

Improving nearest neighbor rule with a simple adaptive distance measure

Jigang Wang^{*}, Predrag Neskovic, Leon N. Cooper

Department of Physics, The Institute for Brain and Neural Systems, Brown University, P.O. Box 1843, Providence, RI 02912, USA

Received 8 March 2006

Available online 24 August 2006

Communicated by R.P.W. Duin

Abstract

The k -nearest neighbor rule is one of the simplest and most attractive pattern classification algorithms. However, it faces serious challenges when patterns of different classes overlap in some regions in the feature space. In the past, many researchers developed various adaptive or discriminant metrics to improve its performance. In this paper, we demonstrate that an extremely simple adaptive distance measure significantly improves the performance of the k -nearest neighbor rule.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Pattern classification; Nearest neighbor rule; Adaptive distance measure; Adaptive metric; Generalization error

1. Introduction

The nearest neighbor (NN) rule, first proposed by Fix and Hodges (1951), is one of the oldest and simplest pattern classification algorithms. Given a set of n labeled examples $D_n = \{(\vec{X}_1, Y_1), \dots, (\vec{X}_n, Y_n)\}$ with input vectors $\vec{X}_i \in \mathbb{R}^d$ and class labels $Y_i \in \{\omega_1, \dots, \omega_M\}$, the NN rule classifies an unseen pattern \vec{X} to the class of its nearest neighbor in the training data D_n . To identify the nearest neighbor of a query pattern, a distance function has to be defined to measure the similarity between two patterns. In the absence of prior knowledge, the Euclidean and Manhattan distance functions have conventionally been used as similarity measures for computational convenience.

The basic rationale for the NN rule is both simple and intuitive: patterns close in the input space \mathbb{R}^d are likely to belong to the same class. This intuition can be justified more rigorously in a probabilistic framework in the large sample limit. Indeed, as one can easily show, as the number

of training examples $n \rightarrow \infty$, the nearest neighbor of a query pattern converges to the query pattern with probability one, independently of the metric used. Therefore, the nearest neighbor and the query pattern have the same *a posteriori* probability distribution asymptotically, which leads to the asymptotic optimality of the NN rule:

$$L^* \leq L_{\text{NN}} \leq L^* \left(2 - \frac{M}{M-1} L^* \right), \quad (1)$$

where L^* is the optimal Bayes probability of error, see Cover and Hart (1967). According to (1), the NN rule is asymptotically optimal when $L^* = 0$, i.e., when different pattern classes do not overlap in the input space. When the classes do overlap, the sub-optimality of the NN rule can be overcome by the k -nearest neighbor (k -NN) rule that classifies \vec{X} to the class that appears most frequently among its k nearest neighbors (Stone, 1977).

It should be noted that the above results are established in the asymptotic limit and essentially rely on averaging over an infinite amount of training examples within an infinitesimal neighborhood to achieve optimality. In reality, one most often only has access to a finite number of training examples, and the performance of the k -NN rule depends

^{*} Corresponding author. Tel.: +1 401 863 3920; fax: +1 401 863 3494.
E-mail addresses: jigang@physics.brown.edu (J. Wang), pedja@brown.edu (P. Neskovic), Leon_Cooper@brown.edu (L.N. Cooper).

crucially on how to choose a suitable metric so that according to the chosen metric the majority of the k nearest neighbors to a query pattern is from the desired class. In the past, many methods have been developed to locally adapt the metric so that a neighborhood of approximately constant *a posteriori* probability can be produced. Examples of these methods include the flexible metric method by Friedman (1994), the discriminant adaptive method developed by Hastie and Tibshirani (1996), and the adaptive metric method by Domeniconi et al. (2002). The common idea underlying these methods is that they estimate feature relevance locally at each query pattern. The locally estimated feature relevance leads to a weighted metric for computing the distance between a query pattern and the training data. As a result, neighborhoods get constricted along the most relevant dimensions and elongated along the less important ones. Although these methods improve the original k -NN rule due to their capability to produce local neighborhoods in which the *a posteriori* probabilities are approximately constant, the computational complexity of such improvements is high. More recently, there has been considerable research interest in directly learning distance metrics from training examples to improve the k -NN rule. For example, Goldberger et al. (2004) proposed a method for learning a Mahalanobis distance measure by directly maximizing a stochastic variant of the leave-one-out k -NN score on the training data. Weinberger et al. (2005) developed a method for learning a Mahalanobis distance metric by semidefinite programming. Many other methods along this line can be found in the references therein.

In our previous work, we proposed a simple adaptive k -nearest neighbor classification algorithm based on the concept of statistical confidence we borrowed from hypothesis testing (Wang et al., 2005, 2006). The proposed adaptive k -NN algorithm involves both a locally adaptive distance measure for identifying the nearest neighbors to a query pattern and a weighting scheme that assigns a weight to each nearest neighbor based on its statistical confidence. We showed that the adaptive k -nearest neighbor algorithm not only outperforms the original k -NN rule with Euclidean distance measure but also achieves comparable or better performance than the Support Vector Machines (SVMs) on real-world datasets. However, due to the presence of both contributing factors, it is hard to know whether it is the locally adaptive distance measure or the weighting scheme that contributes most to the generalization performance improvements. In this paper, we show that, the extremely simple adaptive distance measure, which basically normalizes the ordinary Euclidean or Manhattan distance from a query pattern to each training example by the shortest distance between the corresponding training example to training examples of a different class, is the leading factor for the improvements over the original k -NN rule using the Euclidean or Manhattan distance measure, while the contribution of the weighting scheme is only marginal.

The remainder of the paper is organized as follows. In Section 2, we describe the locally adaptive distance measure

and the k -NN rule using the adaptive distance measure. In Section 3, we present experimental results of the resulting adaptive k -NN rule on several real-world datasets and compare it to the k -NN rule with the Euclidean and Manhattan distance measures and the adaptive k -nearest neighbor algorithm we proposed before. Concluding remarks are given in Section 4.

2. Adaptive nearest neighbor rule

We briefly describe the k -NN rule to introduce notation. Let us assume that patterns to be classified are represented as vectors in a d -dimensional Euclidean space \mathbb{R}^d . Given a set of training examples $\{(\vec{X}_1, Y_1), \dots, (\vec{X}_n, Y_n)\}$ and a query pattern \vec{X} , the k -NN rule first finds the k nearest neighbors of \vec{X} , denoted by $\vec{X}_{(1)}, \dots, \vec{X}_{(k)}$, and assigns \vec{X} to the majority class among $Y_{(1)}, \dots, Y_{(k)}$, where $Y_{(i)}$ are the corresponding class labels of $\vec{X}_{(i)}$. Without prior knowledge, the Euclidean distance (L2)

$$d(\vec{X}, \vec{X}_i) = \left(\sum_{j=1}^d |X^j - X_i^j|^2 \right)^{1/2} \quad (2)$$

and the Manhattan distance (L1)

$$d(\vec{X}, \vec{X}_i) = \sum_{j=1}^d |X^j - X_i^j| \quad (3)$$

have conventionally been used for measuring the similarity between \vec{X} and \vec{X}_i . For a binary classification problem in which $Y \in \{-1, 1\}$, the k -NN rule amounts to the following decision rule:

$$f(\vec{X}) = \text{sgn} \left(\sum_{i=1}^k Y_{(i)} \right) \quad (4)$$

To define the locally adaptive distance between a query pattern \vec{X} and a training example \vec{X}_i , we first construct the largest sphere centered on \vec{X}_i that excludes all training examples from other classes. This can be easily achieved by setting the radius of the sphere to

$$r_i = \min_{l: Y_l \neq Y_i} d(\vec{X}_i, \vec{X}_l) - \epsilon \quad (5)$$

where $\epsilon > 0$ is an arbitrarily small number. Notice that depending on the metric $d(\vec{X}_i, \vec{X}_l)$ that is actually used, the regions defined by points with distance to \vec{X}_i less than r_i may not be a sphere. However, for simplicity, we refer to such defined regions as spheres for convenience when

Table 1
Comparison of error rates

Dataset	NN	A-NN
Breast cancer	04.85 (0.91)	03.09 (0.71)
Ionosphere	12.86 (1.96)	06.86 (1.36)
Pima	31.84 (1.05)	28.16 (1.57)
Liver	37.65 (2.80)	32.94 (2.23)
Sonar	17.00 (2.26)	13.00 (1.70)

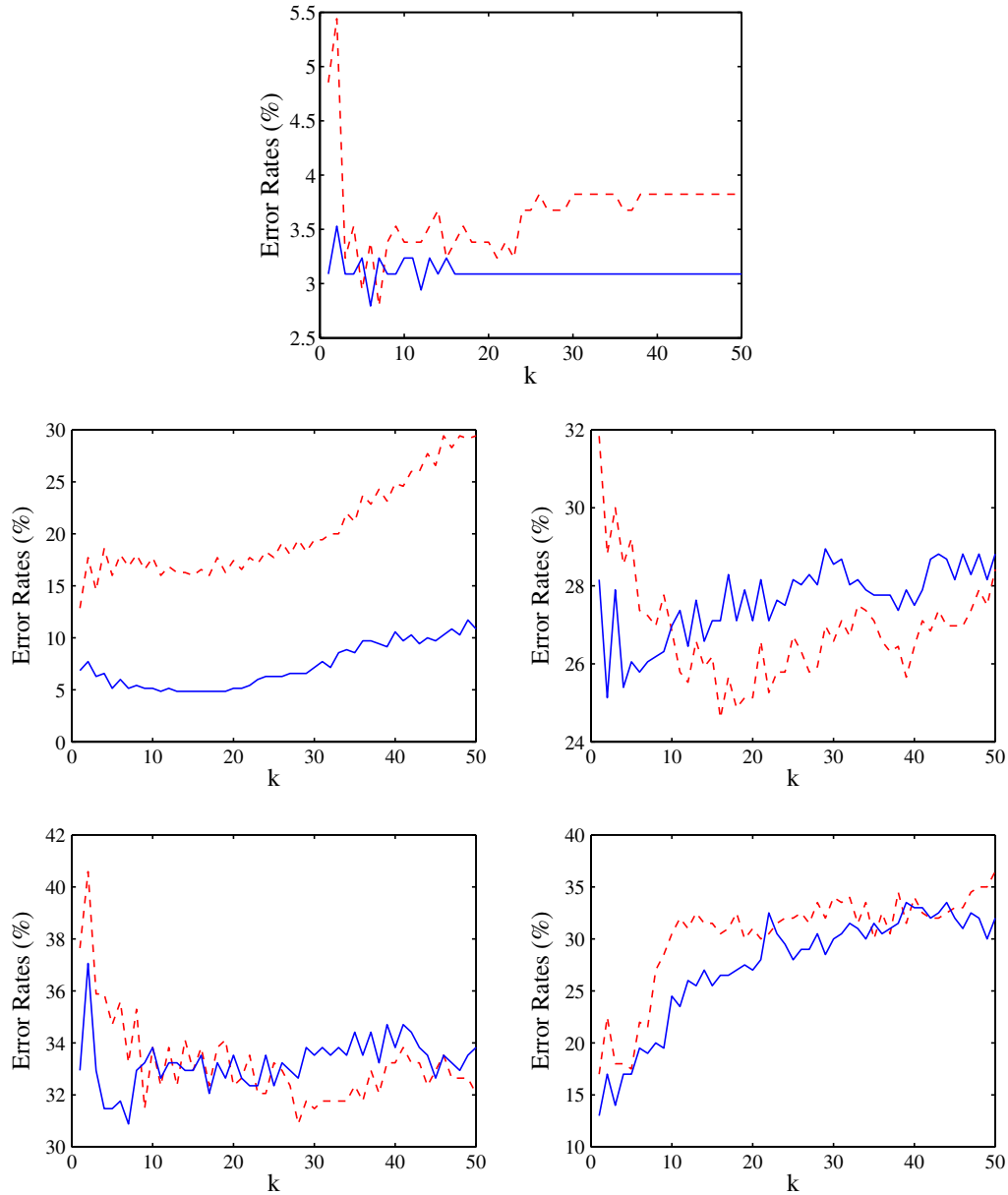


Fig. 1. Error rates at different values of k on the five datasets. The plots are for the Wisconsin Breast Cancer, Ionosphere, Pima, Liver, and Sonar datasets in the top-down, left-right order respectively. Solid lines: the k -NN rule with the adaptive distance measure. Dashed lines: the k -NN rule with the Euclidean distance.

no confusion arises. The locally adaptive distance between \vec{X} and the training example \vec{X}_i is defined as

$$d_{\text{new}}(\vec{X}, \vec{X}_i) = \frac{d(\vec{X}, \vec{X}_i)}{r_i} \quad (6)$$

Several important points are immediately clear from the above definition. First, although the above distance measure (6) is only defined between a query pattern \vec{X} and existing training examples \vec{X}_i , the definition can be easily extended to measure the similarity between \vec{X} and an arbitrary point \vec{X}' by first defining a radius r' associated with \vec{X}' similarly to (5). Secondly, by definition, the distance function (6) is not symmetric. For example,

$$d_{\text{new}}(\vec{X}_i, \vec{X}_j) \neq d_{\text{new}}(\vec{X}_j, \vec{X}_i) \quad (7)$$

if the radii r_i and r_j associated with \vec{X}_i and \vec{X}_j , respectively are not the same. Therefore, the new distance measure is generally not a metric. Finally, according to the new distance measure, the smallest distance between a training example and training examples of other classes is one, and training examples with their distances less than one to a training example all have the same class label.

After adopting the new distance measure (6), the adaptive nearest neighbor rule works exactly the same as the original nearest neighbor rule except that we use the adaptive distance measure to replace the original L2 or L1 distance

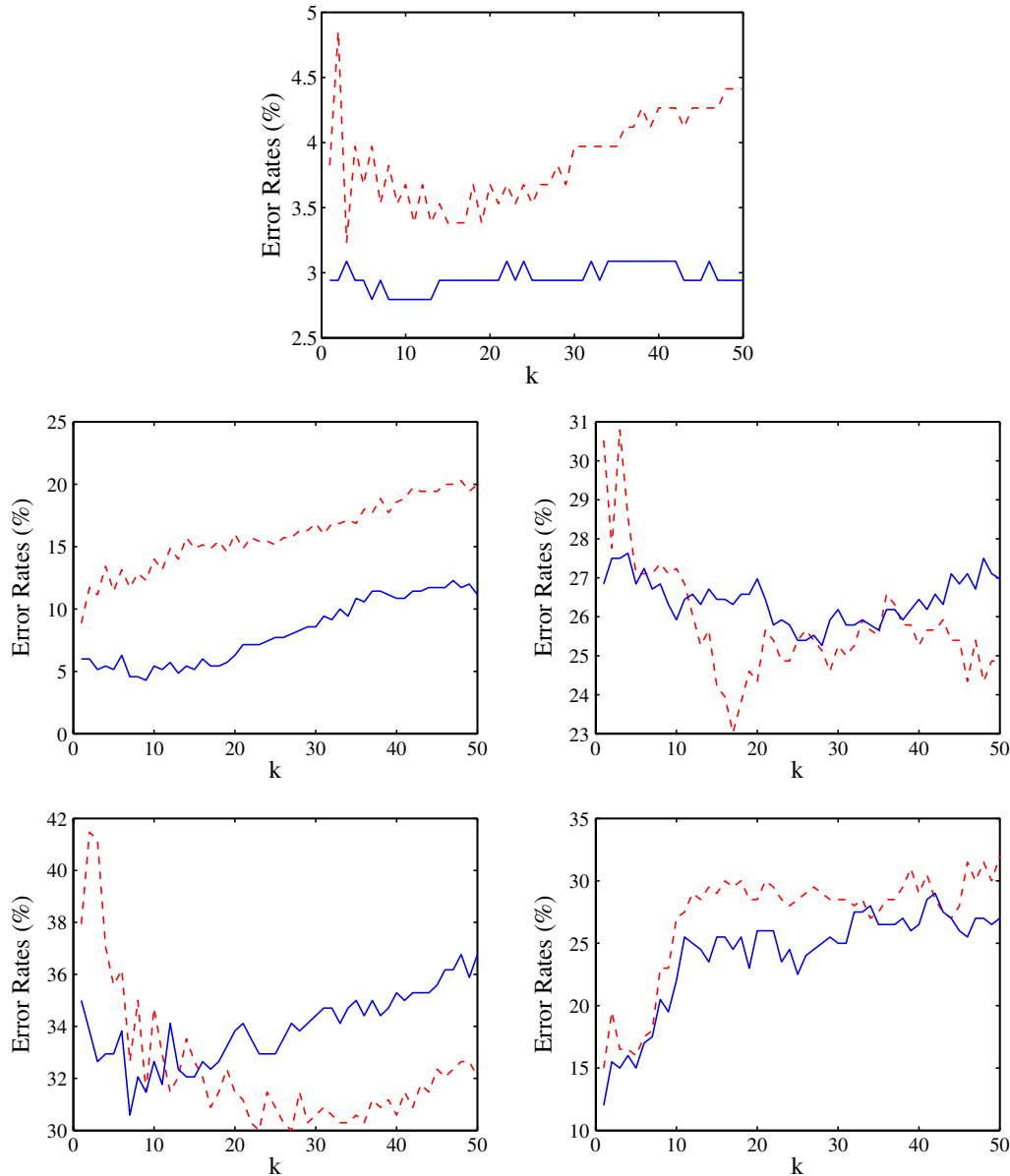


Fig. 2. Error rates at different values of k on the five datasets. Solid lines: the k -NN rule with the adaptive distance measure. Dashed lines: the k -NN rule with the Manhattan distance measure.

measure for identifying the nearest neighbors. Formally, given a query pattern \vec{X} for a binary classification problem, the adaptive nearest neighbor rule first identifies its k nearest neighbors, denoted again by $\vec{X}_{(1)}, \dots, \vec{X}_{(k)}$, according to the new distance measure $d(\vec{X}, \vec{X}_i)/r_i$ for $i = 1, \dots, n$, and classifies \vec{X} to the class

$$f(\vec{X}) = \text{sgn} \left(\sum_{i=1}^k Y_{(i)} \right). \quad (8)$$

3. Results and discussion

In this section, we present experimental results on several real-world benchmark datasets from the UCI machine

learning repository.¹ Throughout our experiments, we used the 10-fold cross validation method to estimate the generalization error. Table 1 shows the error rates and the standard deviations of the NN rule using the Euclidean distance measure (2) and our adaptive nearest neighbor (ANN) rule using the adaptive distance measure (6).

The results show that the adaptive NN rule using the simple adaptive distance measure outperforms the NN rule using the Euclidean distance measure on all five datasets being tested. On most datasets, the improvements of the adaptive NN rule is statistically significant. These results confirm that the first nearest neighbor identified according

¹ <http://www.ics.uci.edu/~mllearn/MLRepository.html>.

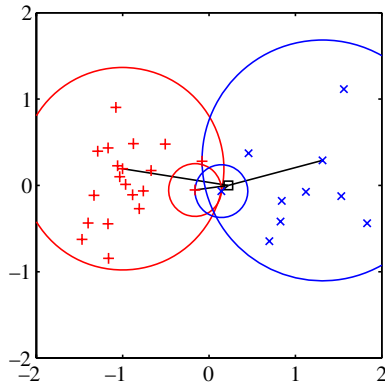


Fig. 3. The largest spheres associated with training examples that are inside the classes or near the class boundaries.

to the adaptive distance measure is more likely to have the same class label as the query pattern than the first nearest neighbor identified according to the Euclidean distance.

Fig. 1 shows the generalization errors of the k -NN rule using the two different distance measures on the five datasets at various k values. The solid lines represent the results of the k -NN rule using the adaptive distance measure, and the dashed lines are the corresponding results of the k -NN rule with the Euclidean distance. From the five plots in Fig. 1, we see that the 1-NN using the adaptive distance measure always outperforms the 1-NN rule using the Euclidean distance measure, just as shown in detail in Table 1. For k greater than 1, the k -NN rule using the adaptive distance measure also outperforms k -NN rule using the Euclidean distance measure on all five datasets when k is less than 9. On some datasets, such as the Breast Cancer, Ionosphere and Sonar datasets, using the adaptive distance measure is almost always better than using the Euclidean distance measure for k up to 50. Although on some specific datasets, there are particular k values at which the Euclidean distance measure actually performs better, it is nevertheless clear from Fig. 1 that the adaptive distance measure significantly improves the k -NN rule in general, especially at the lower k range.

Fig. 2 shows similar results of the k -NN rule using the adaptive distance measure and the Manhattan distance

measure on the five datasets. Note that in this case, the adaptive distance measure is based on the Manhattan metric, i.e., both $d(\vec{X}, \vec{X}_i)$ and the radius r_i in the definition of the adaptive distance measure $d_{\text{new}}(\vec{X}, \vec{X}_i)$ are measured in the Manhattan metric. The five plots are in the same order as in Fig. 1. The solid lines again represent the results of the k -NN rule using the adaptive distance measure, and the dashed lines represent the corresponding results of the k -NN rule with the Manhattan distance measure. Similarly to Fig. 1, we see from the five plots in Fig. 2 that the k -NN rule with the adaptive distance measure significantly outperforms the k -NN rule with the Manhattan distance measure on all five datasets when k is less than 11. On the Breast Cancer, Ionosphere and Sonar datasets, the performance of the k -NN rule using the adaptive distance measure is again always better than using the Manhattan distance measure for k up to 50.

The improvements of the adaptive distance measure on the k -NN rule can be explained as follows. We first note that, according to the locally adaptive distance measure, each training example \vec{X}_i is associated with a scaling factor r_i , which is defined to be the radius of the largest sphere centered on \vec{X}_i that excludes all training examples of other classes. The largest sphere associated with each training example defines the largest spherical region within which its class label can be generalized to other training examples reliably, i.e., without making an error. It is easy to see that spheres associated with training examples inside the classes will have relatively larger radii than those associated with training examples near the class boundaries, as illustrated in Fig. 3 that shows four spheres associated with four training examples from the two classes, with the lines connecting the centers of the spheres to a query pattern. As a result of the locally dependent scaling factor, training examples that are farther away from a query pattern according to the L2 or L1 metric may actually become closer to the query pattern according to the adaptive distance measure if their associated spheres are large enough, as illustrated in Fig. 4, where the left plot shows the nine nearest neighbors according to the Euclidean distance measure and the right plot shows the nine nearest neighbors according to the adaptive distance measure. For a query pattern

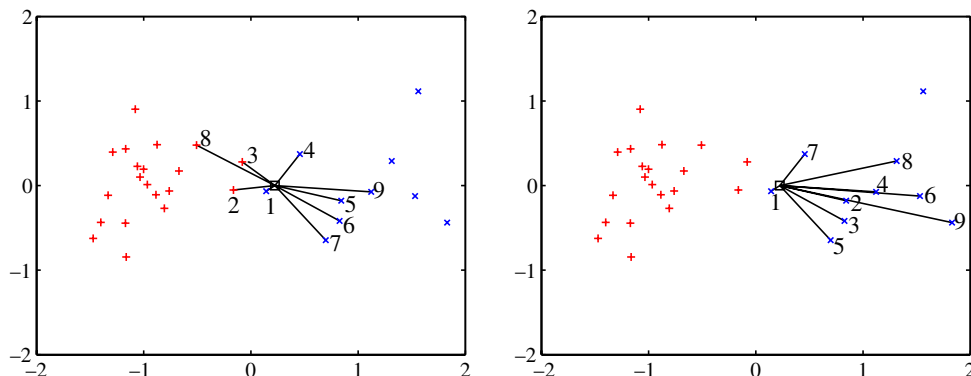


Fig. 4. Nearest neighbors according to different distance measures.

that is located far from other classes, the adaptive distance measure is unlikely to change the classification result because all the close neighbors are from the same class. However, for a query pattern near the class boundaries where different classes may overlap or the noise level is high, this feature is beneficial because it tends to identify training examples with relatively large spheres as its nearest neighbors. Compared to training examples that are near the class boundaries and have smaller spheres, which would otherwise be identified as nearest neighbors should the ordinary L2 or L1 metric be used, training examples with larger spheres are closer to the class centers and their class labels are more reliable, see Figs. 3 and 4. By dividing

the distance $d(\vec{X}, \vec{X}_i)$ from a query pattern to the training example \vec{X}_i by the corresponding sphere radius r_i , the adaptive k -NN rule relies more on training examples with larger spheres to generalize to boundary regions where local information is of high variance and unreliable.

There are two differences between the k -NN rule using the adaptive distance measure and the adaptive nearest neighbor algorithm we proposed previously. First, the radius used in the definition of the adaptive distance measure is different. In this work, the radius r_i associated with training example \vec{X}_i is defined to be simply the smallest distance between \vec{X}_i and training examples of a different class, while in our previous work, the radius r_i associated with \vec{X}_i

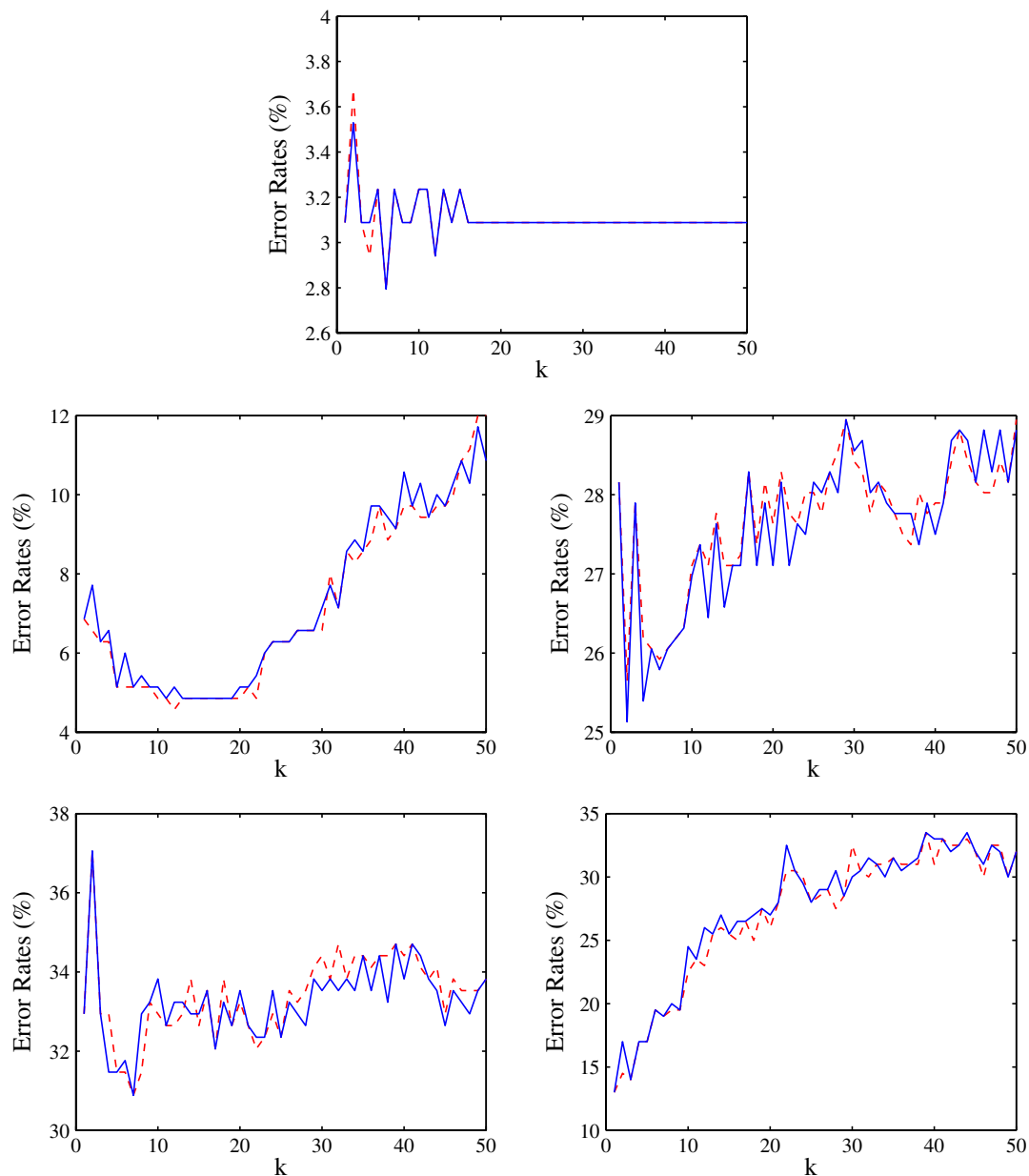


Fig. 5. Error rates at different values of k on the five datasets. The plots are for Breast Cancer, Ionosphere, Pima, Liver, and Sonar datasets in the top-down, left-right order respectively. Solid lines: k -NN rule with the adaptive distance measure. Dashed lines: k -NN rule with both the adaptive distance measure and the weighting scheme.

Table 2
Comparison of results

Dataset	k -NN (L2)	A- k -NN (L2)	k -NN (L1)	A- k -NN (L1)	SVMs
Breast cancer	2.79 (0.67)	2.79 (0.74)	3.24 (0.81)	2.79 (0.60)	3.68 (0.66)
Ionosphere	12.86 (1.96)	4.86 (1.28)	8.86 (1.93)	4.29 (0.88)	4.86 (1.05)
Pima	24.61 (1.36)	25.13 (1.46)	23.03 (1.75)	25.26 (1.54)	27.50 (1.68)
Liver	30.88 (3.32)	30.88 (1.77)	30.00 (2.43)	30.59 (2.37)	31.47 (2.63)
Sonar	17.00 (2.26)	13.00 (1.70)	15.00 (2.47)	12.00 (2.60)	11.00 (2.33)

may have to be increased from the smallest distance to meet a preset statistical confidence requirement. In fact, the two definitions coincide with each other if we set the minimum required statistical confidence level to 50%. Secondly, once the k nearest neighbors are identified, in this work, we simply used the plain k -NN rule, while in our previous work, each nearest neighbor is weighted differently according to its associated statistical confidence, see (Wang et al., 2005) for details. To see if the weighting scheme plays any significant role, we obtained the error rates of the k -NN rule using the adaptive distance measure and the adaptive nearest neighbor algorithm with the weighting scheme at different values of k . To ensure that the only difference is in the weighting scheme, we used the same L2-based radius definition (5) in both algorithms. The results are shown in Fig. 5. The solid lines illustrate the results of the k -NN rule with adaptive distance measure and the dashed lines illustrate the results of the adaptive nearest neighbor algorithm with the weighting scheme. As we can clearly see from the five plots, the weighting scheme does not significantly improve the performance. Therefore, we conclude that the most important factor that leads to the performance improvements is the adaptive distance measure, which is simply the original Euclidean distance measure divided by the smallest distance to other classes.

In Table 2, we report the lowest error rates of the k -NN rule using the Euclidean (L2) and Manhattan (L1) metrics and the corresponding adaptive distance measures and compare them to the lowest error rates of the SVMs with Gaussian kernels. On each dataset, we run the k -NN rule using all four distance measures at various values of k from 1 to 50 and picked the lowest error rate. As we can see from Table 2, the k -NN rule with the adaptive distance measures performs significantly better than the k -NN rule with the Euclidean and Manhattan distance measures on the Breast Cancer and Sonar datasets, making the adaptive k -NN rule overall better than or comparable to the state-of-the-art SVMs.

4. Conclusion

In this paper, we demonstrated that an extremely simple adaptive distance measure significantly improves the performance of the k -NN rule. In our tests on several real-

world datasets, the resulting adaptive k -NN rule actually achieves consistently better or comparable performance to the state-of-the-art Support Vector Machines. The advantage of our adaptive k -NN rule over SVMs and other adaptive metric methods, however, is apparent. The adaptive k -NN rule is simply the k -NN rule with the conventional distance measure, be it the Euclidean or Manhattan metric, divided by the smallest distances from the corresponding training examples to training examples of different classes. We believe the simplicity of this algorithm and its great performance makes it an appealing tool for pattern classification.

Acknowledgement

This work was supported in part by ARO under Grants W911NF-04-1-0357.

References

- Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. *IEEE Trans. Information Theory* IT-13 (1), 21–27.
- Domeniconi, C., Peng, J., Gunopulos, D., 2002. Locally adaptive metric nearest neighbor classification. *IEEE Trans. Pattern Anal. Machine Intell.* 24, 1281–1285.
- Fix, E., Hodges, J., 1951. Discriminatory analysis, nonparametric discrimination: consistency properties. Tech. Rep. 4, USAF School of Aviation Medicine, Randolph Field, Texas.
- Friedman, J., 1994. Flexible metric nearest neighbor classification. Tech. Rep. 113, Stanford University Statistics Department.
- Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R., 2004. Neighbourhood component analysis. In: *Neural Information Processing Systems (NIPS)* 17. pp. 513–520.
- Hastie, T., Tibshirani, R., 1996. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Machine Intell.* 18 (6), 607–615.
- Stone, C.J., 1977. Consistent nonparametric regression. *Ann. Statist.* 5, 595–645.
- Wang, J., Neskovic, P., Cooper, L.N., 2005. An adaptive nearest neighbor rule for classification. In: *Proc. 4th Internat. Conf. on Machine Learning and Cybernetics (ICMLC)*, Lecture Notes in Artificial Intelligence (LNAI).
- Wang, J., Neskovic, P., Cooper, L.N., 2006. Neighborhood selection in the k -nearest neighbor rule using statistical confidence. *Pattern Recognition* 39, 417–423.
- Weinberger, K., Blitzer, J., Saul, L., 2005. Distance metric learning for large margin nearest neighbor classification. In: Y. Weiss, B.S., Platt, J. (Eds.), *Advances in Neural Information Processing Systems (NIPS)* 18.