International Journal of Engineering & Technology, 7 (4.39) (2018) 52-55

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], Jaesung Lee[3*]

[1]School of Computer Science and Engineering, Chung-Ang University, Korea
[2]Div. of Media Software, Sungkyul University, Korea
[4]School of Computer Science and Engineering, 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 the 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 the compared method, our method was able to classify 245 more labels.
Improvements/Applications: Experiments using ten different musical datasets showed that the proposed method demonstrated better performance than the compared methods.

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

1. Introduction

Social network services often include music recommendation applications such as playlist suggestion [1], auto-tagging [2], and music emotion recognition [3], which demonstrate their popularity and importance. To realize a compelling music recommendation service, the classification of music that identifies the relevant tags or labels such as gases and themes and assigns them to individual songs in a large collection of music is paramount. This is because music tags enable users to quickly find the types of music they are looking for [4,5]. 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.
Many researchers reported that it is attracting scientific attention that treating the task of classifying music as a multi-label learning problem [7]. For example, the classification of music emotions is modeled as a multi-label classification because a single song can be associated with multiple labels [8-10]. Also, Naula et al. highlight the importance of minimizing the amount of information required for recommending music clips to users on mobile devices [11]. 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 ten 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 a study of [14], auto music annotation, such as mood, genre,

style classification, is naturally cast as multi-label learning. For instance, not all songs are necessarily categorized as a single genre; they can be a multi-genre, such as a 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 with eight sets of music data collected through a national research project in Korea while *CAL500* and *Emotions* datasets are ready-mades and frequently-used datasets from a music tag annotation application that learns the relationship between sound characteristics and words. Among the ten datasets on the task of audio annotating [15], the eight project datasets summarized as follow.

- **Bugs2664 and BugsEmo:** *Bugs2664* dataset is made up of 2,664 music clips collected from Korea's online music streaming service, mostly K-pop music. Each music clip has 40 tags that are categorized as season, emotion, usage, and location. *MIR Toolbox* was used to extract sound characteristics [19]. The *BugsEmo* dataset is created by subsampling *Bugs2664* dataset, taking into account only seven emotional tags.
- **Style812, Genre3, and Highlight:** In the *Style812* dataset, 812 music clips are categorized into one of three styles: rhythm, romance, and melancholy. The *Genre3* dataset was created by extracting sound characteristics from the same 812 music clip as in *Style812* dataset. This dataset is designed to identify music themes that change over time, including genre, highlights, and emotions. Therefore, instead of selecting a representative chunk from each music clip, including all chunks. Similarly, although the *Highlight* dataset was created using the same procedure as the *Genre3* dataset, each label indicates whether the corresponding chunk can represent the entire clip.
- **KOCCA40**: This dataset was designed to use from the undergraduate course on retrieving music information. To encourage students, 40 music clips were selected, which are found to have been easily learned by machine learning algorithms. Each music clip was labeled with one of four labels: passionate, cool, depressed, and peaceful.
- **MusicEmo-A and MusicEmo-B:** In *MusicEmo-A* dataset, 864 sound characteristics were analyzed from 100 music clips and were labeled approximately 500 times via the online annotation system. With 566 music clips from the *MusicEmo-B* dataset, 346 audio features were extracted and were labeled approximately 3,600 times. Each music clip was tagged with related tags, including excitement, distress, depression, and satisfaction. Previous versions of these two datasets were discussed in the previous study [16]. However, in this study, 21 errors in attribute 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 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. 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].

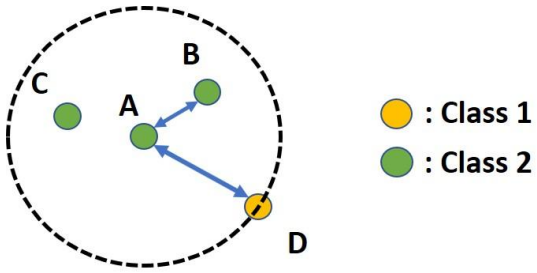

Figure 1: Example of local set on 2 dimensional space

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. Proposed Strategy

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.

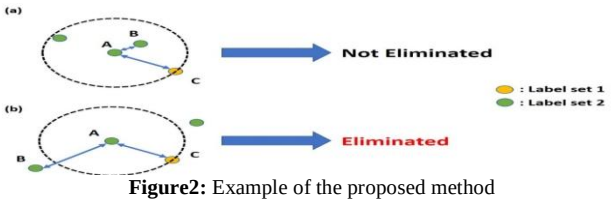

Figure2: Example of the proposed method

3. Results and Discussion

We experimented to determine the performance of the proposed method in ten multi-labeled music datasets. Eight datasets, excluding *CAL500* and *Emotions*, were collected through a national research project in Korea. These eight datasets consist of sound characteristics extracted by *MIR Toolbox* [19]. Sound characteristics are the result of song analysis regarding dynamics, fluctuations, rhythms, spectral, and tonal feature. These characteristics are extracted from a 40-second audio clip for each song. The other two datasets, *CAL500* and *Emotions* datasets, were created in a music tag annotation application where the music search system learns the relationship between words in a dataset on audio tracks with sound characteristics and annotations [15]. See Section 2 for a detailed description of datasets.

3.1. Characteristics of Datasets

Table 1 summarizes the characteristics of employed music datasets. $|W|$, $|F|$, and $|L|$ indicate the number of samples in the dataset, the number of extracted sound characteristics, and the number of labels respectively. The label cardinality *Card.* indicates the average number of assigned labels for each instance or music clip. The label density *Den.* indicates the label cardinality over the total number of labels $|L|$. The number of distinct label sets *Distinct.* indicates the number of label subsets ignoring duplication. *Subject* indicates the target information that the application is try to capture and deliver to the user.

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:

$$loss(T) = \frac{1}{|T|}\sum_{i=1}^{|T|}\frac{1}{|L|}\left|\lambda_i \vartriangle \hat{\lambda}_i\right| \quad (2)$$

where $\hat{\lambda}_i$ 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] Chen, S.-Y., Yu, Y., Da, Q., Tan, J., Huang, H.-K., & Tang, H.-H. (2018). Stabilizing reinforcement learning in dynamic environment with application to online recommendation. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, pp. 1187-1196.

[2] Choi, K., Fazekas, G., Cho, K., & Sandler, M. (2018). The Effects of Noisy Labels on Deep Convolutional Neural Networks for Music Tagging. IEEE Transactions on Emerging Topics in Computational Intelligence, 2(2), 139-149.

[3] Zhang, K., Zhang, H., Li, S., Yang, C., & Sun, L. (2018). The PMEmo Dataset for Music Emotion Recognition. In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, Yokohama, Japan, pp. 135-142.

[4] Murthy, Y.V. & Koolagudi, S.G. (2018). Content-Based Music Information Retrieval (CB-MIR) and Its Applications toward the Music Industry: A Review. ACM Computing Surveys, 51(3), Article No. 45.

[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. (2018). Scalable Multilabel Learning Based on Feature and Label Dimensionality Reduction, Complexity 2018(1), 1-15.

[7] Liu, J., Lin, Y., Li, Y., Weng, W., & Wu, S. (2018). Online multi-label streaming feature selection based on neighborhood rough set. Pattern Recognition, 84, 273-287.

[8] Lee J., Seo. W., Han. H & Kim, D.-W. (2018). Evolutionary Multilabel Feature Selection Using Promising Feature Subset Generation. Journal of Sensors, 2018(1), 1-12.

[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] Panda, R., Malheiro, R.M., & Paiva, R.P. (2018). Novel audio features for music emotion recognition. IEEE Transactions on Affective Computing, doi:10.1109/ TAFFC.2018.2820691.

[11] Lee J., Seo. W., & Kim, D.-W. (2018). Effective Evolutionary Multilabel Feature Selection under a Budget Constraint. Complexity, 2018(1), 1-14.

[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] Arnaiz-Gonzalez, A., Diez-Pastor, J.-F., Rodriguez, J.J., & Garcia-Osorio, C. (2018). Local sets for multi-label instance selection, Applied Soft Computing, 68(1), 651-666.

[14] Huang, S., Zhou, L., Liu, Z., Ni, S., & He, J. (2018). Empirical Research on a Fuzzy Model of Music Emotion Classification Based on Pleasure-Arousal Model, In Proceedings of 2018 37th Chinese Control Conference, Wuhan, China, pp. 3239-3244.

[15] Shao, X., Cheng, Z., & Kankanhalli, M.S. (2018). Music auto-tagging based on the unified latent semantic modeling. Multimedia Tools and Applications, doi:10. 1007/s11042-018-5632-2.

[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] Pereira, R.B., Plastino, A., Zadrozny, B., & Merschmann L.H. (2018). Categorizing feature selection methods for multi-label classification. Artificial Intelligence Review, 49(1), 57-78.

[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] Lange, E.B. & Frieler, K. (2018). T6G: Short Talks 6-Emotion Computing. In proceedings of the 15th International Conference on Music Perception and Cognition / 10th Triennial Conference of the European Society for the Cognitive Sciences of Music, Sydney, Australia, p. 35.