How to quantitatively compare data dissimilarities for unsupervised machine learning?

Bassam Mokbel[1*], Sebastian Gross[2], Markus Lux[1],
Niels Pinkwart[2], Barbara Hammer[1]

[1]) CITEC centre of excellence, Bielefeld University, Germany
[2]) Computer Science Institute, Clausthal University of Technology, Germany
[*]) E-Mail: bmokbel@techfak.uni-bielefeld.de

Abstract

For complex data sets, the pairwise similarity or dissimilarity of data often serves as the interface of the application scenario to the machine learning tool. Hence, the final result of training is severely influenced by the choice of the dissimilarity measure. While dissimilarity measures for supervised settings can eventually be compared by the classification error, the situation is less clear in unsupervised domains where a clear objective is lacking. The question occurs, how to compare dissimilarity measures and their influence on the final result in such cases. In this contribution, we propose to use a recent quantitative measure introduced in the context of unsupervised dimensionality reduction, to compare whether and on which scale dissimilarities coincide for an unsupervised learning task. Essentially, the measure evaluates in how far neighborhood relations are preserved if evaluated based on rankings, this way achieving a robustness of the measure against scaling of data. Apart from a global comparison, local versions allow to highlight regions of the data where two dissimilarity measures induce the same results.

1 Introduction

In many application areas, data are becoming more and more complex such that a representation of data as finite-dimensional vectors and their treatment in terms of the Euclidean distance or norm is no longer appropriate. Examples include structured data such as bioinformatics sequences, graphs, or tree structures as they occur in linguistics, time series data, functional data arising in mass spectrometry, relational data stored in relational databases, etc. In consequence, a variety of techniques has been developed to extend powerful statistical machine learning tools towards non-vectorial data such as kernel methods using structure kernels, recursive and graph networks, functional methods, relational approaches, and similar [9, 12, 5, 27, 6, 26, 10, 11]. One very prominent way to extend statistical machine learning tools is offered by the choice of problem-specific measures of data proximity, which can often directly be used in machine learning tools based on similarities, dissimilarities, distances, or kernels. The latter include popular techniques such as the support vector machine, other kernel approaches such as kernel self-organizing maps or kernel linear discriminant analysis, or distance-based approaches such as k-nearest neighbor techniques or distance-based clustering or visualization, see e.g. [23]. Here, we are interested in dissimilarity-based approaches in general, treating metric distances as a special case of (non-metric) dissimilarities.

With the emergence of more and more complex data structures, several dedicated structure metrics have become popular. Classical examples include alignment for sequences in bioinformatics [22], shape distances [21], or measures motivated by information theory [4]. Often, there exists more than one generic possibility to encode and compare the given data. In addition, dissimilarity measures often come with parameters, the choice of which is not clear a priori. Hence, the question occurs how to choose an appropriate metric in a given setting. More generally, how can we decide whether a change of the metric or its parameters changes the data representation which is relevant for the subsequent machine learning task? Are there possibilities to compare whether and, if so, in which regions two metrics differ if used for machine learning?

Many approaches which are used in machine learning for structures have been proposed in the supervised domain. Here, a clear objective of the task is given by the classification or regression error. Therefore, it is possible to evaluate the difference of dissimilarities by comparing the classification error obtained when using these different data representations. A few extensive comparisons how different dissimilarities influence the outcome have been conducted; see, e.g. [18] for the performance of different dissimilarities for content-based image retrieval, [19] for an according study in the symbolic domain, [2] for the comparison of distances for probability measures, or [3] for the performance of classifiers on differently preprocessed dissimilarities to arrive at a valid kernel.

The situation is less clear when dealing with unsupervised domains. Unsupervised learning is essentially ill-posed and the final objective depends on expert evaluation. The primary mathematical goal is often to cluster or visualize data, such that an underlying structure becomes apparent. Quite a few approaches for unsupervised learning for structures based on general dissimilarities have been proposed in the past: kernel clustering techniques such as kernel self-organizing maps (SOM) or kernel neural gas (NG) [34, 24] or relational clustering such as proposed for fuzzy-k-means, SOM, NG, or the generative topographic mapping (GTM) [13, 7, 8]. Further, many state-of-the art nonlinear visualization techniques such as t-distributed stochastic neighbor embedding are based on pairwise dissimilarities rather than vectors [31, 15].

In this contribution, we will investigate how to compare dissimilarity measures with regard to their influence on unsupervised machine learning tasks, and discuss different possibilities in Sec. 2. Thereafter, we will focus on a principled approach independent of the chosen machine learning technique, rather we will propose a framework which compares two dissimilarity measures based on their induced neighborhood structure in Sec. 3. This way, it is possible to decide prior to learning whether and, if so, in which regions two different dissimilarity measures or different choices of parameters lead to different results, which we will demonstrate on examples in Sec. 4 and 5, concluding with a discussion in Sec. 6.

2 How to compare dissimilarity measures?

We assume that data $\mathbf{x}_i$ are sampled from some underlying data space. These data are input to an unsupervised machine learning algorithm by means of pairwise comparisons $d_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)$. These values constitute dissimilarities, as given by the squared Euclidean distance, provided data are vectorial. We assume that d refers to a general dissimilarity measure for which Euclidean properties are not necessarily guaranteed, maybe even the constraints of a metric are violated. Note, that the dual situation, similarities or kernels, can easily be transferred to this setting, see [23].

Interestingly, albeit the chosen dissimilarity structure crucially determines the output of any machine learning algorithm based thereon, no framework of how to compare different dissimilarities for unsupervised domains is commonly accepted in the literature. The question occurs what is the relevant information contained in a dissimilarity which guides the output of such an algorithm? Interestingly, even slight changes of the dissimilarity such as a shift can severely influence the result of an unsupervised algorithm, as shown in [8]. Apart from generic mathematical considerations, indications for the answer to this question may be taken from attempts to formalize axioms for unsupervised learning [1, 17, 33, 14]. Here, guidelines such as scale-invariance, rank-invariance, or information retrieval perspectives are formalized. Now, we formalize and discuss different possibilities how to compare dissimilarity measures. We assume that pairwise dissimilarities d^1_{ij} and d^2_{ij}, which are to be compared, are given.

Matrix comparison:

The pairwise dissimilarities d^1_{ij} and d^2_{ij} give rise to two square matrices D_1 and D_2 respectively, which could directly be compared using some matrix norm. This possibility, however, is immediately ruled out when considering standard axioms for clustering [1], for example. One natural assumption is scale-invariance of the unsupervised learning algorithm. Scaling the matrix, however, does affect the resulting matrix norm. More generally, virtually any matrix norm severely depends on specific numeric choices of the representation rather than the global properties of the data.

Induced topology:

An alternative measure which ignores numerical details but focuses on basic structures could be connected to the mathematical set-theoretic topology of a data space. Every distance measure induces a topology. Hence, it is possible to compare whether the topological structure induced by two metrics is equivalent. In mathematics, two metrics are called topologically equivalent if the inequality $c \cdot d^1(\mathbf{x}_i, \mathbf{x}_j) \leq d^2(\mathbf{x}_i, \mathbf{x}_j) \leq c' \cdot d^1(\mathbf{x}_i, \mathbf{x}_j)$ holds for all $\mathbf{x}_i, \mathbf{x}_j$ for some constants $0 < c \leq c'$, since they induce the same topology in this case. It can easily be shown that any two metrics in a finite-dimensional real vector space are topologically equivalent. However, this observation shows that this notion is not appropriate to compare metrics with respect to their use for unsupervised learning: topologically equivalent metrics such as the standard Euclidean metric and the maximum-norm yield qualitatively different clusters in practical applications, as we will demonstrate in an example in Sec. 4.

Rank preservation:

One axiom of clustering, as formalized in [1], is the invariance to rank-preserving distortions. Indeed, many clustering or visualization techniques take into account the ranks induced by the given dissimilarity measure only, this way achieving a high robustness of the results. Examples include algorithms based on winner-takes-all schemes or extensions such as vector quantization, NG, SOM, or similar approaches. Also, many visualization techniques try to preserve local neighborhoods as measured by the rank of data. How can rank-preservation be evaluated quantitatively? One way is to transform the matrices D_1 and D_2 into rank matrices, i.e. matrices which contain permutations of the numbers $\{0, \ldots, N-1\}$, N being the number of data points. Then, these two matrices could be compared by their column-wise correlation. However, usually the preservation of all ranks is not as critical as the preservation of a local neighborhood for most machine learning tools, such that different scales of the neighborhood size should be taken into account. In Sec. 3, we will explain the co-ranking framework which can be seen as a way to observe this rank-preservation property according to various neighborhood sizes of interest.

Information retrieval based comparison:

Information retrieval constitutes a typical application area for unsupervised learning. Therefore a comparison of dissimilarity measures based on this perspective would be interesting. Assume a user queries a database for the neighborhood of $\mathbf{x}_i$. What is the precision/recall, if d^2 is used instead of d^1? When defining the notion of neighborhood as the K nearest neighbors, precision and recall for a query $\mathbf{x}_i$ are both given by the term $|\{\mathbf{x}_j | d^1(\mathbf{x}_i, \mathbf{x}_j) \leq K \wedge d^2(\mathbf{x}_i, \mathbf{x}_j) \leq K\}|$ normalized by K. Summing over all $\mathbf{x}_i$ and dividing by N yields an average of all possible queries. In fact, this instantiation of a quality measure coincides with an evaluation within the co-ranking framework which will be introduced in Sec. 3.

3 The co-ranking framework

One very prominent tool in unsupervised learning is given by nonlinear dimensionality reduction and visualization [15]. Although many of the most relevant nonlinear dimensionality reduction methods have been proposed in the last years only, the question of what are appropriate quantitative evaluation tools is still widely unanswered. Interestingly, as reported in [32], a high percentage of publications on data visualization evaluates results in terms of visual impression only – about 40% out of 69 papers referenced in [32] did not use any quantitative evaluation criterion. In the last years, a few formal mathematical evaluation measures of dimensionality reduction have been proposed in the literature. We argue that one of these measures, the co-ranking framework proposed in [14, 16], is directly suitable as a highly flexible and generic tool to evaluate the preservation of pairwise relationships in different dissimilarity measures.

In this section, we give a short overview about the co-ranking framework. Assume points $\mathbf{x}_i$ are mapped to projections $\mathbf{y}_i$ using some dimensionality reduction technique. The co-ranking framework essentially evaluates, in how far neighborhoods in the original space and the projection space correspond to each other. Let δ_{ij} be the distance of $\mathbf{x}_i$ and $\mathbf{x}_j$ and d_{ij} be the distance of $\mathbf{y}_i$ and $\mathbf{y}_j$. The rank of $\mathbf{x}_j$ with respect to $\mathbf{x}_i$ is given by $\rho_{ij} = |\{k \mid \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } k < j)\}|$. Analogously, the rank of r_{ij} for the projections can be defined based on d_{ij}. The co-ranking matrix Q [14] is defined by $Q_{kl} = |\{(i,j) \mid \rho_{ij} = k \text{ and } r_{ij} = l\}|$. Errors of a dimensionality reduction correspond to rank errors, i.e. off-diagonal entries in this matrix. Usually, the focus of dimensionality reduction is on the preservation of local relationships. In [14], an intuitive measure of rank-preservation has been proposed, the *Quality*

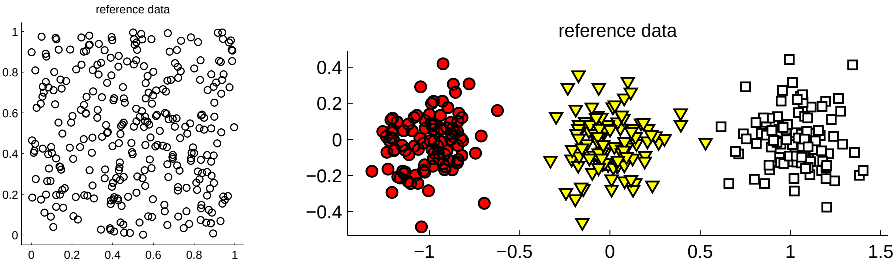
$$Q_{\mathrm{NX}}(K) = \frac{1}{KN} \sum_{k=1}^{K} \sum_{l=1}^{K} Q_{kl}.$$

where N denotes the number of points. This summarizes all 'benevolent' points which change their rank only within a fixed neighborhood K. Essentially, it is the average ratio of all points which stay in a K-neighborhood in the original and the projection space. To get an overall impression of the quality in different neighborhood regimes, usually a curve is plotted for a all possible K or a range thereof. A qualitatively good visualization w.r.t. all K-neighborhoods corresponds to the value $Q_{\mathrm{NX}}(K)$ approaching 1. Interestingly, this framework can be linked to an information theoretic point of view as specified in [33] and it subsumes several previous evaluation criteria, see [14, 20]. It is possible to extend this framework to a point-wise evaluation as introduced in [20]. Here, all neighborhood sizes are considered for one fixed point $\mathbf{x}_i$ only, leading to the local quality curve $Q_{\mathrm{NX}}^{\mathbf{x}_i}(K) = \frac{1}{KN} \sum_{k=1}^{K} \sum_{l=1}^{K} Q_{kl}(\mathbf{x}_i)$. Obviously, $Q_{\mathrm{NX}}(K) = \sum_{\mathbf{x}_i} Q_{\mathrm{NX}}^{\mathbf{x}_i}(K)$.

How can this technique be used to compare two dissimilarities? Since $Q_{\mathrm{NX}}(K)$ essentially evaluates in how far a rank-neighborhood induced by δ_{ij} coincides with a rank-neighborhood induced by d_{ij}, we can directly apply this measurement to two given dissimilarity measures d^1 and d^2, and obtain a quantitative statement about the rank-preservation of d^2 given d^1. Since $Q_{\mathrm{NX}}(K)$ is symmetric, the ordering of the dissimilarities is not important.

4 Comparison of metrics for the Euclidean vector space

We start with an illustrative example which shows that the measure $Q_{\mathrm{NX}}(K)$ allows to identify situations where dissimilarities induce similar/dissimilar results. We restrict to the two-dimensional Euclidean vector space where data are distributed uniformly or in clustered form, respectively, see Fig. 1. For these data, we compare the Euclidean distance to the L_k norm, with $k \in \{1, 3, 6\}$ as well as the maximum-norm as the limit case. We can see the effect of these choices by using a metric multidimensional scaling (MDS) to project the data to the



(a) Uniform data

(b) Clustered data

Figure 1: Original data in the two-dimensional plane with uniform distribution (a) or clustered distribution (b).

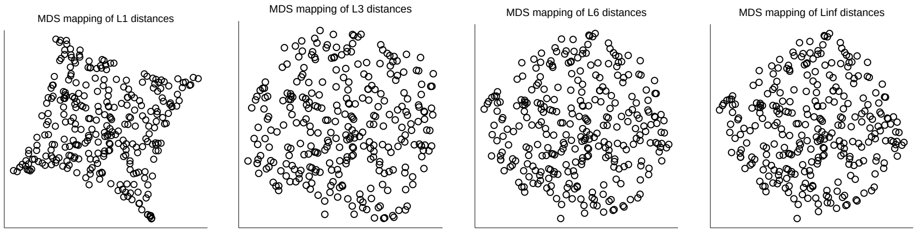


Figure 2: Comparison of L_k-norms on uniform square data. $(L_1, L_3, L_6, L_\infty$ l.t.r.)

Euclidean plane, see Figs. 2 and 3. Obviously, if data is distributed uniformly, a smooth transition from L_1 to L_∞ can be observed, as expected, whereby the global topological form does not change much. This observation is mirrored in the co-ranking evaluation, see Fig. 4. The quality curves change smoothly and have a value near 1, indicating a good agreement of the topologies. Note that these metrics are topologically equivalent in the mathematical sense, which is supported by the observation made in this case.

The situation changes if more realistic settings are considered, i.e. if structure is present in the data. We consider three clusters and the same setting as before. Here, the metric L_1 and L_∞ yield very different behavior, as can be seen in the projection in Fig. 3 as well as in the evaluation in Fig. 4. Thus, mathematical topology equivalence does not imply that the overall topologies are similar for realistic settings displaying structure. The co-ranking framework mirrors the expected differences in these settings. Note, that due to the choice of K, also differences at different scales are displayed. In Fig. 4, clearly the underlying structure with cluster sizes of 100 can be recovered from the quality curves.

5 Comparison of non-Euclidean settings

In the previous sections, we introduced a mathematical approach to compare two dissimilarity measures, and demonstrated it on artificial data sets. In this section, we use two real world data scenarios as a first proof-of-concept study, to show the usefulness of our approach given domain-specific – and possibly non-Euclidean – dissimilarity measures.

App description texts

Current research in the area of semantic web utilizes state-of-the-art machine learning and data visualization techniques, in order to automatically organize and represent vast data collections within user-friendly interfaces. Here, sophisticated data dissimilarity measures for textual content play an important role.

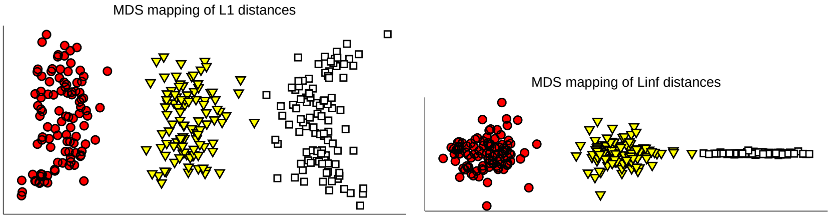


Figure 3: MDS projection using L_p-norms on three clusters data. (L_1, L_∞)

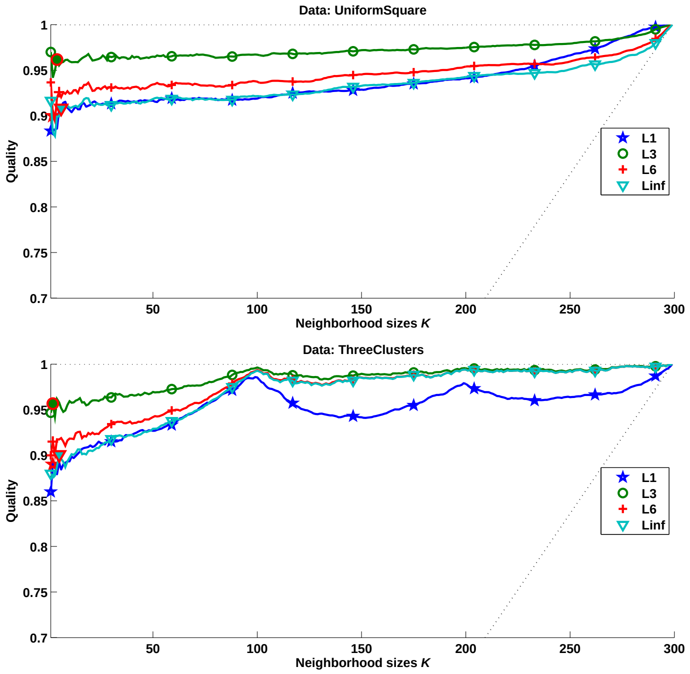


Figure 4: Comparison of the dissimilarities using the co-ranking framework: uniform square (top) and three clusters (bottom).

Our first experimental scenario relates to a typical machine learning task in this context. It consists of descriptions from 500 randomly collected applications, available on the online platform *Google Play* (`http://play.google.com`). Google Play is a large distribution service for digital multimedia content which currently offers about 450.000 downloadable programs (commonly referred to as *apps*) for the mobile operating system *Android*. Each app is attributed to one of 34 categories, while every category belongs to one of the two major branches "Games" or "Applications". The content of every app is summarized in a textual description of about 1200 characters on average. Our 500 apps come from two categories: 293 from "Arcade & Action" (in Games), and 207 from "Travel & Local" (in Applications). In the following they will be referred to as class 1 and 2, respectively. We consider three different measures to calculate dissimilarities between the descriptions:

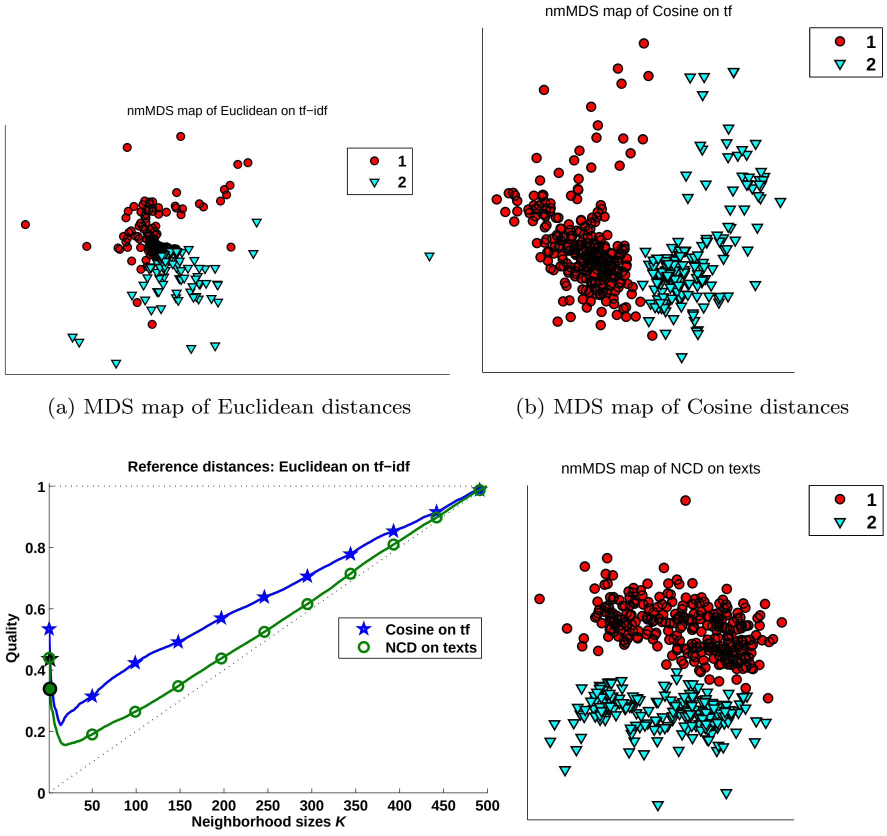
(I) *Euclidean* distances on the tf-idf weights, where weight vectors are calculated from the frequencies of the appearing terms (tf) and their inverse frequency of occurrence in all documents (idf), see [25],

(II) the *Cosine* distance on the term frequencies, which is calculated as $c(\mathbf{a},\mathbf{b}) := 1 - \left((\vec{a}^\intercal \vec{b})/(\pi \left\|\vec{a}\right\| \left\|\vec{b}\right\|)\right)$, where $\vec{a}$ and $\vec{b}$ are vectors of term frequencies for the two respective documents,

(III) the *normalized compression distance* (NCD), which is a string dissimilarity measure based on the Kolmogorov complexity [4], in our case using the Lempel-Ziv-Markov chain compressor (LZMA).

While the first two measures are based on basic word statistics, the NCD also takes structural aspects into account implicitly, since the lossless compressor utilizes recurring patterns in the texts to reduce the description length. Prior to applying the dissimilarity measures, we used a standard preprocessing workflow of stopword reduction and Porter stemming.

Fig. 5 shows MDS visualizations of the three different dissimilarities, as well as evaluation curves from the comparison of Euclidean distances versus the Cosine and the NCD measure. For the visualizations in Fig. 5a, 5b, 5d, we used non-metric MDS with squared stress. From the evaluation curves in Fig. 5c we see that the agreement of the Euclidean distances to the Cosine and NCD measure is low in general, with values below 0.6, even for very small neighborhood sizes. Although the visualizations indicate a qualitatively similar structure, the overall ranks seem to be rather different, which is also reflected in the visualizations to some extent: Fig. 5a shows a small number of outliers, while there are fairly distinct clusters in Fig. 5d; and Fig. 5b shows both characteristics: similarly dense regions and some widespread outliers. In this real world data set, every pair of measures showed a low agreement when compared with the evaluation framework, with $Q_{\mathrm{NX}}(K) < 0.6$ for all $K < 100$.



(a) MDS map of Euclidean distances

(b) MDS map of Cosine distances

(c) $Q_{\mathrm{NX}}(K)$ of Euclidean dist. vs. Cosine & NCD

(d) MDS map of NCD distances

Figure 5: Comparison of the three dissimilarity measures in our first real world showcase scenario consisting of 500 textual descriptions of Android apps.

Java programs

The second example is related to current challenges in the research of *intelligent tutoring systems* (ITS). In general, these educational technology systems are intended to provide intelligent, one-on-one, computer-based support to students in various learning scenarios. Especially in situations where this type of learning support is not available due to scarce (human) resources, the benefits of ITSs become apparent. Since traditional ITSs rely on an exact formalization of the underlying domain knowledge in order to judge whether a given answer from a student is correct or not, they are today mainly applied in well-structured and comparably narrow domains. In order to make future ITSs more flexible, current approaches suggest the application of machine learning techniques to automatically infer models from given sets of student solutions, see [28]. The structural aspects of such data is hard to represent in vectors of numerical features, which would yield an embedding in a Euclidean vector space. Instead, a crucial ingredient of such approaches are domain-specific, and possibly non-metric dissimilarity measures, by which the data can be represented in terms of pairwise relations only. The analysis and development of dissimilarity measures in this area makes a framework for quantitative comparison necessary.

Our data scenario is related to this domain and consists of 169 short Java programs which represent student solutions, originating from a Java programming class of first year students at Clausthal University of Technology, Germany. We used the open source plagiarism detection software *Plaggie* [29] to extract a tokenized representation (a *token stream*) from each given Java source code. Based on the token streams, we consider four different dissimilarity measures:

(I) Euclidean distances on the tf-idf weights like in the previous data set, however, tf and idf now refer to the occurrence of each token instead of term,

(II) the Cosine distance on the token frequencies,

(III) the normalized compression distance (NCD) on the token streams,

(IV) *Greedy String Tiling* (GST) which is the inherent similarity measure that Plaggie uses to compare the given sources [29, 30]; since GST yields a matrix S of pairwise similarities $s(\mathbf{x}^i, \mathbf{x}^j) \in S$, where values are in $(0, 1)$ and self-similarities equal 1, we converted S into a dissimilarity matrix by taking $D := \sqrt{1 - S}$, as proposed in [23].

Fig. 6 shows the quality $Q_{\mathrm{NX}}(K)$ when comparing Euclidean distances to Cosine, GST, and NCD dissimilarities. The curves show the highest similarity to the Cosine distances, especially high in small neighborhood ranges, which is expected due to the fact that both are based on token frequencies. Interestingly, the curves of the Cosine and the GST measure show a similar shape in comparison to Euclidean distances, which may indicate a similar response to certain structural aspects in the data, in contrast to the steadily growing curve for NCD.

Fig. 7 demonstrates our proposed framework for the pointwise comparison of dissimilarity measures on the same data scenario. The coloring in 7c and 7d refers to $Q_{\mathrm{NX}}^{\mathbf{x}_i}(20)$, which is the agreement of the 20-neighborhood for every point $\mathbf{x}_i$ as compared to the other dissimilarity measure. To link the coloring scheme to the evaluation curves, $K = 20$ is highlighted on the graphs in Fig. 6. The pointwise evaluation clearly reveals a region of data which is very close in the Euclidean case, but was considered very dissimilar by the GST measure.

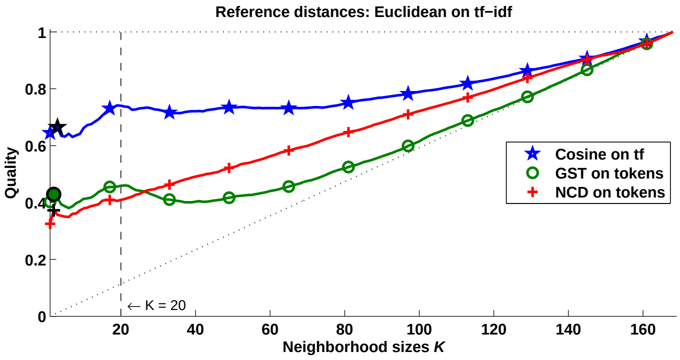
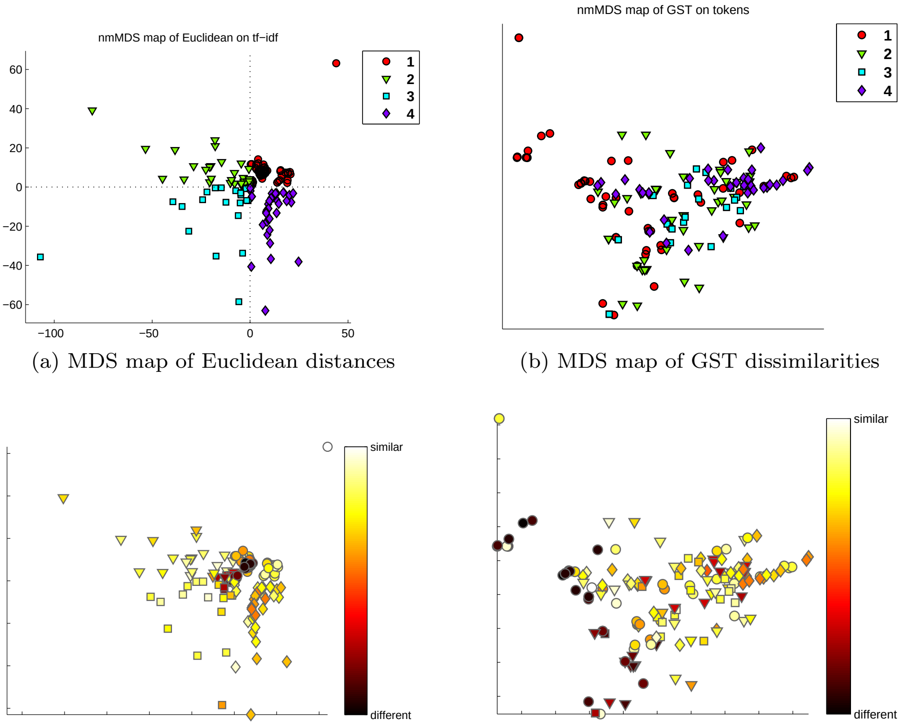


Figure 6: $Q_{\mathrm{NX}}(K)$ when comparing Euclidean distances to Cosine, GST, and NCD dissimilarities used on our second showcase data set consisting of 169 student solutions from a Java programming class.



(a) MDS map of Euclidean distances

(b) MDS map of GST dissimilarities

(c) Euclidean distances, colored by $Q_{\mathrm{NX}}^{\mathbf{x}_i}(20)$ vs. GST dissimilarities

(d) GST dissimilarities, colored by $Q_{\mathrm{NX}}^{\mathbf{x}_i}(20)$ vs. Euclidean distances

Figure 7: Pointwise comparison of dissimilarity measures used on a data set of 169 student solutions from a Java programming class. The dissimilarities from two measures (Euclidean and GST) are mapped to 2D using non-metric MDS. The different symbols for points in the visualizations do not correspond to semantical data classes, but to the quadrants of the cartesian coordinate system in (a), to give some indication of how the point locations differ to the map of GST in (b). The pointwise coloring in (c) and (d) shows for each point, how much the neighbor ranks in the Euclidean case differ to the ranks given by GST.

6 Discussion

We have discussed possibilities to compare dissimilarity measures for unsupervised machine learning tasks. We argued, that rank-preservation or, alternatively, an information retrieval perspective seem very suitable and can be formalized by means of the co-ranking framework taken from the evaluation of dimensionality reduction. We have demonstrated the usefulness in one illustrative artificial example referring to Euclidean vector spaces, as well as two real world examples with problem-specific metrics. The results show that this proposal offers a promising step towards the evaluation, in how far different dissimilarity measures or different choices of metric parameters can lead to substantially different results, when used for unsupervised machine learning.

Naturally, further evaluation techniques are possible such as an evaluation based on the mutual information of the dissimilarities, for example. We conjecture, however, that an information theoretic perspective leads to results which are similar to the co-ranking framework. This is the subject of ongoing work. Further, it is necessary to test whether this a priori comparison of dissimilarity measures coincides with their behavior in typical unsupervised machine learning tasks. Actually, we have already evaluated this behavior to some extent, when visualizing the data in this contribution. The test of further visualization and clustering techniques will be the subject of future work.

Acknowledgment

This work has been supported by the German Science Foundation (DFG) under grants number PI764/6 and HA2719/6-1, and by the center of excellence for cognitive interaction technology (CITEC).

References

[1] M. Ackerman, S. Ben-David, and D. Loker. Towards property-based classification of clustering paradigms. In *NIPS 2010*, 10–18. 2010.

[2] S.-H. Cha. Comprehensive survey on distance/similarity measures between probability density functions. *Int. J. of Mathematical Models and Methods in Appl. Sci.*, 1(4):300–307, 2007.

[3] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti. Similarity-based classification: Concepts and algorithms. *JMLR*, 10:747–776, June 2009.

[4] R. Cilibrasi and P. Vitányi. Clustering by compression. *IEEE Trans. on Information Theory*, 51(4):1523–1545, April 2005.

[5] P. Frasconi, M. Gori, and A. Sperduti. A general framework for adaptive processing of data structures. *IEEE TNN*, 9(5):768–786, 1998.

[6] T. Gärtner. *Kernels for Structured Data.* PhD thesis, Univ. Bonn, 2005.

[7] A. Gisbrecht, B. Mokbel, and B. Hammer. Relational generative topographic mapping. *Neurocomputing*, 74(9):1359–1371, 2011.

[8] B. Hammer and A. Hasenfuss. Topographic mapping of large dissimilarity datasets. *Neural Computation*, 22(9):2229–2284, 2010.

[9] B. Hammer and B. Jain. Neural methods for non-standard data. In *ESANN 2004*, 281–292. 2004.

[10] B. Hammer, A. Micheli, and A. Sperduti. Universal approximation capability of cascade correlation for structures. *Neural Computation*, 17:1109–1159, 2005.

[11] B. Hammer, A. Micheli, and A. Sperduti. Adaptive contextual processing of structured data by recursive neural networks: A survey of computational properties. In B. Hammer and P. Hitzler, editors, *Perspectives of Neural-Symbolic Integration*, vol. 77 of *Studies in computational Intelligence*, pages 67–94. Springer, Berlin, 2007.

[12] B. Hammer, B. Mokbel, F.-M. Schleif, and X. Zhu. White box classification of dissimilarity data. In, *(HAIS 2012)*, vol. 7208 of *LNCS*, 309–321. 2012.

[13] R. J. Hathaway and J. C. Bezdek. Nerf *c*-means: Non-euclidean relational fuzzy clustering. *Pattern Recognition*, 27(3):429–437, 1994.

[14] J. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7-9):1431–1443, 2009.

[15] J. A. Lee and M. Verleysen. *Nonlinear dimensionality redcution*. Springer, 2007.

[16] J. A. Lee and M. Verleysen. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31:2248–2257, 2010.

[17] J. Lewis, M. Ackerman, and V. D. Sa. Human cluster evaluation and formal quality measures. In *Proc. of the 34th Ann. Conf. of the Cog. Sci. Society*, 2012.

[18] H. Liu, D. Song, S. Rüger, R. Hu, and V. Uren. Comparing dissimilarity measures for content-based image retrieval. In *Proc. of AIRS'08*, 44–50, 2008.

[19] D. Malerba, F. Esposito, V. Gioviale, and V. Tamma. Comparing dissimilarity measures for symbolic data analysis. In *Pre-Proc. of ETK-NTTS 2001, HERSONISSOS*, 473–481, 2001.

[20] B. Mokbel, W. Lueks, A. Gisbrecht, M. Biehl, and B. Hammer. Visualizing the quality of dimensionality reduction. In *ESANN 2012*, 179–184, 2012.

[21] M. Neuhaus and H. Bunke. Edit distance-based kernel functions for structural pattern classification. *Pat. Rec.*, 39(10):1852–1863, 2006.

[22] W. R. Pearson and D. J. Lipman. Improved tools for biological sequence comparison. *Proc. of the National Academy of Sciences USA*, 85(8):2444–2448, 1988.

[23] E. Pekalska and R. P. Duin. *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications*. World Scientific, 2005.

[24] A. K. Qin and P. N. Suganthan. Kernel neural gas algorithms with application to cluster analysis. In *ICPR'04 Vol. 4*, 617–620, 2004. IEEE Computer Society.

[25] S. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, 2004.

[26] F. Rossi and N. Villa-Vialaneix. Consistency of functional learning methods based on derivatives. *Pat. Rec. Letters*, 32(8):1197–1209, 2011.

[27] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. Computational capabilities of graph neural networks. *IEEE TNN*, 20(1):81–102, 2009.

[28] B. Hammer, S. Gross, X. Zhu, and N. Pinkwart. Cluster based feedback provision strategies in intelligent tutoring systems. In *Proc. of 11th Int. Conf. on Intelligent Tutoring Systems (ITS)*, 2012.

[29] M. Mozgovoy, S. Karakovskiy, V. Klyuev. Fast and reliable plagiarism detection system. In *Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports. FIE '07. 37th Annual*, 2007.

[30] M.J. Wise. Running Karp-Rabin Matching and Greedy String Tiling. In *Technical report 463 (Univ. of Sydney. Basser Dept. of Comp. Sci.)*, ISBN 0867586699, 1993.

[31] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 9:2579–2605, 2008.

[32] J. Venna. *Dimensionality reduction for Visual Exploration of Similarity Structures*. PhD thesis, Helsinki University of Technology, Espoo, Finland, 2007.

[33] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *JMLR*, 11:451–490, 2010.

[34] H. Yin. On the equivalence between kernel self-organising maps and self-organising mixture density networks. *Neural Netw.*, 19(6):780–784, July 2006.