Journal of New Music Research
2006, Vol. 35, No. 2, pp. 113 – 120

Statistical Evaluation of Music Information Retrieval Experiments

Arthur Flexer

Institute of Medical Cybernetics and Artificial Intelligence, Austria

Abstract

This work concerns the necessity of statistical evaluation of Music Information Retrieval (MIR) experiments. This necessity is motivated by applying fundamental notions of statistical hypotheses testing to MIR research. Minimum requirements concerning statistical evaluation are developed and the appropriate statistical techniques are introduced and exemplified in a genre classification context. Articles from the MIR literature are examined and criticized for the lack of statistical evaluation they contain.

1. Introduction

The general goal of a music information retrieval (MIR) system can be broken down into two major objectives: the automatic structuring and organization of large collections of digital music, and intelligent music retrieval in such structured "music spaces". To achieve these goals, MIR uses a multitude of methods from diverse fields of science: statistics, signal processing, machine learning, artificial intelligence, data mining, etc. The fundamental role of computer experiments and their proper statistical evaluation in all these subfields is now widely accepted and understood although probably not always applied with the necessary strictness. For the general field of artificial intelligence, a whole textbook has been devoted to this problem (Cohen 1995). For the field of neural networks, early critical work on the lack of proper statistical evaluation includes Flexer (1996) and Prechelt (1996). Salzberg (1997) reports on the comparison of machine learning classifiers including a recommended approach for statistical evaluation.

For the field of MIR, although the foundations of the scientific evaluation of MIR systems has already been discussed (e.g. Downie, 2004), the majority of literature shows a lack of proper statistical evaluation of their results.Reading through recent proceedings of the field's premier conference (International Conference on Music Information Retrieval, ISMIR, Buyoli & Loureiro, 2004) we obtained the following statistics: of the total amount of 104 papers in the proceedings, 53 reported some form of quantitative result[1] and should therefore also contain an appropriate statistical evaluation; of those 53 papers more than 80% (45) reported only mean values of performance measures (accuracy, precision, recall, etc.); only 6 papers reported mean performances plus standard deviations to give an idea of the variability of results; only 2 papers employed a statistical test to prove significance of their results. The evaluation of the Audio Description Competitions held at the same conference[2] also relied on comparing mean performances only.

Given the short history of MIR research, such lack of scientific rigour is understandable. Nevertheless it seems about time to raise the standards of empirical work in MIR in order to allow for faster progression of the whole field. Therefore, this paper is concerned with the quality of statistical evaluation of MIR experiments. It is based on an earlier report on the practice of statistical evaluation in neural network research (Flexer 1996). First we will describe data and methods of a standard

[1] Some of the remaining papers should in principle have had a quantitative evaluation but reported only some illustrative examples instead. The rest of the papers reported on topics such as user interface design, data bases, user studies, etc.

[2] ISMIR 2004, 5th International Conference on Music Information Retrieval, Audiovisual Institute, Universitat Pompeu Fabra Barcelona, Spain, 10 – 14 October 2004; Genre classification, Artist identification, Rhythm classification, Melody estimation, and Tempo induction; see http://ismir2004.ismir.net/ISMIRContest.html

Correspondence: Arthur Flexer, Institute of Medical Cybernetics and Artificial Intelligence, Center for Brain Research, Medical University of Vienna, Freyung 6/2, A-1010 Vienna, Austria. E-mail: arthur.flexer@meduniwien.ac.at

DOI: 10.1080/09298210600834946 © 2006 Taylor & Francis

MIR genre classification scenario in Section 2. This genre classification context will be used to illustrate our line of argumentation throughout the rest of the paper. Next we will establish why statistical evaluation of MIR experiments is a must in Section 3. Then we will formulate minimum requirements concerning re-sampling techniques for training and test data as well as statistical evaluation in Section 4. This is followed by discussion and conclusion in Sections 5 and 6.

2. Genre classification

The following approach to genre classification based on spectral similarity pioneered by Logan & Salomon (2001), Aucouturier & Pachet (2002) and Tzanetakis & Cook (2002) is now seen as one of the standard approaches in the field of music information retrieval. This genre classification context will be used to illustrate our line of argumentation throughout the rest of the paper. For a given music collection of S songs, each belonging to one of G music genres, it consists of the following basic steps:

- for each song, divide raw data into overlapping frames of short duration (around 25 ms);
- compute Mel Frequency Cepstrum Coefficients (MFCC) for each frame (up to 20);
- train a Gaussian Mixture Model (GMM, number of mixtures up to 50) for each of the songs;
- compute a similarity matrix between all songs using the likelihood of a song given a GMM;
- based on the genre information, do nearest neighbour classification using the similarity matrix.

For our experiments we used the data set of the ISMIR 2004 genre classification contest[3]. The data base consists of $S = 729$ songs belonging to $G = 6$ genres. The different genres plus the numbers of songs belonging to each genre are given in Table 1.

Table 1. ISMIR 2004 contest data base (genre, number of songs, percentage).

Genre	No.	%
Classical	320	43.9
Electronic	115	15.8
Jazz Blues	26	3.6
Metal Punk	45	6.2
Pop Rock	101	13.9
World	122	16.7
Sum	729	100.0

We divide the raw audio data into overlapping frames of short duration and use Mel Frequency Cepstrum Coefficients (MFCC) to represent the spectrum of each frame. The frame size for computation of MFCCs for our experiments was 23.2 ms (512 samples), with a hop-size of 11.6 ms (256 samples) for the overlap of frames. We used the first eight MFCCs for all our experiments.

A Gaussian Mixture Model (GMM) models the density of the input data x by a mixture model of the form

$$p(x) = \sum_{m=1}^{M} P_m \mathcal{N}[x, \mu_m, U_m], \tag{1}$$

where P_m is the mixture coefficient for the mth mixture, $\mathcal{N}$ is the normal density and μ_m and U_m are the mean vector and covariance matrix of the mth mixture. Please note that we use diagonal covariance matrices only. The dimensionality of an input vector x is eight. The log-likelihood function is given by

$$L = \frac{1}{T}\sum_{t=1}^{T} \log(p(x^t)) \tag{2}$$

for a data set containing T data points. This function is maximized both with respect to the mixing coefficients P_m and with respect to the parameters of the Gaussian basis functions using Expectation-Maximization (see e.g. Hastie et al. 2001). The log-likelihood L is used to compute the similarity between songs for the similarity matrix. A small log-likelihood indicates a high similarity.

The following research question will be used to illustrate our arguments concerning statistical evaluation in the remainder of the paper: *"Do GMMs with mixtures of 30 Gaussians (GMM30) achieve better genre classification accuracy results than GMMs with mixtures of 10 Gaussians (GMM10)?"*.

3. Why statistical evaluation is a must

We do need experiments in MIR research because the methods we employ and the data we want to analyse are too complex for a complete formal treatment. That is, for a given data analysis problem we do not have the formal instruments to decide which of the methods is the optimal one. Of course there is a vast literature in statistics, computational learning theory and the like that does help us in such decisions. But the last word in the decision is always spoken by an empirical check, an experiment, as in any other science that needs empirical evaluation of its theories (see Kibler & Langley (1988) for a comparison of physics and machine learning). In our genre classification context, there is no theoretical result which would allow us to decide whether a mixture

[3] To be more precise, we used the training set of the contest.

of 10 or 30 Gaussians is more suited for the task at hand.

The basic structure of MIR experiments is the same as in any other experimental situation: the question is whether there are effects of the variation of the independent variables (mainly type and certain parameter characteristics of the methods used and type and characteristics of the data set in question) on the dependent variables (various performance measures like accuracy, precision, root mean squared error or training time). In our genre classification context, the independent variable is the size of the mixture of Gaussians and the dependent variable is the achieved accuracy.

The observations (in terms of dependent variables) that we make during our experiments are only a portion of the entirety of experiments and observations that are possible in principle. There are at least three arguments that motivate this restriction: first, the data available for our experimentation are usually assumed to be just a, hopefully representative, sample of a larger number of data. Second, within such data samples we make divisions into data sets for training and testing, again not all those possible in principle but rather those manageable by our restricted computer resources. Third, there are some random influences (e.g. the random initialization of algorithms, the sequence in which training data are presented) of which again only a sample can be computed and observed. In our genre classification context, of course the songs in the data base are only a small sample of the genres they belong to. We will also use only a limited set of divisions into training and test data and Gaussian mixture models are known to be sensitive to initialization of their parameters.

But how can we be sure that the portion that we are able to observe is representative of the whole number of events in question? "The procedures of statistical inference" allow us "to draw conclusions from the evidence provided by samples" (Siegel 1956). Only by statistical testing can it be ensured that the observed effects on the dependent variables are caused by the varied independent variables and not by mere chance "that they represent real differences in the larger group from which only a few events were sampled" (idem) (i.e. whether the phenomena observed in our sample are significant in a statistical sense or not). Therefore, statistical evaluation of MIR research is in fact a must.

4. Minimum requirements

Minimum guidelines of proper MIR experimentation can be divided into how to select training and test data and how to statistically evaluate such experiments, whereas the former is a prerequisite for the latter.

4.1 Re-sampling techniques

Re-sampling techniques enable one to estimate the performance of a classifier in a fair way, i.e. such that it is guaranteed that approximately the same level of performance will be achieved with a new data set of the same domain.

It is not sufficient to use the so-called *resubstitution* method where the performance of a trained classifier is measured on the data set used for training. It is widely known (see e.g. Ripley (1992), Michie et al. (1994) or Hastie et al. (2001) for discussions related to neural networks, machine learning and statistical pattern recognition) that the performance measure estimated with this resubstitution method is usually overoptimistic, i.e. that the same performance measure computed on new, previously unknown, data is very likely to yield worse results. In statistical terms it is said that such error rates tend to be biased.

Using the *resubstitution* method in our classification context yields accuracy results of 99.86% for GMM10 and 100.00% for GMM30. A close look at the nearest neighbour results reveals that, as expected, in most of the cases the log-likelihood of a song is smallest for the GMM which was trained on that very song and is therefore responsible for the correct classification (this is the case for 715 (GMM10) and 716 (GMM30) songs out of the 729).

Therefore it is at least necessary to *use different sets of data for training and testing*. The simple method of dividing the available data into one training and one test set (2/3 and 1/3 of the data which are mutually exclusive) is called the *holdout* method. Since the algorithm employed cannot use all data for training, performance measures are often pessimistic. Additionally, if such a division into a training and a test set is undertaken, it is necessary to *compute multiple runs* of the experiment in order to avoid random influences (e.g. initialization, specific division of the data, sequence of training data). The computation of multiple runs also gives you a better estimate of the true performance.

A simple modification to the holdout method is a *rotation estimator*. The whole data set is divided into K equally sized parts, and each part is used as a test set for a classifier trained with the remaining data. The observed performance measures for the K different runs are averaged. This procedure is usually known as K-fold *cross-validation*. For our genre classification scenario, it is advisable to actually parallelize the samples that are being drawn for the two classifiers GMM10 and GMM30, i.e. to use the same data for training and testing for both types of GMMs during the multiple runs (i.e. to use so-called paired samples). This yields results which allow better comparison and more powerful test statistics later on. *Cross-validation* is still biased in its estimation of performance and there are other techniques like bootstrap (Efron 1982) that are able to reduce this bias further at even greater computational cost by using re-sampling with replacement.

The result of using a 10-fold *cross-validation* with paired samples in our classification context is depicted in Figure 1. The accuracy results for both GMM10 and GMM30 range between about 68% and 82%. For the majority of folds, GMM30 performs better than GMM10, but not for all of them. It is also evident that the performance is quite dependent on the training and test sets used. The correlation of the GMM10 and GMM30 accuracy results across the ten folds is 0.8. This shows how important parallelization of samples is in this context. It is also obvious how overoptimistic the resubstitution method (yielding almost perfect 100% results for both GMM10 and GMM30) is in comparison to the cross-validation results.

Another important issue, often neglected within machine learning research in general, is the fact that *sometimes another third independent data set is needed* for fair performance estimation. Since it is usually necessary to tune some parameters (e.g. learning rate, number of clusters, size of mixtures, etc.) to get the best performance, a division of the available data into three different sets is recommended. Michie et al. (1994) recommend to hold back approximately 20% of the data and divide the remaining data in a set for training and a set for testing, and then tune the parameter using those two sets and an appropriate re-sampling technique. The final algorithm should use both training and test data for learning with the now optimized parameters and should then be finally tested with the remaining, never before used, 20% of the data. If the use of such a third independent data set is omitted, the obtained error rates will again be biased and overoptimistic because the test set used for repeated tuning in fact becomes a training set. Mosteller & Tukey (1977, p. 37) distinguish between the "form" of a method (i.e. type and form of a method, e.g. GMMs versus *k*-means clustering) and the "numerical values" of its parameters and calls the threefold division of data described above "double cross-validation".

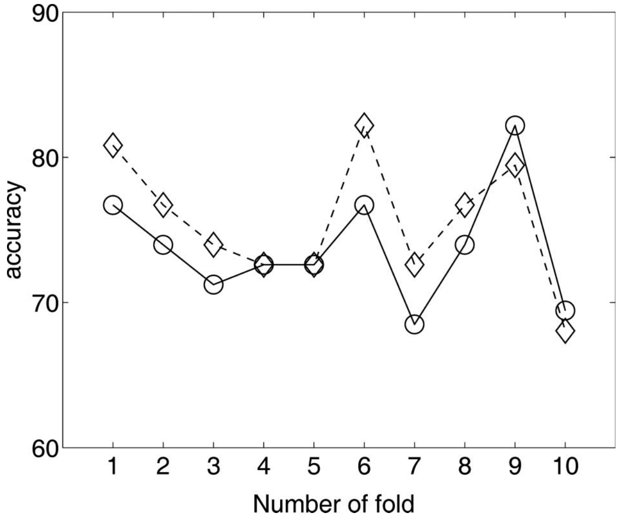

Fig. 1. Results of the 10-fold cross-validation. Number of fold on *x* axis, accuracy in percent on *y* axis, solid line with circles for GMM10 and broken line with diamonds for GMM30.

4.2 Statistical testing

To get more acquainted with the notion of statistical testing, let us start with a brief introduction on hypotheses testing exemplified by our problem in genre classification research. The first step in statistical testing is the formulation of the null hypothesis *H*0. The *H*0 is a hypothesis of no differences that is usually formulated for the express purpose of being rejected, e.g. that there are no performance differences between two classifiers. If *H*0 is rejected, the alternative hypothesis *H*1 may be accepted, in our example that there are performance differences between the two classifiers GMM10 and GMM30. Before actually computing our dependent performance measures, we specify the set of all possible samples of such measures that could occur when *H*0 is true (e.g. the, often only assumed, distribution of performance differences between two classifiers). From these, we specify a subset of possible samples that is so extreme (e.g. very large performance differences) that the probability, if *H*0 is true, to actually observe such a sample is very small. If in our experiments we then observe a sample which was included in that subset (e.g. a rather big performance difference), we reject *H*0. The aforementioned small probability is called the level of significance α, common values are 5 or 1%. The level of significance α is at the same time the probability to falsely reject *H*0, i.e. to come to the conclusion that the observed differences are significant although they are not (this is called "Type I Error"). The second kind of error, "Type II Error", is to accept *H*0 when in fact it is false, i.e. not detecting significant performance differences.

All statistical tests are only valid under certain conditions and can be divided into parametric and non-parametric methods (see e.g. Siegel, 1956). Parametric methods have a variety of strong assumptions (e.g. that of normal distribution of the data) and are therefore more powerful (i.e. it is easier to come to significant results) than non-parametric methods. From what has been outlined above, it should be clear that multiple runs are necessary for classifier experiments and that usually means over the multiple runs are to be evaluated. The use of parametric methods for the evaluation can be justified with the central-limit theorem which suggests that sample means are normally distributed no matter what distribution the samples themselves form. Therefore, parametric tests can be used for both categorical (e.g. accuracy, sensitivity, specificity) and continuous measures (e.g. root mean squared error, training time) of performance. If parametric tests are being used and the assumption of normality does not hold, it can only happen that instances are being judged as "not significant" that otherwise would have been judged as "significant" but not vice versa ("Type II error", see e.g. Mosteller & Tukey (1977), p. 16). If this actually happens, one can still try appropriate non-parametric tests (see Siegel (1956) for a classical treatment).

It is by no means justified to report just the best result of the multiple runs of a classifier. Instead, at least *the mean* of the performance measures (e.g. accuracy) over all those runs *and the corresponding variance σ^2 should be reported* to give a better estimate of the true performance.

The best accuracy result for both GMM10 and GMM30 is 82.19%. The mean accuracy for GMM10 is 73.79% ± 16.00 (mean ± variance) versus 75.57% ± 19.39 for GMM30.

It is even better to *report* the mean over the multiple runs and *the corresponding confidence interval* which can be computed from the standard deviation s. Assuming normal distribution of the data gives:

$$\bar{X} \in [\bar{x} \pm 2.58\hat{\sigma}_{\bar{x}}] \text{ with } \alpha = 1\%, \tag{3}$$

$$\bar{X} \in [\bar{x} \pm 1.96\hat{\sigma}_{\bar{x}}] \text{ with } \alpha = 5\%, \tag{4}$$

$$\hat{\sigma}_{\bar{x}} = s/\sqrt{N}. \tag{5}$$

With a probability of 99% the true value $\bar{X}$ of the observed mean $\bar{x}$ will be within the interval given in equation (3), with a probability of 95% within the interval given in equation (4), where $\hat{\sigma}_{\bar{x}}$ is the standard error estimated from the sample standard deviation s. If the sample is rather small (i.e. number of runs $N \ll 100$, say less than 30), it is no longer justified to assume normality of the distribution of performance measures. Instead, the distribution forms a Student- or t-distribution with degrees of freedom equal $df = N - 1$. The values 2.58 and 1.96 in equations (3) and (4), which were obtained from a standard normal distribution, have to be replaced by the appropriate t-values when computing confidence intervals which in consequence become larger.

Confidence intervals allow us to express the amount of uncertainty that comes with every experiment. They also enable use to compare the outcome of experiments under different conditions, e.g. to compare the accuracy means of two different classifiers applied to one data set by computing the confidence intervals for both of them. If the two confidence intervals do not overlap at all, i.e. if the upper bound of one is below the lower bound of the other, a statistically significant difference is guaranteed (Cohen 1995, p. 137).

The 95% confidence interval for GMM10 is [73.79 ± 2.62 * 1.27] which gives [73.79 ± 2.86]. The 95% confidence interval for GMM30 is [75.57 ± 2.62 * 1.39] which gives [75.57 ± 3.15]. The confidence intervals around the two average accuracies are highly overlapping with each interval containing the other method's mean accuracy result.

Therefore it is advised to use a *t-test*, which *should be computed to test the significance of the difference between means* (see Feelders & Verkooijen (1995) or Egmont-Petersen et al. (1994) for a discussion related to neural nets and classifiers in general). Since in the standard comparative experiment the performance measures are all estimated from the same test samples (identical cross-validation for all methods), which makes them highly correlated, a paired sample t-test should be used which gives a more powerful test statistic. The formulas for the computation of the paired t-test are given in equations (6) to (9). Assume we have two classifiers A and B and we perform a cross-validation with N folds. $x_{A,i}$ and $x_{B,i}$ are the achieved accuracies on the ith test fold.

$$t = \frac{\bar{d}}{s_{\bar{d}}}, \tag{6}$$

$$\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i, \tag{7}$$

$$s_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^{N} d_i^2 - (\sum_{i=1}^{N} d_i)^2/N}{N(N-1)}}, \tag{8}$$

$$d_i = x_{A,i} - x_{B,i}. \tag{9}$$

We compute the t-value and examine the observed performance difference at an appropriate level of significance $\alpha = 1$ or 5% (i.e. a probability of 95 or 99%) and with degrees of freedom $df = N - 1$ for significance with the help of a t-table (for the two-tailed test, i.e. H0 will be rejected when the t-value is either sufficiently small or sufficiently large.).

The difference in genre classification accuracy between GMM10 and GMM30 is not significant: $|\mathrm{t}| = |-2.1077| < t_{(95,df=9)} = 2.26$ (the same of course holds true for the stricter 99% error level).

Dietterich (1998) has argued that the use of a t-test in the above experimental design is not correct since the assumption of independence for the $x_{A,i}$ and $x_{B,i}$ is violated. After all, in a 10-fold cross-validation, each pair of training sets shares more than 80% of the data which may prevent a statistical test from obtaining a good estimate of the amount of variation that would be observed for completely independent training sets. According to Dietterich (1998) this can lead to a slightly increased probability of Type I Error. In case one wants to stay on the safe side, the appropriate non-parametric alternative is McNemars's test (Everitt 1977).

If more than two means of performances are compared via repeated pairwise t-testing one will end up with a high probability to find one or more 'significant' differences when in fact there are none (e.g. for 20 tests with $\alpha = 5\%$, the probability of such an error is 0.64). The simplest approach to deal with this *multiplicity effect* is to divide α through the number of tests that are being performed, which makes it rather hard to come to significant results. Some pointers to more sophisticated solutions can be found in Feelders & Verkooijen (1995) or Cohen (1995, p. 189).

5. Discussion

In the last section we have introduced a range of different resampling techniques plus appropriate statistical measures and tests. All these have been illustrated using a genre classification context and trying to answer a standard MIR research question: *"Do GMMs with mixtures of 30 Gaussians (GMM30) achieve better genre classification accuracy results than GMMs with mixtures of 10 Gaussians (GMM10)?"*. Table 2 sums up the answers different evaluation approaches yielded. The correct answer is that GMM10 and GMM30 do not differ significantly in their accuracy performance. This can be seen by looking at the highly overlapping confidence intervals (last line in Table 2) and by computing a paired sample t-test. Other evaluation strategies can give quite incorrect answers: not using separate data for testing (resubstitution method) gives very overoptimistic results; using cross-validation and reporting the range of results gives too wide intervals compared to the correct confidence interval; reporting just the results from the best folds is again too optimistic; looking just at the means might tempt one to see an advantage on the side of GMM30 when in reality there is none.

Table 2. Summary of different accuracy results depending on evaluation method used. Results are given in percent correct classification.

Evaluation method	GMM10	GMM30
resubstitution	99.86	100.00
cross-validation range	68.49 – 82.19	68.06 – 82.19
cross-validation best	82.19	82.19
cross-validation mean	73.79	75.57
cross-validation confidence interval	70.93 – 76.65	72.42 – 78.72

To sum up what has been said so far, the following can be seen as minimum requirements for proper MIR experimentation:

- the use of different training and test sets;
- the computation of multiple runs using an appropriate re-sampling technique;
- to report mean, variance and confidence intervals;
- to compute a statistical test (e.g. a t-test) for the comparison of performances.

Looking again at our review of recent ISMIR conference proceedings (Buyoli & Loureiro 2004) from Section 1, the following can be noted: the use of different training and test sets as well as computation of multiple runs seems to be well established; but only 6 out of 53 papers reported more than just the mean performance results and only 2 out of 53 papers employed a statistical test.

The minimum requirements for the quality of statistical evaluation which we introduced above should be seen as being necessary but maybe sometimes not sufficient, i.e. if an experiment *does not* meet them, its quality will be impaired, if it *does* meet them, there are still other things that can go wrong. Some of the possible other pitfalls include: the number of data available for training is too little, the dimensionality of the input vectors is too high, general errors in the design of the experiments, etc.

In the case of classification research one should also not forget to compare the results with the so-called *baseline accuracy*. The baseline accuracy is the accuracy that can be trivially achieved by a naive guesser which simply picks the most frequent class all the time. In our genre classification context the most frequent class is "classical" to which 43.9% of all songs belong. Accuracy results should therefore never fall below this threshold. Confidence intervals or statistical tests could be used to compare the classifier performance with the baseline accuracy.

A problem inherent to genre classification experiments in MIR research is the use of songs from the same artist in both training and test sets. It can be argued that in such a scenario one is doing artist classification rather than genre classification. Pampalk et al. (2005) propose to use a so-called "artist filter" ensuring that all songs from an artist are in either the training or the test set. The authors found that the use of such an artist filter can worsen the classification results quite considerably (with one of their music collection even from 71% down to 27%). For the genre classification results reported in this paper no artist filter has been used, i.e. songs from the same artist appeared both in training and test sets. Our music collection contains 729 songs from 128 different artists. Applying an artist filter and repeating the 10-fold cross-validation as described above yields the following results (95% confidence intervals): [59.68 ± 5.49] for GMM10 and [61.60 ± 4.92] for GMM30. The accuracies dropped from around 75% to around 60%. The confidence intervals around the two average accuracies are still highly overlapping with each interval containing the other method's mean accuracy result. The mean performance of GMM30 is still higher than that of GMM10 but the difference is again not significant ($|t| = |-2.1433| < t_{(95, df=9} = 2.26$).

As with any other classification problem, the results in a genre classification experiment are highly dependent on the data base used in the evaluation. Factors which clearly influence the performance are number and types of genres as well as artists but also quality of class labels being used (see e.g. Pampalk et al. (2005) for a comparison of results on four different data sets). To simulate the effect of different types of genres in a data base we did the following experiment: we computed a series of six 10-fold cross-validations using the data base described in Section 2 and deleting the songs from one whole genre for each of the cross-validations; we used GMM10 as a classifier without using an artist filter. This gives us genre classification results with six different albeit overlapping data sets with the following accuracy results (95% confidence intervals): without-classical [66.71 ± 5.74], without-electronic [81.75 ± 5.13], without-jazz-blues [76.85 ± 3.20], without-metal-punk [78.19 ± 2.35], without-pop-rock [83.41 ± 4.29], without-world [84.00 ± 3.01]. As can be seen, dismissing songs of the genre "classical" worsens the classification results whereas deleting all "pop rock" or "world" songs improves them. This is no surprise since classical music is quite distinct from the other genres in terms of its timbre representation captured via MFCCs. The genres "pop rock" and "world" on the other hand are notoriously ill-defined.

6. Conclusion

In this work we have motivated the necessity of statistical evaluation of Music Information Retrieval (MIR) experiments and have given minimum requirements to be met. We have illustrated our line of argumentation using a standard research question of a genre classification context. Our review of the papers published in the field's premier conference has shown a clear lack of proper statistical evaluation of their results. It is our hope that our work will be of help to enhance the quality of MIR experimentation and thereby allow for faster progression of the whole field.

7. Acknowledgments

Parts of the MA Toolbox (Pampalk 2004) and the Netlab Toolbox[4] have been used for this work. Thanks are due to Elias Pampalk for the artist filter code and to all the members of OFAI's "intelligent music processing group" for introducing me to the field of MIR. Part of this research was supported by the EU project FP6-507142 SIMAC[5].

References

Aucouturier, J.-J. & Pachet, F. (2002). Music similarity measures: what's the use? In *Proceedings of the Third International Conference on Music Information Retrieval (ISMIR'02)* pp. 157–163, Paris: IRCAM.

Buyoli, C.L. & Loureiro, R. (Eds.). (2004). *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*. Barcelona: Audiovisual Institute, Pompeu Fabra University.

Cohen, P.R. (1995). *Empirical Methods for Artificial Intelligence*. Cambridge, MA: MIT Press.

Dietterich, T.G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1924.

Downie, J.S. (2004). The scientific evaluation of music information retrieval systems: foundations and future. *Computer Music Journal*, *28*(2), 12–23.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. SIAM CBMS-NSF Monograph 38, Philadelphia.

Egmont-Petersen, M., Talmon, J.L., Brender, J. & McNair, P. (1994). On the quality of neural net classifiers. *Artificial Intelligence in Medicine*, *6*, 359–381.

Everitt, B.S. (1977). *The Analysis of Contingency Tables*. London: Chapman & Hall.

Feelders, A. & Verkooijen W. (1995). Which method learns most from the data? In *Proceedings of the Fifth International Workshop on AI and Statistics*, January 1995, Fort Lauderdale, Florida, pp. 219–225.

Flexer, A. (1996). Statistical evaluation of neural network experiments: minimum requirements and current practice. In: R. Trappl (Ed.), *Cybernetics and Systems '96*. pp. 1005–1008. Wien: Oesterreichische Studiengesellschaft fuer Kybernetik.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. New York/Berlin/Heidelberg: Springer.

Kibler, D. & Langley, P. 1988. Machine learning as an experimental science. *Machine Learning*, *3*(1), 5–8.

Logan, B. & Salomon, A. (2001). A music similarity function based on signal analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, Tokyo, Japan, p. 190.

Michie, D., Spiegelhalter, D.J. & Taylor, C.C. (Eds.). (1994). *Machine Learning, Neural and Statistical Classification*. England: Ellis Horwood.

Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Regression – A Second Course in Statistics*. Reading, MA: Addison-Wesley.

[4] http://www.ncrg.aston.ac.uk/netlab

[5] http://www.semanticaudio.org

Pampalk, E. (2004). A Matlab Toolbox to compute music similarity from audio. In Buyoli & Loureiro (2004). pp. 254–257.

Pampalk, E., Flexer, A. & Widmer, G. (2005). Improvements of audio-based music similarity and genre classification. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR'05)*, London, UK, 11–15 September, pp. 628–633.

Prechelt, L. (1996). A quantative study of experimental evaluations of neural network learning algorithms: current research practice. *Neural Networks*, *9*(3), 457–462.

Ripley, B.D. (1992). *Statistical Aspects of Neural Networks*. Oxford: Department of Statistics, University of Oxford.

Salzberg, S. (1997). On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, *1*(3), 317–328.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. Tokyo: McGraw-Hill.

Tzanetakis, G. & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293–302.