Advanced Metrics for Class-Driven Similarity Search

Paolo Avesani and Enrico Blanzieri and Francesco Ricci
Istituto per la Ricerca Scientifica e Tecnologica
ITC-IRST
via Sommarive, 38050 Povo (TN), Italy
avesani@itc.it, blanzieri@itc.it, ricci@sodalia.it

Abstract

This paper presents two metrics for the Nearest Neighbor Classifier that share the property of being adapted, i.e. learned, on a set of data. Both metrics can be used for similarity search when the retrieval critically depends on a symbolic target feature. The first one is called Local Asymmetrically Weighted Similarity Metric (LASM) and exploits reinforcement learning techniques for the computation of asymmetric weights. Experiments on benchmark datasets show that LASM maintains good accuracy and achieves high compression rates outperforming competitor editing techniques like Condensed Nearest Neighbor. On a completely different perspective the second metric, called Minimum Risk Metric (MRM) is based on probability estimates. MRM can be implemented using different probability estimates and performs comparably to the Bayes classifier based on the same estimates. Both LASM and MRM outperform the NN classifier with the Euclidean metric.

1. Introduction

A Case Based Reasoning (CBR) system derives a solution to a new problem by adapting past solutions given the underlying assumption that similar problems share similar solutions. In the CBR cycle [1] (a loop of retrieve, reuse, revise and retain steps) the retrieval phase is devoted to find the similar case among the cases stored in the case base depending on a similarity criteria. Functionally, the retrieving corresponds to formulate the similarity query by example, i.e. given a partial description of the case retrieve the similar case. Although the amount of cases in an average case base is seldom comparable to the size of a large database the techniques developed for CBR can be useful for similarity queries in databases, e.g for similarity ranking of standard query results.

Nearest Neighbor (NN) techniques support one of the oldest and more popular classifier and are widely used in Case Based Reasoning (CBR) systems for similarity retrieval

Using the NN as a classifier is quite simple. Given a set of examples, described as points in an input space, and associated to a nominal attribute that play the role of the target concept, i.e. the class, the classification problem is to find the unknown target class of a new example. The NN technique provide an hypothesis of the target value taking the class value of the nearest example. The "nearest" relation is computed using a similarity metric defined on the input space.

The NN technique is exploited in the CBR systems: The similar case to the case at hand is the NN case in the case base w.r.t. a similarity metric. In such systems an existing classification over the cases can be useful (class-driven retrieval). In fact, even if the cases are not explicitly classified in a set of finite classes, often they are clustered in sets containing equivalent solutions against some domain-specific criteria. Interpreting those sets as classes, reduce the retrieval step in a CBR system to the nearest neighbor search of a NN classifier.

In a NN algorithm the metric appears to be critical. Therefore several alternatives to the standard Euclidean metric for continuous spaces and to the Hamming metric for nominal spaces have been proposed in the literature [19, 20, 7, 6, 2, 11, 4, 23]. The metrics can be local, i.e metric whose definition varies depending on the position of the points in the input space, or global, implicitly assuming that the similarity evaluation does not depend on the area of the input space the points are in. Metrics parameters can be computed on given data with a batch procedure relying on statistics [22] or incrementally adapting the initial values with feedback learning algorithms [16, 18].

One of the goals of a modified metric is to improve classification accuracy (i.e. the ratio between the number of

correctly classified examples and the number of classified examples) that, in a retrieval context, can be seen related to the measure of precision (i.e. the ratio between the number of correctly retrieved instances and the number of retrieved instances). Another focal issue is the reduction of the retrieval time that can be achieved by KD-trees indexing or alternatively by selection of a subset of the training sample (data compression). Data compression can improve also the accuracy when noisy data and outlyers are discarded. For example, the compression technique of an Edited Nearest Neighbor algorithms selects a subset of prototypes from a training set to improve computational efficiency, i.e. retrieval time, and make the classification more reliable (see [7, Chapter 6]).

In this paper we focus our attention on two advanced metrics for the NN classifier that were introduced very recently and have complementary properties in terms of accuracy and compression and can be useful for class-driven similarity search.

The first one is called Local Asymmetrically Weighted Similarity Metric (LASM) and has been introduced and studied by two of the authors [18]. LASM is a metric with asymmetric weights that are learned with a reinforcement learning procedure and it can achieve compression rates that outperforms standard editing algorithms (theoretical arguments sustaining this statement are presented in [17]). Moreover the generalization accuracy is also improved with respect to 1-NN equipped with local or global metrics [7, 20, 23]. An interesting feature of the proposed framework is that the method for selecting the prototypes can be very simple. We report a set of experiments that shows that more complex prototypes selection policies (clustering) do not improve accuracy on random selection. The rationale is that the metric and the optimization procedure are flexible enough to adapt to a suboptimal prototypes selection.

Two of us have introduced the Minimum Risk Metric (MRM) [3] that is a metric based on probability estimation, i.e. consistent estimates of the conditional probabilities of the classes. Among those metrics the one proposed by Short and Fukunaga [19] has the strongest theoretical foundation but it was introduced only for continuous features. The MRM generalizes to mixed features (both continuos and nominal) and relies its effectiveness on a different and simpler optimality condition than the one suggested by Short and Fukunaga. We report experimental comparison between the classification accuracies of MRM and the performances of other metrics that are available in literature.

The work is organized as follows: Section 2 describes LASM and the experimental results. Section 3 briefly presents the Minimum Risk Metric, describes the optimality criteria, the adopted probability estimators and some experimental results. Finally, Section 4 draws conclusions and future directions.

2. Local Asymmetrically Weighted Similarity Metric

LASM is basically a Minkowsky metric[1] with the difference that for each dimension of the input space, the distance from a reference value to another value varies depending on the fact that the second is greater or less than the first (asymmetry). Moreover the metric is learned: For each example stored in the edited training set, ET, there is a corresponding set of weights that defines the metric attached to that example. When an example in the edited training set must be compared with another example its attached metric is used.

More formally, let $[0,1]^d$ be the input space and $x \in [0,1]^d$ a generic example. For each feature we assume that there is a feature metric $\delta_j : [0,1] \times [0,1] \longrightarrow \mathcal{R}_{\geq 0}$. An obvious feature metric for a real feature space is $\delta_j(x_j, y_j) = |x_j - y_j|$. Given an example $x \in [0,1]^d$, a *set of asymmetric weights* $w(x)$ for x is a $2 \times d$ matrix with values in $[0,1]$. Let $w_j^k(x)$, $k = 0, 1$ and $j = 1, \ldots, d$, be a generic entry of $w(x)$. Let y be another point in the input space, the following notation will be used:

$$w_j(x) \cdot \delta_j(x_j, y_j)^p = \begin{cases} w_j^0(x)\delta_j(x_j, y_j)^p & \text{if } y_j \leq x_j \\ w_j^1(x)\delta_j(x_j, y_j)^p & \text{otherwise} \end{cases}$$

for any positive integer p. Given a set $ET \subset [0,1]^d$ (Edited Training Set) and a set of asymmetric weights for each example in ET, a *local asymmetrically weighted similarity metric* (LASM) is defined as follows:

$$\delta(x, y) = \left(\sum_{j=1}^{d} w_j(x) \cdot \delta_j(x_j, y_j)^p\right)^{1/p} \qquad (1)$$

A learning procedure (see [18] for details), based on reinforcement learning, computes the weights. Two functions $P(w, x, y)$ and $R(w, x, y)$, the punishment and reinforcement respectively, adjust iteratively an initial system of weights. Assuming a boolean classification $c : T \rightarrow 0, 1$, when $c(x)$, $x \in ET$, is equal to $c(y)$, $y \in T$, i.e. the prediction is correct, a reinforcement is given to the system that reduces the distance between the nearest neighbor x and the sample y; whereas if the two values are not equal a punishment step increases the distance between x and y.

The importance of an asymmetric weighting scheme is best understood when, given a training set T and an edited training set ET, a system of weights is to be found in such a way that the nearest neighbor classifier, endowed with the

[1] Minkowsky metrics are defined as $m(x, y) = \sum_{j=1}^{d} (|x_j - y_j|^p)^{1/p}$ where p has integer values. Varying p the metrics generate the Manhattan metric ($p = 1$), the Euclidean metric ($p = 2$) and the Chebychev metric ($p \rightarrow \infty$).

metric defined by that set of weights, is as accurate as possible. Asymmetric local weights, acquired with such a kind of learning procedure, produce a metric that adapt to the data in a flexible way; they enable a freer choice of examples to store in the edited training set; they require less search when the selection is to be optimized, and make possible high compression rates.

2.1. Experimental results

LASM chooses ET at random. The number of examples for each class in ET is made proportional to the number of examples in that class found in the training set. So the selection is random but the probability to have an example in a given class in the training set is equal to that in ET.

The first set of experiments are aimed to evaluate how different techniques for prototypes selection impact on LASM accuracy. A natural choice would be to select those prototypes that are more representative of the clusters in the data. We have therefore defined three new classifiers that differ from LASM in the way the initial set of prototypes are selected. In the first, K-LASM, the prototypes are taken as the centroids of the clusters computed by the k-means algorithm [9, 12]; in the second, H-LASM, the prototypes are the upper nodes of the hierarchy induced on the training sets by the nearest neighbor hierarchical clustering method [12]; in the third, R-LASM, 20 ET sets are randomly generated by taking a fixed percentage of the examples in the training. The set that maximize the accuracy on the training is chosen. All these algorithms, for each data set, use the same number of prototypes.

Table 1. Comparison of the average accuracy and standard deviation obtained on different data sets by using different techniques to select the prototypes. No significative difference is detected at 0.02 level for a paired t-test.

	Algorithm		
Data Set	K-LASM	H-LASM	R-LASM
Balance Scale	86.4 ±2.0	87.1 ±1.7	86.5 ±1.8
Wisconsin	75.9 ±3.7	75.9 ±3.7	75.7 ±3.8
Cleveland	80.6 ±2.8	79.6 ±3.6	80.1 ±3.4
Echocardiogram	70.1 ±9.8	71.8 ±7.4	70.9 ±10.5
Ionosphere	92.7 ±2.2	92.3 ±2.8	90.0 ±2.3
Iris	94.6 ±3.0	95.0 ±2.7	95.0 ±2.9
Liver Disorders	62.6 ±4.4	59.3 ±4.6	63.4 ±4.5
Thyroid	94.2 ±2.8	93.4 ±3.7	95.0 ±2.8
Wine	95.8 ±2.6	96.2 ±2.7	96.3 ±2.6
Pima	73.9 ±4.0	73.8 ±2.4	73.9 ±3.0

Table 1[2] shows the accuracy of the three new classifiers. Each experiment is based on 50 runs with the data set split in two parts: 2/3 for training and 1/3 for test. There is no significant difference in accuracy. These results strongly suggest that the choice of prototypes is not determinant and the behavior of LASM is only due to the particular metric (local and asymmetric) and to the learning procedure.

Another session of experiments compares the compression rates of LASM to other edited nearest neighbor algorithms [7, Chapter 6]. The common objective of the algorithms considered is the computation of a consistent subset, that is a reduction of the original training in such a way that the accuracy of the classifier is not decreased. The edited nearest neighbor algorithms compared with LASM are: RNN, Reduced Nearest Neighbor [10]; ICA, Iterative Condensation Algorithm [21]; PNN, Prototypical Nearest Neighbor [5].

Table 2. Compression rates in percentage of the editing experiments. A "-" means significantly worse than LASM at least at 0.02 level for a paired t-test.

	Algorithm				
Data	1-NN	ICA	RNN	PNN	LASM
Ba	- 78.2 100%	- 75.3 27%	- 71.7 40%	- 67.6 36%	85.5 3%
Wi	- 63.9 100%	- 59.2 48%	- 59.2 48%	- 59.4 51%	75.9 2%
Cl	- 77.2 100%	- 73.8 34%	- 74.1 35%	- 73.7 37%	81.1 4%
Ca	- 63.2 100%	- 64.0 48%	- 62.4 46%	- 59.2 55%	70.4 4%
Io	- 90.0 100%	- 87.0 16%	- 87.6 15%	- 86.7 23%	91.8 8%
Ir	- 93.9 100%	- 92.0 13%	- 93.0 12%	93.8 13%	95.3 9%
Li	62.2 100%	62.6 49%	61.9 50%	60.0 48%	61.4 4%
Th	96.0 100%	- 92.6 11%	- 92.0 10%	93.8 12%	94.7 6%
Win	96.0 100%	92.7 11%	93.5 9%	94.2 11%	95.7 3%
Pi	- 69.5 100%	- 64.4 43%	- 63.7 43%	- 65.2 42%	73.7 2%

The results are shown in table 2. LASM outperforms both in accuracy and compression these classifiers. The most significant improvement of LASM concerns the compression rate. ICA, RNN and PNN store on average 31% of the training examples, whereas LASM stores only 4.5% of the training. It should also noted that while the other algorithms decrease the accuracy of the 1-NN classifier, LASM obtains both better compression rates and outperforms 1-NN in accuracy.

3. Minimum Risk Metric

Minimum Risk Metric (MRM) is a metric based on probability estimation that minimizes the risk of misclassification.

[2]The Cleveland data have been provided by Robert Detrano from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. Breast Cancer data have been provided by M. Zwitter and M. Soklic from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.

Given an example x in class c_i with $i = 1 \dots m$ and a nearest neighbor y the finite risk of misclassifying x is given in terms of conditional probabilities 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 [19] and of Myles and Hand [15] is to subtract the asymptotic risk $r^*(x,y)$ w.r.t the increasing of the dimension of the sample and minimizing $E(r(x,y) - r^*(x,y))$. Instead it is possible to minimize directly the risk $r(x,y)$ and that leads to the metric:

$$mrm(x,y) = r(x,y) = \sum_{i=1}^{m} p(c_i|x)(1 - p(c_i|y)). \quad (2)$$

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

$$p(c_i|x) = \frac{p(x|c_i)p(c_i)}{p(x)} = \frac{p(x|c_i)p(c_i)}{\sum_{k=1}^{m} p(x|c_k)p(c_k)} \quad (3)$$

therefore reducing to the problem of estimating $p(x|c_k)$. The Minimum Risk metric can work with different probability estimators. Here we choose an estimator based on the same estimate of the Naïve Bayes Classifier [14, 8]. The probabilities are estimated with frequencies (discretizing the continuous features). 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. Frequency tends to underestimate the probability if the sample size is small, a viable solution is to adopt the Laplace–corrected estimate or equivalently incrementing artificially the sample size. Following the former option 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. The features' independence assumption is expressed by:

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

and, substituted in the equation (3), leads to the estimates used in the Naïve Bayes Classifier approach.

3.1. Experimental Results

The aim of the experiments is the comparison of of the performances of MRM to the ones of two other metrics available in literature: Value Difference Metric and Combined Euclidean-Overlap Metric.

Value Difference Metric (VDM). Another very common metric based on probabilistic consideration is VDM introduced by Stanfill and Waltz [20] who used it exclusively on input spaces with nominal features. They express their metric by means of frequencies of occurrences of feature values. Interpreting the frequencies as estimates of probability the definition can be written as:

$$vdm(x,y) = \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{m} (p(c_i|x_j)^2} \sum_{i=1}^{m} (p(c_i|x_j) - p(c_i|y_j))^2 \quad (4)$$

In our experiments we used a simplified version of VDM:

$$vdm(x,y) = \sum_{j=1}^{n} \sum_{i=1}^{m} |p(c_i|x_j) - p(c_i|y_j)| \quad (5)$$

Wilson and Martinez extended VDM to instances with numeric attributes [23] essentially by discretization of the numeric attribute (DVDM).

Combined Euclidean–Overlap Metric (HEOM). The metric HEOM was introduced by Wilson and Martinez [23], 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 $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ 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.

MRM, DVDM and HEOM were tested on 27 datasets taken from the Machine Learning Databases Repository at UCI [13] and on two new databases (Derma and Opera). The datasets contain continuous, nominal and mixed features as well as unknown values. The experimental paradigm was a 10-CV cross validation and the results are shown in Tab. 3.

MRM outperforms HEOM (with a notable exception on the sonar data set) and DVDM except for the monks datasets. The latter datasets appear to be a hard task probably as a consequence of the assumption of the independence among features that underlies the Naïve Bayes Estimates. Results not shown here shows how MRM outperforms DVDM and HEOM more convincingly than SF2 and without any local restriction. Moreover other results shown that MRM behaves equivalently to the Bayes Classifier based on the same estimation [3]

4. Conclusion

In this paper we have described two metrics for nearest neighbor classification: Local Asymmetrically–weighted Similarity Metric and Minimum Risk Metric. LASM uses

Table 3. Classification accuracy in percentage of the Minimum Risk Metric with the Naïve Bayes Estimator, DVDM and HEOM. Significative differences ($p < 0.05$) are shown e.g. MRM performs significatively better than both DVDM and HEOM on the breast-cancer data set.

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

an asymmetrical weighted schema and a feedback learning procedure. It achieves good compression rates and outperforms the accuracy of the standard nearest neighbor classifier. Minimum Risk Metric is based on probability estimation and minimizes the risk of misclassification. Its analytical form is simple and well founded, and finally, equipped with a simple Naïve Bayes Estimator, outperforms the other metrics. A direct comparison of the two techniques is scheduled as future work.

Both the metrics are implemented as part of Case Based Exploration Tool, CBR oriented collection of algorithms implemented as a C++ Library and developed at ITC-IRST. As future work we are planning to interface the library to a commercial DBMS for applying the techniques to case bases extracted from data bases via standard queries.

References

[1] A. Aamodt and E. Plaza. Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59, 1994.

[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] E. Blanzieri and F. Ricci. A minimum risk metric for nearest neighbor classification. 1999. Accepted at ICML99.

[4] 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.

[5] C.-L. Chang. Finding prototypes for nearest neighbour classifier. *IEEE Transactions on Computers*, C-23(11):1179–1184, 1974.

[6] 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.

[7] B. V. Dasarathy, editor. *Nearest neighbor (NN) norms: NN pattern classification techniques.* IEEE Computer Society Press, Los Alamitos, CA, 1991.

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

[9] R. O. Duda and P. E. Hart, editors. *Pattern Classification and Scene Analysis*. John Wiley & SonsMIT Press, 1973.

[10] G. W. Gates. The reduced nearest neighbor rule. *IEEE Transaction on Information Theory*, 18(3):431–433, 1972.

[11] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbour classification. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 18(6):607–615, 1996.

[12] A. K. Kain and R. C. Dubes, editors. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[13] 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.

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

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

[16] 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.

[17] F. Ricci and P. Avesani. Exact learning and data compression with a local asymmetrically weighted metric. In *ICML-96 Workshop on Learning in Context-Sensitive Domains, Bari, July 3rd,*, 1996.

[18] F. Ricci and P. Avesani. Data compression and local metrics for nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(4), Apr. 1999.

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

[20] C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communication of ACM*, 29:1213–1229, 1986.

[21] C. W. Swonger. Sample set condensation for a condensed nearest neighbor decision rule for pattern recognition. In S. Watanabe, editor, *Frontiers of Pattern Recognition*, pages 511–519. Academic Press, 1972.

[22] 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.

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