제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

희귀 목적값 분류를 위한 학습 알고리즘

이광호, 이창환
동국대학교 정보통신공학과
e-mail : keyss798@dongguk.edu, chlee@dgu.edu

A New Learning Algorithm for Rare Class Classification

Kwang-Ho Lee, Chang-Hwan Lee
Dept of Information and Communications Engineering,
Dong-Guk University, Seoul, Korea

요 약

본 논문에서는 데이터 마이닝에서 발생되는 희귀 데이터를 분석하기 위한 희귀 목적값 분석의 새로운 알고리즘을 제시한다. 이를 위하여 속성들이 가지는 속성의 가중치 값과 속성값이 목적 속성에 미치는 가중치값을 정보이론에 입각하여 가중치 계산을 하고, 계산된 가중치값을 사용하여 스코어링 함으로써 희귀 목적값에 속한 데이터 예측/분류에 사용하는 방법을 제시하였다. 실험을 통해 본 알고리즘의 성능을 입증함은 물론 제안된 알고리즘이 희귀 데이터의 분류/학습에 좀 더 효과적이다는 것을 보였다.

1. 서론

데이터 마이닝은 기존의 데이터 집합에서 우리가 알고 있는 일반적인 방법으로는 식별할 수 없는 새로운 정보를 추출 하고자 하는 학문이다. 이러한 정보 추출을 통해 새로운 데이터의 예측과 탐색을 할 수 있고, 보다 정확한 데이터의 분류를 할 수 있다. 하지만, 이러한 데이터 마이닝에서 중요하게 다루어지는 문제 중 하나가 바로 희귀 사건(rare events)의 발생으로 인한 탈선 탐색(deviation detection)이다. 탈선 탐색은 일반적이지 않는 아주 소수의 데이터 발생을 탐색하는 것을 말하는데, 이러한 데이터는 총 발생되는 데이터 양에 비해 아주 극소수의 데이터이기 때문에 예측/분류가 어려워지는 것이 문제점이라 할 수 있다. 이러한 희귀 데이터의 예로는, 네트워크 지연문제, 보안관련 사고, 갑작스런 심장질환의 발생, 신용카드나 그 밖의 금융사고 그리고 교통사고 등이라 할 수 있다. 보는 바와 같이 이러한 데이터들의 공통적인 특징은 관련 데이터들이 너무나도 많이 발생하지만, 정작 이러한 사고를 발생시키는 요인은 그 방대한 양의 데이터 중에서 아주 극소수의 데이터만이 그 정보를 가지고 그 원인을 규명해 줄 수 있다는 것이다. 따라서 일반적인 데이터 마이닝 알고리즘으로는 정확한 예측/분류가 힘들 뿐 아니라 그 정확도가 현저하게 떨어지는 것이 사실이다. 그 이유는 앞에서 말한 너무 방대한 일반 데이터(normal data)의 영향보다 희귀 데이터(rare data)의 영향을 더 많이 받으므로, 일반적인 데이터 마이닝 알고리즘으로는 검출할 수 없을 뿐만 아니라 방대한 양의 일반 데이터에 희귀 데이터는 그 중요성을 인식받지 못하고 묻혀버리게 된다. 이러한 희귀 데이터들을 희귀 목적값(rare class)이라 하며, 이러한 데이터를 예측/분류하는 방식으로는 탈선 탐색(deviation detection), 분리물 인식(outlier analysis), 예외탐색(anomaly detection), 예외마이닝(exception mining)등 매우 다양하게 불려지고 있으며, 이러한 방식들을 포괄적으로 희귀 목적값 분류(rare class analysis)라는 용어를 사용하여 나타내고 있다 [2].

최근까지 알려진 희귀 목적값 분류에 포함되는 알고리즘에는 Boosting Algorithm [1]과 Two-Phase Rule Induction [2]이 존재하고 있다. 하지만 아직까지 이 알고리즘들이 완벽하게 희귀 목적값의 데이터들을 완벽하게 예측/분류하기에는 아직도 많은 시간과 발전을 필요로 한다.

본 논문에서는 이러한 희귀 목적값 분류에 포함될 수 있는 속성값 가중치 계산을 통한 새로운 희귀 목적값 분류 학습을 위한 알고리즘을 제시하고자 한다.

2. 관련연구

희귀 목적값 분류에 대한 연구는 아직까지 많은 연구가 진행되지 않은 상태이므로 미지의 개척지라 할 수 있다. 최근 희귀 목적값 분류에 대한 해결책

으로 제시된 알고리즘은 크게 두 가지로 양분된다. 첫 번째 알고리즘은 바로 Boosting Algorithm 이라 한다. 이 알고리즘은 전혀 새로운 알고리즘이라고는 할 수 없으며, 기존의 알고리즘에 추가로 Boosting Algorithm 이라는 개념을 추가함으로써, 회귀 목적값에 대한 보다 정확한 예측/분류를 할 수 있다는 내용이다. 이 알고리즘의 내용은 회귀 목적값에 대해 취약한 대부분의 알고리즘에 선택(voting)의 효과와 가중치 수정(weight update)효과를 사용하여 기존의 현저하게 떨어진 회귀 목적값에 대한 예측/분류의 정확도를 올릴 수 있다는 내용이다. 즉, 기존의 알고리즘으로는 회귀 목적값의 예측/분류 정확도가 낮으므로, 그 정확도의 증가를 위해서 계속적으로 반복적 수행을 하게되며, 이때 선택과 가중치 수정을 수행해줌으로 인해서 좀 전의 정확도를 향상시킬 수 있게 된다. 여기에서 선택이란 좀 전의 훈련 데이터의 정확도가 낮게 나왔으므로, 정확도를 낮추게 만들었던 데이터에 대해서는 낮은 가중치를 부여하고 정확도를 낮출 거라고 생각되어 무시했던 데이터에 대해서는 높은 가중치를 부여하도록 선택하는 것을 뜻한다. 여기에서 선택되어지는 것에 가중치 수정을 수행하게 되므로 결국, 마지막에는 보다 높은 정확도를 나타낼 수 있다는 것이다.

두 번째로 살펴볼 회귀 목적값 분석 중 가장 많이 알려진 알고리즘인 Two-Phase Rule Induction에 대해서 살펴보자. 이 알고리즘은 회귀 목적값 분석에 대한 또 다른 해결책으로 제시된 알고리즘으로 기존의 PNrule [3]을 변형시킨 알고리즘이라 할 수 있다. 이 알고리즘은 올바른 예측과 틀린 예측에 대한 모든 상황을 두 개의 상황으로 분리해서 완벽한 모델 생성을 위해 다시 연결하게 된다. 훈련 데이터에 대한 정확도를 계산하여 두 상황의 연결을 계속 조합하여 만들어지는 모델이라 할 수 있다. 두 개의 각기 다른 상황에서 모아진 모델들을 보다 높은 정확도를 위해 재호출(recall)과 함께 스코어링을 하게 되면서 서로 조합하게 되면 회귀 목적값에 대한 정확도를 높일 수 있게 된다. 이러한 Two-Phase Rule Induction 알고리즘 또한 바로 전의 Boosting Algorithm과 마찬가지로 반복적으로 정확도를 높이기 위한 수행을 하여야 한다.

2. 알고리즘의 내용

본 알고리즘의 수행을 위해서는 크게 두가지 단계로 구분된다.

1. 속성값 가중치 계산
2. 테스트 데이터 스코어링

2.1 속성값 가중치 계산

본 논문에서 제안된 속성의 가중치를 계산하는 방법은 다음의 세 가설에 바탕을 두고 있다: (1) 속성의 특정한 값이 정해지면 이는 목적 속성에 정보를 제공한다. (2) 제공되는 정보의 양은 엔트로피 함수에 의하여 정의될 수 있다. (3) 속성이 제공하는 정보의 양이 많을수록 엔트로피 함수의 값은 커진다. 이분 속성, 카테고리 속성과 같은 이산형(discrete)의 속성에서 속성에 대한 가중치의 계산을 위해서는 먼저 속성의 각 이산형 값에 대한 정보량을 계산한 후 평균값을 해당 속성의 가중치로 사용한다. 연속 속성의 경우에는 데이터를 몇 개의 범위로 분할하여 주는 이산화과정(discretization)을 수행 후 그 결과를 가지고 위의 방법을 적용하는 것이다.

우선 본 알고리즘에서 사용되어지는 속성값에 대한 가중치를 계산하기 이전에 속성에 대한 가중치 계산을 필요로 한다. 바로 '속성의 특정 값이 목적 속성에 제공하는 정보의 양을 어떤 방법으로 측정할 것인가'를 알아야 한다. 이 글에서는 정보의 양을 측정하는 엔트로피 함수로서 Hellinger 변량(divergence)을 사용하는데 이는 Beran[5, 6]에 의해 제안된 후 여러 분야에서 사용되고 있다.

목적 속성을 T라고 하고, '특정 속성 A의 값이 a가 됨'을 '$A=a$'로 표현한다. t_i를 T값 중의 하나로 가정하고, $p(t_i)$와 $p(t_i|a)$를 목적 속성 T의 사전 확률 및 사후 확률로 가정한다. $A=a$가 T에게 제공하는 정보의 양을 $H(T|A=a)$로 표시할 때 그 변량은 다음과 같이 정의된다.

$$H(T|A=a)=\left[\sum_i\left(\sqrt{p(t_i)}-\sqrt{p(t_i|A=a)}\right)^2\right]^{\frac{1}{2}}$$

이 변량은 목적 속성 T의 사전 확률분포와 사후 확률분포간의 차이를 측정한 함수이다. 이 변량의 특성에 대하여 조사하여 보면 우선 모든 경우의 $p(t_i)$와 $p(t_i|a)$ 값에 대하여 그 값이 정의가능(definable)하고 연속(continuous)이다.

또한 위의 Hellinger 변량은 사전 확률분포와 사후 확률분포가 일치할 때만 값이 0이 되며 나머지 경우는 항상 0과 1사이의 값을 가진다. 특정 속성 A의 가중치 계산을 위해 A가 가지는 모든 값에 대하여 위에 정의된 Hellinger 변량 값을 구하고 그 합을 A의 가중치로 사용할 수 있을 것이다. 하지만 이 경우 속성이 가지는 값의 개수가 증가하게 되면 변량의 값도 증가하게 된다. 본 연구에서는 위 변량에 각 속성 값의 발생확률 $P(a)$를 곱해 해결하였다.

$$H(T|A)=\sum_a p(a)\cdot H(T|A=a)$$

따라서 $H(T|A)$의 값은 속성의 값의 개수에 영향을 받지 않게 된다. 끝으로 위에서 정의된 가중치의 범위를 0과 1사이로 제한하기 위하여 모든 속성의 가중치 값의 합에 대한 비율로써 표현하였다. 최

종적인 속성 가중치 값의 식은 다음과 같이 표현된다.

$$\omega_T(A) = \frac{H(T|A)}{\sum_A H(T|A)} = \frac{\sum_a p(a)H(T|A=a)}{\sum_A H(T|A)}$$

위의 속성 가중치식을 이용하여 속성값 가중치를 계산하고자 한다. 위에서 정의된 수식을 이용, 특정 속성 A가 갖는 속성값 a에 대한 가중치 $J(T|A=a)$는 다음과 같이 표현할 수 있다.

$$J(T|A=a) = H(T|A=a) * \frac{\omega_T(A)}{\sum_{a_i \in A} H(T|A=a_i)}$$

2.2 테스트 데이터 스코어링

2.1 과정이 속성값이 목적 속성에 미치는 정도를 계산하기 위한 과정이라면, 테스트 데이터 스코어링는 훈련 데이터에서 얻어진 가중치값을 사용 테스트 데이터의 실질적인 희귀 데이터와 일반 데이터의 구분을 위한 부분으로, 모든 속성값에 대한 가중치가 계산되면 하나의 데이터가 가지는 모든 속성값의 가중치 합을 구할 수 있게 된다. 그 가중치의 합으로 본 알고리즘에서는 희귀 데이터와 그렇지 않은 일반 데이터로 분류할 수 있게 된다. 그 이유는 본 가중치의 값이 가지는 의미가 목적 속성에 얼마만큼의 영향을 미치느냐의 척도를 나타내기 때문이다. 즉, 희귀 데이터 일수록 가중치가 높은 속성값을 가지는 속성값들을 일반 데이터보다 더 많이 갖고 있다고 보는 것이다. 그렇기 때문에, 해당하는 데이터의 가중치의 합이 높으면 높을수록 희귀 데이터에 더 가깝다고 판단할 수 있게 된다.

모든 데이터에 대한 가중치의 합이 구해지면 희귀 데이터에 대한 가중치 값 영역과 일반 데이터에 대한 가중치 값 영역으로 나뉘지게 된다. 그렇게 나누어진 영역에 데이터 분류를 위한 중앙값(middle value)을 구할 수 있다. 중앙값은 데이터 예측/분류를 위한 기준점으로 사용되어 진다. 즉, 중앙값보다 높은 값을 가지는 데이터는 희귀 데이터로 분류하고, 낮은 값은 일반 데이터로 분류하게 된다.

3. 제안된 알고리즘의 특징

관련연구에서 제안된 Boosting Algorithm과 본 논문에서 제안하고자 하는 속성값 가중치 스코어링 알고리즘을 비교해보도록 하자. 간단하게 위에서 설명한 바와 같이 Boosting Algorithm은 전혀 새로운 알고리즘이 아닌 기존의 알고리즘에 모두 사용할 수 있는 장점과 정확도를 높일 수 있다는 장점을 가진다. 그러나 알고리즘이 하는 역할은 기존의 알고리즘을 다시 선택적으로 데이터를 호출하여 가중치를 부여하여 다시 정확도를 계산하는 방식이므로, 본래 알고리즘의 성능이 높지 않다면 그 성능은 장담할 수 없게 된다 [4]. 또, 정확도를 높이기 위해 반복적으로 수행되어져야 하므로 방대한 양의 데이터를 계속해서 학습하여야하는 수행시간 또한 이 알고리즘의 단점이라 할 수 있겠다.

하지만 본 알고리즘에서는 훈련 데이터를 한번만 읽어서 데이터가 가지는 속성값들이 목적 속성에 미치는 가중치를 계산하게 되므로 훈련 데이터를 한번만 사용하게 되어, 몇 번씩 데이터를 읽어야 하는 Boosting Algorithm보다 시간복잡도에서 훨씬 빠르다. 즉, 다음과 같이 본 알고리즘의 특징을 정의할 수 있다.

● 정리 1 :
학습 데이터의 개수를 n이라할 때, 본 알고르즘의 시간복잡도는 $O(n)$이다.

◎ 증명 1 :
지면의 제약으로 생략함.

데이터 마이닝 알고리즘 중에서 훈련 데이터를 한번도 사용하지 않고 예측/분류가 가능한 알고리즘은 없기 때문에 최상의 시간복잡도라 할 수 있으며, 이 특징이 기존의 희귀 목적값 분류의 알고리즘과 가장 두드러진 차이점이라 할 수 있다.

4. 실험 결과

본 알고리즘은 C언어로 구현하였으며, 성능 실험을 위해 임의의 가상 시나리오를 정하고 데이터를 생성하였다. 데이터 생성 시 좀 더 실제 데이터와 비슷하고자 몇 개의 속성값이 일치하는 데이터를 각각 다른 목적 속성을 가진 데이터로 생성하였다. 또, 모든 데이터는 이산화 된 데이터로 생성 실험하였다. 속성값의 수는 모든 속성들마다 랜덤하게 적용하였으며, 목적 속성은 이분 속성으로 데이터를 생성하였다. 본 알고리즘의 테스트는 훈련 집합/테스트 집합의 방법을 이용하며, 임의의 데이터를 생성하여 70%를 훈련 집합으로 나머지는 테스트 집합으로 분류 실험하였다. 테스트 집합의 모든 개체에 대하여 예측결과를 실제 결과와 비교 그 정확도를 계산하였다. 마지막으로 정확도에 대한 편차를 줄이기 위해 훈련 집합과 테스트 집합의 분할을 10회 반복 수행하였다.

그림 1은 희귀 목적값 데이터의 발생 빈도에 따른 전체 정확도를 보여주고 있다. 그림에서도 알 수 있듯이 발생되는 희귀 데이터가 전체 데이터에서 그 비중이 적을수록 더 높은 정확도를 나타내고, 그 반대로 전체 데이터 중 희귀 데이터의 비율이 증가 할수록 성능이 떨어짐을 알 수 있다.

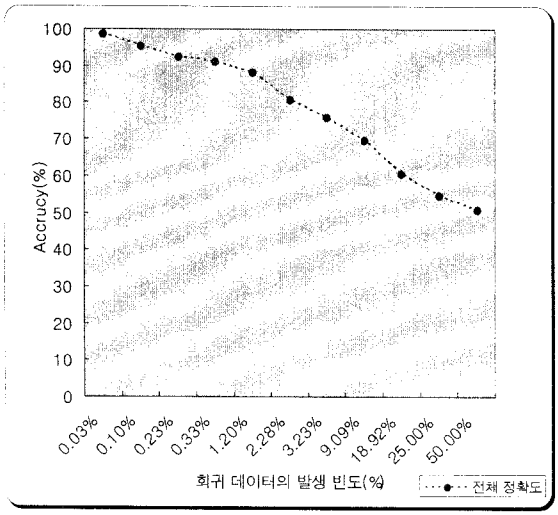

그림 1에서의 회귀 데이터 발생 빈도에 따른 테이트 데이터 스코어링 과정을 그림 2와 3에서 보여주고 있다. 그림에서 나타내고 있는 일반 데이터(normal data, 이하 N/D) 가중치 값의 영역은 점선으로, 희귀 데이터(rare data, 이하 R/D)는 실선으로 나타내고 있다.

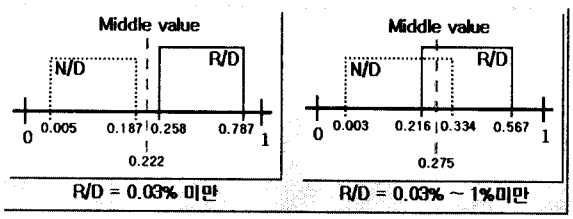

그림 2 : R/D 1% 미만에서의 스코어링

그림 2에서 보는 바와 같이, 0.03%미만의 발생 빈도에서 두 영역의 겹치는 부분이 없어 완벽하게 R/D와 N/D를 구분하고 있다. 또한 1%미만에서도 두 영역의 겹치는 부분이 적어 성능이 높게 나오고 있음을 확인할 수 있다.

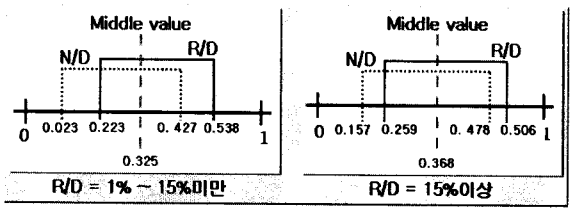

그림 3 : R/D 1% 이상에서의 스코어링

그러나, 1% 이상에서부터 두 영역의 겹치는 부분이 많아져 알고리즘의 전체 성능이 떨어짐을 알 수 있다. 그 이유는, N/D 가중치 값 영역이 점점 더 높은 값을 가지게 되어, 중앙값이 더 높아 지고, 또 중앙값보다 더 높은 값을 가진 N/D 데이터의 수가 많아지는데 있다. 이렇게 N/D 가중치 값이 높아지는 이유는 희귀 데이터를 발생시키는 특정 속성값에 높은 가중치를 주도록 되어 있는 본 알고리즘에서 희귀 데이터가 많아지므로, 제대로 된 가중치 계산을 할 수 없게 되었기 때문이다.

본 실험을 통해서 알 수 있듯이 논문에서 제안하고자하는 알고리즘이 일반 데이터 분석보다는 희귀 목적값 분석에 더 적합하다는 것을 알 수 있다.

5. 결 론

본 논문에서는 정보이론을 바탕으로 한 엔트로피 함수를 사용하여 속성에 대한 가중치를 계산하고, 속성값에 대한 가중치값을 계산하였다. 그리고 계산된 가중치값으로 테스트 집합을 스코어링 하였다. 제안된 알고리즘의 실험결과에서도 알 수 있듯이 일반 데이터에서 발생되어지는 희귀 데이터의 비율이 낮으면 낮을수록 기존의 알고리즘에서는 보여주지 못하는 높은 정확도를 보이고 있다는 것이 증명되었다. 이처럼 제안하는 알고리즘이 일반 데이터와는 다른 희귀 데이터에서의 분석에 더 적합하다는 것을 알 수 있다. 따라서, 희귀 목적값 데이터에 새로운 분류/학습 알고리즘으로 평가되어 질 수 있을 것이다.

앞으로의 과제는 실생활 희귀 데이터에서의 실험을 통해 본 논문에서 제안된 알고리즘이 실생활 데이터에서도 높은 분류/학습을 할 수 있다는 것을 증명하는 것이다.

참고문헌

[1] M. Joshi, R. Agarwal and V. Kumar, "Predicting Rare Classes : Can Boosting Make Any Weak Learner Strong?", ACM SIGKDD

[2] M. Joshi, R. Agarwal and V. Kumar, "Mining Needles in a Haystack : Classifying Rare Classes via Two-Phase Rule Induction", ACM SIGMOD

[3] R. Agarwal and M. Joshi, "PNrule: A new framework for learning classifier models in data mining(a case-study in network intrusion detection).", SIAM

[4] M. Joshi, V. Kumar, R. Agawal, "Evaluatiog Boosting Algorithms to Classify Rare Classes", IEEE

[5] R. J. Beran, "Minimum Hellinger Distance for Parametric Models", Ann. Statistics, Vol. 5, pp.445-463, 1977.

[6] Z. Ying, "Minimum Hellinger Distance Estimation for Censored Data", Annals of Statistics, VoL. 20, No 3, pp. 1361-1390, 1992