Journal of Machine Intelligence 1 (1): 7-12, 2016
sispress.org/journals/jomi
ISSN: 2377-2220
http://dx.doi.org/10.21174/jomi.v1i1.23

Sanford Inter Science Press

Ordered Classifier Chains for Multi-label Classification

Maryam Keikha*, and Sattar Hashemi

Shiraz University, Shiraz, Iran

Abstract— Classifier chains method is introduced recently in multi-label classification scope as a high predictive performance technique aims to exploit label dependencies and in the meantime preserving the computational complexity in a desirable level. In this paper, we present a method for chain's order, called Ordered Classifier Chains (OCC), elaborating that the sequence of labels in the chain plays an important role in predictive performance of corresponding multi-label classifiers. OCC proposes making use of correlation of every class label with that of features. OCC renders an ordering of class labels in their descending order. Once the ordering of labels is determined, the features along with every label are fed to binary classifier. In the classifier chain model the feature space of every binary classifier is extended with the new order of labels. In order to specify association of each sample with the set of class labels, it is given to all of classifiers. Empirical evaluations include an extensive range of multi-label datasets reveal that OCC manages to improve the classification performance compared to existing approaches.

Keywords— Multi-Label Classification; Label Dependency; Classifier Chains; Binary Relevance.

I. Introduction

As opposed to traditional single-label classification that each instance is only associated with a single class label, in multi-label classification the instances are associated with a set of labels. In recent years, the topic of multi-label classification is receiving increased attention, inspired from various range of new applications, including image and video classification [1], text classification [2], and music categorization into emotions [3]. In image classification, for example, models are sought which assign a subset of objects to every image.

Binary relevance (BR) method is the simplest and the most popular technique for dealing with multi-label learning. The BR technique breaks a given multi-label problem with m class labels to |m| binary classification problems. Accordingly, this method trains |m| binary classifiers where every classifier is responsible for predicting the relevance of one class label. The last set of predicted class labels are settled by collecting the classification predictions from all binary classifiers. In spite of linear complexity of BR, its weak point is its assumption of ignorance of label interdependencies which causes information loss.

A fundamental challenge of multi-label learning which is considered in a wide range of researches in recent times is the ability of detecting label correlations that exist in the training data. In fact, these interdependencies among labels are the knowledge which exists in the data and illustrates unique characteristic of every instance.

The label independence assumption of BR approach was introduced in [4]. The authors propose classifier chains method to overcome BR's issue. The classifier chains technique involves |m| binary classifiers as in BR but the difference of these two techniques is the attribute space for every binary classifier. According to the proposed method the attribute space of every binary classifier is extended with the true class labels of all previous labels. The classifier chains method has been shown to improve classification accuracy over the BR method by mentioning label correlations.

The sequence of class labels affects the predictive performance. Ensemble of classifier chains [5] method uses an ensemble of chains; each member of the ensemble applies a random label order to solve chain's selection issue. Moreover, Ensemble of classifier chains as an ensemble method increases overall

* Corresponding author can be contacted via the journal website.

predictive performance. A notable drawback of Ensemble of classifier chains is its high computational time.

In this research, we propose ordered classifier chains (OCC) model to discover a suitable sequence of class labels which improves the predictive performance of classifier chains technique while there is no necessity of high computational time of Ensemble of classifier chains model. In other words, the motivation of this paper is to enhance the classifier chains method's performance through handling chain's order. Our empirical observations over several multi-label benchmarks indicate the importance of chains order on the classifier chains' performance. Moreover, the problem of error propagation down to the chain is a challenging problem for classifier chains method. Our proposed method tries to overcome this problem by selecting the most dependent class labels' order in the first step of chain's generation. The sequence of class labels is exploited with the use of each class label's dependency with the attribute space. This approach is expected to overcome not only the BR's assumption of ignoring the label correlations like classifier chains, but also it solves the challenge of selecting an order of class labels in classifier chains method. The proposed approach improves the accuracy of multi-label classification, as it tries to find the labels interdependencies while considering the dependency of each class label with the set of attribute space. Our empirical results are very promising.

The rest of the paper is organized as follows: the next section presents related work on multi-label classification. Section 3 discusses the motivations and rational of using OCC method. Section 4 produces the experimental settings and the results of our researches. Finally, Section 5 concludes the paper by the conclusion part and presents the future work.

II. Related Work

Multi-label classification approaches can be categorized into two main types [6]. The first group is called algorithm adaptation methods which refer to the kinds of techniques that extend specific learning algorithms in order to handle multi-label data directly. Well-known methods of this group include AdaBoost [7], decision trees [8], and lazy methods [9]. The second group which is in the scope of this paper is called problem transformation methods. These techniques transform the multi-label classification problem into one or more single-label classification problems. In the following paragraphs several problem transformation methods in the literature are presented.

Binary relevance (BR) [6] is a problem transformation method that learns |m| binary classifiers for the training data with m class labels. It transforms the original dataset into m data sets. Every dataset is created from the whole feature space plus one class label. For the classification of a new instance, BR aggregates the results of all classifiers. Despite of its simplicity, BR gained competitive performance among base models of multi-label classification. BR does not take into account the label correlations and it is the challenging disadvantage of this technique.

Classifier chains (CC) method [4] is proposed as a technique that considers label dependencies inspired from BR method. The common point of CC and BR is the number of their classifiers. As mentioned above, the difference between CC and BR is the feature space of classifiers. Despite of BR that considers the original feature space of dataset as an attribute space of each classifier, CC tries to model interdependencies between class labels by considering class labels as attribute space. Therefore, a chain $h = (h_1, \ldots, h_2)$ of binary classifiers is formed. Each classifier h_m in the chain is responsible for learning and predicting the binary association of the m*th* label given the attribute space, augmented by all prior binary relevance predictions in the chain [5]. Ensemble classifier chains (ECC) method [4] is proposed for covering the random arrangement of chain in CC method. It is employed by a random chain order over a set of permutations. It provides higher predictive performance than CC but it increases the time complexity of problem.

Label power set (LP) is a kind problem transformation method. According to this approach, each unique set of labels that exists in the training data is considered as a single label. In despite of BR, it has the advantage of taking label correlations into consideration. LP is suffered from sparseness, this means that there are too many classes with a few number of instances associated with them.

Pruned Sets (PS) method [10, 11] was introduced to overcome the sparseness problem of LP by pruning away instances with infrequently occurring label sets. To prevent loss of data, some of the pruned instances are reintroduced into the data by decomposing them into more frequently occurring label subsets. This method keeps the advantage of considering label correlations of LP method while trying to omit the label sets with low number of associated instances for decreasing the sparseness of LP. PS seems not to be effective in datasets with a large proportion of unique label combinations while the method is proposing to omit rare label sets so the rate of information loss will increase in such a situation.

III. Proposed Method: Ordered Classifier Chain

Ordered classifier chains (OCC) method is presented a supplementary phase to the classifier chains (CC) method. The two main disadvantages of CC method are the order of class labels which is mentioned in [12] and the error propagation down the

chain. Error propagation occurs when one (or more) of the first classifiers down the chain predicts poorly. In order to the error propagation problem, CC method is under question. OCC proposes not only a mechanism for selecting a sequence of class labels, but it also prevents the propagation of noise down to the chain.

A. Preliminaries

Before describing OCC approach, we present the learning task of multi-label classification with formal notations. A vector of X = {x_i: i = 1,... , N} denotes an instance space with |N| attribute values. Let L = {l_j: j = 1,... ,M} be a finite set of |M| class labels for every instance. Each instance X is associated with a subset of label $L \in 2^L$. The set of multi-label training instances can be presented as D = {(X_k, L_k), k = 1,... ,K} where there are |K| instances. X_{ki} represents the value of the *k*th example pertaining to the *i*th attribute. L_{kj} represents the binary relevance of the *k*th example pertaining to the *j*th label. The aim is to induce from D a classifier h that correctly predicts the subset of relevant labels for unlabeled query instances $\hat{X}$.

In the BR method a chain of |M| classifiers h = (h_1, . . ., h_m) is induced. Each of the classifiers uses only the original feature space, in which h_m: $X \rightarrow \{0, 1\}$. So the prediction of a test instance $\hat{X}$ is of the form: $\hat{X}$ = [h_1(x), h_2(x),..., h_M(x)].

In the CC method a chain of |M| classifiers is induced for each class label too, but the feature space is different in the form of h_m: $X \times \{0, 1\}^{(m-1)} \rightarrow \{0, 1\}$, where the previous class labels are considered as the feature space. The prediction for a test instance {X} is of the form: $\hat{X}$ = [h_1(x), h_2(x,h_1(x)),..., h_M(x,h_1(x),h_2(x, h_1(x)), ..., h_{M-1}(x,..., h_{M-2}(x)))].

B. OCC Algorithm

As a single chain should in principle be enough to capture dependencies between all labels, OCC tries to find just a sequence of class labels. Moreover, OCC prevents redundant iterative chain selection of ECC. In other words, OCC tries to find the optimal tradeoff between CC and ECC methods. OCC considers label dependencies, and then it models and learns such dependencies in a non-redundant way.

Formally, the OCC method works as follows in three phases: (i) Finding chain (ii) Training (iii) Testing. OCC is focused on the first step and the two other phases are exactly like the CC method.

First, for finding a sequence of labels, we consider the problem as a single label classification task. In order to achieve this purpose, we put every class label beside the feature space, modeling a classifier with the new dataset. In this case we will create |M| classifiers as many as class labels in the form of: $\hat{X}$ = [h_1(x), h_2(x), ..., h_M(x)].

In this step we will predict every class label with the use of prepared classifiers. Every classifier is responsible for prediction of one class label. After prediction step, the accuracy score of every classifier is evaluated. The accuracy formula is in the form of single-label classification. Accuracy of a dataset with |D| instances, *L* true label sets and *E* predicted label sets, is in the form of:

$$Accuracy = \frac{1}{|D|}\sum_{i=1}^{|D|} |L_i = E_i| \qquad (1)$$

After evaluating of all class label's accuracies, the new order of chain is acquired, according to the descending order of class label's accuracies. Finding chain procedure is outlined in Algorithm 1. If we present the label with the highest accuracy by the notation of L^1, then the new order of |M| class labels can be seen as: L = { L^i_j: i = 1, ..., M}.

Algorithm 1: OCC's finding chain phase for training set D and label set L

$D = \{(X_k, L_k), k = 1,..., K\}$

$L = \{l_j : j = 1,..., M\}$

for $j = 1,..., M$ *do*

$D' = \{\}$

$D' \leftarrow D' \cup (Y, l_j)$

$hj : D' \rightarrow l_j \in \{0,1\}$

$\hat{l}_j \leftarrow h_j(\hat{X})$

〉 A is an array for accuracies

$A(1, j) = Accuracy(h_j)$

end for

Rank (A) descendingly

In other words, we want to find the dependency of every class label with the original feature space. Class labels which have the higher accuracy value are nominated to be more related to the feature space under the condition of using the same classifier type. In fact, we try to model a function to indicate the relation of feature space *X* with the class label *L* as: F: $X \rightarrow L$.

Moreover, the descending order class labels helps to prevent the propagation of noise down to the chain whereas with merging the more dependent class labels in the first levels of chain, we make a more coherent feature space for future classifiers. Therefore, the classifiers that are created up the chain are more consistent. After finding the new chain's order, CC's training and testing phases will run [5].

IV. EXPERIMENTS

In this section we provide details on the collection of multi-label benchmark datasets and the evaluation measures that are used to empirically evaluate the performance of the proposed method.

A. Datasets

The experiments were performed over ten multi-label datasets from a variety of domains. The properties of these benchmarks are provided in Table 1.

Table 1. Properties of ten benchmark datasets together with their characteristics, including, number of labels, number of features, dataset size, domain, and cardinality.

Dataset	No. of labels	No. of features	Dataset size	Domain	Cardinality
Emotion	6	72	593	Music	1.87
Scene	6	294	2407	Image	1.07
Flags	7	19	194	Image	3.33
Yeast	14	103	2417	Biology	4.24
Birds	19	260	645	Audio	1.01
Slashdot	22	1079	3782	Text	1.18
Genbase	27	1185	662	Biology	1.25
Medical	45	1449	978	Text	1.25
Enron	53	1001	1702	Text	3.38
Langlog	75	1004	1460	Text	1.18

We use a variety of different domain datasets to help demonstrate the scalability of the algorithms. All datasets and further information about them can be found online[1]. Properties of these datasets including number of labels, number of features, dataset size, domain, and Label Cardinality (LC) are presented in Table 1. The datasets are ordered by their label set size. Label cardinality of a dataset with |D| instances is the average number of labels per instances as:

$$LC = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=1}^{M} L_{Dj} \qquad (2)$$

[1] http://meka.sourceforge.net\#datasets, http://mlkd.csd.auth.gr/multilabel.html\#Datasets

Label cardinality is a standard measure of multi-labeled-ness [10].

B. Evaluation Measures

We compare the algorithm using three evaluation measures adopted for multi-label classification. The evaluation of methods that are learnt for multi-label data requires different measures than those used in the case of single-label data. Several measures have been proposed in the past for the evaluation of multi-label classifiers that are categorized into label-based evaluations or example-based evaluations [13]. We used two measures of Hamming Loss and Accuracy as example-based evaluations and F-measure macro averaged measure as label-based evaluation.

For a dataset with $|D| = \{(Y_k, L_k), k = 1, ..., K\}$ training instances, let *h* be a multi-label classifier and $E_K = h(Y_k)$ be the set of labels predicted by *h* for instance Y_k. Also p_k and r_k are the precision and recall for the k*th* label. Hamming Loss is defined as:

$$\text{Hamming Loss} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{L_i \Delta E_i}{|L|} \qquad (3)$$

Where Δ stands for the symmetric difference of two sets, which is the set-theoretic equivalent of the XOR operation in Boolean logic. Accuracy is defined as:

$$\text{Accuracy} = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|L_i \wedge E_i|}{|L_i \vee E_i|} \qquad (4)$$

F-measure macro averaged is defined as:

$$\text{F - measure macro} = \frac{1}{|L|}\sum_{i=1}^{|D|}\frac{|2 \times P_i \times r_i|}{|P_i + r_i|} \qquad (5)$$

C. Experimental Settings

In order to evaluate the performance of BR, CC and OCC methods, we use SMO which is implemented in Weka [14] with default parameters as the base classifier.

To evaluate the statistically significant differences of the results from the comparative methods; we conduct the Friedman test [15] with the corresponding post-hoc test which was recommended in [16]. Friedman test ranks all the algorithms for every dataset. Then it compares the average ranks of all techniques. If the null hypothesis, i.e., all techniques are equivalent, is rejected then we proceed with the post-hoc test. In this step, we want to understand which techniques differ. The Bonferroni-Dunn test [17] is used as the post-hoc test method. OCC method is

selected as the control method because we want to prove its significant difference with other techniques. The CD value, which stands for critical difference, is determined by the number of competing algorithms and data sets. If the difference between two average ranks is greater than the CD, the performance of these two algorithms is significantly different. We used F_F test that is based Friedman's χ_F^2 statistic. The values $q_{0.1} = 1.96$ and CD = 0.87 are considered.

We compare BR, CC, and OCC methods under the various evaluation measures from Subsection evaluation measures. We carry out 5×2 fold cross validation (CV) on all datasets and use the average result to display and test for significance. The reason that we did not compare OCC method with other methods in the literature including LP and PS is that OCC is inspired form CC. The OCC model adds a complementary phase to the ordinary CC algorithm. Also, CC method is an extension of BR technique. So we compared our proposed method only with these two Techniques. Moreover, the results of comparing CC with other related methods are provided widely in [5].

V. Results

We compare OCC against the BR, and CC methods. Results for hamming loss, accuracy, and f-measure are shown in Table 2, Table 3, and Table 4, respectively.

In most datasets, OCC outperforms BR and CC in all evaluation measures. OCC is the best method for accuracy to have best accuracy among 9 datasets out of 10 followed by CC. In terms of hamming loss, OCC has the best result to be the first rank of 8 datasets out of 10; moreover it is presented that in those two datasets OCC is in the second rank. From the f-measure point of view expect of Yeast, Enron, and LangLog datasets, OCC outperforms other methods.

In all cases there are critical differences, so we reject the null hypothesis and proceed with the post-hoc. OCC $\succ$ {BR, CC} indicates that OCC has a significant difference over BR and CC techniques for all evaluation measures.

VI. Conclusion and Future Work

This paper presented a mechanism for selecting a suitable class labels order in classifier chain method. This approach is inspired from the binary relevance method. By using the dependency of feature space and each class label, our method counteracts the disadvantage of the classifier chain method and obtains high predictive performance.

Table 2. The results of our comparative methods over ten datasets including BR, CC, and OCC in terms of hamming loss evaluation measure. Statistically significant difference is highlighted in **boldface**.

Dataset	BR(%)	Rank	CC(%)	Rank	OCC(%)	Rank
Emotion	20.15 ± 0.4	1	21.77 ± 0.5	3	20.69 ± 1	2
Scene	11.07 ± 0.3	3	11 ± 0.3	2	10.43 ± 0.4	1
Flags	28.39 ± 1.1	2	28.4 ± 1.4	3	28.3 ± 1.8	1
Yeast	20.17 ± 0.3	1	21.36 ± 0.2	3	21.04 ± 0.2	2
Birds	5.48 ± 0.3	2	5.58 ± 0.3	3	5.4 ± 0.3	1
Slashdot	2.03 ± 0.05	3	2.02 ± 0.04	2	2.01 ± 0.05	1
Genbase	0.13 ± 0.07	2	0.15 ± 0.07	3	0.1 ± 0.04	1
Medical	1.91 ± 0.4	3	1.78 ± 0.6	2	1.16 ± 0.1	1
Enron	6.14 ± 0.3	3	6 ± 0.1	2	5.85 ± 0.1	1
Langlog	19.67 ± 0.2	3	19.52 ± 0.3	2	19.46 ± 0.3	1
Avg. rank	2.3	2	2.5	3	**1.2**	**1**

CD=0.876; Significance: OCC $\succ$ {BR,CC}

Table 3. The results of our comparative methods over ten datasets including BR, CC, and OCC in terms of accuracy evaluation measure. Statistically significant difference is highlighted in **boldface**.

Dataset	BR(%)	Rank	CC(%)	Rank	OCC(%)	Rank
Emotion	51.27 ± 2.06	3	52.13 ± 2.2	2	56.5 ± 2.8	1
Scene	58.55 ± 1.3	3	68 ± 1.2	2	69.35 ± 1.3	1
Flags	56.5 ± 1.4	3	56.58 ± 1.3	2	56.9 ± 2.3	1
Yeast	49.8 ± 0.8	2	48.02 ± 1	3	52.38 ± 0.9	1
Birds	15 ± 1.5	3	15.87 ± 1.06	2	16.17 ± 0.9	1
Slashdot	52.09 ± 0.3	3	52.41 ± 0.5	2	52.53 ± 1.6	1
Genbase	98.6 ± 0.6	2	98.5 ± 0.5	3	98.7 ± 0.5	1
Medical	62.5 ± 3.5	3	64.9 ± 6	2	73.44 ± 1.3	1
Enron	38.26 ± 0.01	3	38.54 ± 1.2	2	40.18 ± 1.1	1
Langlog	37.46 ± 0.7	2	37.63 ± 0.7	1	37.36 ± 1.03	3
Avg. rank	2.7	3	2.1	2	**1.2**	**1**

CD = 0.876; Significance: OCC $\succ$ {BR,CC}

Table 4. The results of our comparative methods over ten datasets including BR, CC, and OCC in terms of f-measure evaluation measure. Statistically significant difference is highlighted in **boldface.**

Dataset	BR(%)	Rank	CC(%)	Rank	OCC(%)	Rank
Emotion	60.74 ± 1.86	2	59.87 ± 2	3	66.36 ± 2.5	1
Scene	67.92 ± 1.2	3	70.12 ± 1	2	71.6 ± 1.1	1
Flags	59.29 ± 4.1	2	58.68 ± 1.9	3	59.5 ± 3.4	1
Yeast	32.68 ± 0.7	3	36.39 ± 1	1	34.13 ± 1.2	2
Birds	29.06 ± 2.6	3	30.8 ± 2.6	2	31.67 ± 2.6	1
Slashdot	10.62 ± 1.1	3	11.44 ± 1.3	2	11.78 ± 1.7	1
Genbase	79.44 ± 3.5	2	77.92 ± 5	3	80.83 ± 3.8	1
Medical	36.5 ± 0.7	3	37.3 ± 1.6	2	37.43 ± 1.6	1
Enron	21.03 ± 1.83	1	20.09 ± 1.2	3	20.34 ± 1.6	2
Langlog	38.41 ± 0.6	1	37.96 ± 1.3	2	37.91 ± 0.6	3
Avg. rank	2.3	3	2.3	2	**1.4**	**1**

CD = 0.876; significance: OCC ≻ {BR,CC}

In an empirical evaluation, we compared our ordered classifier chains method against BR and CC methods using a variety of multi-label datasets. The experiments indicated that the ordered classifier chains (OCC) model performs better than classifier chains (CC). This is due to the fact that the OCC method does not select chain's order randomly.

As future work, we plan to further investigate how to select better order of chain. We use the SMO model as the base learner in the first phase of chain selection; in the further research we can employ other base learners to approximate the dependency of each label with the feature space. Our proposed OCC method can be used as an ensemble technique too.

REFERENCES

[1] Dimou, Anastasios, Grigorios Tsoumakas, Vasileios Mezaris, Ioannis Kompatsiaris, and L. Vlahavas. "An empirical study of multi-label learning methods for video annotation." In Content-Based Multimedia Indexing, 2009. CBMI'09. Seventh International Workshop on, pp. 19-24. IEEE, 2009.

[2] Fürnkranz, Johannes, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. "Multilabel classification via calibrated label ranking." Machine learning 73, no. 2 (2008): 133-153.

[3] Wieczorkowska, Alicja, Piotr Synak, and Zbigniew W. Raś. "Multi-label classification of emotions in music." In Intelligent Information Processing and Web Mining, pp. 307-315. Springer Berlin Heidelberg, 2006.

[4] Read, Jesse, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. "Classifier chains for multi-label classification." In: ECML'09: 20th European Conference on Machine Learning, Springer (2009): 254–269.

[5] Read, Jesse, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. "Classifier chains for multi-label classification." Machine learning 85, no. 3 (2011): 333–359.

[6] Tsoumakas, Grigorios, and Ioannis Katakis. "Multi-label classification: An overview." Dept. of Informatics, Aristotle University of Thessaloniki, Greece (2006).

[7] Schapire, Robert E., and Yoram Singer. "BoosTexter: A boosting-based system for text categorization." Machine learning 39, no. 2-3 (2000): 135-168.Vens, C., Struyf, J., Schietgat, L., Dˇzeroski, S., Blockeel, H.: Decision trees for hierarchical multi-label classification. Machine Learning 2(73), 185–214 (2008).

[8] Vens, Celine, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. "Decision trees for hierarchical multi-label classification." Machine Learning 73, no. 2 (2008): 185-214.

[9] Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." Pattern recognition 40, no. 7 (2007): 2038-2048.

[10] Read, Jesse, Bernhard Pfahringer, and Geoffrey Holmes. "Multi-label classification using ensembles of pruned sets." In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, pp. 995-1000. IEEE, 2008.

[11] Tsoumakas, Grigorios, Anastasios Dimou, Eleftherios Spyromitros, Vasileios Mezaris, Ioannis Kompatsiaris, and Ioannis Vlahavas. "Correlation-based pruning of stacked binary relevance models for multi-label learning." In Proceedings of the 1st International Workshop on Learning from Multi-Label Data, pp. 101-116. 2009.

[12] Cheng, Weiwei, and Eyke Hüllermeier. "Combining instance-based learning and logistic regression for multilabel classification." Machine Learning 76, no. 2-3 (2009): 211-225.Tsoumakas, G., Vlahavas, I.P.: Random k-labelsets: An ensemble method for multilabel classification. In: ECML '07: 18th European Conference on Machine Learning, pp. 406–417. Springer (2007).

[13] Trohidis, Konstantinos, Grigorios Tsoumakas, George Kalliris, and Ioannis P. Vlahavas. "Multi-Label Classification of Music into Emotions." In ISMIR, vol. 8, pp. 325-330. 2008.

[14] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.Friedman, M. :The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32(200) (1973).

[15] Friedman, Milton. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance." Journal of the American Statistical Association 32, no. 200 (1937): 675-701.

[16] Demšar, Janez. "Statistical comparisons of classifiers over multiple data sets." The Journal of Machine Learning Research 7 (2006): 1-30.

[17] Dunn, Olive Jean. "Multiple comparisons among means" Journal of the American Statistical Association 56, no. 293 (1961): 52-64.