MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

A Comparison of Different Heterogeneous Proximity Functions and Euclidean Distance

Janne Lumijärvi, Jorma Laurikkala, Martti Juhola

Department of Computer Sciences, University of Tampere, Tampere, Finland

Abstract

Proximity functions evaluate distances or similarities between objects. Unlike the Euclidean distance, heterogeneous proximity functions process variables differently according to their scale. The correct evaluation of nominal variables, whose values are unordered, is especially important. We compared five heterogeneous functions with the Euclidean distance to study whether functions sensitive to scale are better than a function assuming the same scale. In addition, we were interested of the relative performance of the five heterogeneous functions. The performance of the functions was measured with a nearest neighbor classifier that was applied to 12 medical data sets characterized with different scales. Unexpectedly, the performance of heterogeneous functions did not differ significantly from that of the Euclidean distance. As expected, significant differences between the Heterogeneous Value Difference Metric (HVDM) and the four value-matching-based heterogeneous functions favored HVDM. Additional research is needed to explain why heterogeneous functions did not outperform the Euclidean distance.

Keywords:

Proximity function, Distance, Similarity, Nearest neighbor, Machine learning.

Introduction

Proximity functions [1-5] are an important component of many statistical and machine learning methods. These functions are typically applied to compute distances or similarities between pairs of objects (cases or examples). Methods that employ proximities include many clustering algorithms [1-3] and instance-based learning systems [4] for grouping and classification of objects such as patient cases. The choice of a proximity function is of utmost importance, because an unsuitable function may dramatically degrade the performance of a method that operates on proximities.

This work considers the use of different proximity functions in the classification of heterogeneous data, i.e. data described both with nominal and quantitative (ordinal, interval, and ratio) scaled variables. The scale [2,5] of a variable indicates the information available on relations between variable values. Since nominal values have no meaningful order, nominal scale carries clearly less information than quantitative scales, whose values can be arranged. Of a pair of unordered values one can only observe whether values are equal: A physician may compare two cases, for example, by examining whether the location of tinnitus is the left ear in both cases.

Heterogeneous data is problematic for proximity functions assuming the same scale for all the variables. Application of the well-known Euclidean distance function to nominal values is questionable, because arithmetical operations for the unordered values are meaningless. As an example, consider values "no tinnitus", "left", "right", and "bilateral" with scores 0, 1, 2, and 3. The Euclidean distance would rate "right" twice as distant from "no tinnitus" than "left", which might not be reasonable from the medical point of view.

When a proximity measure is to be applied to the heterogeneous data, there are basically three approaches. Firstly, the nature of the data may simply be ignored. Soundness of results largely depends on the method and the task, but problems described above are likely to emerge.

Secondly, variables can be transformed to meet scale assumptions. If the nominal scale is the simplest in the data, the categorization of interval and ratio scaled variables is needed. Deciding the number of categories is difficult, and more importantly, when categories are treated as nominal the order information is lost. Transformations from a lower to a higher scale are impossible, but a k-valued nominal variable can be coded as k binary or k-1 dummy variables [6] for which arithmetical operations are legal. Dummy coding is often used to facilitate multivariate statistical methods when multi-valued nominal variables are included in models. The obvious drawback of dummy coding is the increased data dimensionality.

Thirdly, one can analyze the heterogeneous data as it is by applying a heterogeneous proximity function that is able to handle different scales. This approach requires no transformations thus avoiding the loss of information, as well as the increase of dimensionality.

In the following, we compare six proximity functions with two aims in mind. Five heterogeneous functions were weighed against the Euclidean distance to assess whether heterogeneous functions are better than a function assuming the same scale. In addition, we wished to study the performance of typical value-matching-based heterogeneous functions with respect to the Heterogeneous Value Difference Metric (HVDM) function [5] of Wilson and Martinez. HVDM utilizes class information and has shown to be a competitive approach to evaluate proximities

[5]. We have found this function useful in classification tasks involving medical data [7,8]. However, there are also tasks, such as clustering, where class labels are unknown and one cannot apply HVDM, but value-matching-based functions are appropriate. Performance of the functions was evaluated using a nearest neighbor classifier and a collection of heterogeneous medical data sets.

Methods

Nearest neighbor classification

Nearest neighbor technique is a classic instance-based machine learning method [4] which classifies the unseen examples in a testing set T into C mutually exclusive classes on the basis of the k nearest examples in a learning set L. Each test example is labeled as belonging to the most frequent class of its neighborhood.

The proximity functions were evaluated with a three-nearest neighbor classifier (3-NN) which is less sensitive to noise than the common 1-NN classifier [4]. To make tests more repeatable, breaking ties randomly was avoided. Most ties were deterministically broken by considering the distances between a test example and its neighbors.

Proximity functions

Function $f: E \times E \; \boldsymbol{R}$ is a proximity function in a space E, if for every example $x, y \in E$, there exists a lower or upper bound f_0 for f so that $f(x,x) = f_0$ and $f(x,y) = f(y,x)$ [3]. Next, we shortly describe the proximity, i.e. distance and similarity, functions evaluated in this study.

Heterogeneous functions were compared to probably the most well-known proximity measure – the Euclidean distance [1-5], a variant of the general Minkowski function. The Euclidean distance (EUCL) is defined as:

$$D_2(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (1)$$

where m is the number of variables and the value of the ith variable for an example x is denoted x_i ($1 \le i \le m$). In these experiments, each quantitative variable was normalized by dividing it by four standard deviations ($4s_i$) as in [5]. The nominal variables were similarly normalized with their standard deviations, to make the Euclidean distance wholly insensitive to scales.

Heterogeneous proximity functions

Aha's Heterogeneous Euclidean-Overlap Metric [5] (HEOM)

$$HEOM(x,y) = \sqrt{\sum_{i=1}^{m} h_i(x_i - y_i)^2} \qquad (2)$$

is a distance function that treats nominal and quantitative variables differently:

$$h_i(a,b) = \begin{cases} 1, & \text{if a or b is missing} \\ I_D(a,b) & \text{if ith variable is nominal} \\ (|a-b|/rng_i) & \text{if ith variable is quantitative} \end{cases}$$

where rng_i is the range of the ith variable and I_D is an overlap function

$$I_D(a,b) = \begin{cases} 0 & \text{if a=b} \\ 1 & \text{otherwise} \end{cases} \qquad (4)$$

Gower's similarity function [1] (GOW) is defined as:

$$GOWER(x,y) = \sum_{i=1}^{m} g_i(x_i,y_i) / \sum_{i=1}^{m} I_o \cdot (x_i,y_i) \qquad (5)$$

where g is

$$g_i(a,b) = \begin{cases} 0, & \text{if a or b is missing} \\ 1 - I_D(a,b) & \text{if ith variable is nominal} \\ 1-(|a-b|/rng_i) & \text{if ith variable is quantitative} \end{cases}$$

and I_O is an indicator function for observed values

$$I_o(a,b) = \begin{cases} 1 & \text{if a and b} \\ 0 & \text{otherwise} \end{cases} \qquad (7)$$

Gower's function differs from HEOM in that the Manhattan distance is computed instead of the Euclidean distance, and values are normalized by definition into range [0,1] by dividing the sum of similarities by the number of observed value pairs. It is assumed that the sum is greater than zero.

Another similarity measure of this study is Estabrook-Rogers similarity function [9] (ER):

$$ER(x,y) = \sum_{i=1}^{m} er_i(x_i,y_i) \qquad (8)$$

where missing values and nominal values of the ith variable are treated as in Gower's function:

$$er_i(a,b) = \begin{cases} 0, & \text{if a or b is missing} \\ 1 - I_D(a,b) & \text{if variable nominal,} \\ \dfrac{2u_i + 1 - |a-b|}{2u_i + 2 + |a-b|u_i}, & \text{if variable quantitative, } (|a-b| \le u_i) \\ 1 & \text{otherwise} \end{cases} \qquad (9)$$

However, the Manhattan distance for quantitative variables has an upper bound u_i, values over which are considered maximally dissimilar. We chose $u_i = rng_i - 2$ which is the largest reasonable value [9].

Ichino-Yaguchi generalized Minkowski metric [10] (GEM) is based on the Cartesian space model and can compute distances for sets and intervals of attribute values by utilizing the Cartesian join and meet operators. Below, GEM is very briefly described in its Euclidean data point form with parameter value g = 0.5:

$$GEM(x,y) = \sqrt{\sum_{i=1}^{m} gem(x_i,y_i)^2} \qquad (10)$$

$$gem_i(a,b) = \begin{cases} 0 & \text{if a=b} \\ 1/dmn_i & \text{if ith variable is norminal} \\ |a-b|/dmn_i & \text{if ith variable discrete} \\ |a-b|/rng_i & \text{if ith variable is continous} \end{cases} \qquad (11)$$

where dmn_i is the size of value domain of the ith variable. In contrast to the other heterogeneous functions, GEM normalizes the difference in a nominal variable with the size of value domain.

Also Heterogeneous Value Difference Metric [5] (HVDM) has a main function that resembles the Euclidean distance:

$$HVDM(x, y) = \sqrt{\sum_{i=1}^{m} hvdm_i(x_i + y_i)^2} \quad (12)$$

Treatment of nominal variables differs greatly from that of the heterogeneous functions described above. Instead of simple value matching, HVDM makes use of the class information to compute conditional probabilities. Function $hvdm_i$ is evaluated for the nominal values with the normalized and simplified Value Difference Metric [5] defined as:

$$vdm(x_i, y_i) = \sqrt{\sum_{c=1}^{C} \left| N_{i,x,c}/N_{i,x} - N_{i,y,c}/N_{i,y} \right|^2} \quad (13)$$

In equation (13) $N_{i,x,c}$ is the number of examples in the learning set L with value x_i and class c, $N_{i,x}$ is the number of examples in L that have value x_i, and C denotes the number of classes. Distance $hvdm_i$ for other than nominal attributes is $|x_i - y_i| / 4s_i$. The missing values are processed as in HEOM.

Table 1: The data sets used in the experiments. N, C, and V are the numbers of cases, classes, and variables, respectively. B, M, and O refer to the numbers of binary, multi-valued nominal, and ordinal variables. Q is the number of interval and ratio scaled variables

Data set	N	C	V	B	M	O	Q
Acute appendicitis	1333	13	16	9	4	1	2
Aids2	2297	2	5	2	1	0	2
Benign breast disease	200	2	11	1	1	2	7
Depression	294	2	14	7	3	2	2
Heart disease	303	2	13	3	3	1	6
Low birth weight	189	2	7	3	1	0	4
Muscular dystrophy	125	2	6	0	1	0	5
PONV (placebo)	141	2	14	8	4	0	2
PONV (prophylaxis)	166	2	14	8	4	0	2
Prostate cancer	380	2	7	2	1	1	3
Vertigo	914	10	38	11	1	10	16
VA lung cancer	137	2	6	2	1	0	3

Materials

Total of 12 heterogeneous medical data sets were used as test material (see Table 1). These data passed the following criteria:

1. At least one multi-valued nominal variable with statistically significant ($p<0.05$) dependency on the class variable as measured with Cramér's V [11].
2. Large enough for 10-fold cross-validation.
3. No missing values or few enough of them to allow the completion of the data.

The aids2 [12], depression [13], low birth weight [14], postoperative nausea and vomiting (PONV) [15], and VA lung cancer [12] data sets were complete. To avoid treatment of missing data in functions, variables having missing values were excluded from the PONV data. For the same reason, the missing values of acute appendicitis [16] (1.3%), benign breast disease [14] (3%), muscular dystrophy carriers [17] (0.7%), prostate cancer [14] (0.2%), and vertigo [8] (11%) data sets were imputed within classes with modes and medians. Missing data in the Cleveland heart disease data [18] (0.2%) were filled in with the modes and medians of the whole data.

Total of 564 and 3 examples were excluded from the aids2 and VA lung cancer data sets, because these cases had been censored before one year and 90 days, respectively.

Table 2: Accuracies (%) of proximity functions in 10-fold cross-validated 3-NN classification

Data set	GEM	ER	EUCL	GOW	HEOM	HVDM
Acute appendicitis	59	62	60	60	60	62
Aids2	59	56	58	57	58	58
Benign breast disease	74	71	75	75	73	72
Depression	79	83	80	83	82	81
Heart disease	83	78	83	79	79	81
Low birth weight	62	68	65	66	66	67
Muscular dystrophy	89	78	86	86	83	89
PONV (placebo)	92	90	92	89	89	92
PONV (prophylaxis)	90	90	90	91	90	89
Prostate cancer	71	70	73	71	71	74
Vertigo	73	80	75	78	78	79
VA lung cancer	66	69	73	69	68	71
Median	73	74	75	77	76	76

Results

The six proximity functions were evaluated experimentally using a 3-NN classifier and 10-fold cross-validation. Cross-validation is a technique where the data is divided into k disjoint sets of equal size and each set is used once as the test set and the union of other k-1 sets as the learning set. Prediction accuracy, the ratio of correctly classified test examples to all the test examples, and true-positive rate (TPR), the ratio of correctly classified

positive test examples to all the positive test examples, were used to evaluate the nearest neighbor classification.

The results were dependent, because the comparison required the same 10-fold partition to be used with all the proximity functions. Due to the small sample size the two-tailed Wilcoxon signed ranks test [11] was used instead of the paired t test to examine whether differences between the pairs of proximity functions were significant ($p<0.05$). Since the Wilcoxon signed ranks test was repeated 15 times, a Bonferroni correction [11] was also made to the probability associated with each test by multiplying it with the number of tests.

Table 2 shows the prediction accuracies of the 3-NN classifier using different proximity functions on the 12 data sets. Medians did not differ much, and neither the original nor the Bonferroni corrected p values of the paired tests were statistically significant.

Table 3: Median TPRs (%) of proximity functions in 10-fold cross-validated 3-NN classification

Data set	GEM	ER	EUCL	GOW	HEOM	HVDM
Acute appendicitis	11	12	14	8	8	14
Aids2	57	54	57	56	57	57
Benign breast disease	62	55	62	61	59	60
Depression	52	57	54	57	55	56
Heart disease	83	78	83	78	79	80
Low birth weight	52	57	57	57	56	59
Muscular dystrophy	85	68	76	79	75	84
PONV (placebo)	92	90	92	89	89	92
PONV (prophylaxis)	87	85	85	87	85	86
Prostate cancer	70	68	72	69	69	73
Vertigo	47	67	55	73	77	81
VA lung cancer	66	69	73	69	68	71
Median	64	67	67	67	66	73

Table 3 presents the median TPRs of the 3-NN classifier using different proximity functions on the 12 data sets. Medians are quite similar (64-67%), except that of HVDM (73%). The paired tests deemed most differences statistically insignificant. Significant differences were in favor of the HVDM function (HVDM > ER and HVDM > HEOM) at the adjusted a level. An additional significant difference at original a level was HVDM > GOW.

Discussion

Five heterogeneous functions and the Euclidean distance were compared by means of the prediction accuracies and TPRs of the 3-NN classifier. Four of the functions (HEOM, Gower, ER, and GEM) treat nominal values simply by comparing their equality. The fifth heterogeneous function (HVDM) is more sophisticated than the four value-matching-based functions. HVDM function evaluates distances between nominal values using conditional probabilities based on the class information.

The first objective of the comparison was to verify the intuitive hypothesis that heterogeneous functions would be more appropriate for computing proximities in the data sets described with a mixture of nominal and quantitative attributers than functions insensitive to scales.

There were no statistically significant differences in the prediction accuracies and the medians of TPRs of the Euclidean distance and the five heterogeneous functions. The results were clearly not in the expected direction and differed from the results of the previous studies. The Euclidean distance outperformed HEOM in [5] and [19], where it also was better than Gower's function. We also found earlier HVDM better than the Euclidean distance [19].

The second objective was to compare the value-matching-based heterogeneous functions with HVDM. The use of the class information is also a weakness, because the HVDM function can be applied only when the classes are known. Therefore, it was of interest to know how the ER and GEM functions, which we have not evaluated earlier, would perform compared to HVDM.

Results showed, as expected [19], that the HEOM produced lower TPRs than HVDM. Furthermore, at the original less stringent a level, HVDM outperformed Gower's function as in the previous study [19]. ER was significantly worse than HVDM at the adjusted a level. Although GEM and HVDM did not differ significantly, it is obvious, knowing the similar nature of the value-matching-based functions, that HVDM was better than the other heterogeneous functions of the present study. The ability to make use of the external information is clearly an important factor behind the good relative performance of HVDM.

Explaining why the relative performance of the Euclidean function was better than expected in the earlier [5,19] and present studies, is an interesting subject for future study. This requires additional, more controlled experiments. For example, normalization should be as similar as possible in the different functions. Future experiments should include data sets with larger numbers of multi-valued nominal attributes. To reach this goal, generation of synthetic data sets is possibly needed. Future work should also include additional classification methods utilizing proximity functions.

The limitation of this study was the small number of data sets. We selected the data more carefully than in the previous studies [5,19], where missing values were allowed to affect the results and the prediction capability of the nominal variables was not considered. Since it was difficult to meet all the requirements, the collection of data sets was not as large as we had wished. Accuracies have earlier [5,19] shown significant differences be-

tween functions, and it is likely that a larger collection would have produced more significant results.

To summarize, the Euclidean distance was as good as, and the HVDM function better than, the value-matching-based HEOM, Gower, ER, and GEM functions.

Acknowledgements

The authors wish to thank Erkki Pesonen, PhD, for providing the acute appendicitis data. The first and second authors gratefully acknowledge the financial support by the Academy of Finland.

References

[1] Everitt BS, Landau S, and Leese M. *Cluster Analysis*. 4th ed. London: Arnold, 2001.

[2] Sharma S. *Applied Multivariate Techniques*. New York: Wiley, 1996.

[3] Boberg J. *Cluster Analysis: A Mathematical Approach with Applications to Protein Structures*. Academic dissertation. Turku: Turku Centre for Computer Science, University of Turku, Finland, 1999.

[4] Mitchell TM. *Machine Learning*. New York: McGraw-Hill, 1997.

[5] Wilson DR and Martinez TR. Improved heterogeneous distance functions. J Artif Intell Res 1997: 6: 1-34.

[6] Agresti A. *An Introduction to Categorical Data Analysis*. New York: Wiley, 1996.

[7] Laurikkala J. *Knowledge Discovery for Female Urinary Incontinence Expert System*. Academic dissertation. Tampere: Department of Computer Sciences, University of Tampere, Finland, 2001.

[8] Viikki K. *Machine Learning on Otoneurological Data: Decision Trees for Vertigo Diseases*. Academic dissertation. Tampere: Department of Computer Sciences, University of Tampere, Finland, 2002.

[9] Estabrook GF and Rogers DJ. A general method of taxonomic description for a computed similarity measure. *BioScience* 1966: 16, Nov: 789-793.

[10]Ichino M and Yaguchi H. Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Trans Syst Man Cybern* 1994: 24, 4 (Apr): 698-708.

[11]Pett MA. *Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions*. Thousand Oaks: SAGE Publications, 1997.

[12]Venables WN and Ripley BD. *Modern Applied Statistics with S-PLUS*. New York: Springer, 1999.

[13]Afifi AA and Clark V. *Computer-Aided Multivariate Analysis*. London: Chapman & Hall, 1996.

[14]Hosmer DW and Lemeshow S. *Applied Logistic Regression*. New York: Wiley, 2000.

[15]Viikki K, Juhola M, Pyykkö I, and Honkavaara P. Evaluating training data suitability for decision tree induction. *J Med Syst* 2001: 25: 133-144.

[16]Pesonen E, Ikonen J, Juhola M, and Eskelinen M. Parameters for a knowledge base for acute appendicitis. *Methods Inf Med* 1994: 33: 220-226.

[17]Percy M. Procedures for the detection of muscular dystrophy carriers. In: Hand DJ, Daly F, Lunn AD, McConway KJ, and Ostrowski E, eds. *A Handbook of Small Data Sets*. London: Chapman & Hall, 1996; pp. 223-228.

[18]Blake CL and Merz CJ. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science, 1998. [http://www.ics.uci.edu/~mlearn/MLRepository.html]

[19]Laurikkala J and Juhola M. Nearest neighbour classification with heterogeneous proximity functions. In: Hasman A, Blobel B, Dudeck J, Engelbrecht R, Gell G, and Prokosch H-U, eds. *Medical Infobahn for Europe: Proceedings of MIE2000 and GMDS2000*. Studies in Health Technology and Informatics, vol. 77. Amsterdam: IOS Press, 2000; pp. 753-757.

Address for correspondence

Janne Lumijärvi,

Department of Computer Sciences, University of Tampere, Kanslerinrinne 1, FIN-33014 University of Tampere, Finland,

Janne.Lumijarvi@uta.fi