# *K* Nearest Neighbor Classification with Local Induction of the Simple Value Difference Metric

Andrzej Skowron and Arkadiusz Wojna

Faculty of Mathematics, Informatics and Mechanics
Warsaw University
ul. Banacha 2, 02-097 Warsaw, Poland
{skowron,wojna}@mimuw.edu.pl

**Abstract.** The classical $k$ nearest neighbor ($k$-nn) classification assumes that a fixed global metric is defined and searching for nearest neighbors is always based on this global metric. In the paper we present a model with local induction of a metric. Any test object induces a local metric from the neighborhood of this object and selects $k$ nearest neighbors according to this locally induced metric. To induce both the global and the local metric we use the weighted Simple Value Difference Metric (SVDM). The experimental results show that the proposed classification model with local induction of a metric reduces classification error up to several times in comparison to the classical k-nn method.

## 1 Introduction

The classical machine learning methods [1, 2] induce a mathematical model of data from training data and apply this model to reason about test objects. The induced model remains invariant while reasoning about different test objects. For many real-life data it is not possible to induce relevant global models. This fact has been recently observed by researches from different areas like data mining, statistics, multiagent systems [3–5]. The main reason is that phenomena described by real-life data are often too complex and we do not have enough knowledge to induce global models or a parameterized class of such models together with searching methods for the relevant global model in such a class. We propose a step toward development of methods dealing with such a real-life data. For any test object $x$ first we use some heuristics (in our example based on distances) that make it possible to eliminate objects not relevant for classifying $x$. From the remaining (rather of small size) neighborhood of $x$ a local model (in our case a distance function) is induced that is relevant for classifying the test object $x$. Hence, our idea is based on extracting for a given test object $x$ its local model that is dependent on $x$ and next using this model for classifying $x$.

To apply this idea we extend the $k$ nearest neighbor ($k$-nn) method [6, 7] with one additional intermediate step (see Figure 1). First it induces a global metric like in the classical $k$-nn. Then for each test object $x$ the extended algorithm selects a neighborhood of $x$ and it induces a local metric based only on the selected neighborhood. After that the $k$ nearest neighbors of the test object $x$

S. Tsumoto et al. (Eds.): RSCTC 2004, LNAI 3066, pp. 229–234, 2004.

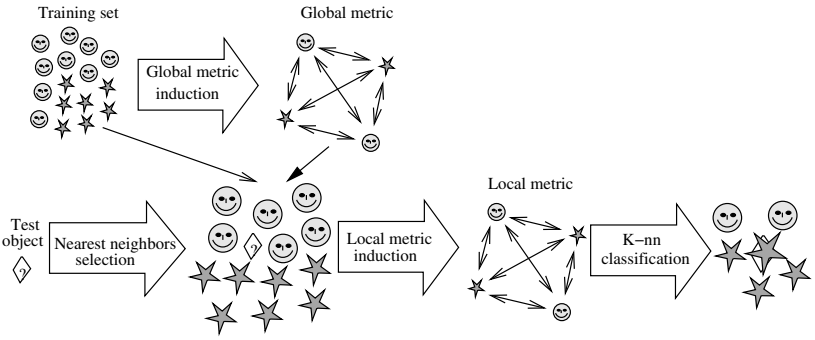


**Fig. 1.** $K$-nn classification with local metric

are selected according to the locally induced metric. Local metric induction is a step towards building a model that depends locally on properties of a test object. In both definitions of metrics: local and global, we use the weighted version of the Simple Value Difference Metric [8] defined for data with nominal attributes.

We have tested three data sets each with several thousand of training objects, and the model with local metric induction has reduced the classification error of the classical k-nn from 30% up to several times depending on the tested set.

## 2 Related Work

The classical nearest neighbor method was introduced by Cover and Hart [6] and the extension of 1-nn to $k$-nn was described by Duda and Hart [7]. Local adaptation of a metric in the $k$-nn method was considered only in the context of a multidimensional space with real value attributes.

Friedman proposed a method that combines $k$-nn with recursive partitioning used in decision trees [9]. For each test object the method starts with the whole training set and it constructs a sequence of partitions. Each partition eliminates a number of training objects. In this way after the last partition a small set of $k$ objects remains to be used for classification. To make a single partition the direction with the greatest decision discernibility is selected.

The algorithm proposed by Hastie and Tibshirani [10] starts with the Euclidean metric and for each test object it iteratively changes the weights of attributes. At each iteration it selects a neighborhood of a test object and it applies local discriminant analysis to shrink the distance in the direction parallel to the boundary between decision classes. Finally it selects $k$ nearest neighbors according to the locally transformed metric.

Domeniconi and Gunopulos use a similar idea but they use support vector machines instead of local discriminant analysis to determine class boundaries and to shrink the distance [11]. Support vectors can be computed during the learning phase what makes this approach much more efficient in comparision to local discriminant analysis.

## 3 K Nearest Neighbors with the Global SVDM Metric

We assume that a training set $\mathbb{U}$ is provided and each object $x \in \mathbb{U}$ is labeled with a decision $dec(x)$ from a finite set $V_d$. The task is to learn from a training set $\mathbb{U}$ how to induce the correct decision for new unlabeled data objects.

$K$-nn is a widely used classification model assuming that data objects are given from a pseudometric space $\mathbb{X}$ with a distance function $\rho : \mathbb{X}^2 \to \mathbb{R}$. The distance function $\rho$ is induced from a training set $\mathbb{U}$ during the learning phase. Then for each data object $x$ to be classified the set $S(x,k)$ of the $k$ nearest neighbors of $x$ is selected from $\mathbb{U}$ according to a distance function $\rho$ and a decision is inferred from the decisions of the nearest neighbors in $S(x,k)$.

In the paper we use one of the most popular procedures to determine a decision for a test object $x$. For each decision value $v \in V_d$ the $Strength$ measure counts the number of the nearest neighbors from $S(x,k)$ with the decision $v$:

$$Strength(x,v) = |\{y \in S(x,k) : dec(y) = v\}|$$

As a decision for a test object $x$ the algorithm assigns the most frequent decision in the set of the k nearest neighbors $S(x,k)$:

$$dec_{k-nn}(x) = \arg\max_{v \in V_d} Strength(x,v)$$

As a distance function $\rho$ we use the weighted version of the Simple Value Difference Metric (SVDM) [12]. It assumes that data objects are represented as vectors of nominal values $x = (x_1, \ldots, x_n)$. The distance between two data objects $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ is defined by

$$\rho(x,y) = \sum_{i=1}^{n} w_i \cdot \rho_i(x_i, y_i)$$

where $\rho_i(\cdot,\cdot)$ is a measure of similarity between two attribute values and $w_i$ are weights computed in the second phase of the metric induction process. Two nominal values $x_i, y_i$ are considered to be similar if they imply similar decision distribution, i.e., if they correlate similarly with the decision on the training set $\mathbb{U}$:

$$\rho_i(x_i, y_i) = \sum_{v \in V_d} |P(dec = v|x_i) - P(dec = v|y_i)|$$

As an attribute weighting procedure we use a procedure described in [8].

## 4 *K* Nearest Neighbors with Local Metrics

In this section we describe an extension of the $k$-nn method with local metric induction. The learning phase of the extended method is analogical to the classical $k$-nn. It induces a global metric $\rho$ from a training set $\mathbb{U}$.

Algorithm 1 presents the classification of a single query object $x$ by the extended method. First the algorithm selects the $n$ nearest neighbors $N(x,n)$

**Algorithm 1** $K$ nearest neighbors with local metric

$\rho$ - the weighted SVDM metric induced from the whole training set $\mathbb{U}$
$x$ - a query object to be classified

$N(x,n) :=$ the set of $n$ nearest neighbors of $x$ from $\mathbb{U}$ according to $\rho$
$\rho^x :=$ local weighted SVDM metric induced from the neighborhood $N(x,n)$
$S(x,k) :=$ the set of $k$ nearest neighbors of $x$ from $N(x,n)$
according to $\rho^x$ $(k \leq n)$
$dec_{local-knn}(x) := \arg\max_{v \in V_d} |\{y \in S(x,k) : \ dec(x) = v\}|$

of $x$ from $\mathbb{U}$ according to the global metric $\rho$. Next it induces a local metric $\rho^x$ using only the selected neighborhood $N(x,n)$. After that the algorithm selects the nearest neighbors of $x$ again but only the $k$ nearest neighbors and only from the neighborhood set $N(x,n)$. The selected set $S(x,k)$ is then used to compute the majority decision $dec_{local-knn}(x)$ that is returned as the final result for the query object $x$. Both for the global and for the local metric definition the algorithm uses the weighted version of the SVDM metric described in Section 4.

To improve classification accuracy the neighborhood size $n$ should be large, at least several hundred. To accelerate selection of a large number of nearest neighbors from a training set we use the advanced hierarchical indexing [8].

The optimal value $k$ can be estimated from a training set. We use the procedure analogical to the estimation procedure proposed for the classical $k$-nn [13].

The classical $k$-nn is called a lazy method: it induced a global metric and it performs the rest of computation at the moment of classification. The described algorithm extends this idea: it repeats metric induction at the moment of classification. The proposed extension allows to use the local properties of data topology in the neighborhood of a query object and to adjust the metric definition to these local properties.

## 5 Experimental Results

We have performed experiments for 3 large benchmark data sets with nominal attributes from the UCI repository [14] (in parenthesis the number of attributes, the training set and the test set size): *chess* (36, 2131, 1065), *nursery* (8, 8640, 4320) and *DNA-splice* (60, 2000, 1186). The data sets provided as a single file (*chess, nursery*) have been randomly split into a training and a test part with the ratio 2 to 1. The data set *splice* have been tested with the original partition.

Each data set has been tested with the classical $k$-nn and with the extended $k$-nn with three different values of the neighborhood size $n$: 100, 200 and 500. To make the results comparable all four classification models have been tested with the same partition of a data set and with the same global metric. Each method estimated the optimal value $k$ from a training set in the range $1 \leq k \leq n$ and it used this value during classification of a test set. Each test has been repeated 3 times for each data set and the average classification error has been calculated.

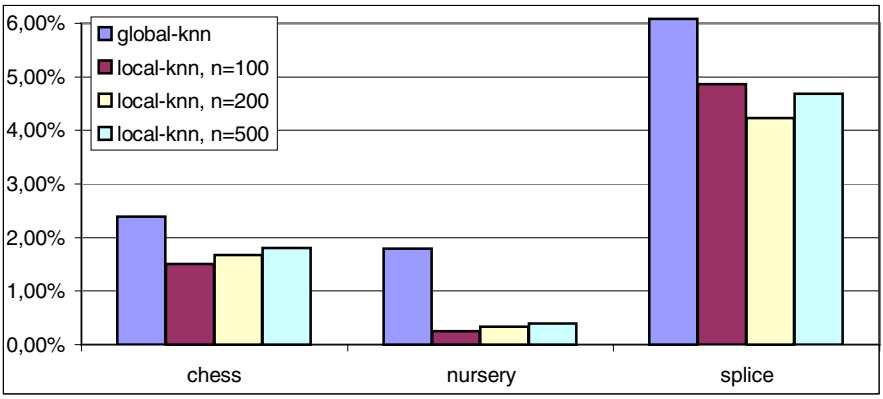


**Fig. 2.** Classification error of the classical $k$-nn and the extended $k$-nn with three different neighborhood sizes: 100, 200 and 500

Figure 2 presents the comparision of the average classification errors obtained from the experiments. The graph shows that for all data sets the k-nn model with local metric induction reduces the classification error significantly in comparision to the classical k-nn based on a global metric. In case of the data sets *chess* and *splice* the reduction is between 20% and 40% depending on the neighborhood size and in case of the data set *nursery* the reduction is severel times (over 7 times in the best case). The presented results prove that a large profit can be obtained if one applies a local approach to data instead of the global one.

The difference between the results for *nursery* and for two other data sets seems to correlate with the data set size. It implies that the larger data set is the more profitable it is to include local metric induction as an intermediate step.

The interesting issue is the dependence between the classification error and the neighborhood size $n$ used to induce a local metric. The best error reduction was obtained for $n = 100$ in case of two data sets and for $n = 200$ in case of the data set *splice*. In particular, the optimal neighborhood size is larger in case of the data set *splice* than in case of *nursery* although the latter data set is about 4 times larger than the former one. It indicates that the optimal neighborhood size depends strongly on the properties of a data set and an advanced technique is necessary to estimate this optimal size.

## 6 Conclusions

In the paper we proposed a new classification model that is an extension of the classical $k$-nn and we compared the accuracy of the new and the original method. The classical $k$-nn assumes that a fixed distance function is defined for the whole data space. The extended model induces a different distance function for each object to be classified and it uses only local information around the object to induce this distance function. This approach allowed us to adapt the metric depending on the local properties of data topology.

We have applied the new method to the classification problem for data with nominal attributes. The experimental results show that the presented approach has an adventage over the original $k$-nn method. The extended method reduces the classification error from several tens percent up to several times.

The $k$-nn model with a local metric corresponds to the idea of transductive reasoning [5]. The transductive approach assumes that a classification model should depend on the objects to be classified and it should be adapted according to the properties of these objects. The presented extension of $k$-nn implements transduction: local metric induction adapts the metric defintion to the local topology in the neighborhood of an object to be classifed.

## Acknowledgments

This work was supported by the grants 4 T11C 040 24 and 3 T11C 002 26 from Ministry of Scientific Research and Information Technology of the Republic of Poland.

## References

1. Mitchell, T.M.: Machine Learning. McGraw-Hill, Portland (1997)
2. Pawlak, Z.: Rough Sets - Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
3. Breiman, L.: Statistical modeling - the two cultures. Statistical Science **16** (2001) 199–231
4. Skowron, A., Stepaniuk, J.: Information granules and rough-neural computing. In: Rough-Neural Computing: Techniques for Computing with Words. Cognitive Technologies. Springer-Verlag, Heidelberg, Germany (2003) 43–84
5. Vapnik, V.: Statistical Learning Theory. Wiley, Chichester, GB (1998)
6. Cover, T.M., Hart, P.E.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **13** (1967) 21–27
7. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York, NY (1973)
8. Wojna, A.G.: Center-based indexing in vector and metric spaces. Fundamenta Informaticae **56** (2003) 285–310
9. Friedman, J.: Flexible metric nearest neighbor classification. Technical Report 113, Department of Statistics, Stanford University, CA (1994)
10. Hastie, T., Tibshirani, R.: Discriminant adaptive nearest neighbor classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **18** (1996) 607–616
11. Domeniconi, C., Gunopulos, D.: Efficient local flexible nearest neighbor classification. In: Proceedings of the Second SIAM International Conference on Data Mining. (2002)
12. Domingos, P.: Unifying instance-based and rule-based induction. Machine Learning **24** (1996) 141–168
13. Góra, G., Wojna, A.G.: RIONA: a new classification system combining rule induction and instance-based learning. Fundamenta Informaticae **51** (2002) 369–390
14. Blake, C.L., Merz, C.J.: UCI repository of machine learning databases. http://www.ics.uci.edu/˜mlearn/MLRepository.html, Department of Information and Computer Science, University of California, Irvine, CA (1998)