Comparative Study of Distance Functions for Nearest Neighbors

Janett Walters-Williams
School of Computing and Information Technology
University of Technology, Jamaica
Kingston 6, Jamaica W.I.
jwalters@utech.edu.jm

Yan Li
Department of Mathematics and Computing
Centre for Systems Biology
University of Southern Queensland
Toowoomba, Australia
liyan@usq.edu.au

***Abstract* - Many learning algorithms rely on distance metrics to receive their input data. Research has shown that these metrics can improve the performance of these algorithms. Over the years an often popular function is the Euclidean function. In this paper, we investigate a number of different metrics proposed by different communities, including Mahalanobis, Euclidean, Kullback-Leibler and Hamming distance. Overall, the best-performing method is the Mahalanobis distance metric.**

Keywords
Kullback-Leibler distance, Euclidean distance, Mahalanobis distance, Manhattan distance, Hamming distance, Minkowski distance, Nearest Neighbor.

I. INTRODUCTION

Nearest Neighbor algorithms are examples of instance-based learning which simply retain the entire training set during learning. Unlike other common classifiers, these algorithms do not build a classifier in advance. When a new sample arrives, the algorithm finds the neighbors nearest to the new sample from the training space based on a distance metric.

Distance functions, or distance metric learning functions are to learn distance metrics for input data from a given collection of pair or similar/dissimilar points that preserves the distance relation among the training data. This paper focuses on local, supervised distance metric learning useful for K nearest neighbor (KNN) classifiers. We are interested in answering the following question: Which distance function should be selected to produce a more accurate output when applied to KNNs? We seek the answer from theoretical analysis and experimental results. Research has shown that Euclidean distance is the mostly widely used function in practice [14, 17, 18], although Cover and Hart [5] state that any function can be used. Choosing the correct function however, ultimately dictates the success or failure of any learning algorithm.

In this paper we focus on distance metrics from two classes: (1) metrics which do not involve any normalization of the components - Euclidean, Mahalanobis, Manhattan (city block), Hamming and Minkowski, and (2) entropy based measures namely Kullback-Leibler, the most widely used theoretical metric [10]. We propose to compare the performance of these six distance metrics when applied to Nearest Neighbor Algorithms. We compute the confusion matrix from each function which is analyzed. We found that the expected performance of each is not the final result. From theoretical analysis and experimental results, we found that there are more similarities among most of the six functions than differences.

The paper is organized as follows. In Section 2, the six distance functions are described. Section 3 highlights work already done in the area. In Section 4 the theoretical analysis and experimental results for the six distances are presented. Finally section 5 discusses the conclusion.

II. DISTANCE

To define a distance is equivalent to defining rules to assign positive numbers between *pairs* of objects. Let, therefore, *a, b*, and *c* be three vectors with *j* elements each. A distance is a function which associates to any pair of vectors a real positive number, denoted $d(\mathbf{a},\mathbf{b})$, which has the following properties [1]:-

$$d(a,a) = 0 \quad (1)$$

K. Elleithy (ed.), *Advanced Techniques in Computing Sciences and Software Engineering*,
DOI 10.1007/978-90-481-3660-5_14, © Springer Science+Business Media B.V. 2010

$$d(a,b) = d(b,a)$$
$$d(a,b) \leq d(a,c)+d(c,b)$$
(1)

There are many learning systems that depend upon a good distance function to be successful. The following defines the six distance metrics used in this paper.

A. *Kullback-Leibler Distance*

The Kullback-Leibler distance is a natural distance function from a true probability distribution *p* to a target probability *q*. It is also known as relative or mutual entropy and is defined as

$$KL(p,q) = \sum_{i=1}^{n} p_i \times \log_2\left(\frac{p_i}{q_i}\right) \tag{2}$$

where *n* is the number of levels of the variables.

B. *Euclidean Distance*

The Euclidean distance computes the real straight line distance between two points, i.e. it measures the 'as-the-crow-flies' distance. If $p = \{p_1, \ldots, p_n\}$ and $q=\{q_1, \ldots, q_n\}$ the Euclidean distance is defined as:

$$EUD(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{3}$$

C. *Manhattan Distance*

The Manhattan distance is also known as the "absolute value" or city block distance. It computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. It is the sum of the differences of their corresponding components. The Manhattan distance is defined as:

$$MN(p,q) = \sum_{i=1}^{n} | p_i - q_i | \tag{4}$$

D. *Hamming Distance*

Hamming distance or the symmetric difference distance is a set of operations which associates to two sets a new set made of the elements of these sets that belong to only one of them. Elements that belong to both sets are excluded. It gives the number of the elements of the symmetric difference set. It is defined as:

$$HAM(p,q) = \sum_{i=1}^{n} | p_i - q_i | \tag{5}$$

if *p* and *q* are vectors consisting of zeroes and ones. Hamming distance is equal to the number of positions where the bit patterns are different.

E. *Mahalanobis Distance*

The Mahalanobis distance is based on the correlations between variables. It is defined as:

$$MD(p,q) = \sqrt{(p_i - q_i)^T V^{-1} (p_i - q_i)}$$
(6)

where V is the covariance matrix of $A_1..A_m$ and A_j is the vector of values for attribute j occurring in the training set instances *1..n*.

F. *Minkowski Distance*

The Minkowski distance or *p*-distance between two strings is the geometric distance between two inputs and uses a variable scaling factor, *r*. It is widely used for measuring similarity between objects (e.g., images) and is defined as:

$$\mathrm{L}_m = \left(\sum_{i=1}^{n} | p_i - q_i |^r\right)^{\frac{1}{r}}$$
(7)

III. RELATED WORK

Since 1981 researchers have tried to compare different similarity measures. Noreault et al. [11] looked at evaluating the performance of measures, empirically. Further Jones and Furnas [9] studied several similarity measures in the field of information retrieval. In particular, they performed a geometric analysis on continuous measures in order to reveal important differences which would affect retrieval performance. Further comparative studies were done by Zwick et al. [19] focusing on Fuzzy sets.

A detailed study of heterogeneous distance functions (for data with categorical and continuous attributes) was carried out by Wilson and Martinez[16]. They did this for instance based learning. Their study was based on a supervised approach where each data instance had class information in addition to a set of categorical/continuous attributes.

The latest set of research has been done by Qian et al.[12] who compared the Euclidean and Cosine Angle distances for nearest neighbor queries in high dimensional data spaces and Boriah et al. [4] who looked at the performance of a variety of similarity measures in the context of a specific data mining task: outlier detection.

IV. COMPARISON ANALYSIS

A. *Theoretical Analysis*

Euclidean and Mahalanobis

The Euclidean norm of *p* yields the equation of a spheroid. This means that all components of an observation *p* contribute equally to the Euclidean distance of *x* from the center. Taking variability of that variable into account we get the distance between *p* and *q* in Euclidean as:

$$ED(p,q)=\sqrt{\left(\frac{p_i-q_i}{s_i}\right)^2+..+\left(\frac{p_n-q_n}{s_n}\right)^2}=\sqrt{(p-q)^T D^{-1}(p-q)} \tag{8}$$

where $D = diag(s_i^2..s_n^2)$.

We then take the correlation between variables into account. To do this the axes of ellipsoid are used to reflect this correlation. This is obtained by allowing the axes of the ellipsoid at constant distance to rotate. This yields the Mahalanobis distance (fig. 1). Thus if *V* in (6) becomes a *d x d* identity matrix the Mahalanobis distance is defined as the Euclidean distance. If *V* is diagonal, then the resulting distance measure is called the normalized Euclidean distance and is defined as:

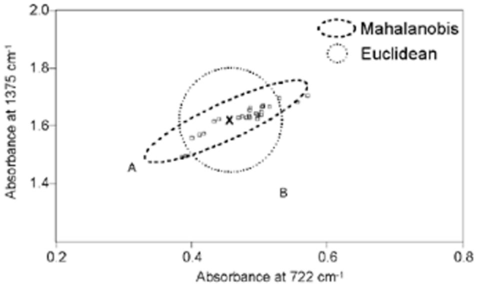


Fig. 1. Conversion of Euclidean to Mahalanobis

$$MD(p,q)=\sqrt{\sum_{i=1}^{n}\frac{(p_i-q_i)^2}{\sigma_i^2}} \tag{9}$$

where σ_i is the standard deviation of the x_i over the sample set. Mahalanobis is different from Euclidean because it takes into account the correlations of the data sets and is not dependent on the scale of measurements. It therefore generalizes the Euclidean function [6].

The Minkowski Relation

The degree *r* in the Minkowski distance (7) can take any number. When *r = 1* the distance function is called the Manhattan distance. If the vectors, when *r = 1*, are binary numbers, the distance becomes the Hamming distance. When *r* is equal to 2, we obtain the usual Euclidean distance. Euclidean and Manhattan are therefore apart of the Minkowski family of distance metrics.

In this family the higher the value of *r*, the greater the importance given to large differences. Thus, when *r = 1* or L_1 there is equal importance to all differences while when *r = 2* or L_2 the distance metric takes into account only that component for which the difference is maximum. These are Manhattan and Euclidean respectively. When calculating Manhattan also deals with the sum of distance along each dimension while Euclidean corresponds to the length of the shortest path between two points as shown in fig. 2.

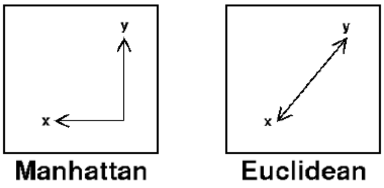


Fig. 2. Difference between Manhattan and Euclidean distances

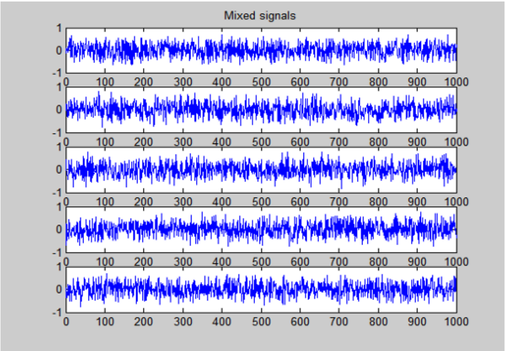


Fig. 3. Sample Signal Set

B. Experimental Analysis

In order to do the study effectively data was collected for analysis. This data, taken from different sites are of two types - real and artificial. All data is comprised of EEG data signals. The artificial data is made of six different data sets, each containing at least 1,000 points per vector (fig 3). The data sets are of two types – mixed with noise and independent from noise. These were taken from the RADICAL ICA algorithm site http://www.cs.umass.edu/~elm/ICA/.

Real data sets comprised EEG signals from both human and animals. These data have been acquired using the Neuroscan or Neurofax software. The human data set is a collection of 32-channel data from 14 subjects (7 males, 7 females) who performed a go-nogo categorization task and a go-no recognition task on natural photographs presented very briefly (20 ms). Each subject responded to a total of 2500 trials. The data is CZ referenced and is sampled at 1000 Hz. This data set can be found at http://sccn.ucsd.edu/~arno/fam2data/publicly_available_EEG_data.html.

Each data set passes through the *k*-nearest neighbor code. This produces a confusion matrix of the set and a classification rate (%). The confusion matrix is used to compute the classification accuracy and to identify misclassified areas. The Friedman test is then preformed on each resulting matrix and the results are passed through a multiple comparison test.

Performance on Nearest Neighbor

Each distance metric is used to determine a resulting nearest neighbor matrix. Each produces such a matrix but at different processing times. Fig. 4 shows results of processing times for all six. We find that as the data sets increase the processing times also increase, however, the rate remains the same. Results also show that the performance rate for each of L_m distances are relatively the same with Euclidean distance having the largest rate.

When the metrics are calculated it is the findings that the matrices that contain the vectors containing the distances between each pair of observations in data matrices are the same for the Minkowski distance and the Euclidean distance metrics. The others except for the Hamming, which produces matrices containing bits, differ.

The Nearest Neighbor code generates a confusion matrix for each distance metric calculation. If the data set is *M x N* this matrix is *N x N* in size showing in Table 1 with the 'true' class in rows and the 'predicted' class in the columns. The diagonal elements represent correctly classified compounds while the cross-diagonal elements represent misclassified compounds. The table also shows the accuracy of the classifier as the percentage of correctly classified compounds in a given class divided by the total number of compounds in that class. Table 1 shows the first ten rows in the first column of a confusion matrix based on the Hamming Distance.

The classification rate is calculated as

$$C_{rate} = \sum diag\left(c_{mat}\right)x100/M \qquad (10)$$

where M is the number of rows and c_{mat} is the confusion matrix. It was found that the rate varied and ranged from -3.7899 in Mahalanobis to 47.9836 in Euclidean, using the 2D mixed data set. It was also found that Kullback-Leibler did not produce the lowest rate; it was13.2525, one of the highest rates.

TABLE 1

FIRST 10 ROWS-CONFUSION MATRIX (2 TYPES OF DATA SET)

2 row independent	2 row mixed
0.0938	-0.0758
0.5272	0.6028
0.1873	0.2171
0.6460	0.6357
-0.6883	-0.6491
0.2188	0.1521
0.2063	0.0671
-0.0467	-0.0509
0.6408	0.5912
0.0006	0.0218

The Friedman Test

The Friedman test is frequently called a two-way analysis on ranks and is used to detect differences in treatments across multiple test attempts. It is at the same time a generalization of the Sign-Test and the Spearman Rank Correlation Test and test models the ratings of *n* (rows) judges on *k* (columns) "treatments". The test is used to test if the means of the distance functions are totally matched when the distribution of the underlying population is not specified.

The hypothesis being tested is that all the methods have equal mean total matches, and the alternative hypothesis is that all methods do not have equal mean total matches. It is our findings that of the six functions Manhattan and Kullback-Leibler produced slightly higher means in each data set (fig. 5). It also shows that as the data sets increase in the number of vectors the error in each increased, however it showed that it increased more in Euclidean, Manhattan and Hamming as the dimensions increased (fig. 4).

It was also seen that the results changed when the data set types changed. Figure 7 shows the Friedman Test table of a independent data set while figure 6 shows a simpler set that was mixed with noise. It can be seen that the probability of having a Chi-square and the error calculated on the mean increase when an independent dataset is used

Multiple Comparison Test

Once the Friedman test is completed the resulting statistics are used in the Multiple Comparison Test. This test is done using the Tukey-Kramer Method. This method is chosen over Scheffé, Bonferoni and Sidák because it produces smaller critical values and it controls the experiment wise error rate at approximation ε very well [7].

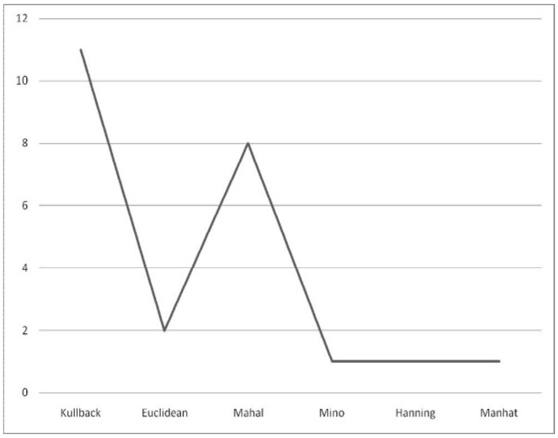


Fig. 4. Performance Rate of The Six Functions

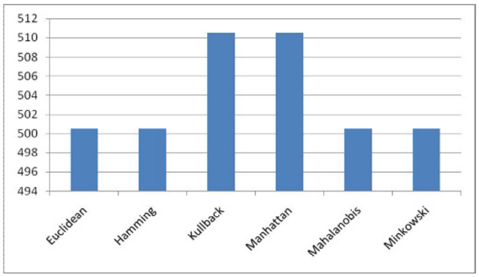


Fig. 5. Mean Values for 2D mixed data set

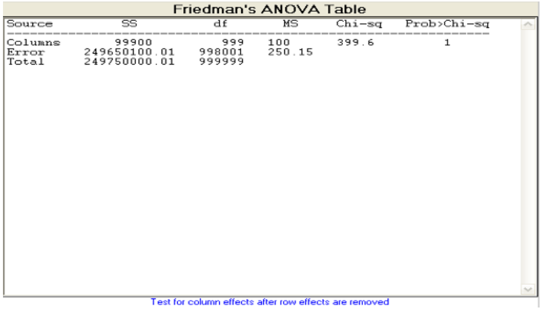
The hypothesis being tested is based on the results from the Friedman Test. We wish to prove that since there are differences in the mean for the Friedman Test it will be the same behavior in the Comparison Test.

In this test the minimum significant difference (MSD) is calculated for each pair of means. If the observed difference between a pair of means is greater than the MSD, the pair of means is significantly different. It was found that for all distance function, except Euclidean, all the mean column ranks were significantly different. For Euclidean there were only significant difference in a few columns.

When the error rates are examined the Kullback-Leibler is considered to have the worst. It is found that the following is the order:- *Mahalanobis < Minkowski family < Kullback-Leibler*. Mahalanobis has the best since the error rate is controlled.

Discussion

Most nearest neighbor algorithms are based on the Euclidean distance function. In this paper, we examine six known distance metrics. Is it really necessary to use the Euclidean function? Our results indicate that Euclidean does not have the fastest performance rate or the best error rate. Overall Mahalanobis metric has the best performance when applied to nearest neighbor. It has (1) a low performance rate; (2) performs well when data is controlled; (3) has a low classification rate; and, (4) its mean value is one which has a low increase rate as the number of vectors increases. Research has also shown that based on Maximum Likelihood criteria Euclidean and Manhattan are proven to be optimal distances for Gaussian and Exponential data, respectively [9]. Mahalanobis on the other hand, is useful for both Gaussian and non-Gaussian data. It is also scale-invariant, i.e. not dependent on the scale of measurements which makes it approximate for applications with different types of measurements.

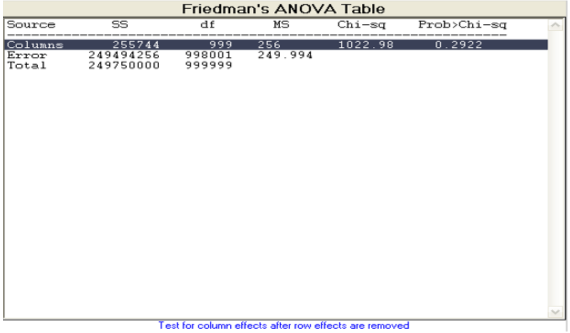


Friedman's ANOVA Table

```
Source          SS          df      MS      Chi-sq    Prob>Chi-sq
-----------------------------------------------------------------
Columns       99900          999    100      399.6           1
Error     249650100.01    998001    250.15
Total     249750000.01    999999
```

Test for column effects after row effects are removed

Fig. 6. Friedman Test on a Kullback-Leibler mixed data set



Friedman's ANOVA Table

```
Source        SS         df       MS        Chi-sq     Prob>Chi-sq
------------------------------------------------------------------
Columns      255744        999   256        1022.98       0.2922
Error     249494256     998001   249.994
Total     249750000     999999
```

Test for column effects after row effects are removed

Fig. 7. Friedman Test on a Kullback-Leibler independent data set

V. CONCLUSION

Over the years Euclidean has been the distance metric of choice by most researchers, however we have observed from our experiments that Mahalanobis has the best performance of the six metrics studied. The Minkowski family of distances is not suitable for all applications. So why is Euclidean the distance of choice? This maybe because of (1) the ease of implementing the Euclidean distance and (2) researchers tend to assume data to be

Gaussian in distribution. Although Mahalanobis is the best of the six researchers may choose their distance metric based on their personal choice and the size and type of the datasets been used. For example if one does not have any prior knowledge the Euclidean function is usually recommended. If there is the need to capitalize on statistical regularities in data that maybe estimated from a large training set then Mahalanobis is best.

REFERENCES

[1] Abdi, H., *Encyclopedia of Measurement and Statistics*, 2007

[2] Bar-Hillel, A., *Learning from Weak Representations using Distance Functions and Generative Models*, Ph.D. Thesis, Hebrew University of Jerusalem, 2006.

[3] Beitao L., Chang, E., Wu, C., DPF – A Perceptual Distance Function for Image Retrieval. In *Proceedings of the IEEE conference on Image Processing,* Sept 2002.

[4] Boriah, S., Chandola, V. Kumar, V. Similarity Measures for Categorical Data: A Comparative Evaluation, In *Proceedings of the 2008 Society of Industrial and Applied Mathematics (SIAM) International Conference on Data Mining.,* pp.23-254, 2008.

[5] Cover, T.M., Hart, P.E., Nearest Neighbor Pattern Classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory, 13,* pp. 21-271, Jan. 1967.

[6] Davis, J.V., Kulis, B., Jain, P., Sra, S., Dhillon, I. S., Information-Theoretic Metric Learning, In the *Proceedings of the 24th International Conference on Machine Learning*, 2007.

[7] Griffiths, R. Multiple Comparison Methods for Data Review of Census for Agriculture Press Releases, In the *Proceedings of the Survey Research Methods Section of the American Statistical Association,* 1992.

[8] Jensen, D.D., Cohen, P.R., Multiple Comparisons in Induction Algorithms, *Klumer Academic Publishers*, pp. 1-33, 2002.

[9] Jones, W.P., Furnas, G.W., Pictures of Relevance: A Geometric Analysis of similarity Measures, *Journal of American Society of Information Science* vol. 38, issue 6, pp. 420-442, 1987.

[10] Kamichety, H.M., Natarajan, P., Rakshit S., An Empirical Framework to Evaluate Performance of Dissimilarity Metrics in Content Based Image Retrieval Systems, *Technical Report, Center of Artificial Intelligence and Robotics, Bangalore,* 2002.

[11] Noreault, T., McGill, M., Koll, M.B., A Performance Evaluation of Similarity Measures, Document Term Weighting Schemes and Representations in a Boolean Environment, In *SIGIR '80 Proceedings of the 3rd Annual ACM Conference on Research and Development in Information Retrieval*, 76, 1981.

[12] Qian, G., Sural, S., Gu, Y., Pramanik, S., Similarity Between Euclidean and Cosine Angle Distance of Nearest Neighbor Queries, In *the Proceedings of the ACM Symposium on Applied Computing,* 2004.

[13] Tumminello, M., Lillo, F., Mantegna, R.N., Kulback-Leiber as a Measure of the Information Filtered from Multivariate Data, Physical Review E. 76, 031123 , 2007.

[14] Weinberger, K.Q., Blitzer, J., Saul, L.K., Distance Metric Learning for Large Margin Nearest Neighbor Classification, *Advances in Neural Information Processing Systems*, MIT Press, 2006.

[15] Weinberger, K. Q., Saul, L. K., Fast Solvers and Efficient Implementations for Distance Metric Learning, Under Review by the I*nternational Conference on Machine Learning (ICML)*, 2007.

[16] Wilson, D.R., Martinez, T.R., Improved Heterogeneous Distance Functions, J*ournal of Artificial Intelligence Research (JAIR),* vol. 6, issue 1, pp. 1-34, 1997.

[17] Wilson, D.R., *Advances in Instance-Based Learning Algorithms*, Ph.D. Thesis, Brigham Young University, 1997.

[18] Wölfel, M., Ekenel,H. K., Feature Weighted Mahalanobis Distance: Improved Robustness for Gaussian Classifiers, In the Proceedings of the 13th European Signal Processing Conference (EUSIPCO 2005), Sept 2005.

[19] Zwick, R., Carlstein, E., Budescu, D.V., *Measures of Similarity among Fuzzy Concepts: A Comparative Analysis, International Journal of Approximate Reasoning* 1, 2, pp. 221-242, 1987.