The Importance of Cross Database Evaluation in Sound Classification

Arie Livshin livshin@ircam.fr
Xavier Rodet rod@ircam.fr
IRCAM Centre Pompidou, 1 place Igor-Stravinsky, 75004 Paris, France

Abstract

In numerous articles (Martin and Kim, 1998; Fraser and Fujinaga, 1999; and many others) sound classification algorithms are evaluated using "self classification" - the learning and test groups are randomly selected out of the same sound database. We will show that "self classification" is not necessarily a good statistic for the ability of a classification algorithm to learn, generalize or classify well. We introduce the alternative "Minus-1 DB" evaluation method and demonstrate that it does not have the shortcomings of "self classification".

1 Testing Platform

The importance of cross database evaluation will be demonstrated through a variety of classification experiments.

1.1 The Test Set

The Sounds. In order to demonstrate well the claims in the paper, we extracted out of 5 sound databases, recorded in various acoustic conditions and different equipment, the samples of 7 instruments common to them, played with a "standard" playing technique. The instruments are: Bassoon, Contrabass, Clarinet, French horn, Flute, Oboe, and Cello. The number of samples extracted out of each database, is: Ircam Studio Online (SOL) - 581, University of Iowa Musical Instrument Samples (IOWA) - 1289, McGill University Master Samples (McGill) - 85, Pro collection - 158, Vi collection - 249. The samples are remixed in mono, 44.1Khz, 16bit and clipped to 2 seconds.

The Feature Descriptors. We use 162 different sound descriptors (Peeters, 2002), normalized to the range 0 - 1.

1.2 The Classification Algorithms

"LDA+KNN". The classified data goes through Linear Discriminant Analysis (McLachlan, 1992) then classified using the K-nearest neighbors algorithm. Instead of selecting constant values for K, we estimate the best K by using the Leave-One-Out Cross Validation method on the learning set with K's in the range of 1 - 20.

"BP80". A back propagation neural network with a single hidden layer of 80 neurons, using "tansig" functions in all the layers, is trained using Conjugate Gradient with Powell/Beale Restarts until a Mean Square Error of 0.004 is reached.

1.3 Evaluation methods

"Self Classification". A single database is used for evaluation of the classification process. The training set consists of 2/3 of the samples from each instrument class, which are randomly selected. To minimize the fluctuations of this random selection, each result shown is the mean of 20/50 tests (depending on the classification algorithm), so the 95% confidence interval is not more than 1% around the mean.

"Mutual Classification". A single complete database is used to classify another single and complete database.

"Minus-1 DB". Several databases are used, each one classified by the rest joined together.

2 Disadvantages of Self Classification

2.1 Claim 1: evaluation using Self Classification is not necessarily a good measure for the generalization abilities of the classification process

In this section we demonstrate several points:

1. Evaluation results using a single database are not necessarily an indication of the generalization abilities of the classification process and its suitability for practical applications of sound classification of musical instruments.

2. Self Classification results do not reflect the classifier ability, after learning the specific database, to deal with new sounds, and thus its performance as a Concept Classifier[1].

3. We shall demonstrate the intuitive claim that enriching the learning database with diverse samples from other databases improves the generalization power of the classifier and makes it more suited for classification of new sounds.

Table 1 mostly consists of the classification success percentage (recognition rate) of classifying each database by every other one. The Minus-1 DB column shows the success results of classifying a database by all the other databases put together. The diagonal shows the mean results of 50 Self Classification rounds for each database, using a learning group of 2/3 of the samples. All classifications are performed using the LDA+KNN classification process.

[1] The ultimate goal of sound classification is to obtain a "Concept Classifier" - such classifier could recognize which instruments are playing regardless of specific recording conditions, a specific performer or a specific instrument.

	SOL	IOWA	McGill	Pro	Vi	**Minus 1 DB**
SOL classified by	(98.24)	39.93	20.14	21.51	58.17	68.5
IOWA by	51.43	(97.75)	35.22	29.17	58.42	65.79
McGill by	51.76	51.76	(60.78)	23.53	48.23	77.65
Pro by	54.43	41.77	26.58	(48.04)	58.86	75.32
Vi by	63.45	48.59	30.12	20.88	(64.42)	75.9

Table 1: Self Classification, Mutual Classification and Minus-1 DB results using LDA+KNN

Table 1 demonstrates that the results of Self Classification of a database are very different from the results of classifying another database by it. This shows that Self Classification results do not predict how a classifier which is trained on one database will classify new samples. **Point 2 demonstrated**.

Comparing the Minus-1 DB column to the Self Classification results, we see that enriching the learning database by samples out of other databases helps the classifier to generalize better and thus get closer to recognizing the Concept Instruments. Even relatively small databases, which do not even contain enough samples for good Self Classification using LDA+KNN (McGill, Pro and Vi), when they are added to the learning set, considerably improve the generalization ability of the classifier. For example, when SOL (a relatively large database) is classified by IOWA (the largest one), still the results are considerably improved, from 39.93% to 68.5%, when the 3 small databases are added to IOWA (Minus-1 DB classification of SOL). **Point 3 demonstrated**.

By comparing Self Classification and Mutual Classification results, it is possible to evaluate for each database its self containment vs. its diversity, thus concluding how well the database is suited for generalized classification, e.g. when examining Table 1, we can see that Vi, while not appearing to be very self contained, seems to be diverse enough and comparatively suited for classification of the other databases.

Let us now examine Table 2, which contains mean success percents of 20 Self Classification and Minus-1 DB classifications using the BP80 neural network:

	Self Classification	**Minus 1 DB**
SOL	(97.93)	87.78
IOWA	(99.35)	74.71
McGill	(77.86)	80
Pro	(87.55)	84.18
Vi	(92.84)	89.16

Table 2: Self Classification and Minus-1 DB results using BP80

If we compare the "Minus-1 DB" columns in Tables 1 and 2, we see that the neural network generalizes much better than LDA+KNN, probably due to its ability to perform nonlinear analysis. We see that by using this net we can get much closer to a Concept Classifier than with LDA+KNN. Yet, if we compare the Self Classification results of the SOL and IOWA databases in Tables 1 and 2, the results are very similar.

Following from that, if we would compare LDA+KNN and BP80 just by using Self Classification of a single large database, we could conclude that there is no considerable difference in the capabilities of these algorithms and that both perform very well. **Point 1 demonstrated**.

2.2 Claim 2: Evaluation using Self Classification of a classification process where specific instruments are being classified, does not necessarily reflect the suitability of the feature descriptors being used, for general classification of these instruments

In this section we shall see that a feature selection algorithm might choose different features for classification of the same instrument types, depending on the sound database being used. This also means that evaluating features using a single database and Self Classification will not necessarily show the suitability of these features for a Concept Classifier.

For feature selection we shall use our GDE (Gradual Descriptor Elimination) algorithm, which repeatedly performs LDA and removes the least important descriptor, until a desired number of the most important descriptors is left.

Demonstrating Claim 2. Using GDE we have (apparently) chosen the best 8 feature descriptors out of 162, for each sound database. Table 3 shows which features were selected using each database. The upper row contains the indices of all the feature descriptors that were chosen. The asterisks indicate the selected features.

Desc.#	18	19	42	44	45	46	47	48	49	50	51	52	73	134	135	136	137	138	140
SOL	*	*						*					*	*		*	*	*	
IOWA	*			*	*	*	*							*				*	*
McGill			*			*	*				*	*				*	*	*	
Pro	*					*	*		*					*		*	*	*	
Vi	*					*	*			*				*	*	*	*		.
Total:	4	1	1	1	1	4	4	1	1	1	1	1	1	4	1	4	4	4	1 .
All DB's merged				*		*	*	*						*		*	*	*	.

Table 3: 8 best feature descriptors provided by GDE

We see that different features are selected for each DB[2], thus evaluation of features using a single database does not necessarily demonstrate the usefulness of these features for a Concept Classifier. **Claim 2 demonstrated**.

Acknowledgement

Many thanks to Geoffroy Peeters for contributing his descriptors computation routines and sharing his knowledge.

References

Fraser, A. & Fujinaga., I. (1999). Toward real-time recognition of acoustic musical instruments. In *Proceedings of the ICMC, 1999*, (pp. 175-177).

Martin, K. D. & Kim, Y. E. (1998). Musical instrument identification: A pattern-recognition approach. Paper read at the *136'th meeting of the Acoustical Society of America*,

McLachlan, G. J. (1992) *Discriminant Analysis and Statistical Pattern Recognition.* New York, NY: Wiley Interscience.

Peeters, G. (2002). WP2.1 Preliminary Audio Descriptors. *Project CUIDADO - Audio Feature Extraction*.

[2] We see 7 descriptor indices that figure a lot. They are Sharpness, Specific Loudness-19, Specific Loudness-20, Spectral Centroid, Spectral Skewness, Spectral Kurtosis and Spectral Slope. Explanation is found in Peeters (2002).