Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition 40 (2007) 346–349

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Rapid and brief communication

Center-based nearest neighbor classifier

Qing-Bin Gao*, Zheng-Zhi Wang

Institute of Automation, National University of Defense Technology, Changsha 410073, Hunan, People's Republic of China

Received 5 November 2005; accepted 14 June 2006

Abstract

In this paper, a novel center-based nearest neighbor (CNN) classifier is proposed to deal with the pattern classification problems. Unlike nearest feature line (NFL) method, CNN considers the line passing through a sample point with known label and the center of the sample class. This line is called the center-based line (CL). These lines seem to have more capacity of representation for sample classes than the original samples and thus can capture more information. Similar to NFL, CNN is based on the nearest distance from an unknown sample point to a certain CL for classification. As a result, the computation time of CNN can be shortened dramatically with less accuracy decrease when compared with NFL. The performance of CNN is demonstrated in one simulation experiment from computational biology and high classification accuracy has been achieved in the leave-one-out test. The comparisons with nearest neighbor (NN) classifier and NFL classifier indicate that this novel classifier achieves competitive performance.

Keywords: Pattern classification; Nearest neighbor; Nearest feature line; Centered-based nearest neighbor; Computational biology

1. Introduction

Pattern classification is a fundamental problem in artificial intelligence and other fields. The problem can be described generally as follows: given N training samples with known class labels, which can be divided into C classes, N_c is the size of class c, $c \in \{1, 2, \ldots, C\}$, how to predict the class label of an unknown sample $\mathbf{x}$? Many methods have been suggested to tackle this problem. Nearest neighbor (NN) classifier is such a simple, yet effective method for performing general, non-parametric pattern classification [1]. The empirical evaluation to data in various fields shows that NN is robust and has asymptotic error rate that is at most twice the Bayes error rate [1]. In literature [2], a nearest feature line (NFL) method is suggested to overcome some limitations of NN as stated by the authors. The NFL classifier uses a linear model to interpolate and extrapolate each pair of sample points within the same class and attempts to generalize the representational capacity of available samples by using the feature line passing through two sample points in the same class. The experimental results show that NFL improves the classification accuracy consistently. In literature [3], a tunable nearest neighbor (TNN) method is proposed to improve the performance of NFL. However, there are still some drawbacks both in NFL and TNN methods that will limit their further application in practice as pointed out in Ref. [4]. One of them is the large computation complexity problem, especially for the task with large training sample set. In literature [4], a nearest neighbor line (NNL) method is introduced to lower the computation cost of NFL method. Face recognition experiments show that the NNL method takes much lower time and achieves competitive performance.

In this paper, a novel center-based nearest neighbor (CNN) method is proposed to lessen the computation burden of the original NFL method by defining another kind of line called center-based line (CL). This CL connects a sample point with known label and the center of the sample class, instead of two labeled sample points in the NFL method. One experiment from computational biology is performed to demonstrate its efficiency. The comparisons with NN classifier and NFL classifier show that this novel classifier is competitive

* Tel.: +86 731 4574 991.
E-mail address: qbgao@nudt.edu.cn (Q.-B. Gao).

0031-3203/$30.00 © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2006.06.033

and can be used as an alternative method for the pattern classification problems.

2. Center-based nearest neighbour (CNN)

Unlike NFL, CNN considers another kind of line for classification. The main idea of CNN is described as follows. Let $\mathbf{x}_i^c$ be a training sample of class c, let $\mathbf{o}^c$ be the center of class c, which can be calculated by

$$\mathbf{o}^c = \frac{\sum_{i=1}^{N_c} \mathbf{x}_i^c}{N_c}. \tag{1}$$

For an unknown sample $\mathbf{x}$, define CL $\overline{\mathbf{x}_i^c \mathbf{o}^c}$, which is the straight line passing through $\mathbf{x}_i^c$ and $\mathbf{o}^c$, as illustrated in Fig. 1. This CL is used to capture the information implied by the interaction between points of $\mathbf{x}_i^c$ and $\mathbf{o}^c$ to achieve better classification performance. Define distance from $\mathbf{x}$ to CL as

$$d(\mathbf{x}, \overline{\mathbf{x}_i^c \mathbf{o}^c}) = \|\mathbf{x} - \mathbf{p}^{c,i}\| \tag{2}$$

which is used as a basic measure for classification, also shown in Fig. 1, where $\mathbf{p}^{c,i}$ is the projection of $\mathbf{x}$ onto the CL, $\|\bullet\|$ is the Euclidean norm. The projection point $\mathbf{p}^{c,i}$ is calculated according to the equation

$$\mathbf{p}^{c,i} = \mathbf{x}_i^c + \mu(\mathbf{o}^c - \mathbf{x}_i^c), \tag{3}$$

where $\mu \in R$, which is called position parameter and formalized as

$$\mu = \frac{(\mathbf{x} - \mathbf{x}_i^c)^{\mathrm{T}}(\mathbf{o}^c - \mathbf{x}_i^c)}{(\mathbf{o}^c - \mathbf{x}_i^c)^{\mathrm{T}}(\mathbf{o}^c - \mathbf{x}_i^c)}, \tag{4}$$

where T is the transpose operator. Then the CNN distance is defined as

$$d_{CNN}^c = \min_{1 \leqslant i \leqslant N_c} d(\mathbf{x}, \overline{\mathbf{x}_i^c \mathbf{o}^c}). \tag{5}$$

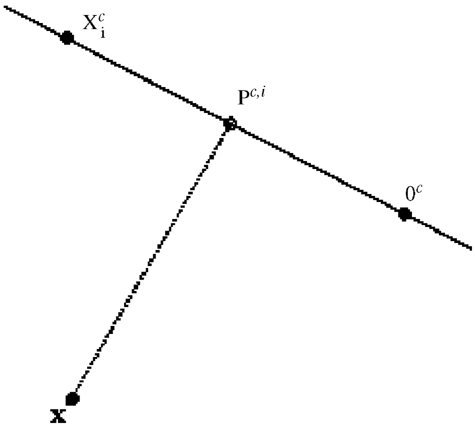

Fig. 1. FL $\mathbf{x}_i^c \mathbf{o}^c$ and CNN distance $d(\mathbf{x}, \mathbf{x}_i^c \mathbf{o}^c)$.

According to the NN rule, the CNN discriminating rule is

IF $d_{CNN}^o = \min_{c=1,2,\ldots,C} d_{CNN}^c$, THEN $\mathbf{x}$ is classified into class o, $o \in \{1, 2, \ldots, C\}$.

3. Experiments and results

To evaluate the performance of CNN classifier, in this paper, one experiment from computational biology is performed. As a comparison, the same experiments based on NN classifier and NFL classifier are also conducted. Protein subcellular location prediction is now a hot topic in computational biology community, which is intrinsically a protein classification problem.

3.1. Data set and feature vector

The benchmark data set constructed by Reinhardt and Hubbard [5] for protein subcellular localization is used in the experiments. All the protein entries in the data set are extracted from Swiss-Prot 33.0. The data set contains 3424 non-redundant proteins with less than 90% sequence identity whose subcellular locations are experimentally determined, in which 2427 are eukaryotic and 997 are prokaryotic proteins. The eukaryotic proteins consist of 684 cytoplasmic, 325 extracellular, 321 mitochondrial and 1097 nuclear proteins, while the prokaryotic proteins consists of 688 cytoplasmic, 107 extracellular and 202 periplasmic proteins.

One typical feature used to predict protein subcellular location is amino acid composition [5,6]. The protein sequence is represented by a 20-dimensional feature vector $\mathbf{x} \in R^{20}$, because there are 20 amino acids in biological proteins. Each element in the feature vector denotes the occurrence frequency of an amino acid.

3.2. Performance measurement

In our work, a leave-one-out test is performed to evaluate the performance of classifiers. Each protein sequence in the data set is singled out in turn as a test sample and the remaining protein sequences are used as the training data set to predict its subcellular location. Compared with other cross-validation methods, the leave-one-out test is considered to be the most effective way and more rigorous and reliable [7]. Total accuracy, subset accuracy and the Matthew's correlation coefficient (MCC) are used to measure the classification performance of our work. MCC provides a single measure of evaluating sensitivity and specificity together [8]. They are defined by

$$\text{Total accuracy }(s) = \frac{p(s)}{\text{sub}(s)}, \tag{6}$$

$$\text{Subset accuracy} = \frac{\sum_{s=1}^{k} p(s)}{N}, \tag{7}$$

Table 1
The comparisons with NN and NFL [9] on eukaryotic proteins

Data set	Location	NN		NFL		CNN	
		Acc (%)	MCC	Acc (%)	MCC	Acc (%)	MCC
Eukaryotic	Cytoplasmic	81.4	0.66	80.1	0.69	82.7	0.70
	Extracellular	85.8	0.87	82.2	0.83	86.2	0.81
	Mitochondrial	60.7	0.59	54.8	0.60	64.5	0.64
	Nuclear	86.8	0.76	92.0	0.77	86.6	0.78
	Total accuracy	81.5	–	82.5	–	82.5	–

Table 2
The comparisons with NN and NFL [9] on prokaryotic proteins

Data set	Location	NN		NFL		CNN	
		Acc (%)	MCC	Acc (%)	MCC	Acc (%)	MCC
Prokaryotic	Cytoplasmic	96.2	0.78	98.5	0.82	96.8	0.83
	Extracellular	80.4	0.81	76.6	0.83	82.2	0.80
	Periplasmic	68.8	0.68	72.3	0.74	74.8	0.75
	Total accuracy	89.0	–	91.0	–	90.8	–

$$\mathrm{MCC}(s) = \frac{p(s)n(s)-u(s)o(s)}{\sqrt{(p(s)+u(s))(p(s)+o(s))(n(s)+u(s))(n(s)+o(s))}}, \quad (8)$$

where N is the total number of proteins in the data set, k is the number of subcellular locations, sub(s) is the number of proteins reside in location s, $p(s)$ is the number of properly predicted proteins in location s, $n(s)$ is the number of correctly predicted proteins not in location s, $u(s)$ is the number of under-predicted and $o(s)$ is the number of over-predicted proteins.

3.3. Results and computational complexity

The experimental results of NN, NFL and CNN methods on eukaryotic and prokaryotic proteins are shown in Tables 1 and 2, respectively. From Tables 1 and 2 we can see that CNN achieves higher total accuracies than NN method. When compared with NFL method, they have nearly the same classification accuracy, but the classification time of CNN is shortened dramatically in our experiment. The CNN classifier only need to construct N CLs to represent a C classes data set, however, the NFL classifier has to construct M FLs for the same purpose, where

$$N = \sum_{c=1}^{C} N_c. \quad (9)$$

$$M = \sum_{c=1}^{C} N_c(N_c - 1)/2. \quad (10)$$

It is clear that the computational time of CNN is reduced dramatically in comparison with that of NFL. Therefore, in the classification task of large training sample set, the CNN method is probably a better choice.

4. Conclusion

In this paper, a novel classifier termed as center-based nearest neighbor (CNN) is proposed to deal with the pattern classification problems. CNN considers the center-based line (CL) passing through a sample point with known label and the center of the sample. These lines seem to have more capacity of representation for example classes than the original samples and thus can capture more information. The experimental results of protein subcellular localization from computational biology demonstrates its efficiency. Moreover, the comparisons with nearest neighbor (NN) and nearest feature line (NFL) classifiers indicate that this novel classifier has a superior performance, especially for the task with large training sample set.

Acknowledgments

The authors would like to thank Dr A. Reinhardt for providing the dataset. This work was partly supported by the National Natural Science Foundation of China (no. 60471003).

References

[1] T.M. Cover, P.E. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1967) 21–27.

[2] S.Z. Li, J. Lu, Face recognition using the nearest feature line method, IEEE Trans. Neural Networks 10 (1999) 439–443.

[3] W. Zheng, L. Zhao, C. Zou, Locally nearest neighbor classifiers for pattern classification, Pattern Recognition 37 (2004) 1307–1309.

[4] Y. Zhou, C. Zhang, J. Wang, Tunable nearest neighbor classifier, Lect. Notes Comput. Sci. 3175 (2004) 204–211.

[5] A. Reinhardt, T. Hubbard, Using neural networks for prediction of the subcellular location of proteins, Nucleic Acids Res. 26 (1998) 2230–2236.

[6] S. Hua, Z. Sun, Support vector machine approach for protein subcellular location prediction, Bioinformatics 17 (2001) 721–728.

[7] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, London, 1979, pp. 322–381.

[8] B.W. Matthews, Comparison of predicted and observed secondary structure of T4 phage lysozyme, Biochim. Biophys. Acta 405 (1975) 442–451.

[9] Q.-B. Gao, Z.-Z. Wang, Using nearest feature line and tunable nearest neighbor methods for prediction of protein subcellular locations, Comput. Biol. Chem. 29 (2005) 388–392.