다층퍼셉트론에 의한 불균형 데이터의 학습 방법

Classification of Imbalanced Data Using Multilayer Perceptrons

오상훈
목원대학교 정보통신공학과

Sang-Hoon Oh(shoh@mokwon.ac.kr)

요약

최근에 클래스 분포의 불균형이 심한 데이터의 학습 문제가 그 중요도에 비하여 만족할만한 성능을 얻기 어려운 관계로 관심이 고조되고 있다. 이 문제에 대한 접근 방법은 데이터 레벨의 불균형 해소, 알고리즘 레벨에서의 비용함수 도입, 인식기의 앙상블에 의한 성능향상 등으로 분류된다. 이 논문은 알고리즘 레벨의 접근 방법으로써, 다층퍼셉트론 신경회로망에 고차의 오차함수를 사용하여 불균형 데이터를 학습하는 방법을 제시한다. 즉, 소수클래스의 학습을 강화시키고 다수 클래스의 학습을 약화시키는 형태로 가중치를 변경시킨다. 클래스 불균형이 심한 유방암 검사와 갑상선 진단 데이터의 학습을 통하여 제안한 방법이 MSE(mean-squaerd error), 2단계 방법 및 문턱조정 방법보다 우수함을 확인한다.

■ 중심어 : | 다층퍼셉트론 | 불균형 데이터 | 오차함수 |

Abstract

Recently there have been many research efforts focused on imbalanced data classification problems, since they are pervasive but hard to be solved. Approaches to the imbalanced data problems can be categorized into data level approach using re-sampling, algorithmic level one using cost functions, and ensembles of basic classifiers for performance improvement. As an algorithmic level approach, this paper proposes to use multilayer perceptrons with higher-order error functions. The error functions intensify the training of minority class patterns and weaken the training of majority class patterns. Mammography and thyroid data-sets are used to verify the superiority of the proposed method over the other methods such as mean-squared error, two-phase, and threshold moving methods.

■ keyword : | Multilayer Perceptrons | Imbalanced Data | Error Function |

I. 서 론

클래스 간의 불균형이 심한 데이터의 인식 문제는 응용분야가 아주 많은 영역에 걸쳐 있음에도 쉽게 해결될 수 없기 때문에 많은 관심을 끌고 있다. 이러한 인식문제는 금융사기 검출[1], 의료진단[1][2], 기름방류 감시[2], 원격감시[3], 신용평가[4], 은행 대출심사[5] 등에 다양하게 나타난다. 일반적으로 인식기들이 불균형 데이터를 다룰 경우 성능이 저하되는 것은, 클래스 간의 데이터 양이 비슷하고 오인식 시 각 클래스의 비용도 동일하다는 가정 하에서 인식기들이 개발되었기 때문이다[6].

접수번호 : #090416-003
접수일자 : 2009년 04월 16일
심사완료일 : 2009년 06월 01일
교신저자 : 오상훈, e-mail : shoh@mokwon.ac.kr

이러한 성능저하의 원인을 살펴보면, 첫째로 인식기 학습의 기준이 되는 오차함수가 부적절한 것을 들 수 있다. 예를 들어 클래스 1이 전체데이터의 1%이고 클래스 2가 99%를 차지하는 경우에 전체 인식률이 향상되도록 인식기를 학습시키면, 클래스 1(소수 클래스)을 무시하고 클래스 2(다수 클래스)만 인식하여도 99%의 인식성능을 얻을 수 있기 때문에 클래스 1의 중요한 정보를 찾아내지 않는다[1]. 두 번째로, 클래스가 분포하고 있는 영역이 침해당하는 것을 들 수 있다. 인식기의 학습 과정에서 다수 클래스의 영역이 소수 클래스의 영역을 침해하여 소수 클래스의 인식 성능이 저하된다[1]. 세 번째는, 소수 클래스에 해당하는 샘플의 수가 충분치 않아서 필요한 정보를 얻을 수 없는 것이다[7]. 네 번째, 클래스들이 겹칠 경우, 소수 클래스가 다수 클래스에 묻혀서 분리해낼 수 없다[7].

이 문제를 해결하기 위한 접근 방법은 크게 데이터 레벨, 알고리즘 레벨, 및 앙상블 접근방법으로 구분할 수 있다[7]. 데이터 레벨 접근방법은 언더샘플링[1][8], 오버샘플링[2][9] 및 이들의 조합 방법[9]을 사용하여 불균형 분포인 데이터를 균형이 되도록 변환시키는 방법이다. 알고리즘 레벨 접근방법은 기본 인식기의 학습에 오차함수를 수정하거나[3] 비용(cost) 개념을 도입하여[5][8] 소수 클래스의 학습에 중요도를 주는 방법이다. 앙상블 방법은 기본 인식기로는 만족할 만한 성능을 얻을 수 없기 때문에 여러 인식기의 앙상블 효과에 의해 성능을 개선시키는 방법이다[1][7].

이러한 방법들 중에서 가장 근본이 되는 것은 기본인식기의 성능 개선이다. 특히, 대부분의 연구가 두 개의 클래스를 인식하는 경우를 대상으로 데이터 불균형 문제를 다루고 있으며[1], 여러 개의 클래스를 다룰 경우는 두 개의 클래스만을 다루는 인식기의 조합에 의하여 전체 인식기를 구현한다[10]. 그렇지만, 여러 개의 클래스를 다룰 수 있는 인식기에서 데이터 불균형 문제를 다루도록 하는 것이 보다 더 직접적인 문제 해결 방법이며, 이 분야의 연구도 중요하다[8].

이 논문은 두 클래스뿐만 아니라 여러 클래스의 데이터 불균형 문제로 확장성도 고려하여, MLP(multilayer perceptron)에 고차의 오차함수를 사용하여 데이터 불균형 문제를 해결하는 방법을 제시한다. 이 논문에서 제시하는 방법을 기본 인식기로 사용하면 데이터 레벨 알고리즘 혹은 앙상블 방법에 적용할 수 있다.

이 논문의 구성은 다음과 같다. 먼저 II 장에서 MLP에 고차의 오차함수를 적용하여 데이터 불균형 문제를 해결하는 방법을 제시하고, III 장에서 시뮬레이션으로 제시한 방법의 효용성을 확인한다. 마지막으로 IV 장에서 결론을 맺겠다.

II. 데이터 불균형 문제의 MLP 학습 방법

1. MLP의 학습

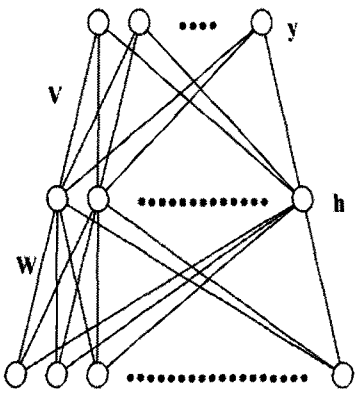

그림 1. MLP(Multilayer Perceptron) 구조

MLP가 [그림 1]과 같이 N개의 입력 노드 $\boldsymbol{x}$와 H개의 중간층 노드 $\boldsymbol{h}$및 M개의 출력 노드 $\boldsymbol{y}$로 구성되어 있다고 하자. 이를 간략히 N-H-M 구조 MLP라 한다. 어떤 N차원의 입력패턴 $\boldsymbol{x}=[x_1,x_2,...,x_N]$이 MLP에 입력되면, j번째 중간층 노드 h_j의 값은

$$h_j = f(\hat{h_j}) = \tanh(\hat{h_j}/2), j = 1,2,...,H \qquad (1)$$

와 같이 주어진다. 여기서 $f(.)$ 는 시그모이드 비선형 함수이며

$$\hat{h_j} = \sum_{i=0}^{N} w_{ji} x_i \qquad (2)$$

는 중간층 노드에 입력되는 가중치 합이다. w_{ji}는 x_i와

h_j를 연결하는 중간층 가중치이며 $x_0=1$로 주어지고 w_{j0}는 바이어스이다. 같은 형태로 k번째 출력 노드에 입력되는 가중치 합은

$$\hat{y_k}=\sum_{j=0}^{H} v_{kj}h_j, k=1,2,...,M \tag{3}$$

이고, v_{kj}는 h_j와 y_k를 연결하는 출력층 가중치이고, $h_0=1$이며 v_{k0}는 바이어스이다. 최종적으로 k번째 출력은

$$y_k=f(\hat{y_k})=\tanh(\hat{y_k}/2), k=1,2,...,M \tag{4}$$

로 주어진다.

이러한 구조의 MLP에 P개의 학습패턴 $\boldsymbol{x}^{(p)}(p=1,2,...,P)$와 이들의 출력층 목표벡터 $t^{(p)}=[t_1^{(p)}, t_2^{(p)}, \ldots, t_M^{(p)}]$가 주어지면 일반적으로

$$E_{MSE}^{out}=\frac{1}{2}\sum_{p=1}^{P}\sum_{k=1}^{M}(t_k^{(p)}-y_k^{(p)})^2 \tag{5}$$

로 주어지는 MSE(mean-squared error)를 최소화시키도록 MLP의 가중치들이 변경된다[11]. 패턴인식 문제에서 출력층 목표벡터의 각 요소 값들은

$$t_k^{(p)}=\begin{cases}+1, \text{if } \boldsymbol{x}^{(p)}\in C_k\\ -1, otherwise\end{cases} \tag{6}$$

와 같이 주어지며, 여기서, C_k 는 클래스 k에 속하는 패턴들의 집합을 나타낸다[12].

그렇지만, 식 (5)로 주어진 MSE를 최소화시키기 위한 EBP(Error Back-Propagation) 학습은 학습속도가 느리다는 단점이 많이 지적되었다[13]. 이의 해결책으로 CE(Cross-Entropy) 오차함수[13]가 제안되었으며, 이 보다 더 나은 성능을 지닌 nCE(n-th order extension of CE) 오차함수[12] 역시 제안되었다. nCE 오차함수는

$$E_{nCE}^{out}=-\sum_{p=1}^{P}\sum_{k=1}^{M}\int\frac{t_k^{(p)^{n+1}}(t_k^{(p)}-y_k^{(p)})^n}{2^{n-2}(1-y_k^{(p)^2})}dy_k^{(p)} \tag{7}$$

와 같이 주어지며, MLP의 출력층 각 노드에 연결된 출력층 가중치는

$$\Delta v_{kj}=-\eta\frac{\partial E_{nCE}^{out}}{\partial v_{kj}}=\eta\delta_k^{(p)}h_j^{(p)} \tag{8}$$

에 따라 학습된다. 여기서,

$$\delta_k^{(p)}=-\frac{\partial E_{nCE}^{out}}{\partial \hat{y_k^{(p)}}}=\frac{t_k^{(p)^{n+1}}(t_k^{(p)}-y_k^{(p)})^n}{2^{n-1}} \tag{9}$$

이다. 또한, 중간층 노드에 연결된 중간층 가중치는

$$\Delta w_{ji}=-\eta\frac{\partial E^{out}}{\partial w_{ji}}=\eta f'(\hat{h_j^{(p)}})x_i^{(p)}\sum_{k=1}^{M}v_{kj}\delta_k^{(p)} \tag{10}$$

에 따라 학습된다. 이렇게 nCE를 최소화 시키는 방식의 학습에 의해 얻어지는 성능이 MSE를 사용한 경우보다 좋은 이유는 바로 식 (9)로 주어진 출력층의 $\delta_k^{(p)}$ 신호 때문이다. 이 특성을 좀 더 고찰하기 위하여 $t_k^{(P)}=1$인 경우의 $\delta_k^{(p)}$를 [그림 2]에 그렸다.

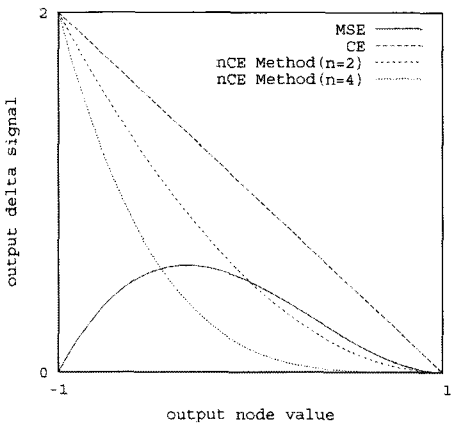

그림 2. 출력층 $\delta_k^{(p)}$ 신호

MSE를 이용하여 학습하는 경우는

$$\delta_k^{(p)}=-\frac{\partial E_{MSE}^{out}}{\partial \hat{y_k^{(p)}}}=(t_k^{(p)}-y_k^{(p)})f'(\hat{y_k^{(p)}}) \tag{11}$$

와 같이 출력노드의 $\delta_k^{(p)}$가 계산된다. 이 경우는 [그림 2]에서 보는 바와 같이 $t_k^{(P)}=1$이고 $y_k^{(P)}\approx -1$인 경우 $(t_k^{(p)}-y_k^{(p)})$ 가 큼에도 불구하고 시그모이드 함수의 기울기 $f'(\hat{y_k^{(p)}})$ 가 작은 값을 지니게 되어 $\delta_k^{(p)}$의 값이 작아진다. 결국 식 (8)에서 Δv_{kj} 가 아주 작게 되어 학습이 잘 되지 않는다. 이 상태를 출력노드가 부적절하게 포화되었다고 한다[14]. 한편, nCE 오차함수를 이용한 경우는 $\delta_k^{(p)}$가 식 (9)와 [그림 2]에서 보는 바와

같이 $(t_k^{(p)}-y_k^{(p)})^n$ 에 비례하므로 부적절한 포화 현상 없이 학습이 잘 이루어진다[12].

2. 불균형 데이터의 인식문제 학습 방법

이 논문은 MLP의 학습에서 nCE 오차함수의 차수 n에 따라 $\delta_k^{(p)}$ 의 모양이 변하는 것을 이용하여 불균형 데이터의 인식 문제에 대한 해결책을 제시하고자 한다. nCE 오차함수에 의해 MLP를 학습시키는 경우, [그림 2]에서 보는 바와 같이 $\delta_k^{(p)}$의 모양은 $(t_k^{(p)}-y_k^{(p)})^n$ 으로 나타난다. 즉, $n=1$인 경우는 $\delta_k^{(p)}=(t_k^{(p)}-y_k^{(p)})$ 이므로 $n\geq 2$인 경우보다 상대적으로 큰 값을 지녀 출력층 가중치가 많이 변하게 된다. n이 커짐에 따라 목표값과 출력값의 차이 $(t_k^{(p)}-y_k^{(p)})$ 는 같더라도 $\delta_k^{(p)}$는 점점 더 작은 값을 지니게 되고, 출력층의 가중치 변경량은 이에 비례하여 줄어들게 된다. 즉, n을 어떻게 정해주느냐에 따라 가중치의 변경량이 증가하기도 하고 감소하기도 한다.

패턴수가 적은 "소수 클래스(클래스 1)" 데이터와 상대적으로 패턴수가 아주 많은 "다수 클래스(클래스 2)" 데이터가 주어진 두 클래스(two-class) 패턴인식 문제를 고려하자. MLP는 두 개의 출력노드를 가지고 있으며, 소수클래스에 속하는 패턴은 출력 목표벡터가 [1,-1]로 주어지고, 다수클래스에 속하는 패턴의 출력 목표 벡터는 [-1,1]로 주어진다고 하자. 이때, 소수 클래스 패턴은 첫 번째 출력노드의 목표값이 1 이므로, 이 첫 번째 출력노드 y_1를 소수 클래스의 목표노드(target node)라고 한다. 역시, y_2는 다수 클래스의 목표노드가 된다. 인식단계에서, 어떤 패턴이 MLP에 입력되면 식 (1)-(4)에 의해 출력노드의 값들을 계산한 후, Max. 법칙에 따라 $y_1 > y_2$이면 클래스 1에 속한다고 결정한다. 반대의 경우에는 클래스2로 결정한다.

데이터 불균형 문제를 MLP에 학습을 시키는 단계에서, 소수 클래스에 해당하는 패턴들은 데이터의 수가 적기 때문에 소수클래스의 목표노드 y_1에 연결된 가중치들의 변경 시 출력층의 δ_k 신호를 크게 발생시켜서 학습을 강화시켜주어야 할 것이다. 그렇지만, CE를 적용하게 되면 목표값에 근접한 경우에도 가중치를 많이 변경시키므로 학습의 안정성이 저하된다[12]. 따라서, y_1에 연결된 가중치들의 학습에는 nCE(n=2)를 적용하는 것이 부적절한 포화 현상이 없이 학습이 안정적으로 이루어지도록 할 것이다.

한편, 다수 클래스의 목표노드 y_2에 연결된 가중치들을 학습시키는 경우에는 nCE 오차함수의 차수 n을 높게 설정해주어 가중치의 변경량이 조금씩 조정되게 하여야 한다.

이러한 방식을 오차함수로 표현하면

$$E=-\sum_{p=1}^{P}\left[\int \frac{t_1^{(p)^{n-1}}(t_1^{(p)}-y_1^{(p)})^n}{2^{n-2}(1-y_1^{(p)^2})}dy_1^{(p)} + \int \frac{t_2^{(p)^{m-1}}(t_2^{(p)}-y_2^{(p)})^m}{2^{m-2}(1-y_2^{(p)^2})}dy_2^{(p)}\right] \tag{12}$$

와 같이 된다. 여기서, n=2는 소수 클래스 목표노드에 적용되는 nCE오차함수의 차수이며, m은 다수 클래스 목표노드에 적용되는 nCE 오차함수의 차수이다. 식 (12)의 $\widehat{y_k^{(p)}}$ 에 대한 미분으로 출력층 노드의 $\delta_k^{(p)}$를 계산하면

$$\delta_k^{(p)}=\begin{cases}\dfrac{t_k^{(p)^{n-1}}(t_k^{(p)}-y_k^{(p)})^n}{2^{n-1}}, & \text{if } k=1\\ \dfrac{t_k^{(p)^{m-1}}(t_k^{(p)}-y_k^{(p)})^m}{2^{m-1}}, & \text{if } k=2\end{cases} \tag{13}$$

와 같이 된다. 따라서, 소수 클래스 목표노드에 해당하는 $\delta_k^{(p)}$는 강하게 발생되고, 다수 클래스 목표노드에 해당하는 $\delta_k^{(p)}$는 약하게 발생되어 $y_1 > y_2$인 영역이 다수 클래스의 목표노드 y_2에 의해 침해를 덜 받도록 한다.

한편, 소수 클래스 목표노드 y_1은 목표값이 "-1"로 주어지는 경우가 "1"로 주어지는 경우보다 훨씬 많으며, 다수 클래스 목표노드 y_2는 목표값이 "1"로 주어지는 경우가 "-1"로 주어지는 경우보다 훨씬 많으므로, 각 $\delta_k^{(p)}$에서 목표값이 1인 경우와 -1인 경우의 수가 불균형을 이루게 된다. 이의 균형을 위하여

$$\gamma=\frac{P_1}{P_2} \tag{14}$$

의 비율만큼 목표값이 "-1"인 $\delta_1^{(p)}$와 목표값이 "1"인 $\delta_2^{(p)}$를 조정한다. 즉,

$$\delta_1^{(p)} \rightarrow \begin{cases} \delta_1^{(p)}, & \text{if } t_1^{(p)} = 1 \\ \gamma\delta_1^{(p)}, & \text{if } t_1^{(p)} = -1 \end{cases} \quad (15)$$

와

$$\delta_2^{(p)} \rightarrow \begin{cases} \gamma\delta_2^{(p)}, & \text{if } t_2^{(p)} = 1 \\ \delta_2^{(p)}, & \text{if } t_2^{(p)} = -1 \end{cases} \quad (16)$$

같이 변경된다. 여기서, P_1은 소수 클래스의 학습패턴 수이며, P_2는 다수 클래스의 학습패턴 수이다. 제안한 학습 방법을 알고리즘 1에 정리하였다.

알고리즘 1. 데이터 불균형 문제의 학습 방법

1. MLP의 초기 가중치를 임의로 설정함.
2. 하나의 학습패턴을 선정함.
3. 선정된 학습패턴을 MLP에 입력함.
4. 식(1)-(4)에 의해 중간층/출력층 노드 값을 구함.
5. 식 (13)에 따라 출력층의 $\delta_k^{(p)}$를 계산함.
6. 식(15)와 (16)에 따라 $\delta_k^{(p)}$를 조정함.
7. 식 (8)과 (10)에 따라 가중치를 변경함.
8. 다음 학습패턴을 선정한 후 과정 3-7을 수행함.

III. 시뮬레이션

1. 시뮬레이션 대상 문제

제안한 방법의 효용성을 확인하기 위하여 데이터의 불균형이 아주 심한 갑상선 진단과 유방암 검사 데이터를 대상으로 시뮬레이션 하였다. 갑상선 진단 데이터(Ann-thyroid)는 UCI 기계학습 데이터 베이스[15]에서 가져왔으며, 유방암 진단(mammography) 데이터는 N. V. Chawla 교수로부터 구하였다[9]. 표 2는 각 데이터베이스의 명칭과 데이터 수를 나타내고 있는데, 이 표에서 보는 바와 같이 이 논문에서 시뮬레이션 대상으로 삼은 문제들은 소수 클래스의 비율이 아주 작다.

표 2. 시뮬레이션 대상 데이터베이스

Data Set	소수 클래스 패턴수	다수 클래스 패턴수	전체 패턴수	소수 클래스 비율
Ann-thyroid13	93	3,488	3,581	2.60%
Ann-thyroid23	191	3,488	3,679	5.19%
Mammography	260	10,923	11,183	2.32%

먼저, 갑상선 진단(Ann-thyroid) 데이터는 3개의 클래스(클래스 1, 2, 3)로 이루어져 있으며, 각 클래스의 데이터는 학습 및 시험 패턴으로 구성되어 있다. 여기서, class3는 정상에 해당한다. 데이터 불균형 문제에 대한 제안한 방법의 효용성 검증을 위하여 이를 각각 소수-다수 클래스로 구성된 두 개의 클래스 인식 문제로 변형하였는데, Ann-thyroid13은 클래스 1과 클래스 3 데이터만을 대상으로 시뮬레이션 하는 것이며, Ann-thyroid23은 클래스 2와 클래스 3 데이터를 대상으로 시뮬레이션 하는 것이다. 물론, 클래스 3이 다수 클래스에 해당한다. [표 2]에 나와 있는 패턴 수는 학습패턴의 수이며, 이와 별도로 주어지는 시험패턴의 수는 각각 클래스 1이 73개, 클래스 2가 177개, 그리고 클래스 3이 3178개이다. Ann-thyroid13과 Ann-thyroid23 문제의 시뮬레이션에서는 학습 패턴으로 MLP를 학습시키며, 시험패턴의 인식률로 각 방법의 성능을 비교한다.

한편, 유방암 검사 (mammography) 데이터는 학습 및 시험패턴의 구분이 없다. 따라서, 이 문제의 시뮬레이션에서는 "1-out of-5 validation 방법"을 사용하여 각 방법의 성능을 비교한다. 여기서, "1-out of-5 validation 방법"이란 주어진 데이터를 임의의 5개 집합으로 구성한 후, 한 집합에 속하는 데이터는 시험패턴으로 사용하고 나머지 집합의 모든 데이터는 학습패턴으로 사용하는 방법을 각각 다른 시험패턴을 대상으로 5번에 걸쳐 실시한 후, 얻어진 데이터의 평균으로 성능을 측정하는 방법이다.

2. 성능비교 학습 방법

성능비교를 위한 MLP의 학습 시뮬레이션 방법은 MSE를 이용한 일반적인 방법, 2단계(two-phase) 방법[3], 문턱 조정(threshold moving) 방법[8], 그리고 이 논문에서 제안한 방법(n=2, m=4)으로 설정하였다. MLP의 구조(N-H-M)는 Ann-thyroid13 및 Ann-thyroid23 문제의 경우 21-16-2 구조이며, Mammography 문제의 경우 6-4-2 구조이다. 각 시뮬레이션 방법에서 공정한 비교를 위하여 $E\{\eta|\delta_k^{(p)}|\}$가 같은 값을 지니도록 학습률 η 를 정하였다. MSE 방법, 2

단계 방법, 문턱조정 방법은 $\eta = 0.006$이며, 이 논문에서 제안한 방법은 $0.001 \times [(n+1)+(m+1)]/2$이다.

2단계 방법[3]의 시뮬레이션에서는 단계 1에서 단계 2로 넘어가는 기준에 해당하는 "*T*"값을 0.4, 0.2, 0.1, 0.05, 0.02, 0.01로 사용하여 시뮬레이션한 후 가장 좋은 결과를 보이는 데이터를 [표 3]-[표 5]에 사용하였다. 문턱 조정 방법의 경우는 여기서 사용한 MLP의 출력이 (-1,+1)사이의 값을 지니므로 (0,1) 사이의 값이 되도록 1차 변환을 한 후에 소수 클래스 목표 노드에 해당하는 출력에 "*TH*"를 곱하였다. 즉,

$$y^*_k = \begin{cases} (y_k+1)/2, & \text{if } k=2 \\ TH \times (y_k+1)/2, & \text{if } k=1 \end{cases} \quad (17)$$

같이 출력노드 값을 변환시킨 후, 그 결과 값을 이용하여 인식결과를 판단하였다. "*TH*"는 2, 3, 4,...로 적용하였으며, G-Mean이 제일 좋은 결과를 보이는 데이터를 표 작성에 사용하였다.

Ann-thyroid13과 Ann-thyroid23 문제의 시뮬레이션에서는 MLP의 초기 가중치를 $[-1 \times 10^{-3}, 1 \times 10^{-3}]$에서 균일분포를 지니도록 임의로 설정하여 MLP를 20000 epoch 동안 학습시키면서 10 epoch 단위로 시험패턴에 대한 인식률을 Max. 법칙에 따라 측정하였다. 이를 초기 가중치가 다르게 설정된 9번에 대하여 실시한 후 그 평균 데이터를 구하였다. Mammography의 경우는 초기 가중치를 위와 같은 방법으로 9번 다르게 설정하여 MLP를 30000 epoch 동안 학습하는 것을 "1-out of-5 validation 방법"에 따라 실시하였다. 즉, 시험패턴으로 설정되는 가지 수가 5 가지이고, 이 각 가지 수에 대하여 9 가지의 다른 초기 가중치를 설정하여 MLP를 학습하였으므로, 각 학습 방법에 대하여 총 45 번의 시뮬레이션 데이터를 얻게 된다. 이 데이터의 평균치로 성능을 비교한다. 여기서 "1 epoch"은 모든 학습패턴에 대하여 가중치 변경이 한번 씩 이루어진 것을 나타내는 단위이다.

일반적으로 패턴인식 문제의 경우 전체패턴에 대한 인식률(total accuracy)을 성능 평가의 기준으로 사용하지만, 데이터 불균형이 심한 경우 이 전체인식률은 각 클래스의 인식률을 제대로 반영한 방법이 아니다. 왜냐하면, 소수 클래스의 인식률이 A_1이고 다수 클래스의 인식률이 A_2인 경우 전체 인식률에서 A_2가 차지하는 비중이 월등이 크므로 A_1이 좋고 나쁜 것이 전체 인식률에 잘 반영되지 않는다. 따라서, 이 논문에서는 A_1과 A_2의 기하학적 평균치(G-Mean: Geometric Mean) $\sqrt{A_1 \times A_2}$를 사용하여 각 방법의 성능을 비교한다[1].

3. 시뮬레이션 결과 비교

Ann-thyroid13 문제에 대한 시뮬레이션 결과를 [표 3]에 정리하였다. 여기서, 표의 각 데이터는 20000 epoch 동안의 학습과정에서 10 epoch 단위로 시험패턴에 대하여 구한 데이터의 평균치를 구한 후, G-Mean이 가장 좋은 값을 보이는 epoch에서 데이터를 추출한 것이다. MSE 방법은 예상한 바와 같이 소수 클래스의 인식률이 낮고 다수 클래스의 인식률이 높게 나타났으며, 그 결과 비록 전체 인식률은 높더라도 G-Mean은 낮다. 2단계 방법은 비록 MSE 보다 개선되어 소수 클래스의 인식률이 향상되었지만 여전히 소수 클래스와 다수 클래스의 인식률 차이가 크다. 문턱조정 방법은 MSE 및 2단계 방법보다 소수클래스 인식률 및 G-Mean이 향상되었다. 한편, 제안한 방법은 소수 클래스의 인식률이 많이 향상되었으며, 소수 클래스와 다수 클래스의 인식률 차이도 많이 줄어들었고 G-Mean도 크게 향상되었음을 볼 수 있다.

표 3. Ann-thyroid13 문제 측정 성능 비교

학습방법	소수 클래스 인식률	다수 클래스 인식률	전체 인식률	G-Mean
MSE	86.1%	99.4%	99.1%	92.5%
Two-Phase (T=0.05)	89.2%	99.1%	98.9%	94.0%
Th. Moving(TH=8)	91.8%	99.0%	98.8%	95.3%
Proposed Method	94.9%	98.8%	98.7%	96.8%

[표 4]의 Ann-thyroid23 문제와 [표 5]의 Mammography 문제에서도 양상이 비슷하게 나타남을 볼 수 있다.

MSE는 각각의 학습패턴을 대상으로 가중치 변경량을 동일한 방식으로 계산한 후 이에 따라 가중치를 변경시키므로, 다수 클래스에 속한 패턴에 관련된 가중치

변경이 많이 이루어진다. 그 결과 다수 클래스의 인식률은 좋은 반면 소수 클래스의 인식률은 낮게 나오며 G-Mean은 가장 낮은 결과를 보였다.

표 4. Ann-thyroid23 문제 측정 성능 비교

학습방법	소수 클래스 인식률	다수 클래스 인식률	전체 인식률	G-Mean
MSE	85.9%	98.5%	97.8%	91.9%
Two-Phase (T=0.05)	84.6%	98.8%	98.0%	91.4%
Th. Moving(TH=8)	97.8%	94.3%	94.4%	96.0%
Proposed Method	97.4%	96.3%	96.3%	96.8%

표 5. Mammography 문제 측정 성능 비교

학습방법	소수 클래스 인식률	다수 클래스 인식률	전체 인식률	G-Mean
MSE	59.4%	99.6%	98.6%	76.8%
Two-Phase(T=0.2)	83.8%	95.3%	95.0%	89.3%
Th. Moving(TH=15)	85.2%	94.5%	94.3%	89.7%
Proposed Method	86.1%	93.6%	93.5%	89.7%

2단계 방법은 단계 1에서 소수 클래스의 패턴 수가 작은 것을 고려하여, 소수 클래스에 속한 패턴의 학습 시 가중치 변경량을 더 강화시켜준다. 그리고, 어느 정도 학습이 이루어져 각 클래스의 오차가 T보다 작아지면, 단계 2로 넘어가서 MSE에 의한 일반적인 방법에 따라 학습이 이루어진다. 비록 단계 1에서 소수 클래스의 학습을 강화시켜주어 소수 클래스의 인식률이 개선되었지만 전체적인 성능 개선은 부족하다.

문턱조정 방법은 일반적인 MSE 방법에 따라 학습하며 인식의 절차에서 소수 클래스 목표 노드에만 TH를 곱하는데, TH 값을 증가시킬수록 소수 클래스의 인식률은 향상되고 다수 클래스의 인식률은 저하될 것이다. 따라서, 적절한 TH 값을 구하기 위하여 여러 TH 값에 대하여 시뮬레이션 한 후 가장 좋은 결과를 취할 수밖에 없다.

여기서 제안한 방법은 $\delta_k^{(p)}$를 소수 클래스 목표노드의 학습에서는 강화시키고 다수 클래스 목표노드의 학습에서는 약화시켜, 출력노드 값으로 이루어진 공간에서 다수 클래스의 영역이 소수 클래스 영역을 침해하는 것을 줄여준다. 또한 목표 값이 "1"과 "-1"인 경우의 균형을 이루어주도록 하는 효과를 추가하여 소수 클래스의 학습이 잘 이루어지도록 하는 효과를 얻은 것이다.

여기서 더 나아가서 다수 클래스의 영역과 소수 클래스의 영역이 서로 침해를 받지 않아서 다수 클래스 인식률의 저하가 없이 소수 클래스의 인식률이 향상되는 학습 방법을 찾아내면 훨씬 더 좋은 인식기를 구현할 수 있을 것이다.

IV. 결 론

이 논문에서는 클래스 간의 데이터 불균형이 심한 패턴인식 문제의 기본 인식기 학습 방법으로 nCE 오차함수를 근거로 소수 클래스와 다수 클래스에 차수를 달리하는 방법을 제안하였다. 제안한 방법은 소수클래스의 인식률을 향상시켜 다수 클래스와 소수 클래스 간의 인식률 격차를 줄여주는 특징이 있다.

갑상선 진단 및 유방암 검사 문제의 시뮬레이션에서 MSE 방법은 가장 낮은 성능을 보였으며, 2단계 방법은 소수 클래스의 인식률을 다소 향상시켰다. 문턱조정 방법은 적절한 문턱 값의 선정으로 소수클래스 인식률이 많이 향상되었다. 이 논문에서 제안한 방법은 소수 클래스의 인식률을 향상시켰으며 기하학적 평균치도 가장 좋은 결과를 보였다.

여기에서 제안한 인식기 학습 방법은 데이터의 처리를 통하여 균형을 이루도록 하는 데이터 레벨 알고리즘에서 인식기로 활용할 수 있다. 또한, 기본 인식기의 성능이 앙상블 방법의 전체 성능에도 영향을 미치므로, 여기에서 제안한 방법을 앙상블 방법의 기본 인식기로도 사용할 수 있다.

참 고 문 헌

[1] P. Kang and S. Cho, "EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems," Proc. ICONIP 2006, pp.837-846.

[2] M. Kubat, R. C. Hilte, and S. Matwin, "Machine

learning for the detection of oil spills in satellite radar images," Machine Learning, Vol.30, pp.195-215, 1998.

[3] L. Bruzzone and S. B. Serpico, "Classification of imbalanced remote-sensing data by neural networks," Pattern Recognition Letters, Vol.18, pp.1323-1328, 1997.

[4] Y.-M. Huang, C.-M. Hung, and H. C. Jiau, "Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem," Nonlinear Analysis, Vol.7, pp.720-747, 2006.

[5] H. Zhao, "Instance weighting versus threshold adjusting for cost-sensitive classification," Knowl. Inf. Syst., Vol.15, pp.321-334, 2008.

[6] F. Provost and T. Fawcett, "Robust classification for imprecise environments," Machine Learning, Vol.42, pp.203-231, 2001.

[7] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," Pattern Recognition, Vol.40, pp.3358-3378, 2007.

[8] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," IEEE Trans. Knowledge and Data Eng., Vol.18, pp.63-77, 2006.

[9] N. V. Chawla, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, Vol.16, pp.321-357, 2002.

[10] H.-C. Kim, "Constructing support vector machine ensemble," Pattern Recognition, Vol.36, pp.2757-2767, 2003.

[11] D. E. Rumelhart and J. L. McClelland, Parallel Distributed Processing. MIT Press, Cambridge, MA, 1986.

[12] S.-H. Oh, "Improving the error back-propagation algorithm with a modified error function," IEEE Trans. Neural Networks, Vol.8, pp.799-803, 1997.

[13] A. van Ooyen and B. Nienhuis, "Improving the convgence of the back-propagation algorithm," Neural Networks, Vol.5, pp.465-471, 1992.

[14] Y. Lee, S.-H. Oh, and M. W. Kim, "An analysis of premature saturation in back-propagation learning," Neural networks, Vol.6, pp.719-728, 1993.

[15] UCI Machine Learning Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html

저 자 소 개

오 상 훈(Sang-Hoon Oh) 정회원

- 1986년 2월 : 부산대학교 전자공학과(공학사)
- 1988년 2월 : 부산대학교 대학원 전자공학과(공학석사)
- 1999년 8월 : 한국과학기술원 전기 및 전자공학과(공학박사)
- 1988년 1월 ~ 1989년 12월 : LG 반도체(주) 사원
- 1990년 1월 ~ 1998년 6월 : 한국전자통신연구원 기초기술연구부 및 이동통신기술연구소 선임연구원
- 1999년 8월 ~ 2000년 3월 : 한국과학기술원 뇌과학연구센터 연구원
- 2000년 4월 ~ 2000년 10월 : 일본 RIKEN, Brain Science Institute, Research Scientist
- 2000년 10월 ~ 20001년 10월 : (주)엑스텔테크놀리지 연구소장
- 2001년 11월 ~ 2002년 2월 : 한국과학기술원 초빙교수
- 2002년 3월 ~ 현재 : 목원대학교 정보통신공학과 부교수
- 2008년 8월 ~ 현재 : 조지아공대 College of Computing, Div. Computational Science and Eng. 방문교수

<관심분야> : 지능정보처리 알고리즘 개발 및 IT에의 응용, 독립성분분석, NMF, 패턴인식, 음성신호 처리