Multi-Label Annotation of Music

Hiba Ahsan
National Institute of Technology Karnataka
Surathkal, Mangalore, India
Email: hiba008@gmail.com

Vijay Kumar and C.V. Jawahar
International Institute of Information Technology
Hyderabad, India
Email: {vijaykumar.r@research., jawahar@}iiit.ac.in

Abstract—**Automatic annotation of an audio or a music piece with multiple labels helps in understanding the composition of a music. Such meta-level information can be very useful in applications such as music transcription, retrieval, organization and personalization. In this work, we formulate the problem of annotation as multi-label classification which is considerably different from that of a popular single (binary or multi-class) label classification. We employ both the nearest neighbour and max-margin (SVM) formulations for the automatic annotation. We consider K-NN and SVM that are adapted for multi-label classification using one-vs-rest strategy and a direct multi-label classification formulation using ML-KNN and M3L. In the case of music, often the signatures of the labels (e.g. instruments and vocal signatures) are fused in the features. We therefore propose a simple feature augmentation technique based on non-negative matrix factorization (NMF) with an intuition to decompose a music piece into its constituent components. We conducted our experiments on two data sets — Indian classical instruments dataset and Emotions dataset [1], and validate the methods.**

Keywords—*Music annotation, multi-label classification.*

I. INTRODUCTION

Rapid advancement in the Internet and multimedia technologies has enabled music listeners to access vast amount of music. People often listen to music from their personal collection or through online music services. Online music streaming services allow listeners to browse and play the music based on their interests such as genre, artist or album and provide them the ability to create their own play lists and share with others. Many cloud based services even allow people for uploading and sharing their own music. In all such scenarios, the organization of music collection becomes difficult as the size of database grows.

One simple way to organize such large collections is by annotating each music piece with different attributes related to composition of music such as style, genre, artist, instrument, mood/emotions, etc. However, manual annotation is a daunting task for large collection of music as they consume lot of time and incur cost. Thus, it is very important to develop automatic methods that perform these tasks.

Automatic annotation of music also helps in providing many additional services to online users such as searching music based on categories such as genre, emotions, instrument rather than primitive searches that rely purely on textual information such as song titles or artists. Music listeners and their demands have evolved over time. Online users, nowadays, want music suggestions that suits their mood or an activity like jogging, studying, traveling, etc. All such personalized services
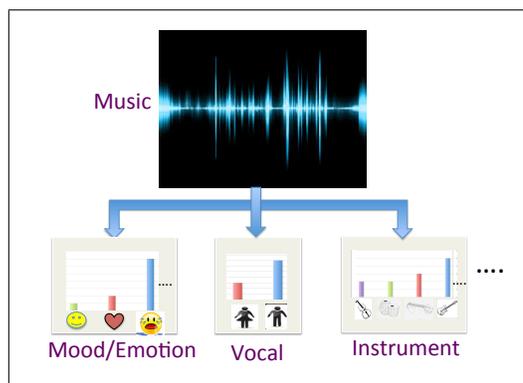
Fig. 1. An illustration of Music Annotation as Multi-label classification. Every segment of the music can have multiple labels.

require an understanding of the composition of music that can be addressed to some extent, through automatic annotation of music data.

In this paper, the task of music annotation is modeled as a multi-label classification problem i.e given a music piece, multiple tags are assigned to it. As shown in Fig 1, a music piece may belong to many different categories with each category specifying certain aspect of the given music. The problem of multi-label classification is significantly different from the popular single label cases. In the single label classification, each sample belongs to the positive set only for a single class. However, in the case of multi-label, each sample can be positive in multiple classes, and the problem becomes more challenging. In this work, we formulate the music annotation problem in nearest neighbour and max-margin frameworks. In particular, we evaluate the performance of Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Multi-Label KNN (ML-KNN) [2] and Max-Margin Multi-Label Classification (M3L) [3]. SVM and KNN are adapted for multi-label classification using one-vs-rest strategy while ML-KNN and M3L are the direct formulations of multi-label classification.

On a careful analysis of the performance of the classification methods, we find that the composite nature of the feature representation is a serious hurdle. We therefore propose a feature augmenting technique based on non-negative matrix factorization (NMF) with an intuition that decomposing a music piece into several components might help in classification. Our experiments on Indian classical instruments dataset and Emotions dataset [1] suggest that all the four algorithms have reasonably similar performance and are promising for multi-label annotation of music. We also noticed that the performance of all these algorithms is further improved by

a simple feature augmentation procedure based on NMF.

II. RELATED WORK

There are many approaches proposed in the literature of music classification. Some of these approaches are focused on single-label classification [4], [5] where a music piece is assigned a particular label such as male, female, guitar, flute, crowd, etc. For Indian Classical music, techniques are proposed in [6], [7], [8] to identify the *raga* being played in a music. In the context of multi-label classification, various techniques are proposed to identify the emotions in music. In [1], [9], authors survey various machine learning techniques for identifying the emotions in the music. Similarly, multiple binary SVM classifiers are trained to tag the emotions in music in [10]. The decisions are made based on whether each classifier output exceeds a pre-determined threshold.

Another closely related work is the use of social tags to annotate the music tracks [11]. A large amount of user *tagged keywords* and their tracks were extracted from commercial services such as *last.fm* and tag-track relationships are learned using latent semantic analysis (LSA), similar to term-document relationships in text retrieval. Once such a co-occurrence matrix is known, it can been applied to applications such as catalogue organization and to provide a browse-by-mood interface [11]. A similar use of social tags for recommendation and music similarity can be found in [12] where an ensemble based filterboost algorithm is used.

The above methods based on social tags focus only on social context while ignoring the acoustic content. Few methods have been proposed that combine the information from both acoustic content and social context. In [13] timbre and harmonic features are used to represent acoustic content while social tags and web documents represent social context. Similar combining approaches are used in [14] for multi-label music style classification and in [15] where a track's significant musical content or musword is considered along with social tags.

Multi-label classification is applied in other domains such as semantic annotation of images [16], [17] where a set of tags are assigned to an image in the context of scene understanding. It is widely applicable in the area of text mining as the textual document can usually belong to multiple categories (see [18] and references therein). Similarly in the area of Bio-informatics for protein function prediction [19] and gene function prediction [20].

III. MULTI-LABEL ANNOTATION OF MUSIC

In this section, we first formulate the problem of music annotation as multi-label classification. We then describe four machine learning techniques can be used for solving this.

A. Annotation as Multi-Label Classification

We start with a music composition and annotate the music as a whole and also in parts. For this, we extract features for different (overlapping) segments of various lengths. Our objective is then to classify these feature representations and assign multiple labels.

Given a training set of instance-label pairs (x_i, y_i), $i = 1, ..., m$, where $x_i \in \mathbb{R}^n$ and y_i is a vector of 0s and 1s and of dimensionality the number of possible labels. During the learning, we are interested in designing a classifier that can assign multiple labels to a given segment of music.

B. Multiple Single Label Classification: KNN and SVM

We first describe the multiple single label classification as a solution to the multi-label classification problem.

K-Nearest Neighbor Algorithm (KNN) is one of the simplest non-parametric methods that can be used for classification. When a new sample has to be classified, its k nearest neighbors in the training samples are identified and the class that is the most common among them is assigned to the new sample. For multi-label classification, one-vs-rest strategy can be used to assign multiple labels to the sample.

Support Vector Machine (SVM) is a popular parametric method that finds the hyperplane defined by parameter w with maximum-margin between two classes of the training data. Given a training set of instance-label pairs $(x_i, \bar{y}_i)$, $i = 1, ..., m$, where $x_i \in \mathbb{R}^n$ and $\bar{y}_i \in \{1, -1\}$, SVM solves the following optimization problem:

$$\min_{\mathbf{w}, \mathbf{b}, \xi} \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{m} \xi_i$$

$$\text{subject to } \bar{y}_i(\mathbf{w}^T x_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0, \ \forall i \qquad (1)$$

where C is the penalty parameter and b is the bias. ξ_i are the slack variables that denote the violations made by the training points when the data is non-separable. During testing, any new sample z is classified based on the sign of $(w^T z + b)$. Similar to K-NN, one-vs-rest strategy can be used for multi-label classification.

C. Multi-Label KNN

Multi-Label KNN (ML-KNN) is the multi-label version of KNN proposed by Zhang *et al.* [2]. Similar to KNN, k-neighbors of new instance are identified and a decision on whether the instance belongs to each category is made based on a maximum-a-posteriori (MAP) estimate. Formally, let L denote the complete set of labels. Given an instance $x_i \in \mathbb{R}^n$ and its associated labels $Y_i \subseteq L$, let y_i be a vector such that its l-th component $y_i(l)$ is 1 if $l \in Y_i$ and 0 otherwise. Then $y_i(l)$ can be obtained as,

$$y_i(l) = \arg \max_{b \in \{0,1\}} P(H_b^l | E_{\vec{C}(l)}^l), \forall l \in L \qquad (2)$$

where $\vec{C}(l)$ denote the number of neighbors of x out of its k-neighbors that belong to class l. H_1^l is the event when x has label l and H_0^l is the event when x does not have the label l. $E_j^l, j \in \{0, 1, ..., k\}$ denote the event that j out of k-neighbors of x have label l. $y_i(l)$ can be easily computed by expanding the Eqn 2 using Bayesian rule and estimating the probabilities $P(H_b^l)$ and $P(E_{\vec{C}(l)}^l | H_b^l)$ from the training data.

D. Max-Margin Multi-Label Classification

Max-Margin Multi-label Classification (M3L) [3] can be viewed as a generalization of SVM formulation for multi-label classification. M3L aims to learn a function f which assigns a set of labels to an instance x_i. Let Δ denote the loss function that imposes different penalties based on the amount of mismatch between expected and predicted output.

The learning function is defined as $f(x, y) = w^T(\phi(x) \otimes \psi(y))$ where ϕ and ψ denote the feature and label space mappings respectively and $\otimes$ is the Kronecker product. Given an instance $x_i \in \mathbb{R}^n$ and its associated labels $Y_i \subseteq L$, let y_i be a vector such that its l-th component $y_i(l)$ is 1 if $l \in Y_i$ and -1 otherwise. M3L is formulated as:

$$\min_f \frac{1}{2}||f||^2 + C\sum_{i=1}^n \xi_i$$
$$s.t. \ \ f(x_i, y_i) \geq f(x_i, y) + \Delta(y_i, y) - \xi_i$$
$$\forall i, \ y \in \{\pm 1\}^{|L|} \setminus y_i \ \ \text{and} \ \ \xi_i \geq 0, \ \forall i \quad (3)$$

where ξ_i are the slack variables as defined earlier and $|L|$ denote the number of labels. During testing, any new sample z is classified as $y^* = \arg\min_y f(z, y)$. The optimization details of the algorithm are beyond the scope of the paper and we refer interested readers to [3] for further details.

IV. Enhancement by Feature Augmentation

Music signals, in real world, are often noisy and composed from a mixture of sound sources. Decomposing the music into individual sources, popularly known as Source separation, often reveals many interesting details about the music. This gives a more compact and interpretable representation that could be helpful in many tasks such as pitch estimation [21], transcription [22], etc. This is usually achieved using matrix decomposition techniques such as non-negative matrix factorization (NMF).

In the case of multi-label music classification, the presence of one source (label) may act as a noise in predicting other source (label). Thus by having a low rank approximation of the music for each of these labels and projecting the music into each of these subspaces, we can hope to get a very compact and discriminative representation.

More formally, given a set of all possible labels L and instance-labels pairs (x_i, Y_i), $x_i \in \mathbb{R}^n$, $Y_i \subseteq L$. Let $M = [m_1, m_2, \ldots, m_j]$, $m_i \in \mathbb{R}^p$ be the set of features extracted from the training samples $[x_{l1}, x_{l2}, \ldots, x_{lj}]$ that contain the label l. We perform the matrix decomposition as follows,

$$\{\hat{W}_l, \hat{H}_l\} = \arg\min_{W_l, H_l} \frac{1}{2}||MM^T - W_l H_l||_F^2$$
$$s.t \ \ W_l \geq 0, H_l \geq 0 \quad (4)$$

where $W_l \in \mathbb{R}^{p \times q}$ and $H_l \in \mathbb{R}^{q \times p}$ are the non-negative matrices and $||.||_F$ denote the Frobenius norm. The final feature vector for instance x_i is obtained as $\phi(x_i) = [h_i^T \ m_i^T]^T$ where h_i is the i-th column in H_l. W_l can be viewed as a set of basis vectors that span the features belonging to label l. By learning separate basis vectors W_l for each of the labels and projecting the instances in each of these subspaces, we

Instrument	#Samples
Female Vocals	187
Flute	135
Harmonium	84
Male Vocals	76
Mridangam	301
Shehnai	90
Sitar	58
Tabla	168
Tanpura	156
Violin	275

get a representation which is discriminative and robust to interference from other labels. It is often difficult to know aprioiri the lower dimension q and thus is selected through cross validation.

V. Experiments

In this section, we describe the data sets used for experiments, evaluation measures and finally report the results.

A. Datasets

Indian Classical Music dataset consists of 18 Indian classical tracks. Each track is divided into samples of 5 seconds resulting in a total of 603 samples with a sampling rate of 22050 Hz. All the samples are manually annotated with multiple labels from a label set of 8 instruments, male and female vocals. The number of samples in each of these categories is shown in Table I.

Emotions dataset [1] consists of 100 songs belonging to the genres- Classical, Reggae, Rock, Pop, Hip-Hop, Techno and Jazz. For each song, a period of 30 seconds after the initial 30 seconds are extracted. These clips are then converted into wav files of 22050Hz sampling rate. All the clips are labeled into 6 emotional clusters as shown in Table II.

B. Features

For Indian Classical music dataset, we extracted standard MFCC features using MIRToolbox [23]. We considered only the top 19 MFCC coefficients. For the emotions dataset, we used the features provided by the authors of [1]. They extracted two kinds of features- rhythmic and timbre. 8 features are derived by extracting periodic changes from a beat histogram. 64 timbre features are derived by extracting the top 13 MFCC coefficients for each frame and obtaining their mean, standard deviation, mean standard deviation and standard deviation of standard deviation. Please refer to [1] for more details on feature extraction.

C. Evaluation Metric

We use the following evaluation measures for multi-label classification, similar to [18]. Let L denote the complete set of labels. Given a set of pairs (x_i, Y_i), $i = 1 \ldots m$, where $x_i \in \mathbb{R}^n$, x_i denotes the ith music sample and Y_i denote a subset of ground-truth labels ($Y_i \subseteq L$). Let Z_i be the set of labels predicted by multi-label classifier for the instance x_i,

TABLE II.   COMPOSITION OF EMOTIONS DATASET

Emotion-cluster	#Samples
amazed-surprised	173
happy-pleased	166
relaxing-calm	264
quiet-still	148
sad-lonely	168
angry-fearful	189

The Hamming Loss is defined as follows:

$$\text{Hamming Loss} = \frac{1}{m} \sum_{i=1}^{m} \frac{|Y_i \; \Delta_d \; Z_i|}{|L|} \tag{5}$$

where Δ_d represents the symmetric difference between two sets.

Accuracy is defined as follows:

$$\text{Accuracy} = \frac{1}{m} \sum_{i=1}^{m} I(Y_i == Z_i) \tag{6}$$

where $I(true) = 1$ and $I(false) = 0$.

Precision is defined as follows:

$$\text{Precision} = \frac{1}{m} \sum_{i=1}^{m} \frac{|Y_i \cap Z_i|}{|Z_i|} \tag{7}$$

Recall is defined as follows:

$$\text{Recall} = \frac{1}{m} \sum_{i=1}^{m} \frac{|Y_i \cap Z_i|}{|Y_i|} \tag{8}$$

D. Results

For all the experiments, we randomly divide half of the data set into training and testing. For reporting the results, we conducted 10 trials and report the average values. We select the parameters C for SVM, k for KNN, q for NMF using cross-validation. Table III and Table IV shows the average accuracy, hamming loss, precision and recall of the classification algorithms for Indian Classical music and Emotions dataset, respectively. Overall the performance of these algorithms is almost similar though KNN performs slightly better for Indian Classical music dataset while SVM for Emotions dataset. We also report the results of these algorithms with feature augmentation using NMF. It can be seen that, the performance of these algorithms is further improved with feature augmentation. Table V and Table VI shows the results of the algorithms for each label on Indian Classical Music dataset and Emotions dataset, respectively. Again, it can be observed that by feature augmentation, the performance is improved for majority of the labels.

VI.   CONCLUSION AND FUTURE WORK

In this paper, we considered the problem of automatic multi-label annotation of a music data. We evaluated the performance of four machine learning algorithms for this task. We further proposed a simple feature augmentation technique based on NMF which further improved the performance of all these algorithms. Future work will focus on incorporating the correlations among the labels in classification and on preparing a comprehensive multi-label classification dataset for Indian classical music.

REFERENCES

[1] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions." in *ISMIR*, vol. 8, 2008, pp. 325–330.

[2] M.-L. Zhang and Z.-H. Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[3] B. Hariharan, S. Vishwanathan, and M. Varma, "Efficient max-margin multi-label classification with applications to zero-shot learning," *Machine learning*, vol. 88, no. 1-2, pp. 127–155, 2012.

[4] T. Li, M. Ogihara, and Q. Li, "A comparative study on content-based music genre classification," in *ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 282–289.

[5] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *Neural Networks, IEEE Transactions on*, vol. 14, no. 1, pp. 209–215, 2003.

[6] K. Gopala-Krishna, S. Gulati, and P. Rao, "A survey of raaga recognition techniques and improvements to the state-of-the-art," in *Sound and Music Computing*, 2011.

[7] R. Sridhar and T. V. Geetha, "Raga identification of carnatic music for music information retrieval," in *International Journal of Recent Trends in Engineering*, 2009.

[8] V. Kumar, H. Pandya, and C. V. Jawahar, "Identifying ragas in indian music," in *International Conference on Pattern Recognition*, 2014.

[9] A. Wieczorkowska, P. Synak, and Z. W. Raś, "Extracting emotions from music data," in *Proceedings of ISMIS Symposium*, 2005.

[10] T. Li and M. Ogihara, "Detecting emotion in music." in *ISMIR*, vol. 3, 2003, pp. 239–240.

[11] M. Levy and M. Sandler, "A semantic space for music derived from social tags," *Austrian Compuer Society*, vol. 1, p. 12, 2007.

[12] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere, "Autotagger: A model for predicting social tags from acoustic features on large music databases," *Journal of New Music Research*, vol. 37, no. 2, pp. 115–135, 2008.

[13] D. R. Turnbull, L. Barrington, G. Lanckriet, and M. Yazdani, "Combining audio content and social context for semantic music discovery," in *ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 387–394.

[14] F. Wang, X. Wang, B. Shao, T. Li, and M. Ogihara, "Tag integrated multi-label music style classification with hypergraph." in *ISMIR*, 2009, pp. 363–368.

[15] M. Levy and M. Sandler, "Music information retrieval using social tags and audio," *Multimedia, IEEE Transactions on*, vol. 11, no. 3, pp. 383–395, 2009.

[16] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[17] Y. Verma and C. V. Jawahar, "Image annotation using metric learning in semantic neighbourhoods," in *ECCV*, 2012, pp. 836–849.

[18] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data mining and knowledge discovery handbook*. Springer, 2010, pp. 667–685.

[19] G. Yu, H. Rangwala, C. Domeniconi, G. Zhang, and Z. Yu, "Protein function prediction using multi-label ensemble classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 10, no. 4, pp. 1–1, 2013.

[20] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya, "Hierarchical multi-label prediction of gene function," *Bioinformatics*, vol. 22, no. 7, pp. 830–836, 2006.

[21] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[22] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[23] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *International Conference on Digital Audio Effects*, 2007, pp. 237–244.

TABLE III.   PERFORMANCE OF VARIOUS ALGORITHMS ON THE TWO MUSIC ANNOTATION DATASETS

Method ↓	Indian Classical Music dataset			
	HammingLoss	Accuracy	Precision	Recall
One-vs-all kNN	0.0748	0.9252	0.8545	0.8506
One-vs-all SVM	0.1139	0.8860	0.8072	0.7250
ML-kNN	0.0833	0.9167	0.8429	0.8255
M3L	0.1119	0.8880	0.8037	0.7362
NMF-SVM	0.1082	0.8917	0.8159	0.7415
NMF-kNN	**0.0743**	**0.9257**	**0.8548**	**0.8524**
NMF-ML-kNN	0.0814	0.9186	0.8449	0.8313
NMF-M3L	0.1093	0.8907	0.8042	0.7493

TABLE IV.   PERFORMANCE OF VARIOUS ALGORITHMS ON EMOTION DATASET

Method ↓	Emotions dataset			
	HammingLoss	Accuracy	Precision	Recall
One-vs-all kNN	0.3202	0.6798	0.4783	0.2962
One-vs-all SVM	0.3074	0.6926	0.6445	0.0494
ML-kNN	0.0833	0.6859	0.5136	0.1347
M3L	0.3126	0.6874	0.7454	0.0101
NMF-SVM	**0.2667**	**0.7333**	0.6827	0.2544
NMF-kNN	0.2893	0.7106	0.5541	**0.3710**
NMF-ML-kNN	0.2886	0.7113	0.5292	0.2436
NMF-M3L	0.3075	0.6925	**0.8000**	0.0215

TABLE V.   PERFORMANCE OF VARIOUS ALGORITHMS ON EACH LABEL FOR INDIAN CLASSICAL MUSIC DATASET

	One-vs-all SVM	One-vs-all KNN	ML-KNN	M3L	NMF-SVM	NMF-KNN	NMF-ML-KNN	NMF-M3L
Female	86.7550	**92.6159**	90.9272	86.0927	87.8808	92.4172	91.3576	87.0861
Flute	96.9536	96.7550	97.9139	97.6821	98.0132	96.7219	**98.1126**	98.0132
Harmonium	98.9735	**99.6689**	99.1391	99.3377	99.0066	99.6358	**99.2384**	99.6689
Male	96.1589	95.0000	95.8609	**97.3510**	96.3576	95.4636	95.9272	97.0199
Mridangam	72.4503	83.6093	82.9139	69.5364	73.8411	**83.8411**	83.2450	70.5298
Shehnai	91.9868	94.2384	**94.9669**	93.3775	93.3113	94.2715	94.7682	93.3775
Sitar	93.5762	95.0662	92.5497	90.3974	93.4437	**95.4967**	93.0795	90.3974
Tabla	91.6225	95.7285	94.4371	92.7152	91.5232	**95.8940**	94.6689	92.7152
Tanpura	79.3377	**86.8874**	84.3709	81.4570	79.3709	86.3245	84.4702	81.7881
Violin	78.2450	**85.6623**	83.6093	80.1325	79.0066	**85.6623**	83.7086	80.1325

TABLE VI.   PERFORMANCE OF VARIOUS ALGORITHMS ON EACH LABEL FOR EMOTIONS DATASET

	One-vs-all SVM	One-vs-all KNN	ML-KNN	M3L	NMF-SVM	NMF-KNN	NMF-ML-KNN	NMF-M3L
Label 1	70.7071	69.3266	69.5623	71.0438	**74.0067**	73.0640	71.6162	71.0438
Label 2	72.0539	67.5421	70.7071	**73.0640**	71.8855	70.2020	70.0673	**73.0640**
Label 3	58.1145	60.2357	56.1279	54.8822	**67.6768**	63.6700	63.8384	57.9125
Label 4	74.9832	74.5118	75.0842	75.0842	**79.8316**	77.8114	78.4512	75.0842
Label 5	71.7508	68.1818	71.2121	69.3603	**74.0741**	71.0101	72.9293	69.3603
Label 6	67.9798	68.0808	68.8552	69.0236	**72.5253**	70.6397	69.9327	69.0236