Transactions on
Case-Based Reasoning
Vol.3, No 1 (2010) 1-16
© ISSN: 1867-366X (Journal),
ISBN: 978-3-940501-17-2,
IBaI Publishing ISSN 1864-9734

ibai Publishing
www.ibai-publishing.org

Learning Discriminative Distance Functions for Case Retrieval and Decision Support

Alexey Tsymbal[1], Martin Huber[1], Shaohua Kevin Zhou[2]

[1]Corporate Technology Div., Siemens AG, Erlangen, Germany
{alexey.tsymbal, martin.huber}@siemens.com
[2]Siemens Corporate Research, Princeton NJ, USA
shaohua.zhou@siemens.com

Abstract. The importance of learning distance functions is gradually being acknowledged by the machine learning community, and different techniques are suggested that can successfully learn a strong distance function in many various contexts. Nevertheless the studies in the area are still rather fragmentary; they lack systematic analysis and focus on a limited circle of application domains. In this paper, two techniques for learning discriminative distance function are evaluated and compared on biomedical data of different kind; learning from equivalence constraints and the intrinsic Random Forest similarity. Both techniques demonstrate competitive results with respect to plain learning; the Random Forest similarity exhibits a more robust behaviour and is shown to be less susceptible to missing data and noise.

1 Introduction

During the last three decades, the importance of the distance function in machine learning has been gradually acknowledged, and recently, a growing body of work has addressed the problem of supervised or semi-supervised learning of customised distance functions [9]. There are several reasons that motivate the studies in the area of learning distance functions and their use in practice [4]. First, learning a distance function helps to combine the power of strong learners with the transparency of case

retrieval and nearest neighbour classification. Besides, learning a proper distance function was shown to be especially helpful for high-dimensional data with many correlated, weakly relevant and irrelevant features, where most traditional techniques would fail. Also, it is easy to show that choosing an optimal distance function makes classifier learning redundant. Next, a distance function "learnt" for a certain context may serve as a basis for data visualization, e.g. with a neighbourhood graph, and ultimately for the discovery of important knowledge. Besides, it fosters the creation of more modular and thus more flexible systems, supporting component reuse. Another important benefit is the opportunity for inductive transfer between similar tasks; this approach is often used in computer vision applications; see e.g. [13].

Historically, the most popular approach in distance function learning is Mahalanobis metric learning, which has received considerable research attention but is however often inferior to alternative non-linear and non-metric distance learning techniques. For example, the Relative Component Analysis algorithm, RCA [9] learns a Mahalanobis distance metric, which is optimal under several criteria, using generative learning with positive equivalence constraints.

While distance metrics and kernels are widely used by various powerful algorithms, they work well only in cases where their axioms hold [9]. For example, in [12] it was shown that distance functions that are robust to outliers and irrelevant features are non-metric, as they tend to violate the triangular inequality. Human similarity judgements were shown to violate both the symmetry and triangular inequality metric properties. Besides, a large number of hand-crafted context-specific distance functions suggested in various application domains are far from being metric. Our focus in this paper is thus on learning non-linear and non-metric discriminative distance functions.

In this paper, we evaluate and compare two techniques for learning distance function; learning from equivalence constraints and the intrinsic Random Forest similarity. These techniques originate from different communities and are commonly applied to different subject domains, but serve the same purpose – learning a strong distance function which is specific for a given classification context.

This paper is organized as follows. Section 2 gives a review of basic concepts and related work. Section 3 introduces our experimental framework implemented for the study, and in Section 4 most important experimental results are presented. The paper finishes with the discussion of experimental findings and potential applications in Section 5 and a summary and directions of ongoing and future work in Section 6.

2 Basic Concepts and Related Work

2.1 Learning from Weak Representations

2.1.1 Equivalence Constraints

While historically the research on distance learning has started from supervised learning of distance functions for *k*-nearest neighbour classification in the original

"feature vector-object label" representation [21], today a more commonly used representation is the one based on so called *equivalence constraints* [9].

Usually, equivalence constraints are represented using triplets (x_1, x_2, y), where x_1, x_2 are data points in the original space and $y \in \{+1,-1\}$ is a label indicating whether the two points are similar (from the same class) or dissimilar. Learning from these triples is also often called learning in the *product space* (i.e. with pairs of points as input); see [8] and [28] for examples. While learning in the product space is perhaps a more popular form of learning from equivalence constraints, yet another common alternative is to learn in the *difference space*, the space of vector differences; see [1] and [27] for examples. The difference space is normally used with homogeneous high-dimensional data, such as pixel intensities or their PCA coefficients in imaging. While both representations demonstrate promising empirical results in different contexts, there is no understanding which representation is better. No comparison was done so far; usually a single representation for the problem is chosen.

There are two essential reasons that motivate the use of equivalence constraints in learning distance functions; their availability in some learning contexts and the fact that they are a natural input for optimal distance function learning [4]. It can be shown that the optimal distance function for classification is of the form $p(y_i \neq y_j \mid x_i, x_j)$. Under the *i.i.d.* assumption the optimal distance measure can be expressed in terms of generative models $p(x \mid y)$ for each class as follows [13]:

$$p(y_i \neq y_j \mid x_i, x_j) = \sum_y p(y \mid x_i)(1 - p(y \mid x_j)) \qquad (1)$$

This function was analytically proven to be no worse than any distance metric and was shown to approach the Bayesian optimal accuracy [13].

Another common representation for learning a distance function, which is even weaker than equivalence constraints are relative comparisons, used to learn distances mainly in retrieval contexts; see [2], [16], [18] for examples. Learning is conducted from triplets of the form "x is more similar to y than to z". Such triplets can always be extracted from labels, and sometimes even from positive and negative constraints (if certain points appear in both positive and negative constraints), but not vice versa. For information retrieval, for which the order of the retrieved items is important, such triplets seem like natural supervision [4].

Even though obtaining equivalence and comparative constraints is often easier than providing real labels, often even such form of supervision is limited and thus *semi-supervised learning* techniques can be useful. For example, the *DistBoost* technique uses semi-supervised learning with Expectation Maximization to learn base weak classifiers which are later combined using boosting in the product space [8], [9]. An important question in this context is *how much* supervision is usually enough. Unfortunately, this is still rather an open question. Usually a fixed number of constraints is used; [2], for example, use *200,000* triples representing comparative constraints in their *BoostMap* technique.

2.1.2 Learning Algorithms

Clearly, as can also be seen from already mentioned techniques, the most popular learning algorithm in the area of distance function learning is *boosting* [7]. BoostMap [2], DistBoost [8], Boosted Distance with Nearest Neighbor [1], and BoostMotion [28] are based on boosting, to mention a few. Different variations of boosting are used with different base learners, though usually all correctly designed boosting-based techniques demonstrate competitive results. The most popular variations are AdaBoost and LogitBoost. Different base classifiers are used with boosting, including decision stumps, 1D and 2D embeddings, Expectation Maximization to learn a mixture of Gaussians, and C4.5 decision trees [9].

Alternative learners considered include SVMs and kernel-based techniques (this is perhaps the second largest group of learners in the area) [9]. Perhaps the most important disadvantage for SVMs is that they cannot be so naturally integrated with feature selection as boosting (with decision stumps), which makes their use in highly dimensional tasks problematic. Different learners are *rarely* compared with each other in literature. There is some empirical evidence though that SVMs may demonstrate strong overfitting behaviour when used for learning distance functions. For example, in [8] boosting-based techniques demonstrate quite stable behaviour in comparison to other learners, being also better on average, and SVMs are significantly unstable, being slightly better that the other learners in a few domains but also being significantly worse in some cases.

As learning with weak representations may often be formulated as a convex optimization problem, corresponding iterative optimization techniques like Newton's method are also used; e.g. for learning how to combine distances [25], for linear programming with comparative constraints [16], and for solving a convex quadratic programming problem based on comparative constraints [18]. These techniques usually learn a Mahalanobis distance function or a linear combination of functions, and linearity has certain limitations in our context as discussed above.

It is important to note that often, instead of learning a single distance function from data, *a combination* of a number of existing distance functions has recently been also attempted; see e.g. [13], [25], [26]. Different combination techniques are used for that, including boosting-based learning and a linear logistic model solved using standard convex optimization techniques. The distance functions being integrated include both canonical elementary functions such as Manhattan and Euclidean, and more sophisticated learning-based techniques. While this is definitely an interesting and promising area of research, the benefit over one-stage distance function learning is not always clear and besides such a scenario assumes specification of a set of base-level distance functions, which may be context-specific and different for different domains. This is especially true for domains where the canonical distance functions may be misleading rather than useful. Often a separate search is conducted in the space of available base-level distance functions, in order to select the best subset for combination. The area of combining distance functions is out of scope of this paper; the focus is on immediate learning a single distance function from the original instances instead.

2.2 The Intrinsic Random Forest Distance Function

For a Random Forest (RF) learnt for a certain classification problem, the proportion of the trees where two instances appear together in the same leaves can be used as a measure of similarity between them [6]. For a given forest f the similarity between two instances x_1 and x_2 is calculated as follows. The instances are propagated down all K trees within f and their terminal positions $\mathbf{z}$ in each of the trees ($\mathbf{z_1}=(z_{11},...,z_{1K})$ for x_1, similarly $\mathbf{z_2}$ for x_2) are recorded. The similarity between the two instances then equals to (I is the indicator function):

$$S(x_1, x_2) = \frac{1}{K}\sum_{i=1}^{K} I(z_{1i} = z_{2i}) \qquad (2)$$

For a detailed description of RFs and their properties the reader is referred to [6]. The most important properties in our context are: (1) their predictive performance was shown to be as good as boosting and sometimes better; (2) they are relatively robust to outliers, noise and missing values; and (3) they are faster than many other techniques, boosting in particular.

Similarity (2) can be used for different tasks related to the classification problem. Thus, [19], [20] successfully use it for hierarchical clustering of tissue microarray data. First, unlabeled data are expanded with a synthetic class of evenly distributed instances, then a RF is learnt and the intrinsic RF similarities are determined as described above and clustered. The resulting clusters are shown to be clinically more meaningful than the Euclidean distance based clustering with regard to post-operative patient survival. Hudak *et al.* [10] use it for nearest neighbour imputation on forestry sensor data. They conclude that the RF distance based imputation is the most robust and flexible among the imputation techniques tested.

It is interesting that using this similarity for the most immediate task, nearest neighbour classification, is rather uncommon, comparing to its use for clustering. In one of related works, Qi *et al.* [15] use it for protein-protein interaction prediction, and the results compare favourably with all previously suggested methods for this task.

2.3 Applications

More than in any other research domain, the problem of learning a better distance function lies today in the core of research in *computer vision* [4]. Different imaging applications have been considered, including image retrieval (with facial images, animal images, hand images, and American Sign Language images), object detection (indoor object detection), motion estimation and image registration; see [4],[9] for an in-depth review.

Besides vision, some other domains were also considered including computational immunology, analysis of neuronal data, protein fingerprints, and text retrieval [9]. Surprisingly, there is relatively few related work in text/document retrieval. One example is [18] which studies the retrieval of text documents from the Web by learning a distance metric from comparative constraints.

The intrinsic RF distance is rather a "dark horse" with respect to learning from equivalence constraints. The number of known applications for it is still limited; perhaps, the most successful application is clustering genetic data [19], [20]. Works on learning equivalence constraints never consider it as a possible alternative. In general, we believe that the circle of applications both for distance learning from equivalence constraints (which is currently applied nearly solely to imaging problems) and for the intrinsic RF distance is still, undeservedly, too narrow and may and should be expanded.

3 Experimental Framework

3.1 Data Sets and Cross Validation

The data sets under study include four clinical data sets; one data set with cardiac aortic valve meshes, *Meshes*, including 63 meshes representing healthy and diseased aortic valves, and three benchmark data sets from the UCI repository, *Liver*, *Thyroid* and *Heart* [5]. All these data sets except Heart include numeric features only, and consist of a relatively small amount of cases (from 63 in Meshes to 345 in Liver), making the extensive computationally expensive experiments involving learning in the product and difference spaces possible. The number of features varies from 5 in Thyroid and 6 in Liver to 13 in Heart and 6,000 in Meshes. Each aortic valve in the Meshes data set is represented as a set of 3D coordinates of the 2,000 characteristic points of the mesh. More details about and some initial experiments with the Mesh data are available in [11]. All the data sets except Thyroid where the number of classes is 3 represent binary classification tasks.

Besides, experiments were conducted on five other high-dimensional data sets, four public microarray gene expression data sets (*Lymphoma*, *Embryonal_Tumours*, *Colon*, and *Leukemia*)[1], and one mass spectrometry data set for cancer identification from the NIPS 2003 challenge (*Arcene*, also available in the UCI repository). All the data sets include numeric features only, and consist of a relatively small amount of cases (from 45 in Lymphoma to 200 in Arcene). The data sets represent binary classification tasks and the number of features varies from 2,000 in Colon to 10,000 in Arcene.

In order to evaluate learning techniques in the study, leave-one-out cross validation was used with the Meshes data and with all the gene expression data sets (Lymphoma, Embr_Tumours, Colon, and Leukemia), and 3 runs of 10-fold cross validation were used for the rest.

[1] The microarray gene expression datasets were taken from and are available at www.upo.es/eps/aguilar/datasets.html.

3.2 Learning Algorithms

Learning discriminative distance functions from equivalence constraints has been implemented with *AdaBoost* and *RF* as the learning algorithms, representing equivalence constraints in the *product* and *difference* spaces. The difference space included simple pairwise L_1 differences for each feature. AdaBoost used C4.5 decision tree as the base learning algorithm. LogitBoost with decision stumps was also tried (as it is often also a technique of choice in this context); its results are not considered in this paper, as it often results in clearly inferior accuracy, sometimes even close to a random guess, due to too weak base models generated with decision stumps in the difference and product spaces.

All the discriminative distance functions are evaluated based on the accuracy of *k*-nearest neighbour classification (*k-NN)*, with k=7, and weighting the votes of the neighbours inversely proportional to the distance. The same settings for *k*-NN were used in all the experiments, in order to avoid unnecessary estimation bias. These settings were shown to lead to best or close to best accuracies for all data sets, in a series of preliminary experiments. The *k*-NN classifier with the customized distance function was compared with the plain *k-NN* using the Euclidean distance, and with the *RF* and *AdaBoost with C4.5* in the original space. 50 component trees were always included in all ensemble models (AdaBoost and RF); it was shown to be enough to achieve competitive accuracy for most data sets in the previous studies.

Among the considered algorithms, only RF-based learning algorithms are able to successfully deal with high-dimensional data, such as the four gene expression data sets, Arcene and Meshes in their raw form, as long as RFs incorporate an explicit feature selection process at each node in the training phase. In order to get rid of this limitation and put each algorithm in the same initial conditions, at each cross-validation run, 200 features were pre-selected according to their Information Gain, for all experiments with the high-dimensional data sets. In a set of preliminary experiments, 200 features were shown to lead to competitive accuracies for all the data; the use of other numbers of features will not change the accuracy substantially.

The learning algorithms mentioned above and the experimental framework for our study were implemented on the basis of machine learning library WEKA 3.4 [24]. All learning algorithms used default settings besides the ones already mentioned.

3.3 Data Imputation Techniques

In order to test the sensitivity of the techniques for learning discriminative distance functions to missing data, different levels of missing data (5%, 10%, 15%, 20%, and 30%) were randomly injected into our complete data sets, and the performance of learning discriminative distances was evaluated after imputing the missing data with four data imputation techniques.

First, the naïve *mean/mode substitution*, which was historically the first and long time the most common method of data imputation, which is however no longer preferred. Second, the simple but more powerful and still often used *k-NN imputation*.

Again, we use k=7 and the inverse distance as the weight for the neighbours in our experiments. Third, *multiple imputation with bootstrap sampling* and fourth, *multiple imputation with random subspacing*. Multiple imputation (MI) is a procedure that replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute [17]. MI has much in common with classification ensembles, and same to them often leads to a better accuracy than the simple single imputation. Random subspaces for MI in the last technique are generated by sampling with each feature having 0.7 probability of being selected. Ten base learners (k-NN) are used in our experiments with MI, which is a common choice.

4 Experimental Studies

4.1 Learning from Equivalence Constraints *vs.* the Intrinsic RF Distance

Table 1 includes classification accuracies for plain learning techniques, such as k-nearest neighbour (*k-NN*), AdaBoost (*AB*), and Random Forest (*RF*); for k-NN with distance learning from equivalence constrains, with AdaBoost and RF in the product and difference spaces (*AB-prod*, *AB-diff*, *RF-prod*, and *RF-diff*, correspondingly); and for k-NN with the intrinsic RF distance (*RF_dist*); for the nine data sets and on average. The best accuracies for every data set are given in bold, and statistically significant differences according to the McNemar's test for the difference between two proportions (this happens for the Embr_tumours and Arcene data sets only) are given in italic. From the table, one may see that in general, k-NN with learning distance functions results in accuracies no worse or even better than the plain techniques. This is also confirmed by the average accuracies. The best accuracies, for the given data sets, are always achieved with distance learning. These results are in line with conclusions made in similar previously reported studies.

Table 1. Classification accuracies for plain learners and learning discriminative distance functions.

Data Set	k-NN	AB	RF	AB-prod	RF-prod	AB-diff	RF-diff	RF_dist
Mesh	.889	.905	.905	.921	**.937**	.905	.905	.921
Lymphoma	.978	.911	.956	**1**	**1**	**1**	**1**	.978
Embr_Tumours	.717	.733	.767	.767	.783	.750	.75	***.800***
Colon	.790	.839	.855	.855	**.871**	**.871**	**.871**	**.871**
Leukemia	.944	.931	.958	.958	**.972**	**.972**	**.972**	**.972**
Arcene	.837	.873	.858	.865	.863	.862	.880	***.898***
Liver	.646	.703	.700	**.721**	.711	.658	.706	.698
Thyroid	.924	.934	.944	.967	.947	.964	.962	**.969**
Heart	.818	.816	.811	.811	.821	.824	.824	**.829**
Average	.838	.849	.861	.874	.878	.867	.874	.882

4.2 Distance Learning with Sampling from Equivalence Constraints

One of the biggest issues with learning from equivalence constraints is its computational complexity. If N is the size of the original training set, it is augmented with a new training set with $O(N^2)$ instances. This may make learning practically infeasible for data sets with already several hundreds of instances, if all instance pairs are used for learning. In order to address this issue, in a separate experiment we have studied the performance of learning when equivalence constraints are sampled randomly from the full set of instance pairs.

In Figure 1 classification performance is shown for RF in the difference space, for random samples of instance pairs, starting from the full set of instance pairs to a 10% sample, on the Liver data set.

As can be seen, performance starts to degrade noticeably when only less than 30% of instance pairs are used. First, the accuracy even increases a little, which can perhaps be explained by overfitting the overcomplete representation. This figure shows a typical pattern characteristic, in general, for all data sets. Only for one data set (Thyroid) the full set of instance pair was needed to achieve the best performance. For other data sets a smaller sample was enough; thus, accuracy started to decrease noticeably for less than 50% samples for Embr_tumours, Leukemia and Arcene, for less than 40% samples for Lymphoma and Colon; and even for less than 10% only for Mesh and Heart. Besides, this "optimal" sample size does not depend much on the type of the space with constraints, be it the product or the difference space.

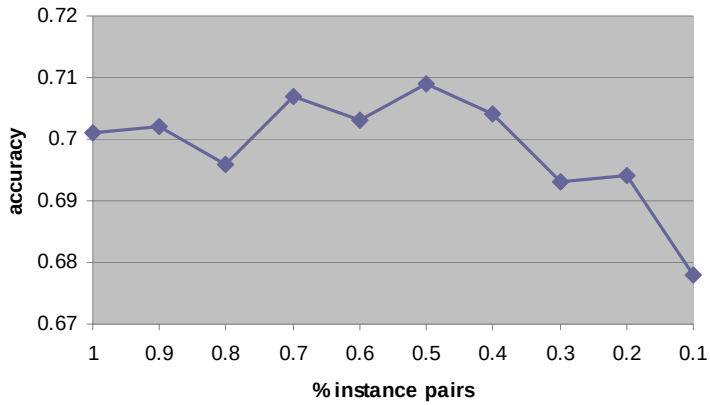

Fig. 1. Classification accuracy for random samples of instance pairs, on the Liver data set

4.3 Sensitivity to Missing Data

In Figure 2 classification accuracy is shown for distance learning with equivalence constraints (*Equivalence*), and the intrinsic RF distance (*RF_dist*), for different levels of missing data, starting from no missing data ("0", the original data sets) up to 30% of missing data values introduced randomly into the data ("0.3"). Ten-model MI with

bootstrap sampling is used for replacing the missing data. First, the training set was processed to replace the missing values, then a similar procedure was used to replace missing data in the test instances, and after that the techniques for learning discriminative distance functions could be evaluated as usual. The accuracies are averaged over the nine data sets.

Learning from equivalence constraints is represented by the best technique for each data set in this experiment. Clearly, this gives a certain optimistic bias to the corresponding estimates, as long as normally validation in advance would be needed in order to determine the most suitable technique out of the four (that is why it outperforms RF_dist here, in contrast to the results in Table 1). Nevertheless, as can be seen from the graph, the intrinsic RF distance exhibits a robust behaviour; its performance reaches and even overpasses the competitor's performance with the increase in the level of missing data.

We assume that this can be explained by the robustness of the RF learner to noise and missing data, which is a fact already reported in many previous studies, starting from [6]. On the contrary, equivalence constraints may propagate noise, due to the nature of the product and difference spaces. Besides the experiments with missing data, a separate series of experiments with artificially introduced noise can be conducted in order to validate this finding; this forms a direction for future work.

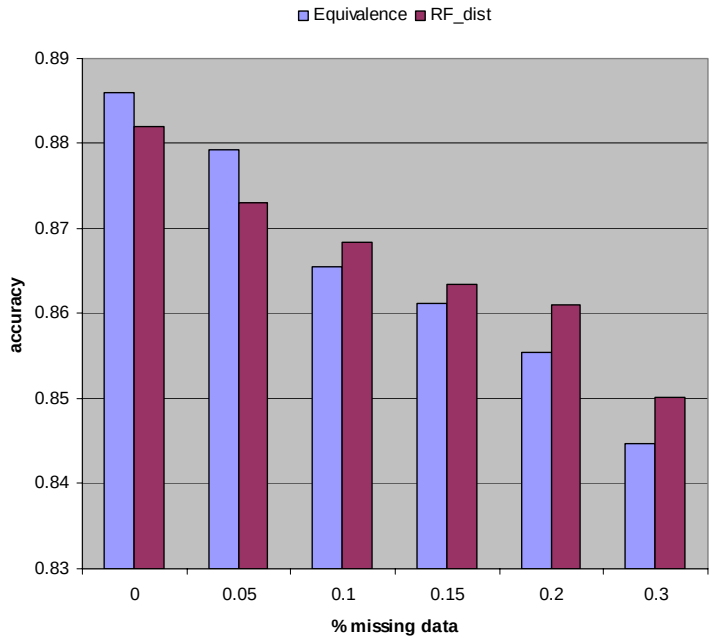

Fig. 2. Classification accuracy for distance learning with equivalence constraints (*Equivalence*), and the intrinsic RF distance (*RF_dist*), for different levels of missing data

4.4 Performance of Imputation Techniques

In Figure 3, classification accuracy is shown for the four imputation techniques that have been used to impute data in the data sets with injected missing values; (1) simple mean/mode substitution, *Mean/mode*, (2) single *k*-NN imputation, *SI*, (3) multiple imputation with random subspacing, *MI (subspacing)*, and (4) multiple imputation with bootstrap sampling, *MI (bootstrap)*. The displayed results are for the intrinsic RF distance, averaged over the nine data sets, for the case with 20% missing values. The performance of the individual imputation techniques in the context of distance function learning is not surprising; mean/mode substitution is the simplest and weakest strategy on average, then goes the single *k*-NN imputation which leads to slightly better results, and the use of multiple imputation helps to improve the imputation accuracy even more, for most data sets. The average accuracy improvement for strong imputation techniques (for multiple imputation with regard to the mean/mode substitution and single imputation) is relatively small, as long as for some data sets the simple mean/mode substitution already gives acceptable results. In multiple imputation in our experiments, *ten* base models only were used, and perhaps a small accuracy improvement can be achieved by increasing this number. Interesting, bootstrap sampling works well in this context of multiple *k*-NN imputation, although bootstrap aggregation (bagging) of lazy learners (such as *k*-NN) is known not to lead to big gains in accuracy for traditional classification.

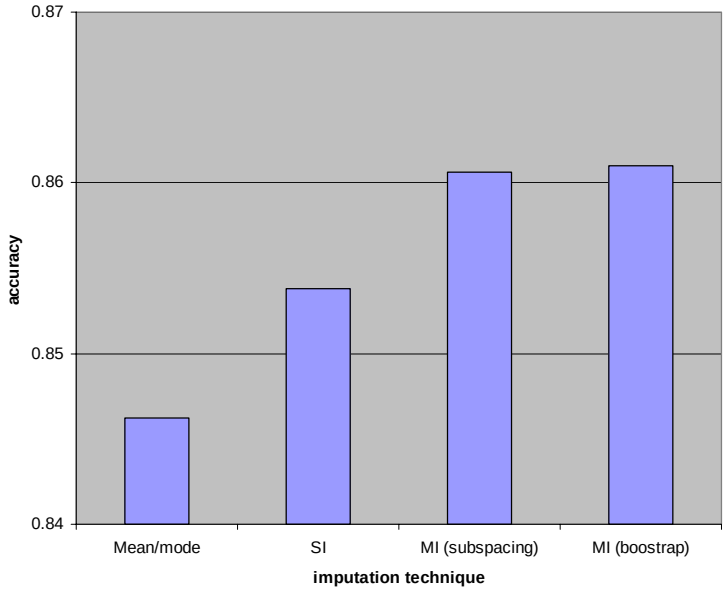

Fig. 3. Classification accuracy for different imputation techniques with 20% of missing data

4.5 Number of Trees in RF for Distance Learning

In order to better understand the behaviour of the two different types of distance learning and to validate the hypothesis regarding the higher sensitivity to overfitting for learning from equivalence constraints, we have conducted a separate series of experiments with RFs including different numbers of trees; 3, 7, 15, 31, and 63 (as a series of powers of 2 minus 1).

In Figure 4, classification accuracy is shown for the plain RF classifier, RF with equivalence constraints (*RF_equiv*), and the intrinsic RF distance (*RF_dist*), for the different numbers of component trees within RF, averaged over the nine data sets. For *RF_equiv*, the best space (difference or product) is selected for every data set to represent equivalence constraints.

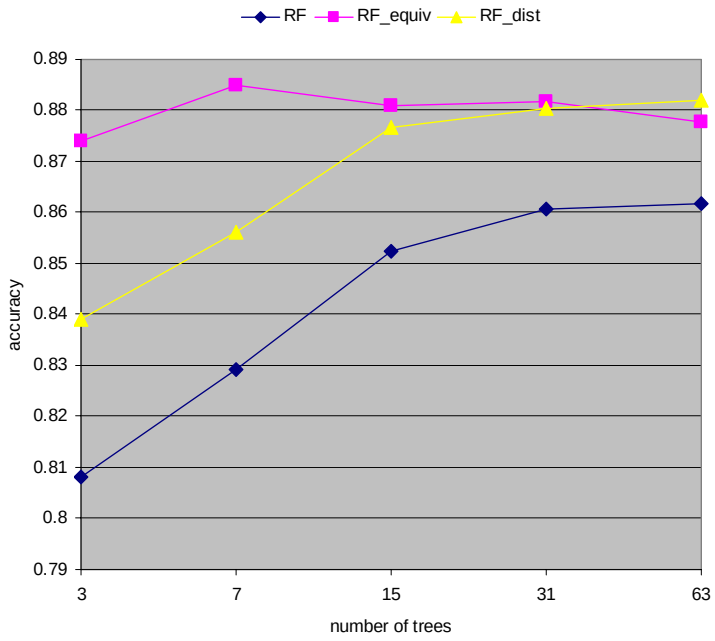

Fig. 4. Classification accuracy for RF, RF with equivalence constraints (*RF_equiv*), and the intrinsic RF distance (*RF_dist*), for different numbers of component trees within RF

The behaviour of the plain RF classifier and the intrinsic RF distance is not surprising and could be expected; the accuracy plateaus with the increase in the number of trees and no significant accuracy increase may be expected for RFs with more than 63 trees. The RF behaviour for learning from equivalence constraints is more interesting and rather unexpected. The peak of accuracy is achieved already with 7 trees and then the accuracy decreases slightly, supporting our hypothesis regarding the greater risk of overfitting for learning from equivalence constraints. It is important to note that this is not only an average trend; a similar trend is exhibited for

every particular data set. Besides, a similar phenomenon is experienced with AdaBoost as well.

The experienced phenomenon that a high accuracy is achieved with smaller numbers of trees in an ensemble, for learning from equivalence constraints, can perhaps be explained by the fact that the new task of distance learning in the transformed space may actually be simpler that the original learning task.

5 Discussion

One of the most interesting findings in our study is the fact that learning from equivalence constraints may be prone to overfitting noise. This was validated in our experiments with missing data and with different amounts of component trees in the ensembles. This contradicts to some extent to the conclusion made in [1], [26], where it is claimed that the big amount of instances in the new training set with equivalence constraints makes AdaBoost more robust to overfitting. However, no empirical or theoretical evidence were given in [1], [26] to support this claim, besides the better accuracies with distance learning.

Another important issue with learning from equivalence constraints is its computational complexity. One way to address this issue, as demonstrated in our experiments, is sampling from the set of equivalence constraints. Another way, not addressed in this paper, is to use online learning techniques. The existing incrementalizations of boosting are far from being lossless, see e.g. [14]. On the other hand, they were shown to converge to the same accuracy for overcomplete representations, and the set of available equivalence constraints is often overcomplete. For RF there is no known good incrementalization, although we believe a lossless or close to lossless algorithm may exist (a nearly lossless algorithm for bagging does exist).

While in our experiments, the intrinsic RF distance was shown to be more robust, a properly configured learning from equivalence constraints may still in many cases be better. However, one would need to have enough training data in order to cross-validate and select the most appropriate representation and a learner for it (this should not be necessarily RF or AdaBoost only).

The techniques and experimental results considered in this paper may be useful for the development of real-world case retrieval and decision support systems. As an example, at present we are developing a system for the retrieval of patient records that may include complex biomedical data; clinical, imaging and genomic data (called *Health-e-Child CaseReasoner*). The underlying inter-patient similarity (that can be calculated via distance functions learnt using techniques considered in this paper) can be visualized using neighbourhood graphs, treemaps, and heatmaps, for better knowledge discovery and decision support [22, 23]. The two considered techniques for learning discriminative distance function, when used in combination with data visualization techniques such as the neighborhood graph or the treemap, may become a powerful and flexible tool for clinical decision support in various classification contexts.

Health-e-Child is an EU-funded Framework Programme 6 (FP6) project aimed at improving personalized healthcare in selected areas of paediatrics, especially focusing on integrating medical data across disciplines, modalities, and vertical levels such as molecular, organ, individual and population. The results presented in this paper contribute to the development of decision support systems in the project.

6 Conclusions

In this paper two techniques for learning discriminative distance functions were compared; learning from equivalence constraints and the intrinsic RF distance. The techniques considered are different by their nature and originate from different communities but have an important commonality in that that they are discriminative distance functions, or functions intended/optimised for case retrieval in a certain classification context. Both techniques demonstrate competitive performance with respect to the plain learning; they help to combine the power of strong "eager" learners with the transparency of case retrieval and nearest neighbour classification. As expected, the techniques of this kind are especially useful for data sets with many correlated, weakly relevant and correlated features and with small sample sizes, which is often the case with biomedical data. With experiments on missing data and with varying the amount of component models in the ensembles, it is shown that learning from equivalent constraints may have a higher risk of overfitting the data. No clear dependency was found between the parameter settings for learning from equivalence constraints, data set characteristics and classification performance. Validation on unseen data can be recommended in order to find a suitable configuration. The intrinsic RF distance was shown to be more robust overall, although finding suitable parameters for learning from equivalence constraints may still be competitive. Our ongoing work in this area includes the application of techniques to other subject domains with complex data and the study of incrementalizing learners for equivalence constraints. The use of product space often leads to an overcomplete representation of the problem so that even non-lossless online implementations of learning algorithms may work well and converge to the accuracy of their batch analogues with this data. Another important direction for future work is the study of how the non-metricity of considered techniques contributes to their success and their comparison with the state-of-the-art metric learning techniques.

Acknowledgments: This work has been done in the framework of the EU project Health-e-Child (IST 2004-027749). The authors wish to acknowledge support provided by all the members of the Health-e-Child consortium in the preparation of this paper.

References

1. Amores, J., Sebe, N., Radeva, P.: Boosting the distance estimation. Application to the *k*-nearest neighbour classifier. Pattern Recognition Letters 27, 201-209 (2006)

2. Athitsos, V., Alon, J., Sclaroff, S., Kollios, G.: BoostMap: A method for efficient approximate similarity rankings. In: Proceedings of the Int. Conf. on Computer Vision and Pattern Recognition, CVPR (2004)
3. Bar-Hillel, A., Weinshall, D.: Learning distance functions by coding similarity. In: Proceedings of the Int. Conf. on Machine Learning, ICML (2007)
4. Bar-Hillel, A.: Learning from weak representations using distance functions and generative models. Doctoral Dissertation, Department of Computer Science, Hebrew University of Jerusalem (2006)
5. Blake, C.L., Keogh, E., Merz, C.J.: UCI repository of machine learning databases. Dept. of Information and Computer Science, University of California at Irvine (1999)
6. Breiman, L.: Random Forests. Machine Learning 45, 15-32 (2001)
7. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceedings of the Int. Conf. on Machine Learning, ICML (1996)
8. Hertz, T., Bar-Hillel, A., Weinshall, D.: Learning distance functions for image retrieval. In: Proceedings of the Int. Conf. on Computer Vision and Pattern Recognition, CVPR (2004)
9. Hertz, T.: Learning distance functions: algorithms and applications. Doctoral Dissertation, Department of Computer Science, Hebrew University of Jerusalem (2006)
10. Hudak, A.T., Crookston, N.L., Evans, J.S., Hall, D.E., Falkowski, M.J.: Nearest neighbour imputation of species-level, plot-scale forest structure attributes from LiDAR data. Remote Sensing of Environment 112(5), 2232-2245 (2008)
11. Ionasec, R.I., Tsymbal, A., Vitanovski, D., Georgescu, B., Zhou, S.K., Navab, N., Comaniciu, D.: Shape-based diagnosis of the aortic valve. In: Proc. SPIE Medical Imaging, vol. 7259 (2009)
12. Jacobs, D.W., Weinshall, D., Gdalyahu, Y.: Classification with non-metric distances: Image retrieval and class representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(6), 583-600 (2000)
13. Mahamud, S., Hebert, M.: The optimal distance measure for object detection. In: Proceedings of the Int. Conf. on Computer Vision and Pattern Recognition, CVPR (2003)
14. Oza, N.C., Russell, S.J.: Experimental comparisons of online and batch versions of bagging and boosting. In: Proceedings of KDD (2001)
15. Qi, Y., Klein-Seetharaman, J., Bar-Joseph, Z.: Random Forest similarity for protein-protein interaction prediction from multiple sources. In: Prooceedings of Pacific Symposium on Biocomputing (2005)
16. Rosales, R., Fung, G.: Learning sparse metrics via linear programming. In: Proceedings of the Int. Conf. on Knowledge Discovery in Data, KDD (2006)
17. Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York (1987)
18. Schulz, M., Joachims, T.: Learning a distance metric from relative comparisons. Advances in Neural Information Processing Systems, NIPS 16 (2003)
19. Shi T., Horvath S.: Unsupervised learning with Random Forest predictors. *Computational and Graphical Statistics*, 15 (1) (2006) 118-138
20. Shi, T., Seligson, D., Belldegrun, A.S., Palotie, A., Horvath, S.: Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod Pathol. 18(4) 547-557 (2005)
21. Short, R.D., Fukunaga, K.: The optimal distance measure for nearest neighbour classification. IEEE Transactions on Information Theory 27 (5) 622-627 (1981)
22. Tsymbal, A., Huber, M., Zillner, S., Hauer, T., Zhou, S.K.: Visualizing patient similarity in clinical decision support. In: Hinneburg, A. (ed.), LWA 2007: Lernen - Wissen - Adaption, Workshop Proceedings, pp. 304-311, Martin-Luther-University Halle-Wittenberg (2007)

23. Tsymbal, A., Zhou, S.K., Huber, M.: The neighbourhood graph for clinical case retrieval and decision support within Health-e-Child CaseReasoner. In IEEE EMBC Conference (2009)
24. Witten, I., Frank, E.: Data mining: Practical Machine Learning Tools with Java Implementations, Morgan Kaufmann, San Francisco (2005)
25. Woznica, A., Kalousis, A., Hilario, M.: Learning to combine distances for complex representations. In: Proceedings of the Int. Conf. on Machine Learning, ICML (2007)
26. Yu, J., Amores, J., Sebe, N., Tian, Q.: A new study on distance metrics as similarity measurement. In: Proceedings of the Int. Conf. on Multimedia and Expo (2006)
27. Yu, J., Amores, J., Sebe, N., Tian, Q.: Toward robust distance metric analysis for similarity estimation. In: Proceedings of the Int. Conf. on Computer Vision and Pattern Recognition, CVPR (2006)
28. Zhou, S.K., Shao, J., Georgescu, B., Comaniciu, D.: Boostmotion: Boosting a discriminative similarity function for motion estimation. In: Proceedings of the Int. Conf. on Computer Vision and Pattern Recognition, CVPR (2006)