Comparing feature-based and distance-based representations for classification similarity learning

Emilia LÓPEZ-IÑESTA, Francisco GRIMALDO and Miguel AREVALILLO-HERRÁEZ

Departament d'Informàtica, Universitat de València
Av. de la Universitat s/n. 46100-Burjassot (Spain)
eloi@alumni.uv.es, francisco.grimaldo@uv.es, miguel.arevalillo@uv.es

Abstract. The last decades have shown an increasing interest in studying how to automatically capture the likeness or proximity among data objects due to its importance in machine learning and pattern recognition. Under this scope, two major approaches have been followed that use either feature-based or distance-based representations to perform learning and classification tasks. This paper presents the first results of a comparative experimental study between these two approaches for computing similarity scores using a classification-based method. In particular, we use the Support Vector Machine, as a flexible combiner both for a high dimensional feature space and for a family of distance measures, to finally learn similarity scores in a CBIR context. We analyze both the influence of the different input data formats and the training size on the performance of the classifier. Then, we found that a low dimensional multidistance-based representation can be convenient for small to medium-size training sets whereas it is detrimental as the training size grows.

Keywords. similarity learning, distance-based representation, training size

1. Introduction

Learning a function that measures the similarity between a pair of objects is a common and important task in applications such as classification, information retrieval, machine learning and pattern recognition. The Euclidean distance has been widely used since it provides a simple and mathematically convenient metric on raw features, even when dealing with a small training set, but it is not always the optimal solution for the problem being tackled [14]. This has led to the development of numerous similarity learning techniques [4,6] aimed to build a model or function that, from pairs of objects, produces a numeric value that indicates some kind of conceptual or semantic similarity and also allows to rank objects in descending or ascending order according to this score.

Some studies have have put their attention into automatically learning a similarity measure that satisfies the properties of a metric distance [19,8] from the avalaible data (e.g. in the form of pairwise constraints obtained from the original labeled information)

and have turned supervised metric learning into a topic of great interest [5]. Under this scope, when the properties of a metric are not required, a similar setting can also be used to train a classifier to decide whether a new pair of unlabeled objects is similar or not, an approach that is named as classification similarity learning.

Classification similarity learning has traditionally represented the annotated objects in the training set as numeric vectors in a multidimensional feature space. It is also known that the performance of many classification algorithms is largely dependent on the size and the dimensionality of the training data. Hence, a question that arises is what the ideal size and dimension should be to obtain a good classification performance, considering that greater values generally yield to a better classification but at the cost of increasing the computational load and the risk of overfitting [13].

To deal with this issue, the dimensionality of the training data has been commonly reduced by using distance-based representations such as pairwise distances or (dis)similarities [12]. It is also a frequent practice to use different (dis)similarity measures, each acting on distinct subsets of the muldimensional available features, that are lately combined to produce a similarity score value. A number of combination techniques then exists, under the name of fusion schemes, that have been categorised either as early or late fusion [20]. While the first one uses a unified measure that merges all the features, the second one computes multiple feature measures on a separate basis and then combines them to obtain the similarity between two objects.

Inspired by the late fusion scheme, in this paper we use a multidistance representation that transforms the original feature space in a distance space resulting from the concatenation of several distance functions computed between pairs of objects. This kind of input data involves an additional knowledge injection to the classifier, because the use of a distance measure is an implicit match between the characteristics of two objects and also because of the usual correlation between semantic similarity and small values of distance. It is worth mentioning that this multidistance space is related to the dissimilarity space defined in [7]. Nevertheless, it differs from it in that the space transformation is carried out at a feature level between freely selected pairs of objects instead of using a fixed representation set.

The aim of this paper is to compare the performance obtained from the feature-based and the multidistance-based representations when applied to a classification similarity learning setting as well as to analyze the influence of different training data sizes. Thus, our goal is twofold: on the one hand, we want to study the ability of a classifier to deal with a high feature dimensionality when the training size grows; on the other hand, we want to test under which circumstances the reduction in dimensionality leads to better results than treating objects in their wholeness.

The proposed experimentation concerns the problem of Content-Based Image Retrieval (CBIR), where image contents are characterized by multidimensional vectors of visual features (e.g. shape, color or texture). By considering pairs of images labeled as similar or dissimilar as training instances, we face a binary classification problem that can be solved through a soft classifier that provides the probability of belonging to each class. This probability value can be considered as the score determining the degree of similarity between the images and it can be used for ranking purposes. In particular, the Support Vector Machine classification algorithm has been selected and we use four different values for the Minkowski distance to construct the multidistance-based representation. Additionally, we use as baseline for our comparison the performances obtained

from the global Euclidean distance and two other traditional score-based normalization methods: the standard Gaussian normalization and the Min-max normalization.

The rest of the paper is organized as follows: Section 2 formulates the problem and describes the multidistance-based representation into detail; Section 3 presents the experimental setting and analyzes the obtained results; finally, Section 4 states the conclusions and discusses future work.

2. Problem Formulation

Let us assume we have a collection of images $\mathcal{X} = \{x_i\}, i = 1, 2 \ldots$, which are conveniently represented in a multidimensional feature space $\mathbb{F}$. Let us also assume that this feature space is defined as the Cartesian product of the vector spaces related to T different descriptors such as color, texture or shape.

$$\mathbb{F} = \mathbb{F}^{(1)} \times \ldots \times \mathbb{F}^{(t)} \times \ldots \times \mathbb{F}^{(T)} \tag{1}$$

Hence, we can represent as $x_i^{(t)}$ the set of features that correspond to descriptor t in x_i. Let us finally consider a classical similarity learning setup [19,8], where k training pairs (x_i, x_j) are available that are accordingly labeled as similar (S) or dissimilar (D). In classification-based learning, these pairs are used to train a classifier that can later be able to classify new sample pairs. Thus, when it comes to using a soft classifier, its output will provide a score that may be used to judge the similarity between objects.

A straigthforward approach that fits this scheme is to concatenate the feature vectors of the objects and use the resulting double-size vector as the input to the classifier (see the arrow labeled "feature-based representation" in Figure 1). However, by following this approach, the learning problem size highly depends on the dimensionality of the feature space $\mathbb{F}$, which is usually rather large. This situation might be specially critical for small sample datasets, which unfortunately are often the case. The dimensionality of the input data can then be reduced by using feature reduction techniques such as Principal or Independent Component Analysis. Another way of tackling this problem is by applying a similarity-based spatial transformation [7]. In this paper we evaluate the performance of a multidistance-based representation resulting from a preprocessing layer that acts before passing the traning data to an SVM (see the arrow labeled "multidistance-based representation" in Figure 1).

The preprocessing layer is composed of two steps. The first one derives from computing a family of N distance functions (e.g. Euclidean, cosine or Mahalanobis) for every training pair. Each distance function is defined in each descriptor vector space as in Equation 2.

$$d_n^{(t)} : \mathbb{F}^{(t)} \times \mathbb{F}^{(t)} \longrightarrow \mathbb{R} \tag{2}$$

Thus, we define a transformation function w as indicated in Equation 3 that, given the feature-based representation of two images x_i and x_j, constructs a tuple of values $\langle d_1^{(1)}, \ldots, d_N^{(1)}, d_1^{(2)}, \ldots, d_N^{(2)}, \ldots, d_1^{(T)}, \ldots, d_N^{(T)} \rangle$, where $d_n^{(t)}$ denotes the distance between $x_i^{(t)}$ and $x_j^{(t)}$.

$$w : \mathbb{F} \times \mathbb{F} \longrightarrow \mathbb{R}^{N \cdot T} \tag{3}$$

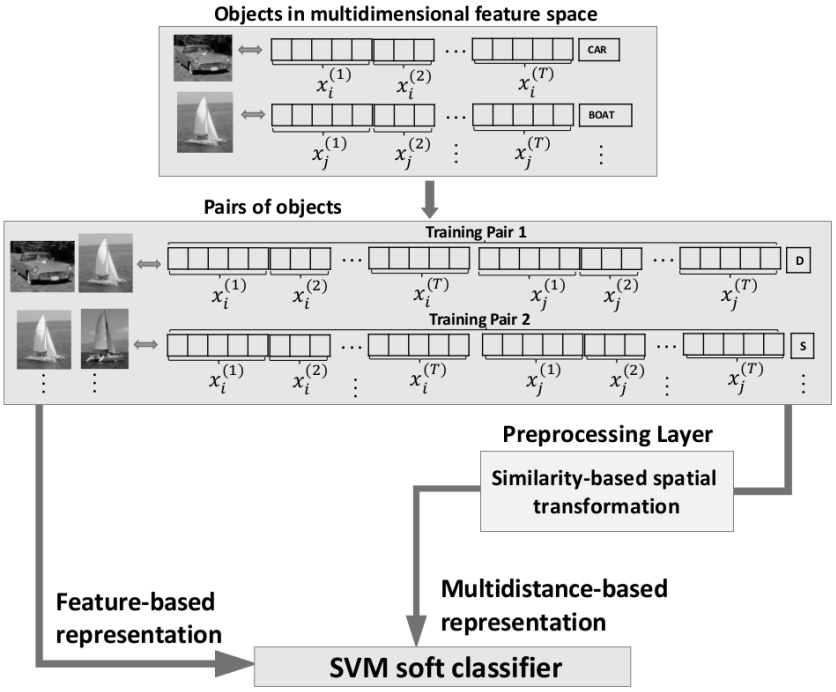

Figure 1. Feature-based and multidistance-based classification similarity learning approaches.

The choice of the most suitable distance function depends on the task at hand and affects the performance of a retrieval system [15]. This has led different authors to analyze the performance of several distance measures for specific tasks [1,9]. Therefore, rather than choosing the most appropriate distance for a task, the proposed multidistance representation aims to boost performance by combining several distance functions simultaneously. This operation transforms the original data into a labeled set of $N \cdot T$-tuples, where each element refers to a distance value, calculated on a particular subset of the features (i.e. the corresponding descriptor).

The second step of the preprocessing layer normalizes the labeled tuples in order to increase the accuracy of the classification [18] by placing equal emphasis on each descriptor space [2]. We denote by $\langle \hat{d}_1^{(1)}, \ldots, \hat{d}_N^{(1)}, \hat{d}_1^{(2)}, \ldots, \hat{d}_N^{(2)}, \ldots, \hat{d}_1^{(T)}, \ldots, \hat{d}_N^{(T)} \rangle$ the normalized tuples, where a simple linear scaling operation into range $[0, 1]$ has been applied. A complete schema of the input data transformation done by the preprocessing layer can be seen in see Figure 2.

Once the SVM soft classifier has been trained, it can be used to provide a score value that can be treated as a similarity estimation between images. For any new pair (x_i, x_j), function w is applied to convert the original features into a tuple of distances (using the same family of functions as for training). After normalization, the resulting vector is used as input to the classifier, that provides a confidence estimation that the pair belongs to any of the classes. This estimate can be used directly for ranking purposes, or converted into a probability value by using the method in [16].

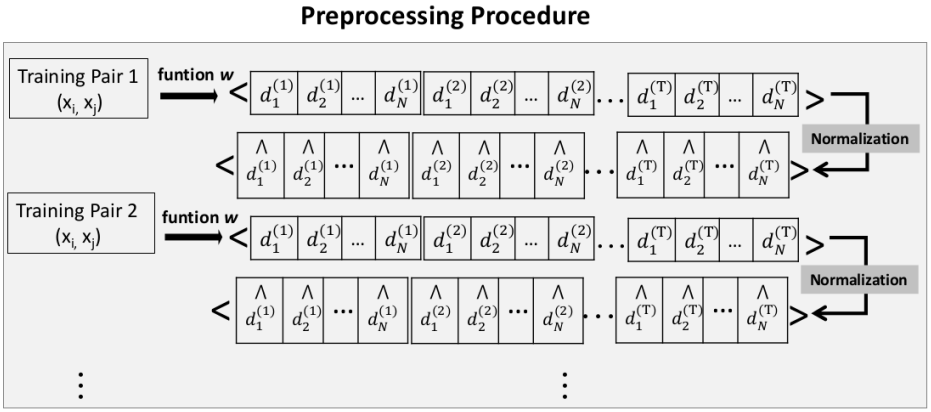

Figure 2. Scheme of the preprocessing layer.

3. Evaluation

3.1. Experimental setting

To analyze the performance of the SVM classifier for the feature-based and the multidistance-based training data formats, a number of experiments have been run on a medium-size image dataset that has also been used in other previous studies (e.g., [3]). The dataset contains a subset of 5476 images from the large commercial collection called "Art Explosion", which is composed of a total of 102894 royalty free photographs that are distributed by the company Nova Development[1]. The images from the repository, originally organized in 201 thematic folders, have been carefully selected and classified into 63 categories so that images in the same category represent a similar semantic concept. Each image in the dataset is described by a label that refers to the semantic concept the image represents according to this manual classification and by a 104-dimensional numeric feature vector defined in a muldimensional feature space through a set of ten visual descriptors[2].

The two classification similarity learning approaches have been also compared with three other traditional score methods, that have been used as baseline. The first one is the global Euclidean distance applied on the entire feature vectors. The second one is the standard Gaussian normalization as described in [10], that consists of a mapping function $d_2^{(t)} \rightarrow (d_2^{(t)} - \mu)/3\sigma$, where μ and σ represent the mean and the standard deviation of the Euclidean distance on each descriptor vector space ($d_2^{(t)}$). The third one is the Min-max normalization, that performs a linear transformation on data computing the minimum and the maximum of the distance $d_2^{(t)}$. The last two approaches are both applied to the individual visual image descriptors and will be referred to as Gaussian normalization and

[1] http://www.novadevelopment.com

[2] The database and details about their content can be found in http://www.uv.es/arevalil/dbImages/

Min-max normalization, respectively. The experiments were run 50 times each and the results were averaged.

To evaluate the influence of the training size, the experiments were run over twelve training sets with increasing sizes. The smallest training set had 100 pairs while the remaining training sets increased their size sequentially from 500 up to 5500 with steps of 500 pairs. In each training set the pairs were labeled as similar (S) when the labels associated with the vectors were the same, and as dissimilar (D) otherwise. It is worth noting that sizes smaller than 100 pairs were discarded as the classification method was outperformed by the baseline methods for such tiny training sets. On other hand, sizes greater than 5500 pair did not show any qualitative difference and followed the trends shown in this paper.

After the training phase, if any, the ranking performance of each algorithm was assessed on a second different and independent test set composed of 5000 pairs randomly selected from the repository. To this end, the Mean Average Precision (MAP), one commonly used evaluation measure in the context of information retrieval [17], was used. The MAP value corresponds to a discrete computation of the area under the precision-recall curve. Thus, by calculating the mean average precision we had a single overall measure that provided a convenient trade-off between precision and recall along the whole ranking.

For the preprocessing layer generating the multidistance-based representation, we have considered a pool composed of four Minkowski distances (L_p norms), with values $p = 0.5, 1, 1.5, 2$. These are widely used dissimilarity measures that have shown relatively large differences in performance on the same data [1,9], and hence suggest that may be combined to obtain improved results. Fractional values of p have been included because they have been reported to provide more meaningful results for high dimensional data, both from the theoretical and empirical perspective [1], a result that has also been confirmed in a CBIR context [9]. In addition, the kernel chosen for the SVM has been a Gaussian radial basis function. The parameters γ and C have been tuned by using an exhaustive grid search on a held out validation set composed of a 30% partition of the training data ($C \in \{10^{-6}, 10^{-5}, ..., 10^{0}, 10^{1}\}$ and $\gamma \in \{10^{-2}, 10^{-1}, ..., 10^{4}, 10^{5}\}$). To compensate the SVM sensitiveness to unbalanced data sets [11], we fix the percentage of similar pairs in the training set to 30%.

3.2. Results

Figure 3 plots the average MAP obtained for the compared similarity learning methods as a function of the training size, where one can distinguish three performance regions. On the left hand side, when the training set is very small (i.e. less than or equal to 100 pairs), the normalization methods perform better than the two classification approaches, that show a limited learning capacity. The reason behind this result is that there is not enough information for training the classifier. Even so, the multidistance-based representation outperforms the feature-based representation since it handles the high dimensionality in a better way.

As the training set increases in size, we identify a second interval (i.e. from 100 to 2000 pairs, approximately) in which the reduction of dimensionality achieved through the concatenation of multiple descriptor distances has positive effects on the classifier performance. Indeed, the multidistance-based representation obtains the highest values in this region of small to medium-size training sets.

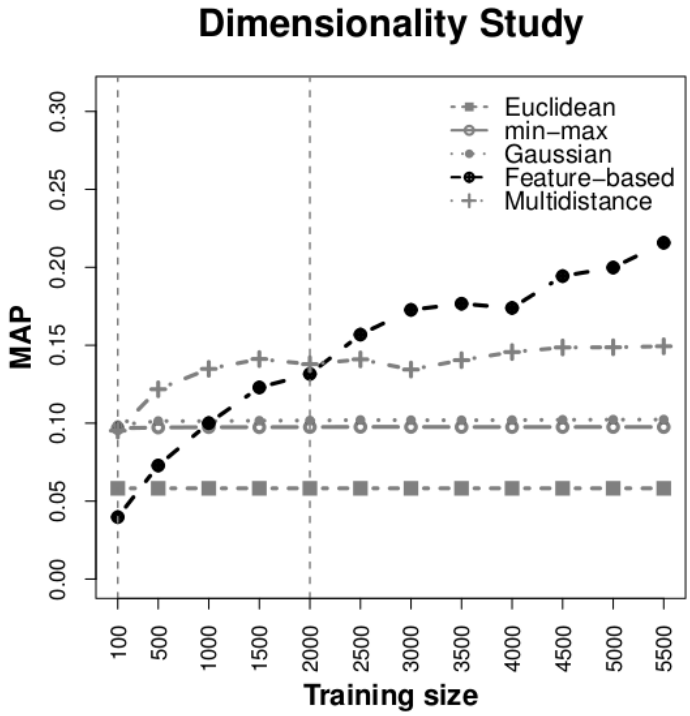

Figure 3. Average MAP vs. training set size.

Notwithstanding, when we allow the training size to grow far enough (i.e. beyond 2000 pairs), the classical feature-based representation improves the results that can be obtained from the rest of the algorithms. This third region demonstrates that the information loss incurred by the multidistance-based representation can be detrimental for big training sizes, since it limits the learning capabilities of the SVM classifier. All in all, these results suggest that the training size can have an important effect on the classification performance when we adopt differents strategies for the representation of data input.

Regarding the performance of the baseline approaches, the average MAP remain constant for all the different training set sizes and show small values for the global Euclidean distance. Values for the Gaussian and Min-max normalization are almost equal and fall generally bellow the average MAP values obtained for both classification methods. Figures 4 and 5 allow us to observe the variability of the MAP values resulting from respectively executing the feature-based and the multidistance-based representations. Each point in these plots correspond to one of the 50 executions run for each training set size, while the solid line shows the linear curve fit. By comparing the plots, we observe that the feature-based representation generates a higher variability while the multidistance-based representation obtain more robust results. Even though these results are still preliminary and need further exploration, the reason behind such an behaviour could again be the difficulty of having enough examples to learn a high dimensional classification model.

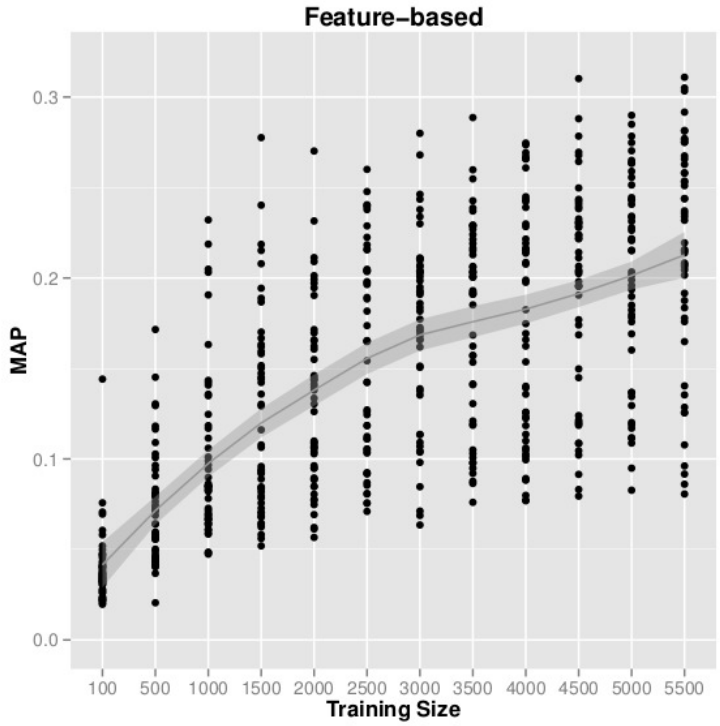

Figure 4. MAP vs. training set size for the feature-based representation.

4. Conclusion

In this paper we have conducted an experimental study comparing two approaches for learning similarity scores in a multidimensional feature space using a classification-based method as the SVM. The difference between these approaches is based on the representation format followed by the sample dataset that is used to train the classifier. On the one hand, a feature-based representation of objects can have as drawback the high dimensionality of the learning problem that it poses to the classifier. On the other hand, a multidistance-based representation can reduce dimensionality by transforming the original multidimensional space in a distance space constructed as the concatenation of a number of distance functions.

A series of performance patterns have been extracted from the analysis of the different input data formats and the training size. We found that a low dimensional multidistance-based representation can be convenient for small to medium-size training sets whereas it is detrimental as the training size grows. The dimensionality reduction (e.g. in the form of distances relations and its combitation) supposes additional information to the classifier and boosts its performance. For large training sets, though, a higher dimensional feature-based representation provides better results for the data base considered. This results can be of value when designing future systems that need to automatically capture the similarity of pairs of objects.

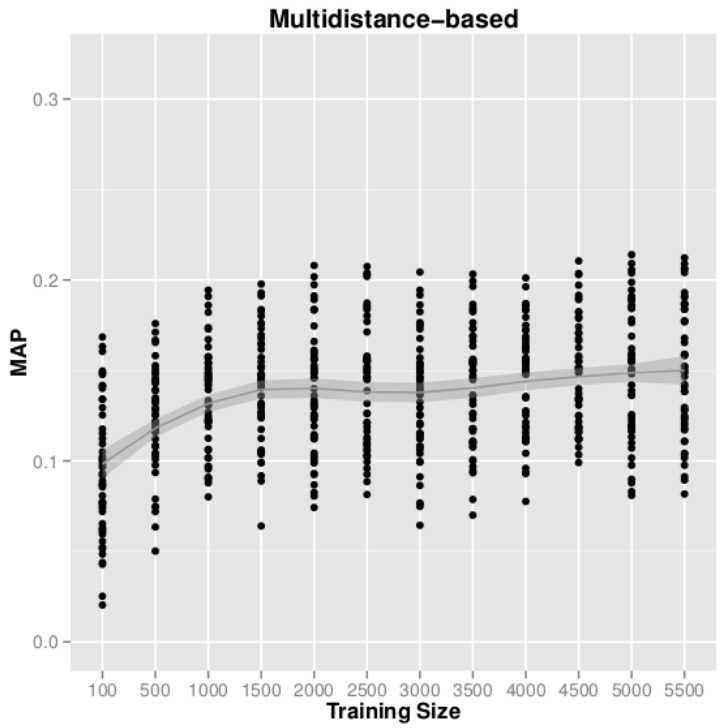

Figure 5. MAP vs. training set size for the multidistance-based representation.

Future work will extend this study by including other databases with different characteristics in size and dimensionality. Besides, further investigation is needed that considers more distance combinations as well as other suitable techniques to reduce the dimensionality of the training set and to finally improve the performance of classifiers.

Acknowledgements

We would like to thank Juan Domingo, for extracting the features from the images in the database. This work has been supported by the Spanish Ministry of Science and Innovation and the Vice-chancellor for Training Policies and Educational Quality at the Universiy of Valencia through projects TIN2011-29221-C03-02 and UV-SFPIE_FO13-147196.

References

[1] C. Aggarwal, A. Hinneburg, and D. Keim. On the surprising behavior of distance metrics in high dimensional space. In J. Bussche and V. Vianu, editors, *Database Theory ICDT*, volume 1973 of *Lecture Notes in Computer Science*, pages 420–434. Springer Berlin Heidelberg, 2001.

[2] S. Ali and K. Smith-Miles. Improved support vector machine generalization using normalized input space. In *AI 2006: Advances in Artificial Intelligence*, volume 4304 of *Lecture Notes in Computer Science*, pages 362–371. Springer Berlin Heidelberg, 2006.

[3] M. Arevalillo-Herráez and F. J. Ferri. An improved distance-based relevance feedback strategy for image retrieval. *Image and Vision Computing*, 31(10):704 – 713, 2013.

[4] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *Proceedings of the 20th International Conference on Machine Learning*, pages 11–18, 2003.

[5] A. Bellet, A. Habrard, and M. Sebban. Similarity Learning for Provably Accurate Sparse Linear Classification. In *ICML*, pages 1871–1878, 2012.

[6] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In Z. Ghahramani, editor, *Machine Learning, Proceedings of the 24th International Conference (ICML)*, volume 227, pages 209–216. ACM, 2007.

[7] R. P. W. Duin and E. Pkalska. The dissimilarity space: Bridging structural and statistical pattern recognition. *Pattern Recognition Letters*, 33(7):826–832, May 2012.

[8] A. Globerson and S. T. Roweis. Metric learning by collapsing classes. In *Advances in Neural Information Processing Systems 18 (NIPS)*. MIT Press, 2005.

[9] P. Howarth and S. Rüger. Fractional distance measures for content-based image retrieval. In *Proceedings of the 27th European conference on Advances in Information Retrieval Research (ECIR)*, pages 447–456, Berlin, Heidelberg, 2005. Springer-Verlag.

[10] Q. Iqbal and J. K. Aggarwal. Combining structure, color and texture for image retrieval: A performance evaluation. In *16th International Conference on Pattern Recognition (ICPR)*, pages 438–443, 2002.

[11] S. Köknar-Tezel and L. J. Latecki. Improving SVM classification on imbalanced time series data sets with ghost points. *Knowledge and Information Systems*, 28(1):1–23, 2011.

[12] W.-J. Lee, R. P. W. Duin, A. Ibba, and M. Loog. An experimental study on combining euclidean distances. In *Proceedings 2nd International Workshop on Cognitive Information Processing (14-16 June, 2010 Elba Island, Tuscany - Italy)*, pages 304–309, 2010.

[13] Y. Liu and Y. F. Zheng. FS_SFS: A novel feature selection method for support vector machines. *Pattern Recognition*, 39(7):1333–1345, 2006.

[14] B. McFee and G. R. G. Lanckriet. Metric learning to rank. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 775–782. Omnipress, 2010.

[15] J. P. Papa, A. X. Falcão, and C. T. N. Suzuki. Supervised pattern classification based on optimum-path forest. *Int. J. Imaging Syst. Technol.*, 19(2):120–131, 2009.

[16] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, pages 61–74. MIT Press, 1999.

[17] B. Thomee and M. S. Lew. Interactive search in image retrieval: a survey. *International Journal of Multimedia Information Retrieval*, 1(2):71–86, 2012.

[18] J. Vert, K. Tsuda, and B. Schölkopf. *A Primer on Kernel Methods*, pages 35–70. MIT Press, Cambridge, MA, USA, 2004.

[19] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell. Distance metric learning with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15 (NIPS)*, pages 505–512. MIT Press, 2002.

[20] J. Zhang and L. Ye. Local aggregation function learning based on support vector machines. *Signal Processing*, 89(11):2291–2295, 2009.