공간 데이터의 분포를 고려한 공간 엔트로피 기반의 의사결정 트리 기법

장 윤 경[†] · 유 병 섭[††] · 이 동 욱[††] · 조 숙 경[†††] · 배 해 영[††††]

요 약

의사결정 트리는 데이터 마이닝의 분류와 예측 작업에 주로 사용되는 기법 중의 하나이다. 실생활에서 공간의사결정을 위한 분류를 수행할 때에는 인접 데이터의 위치와 분산도를 고려하는 것이 매우 중요하다. 기존의 공간 의사결정 트리는 데이터의 공간적 특성을 표현하기 위해 각 객체간의 유클리디안 거리비율을 엔트로피로 반영하여 트리 구축 시 이용하였다. 그러나 이것은 공간 객체간의 거리 비율만을 설명할 뿐 공간 차원에서의 데이터 분산 정도와 각 분류된 클래스간의 연관관계 등은 파악할 수 없다는 한계점이 있었다. 본 논문에서는 분산도와 차별도 기반의 공간 엔트로피를 이용하여 공간 데이터의 분포도를 반영하는 공간 의사결정 트리를 제안한다. 분산도는 분류된 클래스 내의 공간 객체 분포도를 나타내고 차별도는 다른 클래스 내 공간 객체와의 분포도 및 관계성을 나타낸다. 이러한 분산도와 차별도의 비율을 엔트로피 계산 시 이용함으로써 비공간적 속성으로 분류된 각 클래스가 공간적으로는 얼마나 뚜렷하게 분류되는지 알 수 있게 한다. 제안 기법은 정확성과 계산 비용에 있어서 기존 기법보다 각각 약 18%, 11%의 성능 향상을 보였다.

키워드 : 공간 의사결정 트리, 공간 데이터 분산도, 공간 엔트로피

A Spatial Entropy based Decision Tree Method Considering Distribution of Spatial Data

Youn-Kyung Jang[†] · Byeong-Seob You[††] · Dong-Wook Lee[††] · Sook-Kyung Cho[†††] · Hae-Young Bae[††††]

ABSTRACT

Decision trees are mainly used for the classification and prediction in data mining. The distribution of spatial data and relationships with their neighborhoods are very important when conducting classification for spatial data mining in the real world. Spatial decision trees in previous works have been designed for reflecting spatial data characteristic by rating Euclidean distance. But it only explains the distance of objects in spatial dimension so that it is hard to represent the distribution of spatial data and their relationships. This paper proposes a decision tree based on spatial entropy that represents the distribution of spatial data with the dispersion and dissimilarity. The dispersion presents the distribution of spatial objects within the belonged class. And dissimilarity indicates the distribution and its relationship with other classes. The rate of dispersion by dissimilarity presents that how related spatial distribution and classified data with non-spatial attributes are. Our experiment evaluates accuracy and building time of a decision tree as compared to previous methods. We achieve an improvement in performance by about 18%, 11%, respectively.

Key Words : Spatial Decision Tree, Distribution of Spatial Data, Spatial Entropy

1. 서 론

최근, GIS(Geographic Information System)사업의 발달에 따라 공간 정보 시스템이 구축되었고 지리정보에 대한 수요가 급증하였다[1, 3]. 방대한 지리정보는 공간 데이터베이스 또는 공간 데이터 웨어하우스 등에 저장되어 공간 데이터 마이닝에 이용된다. 공간 데이터 마이닝이란 공간 데이터 저장소로부터 함축적인 지식, 공간적 관계, 또는 명시적으로 저장되어 있지 않은 패턴들의 추출을 의미한다[4, 5, 11, 14]. 이러한 유용한 패턴들을 발견하기 위해 연관분석, 분류, 군집 등 여러 가지 데이터 마이닝 기법이 소개되었다. 그 중 분류(Classification)는 토지이용, 입지선정, 상권분석 등 의사결정을 할 때 주로 사용된다.

※ 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성 지원사업의 연구결과로 수행되었음.
† 준 회 원 : 인하대학교 대학원 컴퓨터정보공학과 석사과정
†† 준 회 원 : 인하대학교 컴퓨터정보공학과 박사과정
††† 정 회 원 : 인하대학교 지능형GIS센터 연구원
†††† 종신회원 : 인하대학교 대학원 원장
논문접수 : 2006년 9월 1일, 심사완료 : 2006년 10월 2일

분류기법에는 베이지안, 신경망, 의사결정 트리 등의 기법이 있다[6]. 베이지안 기법은 흥미 있는 변수들 사이의 확률적인 관계를 표현하는 그래프 모델이다. 이는 불완전한 데이터 집합 처리와 인과관계에 대해 학습하는 것을 가능하게 하는 장점이 있지만 모든 경우에 대한 변수를 측정하여 분류를 수행할 경우 입력 변수 중 한 개라도 관찰 되지 않으면 정확한 예측을 수행하지 못하는 단점이 있다. 신경망은 인간 두뇌 신경세포를 마디와 고리로 구성하여 망구조로 모형화 한 분류 기법이다. 이는 과거로부터 수집된 데이터를 사용하여 반복학습과정을 거쳐 데이터의 패턴을 찾아내는 장점이 있지만 정확도를 높일수록 각각의 셀마다 신경망을 구축해야 하는 어려움이 있고 다양하고 많은 입력변수를 필요로 하는 한계점이 있다. 의사결정 트리는 데이터 마이닝의 분류와 예측 작업에 주로 사용되는 기법으로 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴을 분류 모형 트리의 형태로 만드는 것이다. 이 기법은 신경망보다 훈련시간이 짧기 때문에 데이터의 규모가 클 경우 유리하고 결과에 대해 분류나 예측의 근거를 알려주기 때문에 사람이 이해하기 쉽다. 또한 신경망이나 베이지안 기법에 비해 SQL 문으로 바꾸기 쉬워 확장성이 뛰어나다[7, 13, 15].

분류에 유용하게 사용되는 기존의 의사결정 트리는 공간 마이닝에 이용될 경우 공간 속성을 고려하지 않았기 때문에 정확한 분류 결과를 제공하지 못했다. 이러한 문제점을 해결하기 위해 공간 객체간의 패턴 발견을 위해 공간 속성을 고려하는 의사결정 트리가 연구되어 왔다[2, 8, 10]. 기존 의사결정 트리의 대표적인 ID3에 기반하여 이웃 그래프(neighborhood graph)라는 개념을 접목시킨 의사결정 트리는 인접한 공간 객체와의 관련성을 찾는 연구 방법을 제시하였으나 단지 공간 관계에 대해서만 객체를 구분 지었기 때문에 이론과는 달리 실제적으로 효과적인 분류방법이 되지 못하였다[2]. 또 다른 방법으로는 공간 프레디킷을 이용한 2단계 기법 공간 의사결정 트리가 있다[8]. 이 공간 의사결정 트리는 모든 공간관계를 프레디킷으로 만들어야한다는 단점이 있고 공간 객체의 인접한 이웃들만 고려대상으로 한다는 한계점이 있었다. 또한 공간 객체의 유클리디안 거리를 바탕으로 만들어지는 의사결정 트리는 객체의 공간상의 거리에 대한 수치는 나타내지만 공간 차원의 객체의 분포도와 밀집도는 반영하지 못하는 단점이 있었다[10, 12].

본 논문에서는 공간 데이터의 분포를 고려한 공간 엔트로피 기반의 의사결정 트리 기법을 제안한다. 제안 기법은 데이터의 표준편차와 표준정규분포를 이용하여 클래스의 중점과 공간 데이터 간의 분산 정도를 알아내어 비공간적으로 분류된 객체들이 공간적으로는 어떠한 연관성이 있는지를 나타낸다. 또한 클래스 중심점을 거리 비교대상으로 하고 각 중심점을 표준화하여 이용하기 때문에 높은 신뢰성을 제공하고 계산 비용을 절약한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 공간 의사결정 트리 기법과 중심극한정리 등 관련연구에 대해 서술하고, 3장에서는 제안하는 공간 데이터 분포를 고려한 공간 엔트로피 기반의 의사결정 트리 기법에 대해서 알아본다. 4장에서는 제안 기법의 성능평가를 한 뒤, 마지막 5장에서 결론 및 향후 연구를 논한다.

2. 관련연구

본 장에서는 분류를 할 때에 공간 데이터를 고려하는 것이 얼마나 중요한지 알아보고 이러한 공간 속성을 고려하기 위한 기존의 공간 의사결정 트리에 대해 논한다. 또한 의사결정 트리의 공간 속성을 엔트로피에 반영하기 위한 객체간의 거리 계산 방식을 서술한다. 그리고 본 논문에서 분포도가 다른 클래스간의 비교를 위해 사용될 중심극한정리에 대해 알아본다.

2.1 분류에 있어서 공간 데이터의 중요성

실생활에서의 분류 패턴을 찾기 위해서는 공간적 속성을 고려하는 것이 매우 중요하다. 예를 들어 먹이종류, 나무의 종류 그리고 연못, 모래사장, 바다 유무 등에 따라 야생동물 서식에 관한 분류(백로 서식지, 금개구리 서식지, 바비 서식지 등)를 할 때 비공간적 속성만을 고려하면 그 분류의 결과가 정확하지 않다. 일반적인 경우 공간 객체는 균등 분포를 이루기보다는 편중된 분포를 이루기 때문에 그 공간객체의 분포를 비공간 속성만으로는 추출해내기가 어렵다. 또한 입지선정을 위한 분류를 할 때 지가, 인구밀도, 자연환경적 제한요소(홍수, 태풍, 지진 피해 등 빈발 발생 지역), 각종 시설 유무(백화점, 지하철, 학교 등) 등 비공간적 요소만 고려한다면 정확한 분류의 결과를 얻어낼 수 없다.

대형 마트의 입지선정이라고 가정한다면 이는 앞서 말한 비공간적 요소뿐만 아니라 인근 지역의 대형 마트의 위치, 백화점, 상점 등의 위치가 매출에 매우 중요한 요소이기 때문에 필수적으로 공간적 요소를 고려해야 한다. 따라서 비공간적 속성뿐만 아니라 인접한 공간 객체와의 관계까지 고려하는 공간적 분류가 필요하다. 이러한 예로 투기지역모니터링, 토지이용, 상권분석 등 실생활에서의 지리정보와 긴밀하게 연관되어 있는 분류의 예는 가까이서 찾아볼 수 있다.

2.2 공간 의사결정 트리

의사결정 트리는 데이터 마이닝의 분류와 예측 작업에 주로 사용되는 기법으로 과거에 수집된 데이터의 레코드를 분석하여 이들 사이에 존재하는 패턴, 즉 분류모형을 트리의 형태로 만드는 것이다. 하지만 앞서 말한 공간적 분류의 중요성에 따라 공간 데이터의 속성을 고려하는 의사결정 트리가 필요하게 되었다. 공간 의사결정 트리는 비공간 속성들의 패턴뿐만 아니라 비공간과 공간 속성 모두를 고려하여 흥미 있는 패턴을 추출하기 위한 트리 형태의 분류 모델이다.

이러한 공간 의사결정 트리를 위해 많은 연구가 시행되어 왔다. 기존 의사결정 트리의 대표적인 ID3에 기반하여 neighborhood graph라는 개념을 접목시킨 의사결정 트리는

("Topological Relations Between Regions in R2 and Z2") 인접한 공간 객체와의 관련성을 찾는 연구 방법을 제시하였으나 단지 공간 관계에 대해서만 객체를 구분 지었기 때문에 이론과는 달리 실제적으로 효과적인 분류방법이 되지 못하였다.

또 다른 방법으로는 공간 데이터의 분류를 위해 2단계 방법을 제시한 연구[8]는 공간 객체의 근접도나 공간 관계 등을 속성으로 만들어서 각 객체의 공간 분류 패턴을 알아내는 것이다. 이는 연관성있는 공간 객체의 관계를 알아내는 장점이 있지만 모든 공간관계를 프레디킷으로 만들어야한다는 단점이 있고 공간 객체의 인접한 이웃들만 고려대상으로 한다는 한계점이 있었다.

2.3 공간 엔트로피에 이용한 의사결정 트리

최근에 연구된 공간 의사결정 트리 모델의 다른 형태로 기존의 ID3 의사결정 트리의 엔트로피를 공간 엔트로피(spatial entropy)로 확장시킨 연구가 있었다. 이는 엔트로피 계산에 있어서 기존 ID3의 방법을 쓰되, 각 객체를 유클리디안 거리법을 사용해서 공간 차원 상에서의 객체간의 유사성과, 차별성을 반영하여 분류에 이용하는 방법이다. 이때 공간 다양화 계수(spatial diversity coefficient)라는 개념을 사용하여 유사성과 차별성을 계산하였는데 이 때 두 가지 규칙이 이용된다.

[규칙 1]: 다른 클래스에 포함되어있는 객체가 가까이 있으면 공간 다양화계수는 증가한다.

[규칙 2]: 같은 클래스에 포함되어있는 객체가 가까이 있으면 공간 다양화 계수는 감소한다.

이 공간 다양화 계수 값이 작을수록 같은 클래스 내의 유사성이 크고 다른 클래스 간의 차별성이 두드러지는 것이므로 분류가 잘 되었다고 할 수 있다.

유클리디안 거리를 이용하여 각 클래스 내의 객체 거리 비율과 다른 클래스 간의 객체 거리 비율을 구하는 방법은 다음과 같다. 이 식에서 각 파라미터C는 전체 공간 객체의 집합을 나타내고 Ci는 클래스 i에 속하는 객체의 수, dist(j, k)는 객체 j와 k의 유클리디안 거리를 나타낸다. 또한 d^{int}는 클래스 Ci에 포함되어 있는 객체간의 평균 거리이고 d^{ext}는 클래스 Ci와 다른 클래스에 속하는 객체간의 평균 거리를 나타낸다.

$$d_i^{int} = \frac{1}{|Ci| \times (|Ci|-1)} \sum_{j \in Ci} \sum_{k \in Ci; k \neq j} dist(j,k) \quad \text{if } |Ci| > 1; \text{ and } d_i^{int} = \lambda \text{ , otherwise} \tag{1}$$

d^{int}가 높을수록 같은 클래스안의 객체사이의 거리 비율이 높다는 것을 나타내고 이는 같은 클래스에 포함된 객체간의 유사성이 떨어지므로 엔트로피가 증가함을 나타낸다.

$$d_i^{ext} = \frac{1}{|Ci| \times |C - Ci|} \sum_{j \in Ci} \sum_{k \in (C - Ci)} dist(j,k) \quad \text{if Ci} \neq \text{C; and } d_i^{ext} = \beta \text{ , otherwise} \tag{2}$$

또한 d^{ext}가 높은 값을 가질수록 한 클래스와 다른 클래스에 포함된 각각의 객체간의 거리비율이 높다는 것을 나타내고 이는 서로 다른 클래스간의 차별성을 높여주므로 엔트로피를 감소시키는 효과가 있다.

데이터의 무질서 정도를 나타내는 엔트로피는 기존의 ID3의 엔트로피 공식에 d^{int}/d^{ext}의 비율을 반영하는 다음 식으로 나타낸다.

$$\text{Entropy}_S(A) = -\sum_{i=1}^{n} \frac{d_i^{int}}{d_i^{ext}} Pi \log_2 Pi \tag{3}$$

d^{ext}가 작고 d^{int}가 클수록 엔트로피 값이 높아지고 엔트로피가 높을수록 d^{ext}가 크고 d^{int}가 작을수록 엔트로피 값이 작아진다. 한 속성을 선택하였을 때 기대되는 엔트로피 감소율을 나타내는 정보이득률은 공간이 고려된 엔트로피를 이용하여 다음과 같이 표현된다.

$$\text{Gain}_S(GA, SA) = \text{Entropy}_S(GA) - \sum_{v \in Values(SA)} \frac{|GA_v|}{|GA|} Entropy_s(GA_v) \tag{4}$$

이러한 공간 데이터의 거리를 고려하여 엔트로피를 계산한 경우 비공간 속성뿐만 아니라 각 객체의 공간적인 속성까지 고려하므로 공간 데이터 사이에 존재하는 관계를 찾아내는데 유용하다. 본 논문은 기존의 의사결정 트리에서 벗어나 공간 차원의 데이터의 거리를 의사결정 트리 구축에 이용한다는데에 의의가 있다. 그러나 두 객체간의 거리를 계산하는데 있어서 유클리디안 거리 계산 방법을 사용했기 때문에 실생활에 이용할 때에는 공간 객체의 분포와 그 관계를 설명하는데 있어서 한계점이 있었다.

2.4 중심극한정리

중심극한정리는 통계학에서 가장 유용하게 사용되는 정리의 하나이다. 실생활의 사회, 경제, 경영 현상을 수량화하여 분포의 형태를 살펴보면, 그 분포가 정확히 정규분포를 이루는 경우는 드물다. 그러나 중심극한정리에 따르면 모집단의 분포에 관계없이 표본의 크기가 커질 경우 표본평균의 표본분포가 정규분포에 접근하게 되어 모집단이 정규분포라고 하지 않을지라도 표본의 크기가 큰 표본을 추출하면 정규분포의 성질을 이용하여 표본분석을 할 수 있는 것이다[9].

$X_1, X_2, \circ\circ\circ, X_n$을 평균 μ, 분산 σ^2인 임의의 모집단에서 추출한 크기 n인 확률표본이라 할 때, 표본평균 $\overline{X}$의 분포는 표본의 크기 n이 커짐에 따라 정규분포 $N(\mu, \frac{\sigma^2}{n})$ 접

근하게 되는데 이런 현상을 중심극한정리(central limit theorem)라고 한다. 또한 중점과 표준편차가 다른 정규분포를 위해서는 다음과 같이 표준정규분포로 변환을 해준다. 두 개의 확률변수 X, Y가 서로 독립이고 다음과 같이 정규분포를 따를 때 $X \sim N(\mu_1,\sigma_1^2)$, $Y \sim N(\mu_2,\sigma_2^2)$라고 나타낸다면 두 변수의 합과 차이는 각각 정규분포를 따르고 평균은 $\mu_1 \pm \mu_2$, 분산은 $\sigma_1^2+\sigma_2^2$이 된다. 이를 표준정규분포를 위해 변환 시키면 $X \pm Y \sim N(\mu_1 \pm \mu_2,\sigma_1^2+\sigma_2^2)$와 같다.

3. 공간 객체의 분포를 고려한 공간 엔트로피

엔트로피는 데이터의 무질서한 정도를 나타낸 것으로서 각 속성값을 분류 기준으로 채택했을 때 엔트로피 값이 가장 작은 것이 데이터를 지정된 클래스로 가장 잘 구별하는 특성이 된다.

본 장에서는 데이터의 공간 차원에서의 위치와 비공간적 속성들이 서로 얼만큼 영향력이 있는지 알기 위해서 공간 차원에서의 클래스 별 분산도와 차별도를 계산한다. 분산도와 차별도는 다음과 같은 특성이 있다.

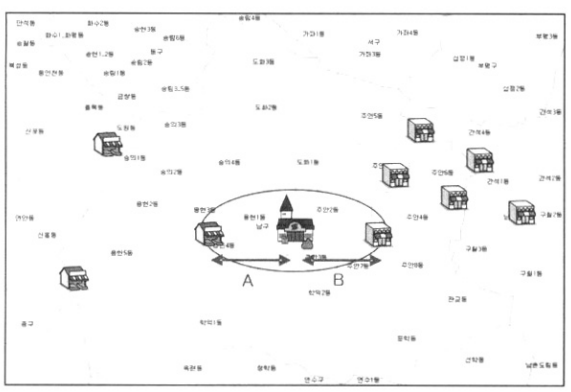
[특성 1]: 클래스 내의 객체들 간의 거리가 짧을수록 그 클래스는 분산도가 낮은 클래스이다.

[특성 2]: 한 클래스에 속하는 객체들이 중심점을 기준으로 퍼짐 정도가 작을 때 그 클래스는 분산도가 낮은 클래스이다.

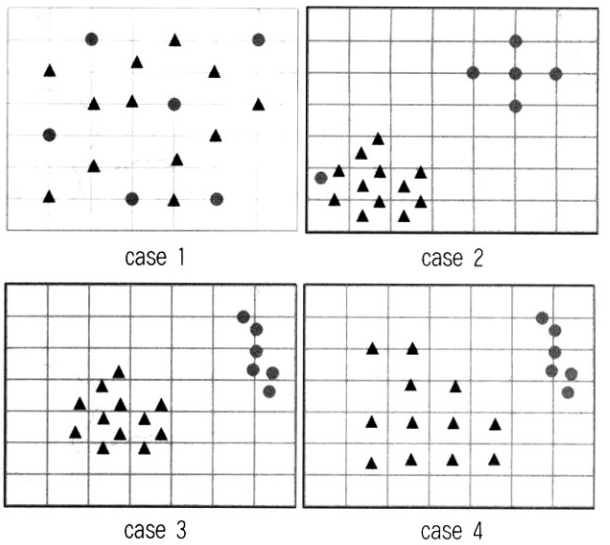
[특성 3]: 해당 클래스의 중심점과 다른 클래스와의 중심점의 거리가 길수록 차별도가 높은 클래스이다.

[특성 4]: 해당 클래스의 객체들이 다른 클래스의 중심점으로부터 멀리 떨어져서 분포되어 있으면 차별도가 높은 클래스이다.

소속된 객체의 위치와 분포도에 따라 클래스의 분산도와 차별도를 계산하는 것은 유클리디안 기반의 공간 의사결정 트리와 같이 객체간의 거리만을 가지고 평가하는 기법보다 높은 신뢰성을 제공한다.



(그림 1) 공간 데이터 밀집도의 중요성



(그림 2) 공간 데이터 분포에 따른 여러 특성

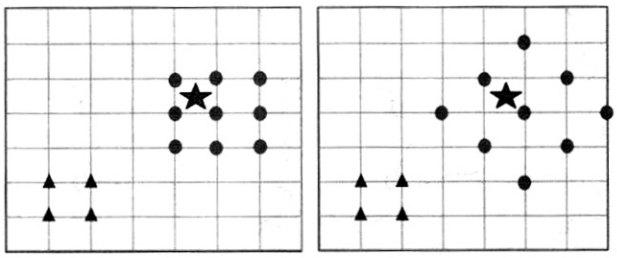
(그림 1)은 인천시 남구 대형 마트의 위치를 표시한 것으로 의사결정을 할 때 공간 객체의 분포도가 갖는 중요성을 설명한다. (그림 1)의 중앙에 있는 마트를 새로운 입지로 선정할 경우 유클리디안 거리 A와 B는 그 크기가 같다. 기존 기법에서는 공간 의사결정 트리를 구축할 때 유클리디안 거리를 공간 엔트로피에 반영한다. 하지만 실생활에서 마트 입지선정을 위한 의사결정을 할 경우 인접 마트와의 거리뿐만 아니라 인접한 지점의 마트 개수 등이 매출에 큰 영향을 미친다. 따라서 공간 객체의 위치와 더불어 공간 객체의 분포를 고려하는 것이 필수적이다. 공간 객체의 분포를 고려하여 (그림1)의 입지선정을 다시 고려해보면 마트가 흩어져 위치한 지역의 거리 A가 마트가 집중적으로 모여있는 지역의 거리 B보다 상대적으로 유리한 것을 알 수 있다.

이와 같이 공간적 위치의 고려가 필수적인 안건에 대한 의사결정을 내릴 때에는 그 객체의 위치, 거리뿐만 아니라 공간적 특성, 즉, 분산과 밀집도 등 다른 공간 객체와의 상관 관계에 대한 정보까지 알아내는 것이 중요하다. 본 논문은 (그림 2)와 같이 공간 객체가 분포되어 있을 경우에 대하여 분산도와 차별도를 제시하고 공간 엔트로피와 정보 이득율을 구하는 과정을 설명한다.

3.1 분산도

분산도는 같은 클래스로 분류된 객체가 서로 공간적으로 얼마나 인접한 거리에 있는지, 클래스 중심점으로부터 얼마나 조밀하게 모여있는지를 반영하는 척도이다. 즉, 같은 클래스에 속한 공간 객체가 중심점으로부터 가까운 거리에 조밀하게 모여있을수록 작은 분산도 값을 가진다. 클래스의 분산도가 작은 것은 비슷한 성질의 객체가 공간적으로 밀집되어 있는 것을 의미하므로 엔트로피를 감소시키는 효과가 있다.

제안 기법에서는 분산도를 구하기 위해 각 클래스의 중심점과 표준편차, 표준편차행렬을 구하고 각 클래스 데이터와



(그림 3) 공간 데이터 분포에 따른 분산도
A(왼쪽 그림), B(오른쪽 그림)경우

중심점의 거리를 구한다. 거리를 구할 시 표준편차행렬을 같이 곱하여 거리를 계산할 경우 데이터의 분포도를 반영하게 된다. (그림 3)과 같이 데이터가 분포되어 있을 경우의 분산도를 구하는 과정은 다음과 같다.

◦ **중심점 구하기**

각 클래스의 중심점($\bar{X}$, $\bar{Y}$)은 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$, $\bar{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i$로 구한다. (그림 3)의 A, B 경우를 보면 클래스(◉)은 같은 중심점과 다른 분포도를 가지고 있고 클래스(△)는 같은 중심점, 같은 분포도를 가지고 있는 것을 알 수 있다.

◦ **표준편차 구하기**

각 클래스에 대한 표준편차는 중심점으로부터 데이터가 얼마나 분산되어있는지 알려주는 척도가 된다. 표준편차 $\sigma_{xx},\sigma_{yy},\sigma_{xy}$는 다음과 같이 구한다.

$$\sigma_{xx}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Xi-\bar{X})^2}\quad \sigma_{yy}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Yi-\bar{Y})^2}\quad \sigma_{xy}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Xi-\bar{X})(Yi-\bar{Y})}$$

표준편차 값이 작을수록 데이터가 중심점으로부터 밀집된 분포로 구성되어 있다는 것을 의미한다.

◦ **데이터의 분포를 고려한 거리**

본 논문에서는 데이터의 분포도를 구려하기 위해 거리를 계산할 때 분산 행렬을 같이 곱한다. (그림 3)의 A, B 경우를 비교해보면 둘 다 같은 중심점을 가지고 있지만 분포도가 다른 경우이다. 이 때 유클리디안 기법으로 거리를 측정하면 데이터의 분포도를 반영하지 못하는 문제점이 있다. 따라서 본 논문에서는 데이터에 대한 표준편차행렬을 같이 곱해서 같은 거리라도 데이터의 분포도를 반영하여 다른 의미를 갖게 한다. 분포도를 고려하는 두 점 (X_1,Y_1), (X_2,Y_2)의 거리는 다음과 같이 구한다.

$$\text{Distance}=\sqrt{(X_2-X_1,Y_2-Y_1)^T S(X_2-X_1,Y_2-Y_1)}$$
$$\text{where } S=\begin{pmatrix}\sigma_{xx} & \sigma_{xy}\\ \sigma_{xy} & \sigma_{yy}\end{pmatrix} \quad (5)$$

(그림 3)의 각 셀의 크기가 2라고 하고 A, B 경우의 클래스(◉)를 예로 들어보자. 그림 A에서 평균은 (10,8), 분산 행렬은 $\begin{pmatrix}1.7 & -1.2\\ -1.2 & 1.7\end{pmatrix}$이 된다. 또 그림 B에서 평균은 (10, 8), 분산 행렬은 $\begin{pmatrix}1.6 & -1.4\\ -1.4 & 2.1\end{pmatrix}$이 된다. 그림A, B에서 중심으로부터 별 좌표(9, 9)의 거리를 구한다고 할 때 유클리디안 기법을 사용하면 그림 A와 B에서 모두 $\sqrt{2}$라는 같은 값을 갖지만 제안 기법을 이용하여 데이터의 분포도를 반영하기 때문에 별 좌표(9, 9)까지의 거리는 그림 A에서는 $\sqrt{5.8}$, B에서는 $\sqrt{6.5}$로 그 값이 다르게 나온다.

◦ **분산도 구하기**

분산도는 클래스의 중심점과 같은 클래스에 속해있는 객체간의 거리 비율을 나타낸다. 클래스 내에서 객체와 중심간의 거리뿐만 아니라 분산까지 알아내기 위해 앞서 말한 데이터 분포를 고려한 거리를 이용한다. 다시 말하면 같은 거리라도 조밀하게 모여있는 곳의 거리는 작게, 산만하게 흩어져있는 곳의 거리는 크게 만들어주어서 그 분산도를 반영할 수 있게 하는 것이다.

Mi는 클래스 i의 중심점, S는 표준편차행렬, X 는 지정된 좌표의 값을 표현하는 위치 벡터, |Ci|는 클래스 Ci에 속하는 데이터의 개수라고 할 때 클래스 i의 분산도는 다음과 같이 구할 수 있다.

$$\text{Dispersion} = \frac{1}{|Ci|}\sum_{X\in Ci}\sqrt{(X-Mi)^T S(X-Mi)}$$
$$\text{where } S=\begin{pmatrix}\sigma_{xx} & \sigma_{xy}\\ \sigma_{xy} & \sigma_{yy}\end{pmatrix} \quad (6)$$

클래스 i의 분산도는 클래스의 중심점(Mi)과 클래스 i의 다른 객체간의 분산 정도를 고려한 거리 비율로 구해진다. 기존 공간 엔트로피 기반의 트리 모형에서의 분산도는 같은 클래스의 각 객체간 거리의 비율로 나타내었다. 앞서 말한 것과 같이 유클리디안 기반 거리 측정법은 데이터의 분포도를 고려하지 않기 때문에 공간 차원에서 중요한 의미를 갖는 공간 데이터의 분포도를 반영할 수 없다.

◦ **데이터 분포도에 따른 기존 기법과의 비교**

(그림 2)의 CASE 1의 분포도의 경우 유클리디안 거리에 기반한 엔트로피와 제안 기법의 엔트로피는 큰 차이가 생기지 않는다. 그 이유는 비공간적으로 잘 분류된 데이터가 공간 차원에서는 특징 없이 흩어져있기 때문이다. 이러한 경우는 한 사람의 신용평가 등급에 대한 의사결정 등 위치공간 데이터와 연관성이 크지 않은 데이터 집합에서 주로 나타난다.

(그림 3)의 CASE 2와 같이 이상치 포함되어 있는 데이터의 엔트로피 계산 을 할 때 기존 기법은 모든 점에 대해서 같은 클래스의 한 데이터와 나머지 데이터의 거리를 모

두 합하기 때문에 ($\sum\sum dist(j,k)$) 기준 데이터가 이상치인 경우에 전체 거리 합에 큰 영향을 주는 약점이 있다. 따라서 이상치가 포함된 데이터 집합의 경우 신뢰성 있는 비교 기준을 제공하지 못한다. 하지만 제안 기법은 기준 데이터가 클래스의 중점이 되고 이를 중심으로 거리를 측정하기 때문에 이상치가 전체 거리 합에 미치는 영향을 줄일 수 있다.

(그림 2)의 CASE 3, 4와 같이 데이터 집합의 각 클래스의 중심점은 비슷하고 분포도는 다를 때 기존 기법은 데이터의 분포를 반영하지 못한다. 따라서 CASE 3과 4의 경우 그 차이를 분명하게 나타낼 수 없었다. 반면에 제안된 기법은 표준편차행렬 S를 이용하여 객체가 Mi를 중심으로 분산된 정도를 나타낸다. 즉, 중심점과 객체간의 길이가 같다고 하더라고 공간 데이터가 많이 퍼져있는 곳은 엔트로피를 증가시키는 효과가 있고 공간 데이터가 밀집되어 있는 곳은 엔트로피가 감소되는 효과가 있다.

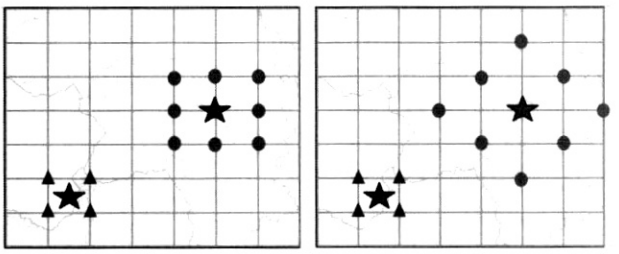
또한 제안 기법은 기존 기법보다 계산 비용을 절감시키는 효과가 있다. 기존 기법은 같은 클래스에 속하는 한 객체와 나머지 객체 모두의 거리 비율을 합하는 방법을 사용한다. 따라서 한 클래스 안의 객체가 n개라고 할 때 n(n-1)번의 거리 계산이 필요하였다. 이것은 높은 거리 계산 비용을 요구해 전체 의사결정 트리의 구축 비용을 증가시키는 요인이다. 그러나 제안 기법은 한 클래스 안의 두 객체간의 거리가 아닌 중심점으로부터의 클래스 내 객체의 거리 비율을 측정한다. 따라서 n개의 데이터가 같은 클래스에 속한다고 가정하였을 경우 (n-1)번의 거리 계산만이 필요하다. 이러한 특성들로 인하여 제안 기법은 계산 비용 절감과 신뢰성 있는 거리 계산 척도를 제공한다는 이점이 있다. 분산도를 고려하면 같은 클래스에 속하는 객체들에 대한 공간적 연관성을 분석할 수 있다. 클래스의 분산도가 작은 것은 비슷한 성질의 객체가 공간적으로도 잘 분류되어 있는 것을 의미하므로 엔트로피를 감소시키는 효과가 있다.

3.2 차별도

차별도는 다른 클래스에 속한 객체들이 공간 상에서 얼마나 특징적으로 군집화 되어있는지를 나타내는 척도이다. 다른 클래스에 속한 객체들이 거리상으로 멀리 떨어져있고 데이터의 퍼짐 정도의 경계선이 뚜렷할 때 클래스간의 차별도가 높다고 말한다. 차별도는 각 클래스의 중심점과 분산을 구하고 각 클래스의 중심점을 분산과 데이터 개수를 통해 Z-정규분포를 따르도록 표준화 시킨 후 거리비교를 통해서 나타난다.

◦ 중심점과 표준편차 구하기

중심점과 표준편차는 분산도를 구할 때 계산되었던 각 클래스의 중심점과 표준편차를 이용한다. (그림 4)의 A, B는 중점이 같고 한 클래스의 분포도가 다른 두 클래스로 구성되어 있다. (그림 4)의 셀 하나의 크기를 2라고 하면 클래스별 중심점과 표준편차행렬은 다음과 같이 구한다.



(그림 4) 공간 데이터 분포에 따른 차별도
A(왼쪽), B(오른쪽)경우

〈표 1〉 분포도에 따른 각 클래스 별 중심점과 표준편차행렬

	그림 A		그림 B	
	Class(▲)	Class(●)	Class(▲)	Class(●)
중점	(3, 3)	(10, 8)	(3, 3)	(10, 8)
표준편차행렬	$\begin{pmatrix} 1.15 & -1.15 \\ -1.15 & 1.15 \end{pmatrix}$	$\begin{pmatrix} 1.7 & -1.2 \\ -1.2 & 1.7 \end{pmatrix}$	$\begin{pmatrix} 1.15 & -1.15 \\ -1.15 & 1.15 \end{pmatrix}$	$\begin{pmatrix} 1.6 & -1.4 \\ -1.4 & 2.1 \end{pmatrix}$

〈표 2〉 표준화 된 클래스의 중심점

	그림 A		그림 B	
	Class(▲)	Class(●)	Class(▲)	Class(●)
표준편차	1.6	2.5	1.6	3.3
분포	X~N((3,3),1.6)	X~N((10,8),2.5)	X~N((3,3),1.6)	X~N ((10, 8),3.3)
표준화	$Z_t = \frac{Mi-(7,5)}{\sqrt{1.6^2/4+2.5^2/9}}$		$Z_t = \frac{Mi-(7,5)}{\sqrt{1.6^2/4+3.3^2/9}}$	

◦ 표준화 하기

각 클래스의 중심을 분산행렬을 이용하여 표준화 시킨다. 분류된 두 클래스는 데이터의 개수가 n_1, n_2이고 중점 및 표준편차가 각각 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$이다. 따라서 클래스의 중점의 표준화 좌표 Z_t는 다음과 같이 구한다.

$$Z_t = \frac{(X-Y)-(\overline{X}-\overline{Y})}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \quad (7)$$

(그림 4)에서 각 클래스에 대한 중점의 표준화는 <표 2>와 같이 이루어진다.

중심점과 분포도가 다른 클래스 A, B를 표준화 시키면 각 클래스는 $Z \sim N(0,1)$인 표준정규분포를 따르게 되고 이것은 동등한 비교 기준을 제시해 주기 때문에 두 집단에 포함된 점들에 대한 연산, 비교가 가능하게 된다.

◦ 차별도 구하기

차별도는 비공간적 속성에 따라 분류된 클래스들이 공간 차원에서 얼마나 각각 특징적으로 군집화 되어있는지를 나타내는 척도이다. 차별도는 표준화된 각 클래스의 중심점간의 거리로 구한다. 각 클래스의 중심점은 표준화 될 때에

이미 중점과 표준편차에 대해 변환이 된 값이므로 다른 클래스 간에도 바로 거리 값 측정 비교가 가능하다. 각 클래스의 표준화된 중심점 Z_t에 따른 차별도는 다음과 같다.

$$\text{Dissimilarity} = \frac{1}{|Zt|\times(|Zt|-1)}\sum_{X=Zt,X'\in Zt;X'\neq X}\sum\sqrt{Dist(X,X')}$$

$$\text{where } Z_t = \frac{(X-Y)-(\bar{X}-\bar{Y})}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}} \quad (8)$$

기존 기법에서는 차별도를 계산하기 위해서 한 클래스에 속하는 각각의 점과 다른 클래스에 속해있는 모든 점들의 거리 비율로 그 값을 나타내었다. 즉, 해당 클래스에 속하는 데이터가 개수가 n개이고 다른 클래스에 속하는 데이터가 각각 k개, m개라고 하면 차별도를 계산하기 위해서는 n*(k+m)번의 거리 계산이 필요했다. 이렇게 클래스에 속한 모든 점들에 대한 거리를 계산하는 것은 클래스를 대표적으로 나타낼 수 있는 점이 없다는 것을 의미한다. 이는 유클리디안 기법을 이용한 엔트로피 계산의 가장 큰 약점이며 이로 인해 높은 계산복잡도와 구축 비용이 필요하다.

본 논문에서 제안한 차별도는 중점과 표준편차에 따라 각 클래스의 중점끼리의 거리 비율을 표준정규분포를 따르도록 정규화 한 것이다. 따라서 각각 n, k, m의 데이터를 가지고 있는 3개의 클래스에 대한 차별도를 계산하기 위해서 단지 3번의 거리계산이 필요하다. 유클리디안 기법과 비교했을 때 표준편차 행렬을 구해야 하는 추가적인 비용이 들어가지만 기존 기법의 계산 복잡도와 비교했을 때 그 비용이 훨씬 작기 때문에 비용을 절감시키는 효과가 있다.

◦ 데이터 분포도에 따른 기존 기법과의 비교

차별도 역시 (그림 2)의 CASE 1과 같이 분류된 데이터 집합이 공간적 특성과 연관이 없는 경우 기존 기법과 큰 차이가 나지 않는다. 그러나 CASE 2와 같이 이상치가 있는 경우에는 각 중심점으로부터 표준편차를 반영하여 거리 계산을 수행하므로 이상치가 전체거리에 미치는 영향을 감소시킬 수 있다.

이렇게 제안 기법이 계산 비용을 줄이고 신뢰성 있는 판단 기준을 제공할 수 있는 것은 클래스의 중심이 중점과 표준편차행렬을 통하여 그 클래스를 타당성 있게 표현할 수 있는 대표 값이기 때문이다. 또한 각 클래스의 중점이 표준정규분포로 표준화 되기 때문에 CASE 3과 CASE 4의 그림 같이 각기 다른 중점과 분포도를 가지고 있다고 하더라도 같은 기준을 가지고 비교를 할 수 있게 된다. 또한 이러한 표준편차행렬은 공간 차원에서의 객체를 독립된 개체로 생각하지 않고 인접 객체와의 공분산도 함께 고려한 것이기 때문에 객체간의 의존성과 연관성까지 반영하게 된다.

이러한 차별도의 계산 방식은 공간적 속성을 고려한 엔트로피 계산시 오류를 줄이고 계산 비용을 절약하여 신뢰성 있고 효과적인 의사결정 트리의 구축을 돕는다. 차별도의 값이 클수록 비공간적 속성으로 분류된 클래스들이 공간적 차원에서도 클래스 별로 특성화되어 나타난다는 것을 의미하므로 엔트로피를 감소시키는 효과가 있다.

3.3 공간 엔트로피와 정보 이득율

데이터의 무질서 정도를 나타내는 엔트로피는 기존의 ID3의 엔트로피 공식에 분산도(Dispersion)/차별도(Dissimilarity)의 비율을 반영하는 다음 식 (8)과 같이 나타낸다.

$$\text{Entropy}_S(A) = -\sum_{i=1}^{n}\frac{Dispersion}{Dissimilarity}Pi\log_2 Pi \quad (8)$$

분산도 값이 크고 차별도 값이 작을수록 엔트로피 값이 높아지고 분산도 값이 작고 차별도 값이 클수록 엔트로피 값이 낮아진다. 즉, 같은 클래스에 속하는 공간 객체들이 흩어져서 위치하고 다른 클래스와의 경계가 불분명할 때 높은 엔트로피를 가진다. 이것은 비공간적 속성으로 분류된 객체들이 공간 차원의 위치와 낮은 상관 관계를 가지고 있음을 나타낸다. 또한 같은 클래스의 공간 객체의 분포가 군집되어 있고 다른 클래스의 공간 객체와 멀리 떨어져있을수록 낮은 엔트로피 값을 갖는다. 낮은 공간 엔트로피 값은 비공간적 속성으로 분류되는 객체들이 공간 차원 상으로도 잘 분류되어 있음을 나타낸다.

공간 의사 결정 트리를 구축할 때 여러 가지 속성 중 목표 속성(GA)의 엔트로피가 지원 속성(SA)를 택했을 경우 감소하는 엔트로피 값이 제일 큰 속성을 우선적으로 의사결정 트리 테스트 속성으로 선택한다.

$$\text{Gain}_S(GA, SA) = \text{Entropy}_S(GA) - \sum_{v\in Values(SA)}\frac{|GA_v|}{|GA|}Entropy_s(GA_v) \quad (9)$$

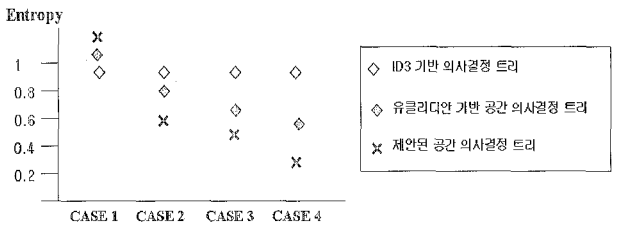
공간 데이터의 분산도와 차별도를 고려하여 계산한 공간 엔트로피를 가지고 의사결정 트리 테스트 속성을 선택할 경우 분류를 하는데 있어서 비공간 속성뿐만 아니라 각 객체의 공간적인 속성까지 고려한다. 그러므로 공간 객체와 밀접하게 관련된 실생활에서 분류를 수행할 때 유용하게 쓰일 수 있다.

4. 실험평가

본 논문의 실험은 2.6GHz의 중앙처리장치, 2GB의 주 기억장치, 120GB 보조 기억장치의 IBM PC 호환 기종에서 MS Window 2000 환경하에 진행되었다. 실험은 ID3 의사결정 트리, 유클리디안 거리 기반 공간 의사결정 트리, 제안된 의사결정 트리에 대해 산사태 발생 여부에 대한 분류를 수행하였다. 실험에 사용된 데이터 집합은 <표 3>과 같이 클래스 레이블을 가진 데이터로써 비공간 속성으로 식물의 종류, 토양의 종류, 비의 양, 지하수 유무, 지질의 종류, 기울기, 지가 등을 가지고 있고 공간 차원에서 (그림2)의 CASE

〈표 3〉 클래스 레이블을 가진 데이터

ID	식물의 종류	토양의 종류	비의 양 (시간 당)	지하수 유무	지질의 종류	기울기	지가	산사태 발생
1	잔디	진흙	≥30mm	있음	현무암	>40°	150	YES
2	관목	모래	≥30mm	있음	대리암	>40°	200	YES
3	잔디	적토	≥30mm	있음	현무암	≤40°	50	NO
…	…	…	…	…	…	…	…	…
19	잔디	진흙		있음	현무암	>40°	45	YES
20	침엽수	모래	≥30mm	있음	대리암	>40°	125	YES
…	…	…	…	…	…	…	…	…

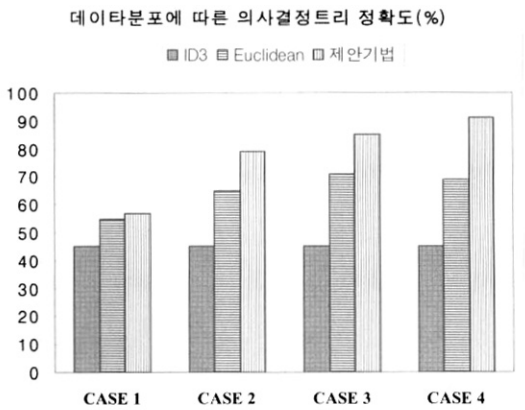


(그림 6) 공간 데이터 분포에 따른 엔트로피 변화

1, 2, 3, 4와 같은 모양으로 분포되어 있다.

(그림 6)은 (그림 2)의 CASE 1, 2, 3, 4와 같이 서로 다른 공간 데이터 위치, 분포도를 가진 데이터 집합에 대하여 ID3 기반 의사결정 트리, 유클리디안 기반 공간 의사결정 트리, 제안된 공간 의사결정 트리의 공간 엔트로피를 비교한 것이다. 실험은 클래스 레이블을 가진 데이터 5000개를 가지고 의사결정 트리를 구축할 때 나타나는 엔트로피를 측정하였다.

기존의 ID3 의사결정 트리는 공간적 속성을 고려하지 않았으므로 CASE 1, 2, 3, 4에 대해 같은 엔트로피 값을 갖는다. 즉 공간 차원에서 객체의 위치에 상관없이 비공간적 속성만을 가지고 분류를 수행하게 된다. 이것은 실생활의 공간 객체에 대한 분류를 수행할 경우 신뢰성 있는 분류 결과를 제공하지 못한다. 유클리디안 기반 공간 의사결정 트리는 공간 객체간의 유클리디안 거리를 가지고 데이터의 공간적 특성을 구분한다. 따라서 객체들이 공간적인 분류 특색 없이 흩어져 있는 CASE 1의 경우에는 공간 차원에 대한 계산 비용으로 인하여 엔트로피가 오히려 증가하지만 공간적으로 특성을 가지는CASE 2, 3, 4의 경우 엔트로피를 감소시켜 ID3보다 신뢰성 있는 분류를 수행한다. 제안된 공간 의사결정 트리는 CASE 1의 경우 공간 데이터에 대한 표준화, 중점과 분산의 계산 등으로 인해 ID3, 유클리디안 기반 공간 의사결정 트리보다 계산 비용이 증가한다. 하지만 CASE 2의 경우 표준화된 데이터 제공과 공분산을 반영함으로써 기존 기법보다 엔트로피가 줄이고 신뢰성 있는 의사결정 트리를 구축한다. 데이터 군집이 타원으로 형성되어 있는 CASE 3의 경우와 데이터가 조밀하게 모여있는 CASE 4의 경우 객체가 공간 차원 상에서 클래스 별로 군집화되어 있기 때문에 분류를 수행할 때 공간 객체의 위치가 비공간 속성과 함께 고려되어야만 한다. 제안기법은 각 클래스 공

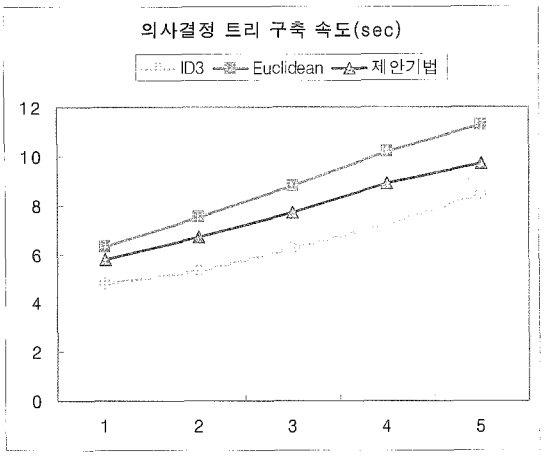


(그림 7) 공간 데이터 분포에 따른 분류 결과 정확도 변화

간 객체들의 중심점과 분산을 고려하여 공간 객체의 분포도가 비공간적 속성과 함께 엔트로피 계산에 참가한다. 제안기법은 ID3, 유클리디안 기반 공간 의사결정 트리보다 작은 엔트로피 값을 가지고 공간, 비공간 속성뿐만 아니라 데이터의 분포도를 고려하여 의사결정 트리를 구축하므로 실생활에 적합한 분류를 수행한다.

(그림 7)은 (그림 2)와 같이 CASE 1, 2, 3, 4와 같이 데이터가 다양하게 분포되어 있는 경우에 따른 의사결정 트리의 정확도를 나타낸 것이다. 실험은 5000개의 클래스 레이블을 가진 데이터 중에서 2000개의 데이터를 훈련 데이터 집합으로 구성하여 의사결정 트리를 구축하고 나머지 3000개의 데이터를 검정 데이터 집합으로 선택하여 구축된 의사결정 트리가 분류를 올바르게 수행하는지를 검증한다. 본래 구축된 트리는 테스트 데이터를 가지고 클래스 레이블을 예측하기 위해 분류를 수행하지만 본 실험에서는 분류 결과의 정확성 확인을 위해 검증 데이터를 가지고 분류를 수행하였다. 검증 데이터는 분류에 사용될 여러 속성과 함께 최종적으로 분류된 클래스 레이블을 가지고 있기 때문에 구축된 의사결정 트리가 올바른 분류를 수행하였는지 확인할 수 있다.

CASE 1의 경우 구축된 의사결정 트리를 가지고 검증 데이터에 대한 분류를 수행하였을 때 의사결정 트리의 분류 수행 결과와 검증 데이터의 클래스 레이블의 일치도가 낮았다. 또한 유클리디안 기반 의사결정 트리와 제안된 의사결정 트리의 분류 정확도의 차이가 크지 않았다. 이것은 CASE 1의 데이터 집합 속성이 공간 차원과 관련이 적기 때문인데 이러한 경우라면 트리 구축 시간, 엔트로피 계산 비용 등을 고려할 때 ID3 방식으로 구축하는 것이 가장 적합하다. CASE 2는 공간 차원에서 클래스 데이터가 잘 분류되어 있지만 이상치가 포함된 경우이다. 유클리디안 기법에서는 모든 데이터가 중심점이 되어 다른 데이터와의 거리 계산을 수행하기 때문에 이상치가 전체 엔트로피에 미치는 영향이 크다. 하지만 제안 기법은 모든 데이터를 대표할 수 있는 중심점으로부터 다른 데이터와의 거리 계산을 수행하기 때문에 이상치가 전체 엔트로피에 미치는 영향을 줄임으로써 의사결정 트리의 정확도를 높인다. CASE 3은 한 클래



(그림 8) 데이터 크기에 따른 의사결정 트리 구축 속도 변화

스의 데이터 분포가 타원형으로 되어있는 경우이다. 유클리디안 거리 계산 방법만으로는 길쭉한 모양으로 구성되어 있는 타원 모양의 데이터 분포를 잘 반영하지 못한다. 즉 고르게 분포되어 있는 데이터에 대하여 높은 정확도를 나타내지만 편중된 분포의 데이터 집합에 대해서는 정확성이 떨어진다. 하지만 실생활의 공간 객체는 고르게 분포되기 보다는 편중된 분포로 위치하는 경우가 더 많으므로 유클리디안 기반 의사결정 트리는 적합하지 않다. 제안 기법은 데이터의 분포도를 반영하기 때문에 편중된 데이터 집합이나 원 모양 분포의 데이터 집합이 아니더라도 신뢰성 있는 분류를 수행한다. CASE 4는 CASE 3과 비슷한 중점을 가진 클래스의 분포도를 다르게 한 경우이다. 분산도와 차별도를 고려한 제안 기법은 이러한 경우에도 정확도의 감소가 없었지만 유클리디안 기법의 경우 데이터의 분포를 반영하지 못하기 때문에 CASE 3보다 정확도가 조금 떨어지는 것을 볼 수 있다.

(그림 8)은 ID3, 유클리디안 기반 공간 의사결정 트리, 제안 기법에 대해서 각각 데이터 1000개, 2000개, 3000개, 4000개, 5000개를 가지고 트리를 구축했을 때 소요되는 시간을 나타난 그래프이다. ID3의 경우 각 속성에 속하는 데이터 개수만를 가지고 엔트로피를 계산하기 때문에 구축 속도가 제일 빠르다. 유클리디안 기반 의사결정 트리의 경우 전체 데이터가 여러 번씩 엔트로피 계산에 이용되기 때문에 의사결정 트리를 구축하는데 있어서 많은 비용이 요구되었다. 제안된 의사결정 트리는 각 클래스를 대표하는 중심점을 가지고 다른 데이터와 비교하면서 엔트로피를 계산하기 때문에 유클리디안 기반 의사결정 트리보다 훨씬 구축 속도가 빠른 것을 볼 수 있었다.

5. 고찰 및 결론

본 논문에서는 표준편차행렬과 표본 정규분포에 기초하여 계산된 엔트로피 기반의 공간 의사결정 트리에 대해서 알아보았다. 제안 기법은 각 클래스의 중심점과의 표준편차행렬을 이용하여 클래스의 분산도를 측정하고 차별도는 각 클래스의 중점을 표본정규분포에 맞게 변형시킨 후 중점 간의 거리 비율로 나타내어진다. 이러한 제안 기법은 데이터의 분포도를 반영하기 때문에 비공간적으로 분류된 객체들이 공간 차원에서는 어떠한 연관 관계가 있는지 알 수 있다. 제안 기법은 기존의 유클리디안 기법 의사결정 트리보다 데이터의 거리와 분포도를 함께 고려하기 때문에 객체들의 공간적 성질을 더욱 잘 반영해줄 수 있는 공간 엔트로피 계산을 수행하고 이로 인해 비공간적으로나 공간적으로 모두 신뢰성 있는 의사결정 트리를 구축하게 된다. 또한 거리를 잴 때 모든 객체간의 거리가 아닌 각 클래스를 대표할 수 있는 중심점으로부터의 거리를 계산하기 때문에 계산 비용을 줄여준다. 실험평가 시 5000개의 클래스 레이블을 가진 데이터를 나누어서 2000개는 트리를 구축하는 훈련 데이터로 사용하고 나머지 3000개는 트리를 검증하는 검정 데이터로 사용했다. 실험 결과 제안 기법은 기존의 ID3 의사결정 트리와 공간 엔트로피 의사결정 트리와 비교하여 정확성 측면에 있어서 약 18%의 성능향상을 보였고 트리 구축에 사용되는 데이터 개수를 늘려가며 기존 기법과 비교한 결과 계산 비용면에서 약 11%의 성능 향상을 보였다.

향후 연구로는 공간 프레디킷을 의사결정 트리의 검사 조건으로 포함시키는 것과 공간 차원에서 장애물을 고려하여 엔트로피를 구하는 방법을 고려하고 있다.

참 고 문 헌

[1] Longley P. A., Goodchild M. F., Maguire D. J., Rhind D. W., Geographical Information Systems - Principles and Technical Issues, John Wiley & Sons, Inc., 1999.

[2] Nadjim Chelghoum, Karine Zeitouni, " Spatial Decision Tree-Application to Traffic Risk Analysis", GeoI100 info Symposium, 2004

[3] Claramunt C 2005 A spatial form of diversity. In Mark D M and Cohn A (eds) Spatial Information Theory: Proceedings of COSIT 2005. Berlin, Springer Lecture Notes in Computer Science No 3693: 218 - 31

[4] Martin Ester, Hans-Peter Kriegel, Jorg Sander, "Spatial Data Mining: A Database Approach", Proceedings of the Fifth Int. Symposium on Large Spatial Databases, 1997

[5] Miller, H. J. and Han, J., 2000, Discovering geographic knowledge in data rich environments: a report on a specialist meeting, ACM SIGKDD Explorations, 1(2), 105-107.

[6] Han, J., Kamber, M., "Data Mining: Concepts and Techniques," Morgan Kaufman, 2001.

[7] Quinlan J R 1986 Introduction of decision tree. Machine Learning 1: 81 - 106

[8] Koperski, K., Han, J., and Stefanovic, N., 1998, An efficient two-step method for classification of spatial data, Proc. International Symposium on Spatial Data Handling

(SDH '98), Vancou-ver, Canada, 45-54.
[9] Kaneko, K., Globally coupled chaos violates the law of large numbers but not the central-limit theorem, Physical Review Letters 65 (12), pp. 1391-1394, 1990
[10] Xiang Li, Christophe Claramunt, "A Spatial Entropy-Based Decision Tree for Classification of Geographical Information", Transactions in GIS, 2006
[11] Ester M, Kriegel H, and Sander J 1997 Spatial data mining: A database approach. In Scholl M and Voisard A (eds) Proceedings of the Fifth International Symposium on Large Spatial Databases (SSD'97). Berlin, Springer Lecture Notes in Computer Science No 1262: 48 - 66
[12] De Maesschalck R., Jouan-Rimbaud D., Massart D.L., "The Mahalanobis distance", Chemomerics and Intelligent Laboratory Systems, Vol, 50, No. 1, 2000
[13] Pal N R and Chakraborty S 2001 Fuzzy rule extraction from ID3-type decision trees for real data. IEEE Transactions on Systems Man and Cybernetics Part B-Cybernetics 31: 745 - 54
[14] Shekhar S, Zhang P, Huang Y, and Vatsavai R 2003 Trends in spatial data mining. In Kargupta H, Joshi A, Sivakumar K and Yesha Y (eds) Data Mining: Next Generation Challenges and Future Directions. London, AAAI Press: 357 - 801
[15] Mitchell T M 1997 *Machine Learning*. New York, McGraw-Hill

장 윤 경

e-mail : ykjang@dblab.inha.ac.kr
2005년 8월 인하대학교 컴퓨터공학부 (공학사)
2005년 8월~현재 인하대학교 대학원 컴퓨터정보공학과(석사과정)
관심분야 : 공간 데이터 마이닝, Stream Data, 공간 데이터 웨어하우스

유 병 섭

e-mail : subi@dblab.inha.ac.kr
2002년 인하대학교 컴퓨터공학부(공학사)
2004년 인하대학교 컴퓨터공학부 (공학석사)
2004년~현재 인하대학교 대학원 컴퓨터정보공학과(박사과정)
관심분야 : 공간데이터베이스, 공간 데이터 웨어하우스, Data Stream, 유비쿼터스 컴퓨팅

이 동 욱

e-mail : dwlee@dblab.inha.ac.kr
1996년~2003년 상지대학교 전자계산공학과 학사
2003년~2005년 인하대학교 컴퓨터정보공학과 석사
2005년~현재 인하대학교 컴퓨터정보공학과 박사과정
관심분야 : Spatial Database Warehouse, Spatial Information Management, Ubiquitous 환경을 위한 SDBMS

조 숙 경

e-mail : skyoe@dreamwiz.com
1990년 인하대학교 전자계산학과(이학사)
1994년 인하대학교 대학원 전자계산공학과(공학석사)
2002년 인하대학교 대학원 전자계산공학과(공학박사)
2003년~2006년 8월 인천대학교 컴퓨터공학과 강의전담교수
2006년 9월~현재 인하대 지능형 GIS 센터 연구원
관심분야 : 데이터베이스, 실시간 데이터베이스 시스템, 스트림 데이터베이스

배 해 영

e-mail : hybae@inha.ac.kr
1974년 인하대학교 응용물리학과 (공학사)
1978년 연세대학교 대학원 전자계산학과 (공학석사)
1989년 숭실대학교 대학원 전자계산학과 (공학박사)
1985년 Univ. of Houston 객원교수
1992년~1994년 인하대학교 전자계산소 소장
1982년~현재 인하대학교 컴퓨터공학부 교수
1999년~현재 지능형GIS연구센터 센터장
2000년~현재 중국 중경우전대학교 대학원 명예교수
2004년~2006년 인하대학교 정보통신대학원 원장
2006년~현재 인하대학교 대학원 원장
관심분야 : 분산 데이터베이스, 공간 데이터베이스, 지리정보시스템, 멀티미디어 데이터베이스