데이터마이닝을 위한 혼합 데이터베이스에서의 속성선택

차운옥[1], 허문열[2]

요 약

데이터마이닝을 위한 대용량 데이터베이스를 축소시키는 방법 중에 속성선택 방법이 많이 사용되고 있다. 본 논문에서는 세 가지 속성선택 방법을 사용하여 조건속성 수를 60%이상 축소시켜 결정나무와 로지스틱 회귀모형에 적용시켜보고 이들의 효율을 비교해 본다. 세 가지 속성선택 방법은 MDI, 정보획득, ReliefF 방법이다. 결정나무 방법은 QUEST, CART, C4.5를 사용하였다. 속성선택 방법들의 분류 정확성은 UCI 데이터베이스에 주어진 Credit 승인 데이터베이스와 German Credit 데이터베이스를 사용하여 10층-교차확인 방법으로 평가하였다.

주요용어: 분류(classification), 결정나무(decision tree), 로지스틱 회귀모형(logistic regression), 속성선택(feature selection)

1. 서론

데이터마이닝 방법 중 분류(classification)는 주어진 대상들을 이미 주어진 몇 개의 클래스 중 하나로 분류하기 위한 분류기(classifier)를 학습하거나 분류를 위한 지식을 생성하기 위해 사용한다. 데이터베이스를 구성하는 대상들은 속성(feature)과 속성의 값(value)으로 묘사되며, 속성들은 조건속성(condition feature)과 결정속성(decision feature)으로 이루어진다. 속성들의 수가 상당히 많을 경우 결정속성에 영향을 미치지 않는 부적절하고 불필요한 속성들이 다수 포함되어 있을 수 있다. 데이터베이스의 크기가 상당히 클 때 미리 부적절하고 불필요한 속성들을 제거하면 학습시간도 줄여줄 수 있고 보다 명료하고 일반적인 지식을 추출할 수 있다. 속성선택의 문제는 이미 통계학 분야에서도 많이 연구되어왔으며, 패턴 인식 분야에서도 많이 다루어진 분야이다 (Devijver and Kittler, 1982, Miller, 1990).

본 논문의 목적은 속성이 연속형과 이산형으로 구성되어 있고 다수의 부적절한 속성이 포함되어 있는 혼합 데이터베이스에서 중요한 속성들을 선택하는 속성선택 방법에 대해 연구하고, 이를 대표적인 분류방법인 결정나무와 로지스틱 회귀모형에 적용하여 분류결과를 분석하고자 하는 것이다.

2. 분류 방법

분류를 위한 방법에는 기계학습, 통계학, 인공신경망 방법 등이 있다. 본 논문에서는 기계학습의 대표적인 방법인 결정나무 모델 중 QUEST (Loh and Shi, 1997), CART(Breiman 외,

1) 한성대학교 공과대학 컴퓨터공학부 교수, 서울시 성북구 삼선동 2가 389. wcha@hansung.ac.kr

2) 성균관대학교 경제학부 교수, 서울시 종로구 명륜동 3가 53번지. myhuh@skku.edu

1988), C4.5 (Quinlan, 1998)와 로지스틱 회귀모형에 대해 속성선택 방법을 적용하였다.

3. 속성선택 방법

속성선택은 데이터베이스에서 부적절하고 불필요한 정보를 가능한 한 많이 찾아내어 제거하는 과정이다. 데이터의 차원을 축소시키면 학습 알고리즘이 보다 빨리 효율적으로 수행될 수 있으며, 목표 개념에 대한 학습 결과가 보다 간결하고 이해하기 쉽게 된다.

속성선택 문제는 결정속성에 영향을 미치는 조건속성에 대해 중요한 속성 순으로 순위를 정하거나, 조건속성들의 집합 중 결정속성에 영향을 미치는 가장 적절한 속성들의 부분집합을 찾는 탐색문제이며, 탐색방법과 속성들에 대한 평가방법에 따라 다양한 속성선택방법들이 연구되었다 (Dash and Liu, 1997, Liu and Motoda, 1998). 속성선택방법은 필터 (filter)와 포장 (wrapper) 방법으로 분류할 수 있는데, 필터방법은 속성들을 평가할 때 학습 알고리즘과는 독립적으로 데이터가 가지고 있는 성질을 이용하는 방법이고, 포장방법은 목표 알고리즘에 의한 분류 정확도를 사용하여 속성들을 평가하는 방법이다. 필터방법에서 평가에 사용되는 척도로는 종속성(dependency), 거리(distance), 정보(information), 일치성(consistency)이 있다 (Das and Liu, 1997, Liu and Motoda, 1998). 종속성은 하나의 조건속성이 결정속성에 얼마나 강하게 연관되어 있는지를 재는 척도이고, 거리는 클래스들을 가능한 한 멀리 분리해 줄 수 있는 조건속성을 찾는 척도이며, 정보는 결정속성의 불확실성을 줄여 줄 수 있는 조건속성을 찾는 척도로 사용된다. 또, 일치성은 모든 속성이 다 사용되었을 때와 동일하게 클래스를 분류할 수 있는 최소수의 조건속성을 찾는데 사용할 수 있는 척도이다.

본 논문에서는 속성들이 가지는 값이 연속형, 이산형인 혼합 데이터베이스에 적용가능한 필터 방법으로서 종속성 척도를 사용하는 MDI 방법 (이승천과 허문열, 2003), 거리 척도를 사용하는 ReliefF 방법 (Kononenko, 1995), 정보척도를 사용하는 정보획득 (information gain) 방법을 적용하고자 한다.

3.1 MDI 방법(MDI)

MDI (Measure of Separture from Independence)는 혼합 데이터베이스에서 두 변수간의 연관성을 측정하기위해 두 변수간의 독립성 검정방법을 사용하고 있다. 두 변수가 모두 연속형인 경우 두 변수간의 독립성 검정은 Pearson 의 순위상관계수 검정을 사용하고 있으며, 연속형과 이산형 자료의 경우는 Kruskal-Wallis 검정을, 그리고 두 변수 모두 이산형인 경우 피어슨의 카이제곱검정을 사용하고 있다. 이들은 독립성 검정을 수행하는 데 사용되는 검정통계량의 유의확률 p를 사용하여 혼합변수들간의 연관성을 측정하여도 좋은 결과가 나타나는 것을 보여주고 있다.

3.2 ReliefF 방법(REL)

Relief방법 (Kira and Rendell, 1992) 에서는 데이터베이스에서 하나의 대상을 임의로 추출하여, 이 대상과 결정속성 값이 같은 클래스와 다른 클래스로부터 최근접 이웃을 찾아내고 이 대상의 속성들의 값과 각 클래스에 속하는 최근접 이웃들에 속하는 속성들의 값을 비교하여 이들 속성들의 점수를 갱신한다. 이 과정을 사용자가 미리 정한 대상들의 수가 될 때까지 반복하여 결정속성에 영향을 미치는 속성들의 순위를 정한다. 이 방법은 유용한 속성이란 서로 다른 클래스에 속하는 대상들에서는 다른 값을 가지고, 같은 클래스에 속하는 대상들에서는 같은 값

을 가지게 될 것이라는 것에 바탕을 둔 것이다. Relief는 결정속성이 가지는 값이 두 클래스일 때 사용할 수 있는 방법이며, 데이터가 잡음을 가지고 있거나 세 개 이상의 클래스인 경우에 사용할 수 있도록 확장한 방법이 ReliefF이다.

3.3 정보획득 방법 (INF)

*C*와 *D*를 조건속성과 결정속성이라 할 때, 조건속성을 관찰하기 전과 후의 결정속성의 엔트로피는 다음 식과 같다.

$$H(D) = -\sum_{d \epsilon D} p(d) \log (p(d))$$

$$H(D|C) = -\sum_{c \epsilon C} p(c) \sum_{d \epsilon D} p(d|c) \log (p(d|c))$$

결정속성의 엔트로피가 감소하는 양은 조건속성에 의해 제공되는 결정속성에 대한 추가적인 정보를 반영하는 것이고 각각의 조건속성 Ci에 대한 정보획득은 $H(D) - H(D|Ci)$와 같다. 이 방법에서는 각각의 조건속성과 결정속성 사이의 정보획득에 근거한 점수가 할당되어 조건속성들의 순위가 결정된다.

4. 혼합 데이터베이스

본 논문의 실험에 사용한 혼합 데이터베이스는 UCI 데이터 창고에 있는 Credit 데이터베이스와 German Credit 데이터베이스로서 이들은 모두 연속형과 이산형 자료가 혼합되어있다 (Merz and Murphy, 1996).

(1) Credit 승인 데이터베이스

명목형 결정속성(클래스의 수는 2)과 15개의 조건속성으로 이루어지고, 데이터베이스의 크기는 690이다.

연속형 값을 갖는 조건속성 : C2, C3, C8, C11, C14, C15
명목형 값을 갖는 조건속성 중 값의 종류가 2개인 속성 : C1, C4, C5, C9, C10, C12, C13
명목형 값을 갖는 조건속성 중 값의 종류가 9개인 속성 : C7
명목형 값을 갖는 조건속성 중 값의 종류가 14개인 속성 : C6

(2) German Credit 데이터베이스

명목형 결정속성(클래스의 수는 2)과 20개의 조건속성으로 이루어지고 데이터베이스의 크기는 1000이다.

연속형 값을 갖는 조건속성 : C2, C5, C8, C11, C13, C16, C18
명목형 값을 갖는 조건속성 중 값의 종류가 2개인 속성 : C19, C20
명목형 값을 갖는 조건속성 중 값의 종류가 3개인 속성 : C10, C14, C15
명목형 값을 갖는 조건속성 중 값의 종류가 4개인 속성 : C1, C12, C17
명목형 값을 갖는 조건속성 중 값의 종류가 5개인 속성 : C3, C6, C7, C9
명목형 값을 갖는 조건속성 중 값의 종류가 9개인 속성 : C4

5. 실험

5.1 실험방법

속성선택을 위하여 MDI 방법은 R (Ihaka and Gentleman, 1967) 을 사용하였으며 ReliefF 방법과 정보획득 방법은 공개 소프트웨어 WEKA (Witten and Frank, 1999)를 사용하였다. 또한 결정나무 QUEST는 QUEST 1.8.19 (Shih, Y, 2003), CART는 CART 5.0 (Salford Systems, 2003) 을 사용하였고, C4.5와 로지스틱 회귀모형은 WEKA를 사용하였다. 결정나무 방법에서 노드의 최소 대상 수를 5로 하였고, 가지치기를 위해서는 QUEST, CART에서는 1*SE* 규칙, C4.5에서는 감소된 오류 가지치기 (reduced error pruning) 방법을 적용하였다. 두 실험 데이터베이스에 대해서 10-층 교차확인(10-fold cross-validation) 방법으로 분류의 정확도를 구하였다.

5.2 실험결과

(1) Credit 승인 데이터베이스

3가지 속성선택방법을 적용하여 15개의 조건속성으로부터 가장 영향력이 있는 6개의 속성들만 택하였다. 선택된 속성들은 다음과 같다 (연속형 속성은 밑줄로 표시).

MDI : C4, C5, C6, C7, C9, C10

REL : C1, C6, C7, C9, C10, C12

INF : C6, C8, C9, C10, C11, C15

15개 조건속성을 다 사용하였을 때 (FULL) 와 선택된 속성만을 사용하였을 때의 10-층 교차확인 방법에 의한 분류행렬은 <표 1>에, 분류정확도와 결정나무에서 생성된 트리의 크기 (종료마디의 수/전체트리의 크기) 는 <표 2>에 나타내었다. • 표시는 모든 조건속성을 다 사용했을 때보다 정확도가 같거나 좋아진 경우이다.

<표 1> 분류행렬

	QUEST		CART		C4.5		Logistic	
FULL	284	23	284	23	267	40	263	44
	77	306	77	306	61	322	56	327
MDI	284	23	284	23	268	39	274	33
	77	306	77	306	58	325	63	320
REL	284	23	284	23	270	37	262	45
	77	306	77	306	63	320	63	320
INF	284	23	284	23	276	31	261	46
	77	306	77	306	74	309	61	322

<표 2> 분류정확도와 트리크기

	QUEST	CART	C4.5	Logistic
FULL	0.855074(2/3)	0.855074(2/3)	0.853623(13/25)	0.855072
MDI	• 0.855074(2/3)	• 0.855074(2/3)	• 0.85942(7/13)	• 0.86087
REL	• 0.855074(2/3)	• 0.855074(2/3)	• 0.855072(2/3)	0.843478
INF	• 0.855074(2/3)	• 0.855074(2/3)	0.847826(5/9)	0.844928

(2) German Credit 데이터베이스

3가지 속성선택방법을 적용하여 20개의 조건속성으로부터 가장 영향력이 있는 6개의 속성들만 택하였다. 선택된 속성들은 다음과 같다 (연속형 속성은 밑줄로 표시).

MDI : C1, C2, C3, C6, C12, C15

REL: C1, C3, C4, C6, C7, C9

INF : C1, C2, C3, C4, C5, C6

20개 조건속성을 다 사용하였을 때 (FULL) 와 선택된 속성만을 사용하였을 때의 10-층 교차확인 방법에 의한 분류행렬은 <표 3>에, 분류정확도와 결정나무에서 생성된 트리의 크기 (종료마디의 수/전체트리의 크기) 는 <표 4>에 나타내었다. • 표시는 모든 조건속성을 다 사용했을 때보다 정확도가 같거나 좋아진 경우이다.

<표 3> 분류행렬

	QUEST		CART		C4.5		Logistic	
FULL	668	32	620	80	589	111	606	94
	233	67	166	134	163	137	150	150
MDI	668	32	613	87	595	105	629	71
	233	67	164	136	167	133	177	123
REL	618	82	658	42	617	83	618	82
	202	98	241	59	185	115	179	121
INF	668	32	616	84	607	93	624	76
	233	67	153	147	182	118	174	126

<표 4> 분류정확도와 트리크기

	QUEST	CART	C4.5	Logistic
FULL	0.735(4/7)	0.754(5/9)	0.726(30/59)	0.756
MDI	• 0.735(4/7)	0.749(5/9)	• 0.728(20/39)	0.752
REL	0.716(10/19)	0.717(3/5)	• 0.732(10/19)	0.739
INF	• 0.735(4/7)	• 0.763(10/19)	0.725(31/61)	0.75

5.3 평가

Credit 승인 데이터베이스의 15개의 조건속성 중 6개와 German Credit 데이터베이스의 20개 조건속성 중 6개를 선택하여 실험해 본 결과, 전체 12번의 실험에서 Credit 승인 데이터베이스의 경우 전체 조건속성을 다 사용하였을 때보다 정확도가 같거나 좋아진 경우가 9번, German Credit 데이터베이스의 경우는 5번 있음을 확인할 수 있고, 트리 크기가 많이 축소됨도 알 수 있다. 세 가지 결정나무 방법의 정확도 결과에서는 큰 차이를 발견할 수 없었고 C4.5 방법이 QUEST, CART보다 더 큰 트리를 생성하였다. 세 가지 속성선택 방법 중에서는 MDI 방법이 ReliefF와 정보획득 방법보다 좋은 결과를 보여준다. 다른 다양한 혼합 데이테베이스에 적용하여 실험해 보는 것이 더 필요하다.

6. 결론

혼합 데이터베이스에 세 가지 속성선택 방법을 사용하여 조건속성 수를 60%이상 축소시켜 결정나무와 로지스틱 회귀모형의 분류정확도를 구한 결과, 전체 속성을 다 사용하는 경우보다 오히려 좋은 결과를 내거나 크게 나쁘지 않은 결과를 얻을 수 있었다. 또한 결정나무의 트리의 크기가 많이 축소됨으로서 결정나무로부터 얻을 수 있는 분류를 위한 규칙의 형태가 보다 간단해 질 수 있다. 따라서 속성선택 방법을 혼합 데이터베이스 분류를 위한 전처리 과정에 사용하여 중요하고 적절한 속성을 선택함으로서 분류가 효율적으로 수행될 수 있도록 하고, 보다 명료하고 간단한 분류를 위한 지식을 얻을 수 있다.

참고문헌

[1] 이승천, 허문열.(2003). "혼합자료에서 독립성검정에 의한 연관성 측정", *응용통계연구*, 제16권, 제 1호, pp151-167

[2] Breman, L. Friedman, J. H. Olshen, R. A. and Stone, C. J. (1984). *Classification and regression trees,* Wadsworth, Belmont, CA.

[3] Dash, M. and Liu, H. (1997). "Feature selection for classification", *Intelligent Data Analysis*, Elsevier Science Inc.

[4] Devijver, P.A., and Kittler, J. (1982), *Pattern Recognition: A Statistical Approach,* Prentice Hall International

[5] Hall, M. A. and Holmes, G. (2000). Benchmarking Attribute Selection Techniques for Data Mining, http://www.cs.waikato.ac.kr/~ml

[6] Ihaka, Ross and Gentleman, Robert (1996) "R: A language for data analysis and graphics". *Journal of Computational and Graphical Statistics*, 5(3):299-314, (http://www.r-project.org).

[7] Kira, K. and Rendell, L.A. (1992), "The feature selection problem: Traditional methods and a new algorithm", Proceed. of Nat'l Conf. of AI, pp. 129-134

[8] Kononenko , I. (1994), "Analysis and extension of RELIEF" In : Proceed. of European Conference on Machine Learning, pp. 171-182

[9] Liu, H. and Motoda, H. (1998). *Feature Selection for Knowledge discovery and Data Mining*, Kluwer Academic Publishers

[10] Loh, W. and Shih, Y.(1997). "Split selection Methods for Classifiacation Trees", *Statistica Sinica*, Vol 7, pp815-840

[11] Merz, C.J., and Murphy, P.M. (1996), UCI Repository of Machine Learning Databases, Department of Information and Computer Science, University of California, Irvine, CA, (http://www.ics.uci.edu/~mlearn/MLRepository).

[12] Miller, A, J. (1990), *Subset Selection in Rgeression,* Chapman and Hall, New York

[13] Quinlan, J. R.(1998), *C4.5: Programs for machine learning,* Morgan Kaufmann Publishers, San Mateo, California

[14] Salford Systems (2003), CART 5.0, (www.salford-systems.com)

[15] Witten, Ian, and Frank, Eibe (1999), *Data Mining*, Morgan and Kaufmann, (http://www.cs.waikato.ac.kr/~ml)

[16] Shih, Y. *QUEST User Manual* (2003), (http://www.stat.wisc.edu/~loh/quest.html)