(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2008년01월02일
(11) 등록번호 10-0789966
(24) 등록일자 2007년12월21일

(51) Int. Cl.
*G06F 19/00* (2006.01)
(21) 출원번호 10-2006-0115668
(22) 출원일자 2006년11월22일
심사청구일자 2006년11월22일
(56) 선행기술조사문헌
10-2003-0022663

(73) 특허권자
**인하대학교 산학협력단**
인천 남구 용현동 253 인하대학교
(72) 발명자
**배해영**
인천 연수구 송도동 2-3 이안아파트 101동 1703호
**이재동**
경기 성남시 분당구 서현동 시범단지우성아파트 225동 1902호
*(뒷면에 계속)*
(74) 대리인
**이은철**

전체 청구항 수 : 총 2 항 심사관 : 박지은

## (54) 공간데이터 분포를 고려한 공간 엔트로피 기반의 의사결정트리 구축방법

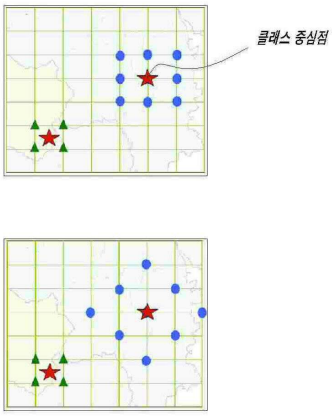
*(57) 요 약*

본 발명은 공간데이터 분포를 고려한 공간 엔트로피 기반의 의사결정 트리 구축방법을 개시한다.

본 발명에 의사결정 트리 구축방법은, 각 클래스의 중심점과 동일 클래스 내에 속한 객체간의 거리비율을 반영한 분산도(dispersion)를 산출하는 제 1 과정과, 각 클래스의 중심점을 분산과 데이터 개수를 통해 Z-정규분포를 따르도록 표준화시킨 후, 클래스 간의 중심점에 대한 거리비교를 수행하여 차별도(dissimilarity)를 산출하는 제 2 과정과, 분산도와 차별도의 비율을 반영하여 공간 엔트로피를 산출하는 제 3 과정으로 구성된다.

이러한 본 발명에 의하면, 공간 차원의 데이터 분포도를 반영하여 비공간적으로 분류된 객체들이 공간 차원에서 어떠한 연관관계(상관관계)가 있는지 추정할 수 있으며, 공간데이터의 거리와 퍼짐 정도를 함께 고려하여 객체들의 공간적 특성을 반영한 엔트로피를 산출하므로 공간적·비공간적으로 신뢰성 있는 의사결정 트리를 구축할 수 있다.

*대표도* - 도4



(72) 발명자

**오영환**

충남 천안시 쌍용동 1538 월봉 벽산 태용아파트 205동 204호

**김경배**

대전 서구 관저2동 느리울마을 1106동 106호

**김호석**

경기 부천시 오정구 원종1동 229-2 8/1

**장용일**

인천 부평구 십정동 316-12

**유병섭**

경기 남양주시 도농동 부영아파트 103-1702

**어상훈**

서울 구로구 개봉동 상우아파트 102동 105호

**이동욱**

서울 강남구 삼성2동 125-26 401호

**조숙경**

인천 남구 주안4동 1537-46 3층

**장윤경**

서울 구로구 구로본동 오피스1 905호

## 특허청구의 범위

### 청구항 1

공간 엔트로피 기반의 의사결정 트리 구축방법에 있어서,

각 클래스의 중심점과 동일 클래스 내에 속한 객체간의 거리비율을 반영한 분산도(dispersion)를 산출하는 과정;

상기 각 클래스의 중심점을 분산과 데이터 개수를 통해 Z-정규분포를 따르도록 표준화시킨 후, 클래스 간의 중심점에 대한 거리비교를 수행하여 차별도(dissimilarity)를 산출하는 과정; 및

상기 분산도와 차별도의 비율을 반영하여 공간 엔트로피를 산출하는 과정; 을 포함하는 것을 특징으로 하는 공간데이터 분포를 고려한 공간 엔트로피 기반의 의사결정 트리 구축방법.

### 청구항 2

청구항 1에 있어서,

상기 분산도 산출 과정은 다음의 [수학식]으로 표현되며, 데이터의 분포도가 반영되는 것을 특징으로 하는 공간데이터 분포를 고려한 공간 엔트로피 기반의 의사결정 트리 구축방법.

$$dispersion = \frac{1}{|c_i|}\sum_{X \in c_i}\sqrt{(X-M_i)^T S(X-M_i)} \quad ; \quad S=\begin{bmatrix}\sigma_{xx} & \sigma_{xy}\\ \sigma_{xy} & \sigma_{xy}\end{bmatrix}$$

여기서, $M_i$는 클래스 $i$의 중심점, $S$는 데이터에 대한 표준편차 행렬, $X$는 지정된 좌표 값의 위치벡터, $|c_i|$는 클래스 $i$에 속하는 객체의 개수이다.

## 명 세 서

## 발명의 상세한 설명

### 발명의 목적

#### 발명이 속하는 기술 및 그 분야의 종래기술

<5> 본 발명은 공간 엔트로피 기반의 의사결정 트리 구축방법에 관한 것으로서, 특히 공간 차원의 데이터 분포도를 반영함으로써 비공간적으로 분류된 객체들이 공간 차원에서 어떠한 연관관계가 있는지 추정할 수 있는 기술에 관한 것이다.

<6> 데이터 마이닝(data mining) 관련 기술이 활발히 연구되고 있다. 데이터 마이닝 시스템에서 필요한 주요 기능적 요소로서 클래스화(classification) 혹은 클러스터링(clustering)이 있다. 클래스화에 대해서는 신경망을 포함한 다수의 기법들이 소개되어 있으며, 특히 의사결정 트리(decision tree)를 이용한 기법이 주목받고 있다. 의사결정 트리에 대한 대표적 툴로서 엔트로피를 이용한 ID3(Interactive Dichotomizer 3)를 들 수 있다.

<7> 최근 의사결정 트리 구축 과정에 공간 엔트로피(spatial entropy)를 적용한 연구가 있었다. 이 연구는 엔트로피 계산에 있어서 기존 ID3의 방법을 이용하되, 각 객체를 유클리디안(Euclidean) 거리법을 사용하여 공간 차원 상에서 객체 간의 유사성 및 차별성을 반영하여 분류에 적용하는 방법을 개시하고 있다. 구체적으로, 공간 다양화 계수(spatial diversity coefficient)라는 개념을 바탕으로 상기 유사성 및 차별성을 계산하며, 이때 다음의 두 가지 규칙을 따른다.

<8> **규칙 1:** 다른 클래스에 포함되어 있는 객체가 가까이 있으면 공간 다양화 계수는 증가한다.

<9> **규칙 2:** 동일 클래스에 포함되어 있는 객체가 가까이 있으면 공간 다양화 계수는 감소한다.

<10> 이 두 가지 규칙을 살피면, 공간 다양화 계수의 값이 작을수록 동일 클래스 내의 유사성이 커지고 다른 클래스 간의 차별성이 현저해지는 것이므로 개념에 대한 분류가 잘 이루어진 것이라 판단할 수 있다.

<11> 한편, 상기 유클리디안 거리법을 이용하여 각 클래스 내의 객체 거리 비율과 다른 클래스간의 객체 거리 비율을

구하는 방법을 아래의 [수학식 1] 및 [수학식 2]를 참조하여 살펴보면 다음과 같다.

<12> $$d_i^{\text{int}} = \frac{1}{|c_i| \times (|c_i| - 1)} \sum_{j \in c_i} \sum_{k \in c_i; k \neq j} dist(j,k)$$ ..... [수학식 1]

<13> $if\ |c_i| > 1;$
$otherwise,\ d_i^{\text{int}} = \lambda$

<14> $$d_i^{ext} = \frac{1}{|c_i| \times |c - c_i|} \sum_{j \in c_i} \sum_{k \in (c_i - c)} dist(j,k)$$ ..... [수학식 2]

<15> $if\ c_i \neq c;$
$otherwise,\ d_i^{\text{ext}} = \beta$

<16> 여기서, $c$는 전체 공간 객체의 집합을 나타내고, $c_i$는 클래스 $i$에 속하는 객체의 수이며, $dist(j,k)$는 객체 $j$와 $k$의 유클리디안 거리를 나타낸다. 또한 $d^{\text{int}}$는 클래스 $c_i$에 포함되어 있는 객체간의 평균거리이고, $d^{ext}$는 클래스 $c_i$와 다른 클래스에 속하는 객체 간의 평균거리를 나타낸다. 위 [수학식 1]은 $|c_i|$가 1보다 큰 경우에 적용되며, 만약 1보다 작거나 같을 경우에는 상수 값 $\lambda$로 지정되어 식의 분모가 0이 되는 것을 방지한다. 또한, [수학식 2]는 $c_i \neq c$의 조건일 때 적용되며, $c_i = c$의 조건일 경우 상수 값 $\beta$로 지정되어 식의 분모가 0이 되는 것을 방지한다.

<17> 위 [수학식 1]에 의하면, $d^{\text{int}}$가 높은 값을 가질수록 동일 클래스 내 객체간의 거리비율이 높음을 나타내고 이는 동일 클래스에 포함된 객체간의 유사성이 떨어지므로 엔트로피가 증가함을 의미한다.

<18> [수학식 2]에 의하면, $d^{ext}$가 높은 값을 가질수록 한 클래스와 다른 클래스에 포함된 각 객체간의 거리비율이 높다는 것을 나타내고, 이는 다른 클래스간의 차별성을 높여준다. 그 결과 엔트로피는 감소된다.

<19> 주지된 바와 같이 엔트로피는 '데이터의 무질서 정도'를 나타내는 것으로, 여기서는 기존 ID3의 엔트로피 수학식에 비율 ($d^{\text{int}}/d^{ext}$)를 반영한다.

<20> $$Entropy_s(A) = -\sum_{i=1}^{n} \frac{d_i^{\text{int}}}{d_i^{ext}} P_i \log_2 P_i$$ ..... [수학식 3]

<21> 여기서, $A$는 데이터의 어느 한 속성(attribute)을 의미하며, $Entropy_s(A)$는 속성 $A$를 데이터 분류 기준으로 삼았을 때 데이터의 무질서 정도를 의미한다.

<22> 위 [수학식 3]을 살피면, $d^{ext}$가 작고 $d^{\text{int}}$가 클수록 엔트로피 값이 높아지고 반대로 $d^{ext}$가 크고 $d^{\text{int}}$가 작을수록 엔트로피 값이 작아짐을 알 수 있다.

<23> 한편, 어느 한 속성을 선택하였을 때 기대되는 엔트로피 감소율을 나타내는 정보이득률은, 공간이 고려된 엔트로피를 이용하여 다음과 같이 표현된다.

<24> $$Gain_s(GA, SA) = Entropy_s(GA) - \sum_{v \in Values(SA)} \frac{|GA_v|}{|GA|} Entropy_s(GA_v)$$ ..... [수학식 4]

<25> 여기서 $GA$는 'Supporting Attribute'($SA$)를 분류 속성으로 선택하였을 때 전체 엔트로피의 기대 감소량이며, 각 속성을 $SA$로 선택하였을 때 $GA$ 값이 가장 큰 $SA$가 전체 엔트로피를 가장 많이 감소시키는 속성인 것을 의미한다. 또한, $GA_v$는 $GA$의 부분 집합으로 $SA$ 속성이 $v$라는 상수 값을 가지고 있는 것을 의미한다. 예를 들어, $SA$가 강남 지역 사람의 월급이라는 속성이라 상정하면, '$GA_{v=50,000,000}$'는 월급이 5,000만원인 사람들을 나타낸다. 그리고 $|GA|$와 $|GA_v|$는 각각 $GA$와 $GA_v$의 집합원 개수(cardinality)를 의미한다.

<26> 이러한 [수학식 4]에 의하면, 공간데이터의 거리를 고려하여 엔트로피를 계산한 경우, 비공간 속성뿐만 아니라 각 객체의 공간적인 속성까지도 고려하게 되는바 공간데이터 사이에 존재하는 소정의 관계를 검색·도출하는데

유용하다.

<27> 상기한 연구는 의사결정 트리 구축에 있어서 공간 차원의 데이터의 거리를 반영한다는 점에 의의가 있다.

## 발명이 이루고자 하는 기술적 과제

<28> 그러나 두 객체간의 거리가 유클리디안 거리법을 따르기 때문에 공간 객체의 분포와 그에 따른 관계를 설명하는데 한계점이 있고, 실생활에 적합하지 않다는 문제점이 있다.

<29> 따라서 본 발명은 공간 차원의 데이터 분포도를 반영하여 비공간적으로 분류된 객체들이 공간 차원에서 어떠한 연관관계가 있는지 추정할 수 있는 의사결정 트리 구축 방법을 제공한다.

<30> 본 발명의 특징 및 이점들은 첨부도면에 의거한 다음의 상세한 설명으로 더욱 명백해질 것이다. 이에 앞서 본 발명에 관련된 공지 기능 및 그 구성에 대한 구체적인 설명이 본 발명의 요지를 불필요하게 흐릴 수 있다고 판단되는 경우에는, 그 구체적인 설명을 생략하였음에 유의해야 할 것이다.

## 발명의 구성 및 작용

<31> 우선, 첨부도면 도 1은 공간데이터의 밀집도에 따른 의미를 설명하기 위한 예시도이다. 앞서 설명한 유클리디안 거리법에 의하면, 중심 지점에 대해 가장 인접한 '대형마트A' 및 '대형마트B'의 거리 A와 거리 B는 그 의미가 동일하다. 그러나 '대형마트A'와 '대형마트B'가 도면과 같이 분포되어 있다면, 상기 거리 A와 거리 B는 다른 의미를 갖는다. 즉, 같은 거리(A=B)라 할지라도 거리 A쪽의 '대형마트A'보다 거리 B쪽에 '대형마트B'가 군집해 있다는 공간적 특성이 있다. 신규 오픈할 대형마트의 입지선정을 예로 들면, 선정한 지점으로부터 인접한 기존 대형마트의 거리 및 개수가 매출에 큰 영향을 주게 될 것임은 자명하다. 기존의 유클리디안 거리법을 이용하여 입지선정을 한다면 상기한 바와 같이 공간적 특성을 반영할 수 없게 된다.

<32> 이와 같이 공간적 위치의 고려가 필수적인 안건에 대한 의사결정을 내릴 때는, 객체의 위치 및 거리뿐만 아니라 공간적 특성, 다시 말해 분산도와 밀집도 등 다른 공간 객체와의 상관관계(연관관계)에 대한 정보까지 감안하는 것이 중요하다.

<33> 도 2는 공간데이터의 분포도에 대한 다양한 예를 보인 예시도이며, 유클리디안 거리법에 의한 정보만으로는 다양한 분포도(distribution)에 따른 공간적 특성을 도출하기 용이치 않음을 알 수 있다.

<34> 도 3은 공간데이터 분포에 따른 여러 가지 특성을 설명하기 위한 예시도이다. 상기 분산도(dispersion)는 동일 클래스로 분류된 객체가 서로 공간적으로 얼마나 인접한 거리에 있는지, 중심점으로부터 얼마나 조밀하게 모여 있는지를 반영하는 척도이다. 동일 클래스에 속한 객체(공간 객체)가 중심점으로부터 가까운 거리에 조밀하게 모여 있을수록 작은 분산도 값을 갖는다. 분산도 값이 작다는 것은 비슷한 성질 또는 유사한 성질의 객체가 공간적으로 밀집(군집)되어 있음을 의미하므로, 엔트로피를 감소시키는 효과가 있다.

<35> 도 3의 [A]와 [B]를 비교해 보면, 양자 모두 동일한 중심점을 가지고 있으나 분포도가 상이하다. 이러한 경우 앞서 언급한 바와 같이 유클리디안 거리법으로는 공간데이터의 분포도를 반영하지 못한다. 따라서 본 발명은 공간데이터의 분포도를 반영하기 위해 객체간의 거리 계산에 데이터에 대한 표준편차 행렬을 곱한다. 이러한 기본 개념은 두 점 $(X_1,Y_1)$, $(X_2,Y_2)$ 간의 거리를 산출하는 다음의 수학식으로 설명된다.

<36> $$\text{Distance} = \sqrt{(X_2 - X_1, Y_2 - Y_1)^T S (X_2 - X_1, Y_2 - Y_1)} \quad \cdots\cdots\cdots\cdots\cdots \text{[수학식 5]}$$

<37> 여기서, $S$는 표준편차 행렬을 의미하며 $S = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yy} & \sigma_{yy} \end{bmatrix}$로 정의된다. 표준편차 행렬 $S$의 각 요소는 다음과 같이 정의된다.

<38> $$\sigma_{xx} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}, \quad \sigma_{yy} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \quad \sigma_{xy} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}$$

<39> 이하, 상기한 바를 토대로 공간데이터 분포를 고려한 공간 엔트로피 기반의 의사결정 트리 구축방법을 살펴본다. 의사결정 트리 구축에 대한 기본 과정에 대해서는 다수의 공지된 논문에서 찾아 볼 수 있으므로 생략키로 하며, 트리 구축에 필수적인 과정, 즉 '공간 엔트로피 산출 과정'만을 중점으로 설명한다.

<40> 본 발명의 공간 엔트로피 산출은 다음의 [수학식 6]으로 표현된다.

<41> $$Entropy_s(A) = -\sum_{i=1}^{n} \frac{\text{차별도}}{\text{분산도}} P_i \log_2 P_i$$ ..... [수학식 6]

<42> 위 엔트로피 산출 수학식의 특징은 차별도(dissimilarity)와 분산도(dispersion)에 대한 비율이 적용된다는 점에 있다.

<43> 먼저, 상기 분산도는 클래스의 중심점과 동일 클래스에 속해 있는 객체간의 거리 비율을 의미한다. 보다 구체적으로 클래스 내에서 객체와 중심점간의 거리뿐만 아니라 분산까지 알아내기 위해서는 앞서 언급한 공간데이터 분포를 고려한 거리를 이용한다. 첨언하면, 동일 거리라도 조밀하게 모여 있는 곳의 거리는 작게, 산만하게 흩어져 있는 곳의 거리는 크게 만들어서 그 분산도를 반영하는 것이다.

<44> 다음의 수학식에서 $M_i$는 클래스 $i$의 중심점이고, $S$는 표준편차 행렬(수학식 5 참조), $X$는 지정된 좌표 값의 위치벡터, $|c_i|$는 클래스 $i$에 속하는 객체의 개수이다.

<45> $$dispersion = \frac{1}{|c_i|} \sum_{X \in c_i} \sqrt{(X - M_i)^T S (X - M_i)}$$ ..... [수학식 7]

<46> 위 [수학식 7]을 살피면, 클래스 $i$의 분산도(dispersion)는 클래스의 중심점($M_i$)과 다른 객체간의 분산 정도를 고려한 거리비율로 산출되며, 이는 공간 차원에서 중요한 의미를 갖는 공간데이터의 분포도(distribution)를 반영하고 있다.

<47> 첨부도면 도 4는 공간데이터 분포에 따른 차별도를 설명하기 위한 예시도이다. 여기서 차별도(dissimilarity)란 다른 클래스에 속한 객체들이 공간상에서 얼마나 특징적으로 군집화되어 있는지를 나타내는 척도를 말한다. 다른 클래스에 속한 객체들이 거리상으로 멀리 이격되어 있고 공간데이터의 퍼짐 정도에 따른 경계가 뚜렷할 경우, 클래스간의 차별도가 높은 것이라 이해할 수 있다.

<48> 상기 차별도는 각 클래스의 중심점과 분산을 산출하고, 각 클래스의 중심점을 분산과 데이터 개수를 통해 Z-정규분포를 따르도록 표준화시킨 후, 거리비교를 수행함으로써 나타난다. 보다 구체적으로, 각 클래스의 중심점을 분산행렬을 이용하여 표준화시킨다. 분류된 두 클래스는 데이터의 개수가 각각 $n_1$, $n_2$이고, 중심점 및 표준편차가 각각 $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$라고 하면, 클래스 중심점의 표준화 좌표 $Z_t$는 다음과 같이 산출된다.

<49> $$Z_t = \frac{(X - Y) - (\bar{X} - \bar{Y})}{\sqrt{\sigma_1^2 / n_1 + \sigma_2^2 / n_2}}$$ ..... [수학식 8]

<50> 중심점과 분포도가 다른 클래스 A, B를 표준화시키면, 각 클래스는 $Z \sim N(0,1)$인 표준정규분포를 따르게 되고, 이것은 동등한 비교기준을 제시해 주기 때문에 두 클래스에 포함된 점들에 대한 연산, 비교를 가능케 한다.

<51> 상기 [수학식 8]에 의해 Z-정규분포를 따르도록 표준화된 클래스의 중심점은 다음의 [수학식 9]에 반영되어, 차별도(dissimilarity)가 산출된다.

<52> $$dissimilarity = \frac{1}{|Z_t| \times (|Z_t| - 1)} \sum_{X \in Z_t} \sum_{X' \in Z_t, X \neq X'} \sqrt{dist(X, X')}$$ ..... [수학식 9]

<53> 만약, 각각 n, k, m개의 데이터를 가지고 있는 3개의 클래스에 대한 차별도를 계산한다면, 단지 3번의 거리 계산만이 요구된다. 유클리디안 거리법과 비교했을 때 표준편차 행렬을 구해야 하는 추가적인 비용이 소요되지만, 계산 복잡도 측면에서는 그 비용이 적기 때문에 결과적으로 비용 절감을 이룰 수 있다.

## 발명의 효과

<54> 상술한 본 발명에 의하면, 공간 차원의 데이터 분포도를 반영하여 비공간적으로 분류된 객체들이 공간 차원에서 어떠한 연관관계(상관관계)가 있는지 추정할 수 있는 주요 효과가 있다.

<55> 또한, 본 발명은 데이터의 거리와 퍼짐 정도를 함께 고려하여 객체들의 공간적 특성을 반영한 엔트로피를 산출하므로, 공간적 및 비공간적으로 신뢰성 있는 의사결정 트리를 구축할 수 있다.

<56> 또한, 거리를 산출할 때 객체간의 거리가 아니라 각 클래스를 대표하는 중심점으로부터의 거리를 산출하기 때문

에 전체적인 계산 비용을 절감할 수 있다.

<57> 이상으로 본 발명의 기술적 사상을 예시하기 위한 바람직한 실시예와 관련하여 설명하고 도시하였지만, 본 발명은 이와 같이 도시되고 설명된 그대로의 구성 및 작용에만 국한되는 것이 아니며, 기술적 사상의 범주를 일탈함이 없이 본 발명에 대해 다수의 변경 및 수정이 가능함을 당업자들은 잘 이해할 수 있을 것이다. 따라서, 그러한 모든 적절한 변경 및 수정과 균등물들도 본 발명의 범위에 속하는 것으로 간주되어야 할 것이다.

## 도면의 간단한 설명

<1> 도 1은 공간데이터의 밀집도에 따른 의미를 설명하기 위한 예시도,

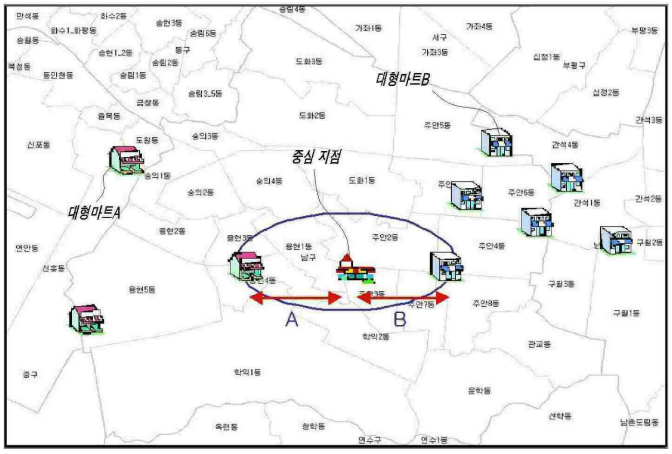
<2> 도 2는 공간데이터의 분포도에 대한 다양한 예를 보인 예시도,

<3> 도 3은 공간데이터 분포에 따른 여러 가지 특성을 설명하기 위한 예시도,

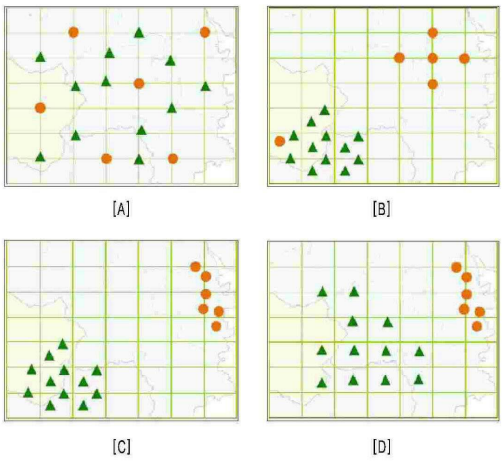
<4> 도 4는 공간데이터 분포에 따른 차별도를 설명하기 위한 예시도.

## 도면

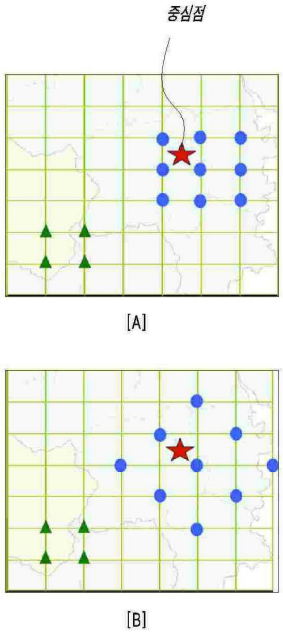
### 도면1




### 도면2



도면3




도면4

