데이터마이닝에서 교사학습에 의한 속성 가중치 최적화

강명구*, 차진호°, 김명원
씨씨매디아*. 숭실대학교 컴퓨터학과
mkang@ccmedia.co.kr, jhcha@info.ssu.co.kr, mkim@computing.ssu.ac.kr

Supervised Feature Weight Optimization for Data Mining

Myung-Ku Kang*, Jin-Ho Cha, Myungwon Kim
CCmedia Inc*. ,School of Computing, Soongsil Univ.

요 약

최근 군집화와 분류기법이 데이터 마이닝에 중요한 도구로 많은 응용분야에 사용되고 있다. 따라서 이러한 기법을 이용하는데 있어서 각각의 속성의 중요도가 달라 중요하지 않은 속성에 의해 중요한 속성이 왜곡되거나 때로는 마이닝의 결과가 잘못되는 결과를 얻을 수 있으며, 또한 전체 데이터를 사용할 경우 마이닝 과정을 저하시키는 문제로 속성 가중치과 속성선택에 관한 연구가 중요한 연구의 대상이 되고 있다. 최근 연구되고 있는 알고리즘들은 사용자의 의도와는 상관없이 데이터간의 관계에만 의존하여 가중치를 설정하므로 사용자가 마이닝 결과를 쉽게 이해하고 분석할 수 없는 문제점을 안고 있다.

본 논문에서는 클래스 정보가 있는 데이터뿐 아니라 클래스 정보가 없는 데이터를 분석할 경우 사용자의 의도에 따라 학습할 수 있도록 각 가중치를 부여하는 속성가중치 알고리즘을 제안한다. 또한 사용자가 의도한 정보를 이용하여 속성간의 가장 최적화 된 가중치를 찾아주며, Cramer's V^2함수를 적합도 함수로 하는 유전자 알고리즘을 사용한다. 알고리즘의 타당성을 검증하기 위해 전자상거래상의 실험 데이터와 몇 가지 벤치마크 데이터를 이용하여 본 논문의 타당성을 보인다.

1. 서론

컴퓨터를 이용한 정보기술과 인터넷의 발달로 데이터의 양이 급속도로 증가되고 있으며, 이와 같은 데이터를 수집하고 사용하기 편리하게 처리하는 DBMS(Database Management System), 데이터 웨어하우징(Data Warehousing), 데이터마이닝(Data Mining) 기법들이 많이 연구되고 있다.

더욱이, 데이터마이닝에 있어서 최근에는 기업들의 치열한 경쟁 속에서 살아남기 위해 영업사원의 고객접촉, 사후관리, 고객으로부터 걸려온 전화, 고객에게 판매촉진을 위해 건 전화 등으로 생기는 수많은 데이터를 고객들의 필요에 초점을 두어 일대일로 차별화 된 마케팅을 실시하는 경영기법인 고객관계관리(CRM : Customer Relationship Management)를 가능케 하는 도구가 되고 있다.

그러나 실제 응용분야에서 수집된 데이터는 시간이 지날수록 데이터의 양이 늘어나게 되고 중복되는 속성과 잡음을 갖게 되어 마이닝의 기법을 이용하는데 많은 시간과 비용이 소요된다. 또한 어느 속성이 중요한지 알 수 없어 중요한 속성이 중요하지 않은 속성에 의해 왜곡되거나 제대로 분석되지 않을 수 있다.

따라서, 최근 마이닝 과정의 속도를 향상시키고 효율을 높이기 위해 중요한 속성을 선택하는 부분속성선택(Feature Subset Selection)과 속성의 중요도에 따라 가중치를 부여하는 속성가중치(Feature Weighting)에 관한 연구가 관심의 대상이 되고 있다.[1]

속성선택의 문제는 클래스 정보를 알고 있는 데이터에 적용하여 좋은 분류 결과를 얻을 수 있는 속성들을 선택하는 방법으로 교사학습인 분류기법에 응용되고 있는 반면, 속성가중치 문제는 클래스 정보가 없는 데이터가 주어졌을 때 어떤 속성이 데이터간의 관계를 잘 표현하고 있는가에 따라 가중치를 부여하는 방법으로 비교사 학습인 클러스터링 등에 적용할 수 있다. 그러나 이 경우 클래스 정보가 없는 상태에서 데이터간의 관계만으로 가중치를 주기 때문에 사용자는 그 결과가 잘되었는지에 대한 여부를 판단할 수 없는 문제점을 안고 있다.

본 연구는 한국과학재단 특정기초연구과제 (과제번호 : 98-0102-01-01-3)와 과학기술처 뇌연구개발사업(과제번호:98-J04-01-01-A-04)에 의해 이루어 졌음.

예를 들어, 다음 <표 1>와 같은 전자상거래에서 판매한 사례를 살펴보자. 속성으로는 '나이', '성별', '구입금액', '방문회수' 등 4개이며 42개의 데이터로 구성되어 있다. ('성별'의 1은 남성, 0는 여성을 뜻하며 '방문회수'는 물건을 구입하지 않은 방문도 포함한 회수이다.)

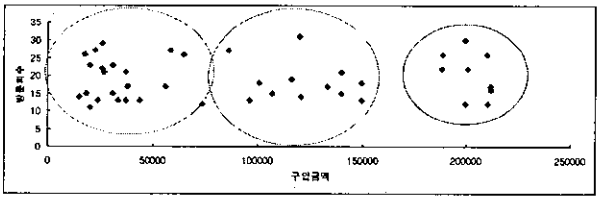
<표 1> 전자상거래 상의 실험 데이터

번호	나이	성별	[illegible]	번호	나이	성별	[illegible]
1	23	1	[illegible]	22	28	1	[illegible]
2	47	0	[illegible]	23	29	0	[illegible]
3	30	1	[illegible]	24	30	1	[illegible]
4	36	1	[illegible]	25	28	1	[illegible]
5	54	1	[illegible]	26	23	1	[illegible]
6	44	1	[illegible]	27	53	1	[illegible]
7	35	1	[illegible]	28	23	0	[illegible]
8	33	0	[illegible]	29	41	1	[illegible]
9	29	0	[illegible]	30	23	1	[illegible]
10	23	1	[illegible]	31	20	0	[illegible]
11	37	1	[illegible]	32	24	1	[illegible]
12	23	1	[illegible]	33	31	1	[illegible]
13	54	1	[illegible]	34	47	0	[illegible]
14	36	1	[illegible]	35	20	0	[illegible]
15	37	0	[illegible]	36	42	1	[illegible]
16	33	0	[illegible]	37	23	0	[illegible]
17	40	0	[illegible]	38	40	0	[illegible]
18	37	1	[illegible]	39	28	1	[illegible]
19	29	0	[illegible]	40	32	1	[illegible]
20	48	1	[illegible]	41	23	0	[illegible]
21	32	0	[illegible]	42	22	0	[illegible]

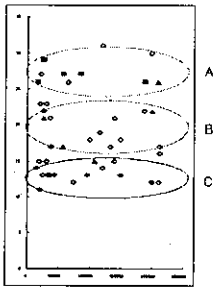
위의 데이터 중에서 '구입금액'과 '방문회수'만을 산점도로 표시하면 아래의 <그림 1>과 같다. 가중치를 전혀 고려하지 않고 클러스터링을 할

경우 '구입금액'의 척도가 '방문회수'의 척도보다 훨씬 크므로 '방문회수'는 클러스터링 결과에 거의 영향을 미치지 못하게 되어 <그림 1>와 같은 결과를 얻게 된다.

만약 사용자가 "*'방문회수'가 많은 고객이 앞으로 판매실적을 높일 것이다*" 라는 가정을 세우고 <그림 2>와 같이 '방문회수'가 많은 고객별로 클러스터링을 하고 싶은 경우에는 다른 속성보다도 '방문회수'에 가중치를 두어야 할 것이다. 그러나 어떤 속성에 얼마정도의 가중치를 주어야 할지 사용자가 판단하는 것은 결코 쉬운 문제가 아니다.



<그림 2> 가중치를 고려하지 않은 클러스터링 결과



<그림 3> 사용자의도에 따라 가중치를 고려한 결과

따라서 본 논문에서는 이와 같이 클래스 정보가 주어진 데이터에 대해서는 클래스 정보대로 잘 분류를 할 수 있는 속성에 높은 가중치를 부여하며, 클래스 정보가 주어지지 않은 데이터를 분석할 경우 사용자가 클래스 정보를 주어 사용자의 의도에 따라 학습할 수 있도록 유전자 알고리즘을 사용하여 각 속성간의 최적화 된 가중치를 부여하는 속성가중치 알고리즘을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 속성선택에 대한 기존의 연구경향과 문제점들을 살펴보고, 3장에서는 유전자 알고리즘과 본 논문이 제안하는 알고리즘을 어떻게 구현하는지에 대해 설명하고 4장에서 몇 가지 실험데이터를 이용하여 제안한 알고리즘의 타당성을 살펴본다. 마지막으로 5장에서는 결론을 맺고 향후 연구 방향을 제시한다.

2. 관련연구

다차원 공간에서 중요성 속성과 그렇지 않은 속성을 구분하여 마이닝에 있어서 데이터 분석의 정확도를 높이며 시간과 비용을 절약하기 위해 중요한 속성을 선택하는 방법들이 제안되었다. 그러나 어떤 속성을 얼마나 선택하느냐에 따라 마이닝의 결과가 다르게 나올 수 있으며 자칫 잡음이 많이 섞인 속성을 취하거나 다른 속성과 반복되는 것을 취할 경우에는 잘못된 결과를 유도할 수 있으므로 어떤 속성을 선택하느냐가 문제가 되고 있다. 속성선택은 크게 부분속성선택(Feature Subset Selection)과 속성가중치(Feature Weighting)로 분류된다.

2.1 부분속성선택

부분속성선택을 정의하면, 다음과 같이 표현할 수 있다. 부분속성선택이란 주어진 d차원 데이터의 속성집합을 Y라하고 속성선택의 판단기준이 되는 함수(Feature Selection criterion function)를 J라 할 때, J의 값을 최대로 하는 부분집합 즉,

$$J(X) = \max_{Z \subseteq Y, |Z| = k} J(Z) \qquad (식.1)$$

을 만족하는 $k(\leq d)$개의 속성을 가진 부분집합 $X(\subseteq Y)$를 찾는 것이다. d개의 속성 중에서 $k(\leq d)$개의 속성을 선택할 경우의 수는

$$\binom{d}{k} = \frac{d!}{k!(d-k)!} \qquad (식.2)$$

와 같으며, 이는 NP-complete문제이다.

이와 같이 차원 d가 커짐에 따라 기하급수적으로 늘어나는 공간을 모두 탐색하는 것은 불가능하기 때문에 최근에는 최적해 문제에 많이 사용되는 유전자 알고리즘을 이용하여 부분속성들을 선택하는 방법들이 많이 제안되고 있다.[1]

최적의 부분속성선택을 구현하기 위해 특별한 학습알고리즘을 적합도 함수로 하는 알고리즘이 많이 제안되고 있으며, 대부분의 부분속성선택 알고리즘은 클래스의 정보를 알고 있는 교사학습법에 의한 속성선택으로 마이닝의 분류기법 등에 응용되고 있다.

2.2 속성 가중치

마이닝의 주요한 기법 중에 하나인 클러스터링은 클래스의 정보가 없는 데이터를 분석하는 기법으로, 클래스 정보를 알아야만 하는 기존의 부분속성선택방법을 적용할 수 없다. 따라서 클러스터링과 같은 비교사학습 알고리즘에 적용하기 위해 클래스 정보가 없이 데이터의 분포와 관계를 나타내는 평가함수을 이용하여 평가함수값이 최고(또는 최소)가 되도록 가중치를 부여하는 속성가중치방법이 제안되고 있다.[2][3][4]

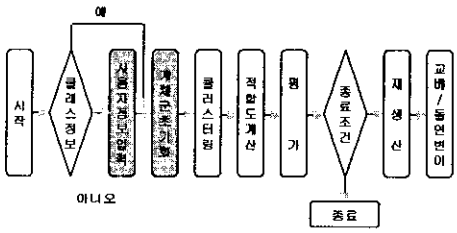
속성가중치는 비교사 학습에 의한 속성선택으로 고전적으로 사용되어져 온 통계적인 방법인 주성분 분석과 데이터간의 관계를 이용한 휴리스틱 함수를 극대화하는 가중치를 찾아주는 방법으로 구분될 수 있다.

2.2.1 휴리스틱 방법

신경망, 퍼지와 더불어 유전자 알고리즘과 같이 인고지능 기법을 이용한 학습, 탐색 알고리즘이 활발히 연구됨에 따라 속성가중치의 문제도 이와 같은 인공지능기법을 이용한 연구가 진행되고 있다.[2][3]

3. 속성가중치 알고리즘

본 알고리즘의 순서도는 <그림 3>과 같다. 전체적인 알고리즘은 유전자 알고리즘의 순서도를 따른다.



<그림 4> 알고리즘의 순서도

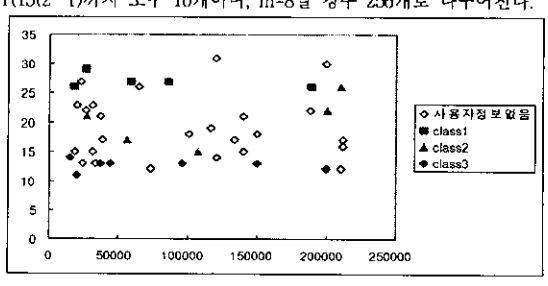
3.1 사용자 정보 입력

<표 1>과 같이 클래스 정보가 없는 데이터를 분석하기 위해서 사용자는 몇 개의 표본 데이터를 추출하여 사용자가 원하는 대로 학습할 수 있도록 각각의 표본 데이터에 대략적인 클래스 정보를 주어야 한다. 본 논문에서는 <그림 4>와 같이 임의로 33%를 취하여 각 표본 데이터에 클래스 정보를 주었다. 0으로 표시된 것은 클래스 정보를 주지 않은 데이터이다.

3.2 개체군 생성

한 차원의 가중치를 표현하기 위한 각각의 염색체는 진화 연산자를 적용하기 쉽게 하기 위해 0, 1로 이루어진 m개의 2진 코드로 표현하였다. d차원인 경우 염색체의 길이는 m×d이다.

예를 들어 10110100의 8비트의 문자열은 0.7의 가중치로 표현할 수 있다. m의 길이가 0과 1사이의 간격을 몇 등분으로 나눌 것인가를 결정한다. 예를 들어 m=4일 경우 표시될 수 있는 가중치의 값은 0000 (0)부터 1111(15(2^4-1)까지 모두 16개이며, m=8일 경우 256개로 나누어진다.



<그림 5> '방문회수'를 기준으로 사용자 정보를 준 산점도

3.3. 적합도 함 [5][6]

본 논문에서 적합도 함수가 의미하는 것은 주어진 가중치(개체) 아래에서 클러스터링의 결과가 사용자가 준 정보대로 잘 되었는지를 나타내는 것이다. 이를 위해 본 논문에서는 클러스터링 알고리즘으로 K-means알고리즘을 사용하였으며, 또한 두 개의 자료의 독립성 검정을 위해 x^2함수를 변형한 Cramer's V2을 적합도 함수로 사용한다.

자료를 둘 이상의 범주에 따라 분류하였을 때 범주들의 상호 독립성(혹은 상호연관성) 여부를 검정하는 통계적인 방법으로 x^2-독립성 검정이 있다.

기본적으로 x^2-독립성 검정은 두 범주의 관찰된 빈도수와 그와 대응되는 기대값을 비교하는 것을 기본으로 한다. 관찰된 빈도수가 기대값에 가까울수록 두 범주가 서로 독립적이라고 말할 수 있다.

그러나, x^2의 값은 데이터의 수에 비례하여 증가하므로 x^2값 자체만으로는 두 범주가 얼마나 연관성이 있는지를 알 수 없다.

Harald Cramer는 x^2값이 총 빈도수(N)와 속성 값의 최소값의 곱에 비례하여 증가하며, 두 범주의 독립성 여부를 0과 100사이의 값으로 나타낼 수 있는 새로운 함수 Cramer's V^2을 (식 3)과 같이 정의하였다.

$$V^2(A,B) = \frac{x^2(A,B)}{N \cdot \{\min(n,m)-1\}} \times 100 \qquad \text{(식 3)}$$

Cramer's V^2의 값은 두 개의 범주가 상호 종속되어 있으며 100에 가까운 값을, 아무런 연관성이 없이 독립적이면 0에 가까운 값을 갖게 된다. 따라서 범주 A를 '(사용자가 준) 클래스 정보'로 범주 B를 '클러스터링의 결과'로 보았을 때, 클러스터링 결과(B)가 클래스 정보(A)대로 되었을 경우 두 범주는 종속적인 관계에 있다고 볼 수 있으므로 100에 가까운 값을 갖게되며, 클러스터링 결과가 사용자가 준 정보대로 되지 않았을 경우 두 범주는 서로 독립적이라고 볼 수 있으므로 0에 가까운 값을 갖게 된다.

3.4 재생산

다음 세대의 개체집단을 선택하기 위한 방법으로 룰렛휠(roulette wheel) 방법과 엘리트(elitist) 보존 방법을 사용한다.

4. 실험 및 분석

4.1 벤치마크 데이터(:클래스 정보가 있는 경우)

4.3.1 아이리스 데이터(Iris Data)

아이리스 데이터로 실험한 결과 <표 2>와 같다. 각 속성의 가중치는 주어진 클래스 정보대로 클러스터링 할 경우 가장 클러스터링이 잘 되는 경우의 가중치를 나타낸 것이다. 결과의 가중치를 적용할 때, 가중치를 적용하지 않은 경우(모든 속성의 가중치가 동일할 경우)보다 적합도 함수 값이 높게 나타남을 볼 수 있다. <표 3>에서는 [][]에서 제안한 휴리스틱 방법의 실험결과와 비교하여 보여주고 있다.

<표 2> 아이리스 데이터의 실험결과

속 성	1(SL)	2(SW)	3(PL)	4(PW)
가 중 치	0.0705	0.0156	0.8705	0.9294
적합도 값	92.38			
가중치를 고려하지 않을 경우의 적합도 값	72.89			

<표 3> 아이리스 데이터의 수행결과 비교

속성	Neuro-Fuzzy [2]		Fuzzy-GA [3]		제안된 알고리즘	
	가중치	순위	가중치	순위	가중치	순위
SL	0.0584	4	0.2235	3	0.0705	3
SW	0.1944	3	0.0776	4	0.0156	4
PL	0.9656	1	0.9741	1	0.8705	2
PW	0.6035	2	0.9278	2	0.9294	1

4.2 전자 상거래 실험 데이터(:클래스 정보가 없는 경우)

클래스 정보가 없는 데이터에 대하여는 1. 서론 에서 예로든 전자상거래 상의 데이터와 같이 사용자가 임의로 클래스 정보를 부여한다. 이 데이터의 속성 중 '판매금액'과 '방문회수'의 두 가지 속성만을 이용하여 적합도 함수가 어떻게 평가되는지를 살펴보자.

① 만약 개체군에 의해 표현된 가중치가 (판매금액 : 1 방문회수 : 1)로 주어졌을 경우 클러스터링의 결과는 아래의 <그림 1>와 같다.

이 클러스터링 결과와 사용자가 준 클래스 정보를 분할표로 만들면 <표 4>과 같으며 이때 적합도 함수(Cramer's V^2)값은 3.6으로 아주 낮은 값을 나타내며 클러스터링의 결과와 사용자 정보가 연관성이 없음(독립)을 나타낸다.

<표 4> 클러스터링 결과에 대한 분할표 1

		클러스터링 결과			
		A	B	C	total
사용자가 준 정보	class 1	3	1	1	5
	class 2	2	1	2	5
	class 3	4	2	1	7
	total	9	4	4	17

② 만약 개체군에 의해 표현된 가중치가 (판매금액 : 0.000121, 방문회수 : 0.999878)로 주어졌을 경우 클러스터링의 결과는 아래의 <그림 2>와 같다. 이 클러스터링 결과와 사용자가 준 클래스 정보를 분할표로 만들면 <표 5>과 같으며 이때 적합도 함수(Cramer's V^2)값은 68.3으로 아주 높은 값을 나타내며 클러스터링의 결과와 사용자 정보가 연관성이 높음을 나타낸다. 즉, 사용자가 준 정보대로 클러스터링이 되도록 가중치가 설정되었다고 볼 수 있다.

<표 5> 클러스터링 결과에 대한 분할표 2

		클러스터링 결과			
		A	B	C	total
사용자가 준 정보	class 1	5	0	0	5
	class 2	1	3	1	5
	class 3	0	0	7	7
	total	6	3	8	17

<표 6>은 4가지 속성을 모두 고려하여 사용자 정보에 맞도록 클러스터링 했을 경우 각 속성의 가중치를 나타내고 있다.

<표 6> 실험 데이터의 결과

속성	나이	성별	판매금액	방문회수
가중치	0.060041	0.320911	0	0.619048
적합도 함수 값	88.571			

5. 결론 및 향후 연구방향

본 논문에서는 대용량의 데이터를 효율적으로 처리하기 위해 사용되는 마이닝을 위한 속성가중치 방법을 제안하였다. 클래스 정보를 알고 있는 데이터에 대해서는 많은 알고리즘들이 연구되어 사용되고 있으며, 클래스 정보가 주어지지 않은 데이터에 대해서 기존의 알고리즘들은 데이터 간의 관계를 고려한 비교사 학습에 의해 중요한 속성을 찾아내었으나, 이러한 비교사 학습은 사용자가 제대로 학습이 되었는지에 대한 여부를 판단하기가 어려운 단점이 있다.

따라서 클래스 정보가 없는 데이터에 대해서는 사용자가 어느 정도의 클래스 정보를 부여하여 그 정보대로 학습이 되도록 함으로서 사용자가 이해할 수 있는 결과를 얻도록 각 속성에 가중치를 부여하는 방법을 제안하였다. 가장 최적화 된 가중치를 찾아주기 위해서 전역적 탐색 방법인 유전자 알고리즘을 사용하였으며, 클래스 정보대로 잘 학습했는지를 알아보기 위한 적합도 함수로 Cramer's V^2 을 이용하였다. 위의 알고리즘의 타당성을 검증하기 위해 전자상거래상의 데이터와 벤치마크 데이터를 사용하여 실험하였다.

앞으로 수치적인 데이터외에 심볼릭 데이터에도 적용할 수 있게 확장할 계획이며, 또한 본 알고리즘에 의해 형성된 가중치를 이용하여 다른 마이닝 기법에 적용 할 계획이다.

참고문헌

[1] Maria J. Martin-Bautista, Maria-Amparo Vila, "A Survey of Genetic Feature Selection in Mining Issues " Proceedings of the 1999 Congress on Evolutionary Computation - Volume 2 ; V.2 ; PP:1314-1321 ; 1999.

[2] Sankar K. Pal. "Unsupervised Feature Evaluation : A Neuro-Fuzzy Approach" IEEE Transaction on Neural Networks, Vol. 11(2) pp 366-376, 2000.

[3] Frank Chung_Hoon Rhee, Y. J. Lee "Unsupervised Feature Selection using a Fuzzy-Genetic Algorithm" IEEE International Fuzzy Systems Conference Proceedings III-1266 - 1269, 1999.

[4] Dallas E. Hohnson, "Applied Multivariate Methods for Data Analysts", Duxbury Press, 1998

[5] B. Liu., W. Hsu and Y. Ma, "Pruning and Summarizing the Discovered Association", In Proceedings of the 5th KDD Conference, pp125-134, San Diego, CA USA, 1999.

[6] Harald Cramer, "Mathematical Methods of Statistics". Princeton University Press, 1973.