IMPROVED AUDIO CLASSIFICATION USING A NOVEL NON-LINEAR DIMENSIONALITY REDUCTION ENSEMBLE APPROACH

Stéphane Dupont
University of Mons
stephane.dupont@umons.ac.be

Thierry Ravet
University of Mons
thierry.ravet@umons.ac.be

ABSTRACT

Two important categories of machine learning methodologies have recently attracted much interest in classification research and its applications. On one side, unsupervised and semi-supervised learning allow to benefit from the availability of larger sets of training data, even if not fully annotated with class labels, and of larger sets of diverse feature representations, through novel dimensionality reduction schemes. On the other side, ensemble methods allow to benefit from more diversity in base learners though larger data and feature sets. In this paper, we propose a novel ensemble learning approach making use of recent non-linear dimensionality reduction methods. More precisely, we apply t-SNE (t-distributed Stochastic Neighbor Embedding) to a large feature set to come up with embeddings of various dimensionality. A k-NN classifier is then obtained for each embedding, leading to an ensemble whose estimates can then be combined, making use of various ensemble combination rules from the literature. The rationale of this approach resides in its potential capacity to better handle manifolds of different dimensionality in different regions of the feature space. We evaluate the approach on a transductive audio classification task, where only part of the whole data set is labeled. We confirm that dimensionality reduction by itself can improve performance (by 40% relative), and that creating an ensemble through the proposed approach further reduces classification error rate by about 10% relative.

1. INTRODUCTION

Feature transformation and dimensionality reduction approaches have attracted a lot of interest as pre-processors in classification problems, including in the area of multimedia information retrieval. In general, they are able to reduce correlation of feature dimensions as well as noise. They also help to tackle the issues related to the curse of dimensionality, to avoid over-fitting on the training data, and to reduce the computational cost of the classification scheme.

Unsupervised non-linear dimensionality reduction schemes have shown their benefit in semi-supervised learning problems for classification [11, 23]. This follows from the so-called cluster and manifold assumptions. In the first, it is assumed that data samples are organized into distinct clusters and that samples from different classes belong to different clusters. In the second, it is assumed that data samples from different classes occupy distinct manifolds of lower dimensionality in the original feature space. If one of these assumptions holds, the more the available data can be used, the better these cluster or manifold structures can be discovered, to the benefit of classification accuracy.

Ensemble approaches constitute another popular research theme in machine learning and classification problems. They consists in training a set of diverse estimators (referred to as base learners) for the same problem, and combine their estimates or decisions when new data samples have to be classified [24]. These have become popular with approaches such as bagging [2] and boosting [18] (f.i. AdaBoost), to name a few.

Intuitively, in order to gain accuracy when combining such estimators, these have to be different. This has led to research into generating base learners that are as diverse as possible, by acting on one or more of the factors that will have an effect on the end result of the learning process [4, 24]. There is literature on using different subsets of the training data for each member of the ensemble, on manipulating the parameters involved in the training (up to the architecture of the individual learners), or on using different output representations.

Methods for altering the input feature space have also been researched. In this paper, we propose to make use of recent developments in the area of unsupervised non-linear dimensionality reduction in order to alter the feature space and extract low-dimensional embedding whose dimensions can be used to define multiple classifiers. Their estimates are then combined through an ensemble scheme.

Previous attempts and popular methods in creating ensemble diversity through feature transformations and dimensionality reduction will first be summarized in Section 2. Section 3 will then describe the proposed approach,

This research has received funding from Région Wallonne (Belgium) through the Numediart long-term research programme (grant nbr. 716631) and the MediaWorkflows project (grant nbr. 1117549), and from the European Union Seventh Framework Programme through the i-Treasures project (grant agreement no. 600676).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.
© 2013 International Society for Music Information Retrieval.

and in particular introduce the non-linear dimensionality reduction algorithm that have been applied (t-SNE, or t-distributed Stochastic Neighbor Embedding), the way it is used to obtain multiple classifiers, and the combination rules used to obtain the ensemble decision. We then apply this method on a use case of interest to the creative community concerned with the classification of musical instrument loops used in rhythmic music composition/production. Section 4 presents the experimental protocol and evaluation metrics, as well as the experimental results, together with a discussion. We conclude the paper in Section 5.

2. ENSEMBLE METHODS AND FEATURE MANIPULATIONS

If earlier proposals in ensemble learning relied on selecting different subset of the training data (bagging or boosting) for each ensemble member, subsequent research has explored various approaches for transforming and manipulating the available feature set, including feature selection or more generally supervised and unsupervised dimensionality reduction. This is summarized in chronological order in the following paragraphs.

With the Random Subspace (RS) method [9, 17], random subsets of the original features are presented to the classifiers. This was followed by a more general approach called Random Forests (RF) [3], combining with the idea of bagging, to end up with ensembles of classifiers (initially decision trees, hence the name "forest") constructed from random samplings on both features and data.

Selecting features after PCA has been proposed in [12], with ensembles where each constituent classifier is trained on a user-determined number of principal components. Supervised dimensionality reduction approaches have also lead to some ensemble learning trials. In [13], Input Decimation (ID) is proposed. Its goal is to decouple the classifiers by exposing them to different features. The method does so by training N classifiers (N being the number of classes of the problem) and selecting for each the input feature dimensions (a user-determined number of them) having the highest absolute correlation to the presence or absence of the corresponding class. In [14], the ID-based approach was shown to compare favorably with the PCA-based approach. One explanation is that unsupervised dimensionality reduction approaches such as PCA are not well suited for finding features useful for classification as they totally disregard class information. Remember however that random or unsupervised feature selection also work in some contexts.

Rather than selecting feature dimensions randomly as in RS, or making use of feature transformations, simply projecting on randomly defined axes as also been proposed with the Random Projections (RP) approach [7, 20].

In [16], the ideas of RS and PCA are combined to lead to Rotation Forests (RotF), where the feature set is randomly split into a number of subsets and PCA is applied to these. Diversity-error diagrams revealed that RotF-based ensembles construct individual classifiers which are more accurate than these in AdaBoost and RF, and more diverse than these in Bagging, sometimes more accurate as well.

More recently [1], it is proposed to make use of Diffusion Maps (DM) [10], a non-linear dimensionality reduction scheme, and to develop classifiers based on the transformed space dimensions. The approach is compared to RP, RS and RF methods cited above, as well as Bagging and Boosting (through the AdaBoost algorithm). A multistrategy approach combining DM and Boosting was shown to be superior to other algorithms in many cases.

Ensemble methods can sometimes look like an art. This is without accounting for the theoretical considerations and developments that participate to this research area. One area concerns the study of so-called diversity metrics and diversity generation approaches. All these are however out of the scope of this paper, and the interested reader may refer to recent literature for more information [19, 24].

3. PROPOSED APPROACH

In this paper, we propose to make use of recent developments in non-linear dimensionality reduction approaches in order to obtain several sets of features enabling the development of an ensemble of classifiers. This hence follows up on the literature summarized in the previous section. An earlier proposal was indeed reported in [1], with the use of Diffusion Maps. Here, we will make use of t-SNE (t-distributed Stochastic Neighbor Embedding), a recent method. As explained in [22], t-SNE is less susceptible than other classical approaches (including Diffusion Maps) to assigning much higher importance to modeling the large pairwise distances than the small ones. Hence, it is better at retaining the local structure, which is definitely thought to be beneficial in visualization but also classification problems.

In a previous paper [6], we showed on two semi-supervised classification tasks that t-SNE (even when reducing to a very low dimensional space) can perform as well and sometimes even better than classification in the original high-dimensional feature space.

3.1 Dimensionality Reduction using t-SNE

The popularity of approaches derived from Multidimensional Scaling (MDS) has inspired variants, in particular through methods attempting to preserve local properties of the data in a "softer" probabilistic fashion. In particular, SNE (Stochastic Neighbor Embedding) tries to preserve neighborhood identity [8]. It does so using a cost function that favors the probability distributions of points belonging to the neighborhoods of other points to be similar in the high-dimensional space and in its low-dimensional embedding. In the original formulation, a Kullback-Leibler (KL) divergence is used to measure that similarity, and probabilities for a sample to belong to a neighborhood of another one is based on Gaussian distributions.

More recently, a symmetric version of SNE has been proposed. It has also been proposed to use a Student-t distribution rather than a Gaussian distribution to compute the similarity between pairs of samples in the low-dimensional space. These modifications have lead to the t-SNE [22] method. The t-Student heavy-tailed distribution in the low-dimensional space significantly alleviate the so-called "crowding" problem observed with SNE where far away data samples, for instance low density areas in between natural clusters, come close together in the low-dimensional embedding.

In details, we first estimate the (symmetric) probability that sample x_i in the high-dimensional space would pick sample x_j as its neighbor using the following expression:

$$p_{ij} = \frac{\exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq l} \exp\left(\frac{-\|x_k - x_l\|^2}{2\sigma_i^2}\right)} \tag{1}$$

where σ_i is the standard deviation of a Gaussian centered on x_i. Similarly, we model the probability that y_i, the low dimensional counterpart of x_i, would take y_j as its neighbor using the following (symmetric) expression:

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}} \tag{2}$$

where the model of proximity is Student-t distributed. t-SNE then proposes to find a representation for which the probabilities q_{ij} are faithful to p_{ij}. This is achieved by minimizing the mismatch between q_{ij} and p_{ij} measured using a KL-divergence:

$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \ln\left(\frac{p_{ij}}{q_{ij}}\right) \tag{3}$$

If P_i represents the probability distribution of p_{ij} over all data points given point x_i, t-SNE first performs a binary search for the value of σ_i producing a P_i with a fixed perplexity specified by the user, where the perplexity is defined based on the Shannon entropy of P_i measured in bits.

The minimization of the cost function in Equation 3 is performed using a gradient descent method.

In our experiments, the perplexity of the conditional probability distribution was set to 20; and we performed 2000 iterations of gradient descent. Also, we used the refinements proposed in [22], including a momentum term in the gradient descent, as well a tricks referred to as "early compression" and "early exageration" in [22].

3.2 Ensemble of t-SNE Features

By varying the parameters involved in dimensionality reduction through t-SNE, it is possible to come with several diverse feature representations of the data samples, and to obtain a classifier for each of these. In particular, it is possible to either (1) alter some of the meta-parameters of t-SNE learning, and in particular the size of the local neighborhood (perplexity of the conditional probability distribution), (2) alter the number of dimensions preserved by t-SNE, (3) alter the high-dimensional input features used as input to t-SNE, and in particular select different subsets from the initial feature set, for instance, as in the Random Projection approach, or through more principled feature groupings.

Here, we have been using the later two approaches. Classification tasks can benefit from dimensionality reduction, which is sometimes presented as enabling the reduction of noise and unimportant details in the data, while preserving the multi-dimensional manifold structures. There is however a tradeoff between more denoising through lower dimensional target spaces, and better preservation of the inherent dimensions of the data. The optimal choice may depend on the selected class of the problem, or on the considered regions of the space. One assumption is that ensemble methods making use of classifiers obtained from various choices of target space dimensionality are able to mitigate this tradeoff. Experimental results on the proposed classification task will show that some classes indeed strongly benefit from dimensionality reduction, while others do much less, or not at all. Combining the obtained classifiers leads to improvement over the single best one.

The details of the experimental setup and of the way the multiple classifiers are obtained are provided in Section 4.

3.3 Combination Rules

As soon as the different classifier are available, several approaches are possible for "combining" the individual estimations they provide. Suppose we have K classifiers available. In this work, we consider classifiers that provide estimates of the posterior probabilities of each class. The "sum rule" consists in averaging these posteriors for each class, possibly using a weight dependent on the classifier. It follows from considering classification as a regression problem on posterior probability estimates, and benefit from the literature on ensemble combination through the averaging of various estimates [24]:

$$P(q|x) = \sum_{k=1}^{K} \alpha_k P_k(q|x) \tag{4}$$

where q is the class label, x the feature vector, and $P_k(q|x)$ the posterior probability for class q assigned by classifier k.

The "product rule" consists of a product of probability estimates for each class. If follows from an independence assumption. When one has available several classifiers making use of distinct and statistically independent feature descriptors, the posterior probability of classes can easily be computed from the a priori probabilities of classes and posteriors estimated by the different classifiers. Let $x = (x_1, ..., x_K)$ be the feature vector built from K independent sub-vectors. Bayes rules tells us:

$$P(q|x) = \frac{P(x|q)P(q)}{P(x)} \tag{5}$$

Assuming that the different subparts of the feature vector are statistically independent, we successively get:

$$\begin{aligned} P(q|x) &= \tfrac{P(q)}{P(x)} \prod_{k=1}^{K} P(x_k|q) \\ &= \tfrac{P(q)}{P(x)} \prod_{k=1}^{K} \tfrac{P(q|x_k)P(x_k)}{P(q)} \\ &= \left[\tfrac{\prod_{k=1}^{K} P(x_k)}{P(x)}\right] \left[\tfrac{\prod_{k=1}^{K} P(q|x_k)}{(P(q))^{(K-1)}}\right] \end{aligned} \tag{6}$$

The first term is independent of k, and if the initial independence assumption holds, its value will be 1. In practice, the posterior probability will be estimated based on the second term normalized in such a way that the sum of estimates for all classes is the unity.

Besides averaging, the use of order statistics has also been proposed in the literature [21]. The ”maximum rule” consists in approximating the posterior probability for each class using the maximum of the various classifier estimates for this class:

$$P(q|x) = \max_{k=1}^{K} P_k(q|x) \tag{7}$$

The ”minimum rule” follows a similar principle:

$$P(q|x) = \min_{k=1}^{K} P_k(q|x) \tag{8}$$

Finally, the ”median rule” is expressed as:

$$P(q|x) = med_{k=1}^{K} P_k(q|x) \tag{9}$$

These five approaches have been compared in this work. Majority voting is another popular approach, but it has not been used in this work as it can not benefit from posterior estimates.

4. EXPERIMENTS

As data set, we used a production music library (ZeroG ProPack). This library contains more than ten thousand ”loops” and samples of various instruments and music styles. Each soundfile is typically a few seconds long of monophonic or polyphonic sound (f.i. in the case of guitars). We manually annotated the files within 7 classes of instruments: Brass, Drums, Vocals, Percussion, Electric Bass, Acoustic Guitar and Electric Guitar. After discarding more complex sounds or effects, we ended up with 4380 samples to be used in our evaluations.

The experimental work that follows is based on a transductive classification task. It hence considers a closed data set that has to be classified with minimal effort. Part of the data is hence annotated with class labels to guide the supervised machine learning, but all the data set can be used in an unsupervised mode. Transductive learning has many interesting applications [5].

4.1 Low-level Features for Audio and Music

A large body of recent work in the music information retrieval literature has been devoted to the design of feature extraction algorithms for the purpose of characterizing, analyzing, searching or classifying audio content. Here, we consider timbral properties. Audio analysis approaches for extracting feature descriptors rely on isolating and analyzing short-term windows of temporal signal (typically around 30 ms long), to end up with one feature vector per window. For representing and be able to classify longer-term signals as used in our experiments, we extracted statistics (up to order 4) from the short-term window feature vectors. From previous research, we ended-up using two groups of features, covering the spectral envelope and the noisiness of the sounds, both being important for characterizing the perceived timbre. The state-of-the-art feature set that we used contains:

- Mel-Frequency Cepstral Coefficients (MFCC) as used in [15], computed using 30 ms frames every 10 ms, using a filterbank of 20 filters covering the audible frequency range, and keeping the first 12 coefficients. To be able to capture the temporal characteristics and statistics of the MFCCs, we actually used as features the MFCCs means along the sample duration, as well as their standard deviation, skewness and kurtosis; the means of the first order temporal derivatives of the MFCCs, as well as their standard deviation, and the means of the second order temporal derivatives of the MFCCs, as well as their standard deviation.
- Spectral Flatness (SF), which is a correlate of the noisiness (opposite of sinusoidality) of the spectrum computed on the same audio frames as MFCCs. It is computed as the ratio between the geometric and arithmetic means of the spectrum energy values. As proposed in [15], the spectrum was divided into 4 sub-bands for computing the flatness: 250-500Hz, 500-1000Hz, 1000-2000Hz and 2000-4000Hz. Here too, we used the mean of the SF over the sound extract duration, as well as its standard deviation, skewness and kurtosis.

4.2 Experimental Protocol

We are interested in semi-supervised transductive classification, where only part of the whole corpus can be annotated with the desired class labels. We hence performed experiments with different percentages of randomly selected labeled data (from 10% to 50%, f.i. 10% means that only 438 samples have been labeled using their instrument class in the production music database). Being unsupervised, t-SNE is always making use of the whole data set however. For each training condition, we ran 100 different training and evaluation batches (with selected labeled data randomized for each of them) for each classification system (either single classifier, or various ensemble configurations). The classifiers (either individual classifiers or classifiers taking part in the ensembles) are using k-NN (with k=5).

http://www.zero-g.co.uk/

Data sets definitions and labels available as supplementary material at http://www.numediart.org/tools/mediacycle/

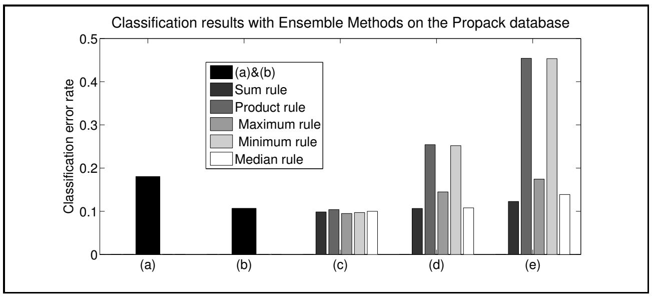

Figure 1. Classification error rates when making use of 10% of labeled data: (a) using the full high-dimensional feature set (b) using a 5-dimensional feature set obtained through t-SNE on the full high-dimensional feature set (best single classifier from different t-SNE based classifiers with various target space dimensionalities) (c) Ensemble composed of 5 classifiers obtained using 1 to 5-dimensional feature sets obtained through t-SNE on the full high-dimensional feature set; various combination rules (d) Ensemble classifier composed of 15 classifiers: 5 target dimensionalities x 3 sub features set (mean, standard deviation, and high-order statistics of the baseline high-dimensional feature set) (e) Ensemble classifier composed of 20 classifiers: 5 target dimensionalities x 4 sub features set (MFCCs, MFCCs first and second derivatives, SF).

The baseline classification system uses the high-dimensional features set described earlier. We normalized each feature to zero-mean and unity-variance. We then created various classifiers used standalone, or involved in ensemble configurations according to three principles:

- creating various classifiers by altering the dimensionality of the t-SNE embedding from 1 to 5. This enabled the design of an ensemble of 5 classifiers.

- creating various classifiers by altering both the dimensionality of the t-SNE embedding and the input features of t-SNE. Rather than selecting random subset of feature dimensions as done in the RD or RS approaches, we partitioned the full feature set according to the order of the statistics used to represent the sound files. More precisely, three subspaces were obtained, one gathering the means of the raw feature vectors, a second one for the standard deviations, and a third one gathering the skewness and kurtosis. This enabled the design of an ensemble of 15 classifiers: 5 target dimensionalities x 3 subsets of features.

- similar to the previous approach but where the category of the feature is use to define the feature partitions. More precisely, we split the individual features into four groups: MFCCs, MFCCs first derivatives, MFCCs second derivatives, and SF. This enabled the design of an ensemble of 20 classifiers: 5 target dimensionalities x 4 subsets of features.

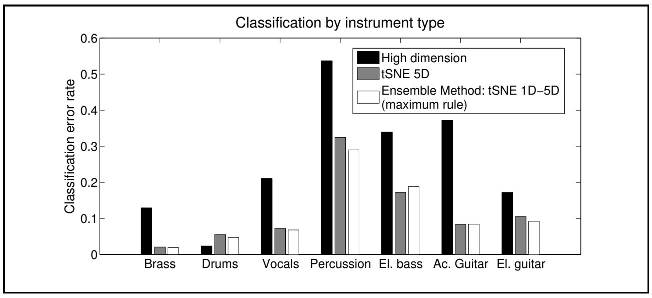

Figure 2. Classification error rates when using 10% of labeled data on each instrument class. Comparison between: (1) using the full high-dimensional feature set, (2) using a 5-dimensional feature set obtained through t-SNE on the full high-dimensional feature set, (3) Ensemble composed of 5 classifiers obtained using 1 to 5-dimensional feature set obtained through t-SNE on the full high-dimensional feature set; maximum rule for combination.

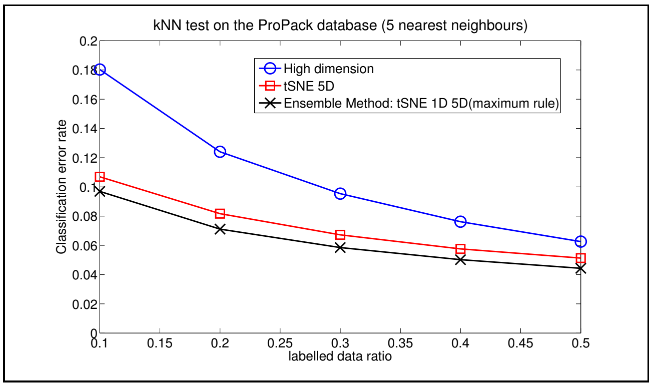

Figure 3. Classification error rates for various proportions of labeled data. Comparison between: (1) using the full high-dimensional feature set, (2) using a 5-dimensional feature set obtained through t-SNE on the full high-dimensional feature set, (3) Ensemble composed of 5 classifiers obtained using 1 to 5-dimensional feature set obtained through t-SNE on the full high-dimensional feature set; maximum rule for ensemble combination.

4.3 Results and Discussion

In Figure 1, we present classification results for the case 10% of the whole data set is labeled. It shows results for the baseline system, for a system where features are first processed using t-SNE (with dimensionality of 5), as well as for various ensembles and combination rules. We can observe that on this kind of data, an efficient dimensionality reduction scheme is a useful pre-processing step for semi-supervised classification. Classification performance on the reduced dimensional space is indeed better. A 40% relative reduction of the error rate is obtained.

The first proposed ensemble approach yields a further error rate reduction of 10% relative. All five combination rules bring some improvement, but best results are obtained using the simple maximum rule, followed by the sum rule. The two other proposed ensembles are unconclusive, and the product and minimum combination rules perform notably much worse. This can be explained by the discrepancy in classification performance of the ensemble members: base classifiers using the standard deviations of features in the first case, and using SF features alone in the second are much worse than the other base classifiers (detailed results not reported here). More complex or weighted combination rules may help in those cases.

Overall, the best single classifier is using a 5-dimensional feature set obtained through t-SNE on the full high-dimensional feature set. The best ensemble classifier is composed of 5 classifiers obtained using 1 to 5-dimensional feature sets obtained through t-SNE on the full high-dimensional feature set; and a maximum rule for ensemble combination. We then present more detailed comparisons of these two with the baseline classifier In Figure 2, we observe that some classes indeed strongly benefit from dimensionality reduction, while others do much less, or not at all. As suggested earlier in the text, the tradeoff between reducing the feature space dimension and preserving its representativity may depend on the class and on the region of the feature space. This also suggests further theoretical and empirical work in ensemble approaches that could account for non-uniform intrinsic dimensionalities of the data set manifolds. In Figure 3, we present results for various proportions of labeled data, showing that our conclusions hold when one can obtain labels for a larger part of the data set. With 50% of labeled date, t-SNE allows to reduce the error rate by 18% relative over high-dimensional features, and the ensemble approach further reduces it by 14% relative.

5. CONCLUSIONS

In this paper, we presented a new method for designing multiple classifiers system relying on non-linear dimensionality reduction through t-SNE, together with an experimental study of its performance on an audio-based musical instruments transductive classification task. We first observed that classification performance can be boosted when applying t-SNE as a pre-processing step, even when going down to as low as a few dimensions. Designing multiple classifiers by altering the dimensionality of the t-SNE embedding and combining them using a simple combination rule further improved the results.

These promising initial results invite further work, in particular in the application of other dimensionality reduction schemes and more complex ensemble combination rules, as well as in understanding how ensembles can be used for mitigating the tradeoff between denoising and feature preservation properties. The application of the proposed approach to larger scale data sets can also be the subject of future work, together with experimental evaluation on non-transductive tasks using out-of-sample extensions.

6. REFERENCES

[1] Amir Amit. Ensemble Classification via Dimensionality Reduction. Master's thesis, Efi Arazi School of Computer Science, Israel, 2011.

[2] Leo Breiman. Bagging predictors. Technical Report 421, Department of Statistics, University of California, Berkeley, California, September 1994.

[3] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[4] Gavin Brown, Jeremy L. Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.

[5] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[6] Stéphane Dupont, Thierry Ravet, Cécile Picard, and Christian Frisson. Nonlinear dimensionality reduction approaches applied to music and textural sounds. In *Proceeding of IEEE International Conference on Multimedia and Expo (ICME)*, San Jose, California, jul 2013.

[7] Xiaoli Zhang Fern and Carla E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *ICML'03*, pages 186–193, 2003.

[8] Geoffrey Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2003.

[9] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(8):832–844, August 1998.

[10] S. Lafon and A.B. Lee. Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(9):1393 –1403, sept. 2006.

[11] John A. Lee and Michel. Verleysen. *Nonlinear dimensionality reduction*. Springer, New York; London, 2007.

[12] Christopher J. Merz and Michael J. Pazzani. A principal components approach to combining regression estimates. *Mach. Learn.*, 36(1-2):9–32, July 1999.

[13] Nikunj C. Oza and Kagan Tumer. Dimensionality reduction through classifier ensembles. Technical Report NASA-ARC-IC-1999-124, National Aeronautics and Space Administration, Moffett Field, CA, 1999.

[14] Nikunj C. Oza and Kagan Tumer. Input decimation ensembles: Decorrelation through dimensionality reduction. In *LNCS*, pages 238–247. Springer, 2001.

[15] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification). in the CUIDADO project. Paris, IRCAM, 2004.

[16] Juan J. Rodriguez, Ludmila I. Kuncheva, and Carlos J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(10):1619–1630, October 2006.

[17] Niall Rooney, David Patterson, Alexey Tsymbal, and Sarab An. Random subspacing for regression ensembles. Technical report, Department of Computer Science, Trinity College Dublin, Ireland, 2004.

[18] Robert E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, July 1990.

[19] Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. The MIT Press, 2012.

[20] Alon Schclar and Lior Rokach. Random projection ensemble classifiers. In Joaquim Filipe and José Cordeiro, editors, *Enterprise Information Systems, 11th International Conference, ICEIS 2009, Milan, Italy, May 6-10, 2009. Proceedings*, volume 24 of *Lecture Notes in Business Information Processing*, pages 309–316. Springer, 2009.

[21] Kagan Tumer and Joydeep Ghosh. Linear and order statistics combiners for pattern classification. *CoRR*, cs.NE/9905012, 1999.

[22] Laurens van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[23] L.J.P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. Technical Report TiCC-TR 2009-005, Tilburg University, 2009.

[24] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, 2012.