그 결과 *k*-means 군집기법과 Decision Tree를 혼합한 모형이 가장 좋은 예측력을 보이는 것으로 나타났다. 이러한 교육 성취집단 예측 모형은 향후 군에서 이루어지는 다양한 교육 프로그램에 적극적으로 이용될 수 있을 것으로 기대된다.

Keywords:

Decision Tree; Neural Network; K-means clustering; 직무교육; 교육성취도

1. 서론

21세기 지식정보사회에 이르러 인적자원개발은 조직전략의 일부분으로서 점차 조직의 핵심적 과제로 떠오르고 있으며, 그 중에서 특히 교육훈련은 조직의 가치창출과 연관되어 조직 목표 달성의 핵심요소로 그 중요성이 더욱 증대되기 시작하였다. 이러한 교육훈련 중에서도 특히 직무에 대한 전문성 향상과 직무몰입을 위해 실시하고 있는 교육 중 하나가 직무교육이다[7]. 육군에서는 이러한 병사 직무교육의 일환으로 신병교육 후 각 특기 별로 전문적인 지식을 교육하는 '후반기 교육' 을 실시하고 있다. '국방개혁에 관한 법률' 에 따르면 현역병에 대한 복무기간이 2014년까지 단계적으로 단축될 예정이다[1]. 이것은 병사들의 직무 숙련도를 향상 시킬 수 있는 시간이 절대적으로 부족해 진다는 것을 의미하는데 이에 따라 후반기 교육을 통한 직무교육의 중요성이 점차 증대 될 것이다.

현재 후반기 교육을 포함한 신병교육은 '교육훈련 성과 증대 제도' (연합뉴스, 2005.5.30)에 따라 차등제 교육으로 실시되고 있다. 차등제 교육이란 훈련병들을 두 집단(일반관리 인원과 집중관리 인원)으로 구분하여 일반관리 인원에게는 원칙에 따라 강도 높은 훈련을 실시하고, 집중관리 인원에게는 1:1 정밀 지도, 정신교육 강화, 그리고 별도 훈련 프로그램 편성 등의 방법을 적용하여, 전반적인 훈련 효과를 증대시키는 것을 말한다. 본 논문에서는 이러한 차등제 교육 방법의 훈련 효과를 향상시키기 위하여 데이터 마이닝 기법을 이용한 교육성취도 예측 모형

을 개발하고자 한다. 이러한 예측은 교육실시 이전에 차등교육의 대상자를 일반 관리 인원과 집중 관리 인원으로 분류함으로써 차등제 교육의 효과를 향상 시킬 수 있을 것으로 기대된다.

지금까지의 교육성취도 예측 모형의 개발에는 통계적 기법을 이용한 연구가 주를 이루었으며 데이터마이닝 기법은 많이 사용되지 않았다. 따라서 본 연구에서는 후반기 교육에 입소하는 신병들의 교육 성취도 예측을 위하여 기초속성 자료를 사용한 예측모형을 개발하려고 하며, KMDT(*k*-means 군집기법과 Decision Tree 기법의 혼합모형), KMNN(*k*-Means 군집기법과 Neural Network의 혼합모형), Neural Networks 모형, Decision Tree 모형, Bayesian 모형, 그리고 Logistic 모형 등 6개 모형을 개발하여 이들의 성능을 비교하여 가장 우수한 모델을 찾고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 직무교육과 교육성취도에 관련된 기존의 연구를 살펴보았으며 데이터 마이닝 기법들, 특히 *k*-means 군집기법과 혼합기법들에 대한 문헌 연구를 수행하였다. 3장에서는 본 연구에서 개발한 분류 예측모형의 개발 과정에 대해서 기술하였다. 4장에서는 개발된 분류 예측 모형을 이용한 실험과 그 결과에 대해서 기술하였고, 마지막으로 5장에서는 본 연구의 결론과 함께 향후 연구방향을 논의하였다.

2. 문헌연구

2.1. 교육 성취도와 관련된 기존연구

교육성취도와 관련된 연구로는 다층모형(Multilevel models)을 이용하여 학업 관련변수들과 교육성취도와의 연관성을 분석하는 연구가 다수 진행되었다[4, 6, 14]. 교육성취도에 데이터마이닝 기법을 적용하려는 연구는 배재호[3]와 김혜숙 등[2]에 의해 수행되었다. 배재호[3]는 고등학생을 대상으로 1학기 성적과 학원수강 여부, 그리고 수업시간의 학습태도를 이용하여 2학기 성적을 예측하는 의사결정나무모형을 개발하였으며, 김혜숙 등[2]은 방학 중 학습변수를 추가적으로 포함시키고, 의사결정나무 기법과 연관규칙을 함께 사용한 모형을 개발하였다.

이상에서와 같이, 지금까지의 관련 연구들은 주로 아동 및 청소년기 학생들에 대한 교육성취도에 집중되어 있으며 직무교육에 대한 성취도 연구는 그 수가 매우 적을 뿐만 아니라 사용된 기법들도 다층 모형과 같은 통계적 기법 또는 제한적인 데이터 마이닝 기법에만 국한되어 있다. 전술한 바와 같이 군 복무기간의 단축이라는 시대적인 변화에 따라, 군에서의 직무교육의 효율성을 높이기 위하여 다양한 데이터 마이닝 기법을 활용할 필요가 충분히 있다.

2.2. 분류모형

분류란 클래스가 알려진 많은 개체들에 대한 정보가 주어진 상황에서 새로운 개체가 기존의 어떤 클래스에 속할 것인가를 예측하는 것을 말한다[22]. 이러한 분류 모형의 개발에 관한 연구에서는 대체적으로 인공지능 기반 기법과 통계적 기반 기법을 이용하거나 이들의 혼합 기법을 이용하였다[13, 17, 21].

k-means 군집 기법은 비계층적 군집 분석의 대표적인 기법으로, 구축하려는 군집의 개수, 즉 *k*가 입력 데이터와 함께 주어지면 *k* 개의 군집을 생성하게 된다[2, 9, 12]. Hanifi 등[15]은 *k*-means기법과 fuzzy C-means, mountain, subtractive 기법 등 3 가지 다른 군집기법을 사용하여 성능비교를 수행하여, fuzzy C-means 기법이 가장 성능이 좋은 것으로 발표했다. 분류 모형에 관한 초기의 연구에서는 decision tree, neural network, support vector machine, rough set theory 등 개별적인 기법들을 활용하였으나 최근에는 이들 기법과 사례기반 추론, 유전자 알고리즘, 또는 군집분석 등 다른 기법들과의 혼합 모형을 활용하고 있다[10, 11, 20]. Nan-Chen 등[16]은 *k*-means 군집기법과 Neural Network 기법을 함께 사용한 혼합기법으로 모형의 성능을 개선시키고자 하였다. Kim 등[18]은 *k*-means 군집기법과 유전자 알고리즘(GA, Genetic Algorithm)을 혼합한 모형을 개발하여 순수 *k*-means 군집기법, SOM 군집기법의 성능을 비교하였다. 그

결과 GA *k*-means 혼합기법이 가장 성능이 좋은 것으로 나타났다. 또한, Kuo 등[19]은 SOM 군집기법과 *k*-means 군집기법을 함께 사용한 혼합기법의 성능을 순수 SOM 군집기법, two-stage method 군집기법의 성능과 비교하여, 제안된 혼합기법이 가장 좋은 성능을 내는 것을 보여주었다.

3. 데이터 및 실험모형

3.1. 실험데이터

실험에 사용된 데이터는 2003년 1년간 군 훈련소에서 특정 특기에 대한 후반기 교육을 받는 훈련병을 대상으로 수집하였다. 4개 입영 기수에서 총 669건의 데이터를 수집하였다. 원천 데이터의 변수는 총 69개인데, 신체상태, 지적상태, 가정환경, 그리고 인적 성향을 나타내는 병사 기초속성과 관련된 변수들로 구성되어 있다.

3.2 실험 절차

본 연구는 <그림1>에서 보인 순서대로 진행하였다. STEP1, STEP2 그리고 STEP3은 3.2절에서 설명하고, STEP4는 4절에서 설명하였다.

3.2.1 입력데이터의 전 처리

원천 변수 중에서 지나치게 하나의 값으로 치중되어 있는 변수들은 실험에서 제외하였으며 세 개의 파생변수를 생성시켜 모형에 사용하였다. 첫 번째 파생변수는 실질적인 훈련성취도에 영향을 미칠 것으로 예상되는 '신장' 과 '체중' 변수를 통합시킨 '비만도(BMI)[1]' 변수이다. 두 번째와 세 번째 파생변수는 출신환경에 의한 영향요소를 알아보기 위해 '주소' 변수에서 추출한 '광역권 주소지[2]', 와 '도시 및 촌락[3]' 변수이다. 또한 해석의 편의를 위하여 수치형 변수인 지능지수와 나이 변수는 범주형 변수로 변환하였다. 이러한 사전처리 과정을 통해 총 15개의 후보 입력변수군이 만들어졌다.

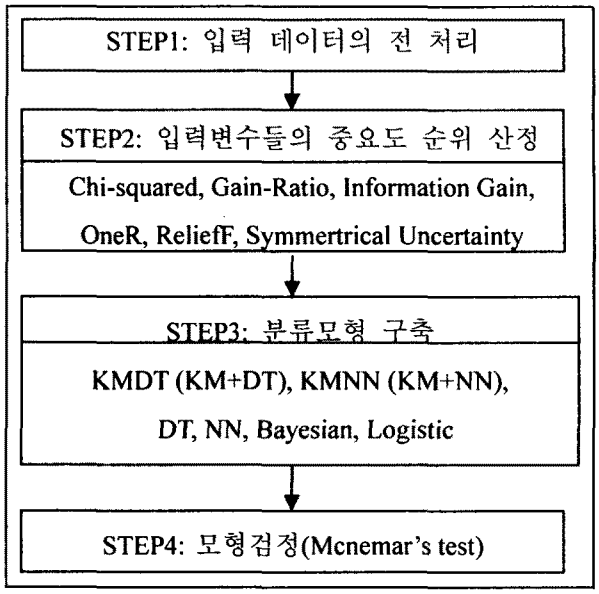

그림 1: 실험절차

목표변수의 값은 최초에는 성적을 나타내는 수치형 변수였으며, 4개 기수간의 상대적 차이를 고려하여 기수간 평균 점수를 기준으로 정규화 시킨 후 이진형 값으로 전환시켰다. 육군 군사교육평가 기준[4]에 따라 성적 값의 상위 80%를 '일반관리 인원'으로, 하위 20%를 '집중관리 인원'으로 정의하였다. 이렇게 목표변수를 이분한 이유는 신병교육 및 후반기 교육의 중점은 특기 직무기술의 습득에 있기 때문에 수준 저조자를 미리 예측하여 이들에 대하여 집중관리함으로써 일반 수준으로 향상시키는 교육이 매우 중요하기 때문이다.

총 669개의 인스턴스 중 20%에 해당하는 135건(집중관리인원)과 나머지 80%에서 무작위 추출한 135건(일반관리인원)을 합쳐 총 270건의 인스턴스를 최종 실험에 사용하였다.

3.2.2 입력변수들의 중요도 순위 산정

[1] BodyMassIndex: 체중/(신장)^2*100,비만도를 측정하는 지수
[2] 현재 주민등록상의 주소지에서 광역권 단위로 추출
[3] 도시(행정구역: 구, 동) / 촌락(행정구역: 읍, 면, 리)
[4] 상(서열 40%이내), 중상(41~80%), 중 (80% 미만)

모형의 성능을 높이면서 좀 더 경제적인 모형을 만들기 위하여 변수선택 단계는 매우 중요하다. 왜냐하면 목표변수와 관련이 적은 입력 변수는 오히려 모형의 성능을 떨어뜨릴 수 있기 때문이다[12]. 따라서 목표변수와 관련된 최적의 입력변수를 적절한 방법으로 선택해 주어야 한다.

표 1: 중요도 순위 목록

순위	변수 명	변수 값
1	생활정도	상, 중, 하
2	시각결함	D1(안경착용), D2(시각장애(색맹, 색약)), D1D2(안경착용 및 시각장애), N(안경 미착용 및 시각장애가 아닌 자)
3	광역권주소지	경기, 강원, 충청, 전라, 경상, 제주
4	신체등급	1등급, 2등급, 3등급
5	입영형태	징집, 모병
6	학력	4년제, 2년제, 고졸 이하
7	신체결함	Y(신체결함이 있음),N(신체결함이 없음)
8	부모	Y(양친), N(편모,편부,고아,계부,계모)
9	혈액형	A형, B형, O형, AB형
10	비만도	저체중, 정상, 과체중, 비만
11	자격면허	Q1, Q2, Q3, Q4, Q5 (자격증 등급에 따른 분류)
12	나이	1(23세 이상), 2(21~22세), 3(20세 이하)
13	인성검사	적격, 부적격
14	지능지수	1(125 초과), 2(115~125), 3(115 미만)
15	도시 및 촌락	도시(구, 동), 촌락(읍, 면, 리)

본 연구에서는 입력변수들의 목표변수와의 관련도를 이용하여 순위를 산정한 후, 순위가 높은 입력변수부터 차례로 하나씩 누적 입력시키며 모형의 성능 변화를 관찰하였다. 먼저, 각 입력변수들의 목표변수와의 관련도 순위를 산정하기 위하여 Chi-squared, Gain Ratio, Information Gain, OneR, ReliefF, 그리고 Symmetrical Uncertainty[5] 등 6가지 변수 선정 기법을 이용하였다. 각 기법들이 산정한 순위의 평균을 최종 순위로 산정하였다 (<표1> 참조).

<표1>을 살펴 보면, 파생 변수 중 하나인 '광역권주소지' 는 목표변수에 대한 중요도가 상대적으로 높은 것으로 볼 수 있다. 또한, 개인의 지적 상태(학력, 지능지수, 자격면허) 보다는 신체상태(시각결함, 신체등급, 신체결함)나 가정환경(생활정도, 광역권주소지, 부모) 요소가 목표변수에 영향을 보다 많이 미치는 것으로 나타났다. 이것은 일반기업에서의 직무교육과는 달리 군(軍)이라는 특수한 상황에서의 교육은 학력, 지능지수 등 개인의 지적 상태 보다는 신체상태나 가정환경에 따른 개인차가 더욱 중요한 요인이 될 수 있음을 보여준다고 할 수 있다.

3.2.3 분류 모형 구축

분류 모형으로 2 가지의 혼합 모형과 4 가지의 순수 모형 등 총 6 가지 모형을 구축하였고, 이들의 성능을 비교하였다.

2 가지의 혼합 모형인 KMDT와 KMNN은 다음과 같이 구축하였다. 입력 데이터를 *k*-means algorithm을 이용하여 3개[6] 군집으로 분류한 후, 각 인스턴스가 속하게 될 군집을 예측하는 Decision Tree 모형과 Neural Network 모형을 생성하였다. 각각의 정확도는 97.03%, 98.52%로 비교적 높게 나왔다. 그리고, 다시 각 군집별로 Decision Tree와 Neural Network 기법을 이용하여 직무교육 성취도 예측 모형을 각각 생성하였다. 생성된 모형의 정확도는 3개 군집 각각의 인스턴스 수[7]와 군집예측모형의 정확도[8]를 고려하여 최종 정확도를 산출하였다.

이들 혼합모형의 경우, 새로운 신병에 대하여 직무교육 성취도를 예측하기 위하여는 먼저 신병이 어떤 군집에 속하는지를 예측하고 이어서 해당 군집에 적용되는 성취도 예측모형으로 이 신병이 일반관리 인원인지 집중관리 인원인지를 예측하게 된다.

4 가지의 순수 모형으로는 Decision Tree 모형, Neural

[5] Symmetrical Uncertainty (Class, Attribute) = 2 * (H(Class) - H(Class | Attribute)) / H(Class) + H(Attribute).

[6] Clementine의 two-step model을 이용한 결과 3개의 군집으로 나눌 것을 결정하였음

[7] Cluster(1): 94개, Cluster(2): 130개, Cluster(3): 46개

[8] DT의 정확도: 97.03%, NN의 정확도: 98.52%

Network 모형, Bayesian 모형, 그리고 Logistic 모형을 구축하였다.

4. 실험결과 및 해석

모형의 개발에는 'Weka V3.4.10' 와 'Clementine V10.1' 을 사용하였다. 전술 한 바대로, 실험에 사용된 인스턴스의 수는 270 건, 변수의 수는 15 개이다. 검증방법으로 10-fold cross validation을 실시하였다[21]. 입력 속성의 중요도 산출과정에서 매겨진 변수를 우선순위가 높은 순서대로 입력변수의 수를 증가시켜 나가면서 예측모형을 만들었다[9]. 실험 결과는 <표2>와 같다. 모형의 성능을 나타내는 정확도는 백분율로 표시하였다.

표2: 분류 예측 정확도 (%)

속성수	KMDT	KMNN	DT	NN	Bayesian	Logistic
2	62.53	66.78	70.37	***70.37***	70.37	70.37
3	70.43	***68.96***	69.63	69.63	69.26	69.26
4	70.80	67.14	69.26	70.37	***71.11***	***71.48***
5	70.80	68.96	70.74	65.93	69.26	68.89
6	70.08	65.32	68.89	64.44	67.41	67.78
7	70.44	66.04	***71.48***	65.56	67.04	68.15
.	...	...	...	...	...	...
15	***72.95***	62.40	69.63	62.59	68.89	66.3

<표2>에서 보는 바와 같이, KMDT 모형이 입력변수의 수가 15개 일 때 모든 모형에서 가장 좋은 72.95%의 정확도를 보였다. 각 모형의 예측 정확도에 있어서도 차이가 있지만, 가장 좋은 결과를 나타낼 때 사용된 입력변수의 수도 각각 다름을 알 수 있다. 또한 인공지능기법이 사용된 KMNN 이나 NN 기법은 의사결정나무기법이 사용된 KMDT 와 DT 기법에 비해 상대적으로 적은 속성만을 가지고도 높은 정확도를 보이고 있다.

[9] 변수선정기법을 filter방식으로 사용하여 입력변수들의 중요도 순위를 산정한 후 wrapper방식으로 변수를 선정한 것임

마지막으로, 6 가지 모형결과 간에 유의한 차이가 있는지를 검정하기 위해 맥니마 검정(McNemar's test)을 실시 하였다[8]. <표 3>은 맥니마 검정의 결과를 보여준다.

표3: 맥니마 검정 결과

	KMDT	KMNN	DT	NN	Bayesian
Logistic	22.23***	21.49***	4.55**	3.27*	0.33
KMDT		0.03	10.67***	29.49***	22.22***
KMMN			10.25***	27.56***	19.93***
DT				15.21***	3.86**
NN					4.57**

*유의수준 10%, **유의수준 5%, *** 유의수준 1%

<표3>에서 보는 바와 같이, 본 연구에서 제안한 KMDT 모형의 경우, 같은 혼합모형인 KMNN 모형을 제외하고 다른 모형들과 유의수준 1% 내에서 그 차이가 유의함을 나타내고 있다. 다른 대부분의 모형간의 비교에서도 유의수준 10%에서 1% 내에서 유의함을 나타내고 있지만 통계적 기법인 Bayesian 모형과 Logistic모형간에는 유의한 차이가 없는 것으로 나타났다.

5. 결론

본 연구에서는 입대 초기에 얻을 수 있는 신병들의 제한된 기초속성 데이터를 이용하여 예상 교육성취도를 예측하는 모형을 제시하였다. 실험 결과 제시한 KMDT 모형이 KMNN 모형이나 기존에 분류 모형에 자주 사용되었던 순수한 DT, NN, Baysian, Logistic 기법보다 성능이 좋았다. 또한 본 연구는 신병들의 극히 제한된 기초속성만을 가지고 70% 이상의 예측률을 얻을 수 있었다.

본 연구는 군 조직을 비롯한 다양한 조직체 내에서 직무교육에 대한 성과를 높일 수 있도록 미리 조직원의 성취도 예측을 시도한 것으로, 이러한 예측은 조직원의 맞춤식교육을 확립하는데 있어 중요한 참고자료로 활용될 수 있을 것으로 기대된다.

본 연구에서는 한 개의 특기 신병만을 대상으로 실험에 사용하였기 때문에 좀 더 다양한 특기의 데이터를 모형에 적용시킬 필요가 있으며, 데이터의 양에 있어서도 보다 많은 데이터를 수정하여 모형의 신뢰도를 향상 시킬 필요가 있다. 본 연구에서는 간부를 대상으로 하는 군사교육평가 기준을 그대로 사용하였으나, 병 교육훈련 실정에 맞게 좀 더 구체적이며 합리적으로 기준을 정할 필요가 있다.

참고문헌

[1] 국방부 (2007), *국방개혁에 관한 법률 시행령*.

[2] 김혜숙, 문양세, 김진호, 노웅기 (2007). "데이터 마이닝을 사용한 방학 중 학습방법과 학업성취도의 관계 분석," *정보과학회논문지: 소프트웨어 및 응용*, 제 34권, pp 40-51.

[3] 배재호 (2001). "데이터 마이닝을 이용한 학업성취도 분석,"경희대학교 교육대학원 석사학위논문.

[4] 오성삼, 구병두 (1999). "메타분석을 통한 한국형 학업성취 관련변인의 탐색," *교육학연구*, 제 39권, pp. 99-122.

[5] 육군본부 (2007), 육군규정.

[6] 차지혜 (2001). "영어과 학업성취도에 영향을 미치는 배경변수에 대한 다차원적 분석," 이화여자대학교 대학원 석사학위논문.

[7] 하영자. (2005). "공무원의 온라인 직무교육에서 자기효능감과 자기조절학습 수행력이 만족도와 성취도에 미치는 영향" , *한국사이버교육학회, e-learning 학술연구*, 제4권 제1호, pp. 31~63.

[8] 허명회 (2005). *비교연구를 위한 통계적 방법론*. 경기, 자유아카데미.

[9] Anderberg, R. (1973), *Cluster analysis for applications, New York*, MA: Academic Press.

[10] Carvalho, D.R., and Freitas, A.A. (2004). "A hybrid decision tree/genetic algorithm method for data mining," *Information Sciences*, Vol. 163, pp 13-35.

[11] Chang, P.C., Lai, C.-Y., and Lai, K.R. (2006). "A hybrid system by evolving case-based reasoning with genetic algorithm in wholesaler ' s returning book forecasting," *Decision Support Systems*, Vol. 42, pp. 1715-1729.

[12] Dash, M., and Liu, H. (1997). "Feature Selection for Classification," *Intelligent Data Analysis*, Vol. 1, pp. 131-156.

[13] Delen, D., Glenn W., and Amit K. (2005). "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, Vol. 34, pp. 113-127.

[14] Fuller, B. (1987). "What school factors raise achievement in the third word," *Review of Educational Research*, Vol. 57, pp. 255-273.

[15] Guldemir, H. and Abdulkadir S. (2006). "Comparison of clustering algorithms for analog modulation classification," *Expert Systems with Applications*, Vol. 30, pp. 642-649.

[16] Hsieh, N.C. (2005). "Hybrid mining approach in the design of credit scoring models," *Expert Systems with Applications*, Vol. 28, pp. 655-665.

[17] Hung, S.Y., David, C.Y., and Wang, H.-Y. (2006). "Applying data mining to telecom churn management," *Expert Systems with Application*, Vol.31, pp. 515-524.

[18] Kim, K.-J., and Ahn, H. (2007). "A recommender system using GA K-means clustering in an online shopping market," Forthcoming.

[19] Kuo, R.J., Ho, L.M., and Hu, C.M. (2002). "Integration of self-organizing feature map and K-means algorithm for market segmentation," *Computers & Operation Research*, Vol. 29, pp. 1475-1493.

[20] Min, S.H., Lee, J., and Han, I. (2006). "Hybrid genetic algorithms and support vector machines for bankruptcy prediction," *Expert Systems with Applications*, Vol. 31, pp 652-660.

[21] Ryu, Y.U., Chandrasekaran, R., and Jacob, V.S. (2007). " Breast cancer prediction using the isotonic separation technique," *European Journal of Operation Research*, Vol. 181, pp. 842-854.

[22] Witten, I.H., and Frank, E. (2005). *DATA MINING: Practical Machine Learning Tools and Techniques*. San Francisco, MA: Morgan Kaufmann.