러프 집합 분류기의 성능평가

Performance Evaluation of Rough Set Classifier

류재홍, 임창균
여수대학교 컴퓨터 공학과
Jae Hung Yoo and Chang-Gyoon Lim
Department Computer Engineering, Yosu National University
Yosu, 550-250, Korea

ABSTRACT

This paper evaluates the performance of a rough set based pattern classifier using the benchmarks in artificial neural nets depository found in internet. The definition of rough set in soft computing paradigm is briefly introduced. Next the design of rough set classifier is suggested. Finally benchmark test results are shown the performance of rough set compare to that of ANNs and decision tree.

I. 서론

1.1 러프 집합 소개

정보 시스템($IS = (U, A)$)은 행렬의 형태로 표현되고 행은 객체($x \in U$)또는 패턴 열은 속성($a \in A$)을 나타낸다. 객체 전체 집합(U)은 속성의 부분 집합(B)에 의하여 분할(partition)된다. 분할 영역이 서로 겹치지 않을 때 각 분할 영역은 동치 관계를 갖는 객체의 집합을 이룬다. 이집합은 객체의 무차별 집합(indiscernibility set) $I_B(x)$ 정의한다.

$$I_B(x) = \{y \in U \mid xR_By \}$$

러프 집합은 객체(U)의 부분집합 X를 속성(A)의 부분집합 B로 정확하게 정의할 수 없을 때 하부 근사 집합($\underline{B}X$)과 상부 근사 집합($\overline{B}X$)등 한 쌍의 집합으로 표현한다.

$$\underline{B} = \{ x \in U \mid I_B(x) \subseteq X \}$$

$$\overline{B} = \{ x \in U \mid I_B(x) \cap X \neq \emptyset \}$$

하부 근사 집합은 확실하게 X에 속하는 객체들의 집합이고 상부 근사 집합은 불확실하게 X에 속하는 객체들의 집합이다. 상부 근사 집합은 하부 근사 집합을 포함한다. 상부와 하부 근사 집합의 차집합은 경계 영역을 나타낸다.

$$\overline{B}X - \underline{B}X = \overline{B}X' - \underline{B}X', \; X' = U - X$$

러프 멤버 함수는 객체가 속한 무차별 집합(indiscernibility set) $I_B(x)$ 와 집합 X의 겹침을 계량화 한다.

$$\mu_B^X : U \rightarrow [0.1]$$

$$\mu_B^X(x) = \frac{|I_B(x) \cap X|}{|I_B(x)|}$$

러프 멤버 함수는 주어진 객체의 속성 B로부터 객체가 집합 X에 속할 확률을 객체 빈도로 추정(estimation)하는 것으로 해석할 수 있다(non-parmetric method). 한편 퍼지 멤버 함수는 경계 영역에 대하여 부분적 멤버쉽을 주어진 함수 모델에 의하여 수치화 한 것이다(parametric method).

1.2 패턴 인식과 학습방법

주어진 정보시스템이 행렬의 형태로 표현된 표로 이루어지고 각 행은 객체 또는 패턴 각 열은 속성을 나타낸다. 한 열이 패턴의 분류속성을 나타내고 나머지 열들은 측정 특성을 나타낸다면 주어진 표로부터 각 패턴 분류에 대한 공통적 특성을 추론하는 방법은 교사 학습(supervised learning)이라 부르는데 이를 체계적으로 연구하는 분야는 전통적으로 패턴 인식론의 영역에 속한다. 최근에는 기계학습이라는 부르는 인공지능의 한 영역으로 접근하는 시각도 있다.

교사학습 방법으로 결정 트리, 신경망 분류기, 퍼지 추론 분류기, 유전자 학습, 기계학습론의 연역학습법 등이 있다. 여러 학습 방법들의 분류 성능 향상과 추론결과에 대한 설명 능력 배양은 둘다 동시에 만족시키는 것이 어려운 것이다.

본 논문은 러프집합에 의한 분류기에 대하여 그 방법론을 소개하고 분류 성능에 대한 간단히 알아 본다.

II. 본론

2.1 러프 집합 분류기

러프 집합 분류기는 다중치 로직 최소화 방법(multivalued logic minimization method)에 의해 최적화된 규칙을 생성하고 객체 분류는 다수결 원칙에 의해 러프 멤버쉽이 최대인 분류 클래스로 정한다. 구체적인 단계는 다음과 같다.

1. 표 완성(Completion)
2. 양자화 (Qunatization)
3. 간략화 (Reduction)
4. 규칙 생성 (Rule generation)
5. 규칙 정리 (Rule filtering)
6. 분류 (Classification)

표 완성은 분실된 특성을 갖고 있는 패턴 객체를 제거하는 법, 각 분류 클래스의 평균값, 비수치 특성인 경우 최빈값으로 치환하느 법, 모든 특성치로 각각 치환하여 객체의 수를 확장하는 법등이 있다. 양자화 방법은 각 특성에 대한 스칼라 양자화(scalar quantization) 방법이 주류를 이룬

다. 균일 밀도 양자화, 엔트로피와 최소 기술 분량 원칙(minimun description length principle)에 의한 방법, 특성치 정렬법, 특성치 정렬과 분류 클래스 사용법 등이 있다. 또한 클러스터링(clustering)에 의한 벡터 양자화(vector quantization) 방법도 생각할 수 있다.
간략화 방법은 유전자 알고리즘을 이용하는 법, 모든 간략체(reducts)를 남김없이 찾는 법(exhaustive method)등이 있고 대규모 표에는 전자가 소규모 표에는 후자가 사용된다. 또한 객체에 따른 간략화 방법과 시스템 전체에 대한 간략화로 구분한다. 규칙 생성은 간략체에 시스템 표를 사용하여 규칙 개체를 만든다. 규칙 생성은 객체에 따른 간략화 방법과 동시에 병행하여 생성하는 것이 효과적이다. 규칙 정리는 간략체의 길이, 지원 객체수, 속성 집합에 의해 제거되고 정리된다. 또한 순수 불함수 최적화 방법을 직접 도입하는 방법도 생각할 수 있다.
분류방법은 한 패턴에 대하여 여러 분류 가능성이 있을 때 러프 집합 멤버쉽이 가장 큰 클래스로 정하는 법, 위험도가 큰 또는 발견 중요성이 큰 특정 클래스에 분류하는 법, 분류를 보류하는 법 등의 선택 요소가 있다. 러프 집합 멤버쉽은 정규화과정(normalization)에 따라 모든 규칙을 고려하는 법, 전제부가 일치한(firing) 규칙을 고려하는 법과 규칙당 가충치가 동일한 계수법, 규칙당 지원 객체수 계수법 등의 조합으로 4가지 경우를 생각할 수 있는 데 전제부가 일치한 규칙 중에서 규칙당 지원 객체수 계수법이나 동일 가중치를 부여하여 것이 클래스 당 개체가 균등할 때와 심하게 불균형을 이룰 때 각각 적용 할 수 있다. 다수결에 의한 방법외에 최대 멤버쉽에 의한 방법도 생각 할 수 있다.

2.2 성능 평가

분류기의 성능 평가는 주어진 데이터 집합에 대하여 k-분할 상호 검증법(k-fold cross validaion), 한 패턴 시험법(leave-one-out testing) 재배치 샘플링(replacemnt sampling)에 의한 자력 시험법(bootstrapping)등 이 체계적으로 패턴 분류 시스템의 성능을 평가하는 방법이다. 이것은 주어진 데이터를 훈련 집합(traing set)과 시험 집합(testing)으로 분리하여 훈련집합으로 패턴 분류 시스템을 학습시키고 시험집합으로 분류 시스템의 성능을 평가 하는 것이다.
성능 평가는 다음의 식으로 표현할 수 있다.

$$P_{cc} \cong \frac{1}{C}\sum_{j=1}^{C}\frac{1}{T_j}\sum_{k=1}^{T_j} b_{jk}$$

여기서 b_{jk}는 j 크래스 kqjsWo 샘플릐 불 함수값에 의한 분류 성적이다. 이것은 모집단이 불균일 크래스 분포를 가질 때 단순한 정확 분류 계수기보다 더 의미를 갖는다. 혼란 행렬(confusion matrix)보다는 축약된 성능 평가 정보를 보여준다.
기존의 결정 트리(decision tree), 인공 신경 회로망(artificial neural networks)의 주어진 패턴의 최종 분류 방법은 다수결에 의한 방법 또는 최고 여기치(maximum activation value) 등이 있다.
학습된 패턴 분류기의 복잡도도 중요한 평가 항목이다. 결정 트리의 평균 말단 노드 깊이(average depth of terminal nodes)와 말단 노드 수, 후 전파(back propagation) 신경 회로망의 비장층 노드 수(number of hidden layer nodes), 생성 규칙 체제(production rule systems)의 규칙의 조건부 평균 길이(average length of antecedent)와 규칙의 개수 등이 평가 항목이다.
본 논문에서는 UCI대학의 기계학습 데이터 베이스에서 가져온 GLASS 데이터에 대한 훈련 성적 평가를 예로 들었다. 근접 이웃 분류기의 인식률은 약 81%이다. GLASS 데이터는 유리의 8개의 화학 조성과 굴절지수(refractive index)의 정보 특성을 갖는 214개 데이터로 이루어 졌다. 6개의 유리 종류가 있고 각각의 패턴 분포가 상이 하다. 많은 분류 클래스의 샘플수가 작으므로 단순히

훈련 데이타로만 사용하였다. 표 2-1에서 SFDT는 단일 특성분류 결정트리, MPA는 수정된 포켓 알고리즘에 의한 신경 트리 분류기, BPN은 후전파 신경망, RSC는 러프집합 분류기를 나타낸다. 평균 크기는 노드수와 평균 길이 또는 깊이를 나타낸다.

	평균 크기	인식률(%)
SFDT	15,4.4	82.81
MPA	9,3.60	85.83
BPN	8	92.05
RSC	85, 4.082	71.03

표 2-1. GLASS 데이터에 대한 성능 시험

III. 결론

본 논문은 퍼지 분류기에 대한 규칙 생성 과정과 분류 기법에 대하여 소개하고 간단한 실험을 통하여 그 성능을 평가를 시도하였으나 충분한 비교 결과는 제시하지 못하였다. 추후 연구 과제로 남겨둔다.

참고 문헌

[1] D.E Rumelhart, G.E Hinton, and R.J. Willliams, "Leaning internal representation bu error propagation," in Parallel Distributed Processing Volume 1:Foundation, Eds. DE. Rumelhart, J.L. McClelland, MIT Press, 1986.

[2] I.K. Sethi and J.H. Yoo, "design of multicategory multisplit decision tree using perceptron learning." Pattern Recognition, Vol. 27, No. 7, pp. 939-947, 1994.

[3] Aleksnader Øhrn, *Rosetta Technical Reference Manual*, Draft version , Knowledeg System Group, Dept. of Computer and Information Science, Norwegian University of Science and Technology, Nov. 6, 1998.