k-최근접 템플릿기반 다중 분류기 결합방법[1)]

민준기° 조성배
연세대학교 컴퓨터과학과
생체인식연구센터
loomlike@sclab.yonsei.ac.kr sbcho@cs.yonsei.ac.kr

Multiple Classifier Fusion Method based on k-Nearest Templates

Jun-Ki Min° Sung-Bae Cho
Department of Computer Science, Yonsei University
Biometrics Engineering Research Center

다중 분류기의 결합은 높고 안정적인 분류성능을 얻기 위한 방법으로 패턴인식 분야에서 많이 연구되어 왔다[1]. 이 중 결정템플릿은 클래스별 학습샘플에 대한 분류기의 출력 벡터들의 중심을 해당 클래스의 분류모형으로 사용하는 방법으로, 높은 분류성능을 보이며 분류기 선택방법과 혼합되어 사용되기도 하였다[2]. 그러나 이 방법은 클래스를 하나의 템플릿으로 모델링하기 때문에 데이터의 특징을 정교하게 표현하는데 어려움이 있다. 이를 해결하기 위해 클러스터링 알고리즘을 이용하여 여러 개의 국부화된 템플릿을 생성하는 다중결정템플릿 방법이 제안되었지만, 이는 하나의 템플릿만을 참조함으로써 클러스터링 결과에 민감할 수 있다[3].

본 논문에서는 다중 분류기를 효과적으로 결합하기 위하여 클래스를 여러 개의 하위클래스로 분해한 뒤 k개의 템플릿을 참조하여 분류성능과 안정성을 높인 k-최근접 템플릿방법을 제안한다. 이를 위해 먼저 개별 분류기를 학습하고, 동일한 학습 데이터로부터 얻은 분류기들의 출력 값을 식 (1)의 결정프로파일 $DP(x_i)$로 구성한다.

$$DP(x_i)=\begin{bmatrix} d_{1,1}(x_i) & \cdots & d_{1,M}(x_i) \\ \vdots & d_{y,z}(x_i) & \vdots \\ d_{L,1}(x_i) & \cdots & d_{L,M}(x_i) \end{bmatrix} \quad (1)$$

여기에서 M과 L은 각각 클래스 수와 분류기 수를 의미한다. 각 클래스의 결정프로파일은 C-Means 알고리즘[4]을 이용하여 클러스터링하고, 식 (2)와 같이 클래스 m의 c번째 클러스터의 지역화된 템플릿 $DT_{m,c}$를 계산한다. 식에서 $u_{m,c,i}$는 샘플 x_i가 클래스 m의 c번째 클러스터에 소속되어있는 경우 1이고 그 외에는 0의 값을 갖는다. 본 논문에서는 클러스터의 수를 데이터 셋에 상관없이 C=20으로 고정시켰다.

$$DT_{m,c}=\begin{bmatrix} dt_{m,c}(1,1) & \cdots & dt_{m,c}(1,M) \\ \vdots & dt_{m,c}(y,z) & \vdots \\ dt_{m,c}(L,1) & \cdots & dt_{m,c}(L,M) \end{bmatrix}, \quad dt_{m,c}(y,z)=\sum_{i=1}^{n} u_{m,c,i} d_{y,z}(x_i) \Big/ \sum_{i=1}^{n} u_{m,c,i} \quad (2)$$

평가샘플의 분류는 샘플의 결정프로파일과 템플릿들 간의 유사도를 계산한 뒤, 가장 유사한 k개의 템플릿들 중 가장 많은 비율을 차지하는 레이블로 할당하는 것으로 수행한다. 본 논문에서는 유클리드거리식을 이용하여 유사도를 계산하였다. 이와같은 방법은 k최근접 이웃(k-Nearest Neighbor)분류기와 마찬가지로 k값에 영향을 받으며, 최적의 k값은 데이터 셋에 의존적이다. 따라서 제안하는 방법은 클래스 내 밀집도 IC와 클래스 간 분리도 IS를 식 (3)을 이용하여 분석하고[5], 식 (4)의 규칙을 이용하여 k값을 결정한다.

1) 본 연구는 생체인식연구센터(BERC)를 통해 한국과학재단(KOSEF)에서 지원받았음.

$$IC = E_1/E_M, \quad E_M = \sum_{i=1}^{n}\sum_{m=1}^{M} u_{m,i}\|x_i - z_m\|, \quad IS = \max_{i,j=1,\ldots,c}\|z_i - z_j\| \tag{3}$$

$$k = \begin{cases} 1 & \text{if } IC \le t_{IC} \text{ and } IS \le t_{IS} \\ C/2 & \text{if } IC > t_{IC} \text{ and } IS > t_{IS} \end{cases} \tag{4}$$

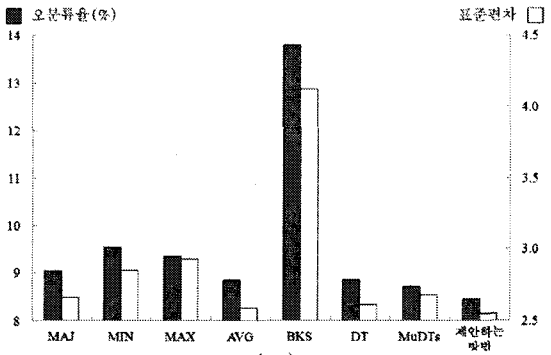
이때 임계값인 t_{IC}와t_{IS}는 실험을 통해 각각 1.5와 2로 결정하였다.

제안하는 방법의 검증을 위해서 UCI(http://mlearn.ics.uci.edu/MLRepository.html)와 ELENA (http://www.dice.ucl.ac.be/mlg/?page=Elena)데이터베이스 중 10개의 데이터 셋(Breast-cancer, Ionosphere, Iris, Satellite, Segmentation, Sonar, Phoneme, Texture, Clouds, Concentric)을 대상으로 실험하였으며, 25개의 신경망을 기본분류기로 사용하였다. 신경망의 학습율과 모멘텀은 각각 0.15와 0.9로 설정하였고, 내부 노드들 간의 초기 연결강도는 -0.5~0.5 사이의 임의 값으로 초기화하였으며, 각 신경망은 Bagging을 이용하여 학습하였다. 신경망의 은닉노드(Hidden node)수와 세대 수는 [1]의 연구와 동일한 기준으로 그림 1(ㄱ)과 같이 설정하였다. 제안하는 방법의 파라메터 k는 식 (4)에 의해 Ionosphere, Sonar, Phoneme, Clouds, Concentric의 5가지 데이터 셋의 경우 k=1, 나머지의 경우는 k=C/2=10으로 선택되었다. 비교 결합방법으로는 MAJ, MIN, MAX, AVG, 결정템플릿, 다중결정템플릿을 사용하였으며, 각 데이터 셋에 대해 10-fold cross validation실험을 수행하였다.

실험결과 제안하는 방법이 10개중 6개의 데이터 셋에서 최고 분류율을 보였으며, 그 외의 데이터 셋에서도 높은 성능을 나타냈다. 그림 1(ㄴ)은 전체 데이터 셋에 대한 평균 오분류율과 표준편차를 보여주는 것으로, 표준편차가 적을수록 성능이 안정적임을 나타낸다. 그림과 같이 제안하는 방법이 다른 방법에 비해 가장 높고 안정적인 성능을 보였으며, 결정템플릿(DT)과 평균 선택 방법(AVG)은 비슷한 성능을 보였다. 다중결정템플릿(MuDTs)의 경우 분류성능은 결정템플릿보다 좋았으나, 클러스터링 결과에 영향을 받기 때문에 안정성이 떨어지는 것을 확인하였다.

데이터 셋	샘플	특징	클래스	NN	
				은닉노드	세대
*Breast-cancer	683	9	2	5	20
*Ionosphere	351	34	2	10	40
*Iris	150	4	3	5	80
*Satellite	6435	36	6	15	30
*Segmentation	2310	19	7	15	20
*Sonar	208	60	2	10	60
+Phoneme	5404	5	2	5	30
+Texture	5500	40	11	20	40
+Clouds	5000	2	2	5	20
+Concentric	2500	2	2	5	20

(ㄱ)



(ㄴ)

그림 1. (ㄱ)실험에 사용된 데이터 셋과 신경망 파라미터(*UCI 데이터베이스, +ELENA 데이터베이스) (ㄴ)결합방법별 모든 데이터 셋에 대한 평균 오분류율과 표준편차

참고문헌

[1] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study," *J. Artificial Intelligence Research*, vol. 11, pp. 169-198, 1999.

[2] L.I. Kuncheva, "Switching between selection and fusion in combining classifiers: An experiment," *IEEE Trans. Systems, Man, and Cybernetics, Part B-Cybernetics*, vol. 32, no. 2, pp. 146-156, 2002.

[3] J.-K. Min, J.-H. Hong, and S.-B. Cho, "Fingerprint classification using multiple decision template with SVM," *J.KorealnformationScienceSociety*,vol.32,no.11,pp.1136-1146,2005.

[4] A.K. Jain and R.C. Dubes, *AlgorithmsforClusteringData*,PrenticeHall,1988.

[5] U. Maulik and S.Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650-1654, 2002.