

Support vector machines for emotion recognition in Chinese speech

Wang Zhiping Zhao Li Zou Cairong

(Department of Radio Engineering, Southeast University, Nanjing 210096, China)

Abstract: Support vector machines (SVMs) are utilized for emotion recognition in Chinese speech in this paper. Both binary-class discrimination and the multi-class discrimination are discussed. It proves that the emotional features construct a nonlinear problem in the input space, and SVMs based on nonlinear mapping can solve it more effectively than other linear methods. Multi-class classification based on SVMs with a soft decision function is constructed to classify the four emotion situations. Compared with principal component analysis (PCA) method and modified PCA method, SVMs perform the best result in multi-class discrimination by using nonlinear kernel mapping.

Key words: speech signal; emotion recognition; support vector machines

Utilizing a new learning method — support vector machines (SVMs), the paper studies the binary classification and multi classification of the emotion in Chinese speech. Being different from other learning machines, SVMs use a structural risk minimization (SRM) principle, while others use an empirical risk minimization principle, thus it has a better generalization performance^[1]. It uses a kernel function for efficiently performing computations in high dimensional spaces and constructs nonlinear decision functions to perform an optimal separating hyperplane in feature space.

In this paper, we introduce SVMs first. Then the parameters are extracted from the materials. The numbers and sorts of classifications of emotions are not the same in different literatures. Happiness, anger, surprise and sadness are used in this paper. Utilizing the SVMs, classifications have been done and conclusions have been drawn to indicate the best performance of the SVMs.

1 Support Vector Machines^[1,2]

Support vector machines are based on the structural risk minimization principle and Vapnik-Chervonenkis (VC) dimension from statistical learning theory developed by Vapnik, et al^[1]. Traditional techniques for pattern recognition are based on the minimization of empirical risk, that is, on the attempt to optimize performance on the training set, SVMs minimize the structural risk to reach a better

performance.

We can suppose that S is a set that is made up of points $\mathbf{x}_i (i = 1, 2, \dots, N)$, which belong to \mathbf{R}^n . These points are divided into two classes, which are separated by an objective function y_i ,

$$y_i = \begin{cases} 1 & \mathbf{x}_i \in S_1 \\ -1 & \mathbf{x}_i \in S_2 \end{cases}$$

where S_1 and S_2 belong to different classes. We want to find a hyperplane to separate two classes, and sort the same class in the same side of the hyperplane as much as possible, and make the margin as far as possible. If S can be separated linearly, there may be $\mathbf{w} \in \mathbf{R}^n, b \in \mathbf{R}$ to satisfy

$$\left. \begin{aligned} \mathbf{w} \cdot \mathbf{x}_i + b &\geq 1 & y_i &= 1 \\ \mathbf{w} \cdot \mathbf{x}_i + b &\leq -1 & y_i &= -1 \end{aligned} \right\} \quad (1)$$

Formula (1) also can be represented by

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad (2)$$

Parameters (\mathbf{w}, b) have determined a hyperplane,

$$\mathbf{w} \cdot \mathbf{x}_i + b = 0 \quad (3)$$

This plane is called the separating hyperplane. The problem of finding the optimal separating hyperplane is converted to an optimal problem as follows.

$$\min \frac{1}{2} \|\mathbf{w}\|^2 \quad (4)$$

$$\text{s.t. } y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad i = 1, 2, \dots, N$$

It is then converted to a dual problem by using Lagrange multiplies,

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

$$\text{s.t. } \sum_{i=1}^N y_i \alpha_i = 0 \quad \alpha \geq 0$$

When S cannot be separated linearly, a nonnegative relax factor $\xi = (\xi_1, \dots, \xi_N)$ is

Received 2003-06-13.

Foundation item: Education Revitalization Program Oriented to the 21st Century under the Chinese Ministry of Education.

Biographies: Wang Zhiping (1977—), male, graduate; Zou Cairong (corresponding author), male, doctor, professor, cairong@seu.edu.cn.

introduced. There is

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i \quad (6)$$

The optimal problem can be described as

$$\max \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \quad (7)$$

$$\text{s.t. } \sum_{i=1}^N y_i \alpha_i = 0 \quad i = 1, 2, \dots, N; 0 \leq \alpha_i \leq C$$

Formula (7) is a general form of the SVM. When C tends to infinite, formula (7) degenerates into a linear separating problem as formula (5). Replacing $y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$ by D_{ij} , the optimal object turns to be the maximum $\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i D_{ij} \alpha_j$. Obviously, this is a quadratic program. We can solve it by using the sequential minimal optimization (SMO) proposed by Platt^[3]. When parameters α_i^* and b^* are obtained, the different classes can be distinguished by objective function

$$y = \text{sgn}(\mathbf{w}^* \cdot \mathbf{x} + b^*) = \text{sgn}\left(\sum_{i=1}^N \alpha_i^* y_i \mathbf{x}_i \cdot \mathbf{x} + b^*\right) \quad (8)$$

In most cases, discrimination is not linear in input space. A higher order function is introduced for mapping a nonlinearly separating problem to a linearly separating problem. Because the optimal problem mentioned above deals with inner product only, a kernel function $K(\mathbf{x}, \mathbf{y})$ can be constructed to substitute the inner product.

$$K(\mathbf{x}_i, \mathbf{y}_i) = \varphi(\mathbf{x}_i) \varphi(\mathbf{y}_i) \quad (9)$$

where $\varphi(\mathbf{x})$ is a nonlinear map from \mathbf{R}^n to feature space. The inner product is converted to a function operation in input space. A kernel function exists when the Mercer condition is satisfied^[1].

When SVM is used for classification, it works like a neural network, which classifies the different classes by inner product between the input vectors and support vectors. Inner product is substituted by kernel function operation. The principle of SVM is shown in Fig.1.

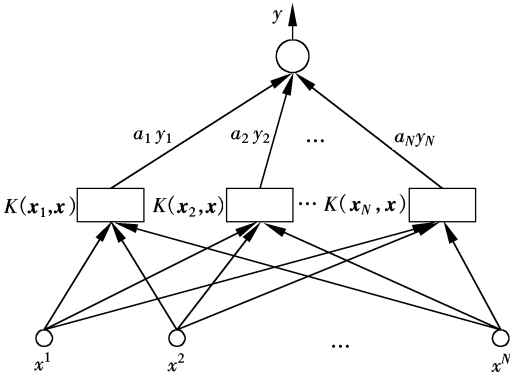


Fig.1 Principle of support vector machines

2 Selection of Samples and Extraction of Feature Parameter

It is important to select the samples of the speech for emotion analysis. Because there is no standard of emotion analysis proposed, two aspects were taken into account in selecting the speech samples in this paper. Firstly, the sentences should not have any emotional tendency. Secondly, several factors should be considered such as the length of the sentence, consonant and auxiliary word structure, and sexual difference. According to these rules, we used four sentences as the speech samples. To obtain the original data of the speech, 15 male speakers who are good at acting spoke the sentences respectively with happiness, anger, surprise and sadness. We also made the speakers speak each sentence in a normal way. Accordingly, we got 900 sentences for experiments. 240 of them were used for training and 480 were used for recognition. The remaining 180 sentences without emotion were used for comparison.

We played all these emotional sentences randomly. Listeners different from speakers judged the type of emotion involved in each sentence. Those samples, which were ambiguous in emotion, were discarded and collected again until they satisfied the conditions.

3 Emotion Recognition in Speech

3.1 Binary classification

SVM is a binary classification essentially, so we study the performance of binary classification of single emotion at first. 720 emotional materials are used for recognition. Each type has 180 materials. 240 are used as training set, and the remaining 480 materials are used for recognition set. One of the emotions is selected as the recognition subject. For the i -th sample, ten features are extracted and compose a feature vector $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^{10})$. If the i -th sample belongs to the given emotion, the objective value is set to 1, or else the value is set to -1 . Training set is composed of the feature vectors extracted from 240 (60×4) speech samples.

Radial basis function machines can be implemented by using a kernel function of the type

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

We get the support vectors \mathbf{x}_i , and corresponding coefficient a_i and bias b when the training is finished. The indicated value of the test material can be obtained

by calculating the indicate a function, if $y_j = 1$, then it belongs to a certain emotion, or else it doesn't. Fig. 2 shows the result classified by the method mentioned above. The emotion considered is sadness. Just for an obvious observation, the input vector, which is ten dimensions, is projected to \mathbf{R}^2 .

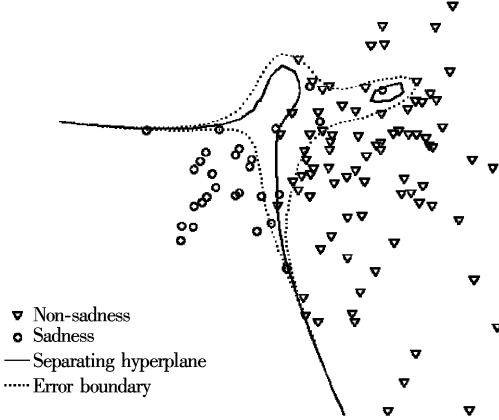


Fig.2 Two-dimensional project of binary classification by SVM when sadness is considered

In Fig.2, a radial basis function machine is used for classification of the sadness emotion. Although it seems that overfit exists in two-dimensional project, it occurred because of the projection from \mathbf{R}^{10} to \mathbf{R}^2 . Experiments have been done with other emotions too, and results are shown in Tab.1. A linear kernel is selected for comparison.

Tab.1 Comparison of different kernels %

Rate	Linear kernel	Gauss (RBF) kernel
Happiness	92.86	87.85
Anger	90.18	91.96
Sadness	92.86	93.75
Surprise	79.46	91.96
Mean rate	88.84	91.38

With the linear kernel method, there is no transform on original space and classification is done in input space directly. With the Gauss kernel method, the input space is mapped to feature space and classification is done in feature space. Comparing with these two methods, better results are obtained by the Gauss kernel method. As recognized before, classification is always done in input space and the problem is conceived as a linear problem. This experiment proves that emotion features are not a linear classification problem; better results can be obtained by nonlinear kernel transform.

3.2 Multi classification

Discriminating for emotional speech is a fuzzy classification. Humans always distinguish emotions by

the levels reflected from different emotions. When one of the emotional levels is stronger than others, it is the main emotion represented in speech. It is a binary classification for single SVM, so investigation is going on for multi-class classification. Now there are some multi-class classification schemes. Some schemes use binary mode combination, others use all mode combination schemes. In binary mode combination, several schemes were proposed such as one-against-all, one-against-one, and DAGSVM. All mode combinations are mainly proposed by Refs.[4,5]. In Ref.[6], all these methods mentioned above are compared and the conclusion is drawn that the precision of all these methods are approximate and that training-time costs by one-against-one and DAGSVM are obviously less than the others. So one-against-one and DAGSVM are fit for implementation.

In our experiment, as precision is the main consideration, and time cost can be ignored, so the one-against-all scheme was selected for recognizing four emotions. N SVMs are constructed for an N class classification problem. The i -th SVM is especially designed to recognize the i -th emotion. The indicated value is set to 1 when the material belongs to the i -th emotion, or it is set to -1 . N SVMs for binary classification can be obtained by training it. A sign function isn't selected as a decision function due to its hard characteristic; a soft decision function is selected. The soft decision function we used in our experiment is illustrated as follows:

$$y_{jk} = \begin{cases} 1 & H \geq 1 \\ \sum_{x \in sv} \alpha_{ik} y_{ik} (K(\mathbf{x}_{ik}, \mathbf{x})) + b & -1 \leq H < 1 \\ -1 & H < -1 \end{cases} \quad (10)$$

where $H = \sum_{x \in sv} \alpha_{ik} y_{ik} (K(\mathbf{x}_{ik}, \mathbf{x})) + b$; j is the label of the test speech sample; k is the label of different emotions. After the feature vector of the material we want to recognize passes through all these N SVMs, the largest one is picked out and the label of the SVM is the emotion which the test material belongs to,

$$y_j = \max(y_{jk}) \quad (11)$$

The method mentioned above is utilized for classifying a test set composed of 160 (40×4) speech samples and results of recognition are shown in Tab.2.

We have classified the same data with principal component analysis (PCA) method and a modified PCA method mentioned in Ref. [7]; results are shown in Tab.3 and Tab.4.

Tab.2 SVMs method

Sample	Result			
	Anger	Happiness	Sadness	Surprise
Anger	97	5	7	11
Happiness	11	85	5	19
Sadness	1	2	116	1
Surprise	26	13	1	80

Tab.3 PCA method

Sample	Result			
	Anger	Happiness	Sadness	Surprise
Anger	73	14	10	23
Happiness	8	49	1	62
Sadness	13	2	71	34
Surprise	14	13	17	76

Tab.4 Modified PCA method

Sample	Result			
	Anger	Happiness	Sadness	Surprise
Anger	91	11	1	17
Happiness	23	61	1	35
Sadness	10	5	95	10
Surprise	23	13	11	73

As we can see, the error rate of the SVMs is lower than some former methods. Combined with the results of the binary classification, the conclusion can be drawn that mapping to feature space can solve the emotion classification more effectively than before.

4 Conclusion

Support vector machines, which are based on statistical learning theory, are used for recognizing the emotion in speech in this paper. Binary classification and multi classification are considered.

A linear kernel method and a nonlinear kernel method are utilized for binary classification. This proves emotion features are not a linear classification problem. Better results can be obtained by nonlinear

kernel transform. In multi-class classification, a one-against-all scheme is utilized. At the same time, PCA method and modified PCA method are used to do classification. Results show that the error rate is reduced when SVMs are utilized, its performance is better than other methods mentioned in former literature. All the results of the experiments have proved that emotion features are a nonlinear classification, better results can be obtained by nonlinear kernel transform, and SVM method is a better method to be utilized.

References

[1] Vapnik V N. *Statistical learning theory* [M]. New York: Wiley, 1998.

[2] Christopher, Burges C J C. A tutorial on support vector machines for pattern recognition [J]. *Data Mining and Knowledge*, 1998, 2(2): 121 - 167.

[3] Platt J. Sequential minimal optimization: a fast algorithm for training support vector machines [R]. Microsoft Research Technical Report MSR-TR-98-14, 1998.

[4] Crammer K, Singer Y. On the learn ability and design of output codes for multiclass problems [A]. In: *Proceeding of the Thirteenth Annual Conference on Computational Learning Theory* [C]. 2000.35 - 46.

[5] Weston J, Watkins C. Multi-class support vector machines [A]. In: Verleysen M, ed. *Proc ESANN99* [C]. Brussels, Belgium, 1999.

[6] Hsu Chih-Wei, Lin Chih-Jen. A comparison of methods for multiclass support vector machines [J]. *IEEE Trans Neural Networks*, 2002, 13(2): 415 - 425.

[7] Zhao Li, Qian Xiangmin, Zou Cairong, et al. A study on emotional feature analysis and recognition in speech signal [J]. *Journal of China Institute of Communication*, 2000, 21 (10): 18 - 24. (in Chinese)

基于支持向量机的语音情感识别

王治平 赵 力 邹采荣

(东南大学无线电工程系, 南京 210096)

摘 要 针对语音情感识别特征识别问题,本文利用支持向量机进行了研究.分析表明语音信号的情感特征参数在输入空间中不完全是一个线性分类的问题,使用非线性的核函数对输入空间进行映射可以有效地提高识别效率.与已有的多模式语音情感识别方式相比,利用高斯(径向基)核函数的支持向量机的识别效果优于其他已有的方法.

关键词 语音信号;情感识别;支持向量机

中图分类号 TN912.34