A Minimum Risk Metric for Nearest Neighbor Classification

Enrico Blanzieri
ITC-IRST
via Sommarive, Povo 38050 Trento Italy
blanzier@itc.it

Francesco Ricci*
ITC-IRST
ricci@sodalia.it

Abstract

Nearest Neighbor is a well-known algorithm extensively studied by the Pattern Recognition and Machine Learning communities and widely exploited in Case Based Reasoning applications. The notion of metric is central to Nearest Neighbor's working and different feature weighting metrics have been proposed in order to increase its performance. In this work we present an original Probability Based Metric, i.e. a metric for classification tasks that relies on estimates of the posterior probabilities, called Minimum Risk Metric (MRM). MRM is optimal but it optimizes directly the finite misclassification risk whereas the Short and Fukunaga Metric minimize the difference between finite risk and asymptotic risk. An experimental comparison of MRM with Short and Fukunaga Metric, Value Difference Metric, and Euclidean–Hamming metrics on benchmark datasets shows that MRM outperforms the other metrics and performs comparably to the Bayes Classifier based on the same probability estimates. The results suggests that MRM can be useful in applications were the retrieval of a nearest neighbor is required (e.g. Case Based Reasoning).

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 in the retrieval phase of almost the totality of the Case Based Reasoning (CBR) systems. Such systems emphasize more the retrieving of a neighbor than the classification accuracy but, as was recently shown by Bellazzi et al. [BMP98], the performance of a CBR system can be improved by driving the retrieval with the information of same relevant classification in the case space, i.e. reducing the retrieval problem to a classification task. In this prospective the problem of improving the classification accuracy for NN algorithms appears to be relevant.

The NN classification procedure is straightforward: Given a set of classified points in an input space, a new point is assigned to the known class of the nearest one with respect to a metric defined on the input space. Many researchers [SF80, SF81, SW86, CMSW92, AG90, AG92, HT95, Fri94, RA95, RA98, CH97, WMA97] 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, in order to outperform systems based on global metrics. The claim appears to be controversial. On one hand the 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 description of the separating surfaces.

Some of the proposed local metrics rely their effectiveness on the optimization of a given criterion and ultimately on the estimation of some probabilities. In this direction Short and Fukunaga [SF81] presented an initial work in the case of continuous multidimen-

* Present Address:Sodalia, via Zambra 1, 38100 Trento Italy

sional 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. In order to achieve this goal they expressed the optimal local metric in terms of a linear estimation of posterior probabilities. More recently, and in the instance based learning context, many proposals of nominal feature metrics also involve probability estimation [SW86, CMSW92, CS93, WD95]. In these cases the estimation is performed computing frequencies of value occurrences. Finally, in the work by Wilson and Martinez [WM97] the estimation of probabilities provides an unifying framework for treating both linear (continuous or discrete) and nominal features. Their heterogeneous distances take advantage of the contemporary elaboration of the two kind of features.

Among the metrics based on probability estimation the one proposed by Short and Fukunaga has the strongest theoretical foundation. In their original proposal it was used only for numeric features problems but it is possible to extend it to nominal or heterogeneous features by considering different and more general probability estimators. In fact, in spite of the centrality of the probability estimation in such type of metrics, little or no attempt has been made to exploit the advanced nonparametric density estimation techniques developed by the applied statistics community [Sco92] and their possible extensions to nominal features.

From our point of view the approach of constructing metrics via combination of well-known probability estimators and optimal metrics presents several advantages.

The resulting 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.

They can be computed using standard density estimation techniques. The issue of the choice of the right degree of locality (bandwidth selection) can rely on the solutions proposed for the choice of the bandwidth in the nonparametric density estimation models.

Finally, they can be defined on uniformly data sets with both numeric and nominal attributes. In real–world databases both continuous and nominal features can be found at the same time. The problem can be tackled in different ways:

- Ordering. Ordering and numbering the values of the nominal features and applying a continuous metric like the Euclidean one. In general this approach introduces fictitious neighborhood.
- Discretization. Discretizing the numeric values and applying a nominal metric, e.g. Hamming or Value Difference Metric [SW86]. With the discretization some information is inevitably lost and parameters of the discretization can be critical.
- Combination. Combining two metrics, a nominal and a numeric, obtaining an heterogeneous metric, e.g. Euclidean and Hamming. The heterogeneous metric is hard to adapt in a consistent way and performs poorly, as Wilson and Martinez have shown [WM97].

The metric based on probability estimation provides a natural unifying framework for dealing with both the kind of features. While the techniques of ordering, discretization and combination listed above can be used for probabilities estimation the optimality of the metric still relies on theoretical arguments.

Our initial hypothesis was that better performances should be achieved by combining Short and Fukunaga metric with more sophisticated and powerful probability estimators. This combination produced unexpected poor experimental results and outperforming of the more standard metrics were obtained only by explicit restricting the scope of application (locality) of the metric or cross–validating the estimator. This led us to a deeper analysis of the optimality condition underlying Short and Fukunaga metric and eventually to the definition of an alternative metric.

In this work we propose a Minimum Risk Metric (MRM) that relies is effectiveness on a different and simpler optimality condition than the one 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. Moreover the accuracies of both these probability based metrics are compared to the results of the Bayes Classifiers (i.e. the classifier that assign to the example to the class with highest probability) based on the same probability estimates.

The work is organized as follows: section (2) describes Short and Fukunaga approach and presents other metrics studied in the work, 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, sections (5) draws conclusions and future directions.

2 Metrics

In this section we will briefly present the four families of metrics studied in this work.

2.1 Short and Fukunaga Metric (SF2)

Short and Fukunaga [SF81] were 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 $r(x,y) = p(c_1|x)p(c_2|y) + p(c_2|x)p(c_1|y)$ be 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 at y) and $r^*(x) = 2p(c_1|x)p(c_2|x)$ (Cover and Hart[CH67]) be 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). 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:

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

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 [MH90] generalize the metric to a multiclass problem and introduce the following two:

$$sf2(x,y) = \sum_{i=1}^{m} |p(c_i|x) - p(c_i|y)| \tag{2}$$

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

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 the VDM introduced by Stanfill and Waltz [SW86] 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 the VDM metric is defined as follow:

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

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 frequencies VDM can also be written in the following form:

$$vdm(x,y) = \sum_{j=1}^{n} \left(\sqrt{\sum_{i=1}^{m} (p^2(c_i|x_j)} \sum_{i=1}^{m} (p(c_i|x_j) - p(c_i|y_j))^2 \right)$$

VDM has no clear justification and seems to assume attributes 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 [CMSW92, CS93, WM97], works quite well on real data sets. Moreover, Wilson and Martinez extended VDM to instances with numeric attributes [WM97]. They essentially discretize the numeric attributes (DVDM) and then smooth the histogram estimation of $p(c_i|x_j)$ by averaging (IVDM). They also suggest an Heterogeneous VDM that combines an Euclidean metric for numeric features with a VDM (HVDM). The version of VDM we adopted in our experiments is the version without weighting factors and with the absolute values:

$$vdm(x,y) = \sum_{j=1}^{n} \sum_{i=1}^{m} |p(c_i|x_j) - p(c_i|y_j)| \tag{4}$$

2.3 Heterogeneous Euclidean–Overlap Metric (HEOM)

HEOM introduced by Wilson and Martinez [WM97], is the combination of the Euclidean metric and the

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 a point x of 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 minimizing $E(r(x, y) - r^*(x, y))$. Instead we propose to minimize directly the risk $r(x, y)$ and that leads to a straightforward metric:

$$mrm(x, y) = r(x, y) = \sum_{i=1}^{m} p(c_i|x)(1 - p(c_i|y)). \quad (5)$$

3 Probability Distribution Estimation

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

$$p(c_i|x) = \frac{p(x|c_i)p(c_i)}{p(x)} = \frac{p(x|c_i)p(c_i)}{\sum_{k=1}^{|C|} p(x|c_k)p(c_k)} \quad (6)$$

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

In the present work we carried on experiments with two different estimators. One is the Naïve Bayes Estimator that is the estimator that is implicit in the Naïve Bayes Classifier. It is a natural estimator for nominal feature and it can be extended to the numeric ones by discretization. The others one the Gaussian Kernel Estimator is a non–parametric density estimator and, in its original formulation, uses a Euclidean metric. In order to extend the density estimation technique to nominal features the Euclidean metric is simply substituted by HEOM and the density are supposed to replace the probability in the expression of the metrics.

3.1 Naïve Bayes Estimator

The simplest probability estimates are the occurrence frequencies. In this way 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 performs poorly if the sample size is small (basically the probabilities result understimated) and so they can be improved adopting the Laplace–corrected estimate or equivalently incrementing artificially the sample size [Mit97]. Following the first possibility leads to the estimate $\hat{p}(x_j|c_i) = \frac{N(c_i)+f}{N+fn_j}$ where n_j is the number of values of the j-th attribute and $f = 1/N$ is a multiplicative factor [DP97].

Asserting the feature independence hypothesis leads 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 (6) are the estimates that are used in the Naïve Bayes Classifier approach:

$$\hat{p}(c_i|x) = \frac{\prod_{j=1}^{n} \frac{N(x_j,c_i)+f}{N(c_i)+fn_j} \frac{N(c_i)}{N}}{\sum_{k=1}^{|C|} \prod_{j=1}^{n} \frac{N(x_j,c_k)+f}{N(c_k)+fn_j} \frac{N(c_k)}{N}} \quad (7)$$

3.2 Gaussian Kernel Estimator

A broad class of nonparametric density estimators is represented by the multivariate fixed kernel estimator [Sco92]:

$$\hat{f}(\bar{x}) = \frac{1}{N} \sum_{l=1}^{n} \frac{1}{h(\bar{x}, \bar{x}_l)^n} K\left(\frac{\bar{x} - \bar{x}_l}{h(\bar{x}, \bar{x}_l)}\right) \quad (8)$$

where n is the dimension of the input space, h is the bandwidth and $K(t)$ is the kernel function. The bandwidth $h(\bar{x}, \bar{x}_l)$ can be fixed overall the input space or it can vary. Depending on whether the bandwidth h depends on the probe point $\bar{x}$ or on the sample point $\bar{x}_l$, the estimators can be divided in two wide groups: *balloon* estimators and *sample point* estimators.

The Gaussian Kernel Estimator is an example of *sam-*

ple point estimator with fixed bandwidth.

$$\hat{f}(\bar{x}) = \frac{1}{N(2\pi)^{n/2}} \sum_{l=1}^{N} \frac{\sqrt{|W|}}{h^n} e^{-\frac{1}{2}\left(\frac{\|\bar{x}-\bar{x}_l\|_W}{h}\right)^2} \qquad (9)$$

where W is a positive definite diagonal matrix, and

$$\begin{aligned} \| \bar{x} - \bar{x}_l \|_W &= \sqrt{(\bar{x} - \bar{x}_l) W (\bar{x} - \bar{x}_l)^T} = \\ &= \sqrt{\sum_{j=1}^{n} w_{jj} (\bar{x} - \bar{x}_i)^2} \end{aligned}$$

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

therefore $\| \, . \, \|_W$ is an Euclidean weighted metric and $w_{jj} = \frac{1}{\hat{\sigma}_j}$ with j ranging over the dimensions of the input space. In this case the optimal bandwidth is $h = \left(\frac{4}{n+2}\right)^{\frac{1}{n+4}} N^{-\frac{1}{n+4}}$.

The estimate $\hat{f}(\bar{x}|c_k)$ takes into account only the samples that belong to class c_k.

4 Experimental Results

The metrics were tested on 28 databases downloaded from the Machine Learning Databases Repository at UCI and on two new databases (Derma and Opera). Derma collects data of images for the diagnosis of skin cancer collected at Santa Chiara Hospital in Trento, Italy and Opera contains the results of a cognitive pragmatics experiment [BB96]. The results on one of the UCI database (Imports-85) was afterwards excluded from the experimentation for a wrong choice of the target. The 29 databases contains continuous, nominal and mixed features. The main characteristics of the databases are presented in tab. 1. We extended to mixed feature databases the estimate of the Naïve 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 use 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 ($p < 0.05$).

The central experiments measure the classification accuracies of of the 1-NN algorithm with SF2 metric (Eq. 2) and MRM (Eq. 5) obtained using Naïve Bayes Estimator (Eq. 7) and the Gaussian Kernel Estimator (Eq. 9). The accuracies are compared to the ones of DVDM (Eq. 4), HEOM (Sec. 2.3) and IVDM (Sec. 2.2) and of the Bayes Classifier based on the same estimators. The application of SF2 can be restricted to h neighbors with respect to the metric HEOM. When the metrics are computed on the whole training set $h = N$ holds. Some of the experiments are led adopting as h the cross-validated value h_{CV}. We also cross–validate the used estimator. When this is the case the estimator is indicated as Est_{CV}. Both the cross–validations are carried on with a 10-fold cross–validation on each training partition.

Table 1: The databases used in the experimentation.

Data Set	Instances	Classes	Features		Unknown
			Number	Cont./Symb.	
Annealing	798	6	38	9C 29S	yes
Audiology(stan.)	200	24	69	69S	yes
Breast-cancer	286	2	9	4C 5S	yes
Bridges	108	6	11	9C 2S	yes
Bridges(dis.)	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

4.1 HEOM and Value Difference Metrics results

A first series of experiments refers to the metrics HEOM, DVDM, IVDM and HVDM. Accuracy results are reported in Tab. 2. Tab. 3 reports the significative differences information that is not present in the previous table. The reading of the Tab. 3 is as follows: When, on a given dataset, a metric performs significantly better than another one, the symbol of the latter compares in the column of the former. All the metrics of the VDM family appears to outperform

the HEOM but they are not clearly one better of the other. This results seems to partially contradict what observed by Wilson and Martinez.

Table 3: Significative differences ($p < 0.05$) of the Tab. 2. Example of reading: IVDM performs significatively better than DVDM on sonar dataset.

	VDMs			
Data Set	IVDM (*I*)	HVDM (*H*)	DVDM (*D*)	HEOM (*E*)
anneal	*E*	*E*	*I*	
audiology	*E*	*E*		
breast-cancer				
bridges1				*H*
bridges2				
crx				
derma				
flag		*E*	*I*	
glass	*D*	*D*		
hepatitis				
horse-colic				
house-votes-84			*H*	*H*
ionosphere				
iris				
led				
led17	*E*	*E*		
liver				
monks-1				
monks-2	*E*	*E*		
monks-3	*E*	*E*		
opera				
pima-indians-diabetes				*H*
post-operative				
promoters	*E*	*E*		
sonar	*D*			
soybean-large				
soybean-small				
wdbc				
zoo				

4.2 Short and Fukunaga Metric

Preliminary results showed a substantial equivalence between SF2 and SFM and therefore we choose the simpler one. The Tab. 4 presents the classification accuracies of SF2 metric with different estimators (Naïve , Gaussian Kernel, and the cross–validated one) with different grades of locality. The application of the SF2 is restricted to the h nearest neighbors with respect to the HEOM metric. The problem of the locality of the SF2 metrics appears to be critical. In fact, Tab. 5 shows how an unrestricted application of the metric leads to very poor results when compared with DVDM and HEOM. The Tab. 4 shows a sharp increasing of the performances when cross–validation of the estimator, of the locality, or of both of them are performed. Finally Tab. 6 shows 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. Experiments not reported here have shown that SF2 often performs worse of the Bayes Classifier based on the same estimation.

Table 5: Significative differences ($p < 0.05$) between SF2 metrics with Naïve and Kernel estimators and DVDM and HEOM.

	SF2 $h = N$			
Data Set	Naïve (*N*)	Kernel (*K*)	DVDM (*D*)	HEOM (*E*)
anneal	*E*			
audiology				
breast-cancer				
bridges1				
bridges2	*D E*			
crx				*N*
derma				
flag	*E*		*K*	
glass				*K N*
hepatitis				
horse-colic			*N*	
house-votes-84			*N*	*N*
ionosphere		*D E*		
iris				
led				
led17	*E*			
liver			*N*	*N*
monks-1		*D E*	*N*	
monks-2	*E*		*K N*	*K*
monks-3				*K*
opera				
pima-indians-diabetes			*K*	*K*
post-operative			*N*	
promoters	*E*			
sonar		*D*		*N*
soybean-large			*K*	
soybean-small				
wdbc			*N*	*N*
zoo				

4.3 Minimum Risk Metric

The Tab. 7 shows classification accuracies achieved by the MRM and by the Bayes Classifier based on the same probability estimate. The results are almost coincident and that does not surprise given the definition of the Bayes Classifier. The significative differences between the best ones and the metric DVDM and HEOM are shown in Tab. 8. The monks datasets appear to be a hard task probably as a consequence of the assumption of the independence among features that underlies the Naïve Estimator. The MRM outperforms DVDM and HEOM more convincingly than SF2 and without any local restriction (compare with Tab. 6). Finally, Tab. 9 reports the significative differences between the best of all the metrics presented. The MRM based on Naïve estimator does not differ from its version with the cross–validated estimator. Both of them, with the exception of monks datasets, outperforms SF2 with cross-validated estimator and cross-validated locality and the IVDM.

Table 2: Classification accuracies in percentage for different metrics. In bold the significative differences ($p < 0.05$) within the couples IVDM/HVDM and DVDM/HEOM.

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

Table 4: Classification accuracies in percentage of the metrics SF2 with Naïve , Kernel and cross-validated estimator with different localities. Bold means a significative ($p < 0.05$) difference between the locally restricted and unrestricted metrics.

	Naïve		Kernel		Est_{CV}	
Data Set	$h = N$	$h = h_{CV}$	$h = N$	$h = h_{CV}$	$h = N$	$h = h_{CV}$
anneal	97.3 ± 2.17	97.9 ± 1.88	96.2 ± 2.88	**97.9 ± 1.47**	97.3 ± 2.17	97.9 ± 1.88
audiology	76.5 ± 6.25	77.5 ± 8.57	76.0 ± 8.43	76.0 ± 10.2	75.5 ± 7.24	75.5 ± 10.1
breast-cancer	60.4 ± 11.7	63.5 ± 9.48	**68.9 ± 5.11**	64.3 ± 8.65	65.7 ± 6.44	62.8 ± 9.85
bridges1	65.8 ± 13.4	64.9 ± 9.95	64.9 ± 9.95	60.9 ± 12.7	64.9 ± 14.5	64.0 ± 8.55
bridges2	70.4 ± 17.9	71.4 ± 19.2	66.6 ± 17.2	66.7 ± 19.1	69.5 ± 19.5	70.5 ± 20.8
crx	78.2 ± 4.32	80.4 ± 2.57	82.3 ± 2.53	82.4 ± 4.75	81.4 ± 4.03	80.8 ± 2.88
derma	74.1 ± 17.4	78.1 ± 13.1	74.8 ± 11.3	73.5 ± 11.8	72.1 ± 15.7	77.5 ± 14.2
flag	63.8 ± 6.58	59.8 ± 10.1	50.0 ± 9.75	**58.9 ± 8.16**	63.8 ± 6.58	59.8 ± 10.1
glass	61.7 ± 14.1	**69.2 ± 11.9**	58.8 ± 14.2	**73.0 ± 10.9**	56.0 ± 11.9	**68.7 ± 11.4**
hepatitis	82.6 ± 6.43	88.5 ± 8.72	82.6 ± 9.27	83.9 ± 7.93	82.0 ± 6.31	**88.5 ± 8.72**
horse-colic	79.3 ± 9.78	**83.6 ± 6.74**	82.6 ± 8.86	84.3 ± 6.85	79.6 ± 11.3	83.3 ± 7.20
house-votes-84	86.4 ± 3.66	**93.0 ± 3.78**	90.5 ± 6.44	93.9 ± 3.67	90.5 ± 6.44	93.7 ± 3.44
ionosphere	87.7 ± 6.13	86.5 ± 3.35	93.4 ± 3.29	92.5 ± 4.50	**93.4 ± 3.29**	89.4 ± 4.28
iris	93.3 ± 5.44	95.3 ± 5.48	92.0 ± 5.25	94.6 ± 6.88	92.0 ± 5.25	95.3 ± 5.48
led	68.5 ± 15.4	68.5 ± 15.4	68.5 ± 15.9	71.5 ± 17.9	68.5 ± 15.9	69.0 ± 14.4
led17	58.5 ± 10.8	58.0 ± 8.23	48.0 ± 13.3	43.5 ± 7.83	58.5 ± 10.8	58.0 ± 8.23
liver	63.4 ± 8.17	66.6 ± 10.7	55.6 ± 10.1	59.4 ± 12.7	62.2 ± 7.36	62.3 ± 12.5
monks-1	65.9 ± 11.4	76.1 ± 10.5	100. ± 0.00	100. ± 0.00	100. ± 0.00	98.1 ± 3.40
monks-2	68.5 ± 11.3	**91.1 ± 7.33**	46.7 ± 5.16	**56.4 ± 7.68**	68.5 ± 11.3	**91.1 ± 7.33**
monks-3	99.7 ± 0.73	100. ± 0.00	99.7 ± 0.73	99.7 ± 0.73	99.7 ± 0.73	100. ± 0.00
opera	49.2 ± 4.63	48.4 ± 4.41	48.9 ± 4.81	48.5 ± 4.78	49.0 ± 4.90	48.6 ± 4.74
pima-indians-diabetes	70.5 ± 4.74	70.3 ± 3.55	65.2 ± 4.66	**70.3 ± 3.17**	67.1 ± 5.94	70.3 ± 3.55
post-operative	47.7 ± 17.4	56.6 ± 18.4	53.3 ± 19.4	53.3 ± 17.9	47.7 ± 17.4	55.5 ± 16.5
promoters	90.3 ± 9.23	88.5 ± 7.81	85.6 ± 11.9	83.8 ± 12.4	89.4 ± 8.60	88.5 ± 7.81
sonar	77.9 ± 7.34	**87.0 ± 7.19**	86.0 ± 7.34	89.3 ± 5.57	86.0 ± 7.34	86.0 ± 7.74
soybean-large	88.2 ± 7.60	92.4 ± 4.94	87.5 ± 5.76	90.5 ± 5.64	88.2 ± 7.60	92.4 ± 4.94
soybean-small	100. ± 0.00	100. ± 0.00	100. ± 0.00	100. ± 0.00	100. ± 0.00	100. ± 0.00
wdbc	92.2 ± 2.35	**95.4 ± 2.36**	94.0 ± 2.21	**96.3 ± 2.53**	94.0 ± 2.21	94.7 ± 2.01
zoo	96.0 ± 5.16	96.0 ± 5.16	97.0 ± 4.83	96.0 ± 5.16	96.0 ± 5.16	96.0 ± 5.16

Table 7: Classification accuracies in percentage of Minimum Risk Metric with different estimators and $h = N$ and Bayes Classifier based on the same estimates.

Data Set	Naïve Bayes	MR Naïve	Kernel Bayes	MR Kernel	Est_{CV} Bayes	MR Est_{CV}
anneal	97.6 ± 1.61	97.6 ± 1.61	76.4 ± 3.96	76.4 ± 3.96	97.6 ± 1.61	97.6 ± 1.61
audiology	76.5 ± 7.47	76.5 ± 7.47	49.0 ± 8.09	47.0 ± 9.77	76.5 ± 7.47	76.5 ± 7.47
breast-cancer	73.4 ± 7.16	73.4 ± 7.16	72.7 ± 6.02	72.7 ± 6.02	72.3 ± 6.58	72.3 ± 6.58
bridges1	63.0 ± 11.0	63.0 ± 11.0	52.9 ± 12.6	54.6 ± 14.4	63.0 ± 11.0	63.0 ± 11.0
bridges2	69.6 ± 19.0	69.6 ± 19.0	71.1 ± 15.3	71.1 ± 15.3	69.3 ± 17.8	69.3 ± 17.8
crx	83.9 ± 1.73	83.9 ± 1.73	62.0 ± 6.02	62.0 ± 6.02	83.9 ± 1.73	83.9 ± 1.73
derma	77.4 ± 17.9	77.4 ± 17.9	75.5 ± 15.8	75.5 ± 15.8	70.8 ± 16.8	70.8 ± 16.8
flag	61.8 ± 7.83	61.8 ± 7.83	55.3 ± 11.9	55.8 ± 11.7	61.8 ± 7.83	61.8 ± 7.83
glass	66.8 ± 13.6	66.8 ± 13.6	49.0 ± 13.6	49.5 ± 14.5	66.8 ± 13.6	66.8 ± 13.6
hepatitis	87.1 ± 7.88	87.1 ± 7.88	81.4 ± 8.62	81.4 ± 8.62	87.1 ± 7.88	87.1 ± 7.88
horse-colic	83.6 ± 7.44	83.6 ± 7.44	84.0 ± 7.16	84.0 ± 7.16	83.6 ± 7.44	83.6 ± 7.44
house-votes-84	90.5 ± 4.30	90.5 ± 4.30	92.1 ± 4.47	92.1 ± 4.47	91.4 ± 5.21	91.4 ± 5.21
ionosphere	91.1 ± 3.42	91.1 ± 3.42	71.5 ± 5.04	71.5 ± 5.04	91.1 ± 3.42	91.1 ± 3.42
iris	95.3 ± 5.48	95.3 ± 5.48	95.3 ± 4.49	95.3 ± 4.49	94.6 ± 5.25	94.6 ± 5.25
led	**78.5 ± 11.5**	72.5 ± 14.5	**78.0 ± 13.1**	68.0 ± 15.3	**77.0 ± 13.3**	72.5 ± 14.5
led17	67.0 ± 9.18	67.0 ± 9.18	41.0 ± 9.06	42.0 ± 9.18	67.0 ± 9.18	67.0 ± 9.18
liver	71.3 ± 9.85	71.3 ± 9.85	60.5 ± 9.07	60.5 ± 9.07	71.3 ± 9.85	71.3 ± 9.85
monks-1	75.0 ± 7.79	66.2 ± 15.0	97.9 ± 2.54	97.9 ± 2.54	97.9 ± 2.54	97.9 ± 2.54
monks-2	66.8 ± 7.51	67.1 ± 7.49	67.1 ± 7.49	67.1 ± 7.49	67.1 ± 7.49	67.1 ± 7.49
monks-3	97.2 ± 2.40	97.2 ± 2.40	98.1 ± 2.40	98.1 ± 2.40	98.1 ± 2.40	98.1 ± 2.40
opera	57.3 ± 4.19	58.0 ± 3.70	**56.7 ± 3.52**	56.1 ± 3.32	57.8 ± 4.16	58.0 ± 3.70
pima-indians-diabetes	75.1 ± 4.76	75.1 ± 4.76	68.6 ± 3.19	68.6 ± 3.19	75.1 ± 4.76	75.1 ± 4.76
post-operative	64.4 ± 17.9	64.4 ± 17.9	71.1 ± 19.0	71.1 ± 19.0	71.1 ± 19.0	71.1 ± 19.0
promoters	90.4 ± 6.38	90.4 ± 6.38	85.6 ± 11.9	85.6 ± 11.9	90.4 ± 6.38	90.4 ± 6.38
sonar	78.3 ± 8.15	78.3 ± 8.15	67.3 ± 7.27	67.3 ± 7.27	78.3 ± 8.15	78.3 ± 8.15
soybean-large	92.5 ± 4.62	92.5 ± 4.62	79.4 ± 6.10	79.4 ± 6.10	92.5 ± 4.62	92.5 ± 4.62
soybean-small	100. ± 0.00	100. ± 0.00	100. ± 0.00	100. ± 0.00	100. ± 0.00	100. ± 0.00
wdbc	93.8 ± 2.22	93.8 ± 2.22	75.9 ± 6.63	75.9 ± 6.63	93.8 ± 2.22	93.8 ± 2.22
zoo	96.0 ± 5.16	96.0 ± 5.16	92.0 ± 7.86	90.0 ± 8.17	96.0 ± 5.16	96.0 ± 5.16

Table 9: Significative differences ($p < 0.05$) between MRM with Naïve Estimator, MRM with cross–validated estimator, SF2 metric with cross-validated estimator and cross-validated locality and IVDM.

	MRM $h = N$		SF2 $h = h_{CV}$	IVDM (I)
Data Set	Naïve (M_N)	Est_{CV} (M_{CV})	Est_{CV} (E^*_{CV})	
anneal				
audiology				
breast-cancer	E^*_{CV} I	E^*_{CV} I		
bridges1				
bridges2				
crx	E^*_{CV} I	E^*_{CV} I		
derma	M_{CV}			
flag				
glass				
hepatitis				
horse-colic				
house-votes-84				
ionosphere				
iris				
led				
led17	E^*_{CV} I	E^*_{CV} I		
liver	E^*_{CV}	E^*_{CV}		
monks-1		M_N I	M_N I	M_N
monks-2			$M_N M_{CV}$	$M_N M_{CV}$
monks-3			$M_N M_{CV}$	$M_N M_{CV}$
opera	E^*_{CV} I	E^*_{CV} I		
pima-indians-diabetes	E^*_{CV} I	E^*_{CV} I		
post-operative		E^*_{CV}		
promoters				
sonar			M_N M_{CV}	
soybean-large				
soybean-small				
wdbc				
zoo				

Table 6: Significative differences ($p < 0.05$) between SF2 with cross-validated estimators and DVDM and HEOM.

	SF2 Est_{CV}			
Data Set	$h = N$ (E_{CV})	$h = h_{CV}$ (E^*_{CV})	DVDM (D)	HEOM (E)
anneal	E	E		
audiology				
breast-cancer				
bridges1				
bridges2	E	D E		
crx				
derma				
flag	E			
glass		D		E_{CV}
hepatitis		E		
horse-colic			E_{CV}	
house-votes-84				
ionosphere	D E			
iris				
led				
led17	E	E		
liver				
monks-1	D E	D E		
monks-2	E	E	E_{CV}	
monks-3	E	E		
opera				
pima-indians-diabetes				E_{CV}
post-operative	D			
promoters	E	E		
sonar	D	D		
soybean-large				
soybean-small				
wdbc				
zoo				

Table 8: Significative differences ($p < 0.05$) between Minimum Risk Metric with Naïve Estimator and cross–validated estimator with DVDM and HEOM.

	MRM $h = N$			
Data Set	Naïve (M_N)	Est_{CV} (M_{CV})	DVDM (D)	HEOM (E)
anneal	E	E		
audiology				
breast-cancer	D E	D E		
bridges1				
bridges2	D E	E		
crx	D	D		
derma				
flag				
glass		D		
hepatitis		E		
horse-colic				
house-votes-84				
ionosphere	E	E		
iris				
led				
led17	D E	D E		
liver	D E	D E		
monks-1		D E	M_N	
monks-2	E	E	M_N M_{CV}	
monks-3	E	E	M_N M_{CV}	
opera	D E	D E		
pima-indians-diabetes	D E	D E		
post-operative	E	E		
promoters	E	E		
sonar		D		M_N M_{CV}
soybean-large				
soybean-small				
wdbc				
zoo				

5 Conclusions

The experiments show that the metric based on the observation by Short and Fukunaga works only if locally restricted. That 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 resulting metric is computationally heavy and it is outperformed by the Bayes Classifier based on the same probability estimate.

The Minimum Risk Metric that we introduced in this paper does not require local restrictions, its performances are comparable to the Bayes Classifier, its analytical form is simple and well founded, and finally, equipped with a simple Naïve Estimator, outperforms the other metrics. The accuracy is equivalent to the one achieved by the Naïve Bayes Classifier that proved to be in general a good classifier, therefore MRM appears to be relevant whenever the retrieval of a neighbor is required. For this reasons MRM seems particularly suitable for Case Based Reasoning application when a relevant classification of the cases is available.

References

[AG90] David W. Aha and Robert 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.

[AG92] David W. Aha and Robert 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.

[BB96] E. Blanzieri and M. Bucciarelli. The evaluation of the communicative effect.

In *XVIII Cognitive Science Conference*, pages 501–506, San Diego, California, 1996.

[BMP98] R. Bellazzi, S. Montani, and L. Portinale. Retrieval in a prototype-based case library: A case study in diabetes therapy revision. *Lecture Notes in Computer Science*, 1488:64–75, 1998.

[CH67] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27, 1967. [Reprinted in [Das91]].

[CH97] C. Cardie and N. Howe. Improving minority class prediction using case-specific feature weight. In *Proceedings of the Fourteenthn International Conference on Machine Learning*, pages 57–65. Morgan Kaufmann, 1997.

[CMSW92] 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.

[CS93] Scott Cost and Steven Salzberg. A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10:57–78, 1993.

[Das91] B. V. Dasarathy, editor. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Los Alamitos, CA, 1991.

[DP97] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[Fri94] Jerome H. Friedman. Flexible metric nearest neighbour classification. Technical report, Stanford University, 1994. Available by anonymous FTP from playfair.stanford.edu.

[HT95] Trevor Hastie and Robert 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.

[MH90] J. P. Myles and D. J. Hand. The multi-class metric problem in nearest neighbour discrimination rules. *Pattern Recognition*, 23(11):1291–1297, 1990.

[Mit97] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[RA95] Francesco Ricci and Paolo 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.

[RA98] Francesco Ricci and Paolo Avesani. Data compression and local metrics for nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998. In press.

[Sco92] D. W. Scott. *Multivariate Density Estimation: Theory , Practice, and Visualization*. John Wiley, New York, 1992.

[SF80] R. D. Short and K. Fukunaga. A new nearest neighbour distance measure. In *Proceeding of the 5th IEEE International Conference on Patter Recognition*, pages 81–86, Miami beach, FL, 1980.

[SF81] R. D. Short and K. Fukunaga. The optimal distance measure for nearest neighbour classification. *IEEE Transactions on Information Theory*, 27:622–627, 1981.

[SW86] Craig Stanfill and David Waltz. Toward memory-based reasoning. *Communication of ACM*, 29:1213–1229, 1986.

[WD95] Dietrich Wettschereck and Thomas G. Dietterich. An experimental comparison of the nearest neighbor and nearest hyperrectangle algorithms. *Machine Learning*, 19:5–28, 1995.

[WM97] D. Randall Wilson and Tony R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 11:1–34, 1997.

[WMA97] Dietrich Wettschereck, Takao Mohri, and David W. Aha. A review and comparative evaluation of feature weighting methods for lazy learning algorithms. *AI Review Journal*, 11:273–314, 1997.