Journal of The Korea Society of Computer and Information
Vol. 18, No. 11, November 2013
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2013.18.11.031

SVM을 위한 교사 랭크 정규화

이 수 종*, 허 경 용**

Supervised Rank Normalization for Support Vector Machines

Soojong Lee*, Gyeongyong Heo**

요 약

특징 정규화는 인식기를 적용하기 이전의 전처리 단계로 특징의 스케일에 따른 오류를 줄이기 위해 널리 사용되고 있다. 하지만 기존 정규화 방법은 특징의 분포를 가정하는 경우가 많으며, 클래스 라벨을 고려하지 않으므로 정규화 결과가 인식률에서 최적임을 보장하지 못하는 문제점이 있다. 이 논문에서는 특징의 분포를 가정하지 않는 랭크 정규화 방법과 클래스 라벨을 사용하는 교사 학습법을 결합한 교사 랭크 정규화 방법을 제안하였다. 제안하는 방법은 데이터의 분포를 바탕으로 특징의 분포를 자동으로 추정하므로 특징의 분포를 가정하지 않으며, 데이터 포인트의 최근접 이웃이 가지는 클래스 라벨을 바탕으로 정규화를 시행하므로 오류의 발생을 최소화할 수 있다. 특히 SVM의 경우 서로 다른 클래스에 속하는 데이터 포인트들이 혼재되어 나타나는 영역에 경계선을 설정하므로 이 영역의 밀도를 줄임으로써 경계선 설정을 보다 용이하게 하고 결과적으로 일반화 오류를 감소시킬 수 있다. 이러한 사실들은 실험 결과를 통해 확인할 수 있다.

▸ Keywords : SVM, 특징 정규화, 랭크 정규화, 교사 학습법

Abstract

Feature normalization as a pre-processing step has been widely used in classification problems to reduce the effect of different scale in each feature dimension and error as a result. Most of the existing methods, however, assume some distribution function on feature distribution. Even worse, existing methods do not use the labels of data points and, as a result, do not guarantee the optimality of the normalization results in classification. In this paper, proposed is a supervised rank normalization which combines rank normalization and a supervised learning technique. The

•제1저자 : 이수종 •교신저자 : 허경용
•투고일 : 2013. 10. 3, 심사일 : 2013. 10. 29, 게재확정일 : 2013. 11. 11.
* 협성대학교 컴퓨터공학과(Dept. of Computer Engineering, Hyupsung University)
** 동의대학교 전자공학과(Dept. of Electronic Engineering, Dong-Eui University)

proposed method does not assume any feature distribution like rank normalization and uses class labels of nearest neighbors in classification to reduce error. SVM, in particular, tries to draw a decision boundary in the middle of class overlapping zone, the reduction of data density in that area helps SVM to find a decision boundary reducing generalized error. All the things mentioned above can be verified through experimental results.

▸ Keywords : Support vector machine, Feature normalization, Rank Normalization, Supervised learning

I. 서 론

SVM(Support Vector Machine)은 구조적인 위험 최소화(structural risk minimization) 방법에 기초한 교사 학습 방법의 일종으로 Vapnik[1]에 의해 소개된 이후 다양한 분야에 적용되어 성공적인 결과를 보여줌으로써 현존하는 단일 분류기 시스템 중에서는 최고의 성능을 보이는 분류기 중 하나로 인정받고 있다[2][3]. 하지만 SVM 역시 특징 공간에서의 정규화는 필수적이다. D 차원 특징 공간에서 특정 차원의 특징값에 상수 T를 곱한다면 상수가 곱해진 특징값은 SVM의 목적 함수에 더 많은 영향을 미치게 되고 결과적으로 SVM의 결정 경계(decision boundary) 형성에 영향을 미치게 된다. 이 논문에서는 SVM에서 사용할 수 있는 대표적인 선형 및 비선형 정규화 방법을 살펴보고 이를 개선한 교사 랭크 정규화 (supervised rank normalization) 방법을 제안한다. 교사 랭크 정규화 방법은 다른 정규화 방법과 달리 데이터 포인트의 클래스 라벨을 활용하여 동일한 클래스에 속하는 데이터 포인트의 특징값 차이는 작게 하고 서로 다른 클래스에 속하는 데이터 포인트의 특징값 차이는 크게 함으로써 SVM의 경계면 설정을 도와 일반화 오류를 줄여주는 역할을 하며 이는 실험 결과를 통해 확인할 수 있다.

이 논문의 구성은 다음과 같다. 먼저 2장에서는 SVM에 관해 간략히 설명한다. 3장에서는 일반적으로 분류기에 적용되는 특징 정규화 방법들을 살펴본 후 개선된 특징 정규화 방법을 제안한다. 4장에서는 기존 방법과 제안하는 방법을 실험을 통해 비교함으로써 제안하는 방법의 우수성을 보이며 5장에서는 결론 및 향후 연구 방향을 제시한다.

II. SVM

N개의 D 차원 특징 벡터 x_i와 클래스 라벨 y_i가 주어졌다고 가정하자. 각 데이터 포인트의 클래스 라벨은 1또는 -1로 주어지는 것으로 가정한다. SVM은 두 클래스 사이의 경계면을 찾는 점에서 다른 분류기와 동일하다. 하지만 SVM은 정확하게 분류된 데이터 포인트들은 경계면에서 멀리 있도록 함과 동시에, 잘못 분류된 데이터 포인트들의 경계면까지의 거리 합을 최소로 하는 목적 함수를 정의함으로써 일반적인 분류기와 달리 데이터의 분포를 가정하지 않고 경계면까지의 거리를 중요시한다. 따라서 SVM은 최대 마진 분류기(maximum margin classifier)라고도 불린다. SVM에서 최소화하는 목적 함수는 식 (1)과 같다.

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \tag{1}$$

식 (1)에서 w는 경계면의 기울기 벡터로 $\|w\|^2$은 정확하게 분류된 데이터 포인트가 가지는 경계면까지의 거리에 반비례한다. ξ_i는 잘못 분류된 데이터 포인트의 경계면까지의 거리를 나타내며 C는 두 항의 비율을 조절하는 상수이다. 식 (1)을 최소화함으로써 SVM은 정확하게 분류된 데이터 포인트들은 경계면에서 최대한 멀리 있도록 하고, 잘못 분류된 데이터 포인트들의 경계면까지의 거리 합은 최소가 되도록 하는 경계면을 찾아낸다.

식 (1)에서 알 수 있듯이 SVM의 목적 함수는 거리에 기초하고 있다. 따라서 특징 공간에서 특정 차원에 상수 T를 곱하면, 즉 특징의 스케일이 달라지면 거리값이 달라지고 식 (1)을 최소화하는 경계면이 달라진다. 따라서 특징값은 전체 분류 오

류를 최소화하는 방향으로 정규화되어야 하지만 지금까지 SVM에서의 특징값 정규화에 대해서는 중요하게 고려되지 않았다. 이는 SVM이 특징 공간을 일반적으로 무한대의 커널 특징 공간으로 사상하여 선형 분류를 시도하는 방식으로 커널의 특성에 따라 일부 정규화 효과를 얻을 수 있기 때문이다.

이 논문에서는 특징의 정규화를 독립적인 데이터의 전처리 과정으로 간주하고 오류를 최소화할 수 있는 특징 정규화 방법을 제시한다. 따라서 이 논문에서 제시하는 특징 정규화 방법은 다른 분류기에서도 사용할 수 있다. 실험을 위해서는 선형 커널(linear kernel)을 사용하였다. 일반적으로 많이 사용되는 다항식 커널, 가우스 커널 등의 비선형 커널은 커널 최적화를 위한 별도의 파라미터들이 존재하므로 특징 정규화에 따른 영향을 파악하기 어려우므로 선형 커널을 사용하였다. 상수 C는 Matlab의 기본 설정을 사용하였다.

III. 특징 벡터 정규화

분류기를 사용함에 있어 특징 정규화는 필수적인 전처리 과정 중 하나로 다양한 정규화 방법이 제안되어 사용되고 있다[4]. 이들 특징 정규화 방법 중 흔히 사용되는 방법으로는 최대-최소 정규화 방법, 평균-분산 정규화 방법, 히스토그램 이퀄라이제이션 방법, 랭크 정규화 방법 등이 있다. 이외에도 여러 가지 방법이 있지만, 전처리 과정을 별도로 분리하기 어려운 경우가 많으며 전처리 과정을 위한 연산이 많이 요구되는 단점이 있다. 이 논문에서는 SVM을 위한 특징 정규화 방법을 다루고 있지만 다른 분류기에도 적용이 가능하고 전처리 과정에 과도한 부담을 주지 않기 위해 간단하면서도 효과적인 위의 4가지 방법을 비교 대상으로 선정하였다. 또한 특징 벡터의 차원이 D차원인 경우 각 차원은 개별적으로 정규화되는 것으로 가정하였다.

특징 벡터 $X=\{x_1,x_2,...,x_N\}$와 정규화된 특징 벡터 $X'=\{x_1',x_2',...,x_N'\}$는 N개의 D 차원 벡터로 구성되며 클래스 라벨 $Y=\{y_1,y_2,...,y_N\}$는 -1 또는 1 값을 가지는 것으로 가정한다. D 차원 특징 벡터는 $x_i=[x_{i1},...,x_{id},...,x_{iD}]^T$로 표시한다.

1. 최대-최소 정규화 방법[5]

최대-최소 정규화 방법은 특징 벡터의 최대값과 최소값을 지정하여 선형으로 사상하는 방법으로 최대값 U, 최소값 L이 주어지는 경우 식 (2)에 의해 정규화를 시행한다.

$$x_{id}'=\frac{x_{id}-\min_j x_{jd}}{\max_j x_{jd}-\min_j x_{jd}}(U-L)+L \tag{2}$$

범위를 [0 1]로 제한하는 경우 식 (2)는 식 (3)과 같이 표현할 수 있으며 흔히 사용되는 정규화 방법이다.

$$x_{id}'=\frac{x_{id}-\min_j x_{jd}}{\max_j x_{jd}-\min_j x_{jd}} \tag{3}$$

최대-최소 정규화는 특징값이 균일한 분포를 가진다는 가정 하에 선형 사상을 수행하는 방법으로 간단하면서도 효과적이어서 많이 사용되는 방법 중 하나이다. 하지만 특징값의 분포를 가정한다는 점, 특히 흔히 볼 수 없는 균일 분포를 가정한다는 점에서 그 한계가 있다.

2. 평균-분산 정규화 방법[5]

평균-분산 정규화 방법은 특징값의 평균과 분산을 이용하여 정규화를 수행하는 방법으로 식 (4)에 의해 정규화를 시행한다.

$$x_{id}'=\frac{x_{id}-\overline{x_d}}{\sigma_d} \tag{4}$$

식 (4)에서 $\overline{x_d}$와 σ_d는 차원 d의 평균과 분산으로 다음과 같이 계산된다.

$$\overline{x_d}=\frac{1}{N}\sum_{i=1}^{N}x_{id} \tag{5}$$

$$\sigma_d=\frac{1}{N}\sum_{i=1}^{N}(x_{id}-\overline{x_d})^2 \tag{6}$$

평균-분산 정규화 방법은 특징값이 가우스 분포를 가진다고 가정하고 평균과 분산으로 선형 사상을 수행하는 방법으로 기본적으로 최대-최소 정규화 방법과 동일하게 특정 범위로 이동시키는 방법이다. 평균-분산 정규화 방법은 서로 다른 차원의 특징들이 평균 0을 중심으로 거의 동일한 다이나믹 레인지를 갖도록 하는 방법이지만 특징들이 서로 다른 특성을 갖는 경우에는 부적당한 방법일 수 있다.

3. 히스토그램 이퀄라이제이션 방법[6]

최대-최소 정규화 방법과 평균-분산 정규화 방법이 특징값의 분포를 가정하고 있는 선형 사상 방법인 반면 히스토그램

이퀄라이제이션(histogram equalization) 방법은 특징값의 분포를 가정하지 않는 비선형 사상 방법인 점에서 차이가 있다. 선형 사상의 경우 잡음에 의해 소수의 특징값이 나머지 값들과 동떨어진 값을 가지는 경우 소수의 특징값으로 인해 나머지 값들이 사상 후 거의 유사한 값을 가지는 경우가 발생할 수 있다. 하지만 히스토그램 이퀄라이제이션 방법은 특징값의 발생 빈도에 의해 정규화를 시행함으로써 잡음에 의한 영향을 줄일 수 있는 장점이 있다.

입력 a에 의해 a'을 출력하는 사상 함수 $a'=f(a)$를 생각해 보자. 사상 함수의 입력이 확률 밀도 함수 (PDF) $p(a)$를 가지는 랜덤 변수 A라고 가정할 때 출력값 a'은 확률 밀도 함수 $p'(a')$을 가지는 새로운 랜덤 변수 A'이 된다. 각각의 누적 분포 함수(CDF)는 다음과 같이 표현된다.

$$P(a)=\int_{-\infty}^{a} p(t)dt \tag{7}$$

$$P'(a')=\int_{-\infty}^{a'} p(t)dt \tag{8}$$

사상 함수가 단조 증가 함수라면 $f(a)$는 누적 분포 함수로부터 얻어낼 수 있다.

$$f(a)=P'^{-1}P(a) \tag{9}$$

$p'(a')$은 레퍼런스 분포(reference distribution)로 사상 이전에 결정되는 함수이며 $p(a)$는 특징값으로 부터 얻어낼 수 있으므로 사상 함수를 쉽게 구할 수 있다. 그림 1은 이러한 사상 과정을 나타낸 것으로 이를 히스토그램 매칭(histogram matching)이라 한다.

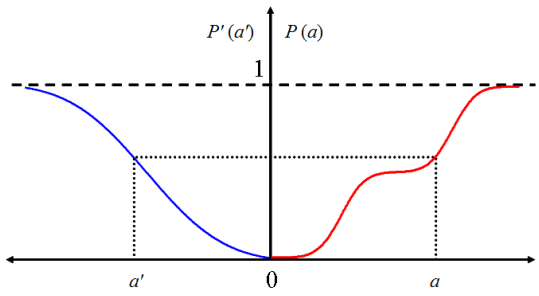


그림 1. 히스토그램 매칭
Fig. 1. Histogram matching

히스토그램 이퀄라이제이션은 히스토그램 매칭에서 레퍼런스 분포가 상수인 $p'(a')=\text{const}$. 특별한 경우에 해당한다.

4. 랭크 정규화 방법[7]

랭크 정규화(rank normalization) 방법은 특정 분포를 가정하지 않는 비모수적(non-parametric) 방법에서 사용되는 대표적인 방법 중 하나로 특징값을 오름차순 또는 내림차순으로 정렬하고 그 순서에 의해 특징값을 정규화하는 방식으로 일반적으로 오름차순에 의해 정렬한다. 랭크 정규화는 식 (10)을 통해 정규화를 시행한다.

$$x_{id}'=\frac{|X_i|}{|X|} \tag{10}$$

식 (10)에서 $|X|$는 집합 X의 크기를 나타내며 특징벡터의 개수 N의 값을 가지게 된다. X_i는 i번째 특징값 보다 작은 특징값의 집합으로 식 (11)과 같이 정의된다.

$$X_i=\{x_{jd} \mid x_{jd}<x_{id}, 1\leq j\leq N\} \tag{11}$$

랭크 정규화는 히스토그램 이퀄라이제이션과 동일한 결과를 가져오는 방법이므로 이 논문에서는 랭크 정규화를 기준으로 설명한다.

5. 클래스 라벨을 이용한 교사 랭크 정규화

기존의 특징 정규화 방법 중 선형 사상 방법인 최대-최소 정규화 방법과 평균-분산 정규화 방법은 선형성으로 인해 잡음의 영향을 많이 받는 단점이 있다. 랭크 정규화는 비선형 사상 방법으로 잡음의 영향을 적게 받으며 특징값이 밀집된 영역에서 특징값의 차이가 확연히 드러나도록 해 줄 수 있어 영상 처리에서는 히스토그램 이퀄라이제이션을 통해 영상의 대비를 향상시키기 위해 사용된다[6]. 하지만 위에서 언급한 기존의 방법들은 클래스 라벨을 이용하지 않는 비교사 방법이라는 한계가 있다. 즉, 사상을 통해 특징값을 새로운 특징값으로 변환하지만 변환된 특징값이 분류를 위해 더 나은 값이라는 보장이 없다. 따라서 이 논문에서는 클래스 라벨을 이용하여 특징값을 사상함으로써 분류에 보다 적합한 특징값으로 변환할 수 있는 교사 랭크 정규화 (supervised rank normalization) 방법을 제안한다. 제안하는 교사 랭크 정규화 방법은 SVM이 두 클래스에 속하는 데이터 포인트 사이에 경계선을 설정하면서 그 폭을 최대로 만들고자 하는 것과 마찬가지로 서로 다른 클래스에 속하는 데이터 포인트의 특징값 차이는 크게 하고 동일한 클래스에 속하는 데이터 포인트의

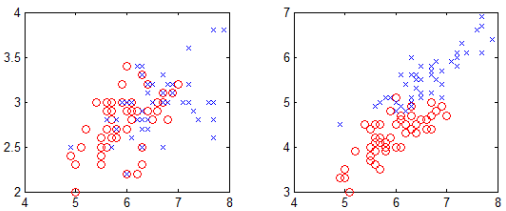
특징값 차이는 작게 함으로써 SVM의 경계면 결정을 도와줄 수 있도록 정규화를 수행한다. 교사 랭크 정규화는 식 (12)를 통해 정규화를 시행한다.

$$x_{id}' = \frac{[X_i]}{[X]} \tag{12}$$

식 (12)는 식 (10)과 기본적으로 동일하다. 하지만 랭크를 계산하는 방식에서 차이가 있다. 랭크 정규화는 단순히 주어진 값보다 작은 값을 가지는 특징값의 개수를 세어서 그 비를 계산하지만 교사 랭크 정규화에서는 개수를 셀 때 클래스 라벨에 따라 가중치를 부여한다. 식 (12)에서 $[X_i]$는 식 (13)과 같이 정의된다.

$$[X_i] = \sum_{j=1}^{N} \delta(x_{jd} < x_{id}) \overline{NN_i} \tag{13}$$

식 (13)에서 $\delta(\cdot)$은 주어진 조건을 만족하는 경우 1을, 만족하지 않는 경우 0의 값을 가지는 지시 함수(indicator function)로 주어진 특징값보다 작은 특징값을 가지는 데이터 포인트 개수를 세기 위함이다. $\overline{NN_i}$는 클래스 라벨에 따른 가중치로 데이터 포인트 x_i의 K개 최근접 이웃(nearest neighbor) 중 x_i와 다른 클래스 라벨을 가지는 데이터 포인트의 개수에 비례하는 값을 가지며 이 논문에서는 x_i와 다른 클래스 라벨을 가지는 데이터 포인트의 개수에 1을 더한 값으로 설정하였다. 즉, x_i의 K개 최근접 이웃 중 클래스 라벨이 x_i와 다른 데이터 포인트가 많은 경우에는 가중치를 증가시켜 이웃한 데이터 포인트들과의 특징값 차이를 크게 만들어 준다. 식 (13)을 이용하여 식 (12)는 식 (14)와 같이 정의할 수 있다.



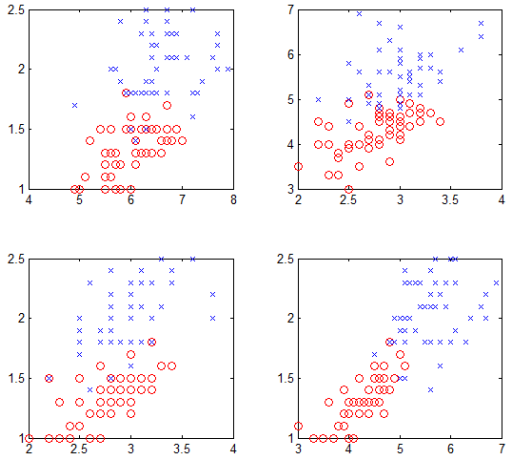
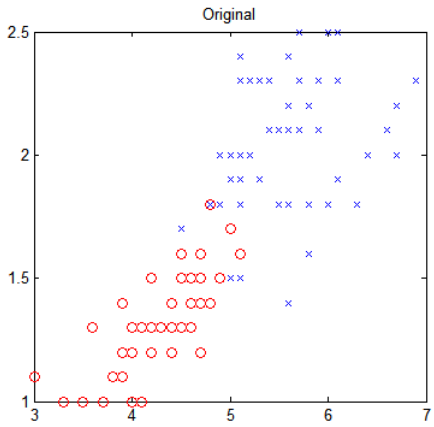
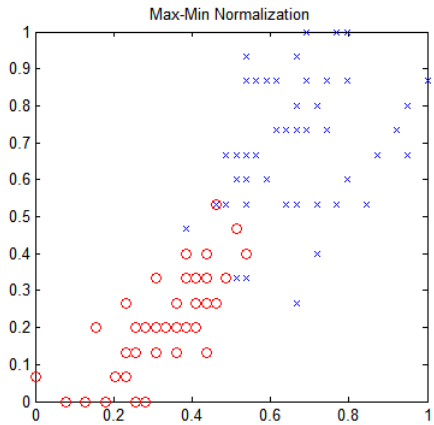


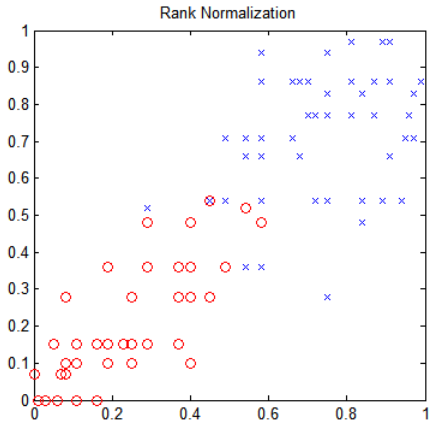
그림 2. 아이리스 데이터의 특징값 분포
Fig. 2. Feature distribution of iris data sets



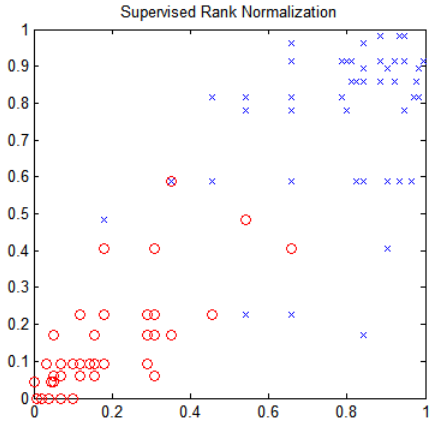
(a) 원본 데이터



(b) 최대-최소 정규화



(c) 랭크 정규화



(d) 교사 랭크 정규화

그림 3. 정규화에 따른 데이터 분포
Fig. 3. Data distribution according to normalization method

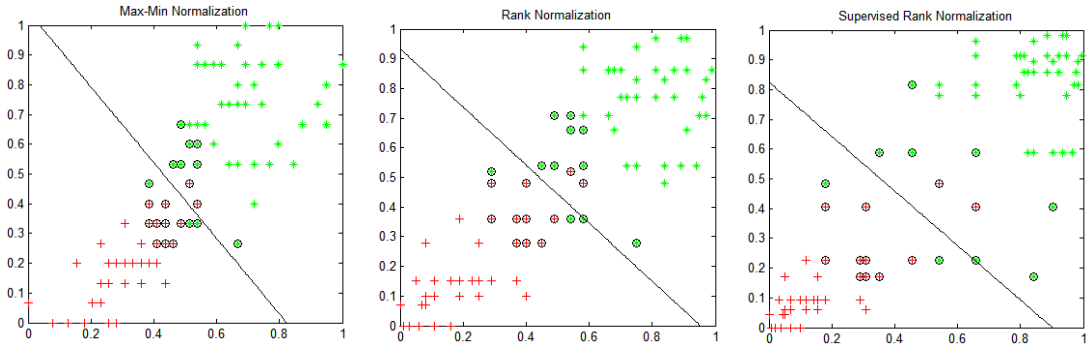
$$[X] = \sum_{j=1}^{N} \overline{NN_i} \tag{14}$$

식 (13)과 식 (14)에서 $\overline{NN_i}$를 상수로 설정하면 식 (12)는 식 (10)과 동일한 식이 된다. 따라서 제안하는 교사 랭크 정규화 방법은 랭크 정규화 방법을 포함하는 일반화된 랭크 정규화 방법이라고 할 수 있다.

IV. 실험 결과

실험에서는 선형 커널을 사용하는 SVM을 사용하였다. 실험에 사용한 데이터는 Fisher의 아이리스 데이터로 아이리스 데이터는 4차원 특징 벡터 150개가 3개의 클래스에 할당되어 있다. 아이리스 데이터의 4차원 특징 중 실험에서는 임의의 2개 특징을 선택하여 서로 다른 2차원 특징 벡터를 가지는 6개의 새로운 아이리스 데이터 집합을 만들어 사용하였다. 아이리스 데이터는 setosa, versicolor, virginica 3개의 클래스로 이루어져 있지만 새롭게 만든 6개의 아이리스 데이터 집합에서 선형으로 분리가 불가능한 경우는 versicolor와 virginica이므로 이들 두 클래스만을 사용하고 setosa는 사용하지 않았다. 따라서 실험에 사용한 데이터는 2차원의 특징 벡터를 가지며 versicolor와 virginica 클래스로 구성된다. 각 데이터는 100개의 샘플로 이루어진다. 그림 2는 6개의 새로운 데이터 집합을 나타낸 것으로 첫 번째 데이터 집합에서 두 클래스가 많이 겹치는 것을 볼 수 있다.

특징 정규화 방법은 정규화 이후 특징값의 범위가 동일하도록 [0 1] 범위의 최대-최소 정규화, 랭크 정규화, 그리고 교사 랭크 정규화 세 가지 방법을 비교하였다. 그림 3은 세 가지 정규화 방법을 6번째 데이터 집합에 적용한 결과를 나타



(a) 최대-최소 정규화 (b) 랭크 정규화 (c) 교사 랭크 정규화

그림 4. 정규화 방법에 따라 SVM이 설정한 경계면
Fig. 4. Decision boundaries w.r.t. a normalization method

낸다. 최대-최소 정규화는 특징값의 범위만이 변경될 뿐 상대적인 위치는 그대로 유지됨을 알 수 있다. 랭크 정규화는 균일한 분포로 특징값을 재배치함으로써 밀집되거나 성긴 곳 없이 균일한 분포를 가지도록 특징값을 정규화한다. 이 논문에서 제안한 교사 랭크 정규화의 경우에는 동일한 클래스에 속하는 데이터 포인트들은 밀집되어 나타나고 두 클래스의 경계 부분에서는 데이터 포인트들이 성기게 나타남으로써 SVM이 경계면을 보다 쉽게 설정할 수 있도록 해준다. 그림 3의 정규화한 데이터에 test-on-train 방법으로 SVM을 적용한 결과는 그림 4와 같으며 각각의 경우 오류는 6%, 6%, 5%로 제안한 방법에서 가장 적은 오류를 보였다. 그림 4에서는 오류가 1% 적게 보이지만 이는 유사한 특징값의 데이터 포인트가 겹쳐져 있기 때문이다.

세 가지 방법의 비교를 위해서 6개의 데이터 집합에 대해 10 fold cross-validation을 시행하였다. 교사 랭크 정규화에서 최근접 이웃의 수는 전체 데이터 포인트 개수의 10%인 10으로 실험적으로 결정하였다. 표 1은 실험 결과를 요약한 것이다.

그림 2에서 알 수 있듯이 첫 번째 데이터 집합은 두 클래스의 겹쳐진 정도가 심해 31%의 오류가 발생하였다. 세 번째 데이터 집합의 경우에는 랭크 정규화가 가장 큰 오류를 보였지만 랭크 정규화의 경우 전반적으로 최대-최소 정규화와 같거나 나은 성능을 보여주었다. 반면 이 논문에서 제안한 교사 랭크 정규화의 경우 모든 데이터 집합에서 최대-최소 정규화나 랭크 정규화와 같거나 나은 성능을 보여주었다.

표 1. 실험 결과 요약
Table 1. Summary of experimental results

데이터 집합	오류 (%)		
	최대-최소 정규화	랭크 정규화	교사 랭크 정규화
1	31.0	31.0	31.0
2	7.0	7.0	6.0
3	7.0	8.0	7.0
4	7.0	6.0	6.0
5	8.0	8.0	8.0
6	7.0	6.0	6.0
평균 오류	11.2	11.0	10.7
평균 오류 (1번 제외)	7.2	7.0	6.6

V. 결론

특징의 정규화는 인식기를 사용하기 위한 전처리 단계로 특징의 스케일 변화에 따른 오인식을 줄이기 위해 사용되고 있다. 하지만 기존 정규화 방법은 특징의 분포를 가정하는 경우가 있으며, 클래스 라벨을 고려하지 않으므로 정규화 결과가 인식률 측면에서 최적임을 보장하지 못하는 문제점이 있다.

이 논문에서는 특징의 분포를 가정하지 않는 랭크 정규화에 교사 학습 방법을 결합한 교사 랭크 정규화 방법을 제안하였다. 교사 랭크 정규화는 데이터 포인트의 최근접 이웃들이 가지는 클래스 라벨을 기반으로 서로 다른 클래스에 속하는 데이터 포인트들이 혼재되어 나타나는 경우 SVM이 경계면 설정을 쉽게 할 수 있도록 밀도를 줄이고, 동일한 클래스에 속하는 데이터 포인트들만 나타나는 경우 SVM의 관심 영역이 아니므로 밀도를 높이는 방향으로 정규화를 수행함으로써 경계면 설정을 도와줄 수 있다. 교사 랭크 정규화는 일반화 오류에서 기존의 방법들에 비해 우수함을 실험 결과로 알 수 있다.

제안한 방법이 기존 방법들에 비해 우수한 성능을 보였지만 몇 가지 개선이 가능하다. 먼저 최근접 이웃의 개수 및 최근접 이웃의 분포에 따른 가중치 $\overline{NN_i}$을 결정하는 방법이 이 논문에서는 실험적으로 결정되었으며 보다 체계적이고 자동화된 방법이 필요할 것으로 생각된다. 두 번째는 특징의 독립성에 대한 가정이다. 이 논문에서는 연산량을 고려해 각 특징들을 개별적으로 정규화하고 있지만 상호 연관성을 고려하여 정규화를 시행할 경우를 고려해볼 필요가 있다. 마지막으로 최근접 이웃의 개수 자동 결정과 최근접 이웃에 가중치 적용 등을 제안한 방법과 결합할 경우 더 나은 성능 향상이 이루어질 수 있을 것으로 기대하며 이에 대한 연구가 진행 중에 있다.

참고문헌

[1] Vladimir Vapnik, Statistical Learning Theory, John Wiley & Sons, New York, 1998.

[2] Ingo Steinwart, Andreas Christmann, "Support Vector Machines," Springer, 2008.

[3] Ashis Pradhan, "Support Vector Machine - A Survey," International Journal of Emerging Technology and Advanced Engineering, Vol. 2,

No. 8, pp. 82-85, Aug. 2012.
[4] Selim Aksoy, Robert M. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," Pattern Recognition Letters, Vol. 22, No. 5, pp. 563-582, Apr. 2001.
[5] Keinosuke Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, SanDiego, AcademicPress, 1990
[6] Rafeel C. Gonzalez, Richard E. Woods, Steven L. Eddins, "Digital Image Processing using MATLAB," McGraw Hill, 2011.
[7] Andreas Stolcke, Sachin Kajarekar, and Luciana Ferrer, "Nonparametric Feature Normalization for SVM-based Speaker Verification," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas NV, pp. 1577-1580, March 2008.

저 자 소 개

이 수 종
1989: 국민대학교 전자공학과 공학사.
1992: 연세대학교 전자공학과 공학석사.
2000: 연세대학교 전기·컴퓨터공학과 공학박사
현 재: 협성대학교 컴퓨터공학과 교수
관심분야: 영상처리,인공지능, 컴퓨터비젼
Email : sjlee@uhs.ac.kr

허 경 용
1994: 연세대학교 전자공학과 공학사.
1996: 연세대학교 전자공학과 공학석사.
2009: University of Florida 컴퓨터공학과 공학박사
현 재: 동의대학교 전자공학과 교수
관심분야: 인공지능, 패턴인식, 로봇공학
Email : hgycap@deu.ac.kr