Instance Selection for Music Genre Classification using Heterogeneous Networks

Angelo Cesar Mendes da Silva[1] **, Paulo Ricardo Viviurka do Carmo**[1] **,**
Ricardo Marcondes Marcacini[1]**, Diego Furtado Silva**[2]

[1]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo, Brazil
Av Trabalhador São-carlense, 400, Centro – 13566-590 São Carlos, SP

[2]Departamento de Computação – Universidade Federal de São Carlos, Brazil
Rod. Washington Luís, km 235, Jardim Guanabara – 13565-905 São Carlos, SP

{angelo.mendes, paulo.carmo, ricardo.marcacini}@usp.br, diegofs@ufscar.br

Abstract. *In scenarios involving musical data, there are usually high-dimensional data and different modalities, such as audio and text, that cost more in machine learning tasks. Instance selection is a promising approach as preprocessing step to reduce these challenges. With the intent to explore the multimodality in music information, we introduce musical data instance selection into heterogeneous network models. We propose and evaluate ten different heterogeneous networks to identify more representative relationships with various musical features related, including songs, artists, genres, and melspectrogram. The results obtained allow us to define which network structure is more appropriate considering the volume of available data and the type of information that the features have. Finally, we analyze the relevance of the musical features, and the relationship does not contribute for instance selection.*

1. Introduction

Data mining and machine learning methods for large volumes of data require expensive computational resources. Moreover, there are usually high-dimensional data and different modalities in scenarios involving musical data, such as audio and text. Instance selection is a promising approach to dealing with these challenges. Such methods aim to select a representative subset S from the complete dataset T, in which the performance function P of a machine learning method M is not significantly reduced, i.e., $P(M,S) \approx P(M,T)$. Although instance selection has been a research topic investigated for decades in the field of data mining and machine learning, most existing methods focus on unimodal scenarios, such as low-level characteristics of music audio signals.

Musical data can be represented by different data modalities such as lyrics, tags, and features extracted directly from audio content [1, 2]. Audio data is widely explored to extract features in different levels that represent the music in a vector space and provide input data for machine learning methods. For example, melspectrogram is used in genre classification [3, 4], and chromagram in the cover identification problem [5, 6]. More recently, deep learning methods have been used to learn representations of music from raw audio data, both for genre classification and instrument detection tasks [7, 8].

Multimodality in music information has been explored as complementary ways to build representations that have more discriminatory power [9, 10]. Exploring musical multimodality is a perceptive activity inherent to humans, so we can easily perceive relationships between different genres, similar artists, or songs that seem to belong to one artist. In general, these relationships are features that can be represented as metadata and are being highlighted as complementary musical information [11, 12].

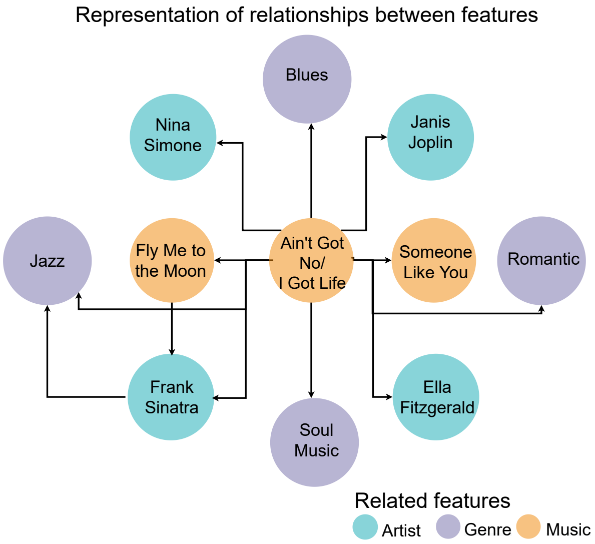

Figure 1: Possible relationships between artists, genres, and songs related to a song.

A promising strategy to abstract and map these relationships is to build a model based on a heterogeneous network architecture for different modalities. Heterogeneous networks can relate the modalities following a specific similarity criterion that allows us to make associations between songs through its metadata, such as related genres, songs, and artists. Following this idea, we hypothesize that if we can build a model on the relationships between musical features, as illustrated in Figure 1, we can also use them to identify the most representative songs and use them as input for machine learning tasks.

In this paper, we introduce musical data instance selection from heterogeneous network models. We present a study on ten different heterogeneous networks to identify more representative relationships from various modalities of the musical data. We evaluate the instance selection for musical genre classification, in which we aim to obtain a reduced subset to train a classification model. Our main contributions are threefold:

- We propose and evaluate heterogeneous networks with the most usual relationships in musical datasets, including songs, artists, genres, and mel-spectrogram features extracted from the audio signal. We discuss the relevance of the relationships between artists, songs, genres, and clusters to emphasize the features' particularities, and we show what kind of relationship does not contribute for instance selection.
- Due to the multimodality (audio and text) of the dataset, we investigate a regularization framework for heterogeneous networks. We explore regularization through label propagation, in which labeled vertices of the training set smoothly propagate their labels across the heterogeneous network. Thus, each vertex of the heterogeneous network has a membership vector with a label, which is then used to identify more representative instances.
- We present a score ranking function using the regularizer membership information obtained by the regularizer. Three different instance selection strategies based on ranking function were proposed and evaluated.

We carried out an experimental evaluation involving the 4MuLA (Multitask, Multimodal, and Multilingual Music Lyrics and Audio features) dataset [13]. The experimental results compare ten different heterogeneous networks and three instance selection strategies. In addition, two traditional methods of instance selection were used as reference models. The experimental results show evidence that heterogeneous networks are competitive for song instance selection in multimodal scenarios, especially with relationships involving cluster-based features from audio signals, as well as related genres and songs.

2. Related works

In this session, we explore existing approaches to the instance selection problem and when no have performance gains in problem-solving. We extend the discussion to discuss the characteristics of the methods and the specific needs for musical data problems. In addition, we discussed the elements of musical datasets and how using instance selection could mitigate their limitations.

In a classification problem, it is expected that data existent in the same class are contained in the same cluster and that there is a similarity between their features. However, there are scenarios where specific data have inconsistencies, and its structure does not match the other data in the group. To reach a good solution to a classification problem, strategies for selecting reduced subsets of instances and removing noisy instances have often been explored. Although classification models are expected to perform better with a higher quality training set, it is also an interesting result to use instance selection to reduce the training set and maintain the classifier performance.

Among the possibilities for applying the instance selection process is isolating a reduced instance subset with a high representation of the data of the class information [14]. Similarly, we can also think that the process aims to remove superfluous or harmful samples that can be used to represent the class. We can understand this representation by observing data that have standard structural features about other data in your cluster, and an association between it and its class can be done objectively. The better strategy for selecting the most representative instances depends on an analysis of this structure [15].

Thus, the reduced set can be used as a representative of the class in several tasks and is expected to maintain or improve performance compared to the original set [16]. Furthermore, there is a wide variety of applications that have the selection process as a preliminary step [17], as an example, we can highlight noise identification [18], undersampling in unbalanced classes [19], or missing data imputation [20].

Defining how to carry out the selection process is a task that requires computational cost evaluation, as there are scenarios where the selection cost is high and does not reflect a high-performance gain in the solution of a task [15, 17, 21]. Initial approaches proposed selection methods based on the analysis of closest instances [22, 23]. Even having a considerable cost, they are simple methods, and their concept is derived from other proposals that incorporate improvements [24, 25, 26]. Other works use models to classify instances to indicate the best candidates to represent the class, either through individual classifiers [27, 28] or through ensembles of models that have been shown to be efficient in identifying the particularities of the dataset [29, 30, 31].

In the context of musical data, problems with real-world scenarios naturally have unbalanced data regardless of the label type. For example, whether for commercial or cultural reasons, issues that have their data organized by genre, artists, or categorical tags, we are always seeing new music appear and be incorporated into categories without a pre-defined proportion [32]. In addition to these characteristics, music data is commonly represented by acoustic information and textual attributes such as lyrics and metadata that can express various information, for example, categories or relationships existing between a song and other songs, artists, or genres [33, 34]. Associating with human's musical perception, we can imagine that these features are complementary [35] and helpful to compose the most robust representation for a music dataset [36, 37].

The exploration of this type of problem using musical data still need further investigation. The granularity of metadata information is a known limitation that makes it difficult to use different data sources to represent songs [38]. However, this difficulty can be mitigated with the availability of new datasets in recent years [39] and direct access to data via proprietary APIs[1,2,3] and with possibilities to design models that explore the relationships between features to build a representation of this information. Models based on heterogeneous networks have this characteristic due to the possibility of incorporating different features such as nodes in their architecture, exploring the relationships between them, and propagating information that allows reaching data with different similarity evaluation criteria.

In this paper, we will explore the existing gap regarding heterogeneous network, music multimodality and instance selection methods. We explore heterogeneous networks to model different types of information e determine relationships among songs that share the same attributes in a way that allows us to measure the relevance of music for its class.

3. Music Instance Selection from Heterogeneous Networks

Our instance selection method is based on two steps. The first step involves the generation of heterogeneous networks from the multiple modalities of the music dataset. The second step consists of a method for heterogeneous network regularization to generate a ranking function for each song on the network.

The generation of the heterogeneous network is performed by processing the music training set. The network relationships are extracted by different music metadata, as illustrated in Table 1. The relationships between songs and their metadata are mainly generated through k-partition network structures. In Figure 2, we can observe a representation of the network #4, where the network has a 3-partite structure connected by a constructed feature using the k-means[4] algorithm partitions from the melspectrogram data and the related artist features.

All cluster-based features are obtained using the k-means algorithm. The k parameter (number of clusters) was chosen by an iterative method that searched the highest silhouette[5] measure considering all the k values between $k = 2$ and $k = (class\_size/2)$.

We used traditional features that can be easily obtained from music metadata, and raw audio signals. They are described as follows:

- **Song**: is the music_id that can be referenced to other features like: melspectogram, metadata, lyrics, and others;
- **k-means cluster**: is the k-means cluster allocated for the song;

Table 1: Overview of the network construction strategy.

#Network	Features	#K
1	Song and k-means cluster	2
2	Song, k-means cluster and closest_cluster	3
3	Song, k-means cluster and sub_cluster	3
4	Song, k-means cluster and related_art	3
5	Song, k-means cluster and related_genre	3
6	Song, k-means cluster and related_music	3
7	Song, k-means cluster, related_art and related_genre	4
8	Song, k-means cluster, related_art and related_music	4
9	Song, k-means cluster, related_genre and related_music	4
10	Song, k-means cluster, related_art, related_music and related_genre	5

- **closest_cluster**: is the k-means cluster with the closest centroid to the song, apart from it's cluster;
- **sub_cluster**: is another k-means cluster that was executed within the genre of the song using the silhouette index to define ideal number of sub clusters;
- **related_art**: are all the related artists to a song according to dataset metadata.
- **related_genre**: are all the related genres to a song according to dataset metadata.
- **related_music**: are all related songs to a song according to dataset metadata.

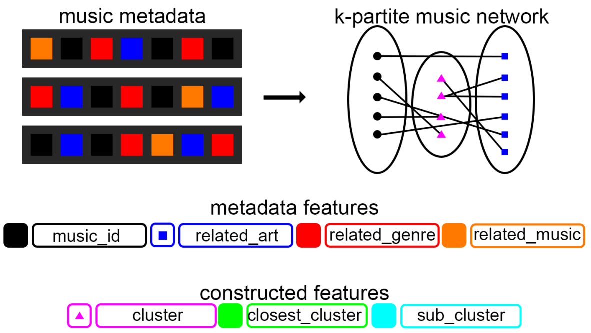

Figure 2: Example of a k-partite network used for feature extraction.

In the generation of heterogeneous networks, both the relationships extracted from cluster-based features and relationships extracted by metadata information on music genres, songs, and artists generate an unweighted network. More formally, we develop a heterogeneous network $N(O, R)$, where O is a set of nodes in the network and R is the relations between the nodes. We use an $O_L \subset O$ node subset (labeled song) for the second step of our method, i.e., the regularization framework defined in Equation 1.

$$Q(\mathbf{F}) = \frac{1}{2}\sum_{o_i,o_j \in O} r_{o_i,o_j}(\mathbf{f}_{o_i} - \mathbf{f}_{o_j})^2 + \mu \sum_{o_i \in O^L} (\mathbf{f}_{o_i} - \mathbf{y}_{o_i})^2 \qquad (1)$$

We want to find a matrix $\mathbf{F}$ that minimizes Equation 1. The node label information of the subset $o_i \in O_L$ is represented by the vector $\mathbf{y}_{o_i}$. A randomly initialized vector $\mathbf{f}_{o_i} \in \mathbf{F}$ is associated with each node o_i in the network.

[1] Last.fm API. Available at https://www.last.fm/api.

[2] Spotify API. Available at https://developer.spotify.com/documentation/web-api/.

[3] Pandora API. Available at https://developer.pandora.com/

[4] K-means is a clustering algorithm that partitions the data minimizing the mean squared error when selecting points near a synthetic center point called centroid [40]

[5] Silhouette is a validation measure for cluster partitions [41]

The first term, $\frac{1}{2}\sum_{o_i,o_j \in O} r_{o_i,o_j}(\mathbf{f}_{o_i} - \mathbf{f}_{o_j})^2$, aims to minimize the distance between the label information of two neighboring vertices in the network according to the relation r_{o_i,o_j}. The second term, $\mu \sum_{o_i \in O^L}(\mathbf{f}_{o_i} - \mathbf{y}_{o_i})^2$, aims to minimize the distance between the vector $\mathbf{f}_{o_i}$ estimated for labeled vertices $o_i \in O_L$. The higher the μ parameter value, the greater the preservation of label information during regularization.

Our regularization framework is a specific case of the LLGC method (Learning with local and global consistency) [42]. We use a label propagation strategy to minimize Equation 1. The strategy is based on random walking, where for each iteration, a node propagates its label information to its neighbors, considering the probability of a path in the network. As a result, the process converges after some iterations, in which the $\mathbf{F}$ matrix shows little change between consecutive iterations.

The matrix $\mathbf{F}$ resulting from the regularization process can be interpreted as membership of the network nodes for each label. Our instance selection process exploits the difference between the membership vector $\mathbf{f}$ and the real label information $\mathbf{y}$ to determine a song's score, as defined in Equation 2. The lower this score, the greater the evidence that the song is well allocated in its original label considering all relationships in the network. On the other hand, the higher the score, the song represents more challenging instance of the classification problem, for example, allocated in the decision boundary between two or more classes.

$$score(o_i) = \|\mathbf{f}_{o_i} - \mathbf{y}_{o_i}\|^2 \qquad (2)$$

For the instance selection, we set up three scenarios, that select only the amount we want for each training step, described as follows:

1. **Instance Selection L-Scores:** we select the m instances with the lowest score values.
2. **Instance Selection H-Score:** we select the m instances with the highest score values.
3. **Instance Selection LH-Score:** We selected m instances composed of the lowest and highest score values, i.e., a combination of the two previous strategies.

The first L-Score scenario aims at instance selection with "easier" instances for the classification model. However, such a strategy can generate classification models with very simple decision functions, which affect the classification of unseen data. On the other hand, the H-Score scenario aims at instance selection with "harder" instances that are more difficult to classify, generating more complex decision functions that are susceptible to overfitting. Thus, the LH-Score scenario is an attempt to obtain a trade-off between the two previous scenarios.

4. Experimental analysis

In this section, we describe the settings of the experiments carried out with the proposal evaluation scenarios. We also discussed the network architectures built and the performance obtained in each scenario.

4.1. Experimental setup

To evaluate our proposal, we use two versions of the 4MuLA [13] music dataset: Tiny (T-4MuLA) and Small (S-4MuLA). The 4MuLA dataset contains melspectrogram information extracted from audio content and various metadata like genres, artists, and related songs. We have chosen it due to the larger musical diversity and the feature structure, like the relationships between artists, songs, and genres required in our network-building process. A summary is shown in Table 2.

Table 2: Overview of the datasets' characteristics.

	T-4MuLA	S-4MuLA
Songs	1988	9661
Artists	93	491
Genres	27	51
Min instances in genre	9	11
Max instances in genre	350	1169
Avg instances in genre	73.62	189.43
Std instances in genre	82.26	260.86

We experimented with all approaches in our proposal, selecting different percentages of training data in both datasets. Thus, we compare our results with specialist and random selection approaches, both using the same input training data. The specialist approach was constructed using a Naive Bayes classifier to indicate an instance's confidence score. The confidence was considered when the class predicted is equal to the true class. The train and test steps occurred using the Leave-One-Out cross-validation strategy to have the max instance by class tagged with the score. We highlight that we created the train and test subsets from the original dataset. However, we showed only the train data to the approach to select the more representatives instances. The test subset was shown only in the evaluation step as unseen data. It is important to mention that all experiments were executed 10 times using different seeds to randomize the division of the data.

To obtain classification performance results, we trained a Random Forest classifier using the more representative instances indicated by all approaches in each data percentage for all scenarios aforementioned. We chose the F1-score as evaluation metric The goal is to solve a multiclass classification to identify the genre of an unknown test split to measure the quality of selected instances, and to discuss the metadata relationship relevance to music representation. We discuss the experimental results considering two aspects: (1) an overview of the F1-score in each heterogeneous network compared to specialist and random approaches; and (2) a behavior analysis for the best scenario of each method with an increase in the training set.

4.2. Network structure

Figure 3 shows the results in 5 scenarios for the T-4Mula dataset. The boxplot contains the F1-score distribution

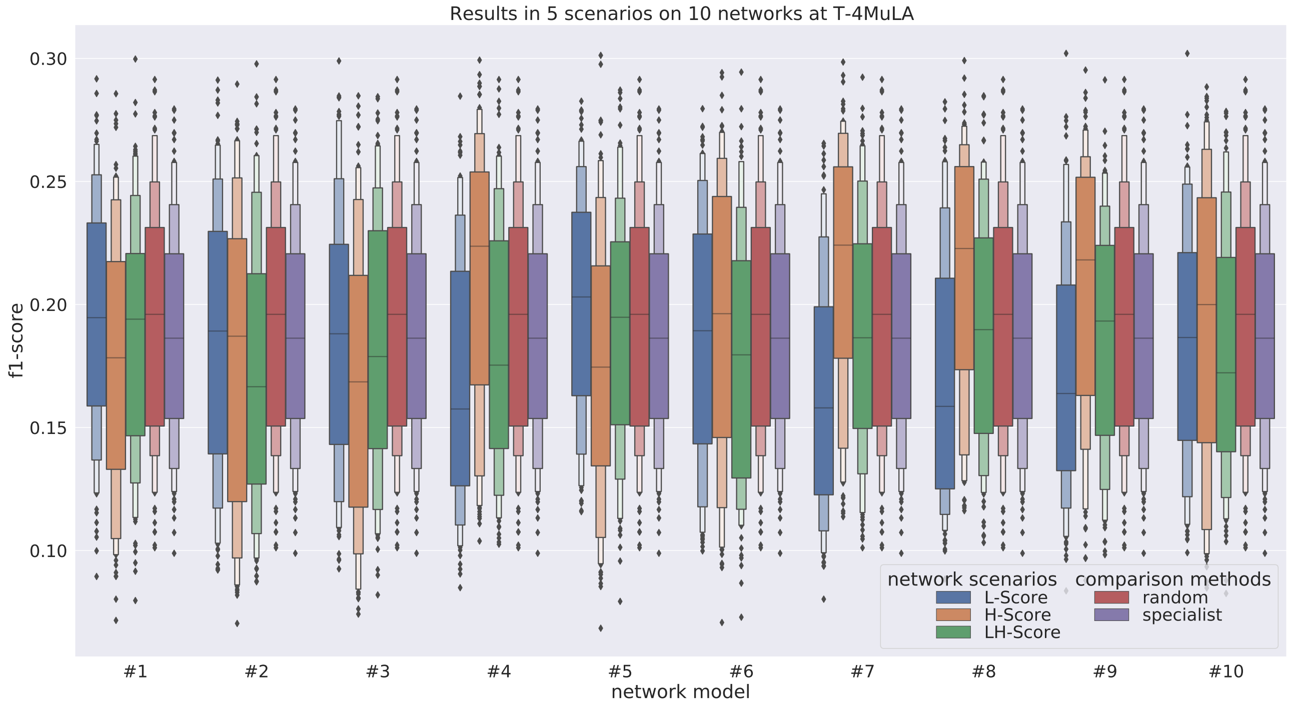

Figure 3: Results overview for the T-4MuLA dataset.

of all networks in the evaluated scenarios together with the two reference approaches. These results disregard the training set size. From this plot, we can observe that most of the methods achieve approximate performance.

The initial three network architectures have nodes formed by cluster information obtained only by melspectrogram. However, due to the difficulty of conceptualizing and assigning a musical genre considering only the audio content, we noticed that only this type of information has limitations highlighting more representative songs for a genre. In this way, randomly selecting instances proved to be more efficient.

In the following networks, we have the metadata relationship information, and we can observe a contribution in the discrimination of the songs. The using the more connected instances to training got better performance, with much higher values the median of comparable methods. It is essential to mention that four networks #4, #7, #8, and #10 have connections with nodes representing the related artists of the songs, indicating that this feature contains helpful information.

The results we obtained for the S-4MuLA dataset differ from T-4MuLA in some key areas. For example, in Figure 4 we can see that the L-Score scenario obtained better results than others scenarios in most of the networks. However, the random sampling got overall better results or was similar to the network approach.

Unlike classification results from T-4MuLA, networks #5, #6, and #10 obtains better classification performance. When analyzing the attributes they have in common, we notice that the relationships between similar genres and similar music are present. These relationships allow us to infer that for a scenario with a large volume of music, the information that indicates the relationship between artists and songs is relevant, and songs and related genres tend to be useful for classification.

4.3. Best scenarios

To provide another point of view from the results, we selected the best scenarios within our general plot to have a deeper look at how each feature changes the performance within the network. In the tables 3 and 4 we show F1-score behavior with an increasing training set. We selected the iteration in which the approach had the highest mean F1-score, and we established the individual value for each percentage.

Both tables show us that when considering a scenario where we have relationships capable of helping in musical discrimination, the instance selection process proves to be valid as a pre-processing step for the genre classification task. With an increase in the volume of training data, the performance of the proposed network increased the difference with the comparison methods. Such results show that instance selection from heterogeneous networks is competitive, but it requires careful tuning of the regularization process and the choice of scenarios. This strategy can be explored with visualization tools, taking advantage of interpretability and visual data exploration in heterogeneous networks.

We performed Friedman's statistical test with Nemenyi's post-test [43] to compare the scenarios to selection instances and the networks proposed. Figure 5 presents the result of the Friedman test with Nemenyi's post-test through the critical difference diagram. The approaches

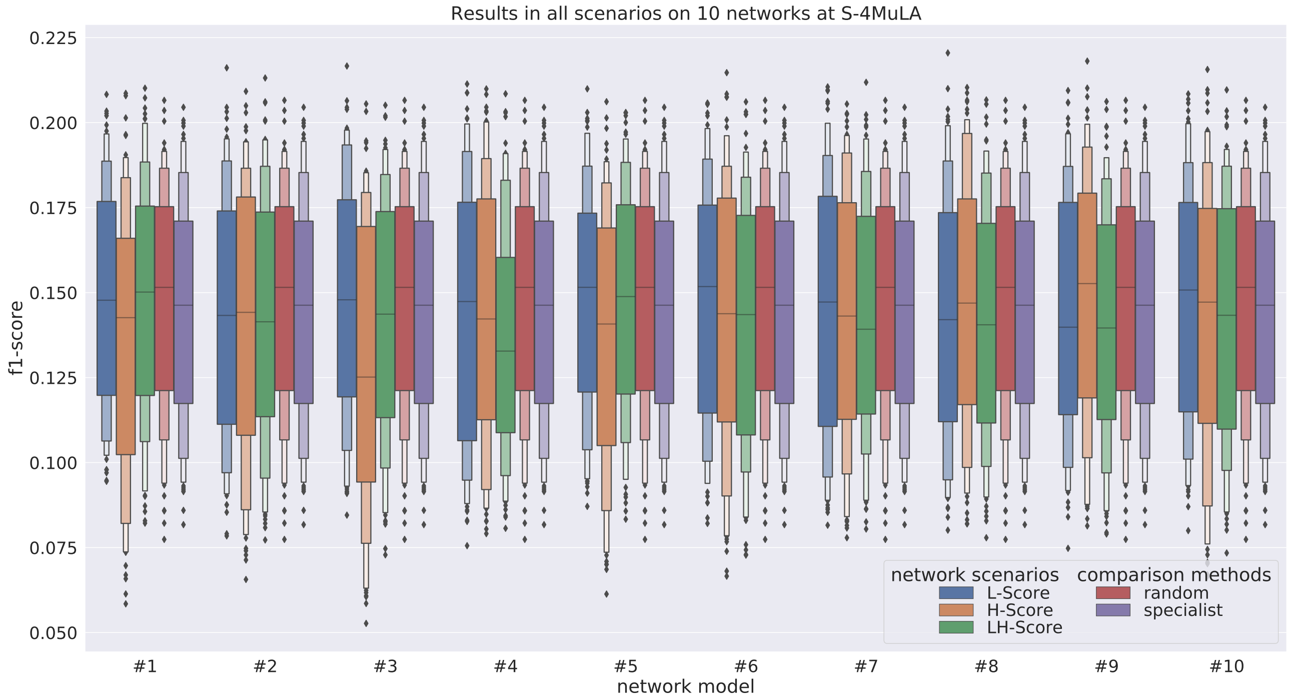

Figure 4: Results overview for the S-4MuLA dataset.

Table 3: Results considering the training size increase in the T-4MuLA dataset

F1-score										
	Percentage of training set for instance selection.									
Approach	10%	20%	30%	40%	50%	60%	70%	80%	90%	Scenario
Network #1	0.0802	0.0983	0.1254	0.1552	0.1737	0.2152	0.2425	0.2461	0.2857	L-Score
Network #2	0.1087	0.1171	0.1224	0.1518	0.2012	0.2007	0.2233	0.2492	0.2642	L-Score
Network #3	0.0807	0.0830	0.1090	0.1456	0.1779	0.2090	0.2398	0.2537	0.2848	L-Score
Network #4	0.1147	0.1624	0.1990	0.2237	0.2472	0.2651	**0.2934**	0.2853	**0.2993**	H-Score
Network #5	0.0684	0.1067	0.1109	0.1385	0.1745	0.1825	0.1953	0.2474	0.2547	L-Score
Network #6	0.1125	0.1597	0.1687	0.1761	0.2236	0.2372	0.2694	0.2850	0.2727	H-Score
Network #7	0.1290	0.1519	0.1965	**0.2493**	**0.2662**	**0.2708**	0.2814	**0.2985**	0.2931	H-Score
Network #8	0.1183	0.1561	0.2082	0.2376	0.2541	0.2695	0.2810	0.2920	0.2991	H-Score
Network #9	**0.1686**	**0.1630**	**0.2170**	0.2381	0.2393	0.2337	0.2555	0.2536	0.2724	H-Score
Network #10	0.1150	0.1398	0.2000	0.2045	0.2166	0.2306	0.2418	0.2802	0.2771	H-Score
specialist	0.1568	0.1492	0.1537	0.1823	0.2131	0.2385	0.2545	0.2327	0.2749	-
random	0.1419	0.1407	0.1913	0.1960	0.2135	0.2395	0.2430	0.2710	0.2616	-

Table 4: Results considering the training size increase in the S-4MuLA dataset

F1-score										
	Percentage of training set for instance selection.									
Approach	10%	20%	30%	40%	50%	60%	70%	80%	90%	Scenario
Network #1	0.0937	**0.1155**	0.1341	0.1453	0.1569	0.1719	0.1820	0.1964	0.2043	LH-Score
Network #2	0.0997	0.1082	**0.1384**	**0.1554**	0.1653	0.1662	0.1817	0.1937	0.2161	L-Score
Network #3	0.0971	0.1139	0.1355	0.1431	**0.1728**	0.1807	0.1928	0.1984	0.2166	L-Score
Network #4	0.0867	0.1034	0.1212	0.1349	0.1540	0.1718	0.1952	0.2005	0.2083	H-Score
Network #5	0.0945	0.1236	0.1322	0.1478	0.1510	0.1690	0.1954	0.2012	0.2099	L-Score
Network #6	0.1013	0.1137	0.1291	0.1437	0.1569	0.1694	0.1841	0.1974	0.2056	L-Score
Network #7	0.0971	0.0954	0.1181	0.1471	0.1607	0.1679	0.1906	0.1966	0.2105	L-Score
Network #8	0.0900	0.1068	0.1240	0.1458	0.1480	**0.1837**	0.1896	0.2028	0.2087	H-Score
Network #9	0.0871	0.1057	0.1237	0.1461	0.1658	0.1781	**0.1980**	**0.2076**	**0.2181**	H-Score
Network #10	**0.1039**	0.1108	0.1286	0.1551	0.1569	0.1753	0.1887	0.1938	0.2057	L-Score
specialist	0.0973	0.1045	0.1254	0.1374	0.1413	0.1647	0.1710	0.1909	0.1999	-
random	0.0951	0.1105	0.1234	0.1343	0.1565	0.1643	0.1740	0.1868	0.1921	-

connected by a line do not have statistically significant differences between them. For the scenarios, as seen in figures 3 and 4, the performance of networks, regardless of the scenario, has similar values. According to this diagram, we cannot point to any superiority on a statistical basis. However, alone, we can note that the LH-Score and random scenarios have a specific difference. We can relate this difference with the definition of both scenarios, where the first is formed by a combination of instances more significant with potential for increase the discrimination, while the second is a random selection of instances. The other scenarios show a more considerable similarity between the decision functions, regardless of the instance selection strategy.

About networks, there is also no statistically superior architecture. However, we can analyze the contribution of the features used by comparing the structure of the networks. For example, we can discuss the relevance of nodes with information related to the closest cluster and subcluster existing only in networks #2 and #3, as well as evaluate the relationships between the metadata of songs, artists, and genres used as an individual or combined feature in the other networks. Due to the similarity of the results, the possibility of discarding or prioritizing a specific type of node enables a cost reduction in the network construction process. Furthermore, it tends to impact in the discrimination of instances.

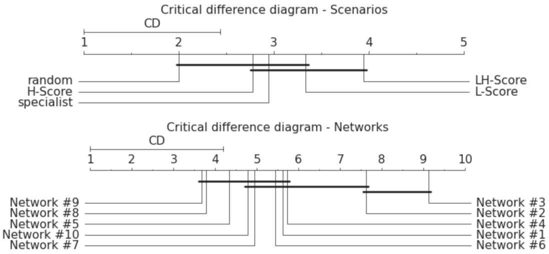

Figure 5: Analysis of the critical difference between scenarios and networks.

5. Conclusion

In this work, we present an evaluative study on the instance selection task using multimodal musical data. We built ten different heterogeneous network architectures that contained information based on melspectrogram and metadata to handle dataset multimodality, indicating the relationship between a song and attributes such as artist, genres, and related pieces. We compare our proposed approach to the (i) method with information from a specialist model to indicate the ideal instances for training and (ii) a random sampling selection process.

In general, the random sampling process for selecting instances proved to be efficient and less costly. However, by analyzing each approach individually, we observed the relevance of metadata and their relationships to compose a musical identification. From this observation, we can identify conditions where there is no large volume of data available and assess what kind of information is helpful knowledge for instance selection.

In future work, we intend to develop an application of this proposal as a web service where users can propose nodes and relationships to qualify the music dataset and create personalized scenarios where networks can contribute to identifying unique characteristics of the musical genres.

6. Acknowledgments

This work was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code PROEX-12049601/D and 88887.513429/2020-00, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) – Finance Code 2019/25010-5 and 2019/07665-4, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) – Finance Code 426663/2018-7.

References

[1] Markus Schedl, Emilia Gómez, Julián Urbano, et al. Music information retrieval: Recent developments and applications. *Foundations and Trends in Information Retrieval*, 8(2-3):127–261, 2014.

[2] YV Murthy and Shashidhar G Koolagudi. Content-based music information retrieval (cb-mir) and its applications toward the music industry: A review. *ACM Computing Surveys*, 51(3):45, 2018.

[3] Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2392–2396, 2016.

[4] Yandre MG Costa, Luiz S Oliveira, and Carlos N Silla Jr. An evaluation of convolutional neural networks for music classification using spectrograms. *Applied soft computing*, 52:28–38, 2017.

[5] Joan Serrà, Emilia Gómez, and Perfecto Herrera. Audio cover song identification and similarity: background, approaches, evaluation, and beyond. In *Advances in Music Information Retrieval*, pages 307–332. Springer, 2010.

[6] Diego F. Silva, Chin-Chia M. Yeh, Gustavo E. A. P. A. Batista, and Eamonn Keogh. SiMPle: assessing music similarity using subsequences joins. In *International Society for Music Information Retrieval Conference*, pages 23–29, 2016.

[7] Jordi Pons, Oriol Nieto, Matthew Prockup, Erik Schmidt, Andreas Ehmann, and Xavier Serra. End-to-end learning for music audio tagging at scale. In *International Society for Music Information Retrieval Conference*, pages 637–644, 09 2018.

[8] Christine Dewi, Rung-Ching Chen, Yan-Ting Liu, et al. Similar music instrument detection via deep convolution yolo-generative adversarial network. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pages 1–6. IEEE, 2019.

[9] W. Guo, J. Wang, and S. Wang. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

[10] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, Mar 2020.

[11] Y. Lin, C. Chung, and H. H. Chen. Playlist-based tag propagation for improving music auto-tagging. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2270–2274, 2018.

[12] Y. Lin and H. Chen. Tag propagation and cost-sensitive learning for music auto-tagging. *IEEE Transactions on Multimedia*, pages 1–1, 2020.

[13] Angelo Cesar Mendes da Silva, Diego Furtado Silva, and Ricardo Marcondes Marcacini. 4mula: A multitask, multimodal, and multilingual dataset of music lyrics and audio features. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '20, page 145–148, New York, NY, USA, 2020. Association for Computing Machinery.

[14] J. Arturo Olvera-López, J. Ariel Carrasco-Ochoa, J. Francisco Martínez-Trinidad, and Josef Kittler. A review of instance selection methods. *Artificial Intelligence Review*, 34:133–143, 2010.

[15] Mellish C Brighton H. Advances in instance selection for instance-based learning algorithms. *Data Mining and Knowledge Discovery*, 6:153–172, 2002.

[16] van der Aalst Wil Sani Mohammadreza Fani, van Zelst Sebastiaan J. Improving the performance of process discovery algorithms by instance selection. *Computer Science and Information Systems*, 17:927–958, 2020.

[17] Marek Grochowski and Norbert Jankowski. Comparison of instance selection algorithms ii. results and comments. In Leszek Rutkowski, Jörg H. Siekmann, Ryszard Tadeusiewicz, and Lotfi A. Zadeh, editors, *Artificial Intelligence and Soft Computing - ICAISC 2004*, pages 580–585, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[18] Jianping Zhang, Yee-Sat Yim, and Junming Yang. *Intelligent Selection of Instances for Prediction Functions in Lazy Learning Algorithms*, pages 175–191. Springer Netherlands, Dordrecht, 1997.

[19] Chih-Fong Tsai, Wei-Chao Lin, Ya-Han Hu, and Guan-Ting Yao. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, 477:47–54, 2019.

[20] Chih-Fong Tsai and Fu-Yu Chang. Combining instance selection for better missing value imputation. *Journal of Systems and Software*, 122:63–71, 2016.

[21] Joel Luís Carbonera and Mara Abel. Efficient instance selection based on spatial abstraction. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 286–292, 2018.

[22] Peter Hart. The condensed nearest neighbor rule. *IEEE transactions on information theory*, 14(3):515–516, 1968.

[23] G Ritter, H Woodruff, S Lowry, and T Isenhour. An algorithm for a selective nearest neighbor decision rule. *IEEE Transactions on Information Theory*, 21(6):665–669, 1975.

[24] Joel Luis Carbonera and Mara Abel. A density-based approach for instance selection. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 768–774, 2015.

[25] B. Ramesh and J.G.R. Sathiaseelan. An advanced multi class instance selection based support vector machine for text classification. *Procedia Computer Science*, 57:1124–1130, 2015. 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015).

[26] Junhai Zhai, Xizhao Wang, and Xiaohe Pang. Voting-based instance selection from large data sets with mapreduce and random weight networks. *Information Sciences*, 367-368:1066–1077, 2016.

[27] Miguel Lopes, Fabien Gouyon, Alessandro L. Koerich, and Luiz E.S. Oliveira. Selection of training instances for music genre classification. In *2010 20th International Conference on Pattern Recognition*, pages 4569–4572, 2010.

[28] Ismail M. Anwar, Khalid M. Salama, and Ashraf M. Abdelbar. Instance selection with ant colony optimization. *Procedia Computer Science*, 53:248–256, 2015. INNS Conference on Big Data 2015 Program San Francisco, CA, USA 8-10 August 2015.

[29] Nicolás García-Pedrajas and Aida de Haro-García. Boosting instance selection algorithms. *Knowledge-Based Systems*, 67:342–360, 2014.

[30] Aida de Haro-García, Gonzalo Cerruela-García, and Nicolás García-Pedrajas. Instance selection based on boosting for instance-based learners. *Pattern Recognition*, 96:106959, 2019.

[31] Marcin Blachnik. Ensembles of instance selection methods: A comparative study. *International Journal of Applied Mathematics and Computer Science*, 29(1):151–168, 2019.

[32] Wenqin Chen, Jessica Keast, Jordan Moody, Corinne Moriarty, Felicia Villalobos, Virtue Winter, Xueqi Zhang, Xuanqi Lyu, Elizabeth Freeman, Jessie Wang, Sherry Kai, and Katherine M. Kinnaird. Data usage in mir: History & future recommendations. In *International Society for Music Information Retrieval Conference*, 2019.

[33] Sergio Oramas, F. Barbieri, Oriol Nieto, and Xavier Serra. Multimodal deep learning for music genre classification. *Transactions of the International Society for Music Information Retrieval*, 1:4–21, 2018.

[34] Ngo Tung Son, Duong Xuan Hoa, and Vu Thanh. Learning sparse representation from multiple-source data for relative similarity in music. In *Proceedings of the 2018 International Conference on Computational Intelligence and Intelligent Systems*, CIIS 2018, page 1–4, New York, NY, USA, 2018. Association for Computing Machinery.

[35] Ho-Hsiang Wu, Chieh-Chi Kao, Qingming Tang, Ming Sun, Brian McFee, Juan Bello, and Chao Wang. Multi-task self-supervised pre-training for music classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 556–560, 06 2021.

[36] Y.R. Pandeya and J Lee. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimed Tools Appl*, 2020.

[37] Changfeng Chen and Qiang Li. A multimodal music emotion classification method based on multifeature combined network classifier. *Mathematical Problems in Engineering*, 2020.

[38] Michael I. Mandel and D. Ellis. Multiple-instance learning for music information retrieval. In *International Society for Music Information Retrieval*, pages 577–582, 2008.

[39] Rachel M Bittner, Magdalena Fuentes, David Rubinstein, Andreas Jansson, Keunwoo Choi, and Thor Kell. mirdata: Software for reproducible usage of datasets. In *International Society for Music Information Retrieval (ISMIR) Conference*, 2019.

[40] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[41] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[42] Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328, 2004.

[43] Bogdan Trawinski, Magdalena Smetek, Zbigniew Telec, and Tadeusz Lasota. Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms. *Applied Mathematics and Computer Science*, 22(4):867–881, 2012.

CNRS - Univ. Paris 6 & PUC-Rio - France