MUSIC CLASSIFICATION BY TRANSDUCTIVE LEARNING USING BIPARTITE HETEROGENEOUS NETWORKS

Diego F. Silva, Rafael G. Rossi, Solange O. Rezende, Gustavo E. A. P. A. Batista
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
{diegofsilva,ragero,solange,gbatista}@icmc.usp.br

ABSTRACT

The popularization of music distribution in electronic format has increased the amount of music with incomplete metadata. The incompleteness of data can hamper some important tasks, such as music and artist recommendation. In this scenario, transductive classification can be used to classify the whole dataset considering just few labeled instances. Usually transductive classification is performed through label propagation, in which data are represented as networks and the examples propagate their labels through their connections. Similarity-based networks are usually applied to model data as network. However, this kind of representation requires the definition of parameters, which significantly affect the classification accuracy, and presents a high cost due to the computation of similarities among all dataset instances. In contrast, bipartite heterogeneous networks have appeared as an alternative to similarity-based networks in text mining applications. In these networks, the words are connected to the documents which they occur. Thus, there is no parameter or additional costs to generate such networks. In this paper, we propose the use of the bipartite network representation to perform transductive classification of music, using a bag-of-frames approach to describe music signals. We demonstrate that the proposed approach outperforms other music classification approaches when few labeled instances are available.

1. INTRODUCTION

The popularity of online music services has dramatically increased in the last decade. The revenue of online services, such as music streaming, has more than triplicate in the last three years. Online services already account for a significant 40% of the overall industry trade revenues [1]. However, the popularization of music and video clips distribution in electronic format has increased the amount of music with incomplete metadata. The incompleteness of the data can hamper some important tasks such as indexing, retrieval and recommendation.

For instance, users of music services commonly define their preferences based on genre information. A recommendation system can make use of such information to suggest other music conditional to the expressed preferences. The lack of genre information on music imposes limits to the capability of the recommendation systems to correctly identify consume patterns as well as to recommend a diverse set of music. Similar statements can be made for music indexing and retrieval.

Due to the academic and commercial importance of digital music, we have witnessed in the last decade a tremendous increase of interest for Music Information Retrieval (MIR) tasks. Most of the proposed MIR methods are based on supervised learning techniques. Supervised learning usually requires a substantial amount of correctly labeled data in order to induce accurate classifiers. Although labeled data can be obtained with human supervision, such process is usually expensive and time consuming. A more practical approach is to employ methods that can avail of both a small number of labeled instances and a large amount of unlabeled data.

Transductive learning directly estimates the labels of unlabeled instances without creating a classification model. A common approach to perform transductive classification is label propagation, in which the dataset is represented as a network and the labels of labeled instances are propagated to the unlabeled instances through the network connections. Similarity-based networks are usually applied to represent data as networks for label propagation [19]. However, they present a high cost due to the computation of the similarities among all dataset instances, and require the definition of several graph construction parameters that can significantly affect the classification accuracy [11].

Bipartite heterogeneous networks have appeared as an alternative to similarity-based networks in sparse domains, such as text mining [9, 10]. In these networks, words are connected to documents in which they occur. Thus, there are no parameters or additional costs to generate such networks. In a similar way, we can represent music collections as a bipartite network though the use of a bag-of-frames (BoF) representation. The BoF is a variation of bag-of-words (BoW) representation used in text analysis and has been applied in studies of genre recognition, music similarity, and others [12].

In this paper, we propose the use of the bipartite network representation to perform transductive classification of music, using a BoF approach to describe music signals. We demonstrate that the proposed approach outperforms other music classification approaches when few labeled instances are available.

2. BACKGROUND & RELATED WORK

Transductive classification is a useful way to classify all dataset instances when just few labeled instances are available [19]. Perhaps the most common and intuitive way to perform transductive classification is through label propagation, which is commonly made by using similarity-based networks to represent the data. Usual ways to generate similarity-based networks are [17–19]: (*i*) fully connected-network, in which every pair of instances are connected; (*ii*) k nearest neighbor, in which an instance is connected with its k most similar instances; and (*iii*) ϵ network, in which two instances are connected if their distance is above a threshold.

Bipartite networks have appeared as an alternative to similarity-based networks in sparse domains such as texts [9, 10]. The use of bipartite networks to represent text collections and the use of algorithms which perform label propagation in bipartite networks obtained results as good as the obtained by similarity-based networks [10]. However, the computation cost to generate similarity-based networks is $O(|I|^2 \times |A|)$, in which $|I|$ is the number of intances and $|A|$ is the number of attributes of a dataset, while the computational cost to generate bipartite networks is $O(|I| \times |A|)$. Morevorer, the generation of bipartite networks is parameter-free.

We can represent music collections as a bipartite network though the generation of a bag-of-frames (BoF). Methods using BoF has became common in different MIR tasks, including similarity, genre, emotion and cover song recognition [12]. Such strategies basically consist of three main steps: feature extraction, codebook generation and learning/classification.

Probably, the most simple and commonly used strategy for the codebook generation is the Vector Quantization (VQ). Basically, the VQ uses clustering algorithms on the frame-level features and consider the center of clusters as the words of a dictionary. The simple k-means is, probably, the most used algorithm in this step and showed to achieve similar results to other methods [8].

Recently, new tools have emerged for creating codebooks. Specifically, strategies based on Sparse Coding [5, 15] and Deep Belief Networks [3, 7] have been widely used. However, even though these strategies often improve the results in different domains, they can present similar performance to simple strategies such as VQ in certain tasks [7].

3. PROPOSED FRAMEWORK: MC-LPBN

In this paper we propose a framework called MC-LPBN (Music Classification through Label Propagation in Bipartite Networks) to perform transductive classification of musics. The proposed framework has three main steps: (*i*) codebook generation, (*ii*) network generation for transductive classification, and (*iii*) transductive classification using bipartite heterogeneous networks. In the next subsections we present the details of each step.

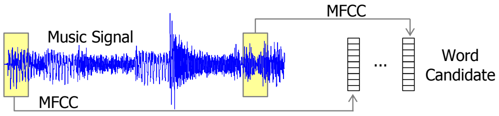


Figure 1. Word candidates generation process

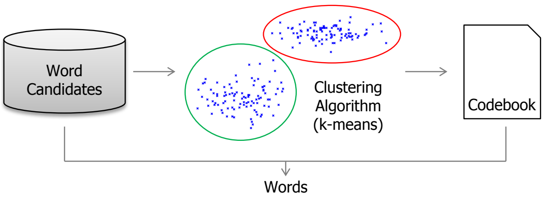


Figure 2. The word candidates are clustered and each centroid is elected as a codeword. The word frequency is directly related to the candidates count in each cluster

3.1 Codebook Generation and Bag-of-Frames

In order to represent a music collection as a BoF it is necessary to extract a set of words. Such procedure starts with the extraction of word candidates. A word candidate is a set of features extracted from a single window. As a sliding window swipes across the entire music signal, each music gives origin to a set of word candidates. In this work we use MFCC as feature extraction procedure. Figure 1 illustrates the word candidates generation process.

The next step is the creation of a codebook. A codebook is a set of codewords used to associate the word candidates to a finite set of words. The idea is to select the most representative codeword for each word candidate. To do this, we use a clustering algorithm with the word candidates and consider the center of each cluster as a codeword. So, each candidate is associated to the codeword that represents the cluster it belongs. In this step, we used the ubiquitous k-means algorithm, due to its simplicity and efficiency. Figure 2 illustrates this procedure.

Finally, there is a step to the generation of a BoF matrix. In such a matrix, each line corresponds to a music recording, each column corresponds to a word and the cells correspond to the frequency of occurrence of the word in the music. The BoF allows the generation of bipartite networks for transductive classification, as we discuss in the next subsection.

3.2 Network Generation for Transductive Classification

Formally, a network is defined by $N = \langle \mathcal{O}, \mathcal{E}, \mathcal{W} \rangle$, in which $\mathcal{O}$ represents the set of objects (entities) of a problem, $\mathcal{E}$ represents the set of connections among the objects and $\mathcal{W}$ represents the weights of the connections. When

$\mathcal{O}$ is composed by a single type of object, the network is called homogeneous network. When $\mathcal{O}$ is composed by h different types of objects, i.e., $\mathcal{O} = \mathcal{O}_1 \cup \cdots \cup \mathcal{O}_h$, the networks is called heterogeneous network [6].

The music collection can be represented by a bipartite heterogeneous network with $\mathcal{O} = \mathcal{M} \cup \mathcal{T}$, in which $\mathcal{M} = \{m_1, m_2, \ldots, m_n\}$ represents the set of music and $\mathcal{T} = \{t_1, t_2, \ldots, t_l\}$ represents the set of words. $\mathcal{M}$ is composed by labeled ($\mathcal{M}^L$) and unlabeled ($\mathcal{M}^U$) music, i.e., $\mathcal{M} = \mathcal{M}^L \cup \mathcal{M}^U$. A music $m_i \in \mathcal{M}$ and a word $t_j \in \mathcal{T}$ are connected if t_j occurs in m_i. The weight of the relation between m_i and t_j (w_{m_i,t_j}) is the frequency of t_j in m_i. Thus, only the words and their frequencies in the music are needed to generate the bipartite network.

For transductive classification based on networks, let $\mathcal{C} = \{c_1, c_2, \ldots, c_l\}$ be the set of possible labels, and let $\mathbf{f}_{o_i} = \{f_1, f_2, \ldots, f_{|\mathcal{C}|}\}^T$ be the weight vector of an object o_i, which determines its weight or relevance for each class. Hence, it is also referred as class information vector. The class information of an object $m_i \in \mathcal{M}$ or an object $t_j \in \mathcal{T}$ is denoted respectively by $\mathbf{f}_{m_i}$ and $\mathbf{f}_{t_j}$. All the class information of objects in $\mathcal{M}$ or $\mathcal{T}$ is denoted by the matrices $\mathbf{F}(\mathcal{M}) = \{\mathbf{f}_{m_1}, \mathbf{f}_{m_2}, \ldots, \mathbf{f}_{m_{|\mathcal{M}|}}\}^T$ and $\mathbf{F}(\mathcal{T}) = \{\mathbf{f}_{t_1}, \mathbf{f}_{t_2}, \ldots, \mathbf{f}_{t_{|\mathcal{T}|}}\}^T$. The class information of a labeled music m_i is stored in a vector $\mathbf{y}_{m_i} = \{y_1, y_2, \ldots, y_{|\mathcal{C}|}\}^T$, which has the value 1 to the corresponding class position and 0 to the others. The weights of connections among objects are stored in a matrix $\mathbf{W}$. A diagonal matrix $\mathbf{D}$ is used to store the degree of the objects, i.e., the sum of the connection weights of the objects. Thus the degree of a music m_i in a bipartite network is ($d_{m_i} = d_{i,i} = \sum_{t_j \in \mathcal{T}} w_{m_i,t_j}$).

3.3 Transductive Classification Using Bipartite Heterogeneous Networks

The main algorithms for transductive classification in data represented as networks are based on regularization [19], which have to satisfy two assumptions: (*i*) the class information of neighbors must be close; and (*ii*) the class information assigned during the classification process must be close to the real class information. In this paper we used three regularization-based algorithms: (*i*) Tag-based classification Model (TM) [16], (*ii*) Label Propagation based on Bipartite Heterogeneous Networks (LPBHN) [10], and (*iii*) GNetMine (GM) [6].

TM algorithm minimizes the differences among (*i*) the real class information and the class information assigned to music ($\mathcal{M}$), (*ii*) the real class information and the class information assigned to and objects from other domains that aid the classification ($\mathcal{A}$), and (*iii*) the class information among words ($\mathcal{T}$) and objects in ($\mathcal{M}$) or ($\mathcal{A}$), as presented in Equation 1.

$$Q(\mathbf{F}) = \alpha \sum_{a_i \in \mathcal{A}} ||\mathbf{f}_{a_i} - \mathbf{y}_{a_i}||^2 + \beta \sum_{m_i \in \mathcal{M}^L} ||\mathbf{f}_{m_i} - \mathbf{y}_{m_i}||^2 \quad (1)$$
$$+ \gamma \sum_{m_i \in \mathcal{M}^U} ||\mathbf{f}_{m_i} - \mathbf{y}_{m_i}||^2 + \sum_{o_i \in \mathcal{M} \cup \mathcal{A}} \sum_{t_j \in \mathcal{T}} w_{o_i,t_j} ||\mathbf{f}_{o_i} - \mathbf{f}_{t_j}||^2$$

The parameters α, β and γ control the importance given for each assumption of TM. Objects are classified using class mass normalization [18].

LPBHN is a parameter-free algorithm based on the Gaussian Fields and Harmonic Functions (GFHF) algorithm [18], which performs label propagation in homogeneous networks. The difference is that LPBHN considers the relations among different types of objects. The function to be minimized by LPBHN is:

$$Q(\mathbf{F}) = \frac{1}{2} \sum_{m_i \in \mathcal{M}} \sum_{t_j \in \mathcal{T}} w_{m_i,t_j} ||\mathbf{f}_{m_i} - \mathbf{f}_{t_j}||^2 \quad (2)$$
$$+ \lim_{\mu \to \infty} \mu \sum_{m_i \in \mathcal{M}^L} ||\mathbf{f}_{m_i} - \mathbf{y}_{m_i}||^2,$$

in which μ tending to infinity means that $\mathbf{f}_{m_i} \equiv \mathbf{y}_{m_i}$, i.e., the information class of labeled musics do not change.

The GM framework is based on the Learning with Local and Global Consistency (LLGC) algorithm [17], which performs label propagation in homogeneous networks. The difference between the algorithms is that GM considers the different types of relations among the different types of objects. For the problem of music classification using bipartite networks, GM minimizes the differences of (*i*) the class information among music and words and (*ii*) the class information assigned to labeled music during the classification and their real class information. The function to be minimized by GM is:

$$Q(\mathbf{F}) = \sum_{m_i \in \mathcal{M}} \sum_{t_j \in \mathcal{T}} w_{m_i,t_j} \left|\left| \frac{\mathbf{f}_{m_i}}{\sqrt{d_{m_i}}} - \frac{\mathbf{f}_{t_j}}{\sqrt{d_{t_j}}} \right|\right|^2 \quad (3)$$
$$+ \sum_{m_i \in \mathcal{M}} \mu ||\mathbf{f}_{m_i} - \mathbf{y}_{m_i}||^2,$$

in which $0 < \mu < 1$.

We highlight that all the algorithms presented above have iterative solutions to minimize the respective equations. This allows to obtain similar results to the closed solutions with a lower computational time.

4. EXPERIMENTAL EVALUATION

To illustrate the generality of our approach, we evaluate our framework in two different scenarios. In this section, we describe the tasks we considered, the experimental setup used in our experiments, as well as the results obtained and a short discussion about them.

4.1 Tasks Description

We evaluate our framework in genre recognition and cover song recognition scenarios. The remaining of this section contains a brief description of each task and the datasets used to each end.

4.1.1 Genre Recognition

Genre recognition is an important task in several applications. Genre is a quality created by the human beings to intuitively characterize music [14]. For humans, the classification of music by genre is relatively simple task, and can be done by listening to a short excerpt of a music.

Therefore, most of the existing data for this task considers a short duration excerpt for each recording. In this work, we use the GTZAN[1] and Homburg[2] datasets. The first has $1,000$ snippets of 30 seconds of ten different genres. The number of instances of each class is equally distributed. The Homburg dataset, in turn, has ten seconds sections of $1,886$ recordings, belonging to nine genres. In this case, the genre with fewer instances has only 47 examples, while the largest has 504.

4.1.2 *Cover Song Recognition*

Cover song may be defined as distinct performances of the same music with differences in tempo, instrumentation, style or other characteristics [4]. Finding reinterpreted music is an important task mainly to commercial ends. For example, it can be used to ensure copyright in websites which allow users to create content.

In this paper, we evaluate our framework in a task similar to the cover song recognition. But, instead find the original recording of a query music, we consider all different interpretations of the same music as the same class.

To evaluate our proposal we used the Mazurkas Project data[3], in which each music has several versions. This dataset contains 2914 recordings of 49 Chopin's mazurkas for piano (from 43 to 97 versions per class).

4.2 Experimental Setup

We evaluated our framework considering different configurations for the 1^{st} and 3^{rd} steps. For the 1^{st} step, we consider variations of parameters of the feature extraction and codebook generation phases. In this work, we use 20 MFCC as frame-level features. This number is a common choice in MIR papers [2]. We use 5 different window sizes, with an overlap of 50% between them: 0.0625, 0.125, 0.25, 0.5 and 0.75 seconds. Finally, we applied the k-means using $k \in \{100, 200, 400, 800, 1600, 3200\}$.

For the 3^{rd}, we consider the algorithms presented in Section 3.3: Label Propagation using Bipartite Heterogeneous Networks (LPBHN), Tag-based Model (TM), and GNetMine (GM). For GM we use the parameter $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. For TM we use $\alpha = 0$, since there are no objects from different domains, $\beta \in \{0.1, 1.0, 10, 100, 1000\}$, and $\gamma \in \{0.1, 1.0, 10.0, 100.0, 1000.0\}$. The iterative solutions proposed by the respective authors were used for all the algorithms. The maximum number of iterations is set to 1000 since this is a common limit value for iterative solutions.

We also carried out experiments using two algorithms for label propagation in similarity-based networks, Learning with Local and Global Consistency (LLGC) [17] and Gaussian Fields and Harmonic Functions (GFHF) [18], and two classical supervised learning algorithms for comparison with the proposed approach, k Nearest Neighbors (kNN) and Support Vector Machines (SVM) [13].

We build similarity-based networks using the fully connected approach with $\sigma = 0.5$ [19] and we set $\alpha = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ for LLGC algorithm. For kNN we set $k = 7$ and weighted vote, and for SVM we used linear kernel and $C = 1$ [9].

The metric used for comparison was the classification accuracy, i.e., the percentage of correctly classified music recordings. The accuracies are obtained considering the average accuracies of 10 runs. In each run we randomize the dataset and select x examples as labeled examples. The remaining $|\mathcal{M}| - x$ examples are used for measuring the accuracy. We carried out experiments using $x = \{1, 10, 20, 30, 40, 50\}$ to analyze the trade-off between the number of labeled documents and classification accuracy. The best accuracies obtained by some set of parameters of the algorithms are used for comparison.

4.3 Results and Discussion

Given the large amount of results obtained in this work, their complete presentation becomes impossible due to space constraints. Thus, we developed a website for this work, where detailed results can be found[4]. In this section, we summarize the results from different points of view.

4.3.1 *Influence of Parameters Variation*

Our first analysis consists in evaluating the influence of the variation of the codebook generation step parameters. Figure 3 presents the variation of accuracy for both genre recognition dataset and each window size according to a different number of words in the dictionary. To do this, we fixed the number of labeled examples in 10. This number represents a good trade-off between the classification accuracy of the algorithms and the human effort to label music. But, we note that the behaviors are similar to other numbers of properly labeled examples.

The results show that the transductive learning methods can achieve similar or even superior results than the obtained by using inductive models. In the case of GTZAN data, there is a clear increasing pattern when the number of words varies. Using higher values to it, both strategies perform well, but the transductive learning obtained the higher accuracies. The results obtained by similarity-based networks were slightly better in most of configurations. But, as mentioned before, similarity-based networks has a high cost to calculate the similarities between all the examples and require the setting of several parameters to construct the network. In the Homburg dataset, transductive learning is better independently of the parameter configuration. In this case, there are no significant differences between bipartite network approaches, but they performed better than the similarity-based networks in most of cases.

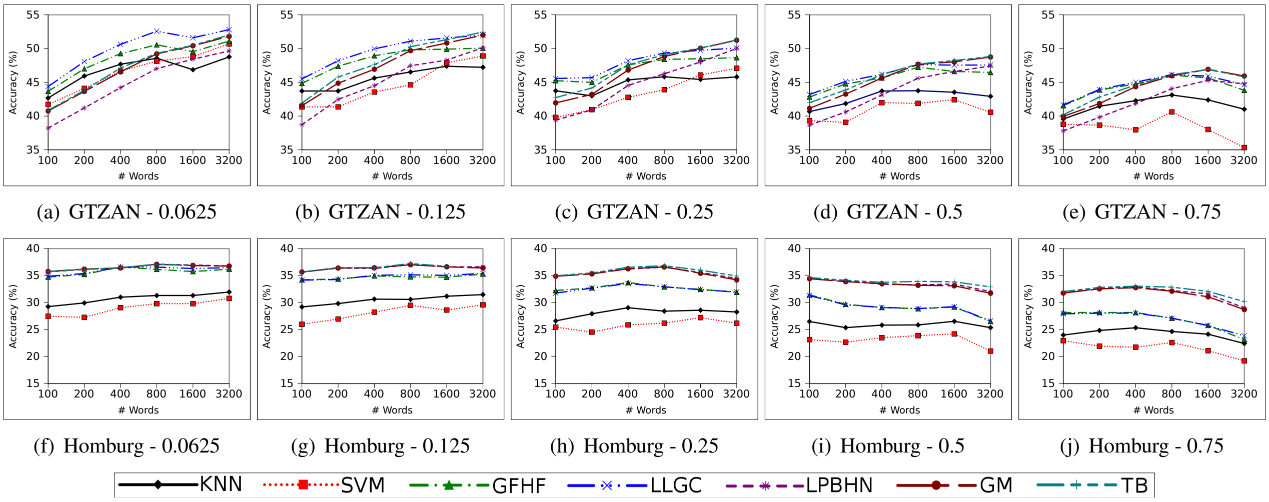
In order to evaluate our framework in the cover song recognition, we used the LPBHN algorithm, that achieved

[1] http://marsyas.info/download/data_sets/

[2] http://www-ai.cs.uni-dortmund.de/audio.html

[3] http://www.mazurka.org.uk/

[4] http://sites.labic.icmc.usp.br/dfs/ismir2014



(a) GTZAN - 0.0625 (b) GTZAN - 0.125 (c) GTZAN - 0.25 (d) GTZAN - 0.5 (e) GTZAN - 0.75

(f) Homburg - 0.0625 (g) Homburg - 0.125 (h) Homburg - 0.25 (i) Homburg - 0.5 (j) Homburg - 0.75

Figure 3. Accuracy for genre recognition by varying the number of words in the codebook. The number of labeled examples per class is fixed in 10.

similar result to other transductive methods and has the advantage of being parameter-free, and both, SVM and kNN, inductive approaches. The results show a high increasing pattern when transductive learning is used, and a more stable pattern to supervised methods. Figure 4 shows the accuracy achieved by fixing the number of labeled examples in 10. We fixed the window size to 0.75 seconds, since Mazurkas is the larger dataset used in this work and this is the fastest configuration to the feature extraction phase. We believe that the performance of transductive learning can overcome the SVM if we increase the number of words.

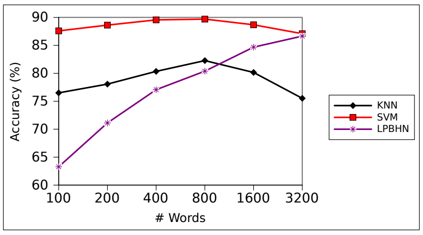


Figure 4. Accuracy for Mazurkas by varying the number of words in the codebook. The number of labeled examples is fixed in 10 and the window length in 0.75 seconds.

4.3.2 *Number of Labeled Examples Variation*

The evaluation of the performance variation according to the number of labeled examples is an important analysis in the context of transductive learning. Figure 5 shows the behavior of accuracy in genre recognition task when there is a variation in the number of labeled examples that belong to each class. The results were obtained by fixing the number of words in 3200, in which good results were achieved in several configurations, and a window size of the middle value of 0.25 seconds, since the results were similar than the obtained with other values to this parameter.

To analyze these graphs, it is interesting to know the proportion that the number of labeled examples represents in each dataset. For example, in the case of GTZAN set, 50 examples correspond to exactly 50% of the examples in each classes. In the case of Homburg, it represents 100% of the minority class, but less than 10% of the majority.

In both cases, the behavior of the accuracies was similar. As the number of labeled samples increases, the performance becomes better. The transductive learning methods performed better across the curve. As the proportion of the number of labeled instances increases, the tendency is that the performance of inductive algorithms approaches the performance of transductive algorithms.

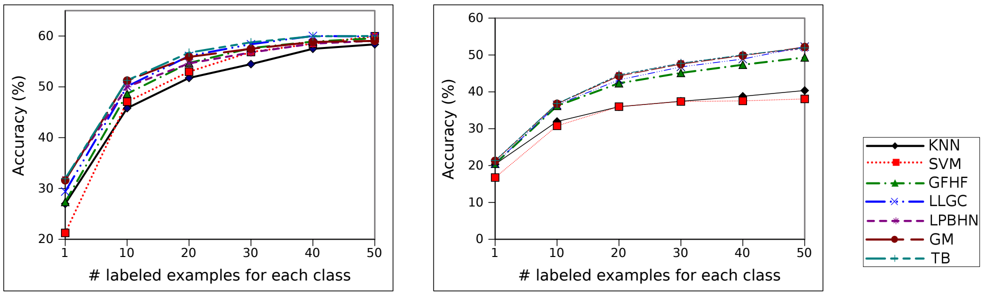
For sake of space limitations, we omitted the results for the cover song recognition task. However, we point out that the behavior of accuracy rates were very similar to obtained in the other task.

5. CONCLUSION

In this paper, we presented a framework for transductive classification of music using bipartite heteregeneous networks. We show that we can have a better performance by using this approach instead the traditional inductive learning. Our results were close or superior to the obtained by similarity-based networks. This kind of network, however, requires several parameters and has a high cost due to the calculation of the similarity between the instances.

We should note that the accuracy rates achieved in this paper are worse than some results presented in related works. For example, there are some papers that achieved accuracy higher than 80% to the GTZAN dataset using BoF approaches. However, these results were probably obtained due to the choice of specific features and parameters. Moreover, these papers used inductive learning approaches, with labels for the entire dataset. Nevertheless, we demonstrated that, for the same parameter set, the use of bipartite heterogeneous network achieved the best results.

As future work we will investigate a better feature configuration and different codebook generation strategies.



(a) GTZAN

(b) Homburg

Figure 5. Accuracy of genre recognition by varying the numbers of labeled examples for each class. The number of words and the windows length are fixed in 3200 and 0.25 seconds, respectively.

ACKNOWLEDGEMENTS: We would like to thank the financial supports by the grants 2011/12823-6, 2012/50714-7, 2013/26151-5, and 2014/08996-0, Sao Paulo Research Foundation (FAPESP).

6. REFERENCES

[1] Recording industry in numbers. Technical report, International Federation of the Phonographic Industry, 2014.

[2] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96(4):668–696, 2008.

[3] S. Dieleman, P. Brakel, and B. Schrauwen. Audio-based music classification with a pretrained convolutional network. In *ISMIR*, pages 669–674, 2011.

[4] D. P. W. Ellis and T. Bertin-Mahieux. Large-scale cover song recognition using the 2d fourier transform magnitude. In *ISMIR*, pages 241–246, 2012.

[5] M. Henaff, K. Jarrett, K. Kavukcuoglu, and Y. LeCun. Unsupervised learning of sparse features for scalable audio classification. In *ISMIR*, pages 681–686, 2011.

[6] M. Ji, Y. Sun, M. Danilevsky, J. Han, and J. Gao. Graph regularized transductive classification on heterogeneous information networks. In *Proc. Eur. Conf. on Machine Learning and Knowledge Discovery in Databases*, pages 570–586. Springer-Verlag, 2010.

[7] J. Nam, J. Herrera, M. Slaney, and J. O. Smith. Learning sparse feature representations for music annotation and retrieval. In *ISMIR*, pages 565–570, 2012.

[8] M. Riley, E. Heinen, and J. Ghosh. A text retrieval approach to content-based audio retrieval. In *ISMIR*, pages 295–300, 2008.

[9] R. G. Rossi, de T. P. Faleiros, de A. A. Lopes, and S. O. Rezende. Inductive model generation for text categorization using a bipartite heterogeneous network. In *Proc. Intl. Conf. on Data Mining*, pages 1086–1091. IEEE, 2012.

[10] R. G. Rossi, A. A. Lopes, and S. O. Rezende. A parameter-free label propagation algorithm using bipartite heterogeneous networks for text classification. In *Proc. Symp. on Applied Computing*, pages 79–84. ACM, 2014.

[11] C. A. R. Sousa, S. O. Rezende, and G. E. A. P. A. Batista. Influence of graph construction on semi-supervised learning. In *Proc. Eur. Conf. Machine Learning and Knowledge Discovery in Databases*, pages 160–175. Springer-Verlag, 2013.

[12] L. Su, C.-C.M. Yeh, J.-Y. Liu, J.-C. Wang, and Y.-H. Yang. A systematic evaluation of the bag-of-frames representation for music information retrieval. *IEEE Transactions on Multimedia*, 16(5):1188–1200, Aug 2014.

[13] P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[14] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[15] C. C. M. Yeh, L. Su, and Y. H. Yang. Dual-layer bag-of-frames model for music genre classification. In *Intl. Conf. on Acoustics, Speech and Signal Processing*, 2013.

[16] Z. Yin, R. Li, Q. Mei, and J. Han. Exploring social tagging graph for web object classification. In *Proc. Intl. Conf. on Knowledge Discovery and Data Mining*, pages 957–966, 2009.

[17] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, pages 321–328, 2004.

[18] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. Intl. Conf. on Machine Learning*, pages 912–919. AAAI Press, 2003.

[19] X. Zhu, A. B. Goldberg, R. Brachman, and T. Dietterich. *Introduction to Semi-Supervised Learning*. Morgan and Claypool Publishers, 2009.