A HIGHER-DIMENSIONAL EXPANSION OF AFFECTIVE NORMS FOR ENGLISH TERMS FOR MUSIC TAGGING

Michele Buccoli[1], **Massimiliano Zanoni**[1], **György Fazekas**[2]
Augusto Sarti[1], **Mark Sandler**[2]
[1] Dipartimento di Elettronica, Informatica e Bioingegneria, Politecnico di Milano, Italy
[2] Centre For Digital Music, Queen Mary, University of London, UK
michele.buccoli@polimi.it, g.fazekas@qmul.ac.uk

ABSTRACT

The Valence, Arousal and Dominance (VAD) model for emotion representation is widely used in music analysis. The ANEW dataset is composed of more than 2000 emotion related descriptors annotated in the VAD space. However, due to the low number of dimensions of the VAD model, the distribution of terms of the ANEW dataset tends to be compact and cluttered. In this work, we aim at finding a possibly higher-dimensional transformation of the VAD space, where the terms of the ANEW dataset are better organised conceptually and bear more relevance to music tagging. Our approach involves the use of a kernel expansion of the ANEW dataset to exploit a higher number of dimensions, and the application of distance learning techniques to find a distance metric that is consistent with the semantic similarity among terms. In order to train the distance learning algorithms, we collect information on the semantic similarity from human annotation and editorial tags. We evaluate the quality of the method by clustering the terms in the found high-dimensional domain. Our approach exhibits promising results with objective and subjective performance metrics, showing that a higher dimensional space could be useful to model semantic similarity among terms of the ANEW dataset.

1. INTRODUCTION

One of the fundamental properties of music is the ability to convey emotions [27]. Consequently, there is a great interest in representing and classifying music according to emotions in areas like music information retrieval, music recommendation and personalisation [11–13, 27]. It has been proven that listeners enjoy looking for and discovering music using mood-based queries, which represents 33% of music queries according to [13]. This is an important reason that urged psychologist and musicologist to investigate different paradigms for the representation and modelling of emotion related descriptors [11–13].

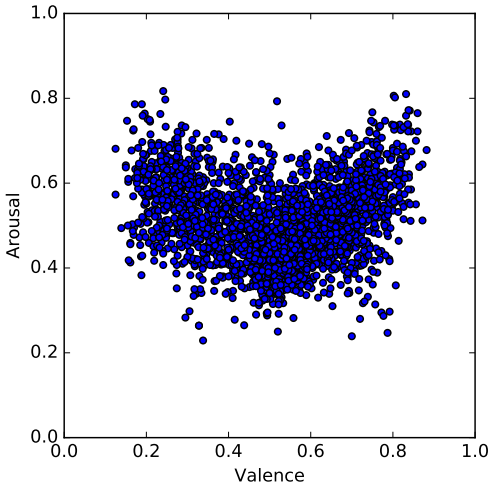

Figure 1: Distribution of the ANEW dataset in the VA space

Dimensional approaches to emotion conceptualisation focus on describing emotional states in a continuous space, where emotion states are represented as points or distributions in an N-dimensional space. Specific emotion terms can be localised in such continuous space and it is possible to define a metric in a way that the distance between points is proportional to the semantic distance between emotions. The most influential dimensional models so far [1, 25] are proposed by Russell [19] and Thayer [24]. Russell devised a *circumplex model of affect* which consists of a two-dimensional, circular structure involving the dimensions of *arousal* (A), linked to the degree of activation or excitement, and *valence* (V), linked to the degree of pleasantness. A third dimension related to *dominance* (D) was later proposed to express the degree of control and to possibly distinguish different and overlapping moods. [8, 21].

The increasing need of a continuous affective model led to the creation of the ANEW dataset [4] for Psychological research. This dataset is composed of 2,476 English words with positions in the VAD space. Despite it is a generic dataset, the large amount of terms in ANEW makes it a powerful resource also for the Music Emotion Recognition (MER) community, including applications for

Type	Symbol	Num Terms	Details
Terms Dataset	$\mathcal{D}_{ANEW}$	2,476	Mean in V, A, D dimensions
Implicit Distance	$\mathbf{M}_{ILM10K}$	240, 450	Tag compact representation from an LSA with $k = 10, 20, 50, 100$ components
Explicit Distance	$\mathbf{M}_{HD}$	180	Human annotation of semantic similarity between terms from ANEW
Explicit Clustering	$\mathbf{M}_{HC}$	100	Human clustering of a subset of ANEW

Table 1: Summary of the collected data

automatic music annotation and retrieval [5, 20]. Unfortunately, ANEW terms in the VAD space tend to have a very uniform and compact distribution concentrated around the centre of the space as shown in Figure 1 for the VA subspace. Although the use of a substantial set of terms provides a very representative model of a large variety of emotions, to deal with a compact and cluttered distribution can be problematic in many musical applications. For this reason, typically only a subset of the terms is used, leading to a loss in the exhaustiveness of the model. A higher-dimensional mood space drawn from the ANEW dataset, where the terms are distributed following some conceptual organisation can benefit several applications. These include semantic music annotation, recommendation and mood-based music retrieval [2].

In this study, we investigate the construction of higher-dimensional emotional spaces from ANEW by means of kernel expansion techniques applied to the VAD space. Although the transformation has the effect to produce a more sparse distribution of terms, it is not clear if the semantic distance between concepts is well represented by the metric in the new space. To solve this problem, we first aim to find a distance reflecting conceptual organisation of terms by applying distance learning techniques [3, 10, 26]. Given some constraints between terms, these methods search for a linear transformation of the space that is semantically relevant. That is, the ideal learnt distance closely correlates with semantic differentiae given by users in a specific task for a subset of the ANEW terms. We generate the set of constraints using "a priori" information collected through a subjective test where participants were asked to specify the semantic similarity between pairs of terms in the context of music. We then perform Latent Semantic Analysis on emotion related annotations for a large set of music pieces.

Under the hypothesis that the terms may be improved by this transformation, we validate the approach by clustering in the new high-dimensional space. We subsequently evaluate the resulting clusters with objective and subjective metrics. Tests show promising results given these new learnt distance metrics and provide an insight into how a better configuration in the high-dimensional space can be achieved.

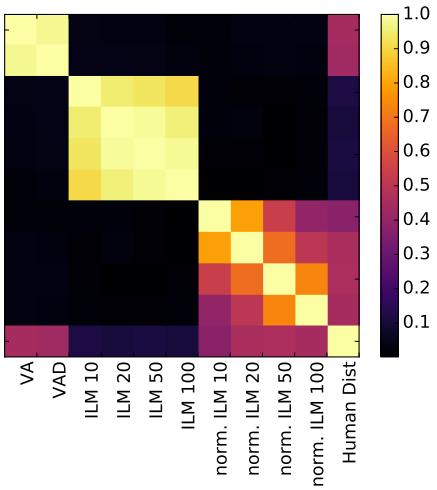

Figure 2: Absolute Pearson Correlation of the self-similarity matrices computed or collected from different sources of data.

2. DATASET CONSTRUCTION

In this section, we describe three datasets created to provide constraints for semi-supervised learning as well as to validate our approach. A summary of the collected data is provided in Table 1.

2.1 Terms in the ANEW Dataset

The terms in the ANEW dataset [4] is already annotated for the Valence, Arousal and Dominance dimensions with a value between 0 and 10 by Psychology class students. For each term, we consider the normalised average (between 0 and 1) of the annotations.

2.2 Implicit Distance Annotation

We compute a similarity distance matrix among emotion related descriptors by performing Latent Semantic Analysis (LSA) on the I-Like-Music [1] (ILM) dataset denoted ILM10K. This is composed of 10,199 tracks from commercial music annotated with crowed sourced editorial and social tags [2, 20] with weights corresponding to the prevalence of each tag.

From the tags in ILM10K, we first discard the tags that are not included in the ANEW dataset. We then filter the tags that are rarely used by means of two thresholds on the number of times the tag is used in ILM10K: 15 times, leading to a set of $T = 240$ terms and 5 times leading to $T = 450$ terms. We build a track-tag matrix using the remaining tags and compute a k-component approximation of this matrix via LSA, keeping different numbers of components k to test different degrees of approximations. We set $k = 10, 20, 50, 100$ components to produce the set of matrices $\mathbf{D}_{ILM10K} \in \mathbb{R}^{T \times k}$.

From $\mathbf{D}_{ILM10K}$, we compute the similarity between tags as the Euclidean distance between the correspondent

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

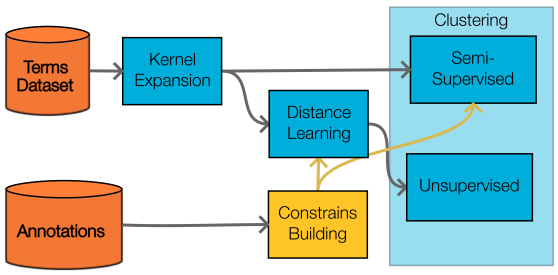

Figure 3: Block diagram of the clustering approach

rows, producing the matrix $\mathbf{M}_{ILM10K}$. We also compute the normalised matrix $\hat{\mathbf{D}}_{ILM10K}$, composed by the L^2 normalised rows of $\mathbf{D}_{ILM10K}$, so that the Euclidean distance of the former is equal to the cosine distance of the latter and compute the (cosine) distance matrix $\hat{\mathbf{M}}_{ILM10K}$.

2.3 Human Distance Annotation

We conducted an online survey and asked annotators to define the perceived mood similarity in the context of music between pairs of descriptors with a value between 0 (very similar) and 1 (not similar at all). 504 people participated in the survey. From the data we kept only the terms that received at least 2 annotations leading to $T = 180$ and we compose the sparse matrix $\mathbf{M}_{HD}$.

2.4 Human Clustering Annotation

We conducted a second online survey and asked annotators to group the set of top 100 descriptors in the dataset. 15 people participated in the survey, leading to the matrix $\mathbf{M}_{HC}$ with $T = 100$ terms, where each entry (i, j) indicates the number of people that grouped together the i-th and j-th terms.

2.5 Further Considerations

In Figure 2 we show the absolute Pearson correlations between the similarity matrices from the different sources of data we have mentioned so far. The human distance annotation exhibits a modest correlation with the Euclidean distance defined in the VA or VAD space. The annotations on distance between terms is also somewhat correlated with the distance that can be inferred from editorial tags. In the rest of this study, we aim to find a space where the defined distance metric provides a better representation of the perceived similarity in musical emotion.

3. HIGH-DIMENSIONAL SPACE LEARNING

The block diagram of our approach is shown in Figure 3. The VA or VAD coordinates for the ANEW dataset are processed using kernel expansion. In order to better represent the perceived similarity, we first rotate and translate the expanded dataset by means of distance learning algorithms.We then use the collected annotations to build constraints for learning the metric distance.

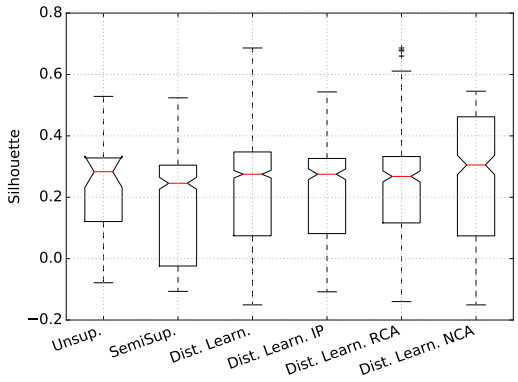

Figure 4: Boxplot of the Silhouette indices for the different scenarios

3.1 Kernel Expansion

Given a generic vector $\mathbf{x} \in \mathbb{R}^N$, we define its kernel expansion the vector $\phi_{\mathbf{x}} = \Phi(\mathbf{x}) \in \mathbb{R}^P$, with $P \geq N$, that is the result of the mapping function Φ. We expand our original dataset (both VA and VAD coordinates) with the following mapping functions: normalisation with respect to the L^2 norm, to include the cosine distance, i.e., $\Phi(\mathbf{x}) = \mathbf{x}/|\mathbf{x}|_2$; polynomial expansion with degrees two and three, to include nonlinear distance, i.e., $\Phi(x)_2 = [x_1, ..., x_N, x_1^2, ..., x_N^2, x_1x_2, ..., x_1x_N, ..., x_{N-1}, x_N]$ for the polynomial expansion of degree two, where $x_1, ..., x_N$ are the components of $\mathbf{x}$ (the third degree polynomial expansion is computed accordingly); approximate feature map of a Radial Basis Function (RBF) kernel by Monte Carlo expansion of its Fourier Transform [17].

3.2 Constrains Building

In order to facilitate the training of distance learning algorithms, we use similarity matrices obtained from human annotations introduced in Sections 2.2, 2.3, 2.4. Our method relies on the definition of a set of Must-Link (ML) and Cannot-Link (CL) constraints. Treating the similarity as a tendency of terms to be grouped together, ML constraints force a pair of terms to be in the same group, whereas CL constraints force a pair of terms to be in different groups. In the distance learning techniques, the space is transformed such that the pairs of points of the ML set are closer together while the pairs of points in the CL set are far apart.

As far as the ILM10K dataset is concerned, we compute the ML and CL constraints by computing the mean value μ and standard deviation σ of the matrices $\mathbf{M}_{ILM10K}$ and empirically defining two thresholds $th_l = \mu - 2\sigma$, $th_h = \mu + 2\sigma$. We compose the set of ML constraints by considering those pairs of terms which are close in the space defined by LSA, hence whose distance is lower than th_l. Similarly, we compose the CL set with those pairs of terms that are far from each other, i.e., whose distance is higher than th_h. We compose another set of constrains in the same way from the matrices $\hat{\mathbf{M}}_{ILM10K}$.

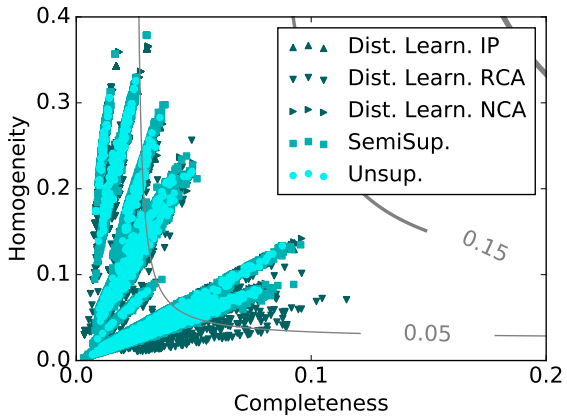

Figure 5: Completeness and Homogeneity results for the comparison of clusters with those obtained from the ILM10K dataset. Gray contour lines indicate the V-score

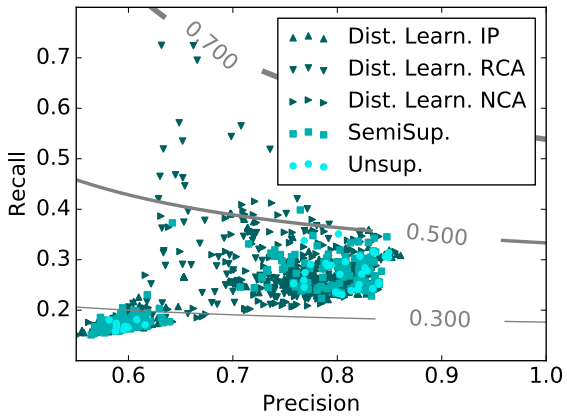

Figure 6: Precision and Recall results for the comparison of clusters with those obtained from human annotation. Gray contour lines indicate the F-measure.

We use a similar approach to compose the ML and CL constraints from the Human Distance annotations. We compute the mean value μ and standard deviation σ of the annotations in the matrix $\mathbf{M}_{HD}$ and we empirically define the thresholds $th_l = \mu - \sigma,\ th_h = \mu + \sigma$.

As far as the Human Clustering annotations are concerned, we compute the mean value μ and standard deviation σ of the non-zero entries of the matrix $\mathbf{M}_{HC}$, i.e., of the number of people who annotated two terms as belonging to the same clusters, and define the soft threshold $th = \{\mu - \sigma\}$. We compose ML with the pair of terms that have been grouped together by more than th people and we compose CL with the zero entries of $\mathbf{M}_{HC}$.

3.3 Distance Learning

Given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ two generic vectors, we can weight the contribution of each components and the inter-correlation among them by defining a Mahalanobis (symmetric, square) matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ and computing:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\intercal \mathbf{A} (\mathbf{x} - \mathbf{y})} = \sqrt{(\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{y})^\intercal (\mathbf{L}\mathbf{x} - \mathbf{L}\mathbf{y})}, \quad (1)$$

that is the Euclidean distance between the projected vectors over the subspace defined by $\mathbf{L}$, with $\mathbf{L}^\intercal \mathbf{L} = \mathbf{A}$.

Distance metric learning is the sub-field of machine learning that aims at finding the best subspace $\mathbf{L}$, from a set of constraints. Using the constraints as described in Section 3.2, we compute a set of subspaces for each combination of input, kernel expansions and constraints. It is worth remembering that with the Mahalanobis distance, only translation and rotation operations are performed. However, the application of kernels on the sample data allows nonlinear transformation of the original space of data.

We employ the following distance metric learning algorithms [6]:

- Iterative Projection [26] (IP), computes the Mahalanobis matrix by means of an iterative minimisation of the distance of the ML data with the constraint to keep CL data far apart
- Relative Components Analysis [3] (RCA), learns a Mahalanobis matrix that assigns large weights to more relevant components and low weights to irrelevant ones, by using chunklets, i.e., subset of data that belong to the same group (i.e., have been defined in the ML set)
- Neighbourhood Components Analysis [10] (NCA) is a component analysis that computes an optimal Mahalanobis matrix for the K-nearest neighbour clustering algorithm.

4. EXPERIMENTAL SETUP AND EVALUATION

As previously discussed, our aim is to find a transformation of the ANEW space with improved conceptual organisation of terms that is relevant in a musical context. We expect terms that are semantically similar in this context to be close and dissimilar terms to be far apart. For this reason, we validate our approach by clustering in the transformed space.

Specifically, we perform three evaluations: i) we evaluate the resulting clusters by mean of the Silhouette quality index as objective metric; ii) we compare the resulting clusters with those obtained by clustering the tags of the ILM10K dataset; iii) we compare the resulting clusters with those obtained from manual annotation.

In the ILM10K dataset we consider the results obtained with both $T = 240$ and 450 tags. This is because the former is less noisy, while the latter provides more information and constrains. We will then specify in the single cases which dataset achieved the mentioned results.

4.1 Unsupervised and Semi-Supervised Clustering

In order to provide a robust evaluation, we apply several clustering techniques. We experimentally choose to retrieve the 6 best representative clusters, given the large

Scenario	Features	Algorithm	Constraints	Silhouette
Unsupervised	Norm. VA	K-Means		0.5287
Semi-Supervised	Norm. VA	SC	$\mathbf{M}_{HC}, th = \mu - \sigma$	0.5240
IP Distance Learning	VAD	kmeans	$\mathbf{M}_{HD}$	0.5432
RCA Distance Learning	VAD, poly 3 degree	AHC	$\mathbf{M}_{HC}, th = \mu - \sigma$	0.6864
NCA Distance Learning	VA, poly 2 degree	kmeans	$\hat{\mathbf{M}}_{ILM10K}, T = 240, k = 10$	0.5456

Table 2: Best results for the Silhouette metric

Scenario	Features	Algorithm	Constraints	ILM10K	Compl.	Hom.	V-score
Unsup.	VA, poly 3 degree	SC		AHC $\hat{\mathbf{M}}_{ILM10K}, k = 100$	0.0877	0.1334	0.1058
Semi-Sup.	VAD poly 2 degree	AHC	$\mathbf{M}_{HD}$	AHC $\hat{\mathbf{M}}_{ILM10K}, k = 100$	0.0955	0.1347	0.1118
IP Dist.	VA	SC	$\mathbf{M}_{HD}$	AHC $\hat{\mathbf{M}}_{ILM10K}, k = 100$	0.0929	0.1371	0.1107
RCA Dist.	VAD	kmeans	$\mathbf{M}_{HD}$	AHC $\hat{\mathbf{M}}_{ILM10K}, k = 100$	0.0869	0.1290	0.1038
NCA Dist.	VA, poly 2 degree	AHC	$\mathbf{M}_{HD}$	AHC $\hat{\mathbf{M}}_{ILM10K}, k = 100$	0.0959	0.1421	0.1145

Table 3: Best results of the Homogeneity, Separation and V-measure metrics for the comparison with the clusters generated by ILM10K dataset (240 terms)

number of configurations we need to test. We employ the following algorithms resulted to be effective in the context of music tag analysis and aggregation (from [16]):

- K-Means [14] is the common unsupervised clustering algorithm;
- Semi-Supervised Non-negative Matrix Factorisation [7] (SS-NMF) applies NMF techniques to the self-distance matrix of the samples;
- Spectral Clustering [22] (SC) employs a low-dimension reduction of the similarity matrix between samples before applying the K-means algorithm;
- Agglomerative Hierarchical Clustering [23] (AHC) performs a bottom-up clustering: initially every sample is a cluster, then the closest clusters merge together until the final number of clusters is reached.

In this study we use the SS-NMF, SC and AHC algorithms in both unsupervised and semi-supervised fashion. In semi-supervised algorithms, the clustering can be guided by constraints. Here we use the ML and CL constraints computed in Section 3.2.

In order to validate the actual contribution of the distance learning techniques, we compare the results of our approach with the results obtained with both unsupervised and semi-supervised clustering of the ANEW dataset. Hence, we have three scenarios: i) unsupervised clustering of the transformed (but not learned) space (Unsup.) ; ii) the semi-supervised clustering of the transformed (but not learned) space (SemiSup.); iii) our approach, that is the unsupervised clustering of the learned space (Dist. Learn). Please note that the first scenario also includes the clustering of non-transformed space using the Euclidean distance.

4.2 Objective metrics

We evaluate the objective quality of clustering by using the Silhouette index that is defined as [16]:

$$\text{silhouette} = \frac{1}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} \frac{b - a}{\max(a, b)} \in [-1, 1], \qquad (2)$$

where a and b are the mean distance between the s-th sample and all other points in the same cluster and in the next nearest cluster, respectively and $\mathcal{D}$ is the dataset of samples. High (positive) values of Silhouette indicate dense well-separated clusters, values around 0 indicate overlapping clusters and low (or negative) values indicate incorrect clusters. In Figure 4 we show the boxplots of Silhouette metric for the different scenarios, while in Table 2 we show the configurations that generate the best results for each scenario.

It is clear that the application of Distance Learning techniques outperforms unsupervised and semi-supervised techniques, even with different configurations of kernels, algorithms and constraints. In particular, the best performance is obtained with the AHC over the third degree polynomial expansion of the VAD dataset, with a translation learned using the RCA technique with Human Clustering.

We can also notice that the semi-supervised scenario performs on average worse than the unsupervised scenario. This is because the Silhouette index evaluates the resulting clusters with respect of the position of the input data, that is not moved from the original position. The estimated clusters are therefore more noisy, which confirms the advantage of distance learning techniques to transform the space.

4.3 Subjective metrics - Comparison with Clustering of ILM10K

We compare the clusters obtained with the different approaches with the clusters obtained from ILM10K data. We consider the homogeneity and completeness metrics [18]. The former evaluates whether each estimated cluster contains only members of a group in the ground truth, while the latter estimates whether the samples of a given group belongs to the same estimated cluster. We also consider the V-measure, i.e., the harmonic mean of homogeneity and completeness. To avoid overfitting issues, we evaluate only the configurations that are not trained with the constraints generated from the ILM10K dataset.

Scenario	Features	Algorithm	Constraints	P	R	F
Unsup.	VAD, RBF	kmeans		0.8012	0.3512	0.4883
Semi-Sup.	VAD, RBF	AHC	$\mathbf{M}_{HD}$	0.7646	0.3986	0.5240
IP Dist.	VAD, RBF	AHC	$\hat{\mathbf{M}}_{ILM10K}, T = 240, k = 100$	0.6979	0.3915	05016
RCA Dist.	VA, poly 3 degree	AHC	$\hat{\mathbf{M}}_{ILM10K}, T = 240, k = 50$	0.6622	0.7241	0.6918
NCA Dist.	VAD, RBF	kmeans	$\hat{\mathbf{M}}_{ILM10K}, T = 240, k = 10$	0.7971	0.4124	0.5289

Table 4: Best results for the Precision, Recall and F-measure metric for the comparison with the human annotated clusters

We show the distribution of results for the different scenarios in Figure 5. We notice the results are fairly low, showing that the organisation given by the expanded and possibly learnt ANEW dataset is very different from that obtained from editorial tags on a real music annotation application. This confirms the necessity to find a space for the ANEW dataset which is more useful for MIR applications. However, it is clear that our approach can only slightly improve the task. In Table 3 we list the configurations that lead to the best performance. These are all obtained with the 240-term subset of the ILM10K dataset. The AHC over the normalised ILM10K dataset with 100 components is the one that best matches all scenarios, from which we can infer that the clustering of the ANEW dataset resembles a clustering based on cosine distance with a large number of components. The best results is reached using NCA technique with a AHC over the polynomial- expanded dataset, by means of annotated distance again.

4.4 Subjective metrics - Comparison with Human Annotations

We finally compare the obtained clusters with those collected from human annotations. Since the correctly grouped terms, i.e., the amount of True Positive (TP) examples are more specific and relevant than the correctly non-grouped ones (True Negative, TN), we consider the Precision (P) and Recall (R) metrics that focus on the number of TP examples. The Precision indicates the ratios of True Positive over the total estimated assignments, i.e., $P = |TP|/(|TP| + |FP|)$, while the Recall defines the number of TP over the total assignments in the ground truth, i.e., $R = |TP|/(|TP| + |FN|)$. We also consider the F-measure (F) as the harmonic mean of the two metrics [15]. To avoid overfitting, we evaluate only configurations that are not trained with the constraints generated from the human annotations on clustering.

In Figure 6 we show the distribution of results for the different scenarios and configurations. In general, Precision is higher than Recall, showing that the estimated clustering is not capable of retrieving all the corrected groups. The F-measures are mostly distributed between 0.3 and 0.5, which is an average result. We can clearly see that some configuration from RCA Distance Learning are able to improve the average Recall and improve the F-measure reaching almost at 0.7. In Table 4 we list the best configurations for each scenarios. The RCA Distance Learning technique clearly outperforms the other approaches and matches very closely the human annotations, i.e., the human way to organise the emotional-related descriptors. We can notice that once again, the best results are obtained by using the AHC over some polynomial expansion of the dataset. However, the constraints based on distance annotations are less helpful for emulating the human organisation of terms than the ones based on editorial and social tags (ILM10K).

5. CONCLUSIONS

We introduced a novel approach consisting of kernel expansion, constraint building from music task specific human annotation and finally distance learning to transform the ANEW dataset and obtain a distribution of terms with better conceptual organisation compared to the conventional VA or VAD space. This facilitates applications that require computer representation of music emotions, including music emotion recognition and music tagging, music recommendation or interactive musical applications such as [9] and [2].

Average and maximum results presented in Section 4 prove that distance learning techniques are effective in the improvement of the organisation of concepts of the ANEW dataset. In particular, hierarchical clustering of the RCA-learned space based on the polynomial expansion of VA/VAD space outperforms the other configurations. The paper also introduces a new task along with a set of techniques that proved successful as a first attempt. This provides useful insight into improving the organisation of mood terms to better reflect application specific semantic spaces. Future work involves the collection of additional "a priori" information for improving the robustness of evaluation, as well as further assessment of the constructed high-dimensional space within other MIR applications.

6. ACKNOWLEDGEMENTS

This work was part funded by the FAST IMPACt EPSRC Grant EP/L019981/1 and the EC H2020 research and innovation grant AudioCommons (688382). Sandler acknowledges the support of the Royal Society as a recipient of a Wolfson Research Merit Award.

7. REFERENCES

[1] A. Aljanaki, Y.-H. Yang, and M. Soleymani. Emotion in music task at mediaeval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop*, 2015.

[2] A. Allik, G. Fazekas, M. Barthet, and M. Sandler. mymoodplay: an interactive mood-based music discovery app. In *proc. 2nd Web Audio Conference (WAC)*, 2016.

[3] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In *Proc. of the 20th International Conference on Machine Learning (ICML)*, 2003.

[4] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Stimuli, instruction manual and affective ratings. Technical report, C-1, The Center for Research in Psychophysiology, University of Florida., Gainesville, FL, USA, 1999.

[5] M. Buccoli, M. Zanoni, A. Sarti, and S. Tubaro. A music search engine based on semantic text-based query. In *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2013.

[6] C.J. Carey and Y. Tang. metric-learn python package. http://github.com/all-umass/metric-learn, 2015.

[7] Y. Chen, M. Rege, M. Dong, and J. Hua. Non-negative matrix factorization for semi-supervised data clustering. *Knowledge and Info. Systems*, 17(3):355–379, 2008.

[8] R. Cowie, G. McKeown, and E. Douglas-Cowie. Tracing emotion: an overview. *International Journal of Synthetic Emotions, Special Issue on Benefits and Limitations of Continuous Representations of Emotions*, pages 1–17, 2012.

[9] G. Fazekas, M. Barthet, and M. Sandler. Novel methods in facilitating audience and performer interaction using the mood conductor framework. *Lecture Notes In Computer Science (LNCS): Sound, Music, and Motion*, 8905:122–147, 2014.

[10] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. Salakhutdinov. Neighbourhood components analysis. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 513–520. MIT Press, 2005.

[11] P. N. Juslin and J. A. Sloboda, editors. *Music and Emotion Theory and Research*. Series in Affective Science. Oxford University Press, 2001.

[12] P. N. Juslin and J. A. Sloboda. *Handbook of Music and Emotion: Theory, Research, Applications*. OUP Oxford, 2011.

[13] J. A. Lee and J. S. Downie. Survey of music information needs, uses, and seeking behaviors: preliminary findings. In *Proc. of the 5th International Society for Music Information Retrieval (ISMIR)*, 2004.

[14] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*. Oakland, CA, USA., 1967.

[15] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] A. Rahimi and B. Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems 22*, 2009.

[18] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proc. of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[19] J. A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[20] P. Saari, G. Fazekas, T. Eerola, M. Barthet, O. Lartillot, and M. Sandler. Genre-adaptive semantic computing and audio-based modelling for music mood annotation. *IEEE Transactions on Affective Computing*, (99):1–1, 2015.

[21] K. R. Scherer. Which emotion can be induced by music? what are the underlying mechanisms? and how can we measure them? *Journal of New Music Research*, 33(5):239–251, 2004.

[22] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[23] R. Sibson. Slink: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

[24] J.F. Thayer. Multiple indicators of affective responses to music. *Dissertation Abst. Int.*, 47(12), 1986.

[25] F. Weninger, F. Eyben, and B. Schuller. On-line continuous-time music mood regression with deep recurrent neural networks. In *proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[26] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 15:505–512, 2003.

[27] Y.-H. Yang and H. H. Chen. *Music Emotion Recognition*. CRC Press, 2011.