Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier

Paul Horton
Computer Science Division 387 Soda Hall
University of California, Berkeley, CA 94720
paulh@cs.berkeley.edu
phone: (510)-704-8895

Kenta Nakai
Institute of Molecular and Cellular Biology
Osaka University,
1-3 Yamada-oka, Suita 565, Japan
nakai@imcb.osaka-u.ac.jp
phone: +81-6-879-7992, fax: +81-6-877-1922

Abstract

We have compared four classifiers on the problem of predicting the cellular localization sites of proteins in yeast and *E.coli*. A set of sequence derived features, such as regions of high hydrophobicity, were used for each classifier. The methods compared were a structured probabilistic model specifically designed for the localization problem, the k nearest neighbors classifier, the binary decision tree classifier, and the naïve Bayes classifier. The result of tests using stratified cross validation shows the k nearest neighbors classifier to perform better than the other methods. In the case of yeast this difference was statistically significant using a cross-validated paired t test. The result is an accuracy of approximately 60% for 10 yeast classes and 86% for 8 *E.coli* classes. The best previously reported accuracies for these datasets were 55% and 81% respectively.

Keywords: Protein Localization, k Nearest Neighbor Classifier, Classification, Yeast, *E.coli*

Introduction

In order to function properly, proteins must be transported to various localization sites within the cell. Conversely, the cellular localization site of a protein affects its potential functionality as well as its accessability to drug treatments. Fortunately the information needed for correct localization is generally found in the protein sequence itself.

The first integrated system for predicting the localization sites of proteins from their amino acid sequences was an expert system (Nakai & Kanehisa 1991; 1992). This system is still useful and popular but it is unable to learn how to predict on its own and therefore very time consuming to update or adapt to new organisms. In more recent work, expert identified features were combined with a probablistic model which could learn its parameters from a set of training data (Horton & Nakai 1996). This model was successful in the sense that it required significantly less labor from the human expert and yielded a similar prediction accuracy to the expert system. However in machine learning, classification (hereafter we will refer to prediction as classification) from labeled examples is a relatively mature field which offers many alternatives to the probabilistic model.

In this paper we investigate the classification accuracy of three standard classification algorithms, namely the k nearest neighbors classifier (kNN), the binary decision tree, and the naïve Bayes classifier; as well as the probabilistic model. To provide an additional baseline we also compare the classification accuracy of using kNN with the PAM120 local alignment distance instead of the expert identified features. In the first section we briefly describe the datasets, classifiers, and testing methodology. In the results section we give the accuracy results of cross-validation tests for the four methods and also for kNN with local alignment distances. Also in the results section, we investigate the affects of varying the k parameter and report which misclassifications are typical of the classifiers. Finally we summarize our results and conclude.

Methods and Materials

Datasets

The datasets used have been submitted to the UCI Machine Learning Data Repository (Murphy & Aha 1996) and are described in (Horton & Nakai 1996), (Nakai & Kanehisa 1991), and (Nakai & Kanehisa 1992). We used two datasets: an *E.coli* dataset with 336 proteins sequences labeled according to 8 classes (localization sites) and a yeast dataset with 1462 sequences labeled according to 10 classes. The occurrence of classes for the datasets are summarized in tables 1 and 2. The 1462 yeast sequences were obtained by removing 12 sequences which occurred twice (with different names) in the original 1484 sequence dataset. The same feature variables were used as those described in (Horton & Nakai 1996).

Classifiers

We investigated the performance of four classifier algorithms. The first method is the probabilistic method (hereafter referred to as HN) specifically designed for the protein localization problem described in (Horton & Nakai 1996). The only modification we applied to that work is that we used Fayyad-Irani binning

Classes for the *E.coli* Dataset

Class	Abbr.	Number
Cytoplasm	cp	143
Inner membrane, no signal sequence	im	77
Periplasm	pp	52
Inner membrane, uncleavable signal sequence	imU	35
Outer membrane non-lipoprotein	om	20
Outer membrane lipoprotein	omL	5
Inner membrane lipoprotein	imL	2
Inner membrane, cleavable signal sequence	imS	2

Table 1: The names, abbreviations and number of occurrences of each class for the *E.coli* dataset are shown.

Classes for the Yeast Dataset

Class	Abbr.	Number
Cytoplasm	CYT	444
Nucleus	NUC	426
Mitochondria	MIT	244
Membrane protein, no N-terminal signal	ME3	163
Membrane protein, uncleaved signal	ME2	51
Membrane protein, cleaved signal	ME1	44
extracellular	EXC	35
Vacuole	VAC	30
Peroxisome	POX	20
Endoplasmic Reticulum	ERL	5

Table 2: The names, abbreviations and number of occurrences of each class for the yeast dataset are shown.

(Fayyad & Irani 1993) for the discretization of continuous feature variables. The other three classifiers are standard classifiers from the fields of machine learning and pattern recognition which we will only describe briefly.

k Nearest Neighbors The k nearest neighbors classifier (Duda & Hart 1973) stores the complete training data. New examples are classified by chosing the majority class among the k closest examples in the training data. For our particular problem, we first used a linear transformation to normalize the feature values to lie within the interval [0,1] and then used the Euclidean, i.e. sum of squares, distance to measure the distance between examples.

Binary Decision Tree

Binary decision trees (Quinlan 1986) recursively split the feature space based on tests that test one feature variable against a threshold value. We used the information gain criteria for choosing the best test, and top-down pruning with a χ^2 value of 0.95 to reduce overfitting.

Naïve Bayes Classifier

The Naïve Bayes classifier (Good 1965), (Langley, Iba, & Thompson 1992) is an approximation to an ideal Bayesian classifier which would classify an example based on the probability of each class given the example's feature variables. The main assumption is that the different features are independent of each other given the class of the example.

Software

The four classifiers were implemented in C and Perl and may be obtained by request from the authors.

Evaluation Methodology

We used stratified cross-validation to estimate the accuracy of the classifiers. In this procedure the dataset is randomly partitioned into equally sized partitions subject to the constraint that the proportion of the classes in each partition is equal. Empirical tests have indicated that this procedure provides more accurate estimates than plain cross-validation (Kohavi 1995).

We employed a cross-validated paired-differences t test to establish the statistical significance of the difference in performance between two classifiers (Kohavi 1995) (a general description of the paired t test for hypothesis testing can be found in introductory textbooks on statistics, for example (Larsen & Marx 1986)). This test makes two assumptions. One assumption is that the difference of the performance of the two algorithms is normally distributed. The second assumption is that the performance difference on different test partitions of the cross-validation is independent. In general both of these assumptions may be violated, in particular the training partitions of the

Results with the *E.coli* Dataset for 4 Classifiers

Partition	kNN	Dec. Tree	Naïve Bayes	HN
0	89.28	83.33	82.14	84.42
1	95.24	80.95	84.52	88.10
2	84.52	88.10	82.14	88.10
3	76.19	69.05	75.00	69.05
mean	86.31	80.36	80.95	82.44
std. dev.	8.04	8.10	4.12	9.08

Table 3: The results of cross-validation are shown in units of percent accuracy, including the mean and sample standard deviation. HN is the probabilistic model of Horton & Nakai. All trials of kNN are for $k = 7$.

cross-validation overlap heavily and thus the trials are not independent (Salzberg 1995). Despite these observations, the t test has been shown empirically to discriminate adequately (Dietterich 1996).

Results

A summary of the accuracies of the different classifiers is given in table 3 for *E.coli* and table 4 for yeast. Accuracies for the smaller *E.coli* dataset were estimated with 4-fold cross-validation to keep the test partitions reasonably large. It can be seen that the mean accuracy of kNN is higher than the other 3 classifiers for both datasets. Using the cross-validated paired t test to test whether the mean accuracy of kNN is different than the other classifiers gives t values of 2.86, 2.59, and 2.88 against the binary decision tree, Naïve Bayes, and HN respectively. For a two-sided t test with nine degrees of freedom the t value corresponding to a confidence level of 0.95 is 2.2622. By the same t test the only significant difference for the *E.coli* dataset is the difference between kNN and Naïve Bayes which has a t value of 3.3169 and is significant at a confidence level of 0.95.

k Parameter

For accuracy estimation we used k values for *E.coli* and yeast datasets of 7 and 21 respectively. We determined those values by doing leave-one-out cross-validation on each training partition (this is a nested cross-validation) and taking the best overall value. Since this procedure averages over all the data and therefore indirectly uses the test data, it is important to know how sensitive the classification accuracy is to the choice of k. figure 1 shows the relationship between the k value and accuracy estimated by cross-validation for the *E.coli* dataset. The accuracy is highest for k values of 5 and 7 but is higher than the other three classifiers from $k = 3$ to $k = 25$. figure 2 shows the corresponding graph for yeast. With the larger yeast dataset the highest accuracy is achieved for k values from 21 to 25, but the accuracy for kNN is higher than the other classifiers for values of k from 9 to 99. We did not calculate the accuracy for $k > 99$.

Results with the Yeast Dataset for 4 Classifiers

Partition	kNN	Dec. Tree	Naïve Bayes	HN
0	55.78	55.10	53.74	55.10
1	59.18	51.02	57.82	55.78
2	60.96	56.16	56.16	58.22
3	65.75	58.22	58.22	55.48
4	48.63	50.00	45.21	47.95
5	62.33	57.53	54.11	53.42
6	68.49	65.75	60.27	67.81
7	58.90	57.53	61.64	56.16
8	56.85	56.85	56.16	55.48
9	58.22	57.53	59.59	57.53
mean	59.51	56.57	56.29	56.29
std. dev.	5.49	4.30	4.66	4.93

Table 4: The results of cross-validation are shown in units of percent accuracy, including the mean and sample standard deviation. HN is the probabilistic model of Horton & Nakai. All trials of kNN are for $k = 21$.

Local Alignment Distance with kNN

To provide a baseline comparison for the effectiveness of the expert identified features used we calculated the accuracy of kNN with the local alignment distances calculated using the PAM120 matrix. Using the same cross-validation partitions and criteria for choosing k we obtained an accuracy of 67.86% on the *E.coli* dataset. This is much higher than the 42.9% accuracy that the majority class classifier achieves but is much lower than the 86.31% accuracy achieved by kNN using the expert identified features.

On the yeast dataset the local alignment distances did relatively better, yielding an accuracy of 52.1%. However, the t test still shows this accuracy to be lower than the accuracy of the four classifiers with the expert identified features at a confidence level of 0.97.

Confusion Matrices

In order to identify common misclassifications we calculated the confusion matrix for both datasets using kNN with the expert identified features. These results are shown in tables 5 and 6.

Discussion

The confusion matrix for *E.coli* is very encouraging in that most of the mistakes can be seen to result from confusing inner membrane proteins without a signal sequence with inner membrane proteins with an uncleavable signal sequence and *vice versa*. We consider this a relatively minor error for two reasons. First, for some uses the distinction between different types of inner membrane proteins may be immaterial. Second, the definition of the presence or absence of an uncleavable signal sequence is somewhat arbitrary and thus the labels for some training examples include some uncertainty. If we collapse the two classes to form a class

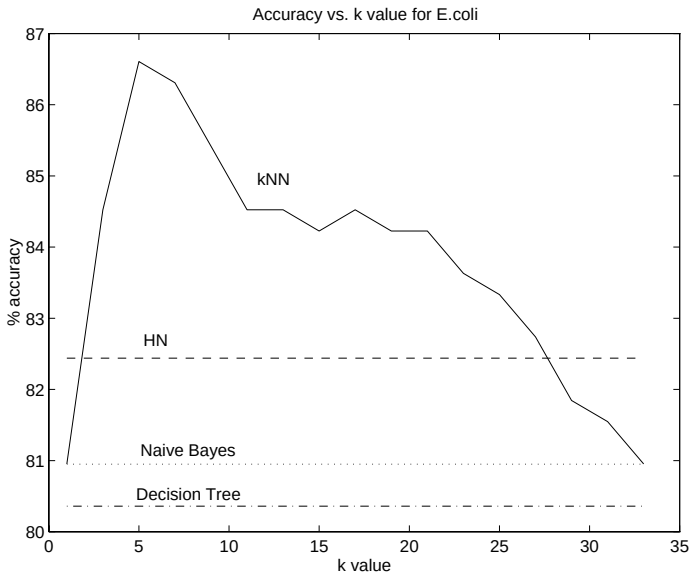

Figure 1: The accuracy of kNN for the *E.coli* dataset is shown for odd k from 1 to 33. The accuracy of the decision tree, Naïve Bayes, and HN is also shown.

"inner membrane protein without a cleavable signal sequence" we attain a surprisingly high accuracy of 94%!

The confusion matrix for the yeast dataset shows that most of the error is due to confusing cytoplasmic proteins with nuclear proteins and *vice versa*. This reflects a fundamental difficulty in identifying nuclear proteins. One component of the difficulty comes from the fact that unlike other localization signals the nuclear localization signal does not appear to be limited to one portion of a protein's primary sequence (Garcia-Bustos, Heitman, & Hall 1991). Another component is the fact that in some cases a protein without a nuclear localization signal may be transported to the nucleus as part of a protein complex if another subunit of the complex contains a nuclear localization signal (Zhao & Padmanabhan 1988).

Another interesting result is the relatively low accuracy of using kNN with the local alignment distance. This is interesting because the common practice of inferring protein function by homology search of the databases is essentially a variant of kNN with local alignment distance. Our results show that localization site prediction is an example of a protein classification problem where domain specific features are much more effective than homology alone.

One question we would like to answer is why kNN was more effective than the other classifiers. It is easy to point out some shortcomings with the other classifiers, the binary decision tree and HN suffer from data fragmentation as the data is repeatedly partitioned. Naïve Bayes has a fixed number of parameters and does not asymptotically approach an optimal classifier as the number of training examples increases. However we do not have a solid answer as to why kNN performs better on this task.

Confusion Matrix for *E.coli* dataset with kNN

	cp	imL	imS	imU	im	omL	om	pp
cp	141	0	0	0	0	0	0	2
imL	0	0	0	0	1	1	0	0
imS	0	0	0	1	0	0	0	1
imU	1	0	0	23	11	0	0	0
im	3	0	0	14	58	0	0	2
omL	0	0	0	0	0	4	1	0
om	0	0	0	0	0	0	18	2
pp	4	0	0	0	1	0	0	47

Table 5: The actual class labels are shown in the vertical column. The predicted class labels are shown in the row across the top. Thus 2 proteins that localize to the cytoplasm were incorrectly predicted to be localized to the periplasm.

In summary we have demonstrated that kNN with expert identified features is superior to three other classifiers for classifying proteins based on their cellular localization sites. For the yeast dataset this difference can be shown to be statistically significant. We have also shown that the expert identified features are much more effective than local alignment distance and that most of the classification errors on the *E.coli* dataset are relatively minor errors. The use of kNN and better testing methodology has allowed us to achieve estimated accuracies of 60% and 86% for the yeast and *E.coli* datasets respectively, exceeding the best previously reported accuracies of 55% and 81%.

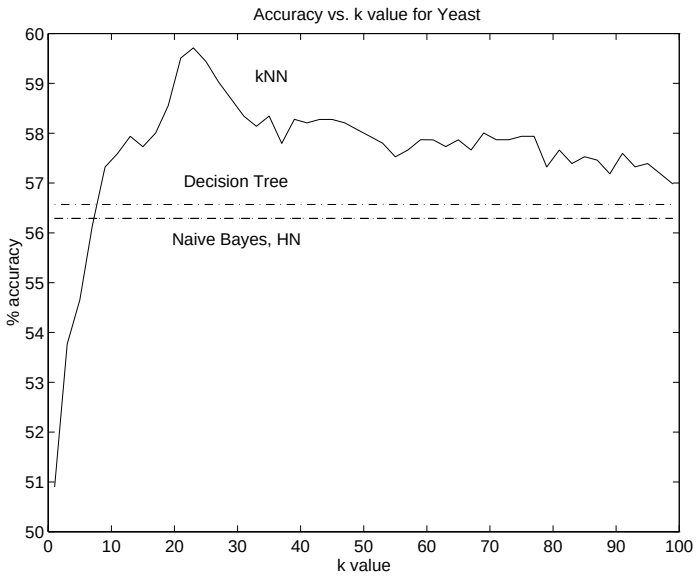

Figure 2: The accuracy of kNN for the yeast dataset is shown for odd k from 1 to 99. The accuracy of the decision tree, Naïve Bayes, and HN is also shown.

Confusion Matrix for yeast dataset with kNN

	cyt	erl	exc	me1	me2	me3	mit	nuc	pox	vac
cyt	314	0	1	0	2	3	32	91	1	0
erl	0	0	3	1	1	0	0	0	0	0
exc	4	0	22	4	2	0	2	1	0	0
me1	0	0	8	33	0	1	2	0	0	0
me2	9	0	7	10	11	3	7	4	0	0
me3	18	0	0	0	1	122	6	16	0	0
mit	62	0	4	2	5	8	141	19	3	0
nuc	171	0	0	0	2	10	27	216	0	0
pox	4	0	1	1	0	0	1	2	11	0
vac	13	0	3	1	1	6	1	5	0	0

Table 6: The actual class labels are shown in the vertical column. The predicted class labels are shown in the row across the top.

Acknowledgements

KN was supported by a grant-in-aid for scientific research on the priority area "Genome Informatics" from the Ministry of Education, Science and Culture, Japan.

References

Dietterich, T. G. 1996. Statistical tests for comparing supervised classification learning algorithms. http://www.CS.ORST.EDU/ tgd/cv/jr.html.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. John Wiley & Sons.

Fayyad, U. M., and Irani, K. B. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In *International Joint Conference on Artificial Intelligence*, 1022–1027.

Garcia-Bustos, J.; Heitman, J.; and Hall, M. N. 1991. Nuclear protein localization. *Biochimica et Biophysica Acta* 1071:83–101.

Good, I. J. 1965. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press.

Horton, P., and Nakai, K. 1996. A probabilistic classification system for predicting the cellular localization sites of proteins. In *Proceeding of the Fourth International Conference on Intelligent Systems for Molecular Biology*, 109–115. St. Louis: AAAI Press.

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*.

Langley, P.; Iba, W.; and Thompson, K. 1992. An analysis of bayesian classifiers. In *Proceedings of the tenth National Conference on Artificial Intelligence*, 223–228. AAAI Press and MIT Press.

Larsen, R. J., and Marx, M. L. 1986. *An Introduction to Mathematical Statistics and its Applications*. Prentice-Hall.

Murphy, P. M., and Aha, D. W. 1996. Uci repository of machine learning databases. http://www.ics.uci.edu/ mlearn.

Nakai, K., and Kanehisa, M. 1991. Expert system for predicting protein localization sites in gram-negative bacteria. *PROTEINS: Structure, Function, and Genetics* 11:95–110.

Nakai, K., and Kanehisa, M. 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14:897–911.

Quinlan, J. R. 1986. Induction of decision trees. *Machine Learning* 1:81–106.

Salzberg, S. 1995. On comparing classifiers: A critique of current research and methods. http://www.cs.jhu.edu/ salzberg.

Zhao, L., and Padmanabhan, R. 1988. Nuclear transport of adenovirus dna polymerase is facilitated by interaction with preterminal protein. *Cell* 55:1005–1015.