# Probability Based Metrics for Nearest Neighbor Classification and Case-Based Reasoning

Enrico Blanzieri and Francesco Ricci*

Istituto per la Ricerca Scientifica e Tecnologica (ITC-IRST)
38050 Povo (TN)
Italy
blanzier@irst.itc.it - ricci@sodalia.it

**Abstract.** This paper is focused on a class of metrics for the Nearest Neighbor classifier, whose definition is based on statistics computed on the case base. We show that these metrics basically rely on a probability estimation phase. In particular, we reconsider a metric proposed in the 80's by Short and Fukunaga, we extend its definition to an input space that includes categorical features and we evaluate empirically its performance. Moreover, we present a novel probability based metric, called Minimum Risk Metric (MRM), i.e. a metric for classification tasks that exploits estimates of the posterior probabilities. MRM is optimal, in the sense that it optimizes the finite misclassification risk, whereas the Short and Fukunaga Metric minimizes the difference between finite risk and asymptotic risk. An experimental comparison of MRM with the Short and Fukunaga Metric, the Value Difference Metric, and Euclidean–Hamming metrics on benchmark datasets shows that MRM outperforms the other metrics. MRM performs comparably to the Bayes Classifier based on the same probability estimates. The results suggest that MRM can be useful in case-based applications where the retrieval of a nearest neighbor is required.

## 1 Introduction

Nearest Neighbor (NN) algorithms are a well-known and intensively studied class of techniques for the solution of Classification and Pattern Recognition problems. Nowadays NNs are widely exploited for the retrieval phase in the majority of Case Based Reasoning (CBR) systems. In CBR, even if cases are not explicitly classified into a set of finite groups (classes), often the solution space can be clustered into a collection of sets each of them containing similar solutions. When such a set of similar solution is labelled with a class tag, it is natural to match the retrieval step in a CBR system with the nearest neighbor search in a NN classifier [3]. In this framework, for example, Bellazzi et al. [4] have shown that the performance of a CBR system can be improved by driving the retrieval with the information about the same relevant classification in the case space, i.e. reducing the retrieval problem to a classification task. In this

* Current Address: Sodalia S.p.A., 38100 Trento, Italy

perspective, improving the classification accuracy for NN algorithms becomes important for CBR.

The NN classification procedure is straightforward: given a set of classified examples, which are described as points in an input space, a new unclassified example is assigned to the known class of the nearest example. The "nearest" relation is computed using a (similarity) metric defined on the input space. Many researchers [22–24, 10, 1, 2, 13, 12, 19, 20, 6, 26] focused their attention on the use of local metrics, i.e. metrics that vary depending on the position of the points in the input space. Conversely, more traditional global metrics assume that similarity evaluation should be independent from the area of the input space the cases to be compared are taken from. There are pros and cons in using local metrics. On the one hand local metrics generate classifiers that are more sensitive to the local changes of the data and hence more accurate. On the other hand global metrics have fewer parameters and consequentially the classifiers are computationally lighter and less prone to the effect of noisy data. The critical point seems to be the grade of locality of the metric: choosing the 'right' locality in different areas of the input space should lead to better descriptions of the separating surfaces.

Some of the proposed local metrics rely for their effectiveness on the optimization of a given criterion and ultimately on the estimation of some probabilities. In this direction Short and Fukunaga [23] presented a seminal work constrained to a multidimensional numerical input space. They proposed to minimize the expected value of the difference between the misclassification error of a two-classes NN classifier with a finite number of samples and the misclassification error hypothetically achievable with an infinite sample. They expressed the optimal local metric in terms of a linear estimation of posterior probabilities. More recently in the instance based learning context, many proposals of nominal feature metrics also involve probability estimation [24, 10, 8, 25]. In these cases the probability estimation is performed computing frequencies of value occurrences. Finally, in the work by Wilson and Martinez [27] the estimation of probabilities provides an unifying framework for treating both linear (continuous or discrete) and nominal features. Their heterogeneous distances, wich extend the VDM metric [24], deal uniformly with both categorical and numerical features. On a different perspective and in the Bayesian framework, Kontaken et al. [15] proposed a matching function explicitly based on probabilistic considerations.

In spite of the centrality of the probability estimation issue in the metrics briefly described above, there is no unifying description in the literature of the impact of different approaches to the solution of this issue (with a notable exception in [14]). Furthermore, little or no attempt has been made to exploit the advanced nonparametric density estimation techniques developed by the applied statistics community [21] and their possible extensions to nominal features.

In this paper we describe a couple of techniques for probability estimation and their use inside two metrics based on this estimation (Short and Fukunaga and Minimum Risk Metric). From our point of view the approach of construct-

ing metrics via combination of well-known probability estimators and optimal metrics presents several advantages.

- *The metrics have a clear analytical expression and motivation.* For example the metric proposed by Short and Fukunaga minimizes the difference between asymptotic and finite risk. That makes these metrics amenable to analytical study.
- *The metrics can be computed using standard density estimation techniques.* Advances in that area can be reused here. For example, the choice of the right degree of locality can rely on the solutions proposed for the choice of the bandwidth in the nonparametric density estimation models.
- *The metrics can be defined uniformly on data sets with both numeric and nominal attributes.* This point is extremely important for CBR applications. Combining different metrics on the categorical and numerical features usually lead to poor performances [27].

Regarding the last item described above, in real world case bases both continuous and nominal features can be useful to describe cases. That poses a new problem: how to sum contributions to the distance evaluation that come from the comparison of pairs of categorical values together with pairs of real numbers? This problem can be tackled in different ways:

- *Ordering.* Ordering and numbering the values of the nominal features and applying a numerical metric like the Euclidean one. In general this approach introduces artificial neighborhood.
- *Discretization.* Discretizing the numeric features and applying a nominal metric to them, e.g. Hamming or Value Difference Metric [24]. With discretization some information is inevitably lost and parameters of the discretization can be critical.
- *Combination.* This is the most common approach in CBR, it consists of combining two metrics, a nominal and a numeric, each one used on the corresponding part of a case. A very common example of a metric in this class is that obtained by combining the Euclidean and Hamming metric. Combined metrics are hard to adapt in a consistent way and perform poorly, as Wilson and Martinez have shown [27].

Conversely, metrics based on probability estimation provide a natural unifying framework for dealing with both kind of features. The same probability estimation technique is used for both type of features. Furthermore, the optimality evaluation that can be done with this type of metric is impossible when the metric is obtained by combination.

Among the metrics based on probability estimation the one proposed by Short and Fukunaga has a strong theoretical foundation. The original definition was applicable only to cases described uniquely by numeric features. In this paper we extend its definition to the most general situation, i.e., with both type of features, by considering different and more general probability estimators than those exploited by the authors. We call this metric SF2.

Experimental results presented in this paper show that SF2 outperforms more standard metrics but only when we explicitly restricting the scope of application of the metric (locality) or cross–validating the estimator. The analysis of SF2 leads us to a deeper evaluation of the optimality condition underlying the Short and Fukunaga metric and eventually to the definition of an alternative metric.

We propose here another metric, called Minimum Risk Metric (MRM), that relies its effectiveness on a different and simpler optimality condition than that suggested by Short and Fukunaga. In fact MRM minimizes directly the finite misclassification risk. In order to test the effectiveness of the approach we run experiments on 29 benchmark datasets and compare the classification accuracies of Short and Fukunaga Metric and MRM with the performances of other metrics available in the literature.

The work is organized as follows: Section 2 describes the metrics studied in this paper, in particular Subsection 2.4 briefly presents the Minimum Risk Metric and its optimality criterium. Section 3 describes the adopted probability estimators. Section 4 presents the experimental results and finally Section 5 draws conclusions and future directions.

## 2 Metrics

In this Section we will briefly present four families of metrics studied in this work. The first was introduced by Short and Fukunaga in the 80's [23] and is not well know in CBR mostly because athe original definition seemed confined to cases with only numerical features. The second family originates from the well known Value Difference Metric (VDM) of Stanfill and Waltz [24] and has stemmed a number of other metrics, most notably those introduced by Wilson and Martinez [27]. Third, we recall the very common metric that combines the Euclidean and the Hamming distances. Fourth, we introduce our novel metric called Minimum Risk Metric (MRM).

### 2.1 Short and Fukunaga Metric (SF2)

Short and Fukunaga [23] were among the first to derive a NN optimal metric relying on probabilistic considerations. In their work they consider a two-class pattern recognition task. Let $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n)$ be two examples in $[0,1]^n$. Let, $p(c_1|x)$ be the probability that the example $x$ be in class $c_1$. Then $r(x,y) = p(c_1|x)p(c_2|y) + p(c_2|x)p(c_1|y)$ is the finite 1-nearest neighbor error rate at $x$ ( i.e., the probability of misclassifying $x$ by the 1-nearest neighbor rule given that the nearest neighbor of $x$ using a particular metric is $y$) and $r^*(x) = 2p(c_1|x)p(c_2|x)$ is the asymptotic 1-nearest neighbor error rate (i.e., the probability of misclassifying $x$ by the 1-nearest neighbor rule, given a hypothetically infinite design set [9]). Short and Fukunaga show that minimizing the expectation $E[(r(x,y) - r^*(x))^2]$ is equivalent to minimizing $E[(p(c_1|x) - p(c_1|y))^2]$, so the best local metric is:

$$\mathrm{SF2}(x,y) = |p(c_1|x) - p(c_1|y)| \tag{1}$$

We shall call this metric SF2. Short and Fukunaga approximate at the first order $|p(c_1|x) - p(c_1|y)| \simeq |\nabla p(1|x)^T(x-y)|$ and therefore their metric in the original formulation can be applied only to numeric features and in a local restriction.

Myles and Hand in [18] generalize that metric to a multiclass problem and introduce the following two:

$$\mathrm{SF2}(x,y) = \sum_{i=1}^{m} |p(c_i|x) - p(c_i|y)| \tag{2}$$

$$\mathrm{SFM}(x,y) = \sum_{i=1}^{m} p(c_i|x)|p(c_i|x) - p(c_i|y)| \tag{3}$$

where the classes $c_i$ are numbered from 1 to $m$. We shall still call the first metric SF2, and SFM the second. It is easy to prove that on a two classes classification problem SF2 and SFM coincide. Myles and Hand use the same technique introduced by Short and Fukunaga to approximate $|p(c_i|x) - p(c_i|y)|$.

### 2.2 Value Difference Metric (VDM)

Another very common metric based on probabilistic consideration is VDM introduced by Stanfill and Waltz [24] and used exclusively on input spaces with nominal attributes. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be two examples in $\prod_{j=1}^{n} F_j$, and $|F_j|$ is finite. The VDM metric is defined as follow:

$$\mathrm{VDM}(x,y) = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} \left( \frac{N(x_j,c_i)}{N(x_j)} \right)^2} \sum_{i=1}^{m} \left( \frac{N(x_j,c_i)}{N(x_j)} - \frac{N(y_j,c_i)}{N(y_j)} \right)^2 \tag{4}$$

where $N(x_j,c_i)$ is the number of examples that have value $x_j$ for the $j$-th attribute and are in class $c_i$, and $N(x_j)$ is the number of examples that have value $x_j$ for the $j$-th attribute. If probabilities are estimated with frequency counts then VDM can also be written in the following form:

$$\mathrm{VDM}(x,y) = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} (p(c_i|x_j)^2} \sum_{i=1}^{m} (p(c_i|x_j) - p(c_i|y_j))^2 \tag{5}$$

VDM has no clear justification and seems to assume attribute independence. It is easy to conceive an ill-formed dataset where all the $p(c_i|x_j)$ are equal (for example the parity bit class) and therefore VDM is not able to distinguish among the classes. Nevertheless VDM, and a set of modified versions [10, 8, 27], works quite well on real data sets.

Wilson and Martinez extended VDM to instances with numeric attributes [27]. They essentially discretize the numeric attributes (DVDM) and then smooth

the histogram estimation of $p(c_i|x_j)$ by averaging. The metric obtained with that procedure is called IVDM. They also suggest a heterogeneous VDM that combines an Euclidean metric for numeric features with a VDM, called HVDM. The version of VDM, and the correspondig metrics IVDM and HVDM, we adopted in our experiments is the version without weighting factors and with the absolute values:

$$\mathrm{VDM}(x,y) = \sum_{j=1}^{n}\sum_{i=1}^{m} |p(c_i|x_j) - p(c_i|y_j)| \tag{6}$$

An empirical study [7] has shown that this version of VDM (IVDM and DVDM) behaves better than the others.

### 2.3 Combined Euclidean–Overlap Metric (HEOM)

The metric HEOM was introduced by Wilson and Martinez [27] and it is the combination of the Euclidean and Hamming metric. Basically HEOM is an heterogeneous distance function that uses different attributes distance functions on different kinds of attributes. If $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are two examples then $heom(x,y) = \sqrt{\sum_{j=1}^{n} d_j(x_j, y_j)^2}$ where $d_j(x_j, y_j)$ is the Hamming distance if the $j$-th feature is nominal and the Euclidean distance if numeric. The numeric features are normalized using the range.

### 2.4 Minimum Risk Metric

Minimum Risk Metric (MRM) is a very simple metric that directly minimizes the risk of misclassification.

Given an example $x$ in class $c_i$ and a nearest neighbor $y$ the finite risk of misclassifying $x$ is given by $p(c_i|x)(1 - p(c_i|y))$. The total finite risk is the sum of the risks extended to all the different classes and is given by $r(x,y) = \sum_{i=1}^{m} p(c_i|x)(1 - p(c_i|y))$. The approach of Short and Fukunaga and followers is to subtract the asymptotic risk $r^*(x,y)$ and minimize $E(r(x,y) - r^*(x,y))$. Instead we propose to minimize directly the risk $r(x,y)$ and that leads to the metric:

$$\mathrm{MRM}(x,y) = r(x,y) = \sum_{i=1}^{m} p(c_i|x)(1 - p(c_i|y)). \tag{7}$$

We observed in some experiments not shown here, that the application of MRM inside a Nearest Neighbor classifier leads to a classifier comparable to the Bayes rule, i.e., "assign $x$ to the class that maximizes $p(c_i|x)$". That points out that the key element in MRM is the estimation of $p(c_i|x)$. This point is dealt with in the next Section.

## 3 Probability Distribution Estimation

The presence of the conditional probabilities $p(c_i|x)$ in both the SF2 metric and MRM requires consistent estimates $\hat{p}(c_i|x)$ and this section illustrates the probability estimation techniques used in the experiments.

We must note that, a classification problem would be solved if the probabilities $p(c_i|x)$ were known. In fact the Bayes optimal classification rule says to choose the class $c_i$ that maximizes $p(c_i|x)$. All the classification methods explicitly or implicitly follow this rule and the estimation of $p(c_i|x)$ is not simpler than computing an optimal metric for NN. For that reason the estimation of the quantities $p(c_i|x)$ is a key issue. Notwithstanding that, we will show that even if many of the metrics here presented are based on the same estimation of the quantities $p(c_i|x)$ the definition of the metric is relevant and different performances can be obtained.

The estimates of $p(c_i|x)$ can be done directly or applying the Bayes theorem

$$p(c_i|x) = \frac{p(x|c_i)p(c_i)}{p(x)} = \frac{p(x|c_i)p(c_i)}{\sum_{k=1}^{|C|} p(x|c_k)p(c_k)} \tag{8}$$

therefore reducing to the problem of estimating $p(x|c_k)$.

In the present work we carried out experiments with two different estimators. The first is the Naive Bayes Estimator that is the estimator that is implicit in the Naive Bayes Classifier. It is a natural estimator for nominal feature and it can be extended to the numeric ones by discretization. The second is the Gaussian Kernel Estimator, a non–parametric density estimator that in its original formulation uses the Euclidean metric. In order to extend the density estimation technique to nominal features the Euclidean metric is simply substituted by HEOM and the densities are assumed to replace the probabilities in the expressions of the metrics.

### 3.1 Naive Bayes Estimator

The simplest probability estimates are based on frequency counts. In this way it is possible to estimate $p(c_i)$ with $\hat{p}(c_i) = \frac{N(c_i)}{N}$ where $N(c_i)$ is the number of cases that are in the $c_i$ class and $N$ is the sample size. Unfortunately, probability estimates based on frequencies perform poorly if the sample size is small (basically the probabilities are understimated) and so they can be improved adopting the Laplace–corrected estimate or equivalently incrementing artificially the sample size [17]. Following the first possibility leads to the estimate $\hat{p}(c_i) = \frac{N(c_i)+f}{N+fn_j}$ where $n_j$ is the number of values of the $j$-th attribute and $f = 1/N$ is a multiplicative factor [11].

Assuming features' independence and adopting the same notation introduced in 2.2, lead to the estimates:

$$\hat{p}(x|c_i) = \prod_{j=1}^{n} \hat{p}(x_j|c_i) = \prod_{j=1}^{n} \frac{N(x_j, c_i) + f}{N(c_i) + fn_j}$$

which, substituted in the equation (8) gives the estimates that are used in the Naive Bayes Classifier approach.

$$\hat{p}(c_i|x) = \frac{\prod_{j=1}^{n} \frac{N(x_j,c_i)+f}{N(c_i)+fn_j} \frac{N(c_i)+f}{N+fn_j}}{\sum_{k=1}^{|C|} \prod_{j=1}^{n} \frac{N(x_j,c_k)+f}{N(c_k)+fn_j} \frac{N(c_k)+f}{N+fn_j}} \tag{9}$$

### 3.2 Gaussian Kernel Estimator

The second type of estimates used in this paper belongs to a broad class of nonparametric density estimators represented by the multivariate fixed kernel [21]:

$$\hat{f}(x) = \frac{1}{N} \sum_{l=1}^{n} \frac{1}{h(x,x_l)^n} K\left(\frac{x - x_l}{h(x,x_l)}\right) \tag{10}$$

where $n$ is the dimension of the input space, $h$ is the bandwidth and $K(t)$ is the kernel function.

The bandwidth $h(x,x_l)$ can be constant on the input space or it can vary. In relation to the bandwidth $h$'s dependency on the probe point $x$ or on the sample point $x_l$, the estimator is called *balloon* or *sample point* respectively.

The Gaussian Kernel Estimator is an example of a *sample point* estimator with fixed bandwidth.

$$\hat{f}(x) = \frac{1}{N(2\pi)^{n/2}} \sum_{l=1}^{N} \frac{\sqrt{|W|}}{h^n} e^{-\frac{1}{2}\left(\frac{\| x - x_l \|_W}{h}\right)^2} \tag{11}$$

where $W$ is a positive definite diagonal matrix, and

$$\| x - x_l \|_W = \sqrt{(x - x_l) W (x - x_l)^T} = \sqrt{\sum_{j=1}^{n} w_{jj} (x_j - x_{lj})^2}$$

$$\sqrt{|W|} = \prod_{j=1}^{n} w_{jj}$$

$\| \cdot \|_W$ is an Euclidean weighted metric and $w_{jj} = \frac{1}{\hat{\sigma}_j}$ where $\hat{\sigma}_j$ is an estimate of the variance on the $j$–th dimension of the input space. In this case the optimal bandwidth is $h = \left(\frac{4}{n+2}\right)^{\frac{1}{n+4}} N^{-\frac{1}{n+4}}$.

## 4 Experimental Results

The metrics presented in Section 2 were tested on 27 databases taken from from the Machine Learning Databases Repository at UCI [16] and on two new databases (Derma and Opera). Derma contains data of images for the diagnosis of melanoma collected in Santa Chiara Hospital in Trento, Italy and Opera

contains the results of a cognitive pragmatics experiment [5]. The 29 databases contain continuous, nominal and mixed features. The main characteristics of the databases are presented in Table 1. We extended to mixed feature databases the estimate of the Naive Bayes Estimator by discretizing the numeric features and the estimate of the Gaussian Kernel Estimator by substituting the Euclidean Metric with HEOM. We normalized the numeric features with their range and used ten intervals for all the discretizations. The unknown values were simply ignored during the computation. The experimental technique is a 10-fold cross–validation and as a significance test we adopted the paired $t$-test ($p < 0.05$).

**Table 1.** The databases used in the experimentation.

| Data Set | Instances | Classes | Features | | Unknown |
|---|---|---|---|---|---|
| | | | Number | Cont/Symb | Attributes |
| Annealing | 798 | 6 | 38 | 9C 29S | yes |
| Audiology(standardized) | 200 | 24 | 69 | 69S | yes |
| Breast-cancer | 286 | 2 | 9 | 4C 5S | yes |
| Bridges | 108 | 6 | 11 | 9C 2S | yes |
| Bridges(discretized) | 108 | 6 | 11 | 11S | yes |
| Credit Screening | 690 | 2 | 15 | 6C 9S | yes |
| Derma | 152 | 2 | 44 | 44C | no |
| Flag | 194 | 8 | 28 | 10C 18S | no |
| Glass | 214 | 7 | 9 | 9C | no |
| Hepatitis | 155 | 2 | 19 | 6C 13S | yes |
| Horse-Colic | 300 | 2 | 27 | 7C 20S | yes |
| House-Votes-84 | 435 | 2 | 16 | 16S | yes |
| Ionosphere | 351 | 2 | 34 | 34C | no |
| Iris | 150 | 3 | 4 | 4C | no |
| Led+17noise | 200 | 10 | 24 | 24S | no |
| Led | 200 | 10 | 7 | 7S | no |
| Liver Disorders | 345 | 2 | 6 | 6C | no |
| Monks-1 | 432 | 2 | 6 | 6S | no |
| Monks-2 | 432 | 2 | 6 | 6S | no |
| Monks-3 | 432 | 2 | 6 | 6S | no |
| Opera | 1216 | 5 | 9 | 9S | no |
| Pima | 768 | 2 | 8 | 8C | no |
| Post-operative | 90 | 3 | 8 | 1C 7S | yes |
| Promoters | 106 | 2 | 57 | 57S | no |
| Sonar | 208 | 2 | 60 | 60C | no |
| Soybean(large) | 307 | 19 | 35 | 35S | yes |
| Soybean(small) | 47 | 4 | 35 | 35S | no |
| WDBC | 569 | 2 | 32 | 32C | no |
| zoo | 101 | 7 | 16 | 16S | no |

The experiments presented here measure the classification accuracies of the 1-NN algorithm with SF2 metric (Eq. (2)) and MRM (Eq. (7)) obtained using Naive Bayes Estimator (Eq. (9)) and the Gaussian Kernel Estimator (Eq. (11)). The accuracies are compared to those of DVDM (Eq. (6)) and HEOM (Section 2.3).

The application of SF2 can be restricted to $h$ neighbors with respect to the metric HEOM. This means that when searching for the SF2 nearest neighbor of an example $x$ only the set of $h$ HEOM neighbors of $x$ are considered (e.g. when the metrics are computed on the whole training set $h = N$ holds).

Some of the experiments are conducted adopting as $h$ the cross-validated value $h_{CV}$. In some cases, we also cross–validate the choice of the estimator. When this is the case the estimator is indicated as $Est_{CV}$. Both the cross–validations are carried out with a 10-fold cross–validation on each training partition.

### 4.1 HEOM and Value Difference Metrics results

**Table 2.** Classification accuracies for different metrics. Significant differences ($p < 0.05$) are shown: for instance, IVDM performs significatively better than DVDM on the sonar dataset.

| Data Set | IVDM ($I$) | HVDM ($H$) | DVDM ($D$) | HEOM ($E$) |
|---|---|---|---|---|
| annealing | $97.4 \pm 1.33 > E$ | $99.1 \pm 1.03 > I, E$ | $98.4 \pm 0.98 > I, E$ | $95.4 \pm 2.59$ |
| audiology | $80.5 \pm 7.24 > E$ | $80.5 \pm 5.98 > E$ | $80.5 \pm 5.98 > E$ | $72.5 \pm 11.3$ |
| breast-cancer | $66.4 \pm 6.92$ | $68.2 \pm 8.21$ | $64.3 \pm 10.0$ | $65.4 \pm 8.54$ |
| bridges1 | $61.1 \pm 7.97$ | $59.3 \pm 11.1$ | $62.3 \pm 16.9$ | $65.9 \pm 13.9 > H$ |
| bridges2 | $62.1 \pm 20.0$ | $59.3 \pm 19.0$ | $59.3 \pm 19.0$ | $55.5 \pm 17.2$ |
| crx | $79.7 \pm 2.36$ | $80.5 \pm 5.21$ | $79.5 \pm 4.06$ | $81.7 \pm 3.36$ |
| derma | $80.0 \pm 12.6$ | $73.0 \pm 12.6$ | $74.8 \pm 13.4$ | $78.1 \pm 10.6$ |
| flag | $57.4 \pm 12.3$ | $66.6 \pm 8.75 > I, E$ | $64.0 \pm 8.34 > I, E$ | $55.8 \pm 12.9$ |
| glass | $72.5 \pm 12.5 > D$ | $69.7 \pm 9.32 > D$ | $62.1 \pm 11.1$ | $71.1 \pm 11.8 > D$ |
| hepatitis | $82.6 \pm 10.1$ | $80.0 \pm 9.94$ | $82.0 \pm 10.8$ | $80.7 \pm 11.8$ |
| horse-colic | $85.6 \pm 5.67$ | $85.6 \pm 7.70$ | $86.6 \pm 7.53$ | $84.6 \pm 4.76$ |
| house-votes-84 | $93.7 \pm 3.10$ | $93.0 \pm 2.45$ | $93.0 \pm 2.45$ | $92.3 \pm 3.82$ |
| ionosphere | $87.4 \pm 3.38 > H$ | $35.9 \pm 4.75$ | $88.8 \pm 4.75 > H$ | $87.1 \pm 2.81 > H$ |
| iris | $94.6 \pm 5.25$ | $96.6 \pm 4.71$ | $92.6 \pm 4.91$ | $95.3 \pm 5.48$ |
| led | $66.5 \pm 13.5$ | $66.5 \pm 13.5$ | $66.5 \pm 13.5$ | $68.0 \pm 12.9$ |
| led17 | $57.5 \pm 12.5 > E$ | $59.5 \pm 11.8 > E$ | $59.5 \pm 11.8 > E$ | $39.0 \pm 9.06$ |
| liver | $63.9 \pm 8.07$ | $59.4 \pm 11.5$ | $64.3 \pm 8.22$ | $63.7 \pm 7.82$ |
| monks-1 | $78.0 \pm 13.4$ | $78.0 \pm 13.4$ | $78.0 \pm 13.4$ | $71.5 \pm 7.54$ |
| monks-2 | $92.6 \pm 8.39 > E$ | $92.6 \pm 8.39 > E$ | $92.6 \pm 8.39 > E$ | $57.1 \pm 7.21$ |
| monks-3 | $100. \pm 0.00 > E$ | $100. \pm 0.00 > E$ | $100. \pm 0.00 > E$ | $79.3 \pm 8.43$ |
| opera | $49.0 \pm 4.78$ | $49.0 \pm 4.78$ | $49.0 \pm 4.78$ | $49.0 \pm 4.84$ |
| pima-indians-diabetes | $70.5 \pm 4.47$ | $68.4 \pm 4.28$ | $70.8 \pm 3.31$ | $71.7 \pm 3.15 > H$ |
| post-operative | $63.3 \pm 14.8$ | $63.3 \pm 13.9$ | $62.2 \pm 14.9$ | $57.7 \pm 22.7$ |
| promoters | $89.7 \pm 10.1 > E$ | $89.7 \pm 8.17 > E$ | $89.7 \pm 8.17 > E$ | $80.1 \pm 9.42$ |
| sonar | $85.0 \pm 8.84 > D$ | $81.6 \pm 6.42$ | $76.9 \pm 6.15$ | $87.0 \pm 7.19 > D$ |
| soybean-large | $92.1 \pm 4.08$ | $90.2 \pm 5.80$ | $90.2 \pm 5.80$ | $91.1 \pm 5.13$ |
| soybean-small | $100. \pm 0.00$ | $100. \pm 0.00$ | $100. \pm 0.00$ | $100. \pm 0.00$ |
| wdbc | $95.2 \pm 2.19$ | $95.7 \pm 2.50$ | $94.9 \pm 3.02$ | $95.2 \pm 2.34$ |
| zoo | $95.0 \pm 7.00$ | $95.0 \pm 7.00$ | $95.0 \pm 7.00$ | $96.0 \pm 5.16$ |

In the first series of experiments we evaluate the metrics HEOM, DVDM, IVDM and HVDM. These metrics will represent a baseline for SF2 and MRM. Accuracy results are reported in Table 2. In this Table, when on a given dataset, a metric $m$ performs significantly better than another one $m'$, the symbol $> m'$ appear in the column of $m$. All the metrics of the VDM family seem to outperform the HEOM but there is not a clear winner among them. This results seem to partially contradict the observation by Wilson and Martinez [27] that a better aproximation of the probabilities $p(c_i|x_j)$ for numerical features, would lead to a better metric. Moreover, HVDM, the combined Euclidean and VDM metric,

performs well even though it simply sums the heterogeneous contributions of the two metrics.

IVDM is more sophisticated than DVDM. In IVDM the estimate of $p(c_i|x_j)$ obtained by discretizing the $j$-th numerical featue is smoothed by interpolation. But this approach seems not to improve DVDM to a great extent. For this reason, in the following experiments we compare SF2 and MRM only with DVDM.

### 4.2 Short and Fukunaga Metric

**Table 3.** Classification accuracies of the metrics SF2 with Naive , Kernel and cross-validated estimator with different localities. Significant differences ($p < 0.05$) are shown: for instance, Naive $h = h_{CV}$ performs significatively better than Kernel $h = h_{CV}$ on led17 dataset.

| Data Set | Naive $h = h_{CV}$ $(N)$ | Kernel $h = h_{CV}$ $(K)$ | $Est_{CV}$ $h = h_{CV}$ $(E^*_{CV})$ |
|---|---|---|---|
| annealing | $97.9 \pm 1.88$ | $97.9 \pm 1.47$ | $97.9 \pm 1.88$ |
| audiology | $77.5 \pm 8.57$ | $76.0 \pm 10.2$ | $75.5 \pm 10.1$ |
| breast-cancer | $63.5 \pm 9.48$ | $64.3 \pm 8.65$ | $62.8 \pm 9.85$ |
| bridges1 | $64.9 \pm 9.95$ | $60.9 \pm 12.7$ | $64.0 \pm 8.55$ |
| bridges2 | $71.4 \pm 19.2$ | $66.7 \pm 19.1$ | $70.5 \pm 20.8$ |
| crx | $80.4 \pm 2.57$ | $82.4 \pm 4.75$ | $80.8 \pm 2.88$ |
| derma | $78.1 \pm 13.1$ | $73.5 \pm 11.8$ | $77.5 \pm 14.2$ |
| flag | $59.8 \pm 10.1$ | $58.9 \pm 8.16$ | $59.8 \pm 10.1$ |
| glass | $69.2 \pm 11.9$ | $73.0 \pm 10.9 > E^*_{CV}$ | $68.7 \pm 11.4$ |
| hepatitis | $88.5 \pm 8.72$ | $83.9 \pm 7.93$ | $88.5 \pm 8.72$ |
| horse-colic | $83.6 \pm 6.74$ | $84.3 \pm 6.85$ | $83.3 \pm 7.20$ |
| house-votes-84 | $93.0 \pm 3.78$ | $93.9 \pm 3.67$ | $93.7 \pm 3.44$ |
| ionosphere | $86.5 \pm 3.35$ | $92.5 \pm 4.50 > N, E^*_{CV}$ | $89.4 \pm 4.28$ |
| iris | $95.3 \pm 5.48$ | $94.6 \pm 6.88$ | $95.3 \pm 5.48$ |
| led | $68.5 \pm 15.4$ | $71.5 \pm 17.9$ | $69.0 \pm 14.4$ |
| led17 | $58.0 \pm 8.23 > K$ | $43.5 \pm 7.83$ | $58.0 \pm 8.23 > K$ |
| liver | $66.6 \pm 10.7$ | $59.4 \pm 12.7$ | $62.3 \pm 12.5$ |
| monks-1 | $76.1 \pm 10.5$ | $100. \pm 0.00 > N$ | $98.1 \pm 3.40 > N$ |
| monks-2 | $91.1 \pm 7.33 > K$ | $56.4 \pm 7.68$ | $91.1 \pm 7.33 > K$ |
| monks-3 | $100. \pm 0.00$ | $99.7 \pm 0.73$ | $100. \pm 0.00$ |
| opera | $48.4 \pm 4.41$ | $48.5 \pm 4.78$ | $48.6 \pm 4.74$ |
| pima-indians-diabetes | $70.3 \pm 3.55$ | $70.3 \pm 3.17$ | $70.3 \pm 3.55$ |
| post-operative | $56.6 \pm 18.4$ | $53.3 \pm 17.9$ | $55.5 \pm 16.5$ |
| promoters | $88.5 \pm 7.81$ | $83.8 \pm 12.4$ | $88.5 \pm 7.81$ |
| sonar | $87.0 \pm 7.19$ | $89.3 \pm 5.57$ | $86.0 \pm 7.74$ |
| soybean-large | $92.4 \pm 4.94$ | $90.5 \pm 5.64$ | $92.4 \pm 4.94$ |
| soybean-small | $100. \pm 0.00$ | $100. \pm 0.00$ | $100. \pm 0.00$ |
| wdbc | $95.4 \pm 2.36$ | $96.3 \pm 2.53 > E^*_{CV}$ | $94.7 \pm 2.01$ |
| zoo | $96.0 \pm 5.16$ | $96.0 \pm 5.16$ | $96.0 \pm 5.16$ |

Preliminary results showed a substantial equivalence between SF2 and SFM and therefore we chose the simpler one. Table 3 presents the classification accuracies of SF2 metric with different estimators (Naive , Gaussian Kernel, and the cross–validated one). Moreover the grade of locality is also cross–validated. This means that in the computation of the SF2 nearest neighbor of an example $x$, the SF2 distance from this example is only taken with examples in a subset of the case base. This subset contains the $h$ nearest neighbors of $x$ with respect to the HEOM metric. In fact, the locality of the SF2 metrics appears to be critical. In

a set of results not showed here we noted that an unrestricted application of the metric leads to poor results when compared with DVDM and HEOM.

In Table 4 we show how the SF2 metric based on cross–validation outperforms significatively DVDM and HEOM. In particular the metric with both estimator and locality cross-validated is never worse of them and outperforms DVDM in 4 datasets and HEOM in 8 datasets. However, in a set of experiments not reported here we noted that SF2 often performs worse than the Bayes Classifier based on the same estimation.

**Table 4.** Classification accuracies of the SF2 with a cross-validated estimator, DVDM and HEOM. Significant differences ($p < 0.05$) are shown.

| Data Set | SF2 $Est_{CV}$ $h = h_{CV}$ $(E^*_{CV})$ | DVDM ($D$) | HEOM ($E$) |
|---|---|---|---|
| annealing | $97.9 \pm 1.88 > E$ | $98.4 \pm 0.98$ | $95.4 \pm 2.59$ |
| audiology | $75.5 \pm 10.1$ | $80.5 \pm 5.98$ | $72.5 \pm 11.3$ |
| breast-cancer | $62.8 \pm 9.85$ | $64.3 \pm 10.0$ | $65.4 \pm 8.54$ |
| bridges1 | $64.0 \pm 8.55$ | $62.3 \pm 16.9$ | $65.9 \pm 13.9$ |
| bridges2 | $70.5 \pm 20.8 > D, E$ | $59.3 \pm 19.0$ | $55.5 \pm 17.2$ |
| crx | $80.8 \pm 2.88$ | $79.5 \pm 4.06$ | $81.7 \pm 3.36$ |
| derma | $77.5 \pm 14.2$ | $74.8 \pm 13.4$ | $78.1 \pm 10.6$ |
| flag | $59.8 \pm 10.1$ | $64.0 \pm 8.34$ | $55.8 \pm 12.9$ |
| glass | $68.7 \pm 11.4 > D$ | $62.1 \pm 11.1$ | $71.1 \pm 11.8$ |
| hepatitis | $88.5 \pm 8.72 > E$ | $82.0 \pm 10.8$ | $80.7 \pm 11.8$ |
| horse-colic | $83.3 \pm 7.20$ | $86.6 \pm 7.53$ | $84.6 \pm 4.76$ |
| house-votes-84 | $93.7 \pm 3.44$ | $93.0 \pm 2.45$ | $92.3 \pm 3.82$ |
| ionosphere | $89.4 \pm 4.28$ | $88.8 \pm 4.75$ | $87.1 \pm 2.81$ |
| iris | $95.3 \pm 5.48$ | $92.6 \pm 4.91$ | $95.3 \pm 5.48$ |
| led | $69.0 \pm 14.4$ | $66.5 \pm 13.5$ | $68.0 \pm 12.9$ |
| led17 | $58.0 \pm 8.23 > E$ | $59.5 \pm 11.8$ | $39.0 \pm 9.06$ |
| liver | $62.3 \pm 12.5$ | $64.3 \pm 8.22$ | $63.7 \pm 7.82$ |
| monks-1 | $98.1 \pm 3.40 > D, E$ | $78.0 \pm 13.4$ | $71.5 \pm 7.54$ |
| monks-2 | $91.1 \pm 7.33 > E$ | $92.6 \pm 8.39$ | $57.1 \pm 7.21$ |
| monks-3 | $100. \pm 0.00 > E$ | $100. \pm 0.00$ | $79.3 \pm 8.43$ |
| opera | $48.6 \pm 4.74$ | $49.0 \pm 4.78$ | $49.0 \pm 4.84$ |
| pima-indians-diabetes | $70.3 \pm 3.55$ | $70.8 \pm 3.31$ | $71.7 \pm 3.15$ |
| post-operative | $55.5 \pm 16.5$ | $62.2 \pm 14.9$ | $57.7 \pm 22.7$ |
| promoters | $88.5 \pm 7.81 > E$ | $89.7 \pm 8.17$ | $80.1 \pm 9.42$ |
| sonar | $86.0 \pm 7.74 > D$ | $76.9 \pm 6.15$ | $87.0 \pm 7.19$ |
| soybean-large | $92.4 \pm 4.94$ | $90.2 \pm 5.80$ | $91.1 \pm 5.13$ |
| soybean-small | $100. \pm 0.00$ | $100. \pm 0.00$ | $100. \pm 0.00$ |
| wdbc | $94.7 \pm 2.01$ | $94.9 \pm 3.02$ | $95.2 \pm 2.34$ |
| zoo | $96.0 \pm 5.16$ | $95.0 \pm 7.00$ | $96.0 \pm 5.16$ |

### 4.3 Minimum Risk Metric

In this Section we evaluate the Minimum Risk Metric introduced in Section 2.4. In this case we used the Naive Bayes estimator, that in a set of experiments not shown here seems to work best for this metric. In Table 5 MRM is compared with the DVDM metric and HEOM metric. MRM compares very favourably with the exception of the monks datasets. These datasets appear to be a hard task probably as a consequence of the assumption of the independence among features that underlies the Naive Estimator. MRM outperforms DVDM and

**Table 5.** Classification accuracy of the Minimum Risk Metric with the Naive Estimator, DVDM and HEOM. Significant differences ($p < 0.05$) are shown.

| Data Set | MRM $h = N$ Naive ($M_N$) | DVDM | HEOM |
|---|---|---|---|
| annealing | $97.6 \pm 1.61 > E$ | $98.4 \pm 0.98$ | $95.4 \pm 2.59$ |
| audiology | $76.5 \pm 7.47$ | $80.5 \pm 5.98$ | $72.5 \pm 11.3$ |
| breast-cancer | $73.4 \pm 7.16 > D, E$ | $64.3 \pm 10.0$ | $65.4 \pm 8.54$ |
| bridges1 | $63.0 \pm 11.0$ | $62.3 \pm 16.9$ | $65.9 \pm 13.9$ |
| bridges2 | $69.6 \pm 19.0 > D, E$ | $59.3 \pm 19.0$ | $55.5 \pm 17.2$ |
| crx | $83.9 \pm 1.73 > D$ | $79.5 \pm 4.06$ | $81.7 \pm 3.36$ |
| derma | $77.4 \pm 17.9$ | $74.8 \pm 13.4$ | $78.1 \pm 10.6$ |
| flag | $61.8 \pm 7.83$ | $64.0 \pm 8.34$ | $55.8 \pm 12.9$ |
| glass | $66.8 \pm 13.6$ | $62.1 \pm 11.1$ | $71.1 \pm 11.8$ |
| hepatitis | $87.1 \pm 7.88$ | $82.0 \pm 10.8$ | $80.7 \pm 11.8$ |
| horse-colic | $83.6 \pm 7.44$ | $86.6 \pm 7.53$ | $84.6 \pm 4.76$ |
| house-votes-84 | $90.5 \pm 4.30$ | $93.0 \pm 2.45$ | $92.3 \pm 3.82$ |
| ionosphere | $91.1 \pm 3.42 > E$ | $88.8 \pm 4.75$ | $87.1 \pm 2.81$ |
| iris | $95.3 \pm 5.48$ | $92.6 \pm 4.91$ | $95.3 \pm 5.48$ |
| led | $72.5 \pm 14.5$ | $66.5 \pm 13.5$ | $68.0 \pm 12.9$ |
| led17 | $67.0 \pm 9.18 > D, E$ | $59.5 \pm 11.8$ | $39.0 \pm 9.06$ |
| liver | $71.3 \pm 9.85 > D, E$ | $64.3 \pm 8.22$ | $63.7 \pm 7.82$ |
| monks-1 | $66.2 \pm 15.0$ | $78.0 \pm 13.4 > M_N$ | $71.5 \pm 7.54$ |
| monks-2 | $67.1 \pm 7.49 > E$ | $92.6 \pm 8.39 > M_N$ | $57.1 \pm 7.21$ |
| monks-3 | $97.2 \pm 2.40 > E$ | $100. \pm 0.00 > M_N$ | $79.3 \pm 8.43$ |
| opera | $58.0 \pm 3.70 > D, E$ | $49.0 \pm 4.78$ | $49.0 \pm 4.84$ |
| pima-indians-diabetes | $75.1 \pm 4.76 > D, E$ | $70.8 \pm 3.31$ | $71.7 \pm 3.15$ |
| post-operative | $64.4 \pm 17.9 > E$ | $62.2 \pm 14.9$ | $57.7 \pm 22.7$ |
| promoters | $90.4 \pm 6.38 > E$ | $89.7 \pm 8.17$ | $80.1 \pm 9.42$ |
| sonar | $78.3 \pm 8.15$ | $76.9 \pm 6.15$ | $87.0 \pm 7.19 > M_N$ |
| soybean-large | $92.5 \pm 4.62$ | $90.2 \pm 5.80$ | $91.1 \pm 5.13$ |
| soybean-small | $100. \pm 0.00$ | $100. \pm 0.00$ | $100. \pm 0.00$ |
| wdbc | $93.8 \pm 2.22$ | $94.9 \pm 3.02$ | $95.2 \pm 2.34$ |
| zoo | $96.0 \pm 5.16$ | $95.0 \pm 7.00$ | $96.0 \pm 5.16$ |

HEOM more convincingly than SF2 and without any local restriction. This is obviously an important feature as it greatly simplifies the computation of the metric.

## 5 Conclusions

In this paper we have introduced two new metrics for nearest neighbor classification that are based on probability estimation. The first, SF2, was originally introduced by Short and Fukunaga [23]. We extended its definition to input spaces with nominal features and introduced a different estimate for the density probability used in this metric. The second, the Minimum Risk Metric (MRM) is very similar to SF2 but optimizes a different criterion, the risk of misclassification. Among the main advantages of these types of metrics is the possibility to manage both nominal and numerical features in an uniform way and the fact that these metrics can be analytically studied.

The experiments show that the metric SF2 works only if locally restricted, i.e., examples used for the SF2 nearest neighbor computation are taken from a set of Euclidean nearest neighbors. This is surprising given the theoretical optimality of the metric and further investigations are required to clarify this point. In fact,

in the original formulation of Short of Fukunaga the locality is not necessary for the optimality argument but only because they adopt a linear approximation of the probability. Nevertheless the combination of cross–validated locality and cross-validated estimator leads to a metric that outperforms VDM and HEOM.

The Minimum Risk Metric does not require any local restriction, its performances are comparable to the Bayes rule with the same probability estimates, its analytical form is simple and well founded, and finally, equipped with a simple Naive Estimator, it outperforms the other metrics. Even if the metric does not improve the performance of the Naive Bayes Classifier (i.e. a quite good classifier) the choice of MRM appears to be relevant whenever the retrieval of a neighbor is required. For this reasons MRM seems particularly suitable for Case Based Reasoning applications when a relevant classification of the cases is available.

Futher work is required to address in depth the relation between MRM and the Bayes rule, to explore the performances of the metrics not only in 1-NN but also in $k$-NN and their sensitivity to noisy data.

## 6 Acknowledgements

We would like to thank M. Cazzani for her contribution to the implementation of CBET, the C++ library used in the experimental evaluation of the metrics presented in this paper.

## References

1. D. W. Aha and R. L. Goldstone. Learning attribute relevance in context in instance-based learning algorithms. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pages 141–148, Cambridge, MA, 1990. Lawrence Earlbaum.
2. D. W. Aha and R. L. Goldstone. Concept learning and flexible weighting. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, pages 534–539, Bloomington, IN, 1992. Lawrence Earlbaum.
3. P. Avesani, A. Perini, and F. Ricci. Interactive case-based planning for forest fire management. *Applied Artificial Intelligence*, 1999. To appear.
4. R. Bellazzi, S. Montani, and L. Portinale. Retrieval in a prototype-based case library: A case study in diabetes therapy revision. In *European Workshop on Case Based Reasoning*, 1998.
5. E. Blanzieri, M. Bucciarelli, and P. Peretti. Modeling human communication. In *First European Workshop on Cognitive Modeling*, Berlin, 1996.
6. C. Cardie and N. Howe. Improving minority class prediction using case-specific feature weight. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pages 57–65. Morgan Kaufmann Publishers, 1997.
7. M. Cazzani. Metriche di similaritá eterogenee per il problema di recupero nei sistemi di ragionamento basato su casi: studio sperimentale. Master's thesis, Univ. of Milano, 1998.
8. S. Cost and S. Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10:57–78, 1993.

9. T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transaction on Information Theory*, 13:21–27, 1967.
10. R. H. Creecy, B. M. Masand, S. J. Smith, and D. L. Waltz. Trading MIPS and memory for knowledge engineering. *Communication of ACM*, 35:48–64, 1992.
11. P. Domingos and M. J. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.
12. J. H. Friedman. Flexible metric nearest neighbour classification. Technical report, Stanford University, 1994. Available by anonymous FTP from playfair.stanford.edu.
13. T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbour classification. In U.M.Fayad and R.Uthurusamy, editors, *KDD-95: Proceedings First International Conference on Knowledge Discovery and Data Mining*, 1995.
14. P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. Bayes optimal instacebased learning. *Lecture Notes in Computer Science*, 1398:77–88, 1998.
15. P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. On Bayesian case matching. *Lecture Notes in Computer Science*, 1488:13–24, 1998.
16. C. J. Merz and P. M. Murphy. *UCI Repository of Machine Learning Databases.* University of California, Department of Information and Computer Science, Irvine, CA, 1996.
17. T. M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.
18. J. P. Myles and D. J. Hand. The multi-class metric problem in nearest neighbour discrimination rules. *Pattern Recognition*, 23(11):1291–1297, 1990.
19. F. Ricci and P. Avesani. Learning a local similarity metric for case-based reasoning. In *International Conference on Case-Based Reasoning (ICCBR-95), Sesimbra, Portugal, Oct. 23-26*, 1995.
20. F. Ricci and P. Avesani. Data compression and local metrics for nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999. To appear.
21. D. W. Scott. *Multivariate Density Estimation: Theory , Practice, and Visualization.* John Wiley, New York, 1992.
22. R. D. Short and K. Fukunaga. A new nearest neighbour distance measure. In *Proceedings of the 5th IEEE International Conference on Patter Recognition*, pages 81–86, Miami beach, FL, 1980.
23. R. D. Short and K. Fukunaga. The optimal distance measure for nearest neighbour classification. *IEEE Transactions on Information Theory*, 27:622–627, 1981.
24. C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communication of ACM*, 29:1213–1229, 1986.
25. D. Wettschereck and T. G. Dietterich. An experimental comparison of the nearest neighbor and nearest hyperrectangle algorithms. *Machine Learning*, 19:5–28, 1995.
26. D. Wettschereck, T. Mohri, and D. W. Aha. A review and empirical comparison of feature weighting methods for a class of lazy learning algorithms. *AI Review Journal*, 11:273–314, 1997.
27. D. R. Wilson and T. R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 11:1–34, 1997.