유전자 알고리즘을 결합한 Support Vector Machine의 화자인증에서의 성능분석

최 우 용(崔 祐 溶), 이 경 희, 반 성 범
한국전자통신연구원 정보보호연구본부
전화 : (042) 860-1680 / 팩스 : (042) 860-5022
H.P 번호 : 018-563-7242

Speaker Verification System Using Support Vector Machine with Genetic Algorithms

Woo-Yong Choi, Kyunghee Lee, Sung Bum Pan
Information Security Research Division,
Electronics and Telecommunications Research Institute
E-mail : {wychoi4, uniromi, sbpan}@etri.re.kr

Abstract

Voice is one of the promising biometrics because it is one of the most convenient ways human would distinguish someone from others. The target of speaker verification is to divide the client from imposters. Support Vector Machine(SVM) is in the limelight as a binary classifier, so it can work well in speaker verification. In this paper, we combined SVM with genetic algorithm(GA) to reduce the dimensionality of input feature.

Experiments were conducted with Korean connected digit database using different feature dimensions. The verification accuracy of SVM with GA is slightly lower than that of SVM, but the proposed algorithm has greater strength in the memory limited systems.

I. 서론

최근 정보통신 기술이 급속도로 발전하고 인터넷의 이용이 확산됨에 따라 사용자 인증에 대한 관심이 높아지고 있다. 90년대까지 사용자 인증 수단으로 많이 사용되던 패스워드나 PIN (Personal Identification Number) 등은 타인에게 노출되거나 잊어버리는 등의 문제점을 가지고 있어 이를 대체하거나 보완하기 위한 방법으로 개인의 고유한 생체정보를 이용한 사용자 인증 방법에 관한 연구가 진행되고 있다.

이러한 생체인식 방법 중에서 사람의 음성을 이용하는 화자인식은 입력장치로 비교적 값이 싸고 손쉽게 구할 수 있는 마이크를 사용하며, 다른 생체인증 방법에 비해서 사용자의 거부감이 적다는 장점이 있다. 전통적으로 많이 사용되어온 화자인증 방법으로는 Dynamic Time Warping(DTW)[1], Hidden Markov Model(HMM)[2], Vector Quantization(VQ)[3], Gaussian Mixture Model(GMM)[4] 등이 있다. DTW와 HMM은 주로 문맥종속 시스템에 많이 쓰이고, VQ와 GMM은 문맥독립 시스템에 많이 쓰이는 방법이다. 일반적으로 이러한 알고리즘들에 사용되는 특징벡터들은 윈도우를 시간축으로 쉬프트하면서 추출한다. 따라서 비밀번호로 사용하는 음성의 길이가 길어지면 특징벡터의 수가 급격히 늘어나게 되므로 스마트 카드와 같은 메모리 제약이 따르는 시스템에는 사용할 수 없게 된다.

본 논문에서는 최근 패턴 분류에서 주목을 받고 있는 Support Vector Machine(SVM)[5]을 화자인증에 적용하였다. 또한 유전자 알고리즘(GA)[6]에 의한 특징선택 과정과 SVM 분류기를 이용한 인증과정을 결합한 새로운 화자 인증 시스템을 제안하였다. SVM 분류기의 입력 벡터로서 화자 특징벡터 전체를 사용하는 대신에, GA과정을 통하여 선택된 우수한 식별력을 가진 특징 집합을 사용함으로써 메모리 사용량을 감소시켰다.

본 논문의 구성은 2장에서 GA를 통한 특징벡터 선택방법 및 SVM을 이용한 화자인증 시스템에 대해서 설명하고, 3장에서는 실험에 사용된 데이터베이스 및 실험 결과를 기술하였으며, 마지막으로 4장에서 결론 및 향후과제를 제시하였다.

II. 화자인식 시스템

본 장에서는 GA를 이용하여 특징들을 선택하는 과정과 선택된 특징들을 이용하여 SVM을 생성하는 과정에 대해서 기술한다.

A. GA를 이용한 특징들의 선택

GA를 이용한 SVM의 생성 과정을 그림 1에 나타내었다. 먼저 학습에 사용할 음성들에 대하여 전처리 및 특징 추출 과정을 수행한다. 다음 과정은 GA를 이용한 진화 과정으로 훈련데이터와 튜닝데이터를 이용하여 각 염색체들을 평가하고 그 평가값으로부터 성능이 우수한 SVM의 입력 특징들을 찾는 과정을 반복하게 된다. 여기서 각 염색체는 특징집합의 부분집합을 나타내고, 따라서 각 염색체의 길이는 전체 특징의 개수와 같다. 즉 각 염색체의 하나의 비트가 각각의 특징의 사용 여부를 나타낸다. 비트의 값 1은 그 비트에 해당하는 특징이 선택되었음을 나타내고, 0은 선택되지 않았음을 나타낸다.

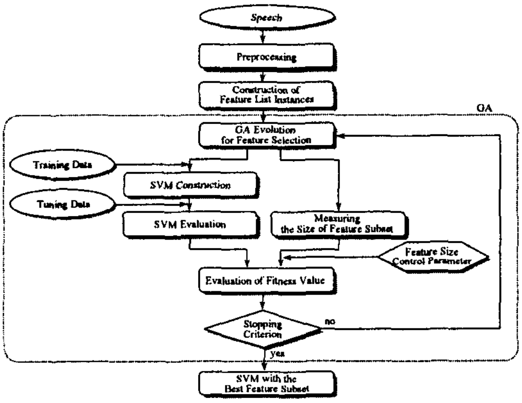

그림 1. GA를 이용한 SVM의 생성 과정

각 염색체의 평가값(Fittness Value)을 계산하기 위하여, 우선 학습 데이터들의 전체 특징값들 중에서 각 염색체로 표현된 특징들의 부분 집합에 대응되는 특징값들로만 이루어진 데이터들을 입력 벡터로 한 SVM을 생성한다. 생성된 SVM에 대하여 튜닝 데이터들을 적용한 인식률과 특징 부분집합의 크기를 평가값으로 이용한다. 즉, 우수한 분리능력과 작은 크기의 특징 부분집합에 더 높은 선호도를 준다. 이러한 과정은 가장 좋은 특징 부분집합을 찾을 때까지 반복을 통하여 진화해 간다. 진화 과정에서 단순교차를 사용하였으며, 엘리트 보존 전략을 이용하여 실험하였다.

B. SVM을 이용한 화자인증 시스템

SVM의 목적은 서로 다른 두개의 class를 분류하는 hyperplane을 설계하는 것으로, structural risk minimization 기법에 그 기초를 두고 있다[7][8]. 최근에 패턴인식 분야에서 활발하게 연구되고 있는 SVM 알고리즘을 간단히 설명하면 아래와 같다.

훈련데이터가 다음과 같이 주어졌다고 가정하자.

$$(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N) \in \mathbf{R}^d \times \{\pm 1\} \quad (1)$$

여기서 $\mathbf{x}_i$는 입력패턴이고, y_i는 그 결과값이다. 만약 두 class가 선형적으로 분리가능하다면 식 (2)와 같은 hyperplane에 의해서 두 class를 구분할 수 있다.

$$\mathbf{w}^T\mathbf{x} + b = 0 \quad (2)$$

SVM은 이러한 *hyperplane* 중에서 hyperplane과 가장 가까운 데이터 포인트와의 거리를 최대로 하는 hyperplane을 찾는 것으로, 식 (4)와 같은 제약조건 하에서 식 (3)을 최소로 하는 가중치벡터 w를 찾는 것이다.

$$\min \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} \quad (3)$$

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, \cdots, N \quad (4)$$

위 식을 라그랑지 승수법을 이용하여 다시 쓰면 다음과 같은 최적화 문제를 푸는 것과 같게 된다.

$$\max \quad Q(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \quad (5)$$

$$\sum_{i=1}^{N}\alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0 \quad \text{for} \quad i = 1, \cdots, N \quad (6)$$

$Q(\alpha)$를 최대로 하는 α를 α_o라고 하면 최적 가중치 벡터 $\mathbf{w}_o$는

$$\mathbf{w}_o = \sum_{i=1}^{N}\alpha_{o,i} y_i \mathbf{x}_i \quad (7)$$

가 되고, 최적 바이어스 b_o는

$$b_o = 1 - \mathbf{w}_o^T\mathbf{x}^{(s)} \quad \text{for} \quad y^{(s)} = 1 \quad (8)$$

를 이용하여 구할 수 있다. 여기서 $\mathbf{x}^{(s)}$는 라그랑지 승수가 0이 아닌 입력패턴이다.

지금까지는 입력공간에서의 데이터들이 선형적으로 분리가능한 경우에 대해서 살펴보았는데, 그렇지 않은 경우에는 보다 높은 차원의 공간(특징공간)으로의 비선형변환을 통하여 데이터를 선형적으로 분리가능하게

만들 수 있다(Cover's theorem on the separability of patterns). 입력공간에서 특징공간로의 비선형 변환을 $\varphi(\mathbf{x})$ 라 두면 특징공간에서의 최적의 hyperplane은

$$\sum_{i=1}^{N} \alpha_i y_i \varphi^T(\mathbf{x})\varphi(\mathbf{x}_i) = 0 \tag{9}$$

와 같이 구할 수 있다. 그러나 특징공간은 매우 높은 차원의 공간이므로 실제적으로는 $\varphi^T(\mathbf{x})\varphi(\mathbf{x}_i)$를 직접 구하지 않고 Mercer의 조건을 만족하는 커널함수로 치환하여 계산하게 된다. 즉, 식 (9)는 다음의 수식으로 표현된다.

$$\sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) = 0 \tag{10}$$

커널함수에는 여러 가지가 있으나 SVM에서 많이 사용되는 커널함수는 다음과 같다.

- Polynomial kernel

$$k(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}^T\mathbf{x}_i + 1)^p \tag{11}$$

- Radial Basis Function (RBF) kernel

$$k(\mathbf{x}, \mathbf{x}_i) = exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}_i\|^2\right) \tag{12}$$

- Tangent hyperbolic kernel

$$k(\mathbf{x}, \mathbf{x}_i) = tanh(\beta_0\mathbf{x}^T\mathbf{x}_i + \beta_1) \tag{13}$$

본 논문에서는 식 (12)의 RBF 커널을 사용하였다.

III. 실험 결과

A. 데이터베이스

실험에 사용된 데이터베이스는 ETRI 음성정보연구센터에서 구축한 한국어 헤드셋 화자인식용 음성DB[9]를 사용하였다. 본 데이터베이스는 사무실 환경에서 중/저가 헤드셋을 이용하여 250명(100명-주차, 100명-월차, 50명-3개월차)의 화자가 발성한 2연 숫자, 4연 숫자, 문장으로 구성된 음성 데이터베이스로, 한 화자당 동일한 목록을 5회 발성하고, 주차/월차/3개월차로 구분하여 4회 반복한 음성을 포함하고 있다. 본 논문에서는 발성목록 1번에 해당하는 28명의 화자가 10개의 4연 숫자를 5회씩 주차별로 4회 반복한 음성데이터를 사용하여 실험하였다.

각각의 데이터들은 SVM을 구성하는 데 사용할 학습 데이터셋, 평가함수를 계산하는 데 사용할 튜닝 데이터셋, 학습이 끝나고 시스템의 성능을 평가할 때 사용할 테스트 데이터셋의 3가지로 나누어서 사용한다. 각 개인별로 학습 데이터셋은 1주차의 본인 음성 5문장과 타인 음성 27문장을 사용하였고, 튜닝 데이터셋은 2주차의 본인 음성 5문장과 타인 음성 27문장을 사용하였다. GA를 통해 선택된 최적의 특징들에 대한 SVM을 구축한 후에 이들의 성능을 평가하는데 사용한 데이터는 3주차 및 4주차의 본인 음성 10문장과 타인 음성 270문장을 사용하였다.

음성 특징벡터로는 24차 Mel-Frequency Cepstral Coefficient(MFCC)와 그 delta 파라메타로 구성된 48차원의 벡터와 36차 MFCC와 그 delta 파라메타로 구성된 72차원 벡터를 사용하였으며, 20ms 윈도우를 10ms씩 이동하면서 추출하였다. 모든 음성데이터는 16kHz로 샘플링되었으며 16bit로 양자화하였다. 실험을 위한 특징벡터로는 모든 프레임의 시간평균을 사용하였다.

B. 실험 결과

표 1에서 두가지 특징벡터에 대해서 RBF 커널의 파라미터(σ)에 따른 SVM의 인식성능을 나타내었다. 36차 MFCC를 사용할 때가 24차 MFCC를 사용했을 때보다 에러율은 약 0.6% 감소하였지만 메모리 사용량이 늘어나므로 메모리 사용량이 제한된 시스템에 적용하기에는 부적절하다고 할 수 있다. 또한 RBF 파라메타 값이 커질수록 FAR은 증가하고 FRR은 감소하며, 두 특징벡터 모두 파라메타 값이 1.6일 때 FAR과 FRR이 비슷한 값을 나타냄을 볼 수 있다.

표 1. RBF 커널의 파라메타에 따른 SVM의 성능 비교

파라메타	24MFCC+delta		36MFCC+delta	
	FAR (%)	FRR (%)	FAR (%)	FRR (%)
1.0	3.3	7.3	2.9	6.3
1.2	3.6	6.5	3.2	5.7
1.4	4.1	5.9	3.5	5.0
1.6	**4.6**	**5.1**	**4.0**	**4.4**
1.8	5.3	4.4	4.7	3.9
2.0	6.1	3.9	5.4	3.4

표 2는 GA과정을 통해 얻은 개인별로 인식을 잘 할 수 있는 특징들의 데이터에 대하여 SVM을 구축한 후에 테스트 데이터로 각 성능을 평가한 표이다. RBF 파라메타 값은 SVM 실험에서 FAR과 FRR이 가장 비슷한 1.6을 사용하였다. GA를 결합한 SVM과 SVM만을 사용한 경우를 비교해 보면, 36차 MFCC를 사용한 경우는 FAR과 FRR이 각각 2.0%, 1.6% 높아졌으나, 특징벡터의 차원은 절반 이상 감소하였다. 또한 24차 MFCC의 경우에는 FAR과 FRR이 모두 1.7% 높아진데 반해 특징벡터의 차원은 절반 이상 감소하였다. 이러한 결과를 이용하여 메모리 제한이 있는 환경에서는 약간의 성능의 손실이 있더라도 특징의 개수를 절반 이하로 줄일 수 있는 본 논문에서 제안한 GA와 SVM

을 결합한 방법을 적용할 수 있다.

표 2. GA를 결합한 SVM의 화자인식 성능

	SVM			GA+SVM		
	특징벡터차원	FAR (%)	FRR (%)	특징벡터차원	FAR (%)	FRR (%)
24MFCC +delta	48	4.6	5.1	22.1	6.3	6.9
36MFCC +delta	72	4.0	4.4	33.6	6.0	6.0

IV. 결론 및 향후 연구

본 논문에서는 GA와 SVM을 결합한 화자인증 시스템을 제안하였다. 인식실험을 위해서 한국어 4연 숫자 데이터베이스를 사용하였으며 특징벡터로는 24차 및 36차 MFCC와 그 delta 파라메타를 사용하였다. RBF 커널 파라메타가 1.6일 때 FAR과 FRR이 비슷한 값을 나타냄을 볼 수 있었고, 36차 MFCC를 사용했을 때가 24차 MFCC를 사용했을 때보다 약 0.6%의 에러율 감소를 나타내었다. 또한 GA를 통하여 각 사람에 대하여 식별력이 우수한 특징을 추출하여 SVM의 입력벡터로 사용하여 인증실험을 한 결과 SVM 단독으로 사용하였을 때에 비해서 약 1.7%의 에러율 증가가 있었으나 입력벡터의 차원은 절반 이상 감소하였다.

본 논문에서는 GA를 이용하여 개인별로 인증에 우수한 특징 집합을 선택함으로써 불필요한 화자 정보의 사용을 배제하여 화자 인증을 위한 메모리 공간의 현저한 감소를 이룰 수 있었다. 또한 GA의 평가함수의 값 계산에 튜닝 데이터셋을 이용함으로써 잡음 및 시간에 따라 변하는 음성의 변화에 덜 민감한 특징들의 선택을 가능하게 하였다. 그러므로 제안된 방법은 스마트 카드 시스템과 같은 메모리 제한적인 환경에서도 유용하게 사용될 수 있다.

참고문헌

[1] Joseph P. Campbell, "Speaker recognition: a tutorial", Proc. of the IEEE, Vol. 85, No. 9, pp. 1437-1462, Sep. 1997

[2] Qi Li, Biing-Hwang Juang, Chin-Hui Lee, Qiru Zhou, and Frank K. Soong, "Recent advancements in automatic speaker authentication", IEEE Robotics and Automation Magazine, pp. 24-34, Mar. 1999

[3] Jialong He, Li Liu, and Gunther Palm, "A New Codebook Training Algorithm for VQ-based Speaker Recognition", Proc. ICASSP, Vol. 2, pp. 1091-1094, 1997

[4] C. Martin del Alamo, F. J. Caminero Gil, C. de la Torre Munilla, and L. Hernandez Gomez, "Discriminative training of GMM for speaker identification", Proc. ICASSP, Vol. 1, pp. 89-92, 1996

[5] Simon Haykin, "Neural Networks", Prentice Hall, 1999

[6] H. Vafaie and K. DeJong, "Feature Space Transformation Using Genetic Algorithms", IEEE Intelligent Systems, pp.57-65, March/April 1998

[7] W. Choi, et. al., "Support Vector Machines for Robust Speaker Verification", Proc. of the AICSST, pp. 262-267, 2002

[8] Bernhard Scholkopf, Christopher J. C. Burges, and Alexander J. Smola, "Advances in Kernel Methods", The MIT Press, 1999

[9] http://voice.etri.re.kr