SVDD 기반 가중치를 이용한 패턴 추출 방법

윤태복*, 이지형**
*서일대학 컴퓨터 소프트웨어과
**성균관대학교 정보통신공학부
e-mail:tbyoon@seoil.ac.kr

Pattern extraction method using SVDD-based weighted

Taebok Yoon* and Jee-Hyong Lee**
*Dept. of Computer Software, Seoil University
**School of ICE, Sungkyunkwan University

요 약

데이터 마이닝은 주어진 데이터로부터 의미 있는 정보를 찾기 위한 방법으로 주로 사용된다. 하지만, 분석을 위한 데이터에 의미 없는 정보가 포함되어 있다면 분석 결과를 신뢰 할 수 없을 것이다. 이를 위해서 의미 없는 데이터를 제거하기 위한 연구 사례가 있으나, 정상적인 데이터도 함께 제거될 수 있다는 단점이 있다. 본 논문은 패턴 추출을 위한 분석 데이터를 SVDD 방법을 이용하여 의미 있는 데이터와 의미 없는 데이터 간에 가중치를 구한다. 생성된 가중치는 의사결정나무 생성에 반영하였고, 실험을 통하여 유효성을 확인하였다.

1. 서론

정보통신기술의 발달과 함께 환경에서 수집된 데이터는 대량성, 불완전성, 패턴의 변화 등의 특성을 가지고 있으며, 이런 데이터들의 특성을 고려한 고급화된 분석 기법이 요구되고 있다. 특히, 현실 세계 데이터의 패턴은 시간의 흐름에 따라 변화하고, 알지 못하는 데이터 즉 이상치(Outlier)를 포함하고 있다[1]. 특히, 수집 데이터를 기반으로 미래를 예측하는 기존의 분석 방법은, 수집된 데이터의 유효성 문제로 인하여 분석의 어려움이 많고, 낮은 예측 결과를 보여 준다. 즉, 분석에 사용되는 데이터가 일관되지 않거나 비정상적인 형태를 가진다면 분석 결과의 신뢰도는 매우 낮을 것이다. 이와 같은 문제를 해결하기 위해 기존에는 패턴 추출 작업 이전에 비정상 데이터를 선별하여 제거하는 연구가 주로 실시되었다[2]. 하지만, 데이터의 특성, 환경의 조건, 수집 데이터의 상대성 등을 고려하지 않고 제거하여 의미 있고 유효한 데이터도 함께 제거 될 수 있다는 단점이 있다. 본 논문은 Support vector data description(SVDD) 방법을 이용하여 의미 있는 데이터와 의미 없는 데이터를 구분하여 가중치를 계산한다. 데이터가 유효성이 낮다고 할지라도 분석에서 제외하는 것이 아니라 낮은 중요도만큼의 의미를 가지며, 분석에 반영되도록 하였다. 실험에서는 기존의 전통적인 분석 방법과 비교하여 제안하는 방법의 유효성을 확인하였다.

2. 관련연구

데이터 분석에서 수집된 데이터에서 유효성을 판단하기 위한 연구로는 다음과 같다. Hwang과 Hahn은 데이터 수집과정의 오류나 결함으로 인하여 부족한 부분 발생하거나 분석자의 의도와 다른 기준으로 수집된 데이터를 불완전한 데이터라 정의하고, k-Means 클러스터링 방법을 이용하여 데이터의 결함을 예측하고 분석자의 의도에 맞게 재분류하는 연구를 하였다[3]. Dick은 이메일 스팸 제거 기술에 불완전 데이터 분류를 위해 SVM을 이용하여 유용한 결과를 얻었다[4]. 또한, Xi는 이상치를 전통적인 이상치와 공간에 기반을 둔 이상치로 구분하고 전통적인 이상치를 확률기반, 거리기반, 표준편차기반, 밀도기반으로 분류하였다[2]. 또한, 공간에 기반을 둔 이상치 방법에는 그래프 기반, 공간 기반으로 분류하여 소개하였다. 그리고 최근 고차원 기반(high dimension-based)에서의 이슈와 SVM 방법을 이용한 이상치 감소 방법을 소개하였다. Kim 등은 지능형 이러닝 시스템에서 수집된 학습자의 학습 행위 데이터에서 이상치 데이터를 감소시키기 위한 연구를 수행하였다. 이 논문에서는 군집화 방법을 이용하여 이상치를 감소시키고 학습자 모델의 성능이 향상된 것을 실험을 통하여 확인하였다.

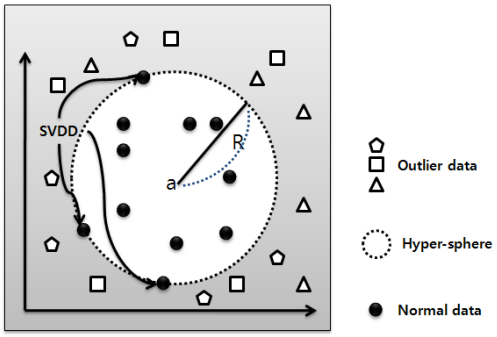
3. SVDD기반 가중치를 이용한 패턴 추출 방법

본 논문에서 이용한 가중치 생성 방법은 SVDD에 기반을 두고 있다. 주어진 데이터를 공간상에 사상시키고, SVDD 기법을 통하여 Hypersphere를 찾는다. 생성된 구의 중심으로부터 각 개체간의 분포를 이용하여 가중치로 활용한다.

3.1 SVDD 기반 가중치

SVDD는 비선형 SVM(Support Vector Machine)의 응용 형태로 단일 클래스 분류(One-class classification)를

위한 대표적인 방법으로 이상치를 분류하기 위한 기법으로 빈번하게 사용된다. SVDD는 Hypersphere를 찾아 데이터를 맵핑하는데, 이 때 최적화된 Hypersphere를 찾는 것이 중요하고, 반지름을 최소화하면서 많은 맵핑 데이터를 둘러싸는 것을 목표이다. SVDD는 중심이 a이고 반경이 R인 구를 찾는 것이 중요한 관건인데, 여기서 집합 D를 가지는 구가 있다고 가정할 때, 이 구는 가능한 크기가 작아야 하고, 가능한 많은 학습 데이터를 포함해야 한다.



(그림 1) SVDD 기본 개념

최적을 구(Hypersphere)를 찾기 위해 많은 과정을 거치며 구체적인 방법은 관련논문을 참고하기 바란다[5,6]. 본 논문은 SVDD 방법을 통하여 얻은 중심점 a로부터 각 데이터들과의 거리를 가중치 값, *svddWeight(i)*라고 정의하고 이를 분석에 활용한다.

3.2 SVDD 가중치를 이용한 의사결정 나무 생성

사례별 가중치를 이용한 의사결정나무를 생성하기 위해서는 기존과 다른 엔트로피 방법을 이용하여야 한다. 데이터 분석에 앞서 소개한 사례별 가중치를 적용할 수 있어야 한다. 다음은 변형된 엔트로피 공식에 대한 설명이다. 여기서 *i*번째 사례가 갖는 가중치를 *svddWeight(i)*로 표시하고, 모든 *i*에 대하여 *svddWeight(i)=1*이라고 한다면, *Pc*는 다음과 같이 다시 나타낼 수 있다. *c(i)*는 *i*번째 사례가 속하는 클래스이다.

$$P_c = \frac{\sum_{c(i)=c} svddWeight(i)}{\sum_{i=1}^{n} svddWeight(i)}$$

기존의 의사결정나무 알고리즘은 모든 사례의 가중치 값을 모두 1로 계산한다. 하지만, 제안하는 방법의 경우 각 사례에 대하여 가중치(*svddWeight*)가 부여되어 있다. *svddWeight*를 의사결정나무 생성에 반영하기 위한 새로운 엔트로피를 산출하는 공식은 다음과 같다.

$$Entropy(s) = -\sum_{c=1}^{T_c} \frac{\sum_{c(i)=c} svddWeight(i)}{\sum_{i=1}^{n} svddWeight(i)} \log_2 \frac{\sum_{c(i)=c} svddWeight(i)}{\sum_{i=1}^{n} svddWeight(i)}$$

4. 실험 및 검증

실험을 위해 UCI Repository에서 제공하는 Breast Cancer Wisconsin (Diagnostic) Data Set를 이용하였다. Breast Cancer Wisconsin (Diagnostic) Data Set은 유방암 진단을 위한 세침 흡인 세포 검사(Fine Needle Aspiration : FNA) 데이터의 집합이다. 전체 데이터의 개수는 569개이며, 이중에서 200개는 양성 종양(benign), 200개는 악성 종양(malignant)을 임의 선별하였다. 비교를 위해 전통적인 기계 학습 방법과 제안하는 방법을 실시하였다. 전통적인 기계 학습 방법으로 의사결정나무 방법을 이용하였고, 모두 RPP(Rule-post pruning) 방법을 이용하여 가지치기를 실시하였다. 학습(Training)데이터의 개수는 200개, 검증(Test) 데이터는 200개로 하였으며, 교차검증방법을 이용하여 5회 실시하였다.

<표 1> 두 실험 환경의 에러율 비교 (%)

	original		proposal method	
	N/A	pruned	N/A	pruned
1	7.0	4.5	7.5	6.3
2	10.0	9.0	8.0	6.9
3	7.5	4.5	7.5	7.2
4	9.0	6.4	7.1	6.0
5	10.0	9.0	7.8	6.1
avg.	8.7	6.7	7.6	6.5

5. 결론 및 향후 연구

본 논문은 SVDD에 기반을 둔 가중치를 이용하여 의사결정나무를 생성하는 방법을 소개하였다. 이 방법은 수집 데이터의 의미 정도에 따라 분석에 반영하며, 데이터의 손실이 없다는 장점이 있다. 향후에는 다양한 실험 환경에 적용하여 유효성을 확인하는 연구가 필요하겠다.

참고문헌

[1] D. Ruan, G. Chen, E. Kerre, and G. Wets, “Intelligent Data Mining: Techniques and Applications”, Springer, 2005.
[2] J. Xi, “Outlier Detection Algorithms in Data Mining”, IEEE Second Int. Symp. on Intelligent IT Application, 2008.
[3] S. Y. Hwang, and H. E. Hahn, “Pre-Adjustment of Incomplete Group Variable via K-Means Clustering”, Journal of Korea Data & Information Science Society, Vol. 15, No. 3, 2004.
[4] U. Dick, P. Haider, and T. Scheffer, “Learning from Incomplete Data with Infinite Imputations”, Proceedings of the 25th International Conference on Machine Learning, 2008.
[5] D. M. J. Tax, and R. P. W. Duin, “Support Vector Data Description”, Machine Learning, Vol. 54, pp. 45-66, 2004.
[6] J. Y. Park, D. S. Kang, J. H. Kim, J. T. Kwok, and I. W. Tsang, “SVDD-Based Pattern Denoising”, Neural Computation, Vol. 19, pp.1919-1938, 2007.