Journal of The Korea Society of Computer and Information
Vol. 20, No. 1, January 2015
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.1.021

학습 샘플 선택을 이용한 교사 랭크 정규화

허 경 용*, 최 훈*, 윤 주 상**

Supervised Rank Normalization with Training Sample Selection

Gyeongyong Heo*, Hun Choi*, Joo-Sang Youn**

요 약

특징 정규화는 인식기를 적용하기 이전의 전처리 단계로 특징 차원에 따라 서로 다른 스케일에 의해 발생하는 오류를 줄이기 위해 널리 사용된다. 하지만 기존 정규화 방법은 클래스 라벨을 고려하지 않으므로 정규화 결과가 인식률에서 최적임을 보장하지 못하는 문제점이 있다. 이를 개선하기 위해 클래스 라벨을 사용하여 정규화를 시행하는 교사 정규화 방법이 제안되었고 기존 정규화 방법에 비해 나은 성능을 보임이 입증되었다. 이 논문에서는 교사 랭크 정규화 방법에 학습 샘플 선택 방법을 적용함으로써 교사 랭크 정규화 방법을 더욱 개선할 수 있는 방법을 제안한다. 학습 샘플 선택은 잡음이 많은 샘플을 학습에서 제외함으로써 잡음에 보다 강한 분류기를 학습시키는 전처리 단계로 많이 사용되며 랭크 정규화에서도 역시 사용될 수 있다. 학습 샘플 선택은 이웃한 샘플이 속하는 클래스와 이웃한 샘플까지의 거리를 바탕으로 하는 두 가지 척도를 제안하였고, 두 가지 척도 모두에서 기존 정규화 방법에 비해 인식률이 향상되었음을 실험 결과를 통해 확인할 수 있었다.

▸ Keywords : 특징 정규화, 랭크 정규화, 교사 학습법, 학습 샘플 선택

Abstract

Feature normalization as a pre-processing step has been widely used to reduce the effect of different scale in each feature dimension and error rate in classification. Most of the existing normalization methods, however, do not use the class labels of data points and, as a result, do not guarantee the optimality of normalization in classification aspect. A supervised rank normalization method, combination of rank normalization and supervised learning technique, was proposed and demonstrated better result than others. In this paper, another technique, training sample selection, is introduced in supervised feature

•제1저자 · 교신저자 : 허경용
•투고일 : 2014. 11. 3, 심사일 : 2014. 12. 16, 게재확정일 : 2015. 1. 6.
* 동의대학교 전자공학과 (Dept. of Electronic Engineering, Dong-eui University)
** 동의대학교 멀티미디어공학과 (Dept. of Multimedia Engineering, Dong-eui University)
※ 이 논문은 2013년 동의대학교 교내연구비 지원으로 연구되었음 (과제번호:2013AA140)

normalization to reduce classification error more. Training sample selection is a common technique for increasing classification accuracy by removing noisy samples and can be applied in supervised normalization method. Two sample selection measures based on the classes of neighboring samples and the distance to neighboring samples were proposed and both of them showed better results than previous supervised rank normalization method.

▸ Keywords : Feature normalization, Rank Normalization, Supervised learning, Training sample selection

I. 서 론

분류기를 사용함에 있어 특징 정규화는 필수적인 전처리 과정 중 하나이다[1]. 분류기에 사용되는 특징은 일반적으로 D 차원으로 구성되며 각 차원은 특징값 범위는 서로 달라 평균값이 크거나 값의 범위가 큰 특징은 일반적으로 분류 척도 계산에 미치는 영향이 커 다른 중요한 특징을 가릴 수 있다. 따라서 특징 정규화에서는 기본적으로 모든 차원의 특징값들이 동일한 분포를 가지도록, 또는 특징의 중요도를 평가하여 중요한 특징이 분류 척도 계산에 미치는 영향이 커지도록 특징값의 분포를 조절하는 과정을 거치게 된다. 하지만 대부분의 정규화 방법은 클래스 라벨을 고려하지 않으므로 정규화가 분류 측면에서 최적임을 보장하지 못하므로, 클래스 라벨을 고려한 교사 정규화 방법을 통해 분류 오류를 줄일 수 있다[2].

학습 샘플 선택 역시 학습 과정에서 흔히 사용되는 기법 중 하나로 잡음의 영향에 의한 비전형적인 샘플을 제거함으로써 오류를 줄이는 방법이다[3]. 비전형적인 샘플의 경우 전형적인 샘플에 비해 학습된 분류기에 미치는 영향이 크기 때문에 비전형적인 샘플을 제거함으로 분류 오류를 줄일 수 있다. 하지만, 비전형적인 샘플은 샘플의 분포를 알지 못하는 경우 판단하기 어려우며 거의 모든 분류 문제에서 샘플의 분포는 알려져 있지 않다. 학습 샘플 선택을 위해 다양한 방법이 제안되었지만 학습 샘플의 분포를 유추할 수 있는 특정 문제에만 사용할 수 있는 방법이 대부분이므로, 이 논문에서는 클래스 라벨을 활용하여 간단하게 전형적인 정도를 판단할 수 있는 방법을 교사 정규화와 함께 사용하여 분류 오류를 줄일 수 있는 방법을 제안하였다. 이 논문에서는 제안한 방법을 SVM(Support Vector Machine)을 위한 전처리 단계로 사용하였지만, SVM과 완전히 분리된 과정으로 구현하였으므로 다른 분류기에서도 사용할 수 있다.

이 논문의 구성은 다음과 같다. 2장에서는 먼저 SVM을 사용함에 있어 정규화 및 학습 샘플 선택에 관에 간략히 설명하고, 3장에서는 이전 정규화 방법들에 대해 살펴본다. 4장에서는 제안하는 방법에 대해 살펴보고 5장에서는 제안하는 방법과 기존 방법을 실험을 통해 비교함으로써 제안하는 방법의 우수성을 보인다. 6장에서는 결론 및 향후 연구 방향을 제시한다.

II. SVM

SVM은 일반적인 분류기가 사용하는 경험적 위험 최소화 방법법이 아닌 구조적인 위험 최소화을 사용한 교사 학습 방법의 일종으로 Vapnik[4]에 의해 소개된 이후 다양한 분야에서 성공적인 결과를 보여줌으로써 단일 분류기 중에서는 최고의 성능을 보여주는 분류기 중 하나로 인정받고 있다[5].

N개의 D 차원 특징 벡터 x_i와 1 또는 -1의 값을 가지는 클래스 라벨 y_i가 주어졌다고 가정하자. SVM이 두 클래스 사이의 분류 경계면을 찾는다는 점에서는 다른 분류기와 동일하다. 하지만 SVM은 정확하게 분류된 데이터 포인트들은 경계면에서 멀리 있도록 함과 동시에, 잘못 분류된 데이터 포인트들은 경계면 가까이에 놓이도록 하는 목적 함수를 정의하고 있다. 이처럼 SVM은 일반적인 분류기와 달리 데이터의 분포를 가정하지 않고 경계면까지의 거리를 중요시하므로 최대 마진 분류기(maximum margin classifier)라고도 불린다.

SVM에서 최소화하는 목적 함수는 식 (1)과 같다.

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

식 (1)에서 w는 경계면 벡터로 $\|w\|^2$은 정확하게 분류된 데이터 포인트가 가지는 경계면까지의 거리에 반비례한다. ξ_i는 잘못 분류된 데이터 포인트의 경계면까지의 거리, 일종의 오류값을 나타내므로 경계면은 가능한 오류가 적은 위치에 위치하게 된다. C는 두 항의 비율을 조절하는 상수로 식 (1)의 최소화는 정확하게 분류된 데이터 포인트들은 경계면에서 최대한 멀리 있도록 하고, 잘못 분류된 데이터 포인트들의 경계면까지의 거리 합은 최소가 되도록 해준다.

식 (1)에서 알 수 있듯이 SVM의 목적 함수는 거리에 기초하고 있다. 특징 공간에서 특정 차원에 상수 $T(>1)$를 곱하면, 거리가 달라지고 식 (1)을 최소화하는 경계면이 달라진다. 따라서 특징값은 전체 분류 오류를 최소화하는 방향으로 정규화 되어야 한다. 하지만 분류 오류를 최소화하는 방향으로 특징을 정규화하기 위해서는 먼저 특징이 분류 오류에 미치는 영향을 파악하여야 한다[6]. 하지만 이 문제 역시 학습 샘플의 선택과 마찬가지로 쉬운 문제는 아니므로 이 논문에서는 모든 특징이 동일한 분포를 가지도록 정규화하여 사용하였다. 학습 샘플 선택을 위해서는 클래스 라벨을 고려한 KNN (k nearest neighbor) 방법을 사용하여 최근접 이웃의 클래스 라벨이 자신과 다른 정도를 비전형적인 샘플의 정도로 나타내는 방법을 사용하였다.

실험을 위한 SVM은 선형 커널(linear kernel)을 사용하였다. 일반적으로 많이 사용되는 다항식 커널이나 가우스 커널 등의 비선형 커널은 커널의 최적화를 위해 별도의 파라미터들이 필요하므로 특징 정규화에 따른 영향을 파악하기 어려워 선형 커널을 통해 비교하였다. 식 (1)에서의 상수 C는 매트랩의 기본 설정을 사용하였다.

III. 특징 벡터 정규화

특징 벡터 정규화는 분류기를 적용하기 이전에 대부분 시행되는 전처리 과정 중 하나이다. 특징 벡터 정규화는 개별적인 차원으로 정규화 하는 방법, 여러 차원을 함께 정규화 하는 방법, 분류 결과를 피드백을 통해 반영하여 정규화 하는 방법 등 다양한 방법이 사용되고 있지만 특징 벡터의 분포를 알지 못하는 경우 그 효과를 보장할 수 없다. 흔히 사용되는 정규화 방법으로는 최대-최소 정규화 방법, 평균-분산 정규화 방법, 히스토그램 이퀄라이제이션 방법, 랭크 정규화 방법 등이 있다[7].

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 가정하자. D 차원 특징 벡터는 $x_i=[x_{i1},...,x_{id},...,x_{iD}]^T$로 표시할 수 있다. 이 논문에서는 특징 벡터의 각 차원을 독립적으로 정규화 하는 방법만을 대상으로 한다. 2개 이상의 차원의 연관성을 고려하는 경우 연산량이 증가하는 문제점이 있으며, 문제 종속적인 정보를 바탕으로 하는 경우가 많으므로 범용적으로 사용하기에는 한계가 있으므로 고려하지 않았다.

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

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

최대-최소 정규화는 특징값이 균일한 분포를 가진다는 가정 에 기반한 간단하면서도 효과적인 방법으로 많이 사용된다. 하지만 특징값의 분포를 가정하고 있으며 가정된 균일 분포를 가지는 특징이 실제로 존재하기 힘들다는 점에서 그 한계가 있다.

평균-분산 정규화 방법은 특징값의 평균과 분산을 이용하여 최대-최소 정규화와 유사하게 정규화를 수행하는 방법으로 식 (3)을 이용한다.

$$x_{id}' = \frac{x_{id}-\overline{x_d}}{\sigma_d} \quad (3)$$

식 (3)에서 $\overline{x_d}$와 σ_d는 차원 d의 평균과 분산을 나타낸다. 평균-분산 정규화 방법은 특징값이 가우스 분포를 가진다고 가정하고 평균과 분산으로 선형 사상을 수행하는 방법이다. 평균-분산 정규화 방법은 특징값이 평균 0을 중심으로 동일한 다이나믹 레인지를 갖도록 해준다.

위의 두 가지 방법은 특징값의 분포를 가정하고 있는 선형 사상 방법인 반면 히스토그램 이퀄라이제이션(histogram equalization) 방법은 특징값의 분포를 가정하지 않는 비선형 사상 방법이다. 선형 사상의 경우 잡음에 민감하며 소수의

잡음이 많이 포함된 값들로 인해 나머지 값들의 변별력이 감소하는 단점이 있다. 하지만 히스토그램 이퀄라이제이션은 특징값의 발생 빈도에 의해 정규화를 시행함으로써 잡음에 의한 영향을 줄일 수 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 하면 출력은 확률 밀도 함수 $p'(a')$을 가지는 새로운 랜덤 변수 A'이 된다. 각각의 누적 분포 함수(CDF)는 식 (4) 및 (5)와 같이 표현된다.

$$P(a)=\int_{-\infty}^{a} p(t)dt \tag{4}$$

$$P'(a')=\int_{-\infty}^{a'} p(t)dt \tag{5}$$

사상 함수가 단조 증가 함수라면 사상함수 $f(a)$는 누적 분포 함수로부터 식 (6)과 같이 얻어낼 수 있다.

$$f(a)=P'^{-1}P(a) \tag{6}$$

이 때 $p'(a')$은 레퍼런스 분포(reference distribution)로 사상 이전에 결정되는 함수이며, $p(a)$는 특징값으로 부터 얻어낼 수 있으므로 사상 함수를 계산할 수 있다. 그림 1은 사상함수를 구하는 과정을 나타낸 것으로 히스토그램 매칭(histogram matching)이라고 한다. 히스토그램 이퀄라이제이션은 히스토그램 매칭에서 레퍼런스 분포가 상수인 경우에 해당한다.

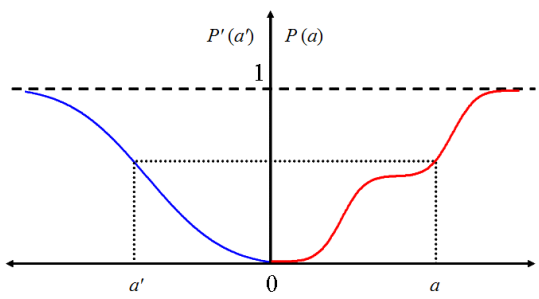


그림 1. 히스토그램 매칭
Fig. 1. Histogram matching

랭크 정규화(rank normalization)[8]는 특징값을 오름차순 또는 내림차순으로 정렬하고 그 순서에 의해 특징값을 정규화하는 방식이다. 랭크 정규화는 히스토그램 이퀄라이제이션과 출발점은 다르지만 동일한 결과를 가져온다. 랭크 정규화는 식 (7)을 통해 정규화를 시행한다.

$$x_{id}'=\frac{|X_i|}{|X|} \tag{7}$$

식 (7)에서 $|X|$는 집합 X의 크기를 나타내며, 집합 X_i는 i번째 특징값 보다 작은 특징값의 집합으로 식 (8)과 같이 정의할 수 있다.

$$X_i=\{x_{jd} \mid x_{jd}<x_{id}, 1\le j\le N\} \tag{8}$$

위의 네 가지 방법들은 클래스 라벨을 사용하지 않는 비교사 학습 방법이라는 한계가 있다. 즉, 사상을 통해 특징값을 새로운 특징값으로 변환하지만 변환된 특징값이 분류를 위해 더 나은 값이라는 보장이 없다. 분류 측면에서 최적의 결과를 보장하기 위해 제안된 방법이 교사 랭크 정규화 방법이다[2]. 교사 랭크 정규화에서는 식 (9)를 통해 정규화를 수행한다.

$$x_{id}'=\frac{[X_i]}{[X]} \tag{9}$$

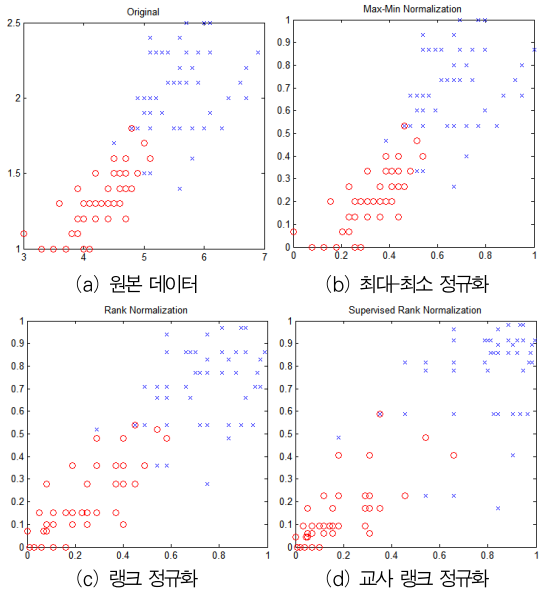
식 (9)는 식 (8)과 기본적으로 동일하지만 랭크를 계산하는 방식에서 차이가 있다. 랭크 정규화는 전체 특징값의 개수에서 주어진 값보다 작은 값을 가지는 특징값의 개수 비율로 사상된 특징값을 계산하지만, 교사 랭크 정규화에서는 특징값의 수를 계산할 때 클래스 라벨에 따라 가중치를 부여한다. 식 (9)에서 $[X_i]$는 식 (10)과 같이 정의된다.

$$[X_i]=\sum_{j=1}^{N}\delta(x_{jd}<x_{id})\overline{NN}_i \tag{10}$$

식 (10)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 1을 더한 값이 사용되었다. 즉, x_i의 K개 최근접 이웃 중 클래스 라벨이 x_i와 다른 데이터 포인트가 많은 경우에는 가중치를 증가시켜 이웃한 데이터 포인트들과의 특

징값 차이를 크게 만들어 줌으로써 서로 다른 클래스에 속하는 샘플들이 섞여서 나타나는 영역에서는 분류를 위한 공간을 넓혀주는 역할을 한다. $\overline{NN}_i$을 상수로 설정하면 식 (9)는 식 (7)과 동일하므로 교사 랭크 정규화는 랭크 정규화를 포함하는 일반화된 랭크 정규화 방법으로 볼 수 있다.

교사 랭크 정규화를 사용하면 그림 2에 나타난 것처럼 서로 다른 클래스에 속하는 데이터 포인트들이 함께 나타나는 영역에서의 데이터 포인트들 간격은 넓어지고, 동일한 클래스에 속하는 데이터 포인트들만이 나타나는 영역에서의 데이터 포인트들 간격은 좁아져서 분류에서 유리하며 실제 교사 랭크 정규화 방법이 위의 5가지 방법 중 가장 나은 결과를 보여주었다[2].



(a) 원본 데이터 (b) 최대-최소 정규화

(c) 랭크 정규화 (d) 교사 랭크 정규화

그림 2. 정규화에 따른 데이터 분포
Fig. 2. Data distribution according to normalization method

IV. 학습 샘플 선택

교사 랭크 정규화가 다른 정규화 방법에 비해 더 나은 결과를 보여주었지만 학습 샘플 선택과 함께 사용되면 더 나은 결과를 얻을 수 있다. 학습 샘플 선택은 각 샘플의 전형성을 정의하고 비전형적인 샘플을 제거하는 단순한 과정이다. 하지만 전형성은 정규화에서와 마찬가지로 샘플의 분포가 알려지지 않은 경우 일반적으로 정의하기 어렵다. 특정 데이터 집합에서만 적용 가능한 방법으로는 보다 면밀한 방법이 존재하지만 이 논문에서는 일반적으로 적용할 수 있는 방법을 제안하고 정규화 이전 단계로 적용하였다. 전형성 판단을 위해서는 두 가지 척도를 사용하였으며 그 중 하나는 K개 최근접 이웃 중 x_i와 다른 클래스 라벨을 가지는 데이터 포인트의 개수로 식 (11)과 같이 정의된다.

$$M_{1i} = \sum_{j=1}^{K} \delta(y_i = y_{NNj}) \tag{11}$$

식 (11)에서 $\delta(\cdot)$는 지시함수이며 y_{NNj}는 K개 최근접 이웃 중 j번째 샘플의 클래스 라벨을 나타낸다.

두 번째 척도는 K개 최근접 이웃까지의 거리 합과, K개 최근접 이웃 중 서로 다른 클래스에 속하는 샘플까지의 거리 합 비율로 식 (12)와 같이 정의된다.

$$M_{2i} = \frac{\sum_{j=1}^{K} \delta(y_i = y_{NNj}) d(x_i, x_{NNj})}{\sum_{j=1}^{K} d(x_i, x_{NNj})} \tag{12}$$

학습 샘플에서 전형성을 판단하는 척도를 계산하고 계산된

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

방법	평균 오류 (개)	오류 분산	최소 오류 (개)	최대 오류 (개)
랭크 정규화	37.10	3.97	32	43
교사 랭크 정규화	36.37	3.79	32	41
교사 랭크 정규화 + M_1	36.14	4.00	31	41
교사 랭크 정규화 + M_2	35.56	4.23	31	42

값이 임계치를 넘어가는 경우 비전형적인 학습 샘플로 판단하고 학습에서 제외하였다. 비전형적인 샘플의 경우 새로운 정보로서의 가치도 있으나, 비전형적인 샘플은 잡음의 영향이 큰 것으로 간주되는 일반적으로 경우를 이 논문에서도 따랐다.

임계치는 두 가지 척도에서 모두 70%로 설정하였으며 이 값은 실험적으로 결정하였다. 학습 샘플 선택 과정과 교사 랭크 정규화 방법은 함께 또는 각각 사용할 수 있도록 독립된 과정으로 설계하였으며, 함께 사용하는 경우 샘플 선택을 먼저 적용하고 이후 교사 랭크 정규화 과정을 적용하였다.

V. 실험 결과

실험에서는 선형 커널을 사용하는 SVM을 사용하였다. SVM은 매트랩에서 제공하는 라이브러리를 사용하였고, SVM과 관련된 파라미터 값들은 매트랩의 기본 설정을 그대로 사용하였다. 실험에 사용한 데이터는 Fisher의 아이리스 데이터이다. 아이리스 데이터는 setosa, versicolor, virginica의 3개 클래스로 구성되지만 제공되는 4개의 특징 중 임의의 2개 특징으로 선형 분리가 가능한 versicolor와 virginica의 두 클래스만을 사용하였다. 또한 4개의 특징 벡터 중 임의의 2개 벡터로 구성되는 6개의 데이터 집합 중에서 1번과 2번 특징값으로 구성되는 데이터 집합의 경우 그림 3에서와 같이 두 클래스를 분리하기가 불가능하므로 이를 제외한 5개 데이터 집합만을 실험에 사용하였다.

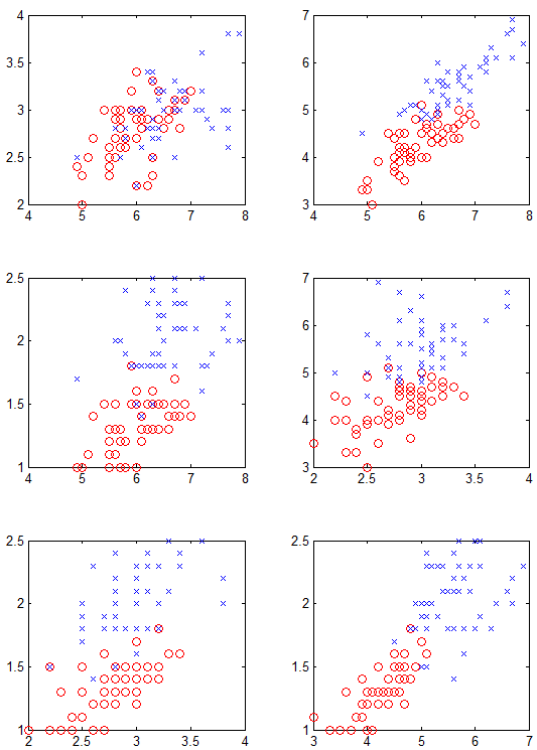
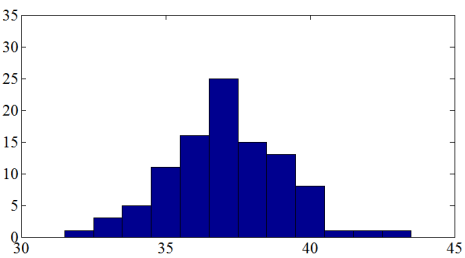


그림 3. 아이리스 데이터의 특징값 분포
Fig. 3. Feature distribution of iris data sets

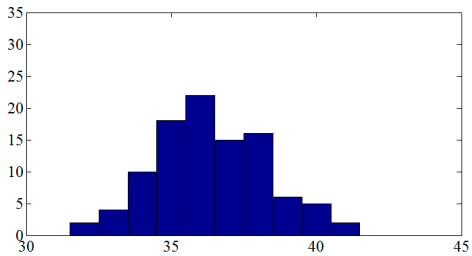
실험에서는 앞 절에서 소개한 5가지 정규화 방법 중 가장 우수한 성능을 보이는 2가지 방법, 랭크 정규화와 교사 랭크 정규화 방법, 그리고 식 (11)과 식 (12)의 척도를 사용한 학습 샘플 선택을 교사 랭크 정규화 방법 이전에 적용한 두 가지 방법, 총 4가지 방법을 비교하였다. 실험은 5가지 데이터 집합에 대해 10 fold cross-validation을 100회 실시하였다.

표 1은 그림 3에 나타난 5가지 데이터 집합을 4가지 전처리 방법을 적용하여 실험한 결과를 요약한 것이다. 표 1에서 알 수 있듯이 학습 샘플 선택과 교사 랭크 정규화를 함께 사용한 경우 이전 방법에 비해 나은 결과를 보였다. 학습 샘플 선택을 위한 척도는 거리 기반의 M_2가 개수 기반의 M_1에 비해 나은 결과를 보여주었다. 최대 최소 오류 역시 학습 샘플 선택을 함께 사용하는 경우 더 나은 결과를 보여준다.

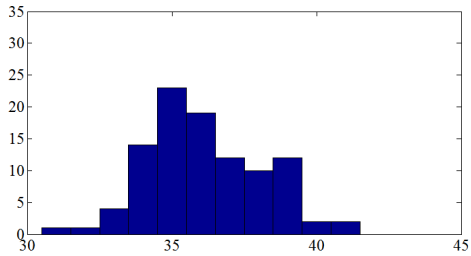
비전형적인 샘플을 제거하게 되면 SVM의 분류 경계면은 단순해진다. 일반적으로 경계면이 단순해지면 일반화 오류는 감소하는 것으로 알려져 있다. 하지만 경계면을 지나치게 단순화하면, 즉, 학습 샘플을 너무 많이 제거하면 학습 샘플 내의 변이를 충분히 반영할 수 없어 오류는 증가할 수 있다. 학습 샘플 선택을 함께 사용한 100회의 실험에서 약 70%는 학습 샘플 선택을 사용하지 않은 경우에 비해 낮은 오류를 보여주었지만, 나머지 약 30%서는 학습 샘플 선택을 사용하지 않은 경우보다 높은 오류를 보였다. 이는 오류의 분산이 학습 샘플 선택을 사용한 경우가 사용하지 않은 경우보다 높게 나타난 것에서도 알 수 있다.



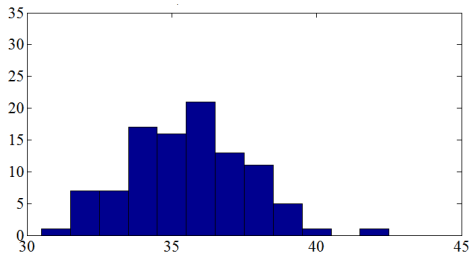
(a) 랭크 정규화



(b) 교사 랭크 정규화



(c) 교사 랭크 정규화 + M_1



(d) 교사 랭크 정규화 + M_2

그림 6. 오류 분포 히스토그램
Fig. 6. Histograms of errors

그림 6은 100회 반복 실험한 오류의 히스토그램을 나타낸 것으로 랭크 정규화의 경우 대부분의 결과가 37 근처에 집중적으로 나타나는 반면, 평균 오류가 감소할수록 넓은 영역에 걸쳐 나타남을 알 수 있다. 실험에서는 고정된 임계치를 사용하였으므로, 학습 샘플을 분석하고 그 결과에 따라 가변적으로 임계치를 설정함으로써 보다 나은 결과를 얻을 수 있을 것으로 판단된다.

VI. 결론

특징 정규화는 분류기를 사용하기 위한 전처리 단계로 특징의 스케일 변화에 따른 오류를 줄이기 위해 널리 사용된다. 하지만 일반적인 정규화 방법은 클래스 라벨을 고려하지 않으므로 정규화 결과가 인식률 측면에서 최적임을 보장하지 못하는 문제점이 있다. 이러한 단점을 보완하고자 클래스 라벨을 고려한 교사 랭크 정규화 방식이 소개되었고 이는 다른 방법에 비해 나은 결과를 보였다.

이 논문에서는 교사 랭크 정규화 방법에 샘플 선택을 추가함으로써 보다 나은 인식률을 얻을 수 있는 방법을 소개하였다. 샘플 선택 역시 최근접 이웃의 클래스 라벨을 고려하는 교사 학습 방법을 사용함으로써 이전 방법에 비해 보다 나은 결과를 얻을 수 있었다.

제안한 방법이 기존 방법들에 비해 우수한 성능을 보였지만 데이터 집합에 따라 학습 샘플 선택이 오히려 오류를 증가시키는 경우가 약 30% 발견되었다. 이는 모든 실험에서 고정된 임계치를 사용한 때문으로 데이터 집합에 따라 임계치를 가변적으로 결정함으로써 고정된 임계치 사용에 비해 더 나은 결과를 얻을 수 있을 것으로 판단되며 현재 이에 대해 연구를 진행 중에 있다.

참고문헌

[1] Eunseog Youn and Myong K. Jeong, "Class dependent feature scaling method using naive Bayes classifier for text datamining," Pattern Recognition Letters, Vol. 30, No. 5, pp. 477-485, Apr. 2009.

[2] Soojong Lee and Gyeongyong Heo, "Supervised Rank Normalization for Support Vector Machines," Journal of The Korea Society of Computer and Information, Vol. 18, No. 11, pp. 31-38, Nov. 2013.

[3] Gyeongyong Heo, Choong-Shik Park, and Chang-Woo Lee, "Training Sample and Feature Selection Methods for Pseudo Sample Neural Networks," Journal of The Korea Society of Computer and Information, Vol. 18, No. 4, pp. 19-26, Apr. 2013.

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

[5] Ashis Pradhan, "Support Vector Machine - A Survey," International Journal of Emerging Technology and Advanced Engineering, Vol. 2, No. 8, pp. 82-85, Aug. 2012.

[6] Yvan Saeys, Inaki Inza and Pedro Larranaga, "A review of feature selection techniques in

bioinformatics," Bioinformatics, Vol. 23, No. 19, pp. 2507-2517, Aug. 2007.

〔7〕 Keinosuke Fukunaga, Introduction to Statistical Pattern Recognition, 2nd edition, SanDiego, AcademicPress, 1990

〔8〕 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

허 경 용 (Gyeongyong Heo)

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

최 훈 (Hun Choi)

1996: 충북대학교
전자전자학과 공학사.
2001: 충북대학교
전자공학과 공학석사.
2006: 충북대학교
전자공학과 공학박사.
현 재: 동의대학교 전자공학과 부교수.
관심분야: 계측신호처리, 적응신호처리
Email : hchoi@deu.ac.kr

윤 주 상 (JooSang Youn)

2001: 고려대학교
전기전자전파공학부 공학사.
2003: 고려대학교
전자공학과 공학석사.
2008: 고려대학교
전자컴퓨터공학과 공학박사.
현 재: 동의대학교 멀티미디어공학과 조교수.
관심분야: 이동통신, 사물지능통신
Email : jsyoun@deu.ac.kr