http://dx.doi.org/10.7848/ksgpc.2013.31.1.23

이종의 공간 데이터 셋에서 매칭 객체 판별을 위한 임계값 산출

Calculation of a Threshold for Decision of Similar Features in Different Spatial Data Sets

김지영[1] · 허 용[2] · 유기윤[3] · 김정옥[4]
Kim, Jiyoung · Huh, Yong · Yu, Kiyun · Kim, Jung Ok

Abstract

The process of a feature matching for two different spatial data sets is similar to the process of classification as a binary class such as matching or non-matching. In this paper, we calculated a threshold by applying an equal error rate (EER) which is widely used in biometrics that classification is a main topic into spatial data sets. In a process of discriminating what's a matching or what's not, a precision and a recall is changed and a trade-off appears between these indexes because the number of matching pairs is changed when a threshold is changed progressively. This trade-off point is EER, that is, threshold. To the result of applying this method into training data, a threshold is estimated at 0.802 of a value of shape similarity. By applying the estimated threshold into test data, F-measure that is a evaluation index of matching method is highly value, 0.940. Therefore we confirmed that an accurate threshold is calculated by EER without person intervention and this is appropriate to matching different spatial data sets.

Keywords : Threshold, Equal Error Rate, Matching, Classification

초 록

이종의 공간 데이터 셋을 매칭하는 과정은 매칭 또는 비 매칭의 이진 클래스로 판별하는 과정과 비슷하다. 이에 이진 클래스의 판별이 중요한 연구주제인 바이오인식 분야에서 임계값을 구하는데 이용되는 동일 오류율을 공간 데이터 셋의 매칭에 적용하여 임계값을 산출하였다. 매칭유무를 판별하는 과정에서 임계값이 계속 바뀌면 매칭으로 판별되는 객체 쌍이 상이해지면서 정확도와 재현율도 바뀌게 되며, 이들 지표 사이에 trade-off가 나타나는 지점이 EER, 즉 임계값이 된다. 동일 오류율 기반의 임계값 산출 방법을 훈련 자료에 적용하여 형상유사도 0.802가 임계값으로 구해졌다. 이를 실험 자료에 적용한 결과, 매칭의 성능을 평가하는 척도인 F-measure가 0.940으로 높게 나타났다. 이를 통하여 동일 오류율을 이용하여 연구자의 개입이 없이 정확한 임계값이 산출되고, 동일 오류율 기반의 임계값 산출이 이종의 공간 데이터 셋 매칭에 적합하다는 것을 알 수 있었다.

핵심어 : 임계값, 동일 오류율, 매칭, 판별

1. 서 론

우리나라는 NGIS(National Geographic Information System) 사업을 통하여 정부 및 지방자치단체 등에 다양한 지리정보시스템이 구축되어 있으며, 생활 깊숙이 들어온 차량용 내비게이션과 구글로 부터 시작된 지도 서비스 등 위치기반서비스의 확대로 공간정보에 대한 사용자의 요구가 증가하고 있다. 따라서 최근에는 NGIS 사업을 통하여 구축된 이종의 공간 데이터 셋을 통합하여 지리정보시스템관련 프로젝트에서 70% 이상을 차지하는 공간정보의 구축비용을 절감하

1) 정회원 · 서울대학교 대학원 공과대학 건설환경공학부 박사과정 (E-mail : soodaq@snu.ac.kr)
2) 정회원 · 서울대학교 공학연구소 선임연구원 공학박사 (E-mail : huhyong78@gmail.com)
3) 정회원 · 서울대학교 공과대학 건설환경공학부 교수 (E-mail : kiyun@snu.ac.kr)
4) 교신저자 · 정회원 · 서울대학교 공학연구소 선임연구원 공학박사 (E-mail : geostar1@snu.ac.kr)

고, 최신의 공간데이터를 확보하려는 움직임이 일고 있다. 특히 위치기반서비스에서 중요한 관심정보(Point of Interests)와 연관된 건물 정보를 융합하려는 연구가 국내외에서 꾸준히 연구되고 있다(Bel Hadj Ali, 2001; Huh and Yu, 2012; Kim, 2010; Kim et al., 2011a; Kim et al., 2011b; Moon et al., 2011; Qi et al., 2010; Samal et al., 2004).

이종의 공간데이터 셋의 매칭을 수행하는 선행연구를 보면 두 데이터 셋에서 매칭이 될 수 있는 매칭 후보객체 쌍을 찾은 후 이 후보객체 쌍의 유사한 정도를 유사도나 비유사도로 표현하고 그 값이 일정 값 이상인 경우 매칭 후보객체 쌍은 매칭으로 판별한다. 이 과정에서 매칭유무를 판별 시 사용되는 일정 값 즉, 임계값은 대부분의 선행연구에서 훈련 자료에서 학습을 통하여 연구자가 그 값을 정하였다(Bel Hadj Ali, 2001; Huh and Yu, 2012; Kim et al., 2011a; Kim, 2010; Qi et al., 2010; Samal et al., 2004). 그러나 이종의 공간데이터 셋의 매칭에서 적용되는 이 방법은 연구자의 개입이 요구되며, 그로 인하여 대용량의 공간데이터 셋의 매칭을 수행할 때 수행시간이 더 소요되는 한계가 있다. Kim et al.(2011b)은 훈련 자료에서 임계값을 산정하기 위하여 치우친 분포(skewed distribution)로 나타나는 매칭 객체 쌍과 비 매칭 객체 쌍의 조정된 상자도표(adjusted box plot)를 적용하여 각각의 분포에서 각 펜스 밖의 영역 중에서 공통인 영역을 특이값(outlier)이 관찰되는 구간으로 정의하였다. 이때, 해당 공통 특이값 구간에서 매칭 객체 쌍의 빈도가 비 매칭 객체 쌍의 빈도보다 낮아지는 구간을 찾고, 해당 구간의 평균을 임계값으로 설정하였다. 그러나 히스토그램은 구간의 범위에 따라 그 분포가 상이하므로 매칭 객체 쌍의 빈도가 비 매칭 객체 쌍의 빈도보다 낮아지는 구간이 공통 특이값 구간에 존재하지 않을 수도 있으며, 이런 구간이 하나의 히스토그램에서 여러 번 나타날 수도 있다(Kim et al., 2013).

따라서 본 연구에서는 바이오인식 분야의 연구자의 개입이 없이 임계값을 산출하는 방법을 이종의 공간 데이터 셋의 매칭에 적용하고자 한다. 바이오인식 분야에서는 지문, 홍채, 안면 등 2개 이상의 생체 인식 시스템에서 측정된 매칭 값(matching score)을 정규화(normalization)하고, 정규화된 매칭 값을 단일 매칭 값으로 결합(score fusion)한 후 임계값을 통하여 본인 여부를 판별한다(Snelick et al., 2005). 앞서 살펴본 공간정보 분야의 선행 연구에서도 이와 비슷한 과정으로 이종의 공간 데이터 셋의 매칭을 수행한다. 이종의 공간 데이터 셋을 중첩하여 교차하는 면 객체를 매칭 후보객체 쌍으로 추출한다. 이들 매칭 후보객체 쌍의 유사 정도를 측정하기 위하여 면 객체의 기하학적 성질이 반영된 매칭 기준으로부터 매칭 후보객체 쌍의 거리를 산출하고, 매칭 기준별로 산출된 거리는 그 값의 최대값으로 정규화하여 유사도를 산출한다. 마지막으로 매칭 기준별 유사도를 결합하여 통합 유사도를 생성하고, 임계값을 통하여 매칭유무를 판별한다. 이에 바이오인식 분야의 임계값 결정 방법인 동일 오류율(Equal Error Rate, EER)을 Kim et al.(2011b)이 제시한 매칭 기법에 적용해 봄으로써 이종의 공간 데이터 셋 매칭에 적용 가능한지 여부를 평가하였다. 이를 통하여 매칭에서 연구자의 개입을 줄이고, 보다 정확하고 객관적인 임계값이 산정되기를 기대한다.

2. EER 기반의 임계값 산출 방법

이종의 데이터 셋에서 매칭인 객체를 탐지하는 것은 이진 클래스로 판별하는 문제와 비슷한 과정을 거친다. 그림 1과 같이 훈련 자료를 이용하여 판별모델을 세우고, 이를 실험 자료에 적용함으로써 판별모델의 성능을 평가한다. 즉, 이종의 데이터 셋을 매칭하는 것은 매칭 후보 객체들을 '매칭이다(매칭)' 또는 '매칭이 아니다(비 매칭)'의 2개의 클래스로 판별하는 것으로, 훈련 자료를 이용하여 매칭을 위한 기준(criterion)을 세우고 이들 매칭 기준을 융합하여 유사도를 산출하고 이 유사도를 임계값을 이용하여 판별하는 단계가 요구된다. 이렇

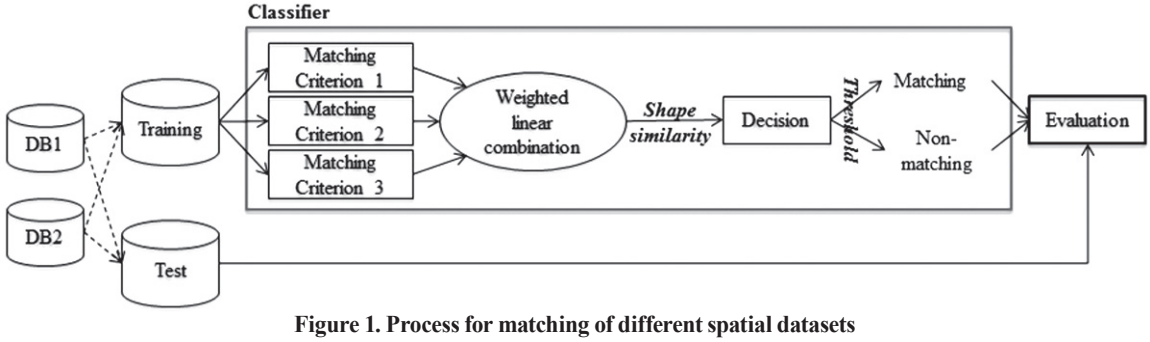


Figure 1. Process for matching of different spatial datasets

게 도출된 매칭 기법, 즉 분류모델을 실험 자료에 적용하여 매칭의 정확도를 평가하는 것이 이종의 공간 데이터 셋의 매칭 과정이다. 이를 수식으로 표현하면 매칭 후보 객체들의 유사정도를 나타내는 기준이 x_i고, 클래스가 $y_i(\in\{0,1\})$인 훈련 자료 집합(x_i, y_i)에서 유사도 함수($f(\bullet)$)와 임계값(θ)을 결정하고, 이를 이용하여 모든 i에 대하여 $f(x_i) > \theta$인 경우 $y_i = 1$, $f(x_i) \leq \theta$인 경우 $y_i = 0$으로 판별하는 과정이다. 결과적으로 이종의 데이터 셋의 매칭을 위해서는 유사도 함수와 임계값이 요구된다.

2.1 유사도 함수

본 연구에서는 매칭 후보 객체 쌍의 유사도 함수는 Kim et al.(2011b)이 제안한 매칭 기법을 적용하였다. 이 기법은 이종의 면 객체 데이터 셋에 대하여 위치 기준, 형상 기준, 면적 기준의 세 가지 매칭 기준을 정의하고, 매칭 기준 각각의 표준편차와 기준들 간의 상관관계를 고려하여 가중치를 자동으로 산출하였다. 최종 유사도 함수는 식 (1)과 같이 매칭 기준과 가중치의 가중 선형 조합으로 구성되며, 이것을 형상유사도($Sim_S(A,B)$)라 한다.

$$Sim_S(A,B) = \omega_1 \times C_P(A,B) + \omega_2 \times C_S(A,B) + \omega_3 \times C_A(A,B) \quad (1)$$

식 (1)에서 A는 참조 데이터 셋의 매칭 후보객체, B는 목표 데이터 셋의 매칭 후보객체로, 위치 기준($C_P(A,B)$), 형상 기준($C_S(A,B)$), 면적 기준($C_A(A,B)$)의 세 가지 매칭 기준은 순서대로 식 (2)부터 식 (4), 가중치(ω_1, ω_2, ω_3)는 식 (5)로 유도된다.

$$C_P(A,B) = 1 - \frac{D_d(P_A, P_B)}{max(D_d(P_{A_{all}}, P_{B_{all}}))} \quad (2)$$

여기서, $D_d(P_A, P_B) = \sqrt{(X_A - X_B)^2 + (Y_A - Y_B)^2}$

$$C_S(A,B) = 1 - \frac{D_s(A,B)}{max(D_s(A_{all}, B_{all}))} \quad (3)$$

여기서, $D_s(A,B) = \left| \frac{perimeter(A)}{2\sqrt{\pi \times area(A)}} - \frac{perimeter(B)}{2\sqrt{\pi \times area(B)}} \right|$

$$C_A(A,B) = 1 - \frac{D_{ov}(A,B)}{max(D_{ov}(A_{all}, B_{all}))} \quad (4)$$

여기서, $D_{ov}(A,B) = \left| \frac{area(A \cup B) - area(A \cap B)}{area(A) + area(B)} \right|$

$$\omega_j = \frac{C_j}{\sum_{k=1}^{m} C_k} \quad (5)$$

여기서, $C_j = \sigma_j \times \sum_{k=1} (1 - r_{jk})$

C_j : 매칭 기준별 정보량, σ_j : 매칭 기준별 표준편차, r_{jk} : 매칭 기준간 상관관계

2.2 임계값 산출

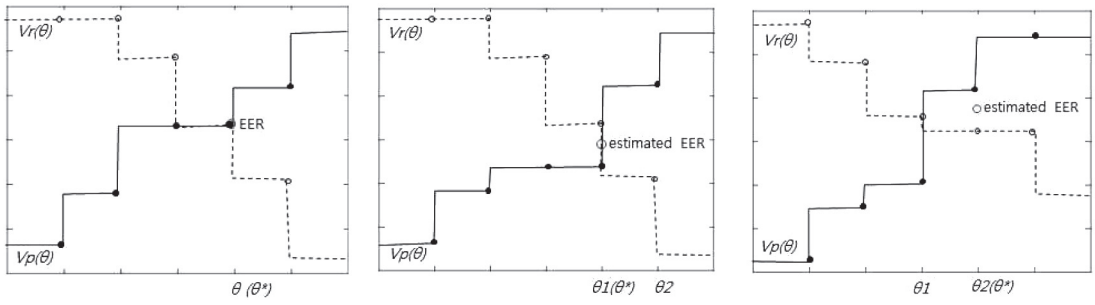
유사도 함수를 이용하여 매칭 후보 객체 쌍의 형상유사도가 산출되면 일정 임계값 이상인 매칭 후보 객체 쌍의 매칭유무를 판별하기 위하여 임계값이 요구된다. 이를 위해서 바이오인식 분야에서 일반적으로 사용되는 EER을 적용하여 임계값을 산출한다.

이진 클래스의 판별 문제에서 판별모델의 성능을 평가하기 위하여 다양한 지수(index)가 사용된다(Han et al., 2011). 그러나 이종의 공간데이터 셋의 매칭과 같이 클래스의 분포가 한쪽으로 꼬리가 긴 분포인 경우 기존의 지표들 중에서 문헌정보나 온톨로지 매칭 분야에서 사용되는 정확도(precision)와 재현율(recall)이 적합하다(Davis and Goadrich, 2006). 식 (6)으로 계산되는 정확도(V_p)는 판별모델을 이용하여 매칭이라고 예측된 객체 쌍 중에서 실제 매칭인 객체 쌍의 비율을 의미하고, 식 (7)의 재현율(V_r)은 실제 매칭인 객체 쌍 중 얼마나 많은 객체 쌍이 제안된 판별모델로 매칭인 객체 쌍으로 예측되어졌는지를 의미한다.

$$V_p = \frac{num(\text{예측된 매칭 객체 쌍 중 실제 매칭 객체쌍})}{num(\text{예측된 매칭 객체쌍})} \quad (6)$$

$$V_r = \frac{num(\text{예측된 매칭 객체 쌍 중 실제 매칭 객체쌍})}{num(\text{예측된 매칭 객체쌍})} \quad (7)$$

매칭유무를 판별하는 과정에서 임계값이 계속 바뀌면 매칭으로 판별되는 객체 쌍이 상이해지면서 정확도와 재현율도 바뀌게 되며, 이들 지표 사이에 trade-off가 나타나게 된다. 그림 2(a)와 같이 두 지표가 trade-off되는 지점, 즉 정확도와 재현율이 서로 같아지는 지점이 EER이고, 일반적으로 해당 지점을 임계값으로 산정한다(Bengio et al., 2005). 그러나 실제 데이터에서는 그림 2(b)나 (c)와 같이 정확도와 재현율의 분포가 연속적이거나 trade-off되는 지점이 존재하지 않는 경우가 있을 수 있다. 따라서 EER, 즉 임계값(θ)은 식 (8)로 구해진다.



(a) EER exits (b) EER is estimated at θ_1 (c) EER is estimated at θ_2

Figure 2. Calculation of a threshold in Precision-Recall graph

$$\theta^* = \begin{cases} \arg min_\theta |V_p(\theta) - V_r(\theta)| & \text{if } V_p(\theta) = V_r(\theta) \\ \dfrac{V_p(\theta_1) + V_r(\theta_1)}{2} & \text{if } V_r(\theta_1) - V_p(\theta_1) \le V_p(\theta_2) - V_r(\theta_2) \\ \dfrac{V_p(\theta_2) + V_r(\theta_2)}{2} & otherwise \end{cases} \quad (8)$$

여기서, $\theta_1 = \max_\theta(V_p(\theta) \le V_r(\theta))$, $\theta_2 = \min_\theta(V_p(\theta) \ge V_r(\theta))$

3. 적용 및 평가

3.1 실험 방법 및 자료

2절에서 설명한 유사도 함수로 훈련 자료에서 형상유사도를 구하고, EER 기반으로 임계값을 산출한다. 다음으로, EER을 이용하여 도출된 임계값을 실험 자료에 적용해 봄으로써 성능을 평가한다. 이때 EER 기반으로 산출된 임계값을 적용하여 이종의 공간 데이터 셋의 매칭을 수행한 분류모델은 F-measure로 성능이 평가된다(Yatskevich et al., 2006). F-measure는 정확도(V_p)와 재현율(V_r)을 같은 가중치로 두고 통합한 값으로 식 (9)로 유도되고, 그 값이 클수록 더 좋은 매칭결과를 나타낸다.

$$F-measure = \frac{V_p \times V_r}{0.5 \times V_p + 0.5 \times V_r} \quad (9)$$

본 연구에서는 매달 갱신되는 도로명주소기본도의 건물과 2년 주기로 갱신되는 축척이 1:5,000인 수치지도2.0 건물 데이터 셋을 매칭하기 위하여 2007년에 갱신된 수치지도2.0 37709094도엽의 일부를 참조 자료로 하고, 해당 도엽과 동일한 영역의 2010년 8월 갱신된 도로명주소기본도를 목표 자료로 하여 매칭인 객체 쌍과 비 매칭인 객체 쌍을 수동으로 추출하여 훈련 자료를 구축하였다. 그 결과 매칭 객체 쌍 451개와 비 매칭 객체 쌍 132개가 훈련 자료로 사용되었다. 실험 자료는 2007년에 갱신된 수치지도2.0 37709081도엽 1,051개의 건물 면 객체와 동일한 영역의 2010년 8월 갱신된 도로명주소기본도 1,156개의 건물 면 객체가 이용되었다.

3.2 실험 결과 및 평가

훈련 자료의 매칭 객체 쌍과 비 매칭 객체 쌍의 형상유사도를 구하면 그 분포가 그림 3과 같으며, 매칭 객체 쌍과 비 매칭 객체 쌍의 형상유사도의 평균은 순서대로 0.908, 0.691이고,

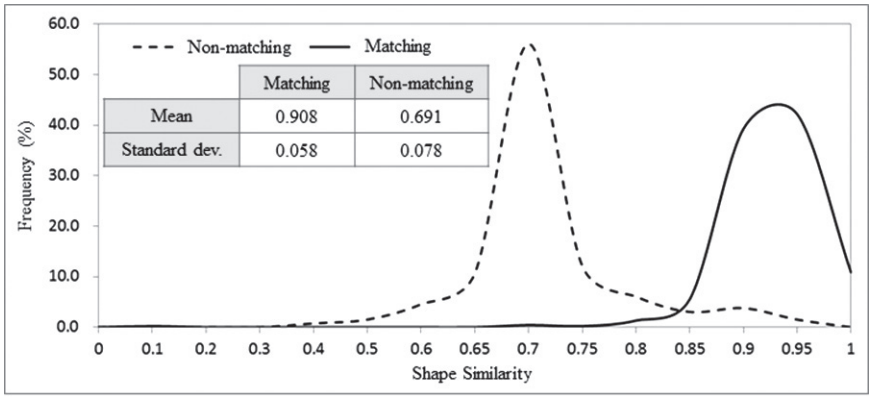


Figure 3. Histogram of a shape similarity in training data

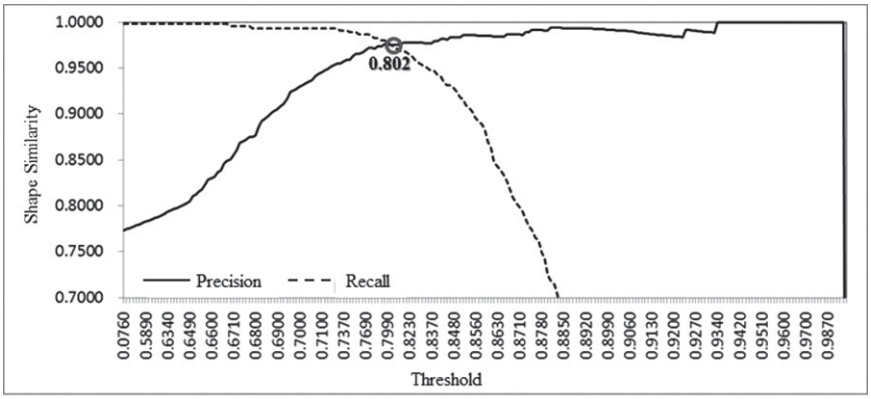


Figure 4. Precision-Recall graph to calculate a threshold in training data

Table 1. Evaluation of a proposed matching method

Predicted matching pairs	Actual matching pairs	Actual matching pairs among predicted matching pairs	Precision	Recall	F-measure
842	846	793	0.942	0.937	0.940

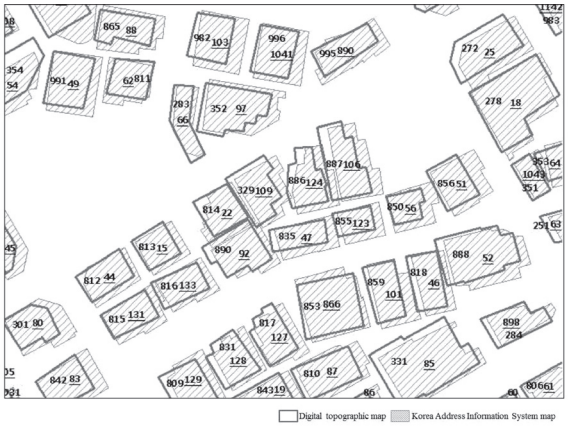


Figure 5. Actual matching pairs among predicted matching pairs in test data

표준편차는 0.058, 0.078로 나타났다.

이들 훈련 자료에서 EER 기반의 임계값을 산출한 결과, 정확도와 재현율의 변화는 그림 4와 같이 나타났으며, 그림 4에서 원으로 표시된 형상유사도 0.802에서 정확도와 재현율이 trade-off되었다. 즉 이 trade-off가 되는 지점이 EER로, 매칭 유무를 판별할 때 기준이 되는 임계값을 의미한다. 따라서 이종의 데이터 셋의 매칭과정에서 매칭 후보객체 쌍의 형상유사도가 0.802보다 크거나 같으면 해당 객체 쌍은 매칭, 그렇지 않은 경우는 비 매칭으로 판별될 수 있다.

본 연구에서 정확도 평가에 사용될 실제 매칭 쌍은 실험 자료에서 수동으로 표 1에 나타난 것과 같이 846쌍을 추출하였다. 훈련 자료에서 산정된 임계값 0.802를 적용하여 매칭으로 예측된 객체 쌍 중에서 실제 매칭 쌍은 793쌍이다. 이를 시각적으로 살펴본 결과, 그림 5와 같이 두 데이터 셋의 위치오차가 큰 경우라도 면 객체의 형상이 유사하거나 중복되는 면적이 넓은 객체가 매칭된 것을 알 수 있었다(그림 5에서 숫자는 수치지도2.0 인덱스, 밑줄 쳐진 숫자는 도로명주소기본도의 인덱스를 의미).

마지막으로 제안된 매칭 기법의 성능은 표 1에 나타난 것처럼 F-measure가 0.940으로 높게 나타났다. 이는 본 연구에서 적용된 EER 기반의 임계값 산출 방법으로 이종의 공간 데이터 셋 매칭을 수행한 결과, 정확한 매칭이 수행됨을 의미한다.

4. 결 론

이종의 공간 데이터 셋을 매칭은 매칭 또는 비 매칭의 이진 클래스로 매칭 후보 객체 쌍을 판별하는 과정과 유사하다. 이종의 공간 데이터 셋을 비교하기 위하여 매칭 기준을 정의하고, 이 매칭 기준들의 가중 선형 조합으로 유사도 함수를 정의한다. 유사도 함수로 산출된 형상유사도가 임계값 이상이면 매칭, 그렇지 않은 경우는 비 매칭으로 판별된다. 따라서 본 연구에서는 바이오인식 분야에서 본인과 침입자를 판별하는

데 적용되는 EER을 이용하여 임계값을 산정하였다. 훈련 자료에 이를 적용한 결과 정확도와 재현율이 trade-off되는 형상 유사도 0.802가 임계값으로 구해졌다. 유사도 함수와 구해진 임계값을 실험 자료에 적용하여 842쌍의 매칭 객체 쌍을 탐지하고, 이들 중 실제 매칭 객체 쌍과 일치하는 것이 793쌍이었으며, F-measure가 0.940으로 높게 나타났다. 따라서 이종의 공간 데이터 셋의 매칭에서 연구자의 개입이 없이 정확한 임계값이 산출된다고 판단되었다.

그러나 향후 다양한 이종의 공간 데이터 셋에 EER 기반의 임계값 산출을 적용하여 공간 데이터 셋의 매칭유무를 판별하는데 적합한지 면밀히 검토할 필요가 있을 것이다.

감사의 글

본 연구는 중소기업청에서 지원하는 2011년도 산학연공동기술개발사업 (No.00045395)의 연구수행으로 인한 결과물임을 밝힙니다. 또한 본 연구는 서울대학교 건설환경종합연구소의 연구비 지원으로 수행되었습니다.

References

Bel Hadj Ali, A. (2001), Positional and shape quality of areal entities in geographic databases: quality information aggregation versus measures classification, *Proceeding of ECSQARU''2001 Workshop on Spatio-Temporal Reasoning and Geographic Information Systems*, Toulouse, pp. 1–16.

Bengio, S., Maréthoz, J. and Keller, M. (2005), The expected performance curve, *Proceedings of the ICML'05 workshop on ROC analysis in machine learning*, Germany, pp. 43–50.

Davis, J. and Goadrich, M. (2006), The relationship between precision-recall and ROC curves, *Proceedings of the 23rd International Conference on Machine Learning*, USA, pp. 233–240.

Han, J., Kamber, M. and Pei, J. (2011), *Data Mining: Concepts and Techniques, Third Edition*, Morgan Kaufmann, USA, pp. 364–370.

Huh, Y. and Yu, K. (2012), Shape similarity measure for M:N areal object pairs using the Zernike moment descriptor, *Korean Journal of Geomatics*, Vol. 30, No. 2, pp. 153–162.

Kim, K., Huh, Y. and Yu, K. (2011a), Study on building data Set matching considering position error, *Korea Spatial Information Society*, Vol. 19, No. 2, pp. 37–46.

Kim, J. (2010), *Method of feature matching for geospatial datasets using the geographic context*, PhD dissertation, Seoul National University, Seoul, Korea, pp. 28–37.

Kim, J., Kim, D., Huh, Y. and Yu, K. (2011b), A new method for automatic areal feature matching based on shape similarity using CRITIC method, *Korean Journal of Geomatics*, Vol. 28, No. 2, pp. 113–121.

Kim, J., Kim, J., Yu, K. and Huh, Y. (2013), Evaluation of classifiers performance for areal feature matching, *Korean Journal of Geomatics* (Under review)

Moon, Y., Park, K. and Choi, S. (2011), The research of effectively matching method of building objects to register UFID, *The Korea Society For GeospatIal Information System*, Vol. 19, No. 2, pp. 75–83.

Qi, H. B., Li, Z. L. and Chen, J. (2010), Automated change detection for updating settlements at smaller-scale maps from updated larger-scale maps, *Journal of Spatial Science*, Taylor & Francis, Vol. 51, No.1, pp. 133–146.

Samal, A., Seth, S. and Cueto, K. (2004), A feature-based approach to conflation of geospatial sources, *International Journal of Geographical Information science*, Taylor & Francis, Vol. 18, No. 5, pp. 459–489.

Snelick, R., Uludag, U., Mink, A., Indovina, M. and Jain, A. (2005), Large scale evaluation of multimodal biometric authentication using state-of-the-art systems, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 3, pp. 450–455.

Yatskevich, M., Giunchiglia, F. and Avesani, P. (2006), A large scale dataset for the evaluation of matching systems, URL: http://eprints.biblio.unitn.it/1015/, *University of Trento*, Italia, (last date accessed: 7 February 2013).

(접수일 2013. 01. 09, 심사일 2013. 02. 04, 심사완료일 2013. 02. 15)