Feature Selection Pitfalls and Music Classification

Rebecca Fiebrink and Ichiro Fujinaga
Music Technology, McGill University
Montreal, Canada
rfiebrink@acm.org, ich@music.mcgill.ca

Abstract

Previous work has employed an approach to the evaluation of wrapper feature selection methods that may overstate their ability to improve classification accuracy, because of a phenomenon akin to overfitting. This paper discusses this phenomenon in the context of recent work in machine learning, demonstrates that previous work in MIR has indeed exaggerated the efficacy of feature selection for music classification, and presents new testing providing a more realistic analysis of feature selection's impact on music classification accuracy.

Keywords: Feature selection, classification.

1. Introduction

Music classification can employ a palette of hundreds of low-level features (e.g., zero-crossing rate, MFCCs, LPC coefficients) and higher-order variations on these (e.g., standard deviation, first-order difference) [1]. Extracting hundreds of features from a large music collection, however, is costly in terms of both time and space. Furthermore, ideally, the size of a classifier's training set should grow exponentially with the number of features [2]. However, it is not necessarily intuitive which of the possible features will be most relevant to a high-level music classification task, such as genre or artist identification, so it is sensible to look for an automated way of selecting a good subset of the available features.

Unfortunately, some proponents of one such technique, wrapper feature selection, have employed poor evaluation methods that may lead to an exaggeration of its benefits. We discuss recent work re-examining feature selection's efficacy, and we temper our previous claims in [3] with results of recent, better-designed testing on the same data.

2. Wrapper Feature Selection

Wrapper feature selection refers to a subset of feature selection techniques wherein each candidate feature subset visited in the algorithm's search is evaluated by training and testing an induction algorithm (a classifier) using only that feature subset [4]. The classification accuracy for the visited candidate subsets is used to guide the search to new subsets, and the output of the algorithm at termination is the visited candidate subset with the best accuracy. Many wrapper algorithms have been proposed (e.g., forward selection, genetic algorithms), and there exists a body of literature claiming that particular methods work well, or that certain methods are superior to others.

Reunanen [5,6] points out a problem with several such studies (e.g., [7]): they do not use an independent evaluation set to evaluate feature selection's efficacy. In such studies, candidate feature subsets are evaluated using *n*-fold cross-validation accuracy on the entire dataset, computed by a classifier using the feature subset. Unfortunately, the cross-validation score of the best subset is typically not representative of the classification accuracy one can expect for new data. Moore and Lee [8] provide further insight: "… a naïve intensive use of cross validation, perhaps over many thousands of models, may produce a deceptively good lowest-error model, in a manner similar to overfitting of data."

Reunanen [5,6] recommends avoiding this pitfall by partitioning the dataset into mutually exclusive "outer" training and testing sets. Feature selection uses cross-validation accuracy on the outer training set to find the "best" feature subset. The efficacy of selection is evaluated by training a classifier on the outer training set, using the chosen feature subset, and examining classification accuracy on the independent outer testing set. Reunanen [5] uses this evaluation method to defend the use of simple feature selection methods such as forward selection over more intense search methods, whose efficacy may be more greatly exaggerated by poor evaluation techniques. Reunanen [6] further demonstrates that, when evaluated properly, feature selection is often actually ineffective at improving classification accuracy. These findings run contrary to claims of previously published studies not using independent test sets.

3. Feature Selection in MIR

3.1 Previous Work

Previous work in MIR, namely [3], has used the aforementioned poor evaluation methodology for feature selection. That work used the evaluation methodology of [7] to show that feature selection was effective in improving classification accuracy on both musical and non-musical data.

3.2 Re-evaluating Feature Selection in MIR

We re-evaluated the claims of [3] using the recommendations of [5] and [6]. We ran forward selection on several datasets used in [3]. All tests were done using a nearest-neighbor classifier. Table 1 shows classification accuracy on the training and testing sets for the best feature subset. It illustrates that feature selection offers little or no benefit when measured using an independent test set, for these datasets and this classifier.

Table 1. Classification accuracy on testing set with no selection, and accuracy on training and testing sets using forward selection, for timbre recognition tasks from [3]

Timbre recognition task		No selection (Testing set)	Forward selection	
			Training	Testing
Snare	All attack	94.9	100	93.8
	All 512	92.5	98.9	92.1
	Time attack	88.9	96.5	89.4
	Time 512	91.0	95.1	86.2
Beat-box		91.6	98.3	91.1

We wondered whether feature selection would still offer any benefit for other problems in music involving many features, such as audio genre classification. We extracted 74 low-level features from the Magnatune database (4476 songs, 24 genres) [9] using jAudio [1]. We split the data into training and testing sets of equal size, stratified by album. Results appear in Table 2.

Perhaps unsurprisingly, forward selection does boost test set accuracy, though not to the extent suggested by the training set performance. It is reasonable to assume that many of the original low-level features contained redundant or irrelevant information. Additionally, the training and testing sets contained near-identical numbers of songs from each album, so the "album effect" ([10]) suggests that classification accuracy would tend to correlate well.

We compared the above results to those obtained using principle components analysis (PCA) for dimensionality reduction, which re-mapped the data into a new 36-dimensional feature space expressed as a linear combination of the original features. The new features accounted for 95% of the variance of the original features. Results show that, using the same classifier and partitioning of the data as for forward selection, PCA provides a similar increase in accuracy in a fraction of the time (Table 2).

Table 2. Classification accuracy for no selection, forward selection, and PCA on Magnatune genre classification

Method	Time	Training Accuracy	Testing Accuracy
No selection	—	61.9	61.2
Forward	6.3 days	97.7	69.8
PCA	1 minute	70.6	71.0

4. Conclusions

Our testing supports the conclusions of [6]: namely, the efficacy of feature selection in improving classification accuracy has been overstated. Our work in [3] fell into a common pitfall of poor evaluation methodology. New tests suggest that feature selection is in fact unable to significantly improve accuracy on the musical problems used in that work. Further testing on a musical genre classification problem suggests that feature selection can still improve classification accuracy under some circumstances, but PCA results in the same increase in accuracy, in a fraction of the time.

Our results do not preclude the possibility that feature selection will work well for other problems and classifiers in MIR. Also, feature selection can provide other benefits, such as reduced feature extraction time with comparable (or only slightly worse) classification accuracy. In any case, one must employ sound evaluation methods to obtain a clear picture of the impact of feature selection on classification accuracy for a particular problem and classifier.

5. Acknowledgments

We gratefully acknowledge support from the McGill University Max Stern Fellowship in Music and the Canada Foundation for Innovation.

References

[1] D. McEnnis, C. McKay, I. Fujinaga, and P. Depalle, "jAudio: A Feature Extraction Library," in *Int. Conf. on Music Inf. Retr. Proc.*, 2005.

[2] R. O. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2^{nd} ed. New York: Wiley & Sons, Inc., 2001, pp. 169–170.

[3] R. Fiebrink, C. McKay, and I. Fujinaga, "Combining D2K and JGAP for Efficient Feature Weighting for Classification Tasks in Music Information Retrieval," in *Int. Conf. on Music Inf. Retr. Proc.,* 2005.

[4] R. Kohavi and G. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.

[5] J. Reunanen, "Overfitting in Making Comparisons Between Variable Selection Methods," *Journal of Machine Learning Research*, vol. 3, pp. 1371–1382, 2003.

[6] J. Reunanen, "A Pitfall in Determining the Optimal Feature Subset Size," in *Int. Workshop on Pattern Recognition in Information Systems Proc.,* 2004.

[7] M. Kudo and J. Sklansky, "Comparison of Algorithms that Select Features for Pattern Classifiers," *Pattern Recognition*, vol. 33, pp. 25–41, 2000.

[8] A. Moore and M. Lee, "Efficient Algorithms for Minimizing Cross Validation Error," in *Int. Conf. on Machine Learning Proc.,* 1994.

[9] Magnatune, "Magnatune: MP3 Music and Music Licensing (Royalty Free Music and License Music)," [Web site] 2006, Available: http://www.magnatune.com

[10] B. Whitman, G. Flake, and S. Lawrence, "Artist Detection in Music with Minnowmatch," in *IEEE Workshop on Neural Networks for Signal Processing Proc.*, 2001.