귀납적 학습방법들의 분류성능 비교

Classification Performance Comparison of Inductive Learning Methods

이 상 호, 지 원 철

홍익대학교 산업공학과
서울 마포구 상수동 72-1
(T) 320-1684 (F) 336-1130 Email: jhee@wow.hongik.ac.kr

Abstract[1)]

In this paper, the classification performances of inductive learning methods are investigated using the credit rating data. The adopted classifiers are Multiple Discriminant Analysis (MDA), C4.5 of Quilan, Multi-Layer Perceptron (MLP) and Cascade Correlation Network (CCN). The data used in this analysis is obtained using the publicly announced rating reports from the three korean rating agencies. The performances of 4 classifiers are analyzed in term of prediction accuracy. The results show that no classifier is dominated by the other classfiers.

1. 서 론

귀납적 추론은 인류가 오랜 세월에 거쳐 자연스럽게 사용 발전시켜 온 추론 방법의 하나이다. 따라서, 전문가시스템 분야에서 지식획득의 어려움을 해결하기 위해 자동학습방법의 개발에 많은 노력을 기울여 온 사실을 감안할 때, 귀납적 추론을 사용하는 학습방법들의 개발 및 사용이 보편화된 것은 당연하다 하겠다.

귀납적 학습방법에 대해 공식적 정의를 한다면 다음과 같다. 우선 주어진 한 쌍의 입출력 $(x, f(x))$을 예제(Example)라고 하면 $f(x)$는 입력값 x가 주어졌을 때 함수 f의 출력값이다. 여기서 귀납적 학습방법이란 함수 f의 예제들이 주어졌을 때, f의 근사함수(Approximator) h를 구하는 것이다. 추정하여야 할 함수 h는 가설(Hypothesis)이라고도 불리운다.

이와 같은 귀납적 학습방법의 정의는 매우 폭넓게 적용할 수 있다. 즉, 인공지능에서 귀납적 학습방법을 대표하는 의사결정수(Decision Tree) 기법들외에도, 학습전략에 의한 학습방법의 분류에서 사례 또는 예제로 부터의 학습 (Learning From Examples)에 해당되는 모든 학습방법들은 귀납적 학습방법에 속한다. 따라서, 신경망은 대표적인 귀납적 학습방법으로 분류될 수 있으며, 통계학의 많은 분석방법들도 귀납적 학습방법들이라고 볼 수 있다.

본 논문에서는 분류작업에 사용할 수 있는 네가지의 귀납적 학습방법들, 즉 통계학의 다변량 판별분석(MDA), 자동학습 분야에서 발전된 의사결정수 기법인 C4.5, 및 백프로파게이션 학습알고리즘을 사용하는 다계층 퍼셉트론(MLP)과 Cascade Correlation Networks(CCN) 두 개의 신경망 모형들에 대해 분류성능을 비교 분석하고자 한다. C4.5와 CCN에 대해서는 아직 국내에서 분류성능에 대한 검증작업이 이루어지지 않았다.

따라서, 본 연구에서는 국내 신용평가 자료들을 사용하여 네가지 귀납적 학습방법들의 분류성능을 비교분석함으로써 C4.5 와 CCN의 국내 자료에의 적용가능성 및 적용시 유의사항들을 검토하고자 한다. 다음 절에서는 일반적 분류모형에 대한 설명과 본 연구에서 사용된 분류방법들에 대해 간단히 설명한다. 제 3절에서는 본 연구에 사용된 국내 신용평가 자료의 성격 및 자료의 전처리에 관련된 사항들을 설명하고, 제 4절에서는 채택된 귀납적 학습방법들의 분류성능에 대해 분석한다.

2. 분류함수로서의 귀납적 학습방법론들

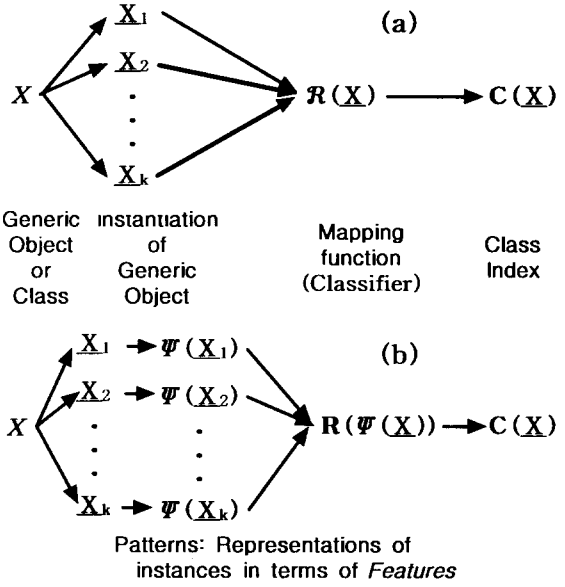
2.1. 분류작업의 일반적 모형

분류작업은 인류의 지식축적에 많은 공헌을 해왔으며 현재도 수많은 형태의 분류작업이 이루어지고 있다. <그림 1-(a)>는 인간이 자연스럽게 행하는 분류작업을 모형화한 것이다. 인간이 분류작업을 행할 때는 분류할 집단내의 각 개체들을 인식하고 표현할 수는 없지만 어떤 기준, 즉 해당 개체를 사전에 정해진 등급중의 하나(Class Index)로 대응시키는 함수(Mapping Function, $\Re(\underline{X})$)를 사용한다.

<그림 1-(b)>는 분류작업을 컴퓨터내에 구현했을 때의 모형으로서 <그림 1-(a)>와의 차이점은 개체의 인식을 위하여 개체의 특성치들을 확정지워주는 특성추출기(Feature Extracture, $\Psi(\underline{X})$)들이 필요하다는 것이다. 즉, 대응함수 또는 분류함수에의 입력값들을 제공하는 별도의 기능이 필요하며, 대응함수($\Re(\Psi(\underline{X}))$)도 구체적인 함수의 형태로 표현되어야 한다는 것이다.

<그림 1-(b)>는 분류작업을 컴퓨터에 의해 행할 경우 두가지 요소, 즉 특성추출기와 대응함수

* 정보통신부 '97 국책기술개발사업의 지원을 받았음



<그림 1> 분류작업의 일반적 모형
(a) 인간의 분류작업 (b) 컴퓨터구현 분류작업

또는 분류기(Classifier)의 성능이 바로 분류작업의 성능을 좌우하며, 이 양자는 서로 보완적인 관계에 있다는 사실을 말해준다. 즉, 원하는 분류성과를 얻기 위해 사용된 특성추출기의 기능이 약할 경우에는 강력한 분류기를 사용하여야 하며, 역의 관계도 마찬가지로 성립된다. 따라서, 분류기들의 공정한 성능평가를 위해서는 가능한 동일한 특성추출기들이 사용되어야 한다.

2.2. 다변량 판별분석

다변량 판별분석은 (1) 각 집단은 다변량 정규분포를 따르는 모집단으로 부터 추출되었으며, (2) 각 집단은 동일한 공분산 행렬(Covariance Matrices)를 가져야 한다는 매우 엄격한 통계적 가정을 요구한다. 하지만 다변량 판별분석은 오랜기간 다양한 응용분야에 이용되어 왔으며 아직도 기본적인 분류함수로서 이용되고 있다. 본 연구에서는 다변량 판별분석에 있어 Step-wise 절차를 사용하여 많은 입력변수들 중에서 판별력있는 입력변수를 선정할 수 있도록 하였다.

2.3. C4.5

C4.5는 Quinlan(1993)에 의해 ID3에 이어 개발된 것으로써 주어진 사례들을 사전에 정의된 범주(Classes)와 속성(Properties)들의 관계를 파악하여 의사결정수(Decision Tree)를 형성해 주는데 ID3에서 생길 수 있는 속성들의 수 및 형태에 의한 편향성을 보완한 프로그램이라 할 수 있다.

C4.5는 의사결정수의 해석이 복잡하다는 점에 착안하여 의사결정수로부터 자동적으로 IF-THEN 규칙을 생성해낼 수 있는 기능과, 학습된 의사결정수의 일반화 능력을 제고시킬 수 있는 가지치기(Pruning) 기능을 제공한다. C4.5는 분류성능이 매우 강력하면서도 사용이 편리한 분류함수이다.

2.4. 다계층 퍼셉트론

다계층 퍼셉트론은 다양한 신경망 모형들 중에서도 가장 많이 사용되고 있는 모형으로 백프로파게이션 알고리즘에 의해 학습되어진다. 본 연구에서는 다음에 설명할 CCN의 분류성능을 객관적으로 평가하기 위해 사용하였다.

2.5. Cascade Correlation Networks

Fahlman과 Lebiere[90]는 백프로파게이션학습의 속도가 느린 원인으로 학습과정에서 Step-size 결정와 Moving Target의 두가지 문제를 지적하고 있다. Step-size 결정문제는 표준 백프로파게이션학습에 있어 전역오차함수에 대해 1次 부분도함수만을 사용하여 gradient, 즉 가중치의 갱신량을 계산하려 하기 때문에 발생한다. 이문제를 완화하기위해 관성항을 사용하거나 2次 도함수를 사용하는 방법들이 많이 제시되었다. 두번째로 Moving Target 문제는 다계층퍼셉트론의 경우 신경망의 구조가 고정되어 있고 변화하는 입출력 패턴에 대해 내부가중치들이 동시에 모두 변화하게되므로 생기는 문제이다. 즉 신경망 내부의 모든 노드들이 동시에 유용한 'feature detector'가 되려하기 때문에 오히려 학습속도를 늦추고 경우에 따라서는 학습결과의 편차가 심해지는 현상을 보이게 된다는 것이다.

CCN은 백프로파게이션 학습방법에 의한 다계층 퍼셉트론의 한계를 벗어나기 위하여 Step-size 결정문제는 Quickprop 학습방법에 의해 해결하고, Moving Target 문제에 대해서는 새로운 신경망 구조를 채택하였다. CCN의 구조는 학습초기 은닉층이 없는 상태로 시작하여 학습이 진행되면서 은닉층의 노드를 하나씩 추가해 나가는 데, 추가되는 은닉층 노드에의 입력으로 입력층의 모든 노드와 은닉층내의 기존의 노드들이 모두 사용하는 층화구조를 가진다. 이러한 구조의 장점은 은닉층 노드들이 입력에 대한 고차원 특성추출기(High-order Feature Detector)로서 작용한다는 점과 사전에 은닉층의 노드 수를 결정할 필요가 없다는 것이다.

3. 분석대상 자료

네 개의 귀납적 학습방법들의 분류성능을 평가하기 위하여 국내 기업어음(CP)의 신용평가 등급자료를 학습자료로 사용하여 분류함수들의 등급예측력을 평가하였다.

분석에 사용된 자료들은 국내 3대 신용평가기관들인 한국신용평가(주), 한국신용정보(주), 한국기업평가(주)가 발표한 신용등급과 한국신용평가(주)의 재무정보프로그램인 FAS에서 해당기업의 재무정보를 얻을 수 있는 경우로 한정하였으며 분석대상기간은 1992년, 1993년, 및 1994년의 3개년이다.

현재 3개 신용평가 기관 모두 동일한 신용등급으로 A1, A2, A3, B, C, D의 6개 등급을 사용하며

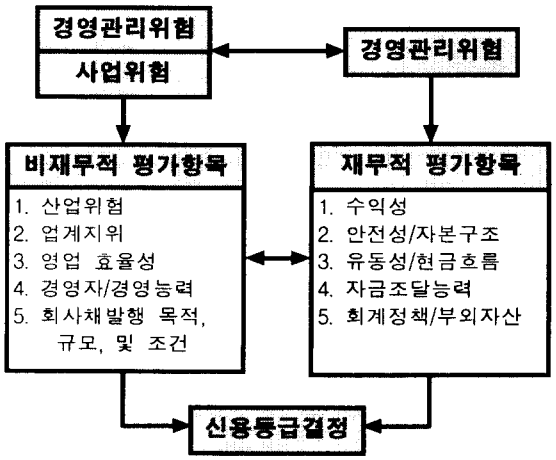
동일 등급내에서의 차이를 나타내기 위해 ± 기호를 겸용하고 있다. 본 연구에서는 ±에 의한 구분은 사용하지 않았으며, C등급과 D등급의 경우는 분석기간 동안의 데이터 수가 총 10개 미만이어서 제외하였다. 업종구분은 1992년도 한국신용평가(주)에서 제공된 자료에는 10개의 업종으로 구분되어 있지만 업종별로 분석에 필요한 데이터 수를 확보할 수 없어 제조업으로 한정하였으나 제조업내에서 1992년 기준으로 11개의 업종을 다시 8개로 최종 정리하여 사용하였다.

수집된 자료들은 분류함수들을 추정하기 위한 훈련집합과 추정된 분류함수의 성능을 평가하기 위한 검증집합의 두 개의 집합으로 분류하였는데, 임의추출 방법에 의해 다섯개의 훈련집합 및 검증집합의 쌍을 생성하였다. <표 1>은 연도별 훈련집합과 검증집합내의 자료수를 정리한 것이다.

<표 1> 분석에 사용된 연도별 자료수

연도	자료구분	A1	A2	A3	B	합계
92	훈련집합	16	68	59	28	171
	검증집합	6	23	20	10	59
	합계	22	91	79	38	230
93	훈련집합	21	69	59	25	174
	검증집합	8	23	20	9	60
	합계	29	92	79	34	234
94	훈련집합	18	50	48	31	147
	검증집합	6	17	17	11	51
	합계	24	67	65	42	198

신용평가에 사용될 특성추출기들 다시 말해서 분류함수들을 위한 입력변수들은 <그림 2>에서와 같이 재무변수와 비재무적변수의 두가지로 크게 나눌 수 있으며 각각의 항목들은 다시 많은 특성치들이나 재무비율들에 의해 구성된다. 본 연구에서는 재무변수의 경우 (1)성장성 지표 9개 (2)기업규모 지표 12개 (3)수익성 지표 37개 (4)안전성 지표 17개 (5)활동성 지표 13개 (6)생산성 지표 16개 (7)현금흐름 지표 6개 등 총 110개의 재무지표들이 사용되었고 비재무적 변수로는 업종만이 고려되었다.



<그림 2> 기업신용평가의 과정

수집된 자료들을 분류함수의 학습에 사용하기 전에 자료의 전처리 과정을 수행하였다. SAS에서 제공하는 기능인 단일변수 Box-Plot을 이용하여 각 재무변수들이 갖는 이상치들을 제거하였다. Missing Value의 경우에는 이상치를 제거한 후 동종업종, 동일 신용등급의 평균값으로 설정하였다. 11개의 규모지표와 4개의 생산성지표들은 변수값들의 크기와 범위가 타변수에 비해 상대적으로 매우 크므로 범주형(Categorical)변수로 변환하였다.

4. 분류성능의 분석

준비된 학습자료들에 대해 네가지 귀납적 학습방법들에 의한 분류함수를 추정하는데 있어 Stepwise MDA와 C4.5는 사전에 입력변수의 수를 확정할 필요가 없다. 즉 이 두 방법론은 입력가능한 모든 변수들을 검토하여 판별력이 높은 변수들만으로 분류함수를 구성하기 때문이다. 하지만 신경망의 경우는 사전에 입력으로 사용될 변수들를 확정하여야 한다.

본 연구에서는 Stepwise MDA와 C4.5에서 선정된 변수들을 함께 고려하여 신경망에 사용할 입력변수들을 확정하였다. <표 2>는 5개의 훈련집합들에 대해 각각 추정된 MDA와 C4.5의 분류함수에 선정된 입력변수들의 개수를 지표별로 분류하여 평균값을 연도별로 구한 것이다. 또, 신경망에 해당하는 열은 MDA와 C4.5의 결과를 이용하여 신경망에의 입력변수로 확정한 변수들의 개수를 나타낸다.

<표 3과 4>는 실험결과를 정리한 것이다. 훈련집합에 대해서는 MDA에 비해 나머지 방법들이 높은 학습률을 보인 반면 검증집합에 대해서는 MDA에 비해 우월한 결과를 보이지 못했다. 비록 CCN의 결과가 가장 좋은 것으로 나타났지만 통계적으로 유의한 수준은 아니었다. 이러한 현상에 대한 해석은 본 연구에 사용된 변수들이 재무변수에 치우쳐 있고, 사용된 자료들이 등급발표 시점에서의 해당기업의 재무상태를 반영하지 못했다는 사실을 감안하면 실망스러운 결과는 아니다.

<표 2> 선정된 입력변수 수의 평균

	92년			93년			94년		
	MDA	C4.5	신경망	MDA	C4.5	신경망	MDA	C4.5	신경망
성장성	1.8	2.6	1	3.2	2.4	3	1.4	1.8	1
규모	2.6	4.8	3	2.4	5.0	3	2.6	4.6	4
수익성	5.4	4.8	5	6.2	6.4	6	5.6	5.4	6
안전성	2.6	0.6	1	1.8	0.8	1	2.4	1.8	1
활동성	2.2	1.0	1	1.8	0.6	1	2.0	0.8	1
생산성	2.4	2.0	5	3.0	2.4	4	2.6	2.4	0
현금흐름	1.6	0.4	1	1.2	0.6	0	0.4	0.2	0
합계	18.6	16.2	17	19.6	18.2	18	17.0	17.0	15

실험결과중 흥미있는 사실들은 C4.5의 경우 가지치기를 한 의사결정수의 경우가 일반화 능력이 좋았으며 프로덕션 규칙의 생성방법론은 보완될 필요가 있는 것으로 생각된다. 또, CCN의 경우에는 주어진 훈련집합에의 학습능력이 매우 뛰어나므로 과잉적합의 문제를 적절히 제어하는 것이 쉽지 않다는 것이다.

5. 결 론

MDA, C4.5, MLP 및 CCN의 네 모형을 분류함수로 사용하여 국내 신용평가자료에 적용한 결과 CCN이 가장 좋은 예측력을 보였지만 통계적으로 유의하지는 않았다. 하지만 C4.5와 CCN의 분류함수로서의 유용성은 입증되었다. 특히 사용상의 어려움은 있지만 CCN의 활용가능성은 앞으로 계속 연구되어야할 과제이다.

분류작업에 있어서 특성추출기와 분류기의 상호보완 작용을 감안하여 귀납적 학습방법들에 대한 입력변수의 선정에 있어 가능하면 공정한 특성추출기로서의 역할을 할 수 있도록 본 연구에서 사용된 재무변수들에 대해 전처리 작업을 시도하였다. 하지만 본 연구에서도 귀납적 학습방법은 사용된 자료가 갖는 한계를 넘어설 수 있는 없다는 사실은 확인하였다.

<표 3> 훈련집합에 대한 실험결과

연도	실험방법		평균 예측률
92년	MDA		78.8%
	C4.5	Unpruning Decision Tree	87.2%
		Pruning Decision Tree	82.2%
		Production Rule	64.6%
	신경망	Backpropergation (은닉노드=입력노드)	92.6%
		Backpropergation(은닉노드=20)	90.6%
		Cascade-correlation	90.5%
93년	MDA		79.8%
	C4.5	Unpruning Decision Tree	92.2%
		Pruning Decision Tree	87.6%
		Production Rule	73.2%
	신경망	Backpropergation (은닉노드=입력노드)	88.0%
		Backpropergation(은닉노드=20)	92.2%
		Cascade-correlation	83.6%
94년	MDA		85.2%
	C4.5	Unpruning Decision Tree	93.8%
		Pruning Decision Tree	89.8%
		Production Rule	75.8%
	신경망	Backpropergation (은닉노드=입력노드)	90.4%
		Backpropergation(은닉노드=20)	92.2%
		Cascade-correlation	95.6%

6. 참고문헌

Falman S.E. & Lebiere, C., "The Cascade Correlation Learning Architecture", Technical Report, CMU-CS-90-100,, 1990.

Robert S. Kaplan, Gabriel Urwitz, "Statistical Models of Bond Ratings: a Methodological Inquiry", Journal of Business Vol.52 No.2, 1979

Pao, Y.H., Adaptive Pattern recognition and Neural Networks., Addison-Wesley, 1988.

J.Ross Quinlan, "C4.5 : Programs for Machine Learning", Morgan Kanfmann Publishers, San Mateo, Califonia, 1992

Kar Yan Tam, Melody Y. Kiang, "Managerial Applications of Neural Networks : The Case of Bank Failure Predictions", Journal of Management Science Vol.38 No.7, July 1992

J. Utans, J. Moody, "Architecture Selection Strategies for Neural Networks: Application to Corporate Bond Rating Prediction", Neural Networks in the Capital Markets, John Wiley & Sons, 1994

한국기업평가, "기업어음 신용등급가이드", 한국기업평가주식회사, 1992-1994

한국신용평가, "기업어음 신용등급가이드", 한국신용평가주식회사, 1992-1994

한국상장회사위원회, "상장회사총람", 한국상장회사위원회, 1996

<표 4> 검증집합에 대한 실험결과

연도	실험방법		평균예측률
92년	MDA		56.8%
	C4.5	Unpruning Decision Tree	57.8%
		Pruning Decision Tree	59.6%
		Production Rule	49.6%
	신경망	Backpropergation (은닉노드=입력노드)	53.4%
		Backpropergation(은닉노드=20)	53.6%
		Cascade-correlation	58.0%
93년	MDA		57.4%
	C4.5	Unpruning Decision Tree	55.0%
		Pruning Decision Tree	56.4%
		Production Rule	55.6%
	신경망	Backpropergation (은닉노드=입력노드)	55.2%
		Backpropergation(은닉노드=20)	56.4%
		Cascade-correlation	57.6%
94년	MDA		58.8%
	C4.5	Unpruning Decision Tree	54.2%
		Pruning Decision Tree	55.4%
		Production Rule	55.0%
	신경망	Backpropergation (은닉노드=입력노드)	54.6%
		Backpropergation(은닉노드=20)	55.8%
		Cascade-correlation	60.6%