2015-08

Concentric RadViz: visual exploration of multi-task classification

Conference on Graphics, Patterns and Images, 28th, 2015, Salvador.
http://www.producao.usp.br/handle/BDPI/49478

Concentric RadViz: Visual Exploration of Multi-Task Classification

Jorge Henrique Piazentin Ono, Fabio Sikansi, Débora Cristina Corrêa
Fernando Vieira Paulovich, Afonso Paiva, Luis Gustavo Nonato
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
São Carlos, Brazil

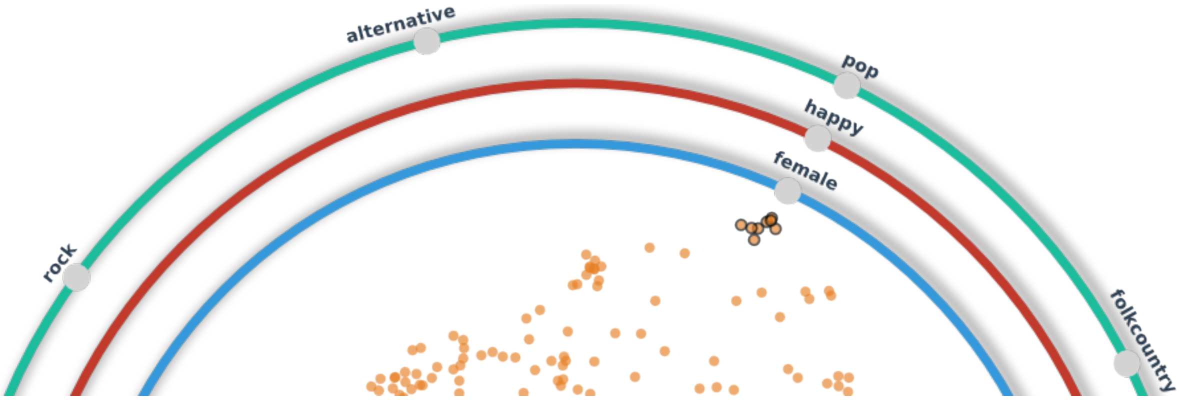


Fig. 1. Concentric RadViz visualization of a song data set. The user can arrange classification tasks (groups of dimensions) in concentric circles and explore the database by moving dimensions on the screen. A group of happy pop songs by female singers is selected.

***Abstract*—The discovery of patterns in large data collections is a difficult task. Visualization and machine learning techniques have emerged as a way to facilitate data analysis, providing tools to uncover relevant patterns from the data. This paper presents Concentric RadViz, a general purpose class visualization system that takes into account multi-class, multi-label and multi-task classifiers. Concentric RadViz uses a force attenuation scheme, which minimizes cluttering and ambiguity in the visual layout. In addition, the user can add concentric circles to the layout in order to represent classification tasks. Our validation results and the application of Concentric RadViz for two real collections suggest that this tool can reveal important data patterns and relations. In our application, the user can interact with the visualization by selecting regions of interest according to specific criteria and changing projection parameters.**

***Keywords*-RadViz; Multi-task classification; Information Visualization;**

I. Introduction

Projection-based layouts have become a fundamental entity in multidimensional data visualization. The main reason for the growing interest in projection-based visual mechanisms is that users can rely on the layout produced by those methods to identify clusters, patterns, and trends in the data [1]. Up-to-date visualization systems combine projection methods with other mathematical and computational resources to further improve the capability of visual analytics tools in revealing important information hidden in the data.

In this context, there has been much effort towards joining projection-based methods with machine learning tools [2], mainly for Active Learning (AL) purposes. For instance, Teoh *et al.* [3] developed PaintingClass, a system that projects multidimensional data onto a visual space while providing interactive resources for users to steer the construction of a decision tree, from which one can uncover important information. Geng *et al.* [4] proposed Supervised Isomap, a visualization tool that takes into account class information to define the distance metric used in the projection process, resulting in layouts with good class separation. Seifert *et al.* [5] adopted RadViz [6] to visualize uncertainties in the results of a classifier. The visualization assists a user-driven active learning system where users can select points to be labeled according to their position in the RadViz circle. In the context of text retrieval, Heimerl *et al.* [7] developed a system that projects document databases in the visual space using Principal Component Analysis and enables non-experts to interactively train classifiers. A common characteristic of the above methods is the use of one classification task to assist the visual analytic process. More specifically, those methods assign a single label to each data instance, therefore enforcing each instance to belong to only one class.

However, in many applications, data instances are allowed to belong to more than one class simultaneously, which is known in the literature as *multi-label classification problem.*

For instance, a movie can be labeled as both Comedy and Romance [8], while text documents can simultaneously be labeled as a scientific paper and as a historical evidence [9]. A closely related problem is the so-called *multi-task classification*, in which several joint classifications are performed simultaneously [10]. A typical example is the prediction of user's musical preferences [11], where a classifier is trained on low-level sound features to simultaneously predict musical genre, mood, and "danceability". In addition, most multi-label/multi-task classifiers perform a "fuzzy" classification by assigning probabilities to instances rather than a hard set of labels. More specifically, given an instance and a set of classes the classifier estimates the probability of the instance to belong to each class from the set. Visualization techniques that relies on the good properties of projection schemes while still being able to deal with multi-label/multi-task classifiers are not so abundant, and there is a clear need for new solutions able to operate in this scenario.

In this work we present *Concentric RadViz*, a novel projection-based technique designed to operate with multi-label/multi-task classifiers during the visualization process. The proposed method uses RadViz as projection mechanism. However, in contrast to other RadViz-based approaches, our method is able to handle multi-task data while producing better layouts in terms of cluttering and ambiguities, which are commonly present in projection-based visualizations.

The main contributions of our methodology can be summarized as follows:

1) A new filtering mechanism that improves RadViz output considerably, reducing visual cluttering and ambiguities in the projection;
2) An interactive concentric circles scheme that allows user to properly visualize, explore, and organize multi-task data in the visual space;
3) A set of experiments and case studies that show the effectiveness of our methodology as a visual analytic tool.

II. Related Work

The literature presents several methods devoted to improve the original RadViz visualization scheme proposed by Hoffman *et al.* [6]. In this section we first review the main aspects of the original RadViz method and then discuss relevant extensions described in the literature.

A. Algorithm and properties

RadViz is a visualization technique inspired by the mass-spring force mechanism to map multidimensional data into a two-dimensional visual space. The algorithm operates as follows: each data attribute (coordinate) is initially normalized to the interval $[0, 1]$. Reference points, called Dimensional Anchors (DA), are then placed over a circle, where each DA represents an attribute dimension of the data. Therefore, n DAs are needed to map n-dimensional data. Each instance is then attached to the DAs using springs. The strength of the spring connecting an instance i to a DA j is proportional to the value of the j^{th} attribute from instance i. The location of each instance in the visual space is given by the position where the system of forces reaches equilibrium. Letting v_{ij} be the value of the j^{th} attribute of the instance i and $\vec{S}_j$ be the position of the j^{th} DA, the location of the instance i is given by

$$\vec{x}_i = \frac{\sum\limits_{j=0}^{d} \vec{S}_j v_{ij}}{\sum\limits_{j=0}^{d} v_{ij}}.$$

Several geometrical properties of normalized radial visualizations like RadViz were proven by Daniels *et al.* [12], as for example: d-dimensional lines maps to lines or points, hyperspheres are mapped to ellipses, and hyperplanes are mapped to bounded polygons.

Other two properties render RadViz an important visualization technique. RadViz can map high-dimensional data with thousands of dimensions to the visual space in a very robust manner [2]. Furthermore, RadViz is inherently interactive: the DAs may be moved freely over the circle, thus the mapping can be updated according to user interaction [13].

The arrangement of the DAs is crucial to the quality of the final layout. Ankerst *et al.* [14] formulated the optimal Dimensional Arrangement Problem as an adaptation of the Travelling Salesman Problem (TSP), showing its NP-completeness. In order to solve it, they proposed different methods to measure the dissimilarity between dimensions, then used an ant colony optimization heuristic to solve the resulting TSP. McCarthy *et al.* [2] proposed a t-statistic metric to compute how effectively each dimension discriminates classes, arranging the DAs so as to optimize the metric. Recently, Di Caro *et al.* [15] presented two different formulations to the Dimensional Arrangement Problem which have a higher probability of finding the global optimum than the original formulation.

B. RadViz Extensions

One of the main drawbacks of RadViz is that many data instances can be mapped to the same point. In fact, instances lying on a line through the origin in the original space will be mapped to the same point in the visual space. This fact hampers the visualization of cluster whose centroids are close to a line through the origin, since they will not be separated in the visual space [12]. Nováková *et al.* [16] proposed an interactive preprocessing alternative to handle this problem: users are allowed to rotate/mirror some dimensions in the original space before performing the RadViz mapping, which can result in better projections. Daniels *et al.* [12] automated this process with a genetic algorithm that performs a large number of rotations and selects the one with best cluster separation. Although these approaches do improve cluster separation, DAs in the resulting visualization do no correspond to the original dimensions, hindering the direct interpretation of the layout.

Sharko *et al.* [13] proposed the Vectorized RadViz, a methodology that enables the visual evaluation of cluster ensembles. With this technique, it is possible to identify patterns such as similar clusters obtained from different techniques as well as clusters that are unstable in the data set. Vectorized RadViz works by partitioning categorical dimensions (attributes with n possible values are transformed in n binary attributes) and positioning each new dimension independently as a DA.

Three-dimensional extensions of RadViz have also been proposed. In [16], a third dimension is added to the RadViz circle and the distance of each instance from the origin in the original data space is mapped to this new dimension. SphereViz [17] maps high dimensional data to the interior of a sphere using a virtual reality environment. The mapping is performed similarly to RadViz: DAs are positioned over the surface of a sphere and the normalized data is interpolated using a spring-based layout.

III. Method

In this section we describe two modifications we propose to the traditional RadViz algorithm, which are designed to enable the visual exploration of multi-class, multi-label and multi-task classification results. Similarly to [5], we use RadViz to visualize class probability estimations. However, we are interested in exploring classification results rather than using an active learning-like procedure to update classification models.

A. Concentric RadViz

Multi-task classification is the problem of performing multiple joint classifications simultaneously [10]. In this context, RadViz can be used to visualize the result of a classification by using classes as DAs and the probabilities of instances belonging to each class as weigths. Traditional RadViz cannot represent multiple classifications at the same time, even if the DAs are sorted according to a similarity measure, like in [15]. Consider, for example, a multi-task song classification data consisting of song genres (eg. rock, pop, etc.), singer gender (male, female) and mood (party, aggressive, or relaxed). RadViz would display the three classification tasks in a single circular visualization, which leads to several problems:

1) Binary classification tasks would exert more influence in the position of each point than multi-class classification tasks (on average, class probabilities are higher in the binary classification problem);
2) Classes from different tasks would be mixed in a single circle, thus hampering visual interpretation and interaction;
3) DAs with opposite meanings could be placed close to each other on the circumference of the circle, presenting incorrect information to the user.

In order to solve these problems and visualize more than one prediction at the same time, we proposed Concentric RadViz, an adaptation of RadViz in which each classification task is represented by a particular circle, hereby called Dimension Group (DG). The DGs are depicted as concentric circles in the visual space, as illustrated in Fig. 1. Each instance is associated to a m-dimensional vector where each coordinate corresponds to a class, therefore, m is total number of classes in all DGs. The value of each coordinate is the probability of the instance to belong to the corresponding class. Therefore, the instance is mapped to the visual space using the corresponding m-dimensional vector, but taking into account the position of the DAs on the concentric circles. DAs are positioned on each circle by solving a Traveling Salesman Problem for each DG, as proposed by [15].

Each dimension of the m-dimensional vector of each instance must be normalized as done in the traditional RadViz method. Moreover, in order to deal with multiple DGs, the m-dimensional vector is also normalized by parts according to the DGs (the normalization is performed per instance of data). More specifically, each group of DG coordinates/classes is normalized so as to ensure that the maximum value for that group of coordinates is equal one. This second normalization guarantees that each individual DG influences equally in the projection.

With Concentric RadViz, the user can combine classes from different groups in order to explore the data and easily extract relevant information. For instance, in Fig. 1, classes “pop”, “happy” and “female” are aligned in order to identify pop songs by female singers with happy mood.

We now discuss an improvement to RadViz algorithm, which reduces clutter in the center of the circle and ambiguity in the visualization of multi-class and multi-label data sets.

B. RadViz with Sigmoidal Weighting

In the traditional RadViz, all DAs contribute to the position of the mapped points. In the context of class probabilities visualization, this may not be interesting. Consider the case of a multi-class problem where an instance i belongs to class A with probability 0.5, therefore, the probability of i to belong to all other classes sum up to 0.5. The classifier is fairly certain the instance of data is of class A, but in RadViz, the point can be pulled towards the center of the circle, as illustrated in Fig. 2.

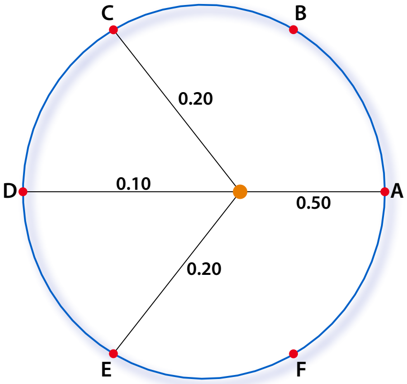


Fig. 2. Cancellation of forces in RadViz with class probabilities.

In order to tackle this problem, we propose a filtering mechanism that excludes low class probabilities from the visualization. For each instance, we scale DA values from all DGs so that the maximum value becomes equal to one. The filter operates by multiplying each dimension v_{ij} with a zero-one normalized sigmoid function

$$\hat{\sigma}(x,s,t) = \begin{cases} \frac{\sigma(x,s,t)-\sigma(0,s,t)}{\sigma(1,s,t)-\sigma(0,s,t)} & \text{, if } \sigma(1,s,t) \neq \sigma(0,s,t) \\ 1 & \text{, otherwise} \end{cases}$$

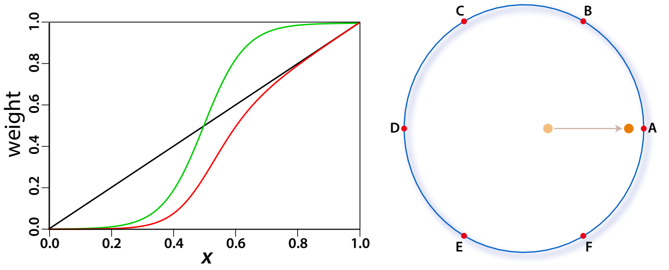
$$\text{with} \quad \sigma(x,s,t) = \frac{1}{1+\exp(-s(x+t))}\,.$$

The scale parameter s, $s \geq 0$, and the translation parameter t, $-1 \leq t \leq 1$, control the threshold after which dimensions with low values are canceled. Finally, the mapping weighted by normalized sigmoid function $\hat{\sigma}$ is given by

$$\vec{x}_i = \frac{\sum_{j=0}^{d} \vec{S}_j v_{ij} \hat{\sigma}(v_{ij},s,t)}{\sum_{j=0}^{d} v_{ij} \hat{\sigma}(v_{ij},s,t)}\,,$$

where $\vec{S}_{ij}$ is the DA position and v_{ij} is the value of coordinate j from instance i.

The user can interactively control the scale and translation parameters of the sigmoid. As the area under the normalized sigmoid curve becomes smaller, more cluttered is the visualization. However, clutter tends to occur close to the DAs, which reduce ambiguities of points in the center of the circles. To illustrate this situation, consider Fig. 2: the original weights are used to place the point, therefore it is a situation equivalent to projecting the instances with parameters $s = 0$ and $t = 1$. If we cancel some of the springs by setting parameters $s = 15$ and $t = -0.5$, the point is attracted towards DA A. Fig. 3 shows the effect of a sigmoidal weighting: Fig. 3a shows the original weight x, the normalized sigmoid function $\hat{\sigma}(x)$ and the influence of $\hat{\sigma}(x)$ on the original weight given by $x\hat{\sigma}(x)$. Fig. 3b shows the new point position after sigmoidal weighting.



(a) x (black), $\hat{\sigma}$ (green) and $x\hat{\sigma}$ (red).

(b) RadViz projection.

Fig. 3. RadViz with sigmoidal weighting: $s = 15$ and $t = -0.5$.

IV. Results and Validation

In this section we present some results towards validating the proposed method. In order to focus our analysis only on the visualization tool, we assume that the data under analysis has been classified by a method that satisfies the following assumption: *the classifier is highly accurate in terms of estimating class probabilities for every instance*. Note that no such classifier exists and hence, the proposed visualization can be at most just as accurate as the algorithm employed in the data analysis task. Therefore, our validation approach measures how well the visualization represents the classifier output.

We validate Concentric RadViz using two well-known information retrieval measures, namely, the Mean Average Precision (MAP), and the R-Precision [18]. MAP is used to assess the quality of neighborhoods present in the projection layout, that is, we measure how similar neighbor points are in terms of their multiple classification. R-Precision is employed to measure the alignment of projected points with respect to the DAs.

We perform the experiments using two distinct data sets: the Dortmund audio benchmark data set [19], and the multi-task facial landmark (MTLF) data set [20]. The Dortmund data set contains 1,886 audio recordings divided into nine genres: Alternative (145), Blues (120), Electronic (113), Folk-Country (222), Funk-Soul-R&B (47), Jazz (319), Pop (116), Rap-Hiphop (300) and Rock (504). We use the Essentia library pre-trained models [21] to classify and obtain the probabilities of each instance with respect to genre, gender (male|female), mood (happy|sad), instrument (voice|instrumental), type (party|aggressive|relaxed) and timbre (dark|bright). The accuracy of Essentia in performing such classifications is reported in [22], [23].

The MTFL data set contains 12,995 face images, which are annotated with four attributes: gender (male, female), head pose (left profile, left, frontal, right, right profile), wearing glasses (with glasses, without glasses), and smiling (smiling, not smiling). In order to predict class probabilities in this data set, we used the Random Forest Classifier [24], implemented in the *Scikit-Learn* Machine Learning Toolbox [25], with default parameters.

The MAP is used to assess if neighbor points in the visual space tend to belong to the same class. In order to be fair when computing the MAP, we aim to project points so as to generate a layout where points are as spread as possible. A satisfactory spreading can be obtained by rotating the concentric circles such that the sum of the distances between DAs from distinct circles is maximal (recall that the DAs are arranged in each circle by solving the Traveling Salesman Problem). Once the concentric circles are properly rotated, we computed the MAP by ranking the neighbors of each point p according to their Euclidean distance to p. The average precision (AP) of p is then computed and the final MAP score obtained by averaging the AP of all points. A neighbor of p is considered relevant (the same class as p) if both set of labels match exactly.

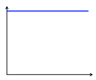
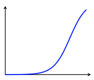
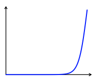
Table I presents the MAP obtained for both data sets. In order to compare our results with the original RadViz, we performed eight experiments with distinct sigmoid parameters. The experiments were run varying the number of DGs used in the visualization as follows:

1) Dortmund data set:
   a) Genre;

   b) Genre & Mood;
   c) Genre & Mood & Gender;
   d) Genre & Mood & Gender & Instrument;

2) MTFL data set:
   a) Head Pose;
   b) Head Pose & Gender;
   c) Head Pose & Gender & Glasses;
   d) Head Pose & Gender & Glasses & Smile.

TABLE I
MEAN AVERAGE PRECISION FOR RADVIZ AND CONCENTRIC RADVIZ.

	RadViz	Concentric RadViz		
		$\hat{\sigma}(x, 0, 1)$ 	$\hat{\sigma}(x, 10, -0.8)$ 	$\hat{\sigma}(x, 20, -1)$ 
1-a)	0.7141	0.7141	0.8846	0.9062
1-b)	0.5377	0.7115	0.8201	0.8989
1-c)	0.3842	0.6567	0.7128	0.8464
1-d)	0.2783	0.6905	0.7257	0.7329
2-a)	0.9150	0.9150	0.9787	0.9852
2-b)	0.6350	0.7010	0.8840	0.9110
2-c)	0.5709	0.5937	0.8113	0.8368
2-d)	0.3574	0.3921	0.6975	0.7500

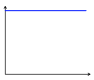
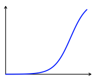
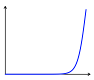
For the sake of comparison, we run the original RadViz with DAs, from all DGs, arranged on a single circle by solving the TSP optimization scheme. The results presented in Table I show that the original RadViz does not properly handle multiple classification tasks as Concentric RadViz does. For example, when 4 DGs are visualized, the MAP of the original RadViz decreases considerably, while the MAP of Concentric RadViz is approximately twice better when low probability classes are filtered out.

The second validation tests make use of the R-Precision metric to assess how closely projected points aligns with the DAs. In this case, the data is projected using Concentric RadViz and queries are accomplished by rotating the dimension groups so as to align specific classes. For instance, in the music data set, a query can be triggered by aligning pop, female, and happy from the DGs genre, gender, and mood, respectively, as illustrated in Fig. 1. Instances whose labels matches the triple (pop,female,happy) are considered relevant, that is, a valid retrieved element. Table II shows R-precision values for eight queries/alignments with RadViz and Concentric RadViz with three sigmoid parameters. For the original RadViz, all classes were placed on the same circle using the TSP independently for each DG. Then, to enforce the alignment, the chosen DAs are placed on top of each other. We have performed tests using the following alignments:

1) Dortmund data set:
   a) Genre: Pop;
   b) Genre: Alternative; Instrument: Instrumental;
   c) Genre: Rock; Gender: Male; Mood: Sad;
   d) Genre: Rap; Type: Aggressive; Gender: Male; Mood: Happy

2) MTFL data set:
   a) Head Pose: Right Profile;
   b) Head Pose: Left Profile; Gender: Male;
   c) Head Pose: Left; Gender: Female; Glasses: Not Wearing;
   d) Head Pose: Frontal; Gender: Male; Smile: Smiling; Glasses: Not Wearing.

TABLE II
R-PRECISION VALUES FOR QUERIES WITH RADVIZ AND CONCENTRIC RADVIZ.

	RadViz	Concentric RadViz		
		$\hat{\sigma}(x, 0, 1)$ 	$\hat{\sigma}(x, 10, -0.8)$ 	$\hat{\sigma}(x, 20, -1)$ 
1-a)	0.1000	0.1000	0.9250	0.9650
1-b)	0.2307	0.4146	0.8780	0.9024
1-c)	0.4062	0.5833	0.6875	0.8541
1-d)	0.6388	0.6111	0.7777	0.9166
2-a)	0.8260	0.8260	1.0000	1.0000
2-b)	0.3750	0.5416	0.83333	0.8333
2-c)	0.6173	0.7198	0.90660	0.9339
2-d)	0.5894	0.5621	0.77052	0.8063

Similarly to the previous experiment, as the sigmoid is translated to the right, the points are positioned according to the highest probability class, which results in higher R-Precision values. This shows that Concentric RadViz also outperforms the original RadViz in the execution of ranked queries.

V. APPLICATIONS

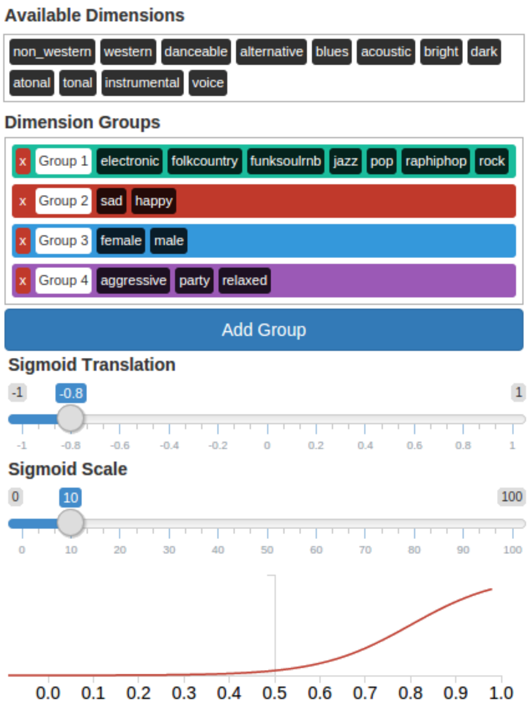
Our prototype has been implemented in JavaScript with Data Driven Documents (D3) library [26]. The Traveling Salesman Problem (DAs placement) is solved by Concorde library [27]. The visualization parameters are controlled by a graphical user interface illustrated in Fig. 4. The menu displays available dimensions, enables the creation of dimension groups (drag and drop interface) and the interactive control of sigmoid translation and scale.

In this section, we employ Concentric RadViz in two visual analytics applications involving music and film data sets. The effectiveness of Concentric RadViz is clearly evidenced in both applications, enabling users easily retrieve information of interest.

A. Exploring songs from multi-task classification

A typical manner of organizing music collections is to identify groups of songs that share similar features. In the context of music information retrieval, songs can be characterized by timbre, melody, rhythm, harmony, instrumentation, among other structures. A large amount of techniques that rely on those features to perform music classification and visualization have been proposed in the literature [28].

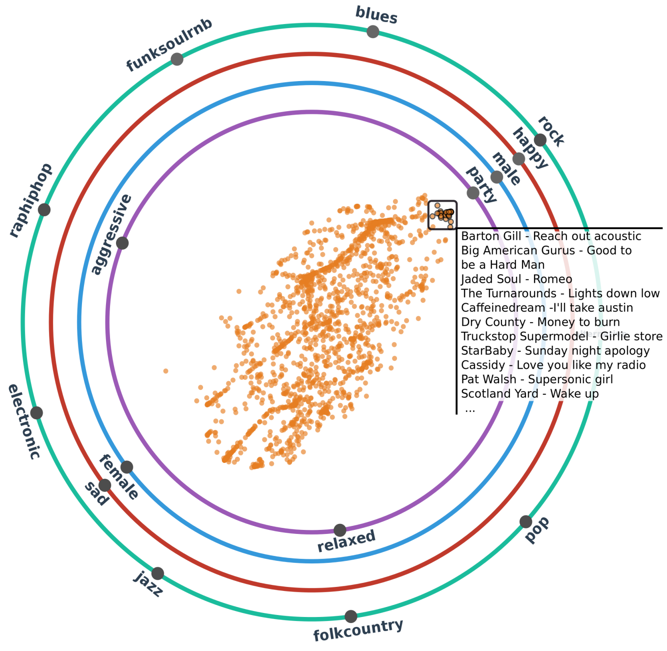
In the present work, music features are used to perform classification tasks, allowing the use of Concentric RadViz to visually identify songs with similar patterns. We make use of high-level descriptors implemented in Essentia 2.0, a recent



Sigmoid Translation -0.8

Sigmoid Scale 10

Fig. 4. Application menu.



Barton Gill - Reach out acoustic
Big American Gurus - Good to be a Hard Man
Jaded Soul - Romeo
The Turnarounds - Lights down low
Caffeinedream -I'll take austin
Dry County - Money to burn
Truckstop Supermodel - Girlie store
StarBaby - Sunday night apology
Cassidy - Love you like my radio
Pat Walsh - Supersonic girl
Scotland Yard - Wake up
...

Fig. 5. Concentric RadViz with the Dortmund music dataset.

open-source C++ library for audio analysis and retrieval of music content [21]. Each high-level descriptor is a classification task obtained from pre-trained SVM classifiers. We employ LIBSVM [29] which implements multi-class SVMs with an one-vs-one strategy and class probability estimation. We refer readers to [22], [23] for a more detailed discussion about the adopted features and classification strategies. The following classification tasks were used:

- **Genre**: Alternative, Blues, Electronic, Folk-Country, Funk-Soul-R&B, Jazz, Pop, Rap-Hiphop, Rock;
- **Mood**: Happy, non-happy;
- **Mood**: Sad, non-sad;
- **Type**: Party, non-party;
- **Type**: Relaxed, non-relaxed;
- **Type**: Aggressive, non-aggressive;
- **Gender**: Male, Female;
- **Instrument**: Instrumental, Voice.

We employ Concentric RadViz to explore the Dortmund audio benchmark data set [19], which was described in section IV. Fig. 5 shows the projection resulting from Concentric RadViz when applied to the Dortmund data set. The DGs refer to the semantic descriptors of music as provided by the Essentia library. Although all dimensions are available, the user can select the ones of most relevance. For instance, the user might be interested in exploring songs of a specific set of genres that bear a specific mood.

The Concentric RadViz layout depicted in Fig. 5 is filtering out low probabilities using the sigmoidal weighting discussed in section III-B with parameters scale $s = 10$ and translation $t = -0.8$.

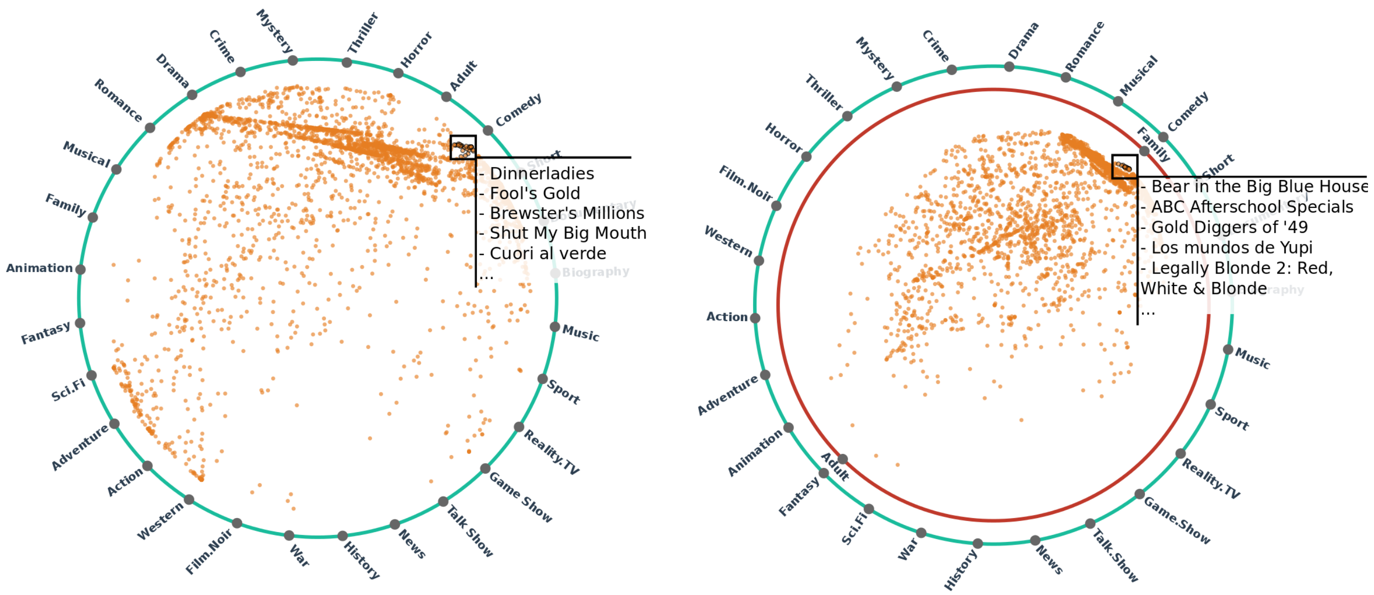
Notice from Fig. 5 that four dimensions of interest are aligned: rock, happy, male and party. Concentric RadViz are thus grouping songs that, according to the output of the classifier, belong to genre rock, are performed by a male, express the emotion happy and have a context of party, closer to the corresponding aligned DAs. By selecting and inspecting a small subset of the songs that are closer to the aligned DAs, one can clearly see a set of rock songs that matches the patterns male, happy and party. This interaction can be used to create a playlist by selecting regions of interest in the visualization.

The original classes indicate if the song is “happy” or “non-happy”, as well as “sad” or “non-sad”. However, the user can place just the “happy” and “sad” labels in one DG to simplify the visualization. Due to the normalization schemes, projected instances will lay closer to the class with higher probability, even though “happy” and “sad” do not belong the same classification task.

B. Exploring films from multi-task classification

In order to perform the classification required by Concentric RadViz, we pre-process the text meta-data associated to the films towards extracting features used by the classifiers. More specifically, word count, term extraction, and automatic natural language summarizations are employed to generate bag-of-words used by the classifiers [30].

The IMDB-F multi-label dataset [31] is used in our application. This data set contains 120,919 titles comprising 28 classes such as “Romance”, “Comedy”, “Adventure”, “Drama”, and “Fantasy”. The bag-of-words is obtained from the film synopsis and input into a multinomial naive-Bayes (MultinomialNB) algorithm [32], largely used for text classification [33]. We adopted the *Scikit-learn* Machine Learning



(a) RadViz projection considering one dimension group: film genre.

(b) Inclusion of an additional dimension group allowing the user to set a second restriction according to his interest.

Fig. 6. Concentric RadViz with the IMDB-F film dataset.

Toolbox, which implements the referred classifier (default parameters were used).

Fig. 6 illustrates the Concentric RadViz projection for the IMDB-F film data set with sigmoid parameters $s = 30$ and $t = -0.6$. Since IMDB-F is not multi-task, only dimensions related to film genre are available. In Fig. 6a, we select a region in which the grouped films belong to the “comedy” category and list some of their names.

Users can further explore the data by imposing specific interests. Suppose a user wishes to identify “commedy” films which are not “adult”. An alternative to perform such task is to create an additional DG with dimensions “family” and “adult”, then align “comedy” with “family”, as shown in Fig. 6b. Projected points close to the aligned DAs should match users expectations. Therefore, users can create new possibilities of data mining and discovery by easily playing with Concentric RadViz, even if a multi-task classification is not available.

VI. Discussion and Limitations

In this paper, we introduced Concentric RadViz, a general purpose visualization scheme that enables users to explore data sets in terms of similarity relations among semantic descriptors (output of classifiers previously trained on the data). The main differences to the original RadViz [5] algorithm are the possibility of using concentric circles to represent groups of dimensions under analysis and a filtering scheme based on sigmoidal weighting to reduce cluttering and ambiguity in the visualization. This kind of analysis allows users to identify regions of interest in the data set according to descriptors related to semantic properties as perceived by humans.

Our technique was validated with two information retrieval metrics, which indicate it outperforms the original RadViz in terms of neighborhood preservation (evaluated with MAP) and ranked queries (evaluated with R-Precision). As future work, we intend to investigate the quality of the visualization from the user perspective, as well as analyze its correlation with other evaluation measures.

Due to limited display space, we constraint the analysis to six simultaneous dimension groups. Another future work is the study of other visual metaphors that can be combined with Concentric RadViz in order to make possible the use of a larger number of concentric circles.

VII. Conclusions

We presented a novel projection-based approach for the visualization of multi-task/multi-label classification data and showed its usefulness in the context of two applications: the exploration of song and film data sets. We conducted the validation of our method with traditional information retrieval metrics, and demonstrated that Concentric RadViz can provide a suitable projection of data. Future works include a user-based case study for the evaluation of the technique and the development of visual metaphors to summarize dimension groups.

Acknowledgment

This work is supported by FAPESP (#2011/22749-8, #2012/17961-0, #2012/24801-0, #2014/09546-9, #2014/18665-1), and CNPq (#132239/2013-2, #305796/2013-5, #302643/2013-3), Brazil. We would like to thank the anonymous reviewers for their valuable comments and suggestions.

References

[1] B. Kovalerchuk and V. Grishin, "Collaborative Lossless Visualization of n-D Data by Collocated Paired Coordinates," in *Cooperative Design, Visualization, and Engineering*, ser. Lecture Notes in Computer Science, Y. Luo, Ed. Springer International Publishing, 2014, no. 8683, pp. 19–26.

[2] J. F. Mccarthy, K. A. Marx, P. E. Hoffman, A. G. Gee, P. O'neil, M. L. Ujwal, and J. Hotchkiss, "Applications of Machine Learning and High-Dimensional Visualization in Cancer Detection, Diagnosis, and Management," *Annals of the New York Academy of Sciences*, vol. 1020, no. 1, pp. 239–262, 2004.

[3] S. T. Teoh and K.-L. Ma, "PaintingClass: Interactive Construction, Visualization and Exploration of Decision Trees," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '03. New York, NY, USA: ACM, 2003, pp. 667–672.

[4] X. Geng, D.-C. Zhan, and Z.-H. Zhou, "Supervised nonlinear dimensionality reduction for visualization and classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, no. 6, pp. 1098–1107, 2005.

[5] C. Seifert and M. Granitzer, "User-Based Active Learning," in *2010 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2010, pp. 418–425.

[6] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley, "DNA visual and analytic data mining," in *Visualization '97., Proceedings*, 1997, pp. 437–441.

[7] F. Heimerl, S. Koch, H. Bosch, and T. Ertl, "Visual Classifier Training for Text Document Retrieval," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2839–2848, 2012.

[8] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011.

[9] R. E. Schapire and Y. Singer, "BoosTexter: A Boosting-based System for Text Categorization," *Machine Learning*, vol. 39, no. 2-3, pp. 135–168, 2000.

[10] T. Evgeniou and M. Pontil, "Regularized Multitask Learning," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 109–117.

[11] Bogdanov, "Semantic audio content-based music recommendation and visualization based on user preference examples," *Information Processing &amp; Management*, vol. 49, no. 1, pp. 13–33, 2013.

[12] K. Daniels, G. Grinstein, A. Russell, and M. Glidden, "Properties of normalized radial visualizations," *Information Visualization*, p. 1473871612439357, 2012.

[13] J. Sharko, G. Grinstein, and K. A. Marx, "Vectorized radviz and its application to multiple cluster datasets," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, no. 6, pp. 1444–1427, 2008.

[14] M. Ankerst, S. Berchtold, and D. Keim, "Similarity clustering of dimensions for an enhanced visualization of multidimensional data," in *IEEE Symposium on Information Visualization, 1998. Proceedings*, 1998, pp. 52–60, 153.

[15] L. D. Caro, V. Frias-Martinez, and E. Frias-Martinez, "Analyzing the Role of Dimension Arrangement for Data Visualization in Radviz," in *Advances in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, M. J. Zaki, J. X. Yu, B. Ravindran, and V. Pudi, Eds. Springer Berlin Heidelberg, 2010, no. 6119, pp. 125–132.

[16] L. Novakova and O. Stepankova, "RadViz and Identification of Clusters in Multidimensional Data," in *Information Visualisation, 2009 13th International Conference*, 2009, pp. 104–109.

[17] M. Doulis, M. Soldati, and A. Csillaghy, "SphereViz - Data Exploration in a Virtual Reality Environment," in *Information Visualization, 2007. IV '07. 11th International Conference*, 2007, pp. 680–683.

[18] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1.

[19] H. Homburg, I. Mierswa, B. Möller, K. Morik, and M. Wurst, "A benchmark dataset for audio classification and clustering." in *ISMIR*, vol. 2005, 2005, pp. 528–31.

[20] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 94–108.

[21] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "Essentia: an audio analysis library for music information retrieval," in *International Society for Music Information Retrieval Conference (ISMIR'13)*, Curitiba, Brazil, 2013, pp. 493–498.

[22] D. Bogdanov, M. Haro, F. Fuhrmann, A. Xambó, E. Gómez, and P. Herrera, "Semantic audio content-based music recommendation and visualization based on user preference examples," *Information Processing & Management*, vol. 49, pp. 13–33, 2013.

[23] D. Bogdanov, "From music similarity to music recommendation: Computational approaches based on audio features and metadata," Ph.D. dissertation, Universitat Pompeu Fabra, Barcelona, Spain, 2013.

[24] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[25] 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*, vol. 12, pp. 2825–2830, 2011.

[26] M. Bostock, V. Ogievetsky, and J. Heer, "D3 Data-Driven Documents," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2301–2309, 2011.

[27] D. Applegate, R. Bixby, W. Cook, and V. Chvtal, *On the solution of traveling salesman problems*. Rheinische Friedrich-Wilhelms-Universitt Bonn, 1998, vol. Extra Volume.

[28] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation." *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.

[29] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[30] U. M. Fayyad, A. Wierse, and G. G. Grinstein, *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann, 2002.

[31] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, vol. 85, no. 3, pp. 333–359, 2011.

[32] J. D. Rennie, L. Shih, J. Teevan, D. R. Karger *et al.*, "Tackling the poor assumptions of naive bayes text classifiers," in *ICML*, vol. 3. Washington DC), 2003, pp. 616–623.

[33] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine learning*, vol. 39, no. 2-3, pp. 103–134, 2000.