International Journal of Engineering & Technology, 7 (3.24) (2018) 782-786

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Toward Accurate Music Classification Using Local Set-based Multi-label Prototype Selection

Wangduk Seo[1], Sanghyun Seo[2], Mucheol Kim[3], Jaesung Lee[4*]

[1]School of Software, Chung-Ang University, Korea
[2]Div. of Media Software, Sungkyul University, Korea
[3]Dept. of Computer & Software Engineering, Wonkwang University, Korea
[4]School of Software, Chung-Ang University, Korea
[*]Corresponding author E-mail: curseor@cau.ac.kr

Abstract

Background/Objectives: Multiple music tags enable quick searching and selection of music clips for by end-users to listen to. Our goal is to improve the accuracy of automatic music categorization.

Methods/Statistical analysis: We propose a local set-based multi-label prototype selection to remove noisy samples in datasets without transforming multi-label datasets to single-label datasets by searching the local set of each sample. To validate the superiority of proposed method, we use ten multi-label music datasets and Hamming loss as a performance measurement, which counts the symmetric difference between predicted labels and ground truth labels.

Findings: Considering time and cost, manual categorization of a large collection of music clips is generally impractical. As such, an automated approach for addressing this task through the training of music tags annotated from an online system is employed. In the real world, multiple labels can be annotated to a music clip by users of an online system, resulting in unintended noisy samples due to inaccurate annotations. Conventional methods attempt to transform multi-label datasets to single-label datasets that can yield additional computational cost and unintended removal of non-noisy samples. In this paper, we propose a novel prototype selection method for multi-label music categorization. Experimental results indicate that the proposed method performed the best performance on nine music datasets. From the experiment of CAL500 dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.1402, which indicates that 15,020 labels on average were correctly classified for 100 test samples. Compared to the second best performance by compared method, our method was able to classify 245 more labels.

Improvements/Applications: Experimental results using ten music datasets with different subjects revealed that the proposed method yields better performance when compared to conventional methods.

***Keywords*:** Auto Music Tags Annotation, Multi-label Learning, Prototype Selection, Pruned Problem Transformation, Local Set-based Smoother

1. Introduction

Recently, music recommendation applications such as playlist recommendations systems [1], automated tagging systems [2], and emotion recognition systems [3] have examined the popularity of social network services and smartphones [4]. Music categorization that identifies relevant tags or labels (such as mood, theme, and usage) and assigns them to individual songs in a large music collection is the most important task for the realization of an appealing music recommendation service. This is because music tags enable users to quickly find the types of music they are looking for [5]. In practice, a variety of relevant tags can be assigned to a song, which makes manual categorization of a large music collection difficult [6]. Manual categorization requires considerable cost and time, which means user demands and song publishing deadlines may not be met. Manual categorization of a huge music corpus can be prevented by adopting a machine learning method for music categorization that is trained using user-allocated tags from an online system.

In the music information retrieval community, considering the task of music categorization as a multi-label learning problem has attracted significant scientific interest [7]. In the works of [8-10], music emotion categorization is modeled as a multi-label classification because one song can be associated with multiple labels. Further, [11] highlights the importance of minimizing the amount of acoustic information required for recommending music clips to users on mobile devices. A common drawback of implementing an automatic music categorization system using the tags from an online system is the existence of noisy samples, because of the subjectivity of ratings and labels that are assigned by non-expert users [12]. This can lead to low performance of automatic label assignment by a trained online system. To remove noisy samples, many studies report that prototype selection, which selects and removes unimportant samples for accurate classification, is effective [13].To achieve an accurate automatic label assignment system for music, we propose an automatic music categorization system using a local set-based multi-label prototype selection method to identify and remove unwanted samples from a training set. We conducted experiments on 10 multi-label music datasets and demonstrate that our proposed technique can improve the accuracy of music categorization by denoising data during the training phase.

2. Review of proposed technique

In study of [14], auto music annotation, such as mood, genre, style classification, is naturally cast as a multi-label learning. For instance, not all songs are necessarily categorized as a single genre; they can be a multi-genre, such as ballad rock. In mood classification, different parts of the same song can have a different mood.

Assume that$W \subset \mathbb{R}^d$ represents a set of training patterns that are constructed from a group of musical features. Each music clip or training pattern $w_i \in W$, where $1 \leq i \leq |W|$, is then assigned to a certain label subset $\lambda_i \subseteq L$, where $L = \{l_1, \ldots, l_{|L|}\}$is a finite set of tags or labels. In practice, a subset of multi-labeled patterns or samples can be allocated to less relevant label subsets. Thus, our goal is to identify $S \subset W$ that contains important patterns for accurate training.

2.1. Description of multi-label music dataset

We experimented on 10 music datasets, 8 of which were collected through a Korean national research project, while CAL500 and Emotions datasets were gathered from a music tag annotation application in which the music retrieval system learns a relation between acoustic features and words from a dataset of annotated audio tracks [15]. The 8 research project datasets have not been discussed in the music information retrieval community, so we summarize their details as follows:

- **Bugs2664:** Bugs2664 created by gathering 2,664 music clips from an online music streaming service in Korea, and most of them correspond to K-pop music. Each music clips is assigned 40 tags that are categorized into seasons, emotions, usage, and places.

- **BugsEmo:** BugsEmo created by subsampling the Bugs2664 dataset by only considering seven emotion tags. The acoustic features of the Bugs datasets were extracted by the MIR toolbox [19].

- **Style812, Genre3, and Highlight:** In the Style812 dataset, 812 music clips are labeled as one of three music styles: rhythmic, romantic, and melancholy. The Genre3 dataset was created by extracting acoustic features from the same 812 music clips as in the Style812 dataset; this dataset was created to identify time-variant musical themes, including genres, highlights, and emotions. Thus, it contains all music pieces as opposed to simply selecting a representative piece from each music clip. Similarly, the Highlight dataset was created by using the same procedure as Genre3, but each label indicates whether each piece of a music clip is a highlight. The acoustic features of these three datasets were extracted by the MIR toolbox.

- **KOCCA40**: This dataset was created with information from undergraduate classes on music information retrieval. In order to encourage students, 40 music clips that are confirmed as easily learned by a machine learning algorithm were selected. Each music clip is assigned one of four different labels: passionate, breezy, depressed, and peaceful. The acoustic features of the KOCCA40 dataset were extracted by the MIR toolbox.

- **MusicEmo-A and MusicEmo-B:** In MusicEmo-A, 864 acoustic features were extracted by the MIR toolbox from 100 music clips and were labeled approximately 500 times through an online annotation system. In MusicEmo-B, 346 audio features were extracted by the MIR toolbox from 565 music clips and labeled approximately 3600 times. Each music clip was assigned relevant tags, including excitement, distress, depression, and contentment. Earlier versions of these two datasets were discussed in one of our previous studies [16]. In this study, 21 errors in feature values were corrected.

2.2. Conventional methods

A major trend in multi-label machine learning studies is the application of conventional classification methods after transforming the label sets in one or multiple ways [6]. Two well-known transformation approaches are pruned problem transformation (PPT) and label powerset (LP) [17]. The LP approach implements indices for each unique combination of labels in multi-label datasets and assigns all samples' labels to such indices in order to change one class value. PPT is a modified version of LP that discards samples that have been assigned to rare or less utilized label subsets during the training phase. LP and PPT have an advantage in that they can use conventional methods of prototype selection for single-label datasets; however, they have unwanted side effects [18], such as imbalance in transformed single label datasets. Additionally, they commonly suffer from poor multi-label classification due to the lack of interaction with multi-label classifiers. After the transformation procedure is completed, conventional prototype selection methods for single-label dataset, such as local set-based smoother (LSSm) or Wilson's method (WM), can then be applied to identify important samples [13].

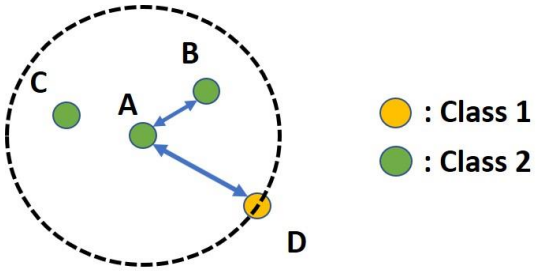

Figure 1: Example of local set on 2 dimensional space

LSSm was the first algorithm proposed for the prototype selection and involves the use of local sets. A local set is a set of samples contained within the hypersphere of largest area that is centered on the target sample, such that it does not contain instances from any other class. The nearest instance of a different class is called the nearest enemy. Figure 1 shows an example of a local set; if *A* in Class 1 is the target sample, its nearest enemy is sample *D*, and the cardinality of the local set (the number of samples inside the local set) is 2. LSSm takes the cardinality of the target sample's local set and the number of samples that have the target sample as their nearest enemy. If the former is larger than the later, then the target sample remains in the training data. Otherwise, the sample is removed.

WM uses the three-nearest neighbor rule for prototype selection. LSSm and WM were originally proposed for use on single-label datasets and can only be applied to multi-label datasets with PPT or LP. Thus, combined versions such as PPT + LSSm or PPT + WM can be used for prototype selection on multi-label datasets. However, doing this causes additional computational cost, which makes it hard to quickly and accurately automatically allocate labels. To tackle this problem, our proposed method does not involve a transformation process.

2.3. The proposed methods

To avoid a transforming process of multi-label datasets into single label datasets, we directly calculate the distance between the samples inside the local set and nearest enemy. Let $NE(w_i) = w_j$ be the sample nearest to w_i, but assigned to λ_j, where $\lambda_i \neq \lambda_j$. The local set of w_i can then be defined as:

$$LS(w_i) = \{w_k | dist(w_i, w_k) < dist(w_i, w_j)\} \quad (1)$$

where $dist(\cdot,\cdot)$ is the Euclidean distance between two samples and $\lambda_k = \lambda_i$. The proposed method eliminates w_i from the training

samples if $|NE(w_i)| > |LS(w_i)|$.

Figure 2 shows an example of our proposed method. In Figure 2(a), if *A* is the target sample, the nearest sample that shares a label set is *B*and the nearest enemy is *C*. Since the distance between *A* and *B* is shorter than that between *A* and *C*, the target sample *A* is not eliminated. However, in Figure 2(b), the opposite is true, so target sample *A* is eliminated as a noisy sample. By repeatedly applying this process to all the patterns in the training dataset, noisy samples can be identified and eliminated in a batch process.

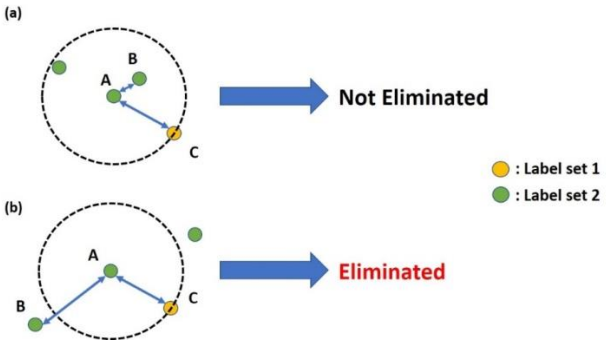

Figure2: Example of the proposed method

3. Results and Discussion

We conducted experiments to determine the performance of our proposed method on 10 multi-label music datasets. The 8 datasets except CAL500 and Emotions were collected through a national research project in Korea. These 8 datasets are composed of acoustic features extracted by MIR toolbox [19]. The acoustic features are analysis data of dynamics, fluctuation, rhythm, spectral, and tonal features. These features are extracted from 40 seconds audio clips from each song. The other 2 datasets, CAL500 and Emotions datasets, were generated from a music tag annotation application in which the music retrieval system learns a relation between acoustic features and words from a dataset of annotated audio tracks [15]. These datasets can be downloaded from our website (http://mi.cau.ac.kr/?f=teaching&m=proc_amc). For a more detailed description of these datasets, refer back to Section 2.

3.1. Characteristics of datasets

Table 1 lists the standard characteristics of the multi-label music datasets employed in our experiments. $|W|$ denotes the number of samples in the dataset, and $|L|$denotes the number of labels. The label cardinality *Card.* represents the average number of labels for each sample. The label density *Den.* denotes the label cardinality with respect to the total number of labels. The number of distinct label sets *Distinct.* indicates the number of unique label subsets in *L*. *Subject* represents the application that the label of each dataset is related to.

Table 1: Standard characteristics of the multi-label music datasets

Datasets	$\|W\|$	$\|F\|$	$\|L\|$	*Card.*	*Den.*	*Distinct.*	*Subject*
Bugs2664	2664	137	40	1.917	0.048	666	Tag
BugsEmo	753	109	7	1.000	0.143	7	Emotion
CAL500	502	68	174	26.044	0.150	502	Tag
Emotions	593	72	6	1.868	0.311	27	Emotion
Genre3	2597	365	3	1.000	0.333	3	Genre
Highlight	2597	365	2	1.000	0.500	2	Highlight
KOCCA40	40	123	4	1.000	0.250	4	Emotion
MusicEmo-A	100	864	4	1.530	0.383	11	Emotion
MusicEmo-B	565	346	4	1.292	0.323	9	Emotion
Style812	812	348	3	1.000	0.333	3	Style

3.2. Performance measurement

We compared our method with PPT+LSSm and PPT+WM using Hamming loss values with the Multi-label k-Nearest Neighbor classifier, which was trained with the prototypes identified using the three different methods [5]. Let $T = \{(t_i, \lambda_i)|1 \le i \le |T|\}$ be a set of test samples, where λ_i is a true label set for t_i. The Hamming loss is therefore defined as:

$$hloss(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{1}{|L|}\left|\lambda_i \vartriangle \widehat{\lambda_i}\right| \quad (2)$$

where $\widehat{\lambda_i}$ is the predicted label subset, and $\vartriangle$ denotes the symmetric difference between the two label sets. For fairness, we conducted a holdout cross-validation for each experiment [8]. In each iteration of each dataset, 80 % of the randomly chosen samples from a given dataset are used for training data, and the remaining 20% of the samples were used as the test set to obtain the hamming loss performance that we report. Each experiment on 10 datasets was repeated for 10 iterations, and the average value was used to represent the classification performance according to each prototype selection method. A low Hamming loss value indicates better multi-label classification accuracy.

3.3. Experimental results

Table 2 lists the experimental results for the proposed method and the conventional methods in terms of the Hamming loss and the average rank of all datasets. The best performance is indicated in a bold font. Experimental results indicate that the proposed method performed the best performance on the nine datasets except the Highlight dataset. It should be noted that the Hamming loss performance indicates the average number of labels that are incorrectly classified. For example, for the CAL500, the difference in Hamming loss value between our proposed method and PPT + LSSm is only 0.0140, which is a relatively insignificant difference. However, this means that approximately 245 more labels were correctly classified by the classifier trained with the proposed method compared to the one trained by PPT + LSSm. This confirms that the proposed method significantly outperformed PPT + LSSm for the CAL500 dataset and the eight other multi-label music datasets: Bugs2664, BugsEmo, Emotions, Genre3, KOCCA40, MusicEmo-A, MusicEmo-B, and Style812.

Table 2: Comparison results for prototype selection in terms of the hamming loss

Datasets	Proposed	PPT+LSSm	PPT+WM
Bugs2664	**0.0481**	0.0482	0.0487
BugsEmo	**0.1073**	0.1075	0.1214
CAL500	**0.1402**	0.1542	0.1542
Emotions	**0.2088**	0.2106	0.2364
Genre3	**0.0131**	0.0135	0.0143
Highlight	0.2657	0.2659	**0.2655**
KOCCA40	**0.2281**	0.2283	0.3656
MusicEmo-A	**0.2538**	0.2575	0.3213
MusicEmo-B	**0.1927**	0.1951	0.2192
Style812	**0.0805**	0.0824	0.0901
Avg. Rank	**1.10**	2.15	2.75

To show the superiority of our proposed method, we compared its Hamming loss performance with other methods based on the experimental results for each music dataset:

- From the experiment on the Bugs2664 dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.0481, which indicates that 20,286 labels on average were correctly classified for 533 test samples.
- From the experiment on the BugsEmo dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.1073, which indicates that 941 labels on average were correctly classified for 151 test samples.
- From the experiment of CAL500 dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.1402, which indicates that 15,020 labels on average were correctly classified for 100 test samples. Compared to the performance of PPT + WM, our method was able to classify 245 more labels.
- From the experiment of Emotions dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.2088, which indicates that 563 labels on average were correctly classified in average for the classification of for 119 test samples.
- From the experiment of Genre3 dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.0131, which indicates that 1,538 labels on average were correctly classified for 519 test samples.
- From the experiment of Highlight dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.2657, which is the second best Hamming loss performance out of the three models that we compared. This is the only dataset for which our method did not perform best.
- From the experiment of KOCCA40 dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.2281, which indicates that 25 labels on average were correctly classified for eight test samples.
- From the experiment of MusicEmo-A dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.2538, which indicates that 60 labels on average were correctly classified for 20 test samples.
- From the experiment of MusicEmo-B dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.1927, which indicates that 365 labels on average were correctly classified for 113 test samples.
- From the experiment of Style812 dataset, multi-label classification performance based on the training samples selected by our proposed method was 0.0805, which indicates that 448 labels on average were correctly classified for162 test samples.

Although the number of additional correctly classified labels varies according to the characteristics of each dataset, our detailed analysis on the experimental results indicates that the effectiveness of our proposed method become more significant as the number of labels increases.

4. Conclusion

An accurate annotation system for music annotation is required to reduce the costs of manual categorization of large music collections. For an accurate music classification, prototype selection for removing noisy samples can be effective. The conventional methods for implementing this have additional computational cost in transforming multi-label datasets into single-label datasets. Thus, we proposed an accurate music classification method that uses local set-based prototype selection for multi-label datasets. Experimental results showed that our proposed method offers superior performance compared to other prototype methods.

Our proposed method's target data domain is music; however, it could be applied to other domains. In future research, we will consider datasets from different domains, such as medical and text datasets. We would like to study this issue further.

Acknowledgment

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2017S1A6A3A01078538)

References

[1] Liebman, E., Saar-Tsechansky, M., & P. Stone. Dj-mc (2015). A reinforcement-learning agent for music playlist recommendation. In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, Istanbul, Turkey, pp. 591-599.

[2] Yan, Q., Ding, C., Yin, J., & Lv., Y. (2015). Improving music auto-tagging with trigger-based context model. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, pp. 434-438

[3] Bai, J., Feng, L., Peng, J., Shi, J., Luo, K., Li, Z., Liao, L., & Wang. Y. (2016). Dimensional music emotion recognition by machine learning. ***International Journal of Cognitive Informatics and Natural Intelligence***, 10(4), 74–89.

[4] Fu, Z., Lu, G., Ting, K. M., & Zhang, D. (2011). A survey of audio-based music classification and annotation. ***IEEE Transactions on Multimedia***, 13(2), 303-319.

[5] Nanni, L., Costa, Y., Lumini, A., & Kim, M. Y. (2017). Combining visual and acoustic features for music genre classification, ***Expert Systems with Applications***, 99(1), 987-996.

[6] Lee, J. & Kim, D.-W. (2016). Efficient Multi-Label Feature Selection Using Entropy-Based Label Selection, ***Entropy*** 18(1), 1-18.

[7] Papanikolaou, Y., Katakis, I., & Tsoumakas, G. (2016). Hierarchical partitioning of the output space in multi-label data. Retrieved from https://arxiv.org/abs/1612.06083(arXiv:1612.06083)

[8] Lee J. & Kim, D.-W. (2015). Memetic feature selection algorithm for multi-label classification. ***Information Sciences***, 293(1), 80-96.

[9] Lee J. & Kim, D.-W. (2017). SCLS: Multi-label feature selection based on scalable criterion for large label set. ***Pattern Recognition***, 66(1), 342-352.

[10] Zhang, Y., Gong, D.-W., Sun, X.-Y., & Y.-N. Guo. (2017). A PSO-based multi-objective multi-label feature selection method in classification. ***Scientific Reports***, 7(376), 1-12.

[11] Naula, P., Airola, A., Salakoski, T., and Pahikkala, T. (2014). Multi-label learning under feature extraction budgets. ***Pattern Recognition Letters***, 40(1), 56-65.

[12] Zhai, E., Li, Z. Li, Z. & Chen, G. (2016). Resisting tag spam by leveraging implicit user behaviors, ***Proceedings of the VLDB Endowment***, 10(3), 241-252.

[13] Leyva, E., Gonzalez, A. & Perez, R. (2015). Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective, ***Pattern Recognition***, 48(1), 1523-1537.

[14] Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2008). Multi-label classification of music into emotions, In Proceedings 9th International Society Music Information Retrieval, Philadelphia, USA, pp. 325-330.

[15] Ness, S. R., Theocharis, A., Tzanetakis, G., & Martins, L. G. (2009). Improving automatic music tag annotation using stacked generalization of probabilistic SVM outputs. In Proceedings of the 17th ACM International Conference on Multimedia, Beijing, China, pp. 705-708.

[16] Lee, J., Jo, J.-H., Lim, H., Chae, J.-H., Lee, S.-U., & Kim, D.-W. (2015). Investigating relation of music data: Emotion and audio signals. ***Lecture Notes in Electrical Engineering***, 330(1), 251-256.

[17] Read, J. (2008). A pruned problem transformation method for multi-

label classification. In Proceedings of New Zealand Computer Science Research Student Conference, Christchurch, New Zealand, pp. 143-150.

[18] Lee, J. & Kim, D.-W. (2013). Feature Selection for Multi-label Classification using Multivariate Mutual Information. ***Pattern Recognition Letters***, 34(3), 349-357.

[19] Lartillot, O. & Toiviainen, P. (2007). A MATALB toolbox for musical feature extraction from audio. In Proceedings of the 10th International Conference on Digital Audio Effects, Bordeaux, France, pp. 237-244.