데이터마이닝 기법 비교 연구: 단일 및 복수 의사결정나무

신은주*, 장남식**

Comparisons of Tree-based Data Mining Techniques: Single vs. Multi-Decision Trees

Shin, Eun-Joo, Chang, Namsik

One widely used knowledge discovery technique is a decision tree inducer that generates classifiers in the form of a single decision tree. As the number of pre-specified decision outcome classes increases, however, the trees generated by this approach usually become more complex than necessary with regard to the number of leaves and nodes, and the predicting accuracy consequently drops. This paper suggests an approach based on multi-decision-tree induction and compares it with a traditional single-decision-tree induction approach. Seven empirical experiments have shown that the multi-decision-tree approach outperformed the traditional approach in terms of prediction accuracy and rule conciseness.

* 서울시립대학교 대학원 경영학과 (m9821003@sidae.uos.ac.kr)
** 서울시립대학교 경상대학 경영학부 교수 (nchang@uoscc.uos.ac.kr)

I. 서 론

정보기술의 빠른 발전은 기업환경의 다변화와 정보시스템을 통한 업무의 자동화를 촉진시켜 엄청난 양의 자료의 수집이나 저장은 물론이고 자료의 분석을 통한 정보의 획득을 가능하게 하였다. 그러나 실제로는 자료(레코드)의 양이 증가하고 차원(속성)의 수가 증가할수록 분석의 효율성은 떨어지고 분석을 통해 얻을 수 있는 정보의 품질도 기대하기 어렵게 된다. 따라서 효과적으로 자료를 여과하고 분석하여 숨겨진 정보를 포함한 다양한 정보를 제공할 수 있는 데이터마이닝(data mining)에 대한 다양한 연구가 진행되고 있다. 특히 귀납적 학습(inductive learning)에 근거한 인공신경망(neural network)이나 의사결정나무(decision tree)에 대한 관련 기법이나 툴을 개발하거나 이들을 현업에 적용하는 프로젝트가 핵심 연구 분야 중의 하나이다. 인공신경망은 일반적으로 가장 널리 사용되는 데이터마이닝 기법으로 다른 기법들에 비해 정확한 분류나 예측력을 제공하는 것으로 알려져 있다[M. J. A. Berry, 1997; V. Ciesielski, 1996; K. Y. Tam, 1992]. 그럼에도 불구하고 의사결정나무 기법에 관한 연구는 여전히 높은 관심의 대상이 되고 있는데 가장 큰 이유는 제공하는 정보(규칙이나 나무)를 우리가 쉽게 이해할 수 있기 때문이다. 이렇듯 의사결정의 이유를 설명할 수 있는 능력은 신용평가, 보험요율 설정, 데이터베이스 마케팅 등의 데이터마이닝 응용분야에서 중요한 역할을 하는 경우가 많다[W. J. Frawley, 1991].

지금까지 소개된 대부분의 의사결정나무 기법들은 데이터에 포함된 부류(클래스 및 목표변수) 값의 수에 관계없이 하나의 모형(나무), 즉 단일 의사결정나무를 만든다[B. Cestnik, 1987; J. Cheng, 1988; K. J. Cherkauer, 1996; J. R. Quinlan, 1986; J. R. Quinlan, 1993]. 부류 값의 수가 증가하게 되면 최종적으로 나무의 너비와 깊이, 즉 모형을 구성하는 내부 마디와 잎의 수는 불필요하게 많아지고 복잡하게 되기 때문에 모형의 예측력과 설명력도 낮아지게 된다[N. Chang, 1995; N. Chang and O. R. L. Sheng, 1995].

본 연구의 목적은 3 개 이상의 부류 값으로 구성된 데이터에 의사결정나무 기법을 적용하여 부류 값별로 복수의 나무를 만들어 모형을 구축하는 것이 단일 의사결정나무 모형을 구축하는 일반적인 방법에 비해 정보의 품질(예측력 및 설명력)면에서 우수하다는 것을 이론적인 측면에서 분석하고 실제 데이터를 통해 검증하는데 있다. 본 논문의 구성은 다음과 같다. 2 장에서는 단일 의사결정나무 기법과 복수 의사결정나무 기법의 정의 및 개요, 속성 선택 과정, 가지치기 과정을 비교 설명한다. 3 장에서는 캘리포니아대학 어바인분교의 인공지능 연구소에서 수집한 7 개의 데이터 세트를 이용하여 위의 기법들을 비교한다. 마지막으로 4 장에서는 결과를 요약하고 향후 연구 방향을 논한다.

II. 단일 vs. 복수 의사결정나무

2.1 정의 및 개요

의사결정나무 기법은 개체의 속성(입력변수)과 부류로 구성된 데이터로부터 순환적 분할(recursive partitioning)방식을 이용하여 나무를 구축하는 기법으로, 구축 되어진 나무는 나무의 가장 상단에 위치하는 뿌리마디(root node), 속성의 분리기준을 포함하는 내부마디(internal nodes), 마디와 마디를 이어주는 가지(link), 그리고 최종 분류를 의미하는 잎(leaves)으로 구성되며, 분류나 예측에 주로 사용된다.

<그림 1>에서 보듯이 지금까지 소개된 대부분의 의사결정나무 기법들은 데이터에 포함된 부류 값의 수에 관계없이 하나의 모형(나무), 즉 단일 의사결정나무를 만든다. 따라서 부류 값의 수가 증가하게 되면 최종적으로 나무의 너비와 깊이, 즉 모형을 구성하는 내부마디와 잎의 수는 불필요하게 많아지고 복잡하게 된다.

```
If 모든 사례가 단일 부류에 속함
then 잎에 해당 부류값을 지정하고 종료
else
   각각의 속성별로 IGR(Information Gain
     Ratio)을 계산하고 그중 가장 큰
     IGR 값을 갖는 속성을 선택하여 나
     무를 확장한다.
   선택된 속성의 값에 따라 사례를 분
     할한다.
   for 각 속성값별로 do
     하위 나무모형을 구축한다.
```

<그림 1> 단일 의사결정나무 구축 절차

이에 반해 복수 의사결정나무 기법은 데이터에 포함된 부류 값별로 의사결정나무를 구축하고 이들로부터 만들어지는 규칙들을 통합한다는 개념이다. <그림 2>는 이러한 개념을 나타내고 있다.

```
For 각 부류값(Ci)별로(1≤i≤c)
  For 각 사례별로
    if 사례들의 부류값이 Ci
    then Pi 로 대체
    else Ni 로 대체
  부류값 Ci 에 대한 나무 구축
규칙 병합
```

<그림 2> 복수 의사결정나무 구축 절차

2.2 속성 선택 과정

<표 1>은 3 개의 부류 값으로 구성된 12 개 사례의 속성 및 해당 부류 값, 그리고 복수 의사결정나무의 구축을 위해 수정된 부류 값들을 보여주고 있다.

<표 1> 3개의 부류 값으로 구성된 표본 데이터

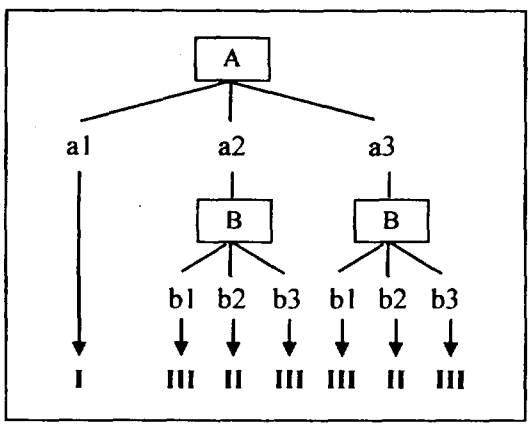
사례	속성			해 당 부류값	수정된 부류값		
	A	B	C		부류 I	부류 II	부류 III
1	a1	b1	c1	I	Y	N	N
2	a1	b3	c2	I	Y	N	N
3	a1	b1	c2	I	Y	N	N
4	a1	b3	c1	I	Y	N	N
5	a2	b2	c1	II	N	Y	N
6	a2	b2	c2	II	N	Y	N
7	a3	b2	c2	II	N	Y	N
8	a3	b2	c1	II	N	Y	N
9	a2	b1	c3	III	N	N	Y
10	a2	b3	c3	III	N	N	Y
11	a3	b1	c3	III	N	N	Y
12	a3	b3	c3	III	N	N	Y

<그림 3>은 <표 1>의 데이터를 이용하여 단일 의사결정나무를 구축한 예이다. 이 그림에서 보듯이 속성 A 가 나무의 뿌리마디(root node)에 위치하며 속성 값 a1, a2, a3 에 따라 사례들이 분할되었다.

그러나 속성 A의 값 a2, a3은 부류 II나 부류 III에 속하는 사례들을 분류함에 있어 주요 역할을 하지 못함에도 불구하고, a1이 부류 I에 속하는 사례들을 분류하는데 결정적인 역할을 하기 때문에 부수적으로 나무의 상단에 위치하게 된다. 이로인해 사례들을 부류 III으로 분류한 4가지 규칙

규칙 1 : If A=a2 and B=b1 then class=III
규칙 2 : If A=a2 and B=b3 then class=III
규칙 3 : If A=a3 and B=b1 then class=III
규칙 4 : If A=a3 and B=b3 then class=III

들은 사례들을 부류 III으로 분류하는 결정적인 규칙, 즉 C=c3라는 규칙을 찾지 못하고 있으며, 이로 인해 정보의 예측력과 설명력을 떨어뜨리는 결과를 초래할 수 있다. <그림 4>는 <표 1>의 데이터를 이용하여 복수 의사결정나무를 구축한 경우로 각 부류 값에 해당하는 나무에 가장 분별력이 있는 속성과 해당 값, 즉 부류 I의 경우 A=a1, 부류 II의 경우 B=b2, 부류 III의 경우 C=c3 만이 포함되어 있음을 알 수 있다.



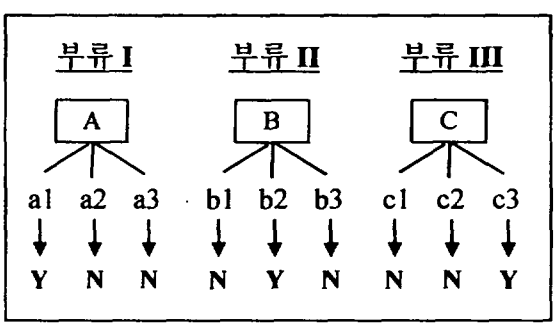
<그림 3> 단일 의사결정나무를 적용한 경우

위의 사례는 아주 극단적인 예로써 실제 데이터의 초기모습에서 나타나기 어려운 경우가 대부분이나 의사결정나무를 구축하는 과정에서 일부 가지에 해당되는 데이터에서 자주 찾아볼 수 있다.

2.3 가지치기(Pruning) 과정

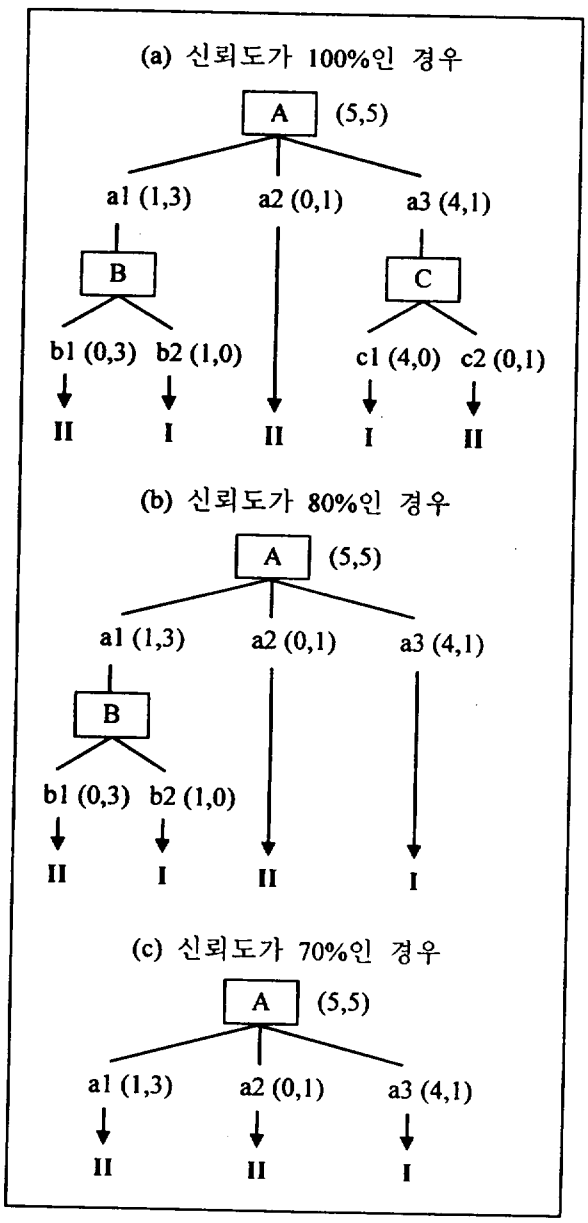
나무의 가지치기는 데이터의 노이즈와 모형 추정용 데이터의 과대맞춤(over-fitting) 현상을 줄이기 위해서 사용되는 대표적인 방법이다. 의사결정나무는 나무마디에 속한 사례들의 부류 값이 모두 동일하거나 또는 마디에 오직 1개의 사례가 남을 때까지 확장된다. 노이즈가 포함된 데이터를 이용하여 구축된 의사결정나무 모형은 모형 추정용 데이터를 분류함에 있어서는 정확한 예측력을 제공하지만 과대맞춤으로 인해 모형 시험용 데이터에 대한 예측력을 상당히 떨어뜨리는 결과를 초래하곤 한다. 따라서 특정한 수준에서 나무의 확장을 중단할 필요가 있다.

일반적으로 가장 널리 사용되는 가지치기 기법이 신뢰도에 근거한 가지치기



<그림 4> 부류 수만큼 의사결정나무를 만들어 적용한 경우

이며, 신뢰도의 값은 모형 구축 전에 미리 정하며 백분율로 표시된다[J. Mingers, 1989]. 이 기법은 마디를 구성하는 사례들에서 한 종류의 부류 값에 속한 사례의 비율이 신뢰도 보다 커질 때까지 계속 나무를 확장한다. 따라서 <그림 5>에서 보듯이 신뢰도의 값을 낮출수록 나무의 구조는 더욱 단순화된다.



<그림 5> 서로 다른 신뢰수준의 효과

단일 의사결정나무 기법을 사용하여 모형을 구축하면 부류 값별 주요 속성들은 <그림 3>의 부류 II 에 대한 주요 속성 B 나 <그림 5(a)>의 부류 I 에 대한 주요 속성 C 와 같이 나무의 하단에 위치할 수 있다. 이러한 속성들은 가지치기 방법의 상향식 성질에 의해서 가지치기 과정에서 제거될 수도 있다. 그러나 복수 의사결정나무 기법을 사용하면 사례를 분류하는 데 있어서 가장 중요한 속성을 우선적으로 고려하기 때문에 이들은 의사결정나무의 윗부분에 위치하게 되고 중요도가 떨어지는 속성들이 우선적으로 제거된다. 따라서 복수 의사결정나무 기법을 이용하여 모형을 구축하게 되면 가지치기로 인한 주요 속성의 잘못된 제거를 줄일 수 있다.

결론적으로 데이터에 각 부류 값별로 서로 다른 중요한 분류 속성이 존재할 경우나 가지치기를 적용할 경우 복수 의사결정나무 기법을 통한 모형은 단일 의사결정나무 기법의 모형보다 규칙의 신뢰도나 설명력 그리고 노이즈 처리에 있어서 우수하다고 할 수 있다. 이와 같은 결과는 다음 장에서 실증적으로 비교해 보겠다.

III. 연구 방법 및 결과

3.1 자료의 수집 및 성격

University of California at Irvine 의 Machine Learning Repository 에 관리되고 있는 데이터 중에서 3 개 이상의 부류 값을 포함하는 7 개의 데이터 세트를 수

집하였다. 이들 데이터의 특성은 <표 2>와 같다.

- Derm : 피부의학에서 병의 유형을 결정하기 위한 데이터이다. 6 개의 부류 값으로 구성되고, 34 개 속성 중에서 33 개는 이산형(discrete)이고 1 개는 연속형(coutinuous)이다.
- Ecoli : 단백질의 세포 분류를 예측하는 데이터이다. 5 개의 부류 값으로 구성되고, 7 개의 속성은 모두 연속형이다.
- Glass : 교통 사고 조사 시, 유리 파편의 잔해를 정확히 조사하면 사건의 증거로 사용되어질 수 있다. 따라서 이 데이터에서는 유리 파편의 잔해를 6 개의 부류 값으로 나누고, 9 개의 연속형 속성을 지니고 있다.
- LED : 숫자인식문제(digit recognition problem)에 관한 클래식 제조 시험 데이터이다. 10 개의 부류 값은 각각 0~9 의 값을 갖는다.
- Lymph : 종양학에서 병의 증상을 예측하는 자료이다. 부류 값은 normal, metastases, malignant, fibrosis 의 4 개이며, 18 개의 속성은 림프관과 림프노드에 대한 세부항목을 제공한다.
- Post-opr : 환자를 수술 후의 회복 정도에 따라 응급실, 집, 일반병실 중에서 어디로 보낼지를 판정하는 작업에 관한 데이터이다. 3 개의 부류 값을 갖고 신체의 체온과 관련된 8 개의 속성으로 구성된다.
- Zoo : 불린(boolean)값을 갖는 16 개의 속성의 조합에 따라 동물의 부류를 판정하는 작업에 관한 데이터로 7 개의 부류 값으로 나누어진다 (예: 포유류, 해조류, 갑각류, 등).

<표 2>에서 'Lymph' 데이터는 4 개의 부류 값으로 분류되지만 148 개의 사례들 중에서 부류 I 과 부류 IV 에 속하는 사례는 6 개이며 나머지는 부류 II 와 부류 III 으로 분류된다. 따라서 2 개의 부류 값만 있는 경우와 차이가 거의 없다. 마찬가지로 'Post-opr' 데이터도 3 개의 부류 값으로 구성되나 90 개의 사례 중에 오직 2 개만이 부류 I 에 속하기 때문에 2 개의 부류 값만으로 이루어진 데이터와 차이가 거의 없다고 판단된다. 이에반해 나머지 데이터 세트의 사례들은 부류 값에 따라 비교적 고르게 분포하고 있으며, 특히 'LED' 데이터의 경우 사례들이 균등하게 분포되었다.

<표 2> 데이터의 특성

데이터	개 수				부류 분포
	모형추정용 사례	모형시험용 사례	속성	부류	
Derm	200	100	34	6	{91, 50, 61, 39, 39, 20}
Ecoli	200	100	7	5	{127, 71, 50, 33, 19}
Glass	144	70	9	6	{70, 75, 18, 13, 9, 29}
LED	200	100	7	10	균등 분포
Lymph	103	45	18	4	{2, 81, 61, 4}
Post-opr	60	30	8	3	{2, 24, 64}
Zoo	71	30	16	7	{41, 20, 5, 13, 4, 8, 10}

<표 3> 평균 예측 오류율

데이터	Blind-guess**	단일	복수	차이*
		평균 오류율(표준편차)		
Derm	69.6	7.3 (1.9)	6.7 (1.8)	단일 = 복수
Ecoli	57.7	18.6 (3.7)	18.7 (2.9)	단일 = 복수
Glass	65.0	34.4 (7.6)	31.4 (3.9)	단일 > 복수
LED	90.0	27.1 (3.5)	24.9 (3.6)	단일 > 복수
Lymph	45.2	22.7 (4.8)	22.4 (4.2)	단일 = 복수
Post-opr	28.8	30.0 (3.5)	29.7 (3.3)	단일 = 복수
Zoo	59.4	6.7 (5.5)	6.3 (3.3)	단일 = 복수

* 쌍대 t-검정 결과 ($\alpha = 0.05$)

** 데이터의 부류 중 최빈 수의 값에 따른 분류

<표 4> 평균 규칙의 간결성

데이터	단일	복수	차이*
	평균 간결성(표준편차)		
Derm	4.3 (0.3)	2.7 (1.0)	단일 > 복수
Ecoli	3.2 (0.3)	2.9 (0.2)	단일 > 복수
Glass	4.4 (0.6)	3.8 (0.5)	단일 > 복수
LED	4.6 (0.4)	4.1 (0.3)	단일 > 복수
Lymph	3.6 (0.4)	3.4 (0.5)	단일 = 복수
Post-opr	0.0 (0.0)**	1.1 (0.2)	단일 = 복수
Zoo	4.0 (0.7)	2.4 (0.5)	단일 > 복수

* 쌍대 t-검정 결과 ($\alpha = 0.05$)

** 조건이 없는 하나의 룰만 발견

3.2 평가 기준

위에서 수집된 데이터 세트에 단일 의사결정나무 기법과 복수 의사결정나무 기법을 적용하여 아래의 기준에 따라 비교하였다.

- 예측 오류율 = (오분류된 사례의 수/시험용 사례의 수)×100
- 정보의 간결성 $= \frac{1}{k} \sum_{i=1}^{k} DS_i$,

  k: 의사결정나무를 통해 만들어진 규칙의 수

  DS_i : i번째 규칙에 포함된 조건의 수

3.3 결과 및 해석

전체자료의 2/3 정도는 모형을 추정하는데, 그리고 나머지는 모형을 시험하는데 사용하였다. 각 데이터 세트 당 모형 추정용과 모형 시험용 사례를 10번씩 무작위로 추출하여 모형을 구축한 후 결과의 평균을 비교하였다. <표 2>는 각 데이터 세트에 대한 사례의 수를 보여주고 있다.

<표 3>과 <표 4>는 각 데이터 세트에

단일 의사결정나무 기법과 복수 의사결정나무 기법을 적용하여 측정된 예측 오류율과 규칙의 간결성을 평균과 표준편차를 이용하여 나온 통계 결과이다. 예를들어 <표 4>의 '4.3(0.3)'은 'Derm' 데이터에서 단일 의사결정나무 기법을 적용하여 구축된 10 개 모형이 제공하는 규칙의 간결성의 평균은 4.3 이고 표준편차는 0.3 임을 의미한다. 표의 마지막 열에 나타난 '차이'는 가설을 시험하기 위하여 모집단의 평균을 쌍대 t-검정하여 나온 결과이다.

<표 3>과 <표 4>에서 보듯이 규칙의 정확도와 간결성에 있어서 복수 의사결정나무 기법은 단일 의사결정나무 기법에 비해 같거나 좋은 결과를 제공했으며, 이것은 쌍대 t-검정에 의해 검증되었다.

'Lymph' 데이터의 경우 4 개의 부류 값이 존재하지만 부류 I 과 부류 IV 에 속하는 사례가 각각 2, 4 개로서 이는 전체 사례의 4%이고, 'Post-opr' 데이터의 경우도 3 개의 부류 값이 존재하지만 부류 I 에 속하는 사례는 2 개로서 전체 사례의 2%에 불과했다. 따라서 오직 2 개의 부류 값만을 갖는 데이터와 거의 차이가 없다고 볼 수 있다. 따라서 예측 오류율이나 규칙의 간결성 측면에서 복수 의사결정나무 기법을 사용하여 얻을 수 있는 기대효과를 제공하지 못했다.

반면 'Glass' 데이터의 경우 복수 의사결정나무 기법을 적용한 결과 각 부류 값별로 주요 속성 및 속성값들이 발견되었으며 이는 앞에서 설명했듯이 데이터에 각 부류 값별로 서로 다른 주요 분류 속성이 존재할 경우 복수 의사결정나무 방식은 단일 의사결정나무 방식보다 모든 기준에 있어서 우수하다는 것과 일치하고 있다.

'Post-opr' 데이터의 경우 데이터 자체에 분류의 기준이 될만한 속성이 존재하지 않았으며 결과적으로 blind-guess 보다도 못한 결과를 초래하였다. 'Derm' 과 'Zoo' 데이터의 경우에는 예측 오류율에 차이가 없었으며 이는 상대적으로 노이즈가 없었기 때문인 것으로 판단된다.

IV. 결론 및 향후 연구 방향

이 논문은 데이터에 포함된 부류 값의 종류가 3 가지 이상일 경우 복수 의사결정나무 기법에 의한 모형을 단일 의사결정나무 기법에 의한 모형보다 규칙의 신뢰도와 설명력 측면에서 우수하다는 것을 이론적으로 설명하였고 실제 데이터를 이용해 검정하였다. 이러한 결과는 향후 복수 의사결정나무 기법이 다른 데이터마이닝의 분류/예측 기법과의 비교의 대상이나 대안으로서 사용할 수 있다는 것을 의미한다.

<참 고 문 헌>

[1] M. J. A. Berry and G. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, 1997.

[2] B. Cestnik, I. Kononenko, and I. Bratko, ASSISTANT 86: A knowledge-elicitation tool for sophisticated users, *Proceedings of the Second European Working Session on Learning*, 1987, pp.31-45.

[3] N. Chang, *Knowledge Discovery in Databases with Joint Decision Outcomes: A Decision-tree Induction Approach*, Doctoral Dissertation, The University of Arizona, 1995.

[4] N. Chang and O. R. L. Sheng, Automated Decision Rule Discovery from Domains with Joint Decision Outcomes: A Decision Tree Induction Approach, *Proceedings of the Third International Conference of the ISDSS*, Vol. 2, 1995, pp.259-267.

[5] J. Cheng, U. Fayyad, K. Irani, and Z. Qian, Improved decision trees: A generalized version of ID3, *Proceedings of the Fifth International Conference on Machine Learning*, 1998, pp.100-106.

[6] K. J. Cherkauer and J. W. Shavlik, Growing Simpler Decision Trees to Facilitate Knowledge Discovery, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp.315-318.

[7] V. Ciesielski and G. Palstra, Using a Hybrid Neural/Expert System for Data Base Mining in Marketing Survey Data, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp.38-43.

[8] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, Knowledge Discovery in Databases: An Overview, in W. J. Frawley, G. P. Shapiro, & C. J. Matheus (eds.), *Knowledge Discovery in Databases*, 1991, pp.1-27.

[9] J. Mingers, An Empirical Comparison of Selection Measures for Decision-Tree Induction, *Machine Learning*, 1989, pp.319-342.

[10] J. R. Quinlan, Induction of Decision Trees, *Machine Learning*, 1986, pp.81-106.

[11] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA., 1993.

[12] K. Y. Tam and M. Y. Kiang, Managerial Applications of Neural Networks: The Case of Bank Failure Predictions, *Management Science*, Vol. 38, 1992, pp.926-947.