제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

의사결정 트리 앙상블을 구축하기 위한 상관성 기반 기법을 이용한 속성 중복성 제거1)

박영준*, 박명호*, 손호선*, 류근호*
*충북대학교 데이터베이스 및 바이오인포메틱스 연구실
e-mail:{pyz, bluemhp, shon0621, khryu}@dblab.chungbuk.ac.kr

Removing the Feature Redundancy using Correlation-Based Approach for Decision Tree Ensemble

Yongjun Piao*, Minghao Piao*, Ho Sun Shon*, Keun Ho Ryu*
*Database/Bioinformatics Laboratory, Chungbuk National University, Korea

요 약

대량의 분류 규칙 탐사 과정은 앙상블기법을 사용하여 다양한 연구들이 이루어지고 있다. 본 논문에서는 의사결정 트리의 분열 문제와 singleton 포함 한계를 해결하기 위하여 Cascading-and-Sharing 앙상블 기법을 적용하여 점진적 다중 의사결정 트리를 구축하였다. 또한 분류의 정확도를 향상시키고, 트리의 복잡도와 모델 과잉접합을 피하기 위하여 다중 트리 구축과정에서 선형 상관분석기법을 기반으로 훈련 데이터 속성들의 중복성을 제거하였다. 실험 결과, 속성들의 중복성을 제거하여 구축한 트리들은 원래 기법보다 더 좋은 결과를 보여주었다.

1. 서론

의사결정 트리는 데이터를 분류하는 목적으로 인공지능, 기계학습, 패턴인식, 통계분석 등 많은 분야에서 이용되고 있는 분류기법이다. 의사결정 트리는 아래와 같은 이유로 많은 인기를 가지고 있다[1]. 첫째, 의사결정 트리 귀납은 분류 모델을 구축하기 위한 매개변수가 없는 접근법이다. 즉, 그 클래스에 의해 만족되는 확률 분포의 유형과 그 밖의 다른 속성들에 대한 어떤 전제도 필요하지 않다. 둘째, 의사결정 트리를 구축하기 위하여 개발된 알고리즘들은 계산비용이 크지 않기 때문에 방대한 훈련 데이터의 크기에도 모델을 빠르게 구축할 수 있다. 셋째, 의사결정 트리, 특히 작은 크기의 트리는 상대적으로 이해하기 쉬우며 대부분의 단순 데이터 집합들의 경우 트리의 정확성도 다른 분류 기법들과 비슷하다. 결정 트리 귀납 알고리즘의 하나로 헌트 알고리즘(Hunt's algorithm)이 있고, ID3[2], C4.5[3], C5.0[4] 등을 포함하여 수많은 의사결정 트리 귀납 알고리즘들의 기초가 되고 있다. 하지만 의사결정 트리는 부분 트리의 루트노드를 구성하는 훈련 데이터들이 점점 더 적어지는 분열(fragmentation)문제가 존재한다.

분류기의 정확성을 향상시키기 위하여 많은 앙상블 기법들이 개발 되었다. Bagging과 boosting은 다중 의사결정 트리를 구축하기 위한 앙상블 기법으로써 일정한 확률분포에 따라 본 데이터로부터 반복적으로 샘플링을 하고 각각의 부트스트랩(bootstrap) 샘플로 기본 분류기들을 생성한다. 하지만 [5]에서는 다양한 앙상블기법의 연구에 근거하여 아래와 같은 문제점들을 제시하였다. 첫째, boosting 방법으로 생성된 많은 앙상블들은 singleton이다. 둘째, 최상위로 랭킹 된 많은 속성들은 분류를 함에 있어서 비슷한 분류 능력을 가지고 있다. 이것은 다중 의사결정 트리를 구축함에 있어서 최상위로 랭킹 된 서로 다른 속성들이 루트노드가 될 수 있음을 의미한다.

속성선택(feature selection)은 패턴 분류 문제에서 분류기의 성능을 향상시킬 수 있는 중요한 기법이다. 특히, 많은 속성들을 가지는 데이터의 분류문제에서 관련이 적은 데이터(irrelevant feature), 중복된 데이터(redundant feature)를 제거한 속성의 부분집합을 선택하여 이용함으로써 분류기의 정확도를 향상시킬 수 있다. 또한 다중 트리를 구축함에 있어서 중복된 속성들은 의사결정 트리의 복잡도를 증가시켜 모델 과잉접합(overfitting)에 빠질 수 있다.

따라서 본 논문에서는 의사결정 트리의 분열문제와 singleton 포함 한계를 해결하기 위하여 cascading-and-sharing 앙상블기법을 이용하여 점진적 다중 의사결정 트리를 구축하였다. 또한 트리를 구축하는 과정에서 비관련 속성(irrelevant feature)들을 제거하고 상관성 기반 기법을

1) 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 (No. 한국연구 2011-0001044) 지원과 2011년 교육과학기술부로부터 지원받아 수행된 연구임" (지역거점연구단육성사업 / 충북BIT연구중심대학사업단)

이용하여 속성들 사이의 상관관계를 측정하여 강하게 중복된 속성들을 각각 서로 다른 다중 트리 구축에 사용되도록 하였다.

2. 중복성 제거 및 다중 의사결정 트리 구축

속성선택은 데이터에서 부적절하고 불필요한 정보를 제거하고 중요한 속성만을 선택하여 데이터분석의 효율성을 높이는 처리과정이다. 일반적으로, "좋은(good)" 속성은 예측 작업에 관련이 있을 뿐만 아니라 다른 속성들과 중복이 되지 않는 속성들을 의미한다. 만약 우리가 두 속성 사이의 상관성을 속성 평가의 척도로 한다면, 한 속성이 클래스와 높은 상관성을 가지고 있고 다른 속성들과 높게 상관되어있지 않다면 이 속성을 좋은 속성이라고 정의 할 수 있다.

두 속성 사이의 상관성을 측정하는 척도로 비교적 널리 알려진 선형 상관계수(linear correlation coefficient)[6]가 있다. 한 쌍의 속성(X,Y)에 대한 선형 상관계수 r은 아래 식(1)과 같이 계산된다.

$$r = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2}\sqrt{\sum_i (y_i - \overline{y_i})^2}} \quad (1)$$

이식에서 $\overline{x_i}$ 는X의 평균값이고 $\overline{y_i}$는 Y의 평균값이며 r값은 -1부터 1사이 구간에 놓이게 된다. 만약 속성 X와 Y가 완전히 상관되어 있으면, r값은 1 혹은 -1을 가지게 되고, 만약 X와 Y가 서로 독립적이면, r값은 0을 가지게 된다.

따라서 우리는 다중 의사결정 트리 구축과정에서 선형 상관계수 척도로 속성들 사이의 중복성을 제거하고 CS4 알고리즘의 cascading-and-sharing 앙상블기법을 적용하여 분류규칙의 singleton 문제를 해결하여 다양한 분류 규칙들을 생성하고, 생성된 규칙들을 총합스코어(aggregate score)[7]를 통하여 예측 프로세스를 완성한다. 총합스코어는 아래의 식 2와 같은 방법으로 산출할 수 있다. 여기서 C는 특정된 클래스의 분류 스코어를 의미한다.

$$Score^C(T) = \sum_{i=1}^{K_c} Coverage\,(rule_i^C) \quad (2)$$

3. 실험 결과

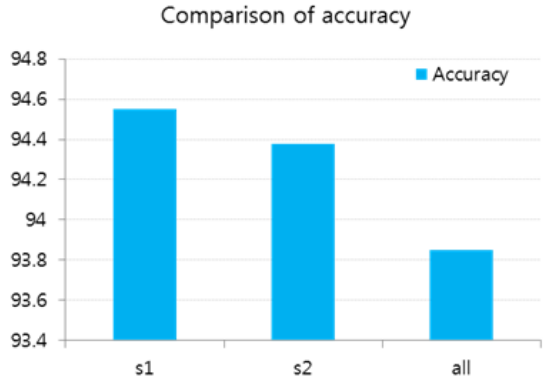
실험 데이터로는 캘리포니아 대학의 기계학습 저장소[6]의Wisconsin Breast Cancer Dataset을 이용하였으며. 이 데이터 집합은 총 569개의 레코드와 32개의 속성으로 구성되었다. 우리는 다중 트리 구축과정에서 이 속성들 사이의 중복성을 선형 상관계수 척도로 측정하여 중복성이 강한 속성들을 분리하여 각각 서로 다른 의사결정 트리를 구축하였고 각 트리의 예측을 종합하여 분류 작업을 수행하였다. 그 예측 결과는 (그림 1)에서 나타내고 있다. (그림 1)을 통해 알 수 있듯이, 중복된 속성들을 제거한 데이터 부분집합을 이용하여 구축한 트리는 전체 데이터를 사용 하였을 경우보다 좀 더 높은 정확도를 나타내는 것을 확인 할 수 있고 또 <표 3>과 같이 트리의 복잡도도 줄여줄 수 있다. 여기서 s1과 s2는 각각 중복된 속성들을 분리하여 구축된 트리이고 all은 중복된 속성들을 분리하지 않고 구축된 트리이다.

<표 1> 정확도 상세 정보

	FP rate	Precision	Recall	F-measure	Class
s1	0.031	0.946	0.906	0.925	M
	0.094	0.945	0.969	0.957	B
s2	0.045	0.925	0.925	0.925	M
	0.075	0.955	0.955	0.955	B
all	0.053	0.912	0.925	0.918	M
	0.075	0.955	0.947	0.951	B

<표 2> 트리 크기 비교

Tree	Number of Leaves	Size of the tree
s1	10	19
s2	12	23
all	13	25



(그림 1) 정확도 비교

4. 결론 및 향후 연구

본 논문에서는 중복된 속성들을 분리하고 cascading-and-sharing 앙상블기법을 적용하여 점진적으로 여러개의 다중 의사결정 트리를 구축하였다. 성능 평가를 위해서는 Wisconsin Breast Cancer Dataset을 사용하여 실험을 하였다. 그 결과 점진적 다중 의사결정 트리를 구축함에 있어서 중복성을 나타내는 속성들을 하나의 다중 의사결정 트리를 구축하는데 사용하는 것 보다 서로 다른 다중 트리 구축에 사용하면 분류기의 정확도를 높일 수 있음을 알 수 있었다.

향후 연구로는 속성들의 중복성을 측정함에 있어서 연속형 데이터뿐만 아니라 범주형 데이터들의 중복성도 측정 할 수 있는 상관성 척도를 이용하여 비선형 분포를 가지는 데이터에 대해서도 분석하고 중복성 기준값에 대해 더 명확한 방향을 제시할 것이다.

참고문헌

[1] Tan, P.N., Steinbach, M., Kumar, V.: Classification: Basic Concepts, Decision Trees, and Model Evaluation. In: Introduction to data mining, pp. 168-172. Addison Wesley, Reading (2006)

[2] Quinlan, J.R.: Induction of Decision Trees. Machine Learning, 81-106 (1986)

[3] Gao, Z., etc: Osteoporosis Diagnosis Based on the Multifractal Spectrum Features of Micro-CT Images and C4.5 Decision Tree(2010)

[4] Wang, Z. H., etc: Methodological Study on the Detection of the Variations of Forest Resources Based on C5.0 Algorithm- A Case of Culai Forest in Shandong(2011)

[5] Li, J. Y., Liu, H. A., See-Kiong Ng, Limsoon Wong: Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics, vol. 19, 93-102(2003)

[6] Lei Yu, Huan Liu: Proceedings of the Twentieth International Conference on Machine Learning(ICML-2003), Washington DC(2003)

[7] Utgoff, P. E.: Decision Tree Induction Based on Efficient Tree Restructuring. Technical report, University of Massachusetts(1994)

[8] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.html