Transactions on
Case-Based Reasoning
Vol. 2, No. 1 (2009) 3 - 14
©ISSN:1867-366X (Journal),
ISBN: 978-3-940501-8-0,
IBaI Publishing ISSN 1864-9734

ibai Publishing
www.ibai-publishing.org

Distances in Classification

Claus Weihs and Gero Szepannek

Department of Statistics
University of Dortmund
44227 Dortmund

Abstract. The notion of distance is the most important basis for classification. This is especially true for unsupervised learning, i.e. clustering, since there is no validation mechanism by means of objects with known groups. But also for supervised learning standard distances often do not lead to appropriate results. For every individual problem the adequate distance is to be decided upon. This is demonstrated by means of three practical examples from very different application areas, namely social science, music science, and production economics. In social science, clustering is applied to spatial regions resulting in unconnected clusters. However, connectedness is sometimes important for interpretation, and may have to be taken into account for clustering. In statistical musicology the main problem is often to find an adequate transformation of the input time series as an adequate basis for distance definition. Also, local modelling is proposed in order to account for different subpopulations, e.g. instruments. In production economics often many quality criteria have to be taken into account with very different scaling. In order to find a compromise optimum classification, this leads to a pre-transformation onto the same scale, called desirability.

1 Introduction

The notion of distance is the most important basis for classification. This is especially true for unsupervised learning, i.e. clustering, since there is no validation mechanism by means of objects with known groups. But also for supervised learning standard distances often do not lead to appropriate results. For every individual problem the adequate distance is to be decided upon. Obviously, the choice of the distance measure determines whether two objects naturally go together (Anderberg, 1973). Therefore, the right choice of the distance measure is

one of the most decisive steps for the determination of cluster properties. The distance measure should not only adequately represent the relevant scaling of the data, but also the study target to obtain interpretable results.
Some classical distance measures in classification are discussed in the following. In supervised statistical classification distances are often determined by distributions. A possible distance measure treats each centroid and covariance matrix as the characteristics of a normal distribution for that class. For each new data point we calculate the probability that that point came from each class; the data point is then assigned to the class with the highest probability. A simplified distance measure assumes that the covariance matrices of each class are the same. This is obviously justified if the data is similarly distributed for each class, however, nothing prevents from using this assumption if this is unclear. Examples for the application of such measures are **Quadratic and Linear Discriminant Analysis** (QDA and LDA) (Hastie et al., 2001, pp. 84). For a more general discussion of distance measures in supervised classification see Gnanadesikan (1977).
With so-called kernels, standard transformations are explicitly introduced in classification methods, e.g., like in **Support Vector Machines** (SVM) (Hastie et al., 2001, p. 378), in order to transform the data so that it can be separated linearly as with LDA.
For **decision trees**, e.g., a measure for the distance between partitions is proposed such that the selected split attribute in a node induces the partition which is closest to the correct partition of the subset of training examples corresponding to this node (Lopez De Mantaras, 1991). However, also the standard CART decision tree (Breiman et al., 1983) can be interpreted to be associated with a distance measure as follows: Identify splits with maximally distant nodes by maximizing

$$d(split(node)) = i(node) - p_L \cdot i(node_L) - p_R \cdot i(node_R)$$

with the Gini-index or the entropy as **impurity measures** i, and p_L, p_R as the proportion of elements of *node* split into the left node $node_L$ and the right node $node_R$, respectively.
In unsupervised classification Euclidean distance is by far the most chosen distance for metric variables. One should notice, however, that the Euclidean distance is well-known for being outlier sensitive. This might motivate switching to another distance measure like, e.g., the **Manhattan-distance** (Tan et al., 2005). Moreover, one might want to discard correlations between the variables and to restrict the influence of single variables. This might lead to transformations by means of the covariance or correlation matrices, i.e. to **Mahalanobis-distances** (Tan et al., 2005). Any of these distances between two data points can then be used for defining the distance between groups of data. Examples are the minimum distance between the elements of the groups (**single linkage**), the maximum distance (**complete linkage**), and the average distance (**average linkage**) (Hastie et al., 2001, p. 476). For non-metric variables often methods are in use, which, e.g., count the number of variables with matching values in

the compared objects, examples are the **Hamming-, the Jaccard-** and **the simple matching distances** (Tan et al., 2005). E.g., the Jaccard-distance is defined for binary problems by

$$d_{Jac}(X,Y) = \frac{\text{no.(non-matching entries in X and Y)}}{\text{no.(double-positives + non-matching)}}.$$

Thus, data type is an important indicator for distance selection. E.g., in Perner (2002), distance measures for image data are discussed. However, distance measures can also be related to other aspects like, e.g., application. E.g. time-series representing music pieces need special distances (Weihs et al. 2007). Other important aspects of distance are translation, size, scale and rotation invariance, e.g. when technical systems are analysed (Perner, 2008).

Last but not least, **variable selection** is a good candidate to identify the adequate space for distance determination for both supervised and unsupervised classification. For an overview over variable selection methods in classification see, e.g., Dash and Liu (1997).

In practice, most of the time there are different plausible distance measures for an application. Then, quality criteria are needed for distance measure selection. In supervised classification the misclassification error rate estimated, e.g., on learning set independent test sets, is the most accepted choice. In unsupervised learning, one might want to use background information about reasonable groupings to judge the partitions, or one might want to use indices like the ratio between within and between cluster variances which is also optimized in discriminant analysis in the supervised case.

In what follows examples are given for problem specific distances. The main ideas are as follows. Clusters should often have specific properties which are not related to the variables that are clustered, but to the space where the clusters are represented. As an example city districts are clustered by means of social variables, but represented on a city map. Then, e.g., the connection of the individual clusters may play an important role for interpretation. This may lead to an additional objective function for clustering which could be represented by a distance measure for unconnected cluster parts. These two objective functions or distance measures could be combined to a new measure. Another, much simpler, possibility would be, however, just to include new variables in the analysis representing the district centres. By differently weighting the influence of these variables the effect of the variables can be demonstrated. This will be further discussed in section 2.1.

Often, the observed variables are not ideal as a basis for classification. Instead, transformations may be much more sensible which directly relate to a re-definition of the distance measure. Also, in supervised classification the observed classes may not have the right granularity for assuming one simple distance measure per class. Instead, such distances may be more adequate for subclasses, which may be, e.g., defined by known subpopulations across the classes or by unknown subclasses of the classes. Distances then relate to, e.g., distributions in subclasses, i.e. to mixtures of distributions in classes. This will be further discussed in section 2.2.

Another example for more than one objective function is given for production economics. Typically, for more than one objective function there is the problem of weighting the different targets. In contrast to section 2.1 this can also be achieved by transformation to a common scale by means of different so-called desirability functions. The overall distance is then typically related to some combination of the different desirabilities in a so-called desirability index. This will be further discussed in section 2.3.

2 Case-based distance measures

2.1 Additional variables

In social science clustering is often applied to spatial regions with very irregular borders. Then adequate spatial distances may have to be taken into account for clustering. Clusters of spatial regions should most of the time represent similar properties of the predefined regions. However, for better interpretation the question arises as well whether the resulting clusters are connected in space. Then, two different kinds of distances have to be compared, namely the distance of regions in clusters related to given properties and the spatial dispersion of the clusters.
Assume that spatial regions are predefined, e.g. as city districts. Consider the case where some clusters are already defined, e.g. by means of social properties in the regions. In more detail, social milieus were clustered by means of six social variables (after variable selection), namely "fraction of population of 60-65", "moves to district per inhabitant", "apartments per house", "people per apartment", "fraction of welfare recipients" and "foreigners share of employed people". Then, the question arises whether clusters represent connected regions in space.

In Roever and Szepannek, 2005, cluster dispersion was not explicitly utilized as a criterion for good clustering. They just minimize the Classification Entropy

$$CE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}(u_{ij}\log_2 u_{ij}),$$

where N = number of observations, k = number of clusters, u_{ij} = probability that observation i belongs to cluster j. Using this fitness function and some variables' subgrouping, $k = 4$ clusters were produced similar to Figure 1 by means of genetic programming. Note that the white areas were not clustered.
In order to represent the connectedness of the individual clusters, an additional objective function for clustering could represent a distance measure for unconnected cluster parts. The then resulting two objective functions or distance measures could be combined to a new measure. In this paper, however, we have tried to take into account cluster dispersion explicitly in that we introduced new variables representing the x- and y-coordinates of the district centres. By this,

Fig. 1. Clusters of districts of the City of Dortmund (Germany)

distance of district centres are also taken into account with clustering. When these centre variables were weighted only 20% of the other variables the result was hardly influenced (Figure 2, left). After they were weighted twice as much as the other variables, however, the result was totally different and the clusters were much more connected (Figure 2, right).

2.2 Transformations and local modelling

In statistical musicology the main problem is often to find the right transformation of the input time series adequate for analysis. Also, local modelling is proposed in order to account for different subpopulations, e.g. instruments.
This example of distance definition concerns supervised classification. In music classification the raw input time series are seldom the right basis for analysis. Instead, various transformations are in use (see, e.g., Weihs et al., 2007). Since with music frequencies play a dominant role, periodograms are a natural representation for observations. From the periodogram corresponding to each tone, voice characteristics are derived (cp. Weihs and Ligges, 2003). For our purpose we only use the mass and the shape corresponding to the first 13 partials, i.e. to the fundamental frequency (FF) and the first 12 overtones (OTs), in a pitch independent periodogram (cp. Figure 3). Mass is measured as the sum of the percentage share (%) of the peak, shape as the width of the peak in parts of half tones (pht) between the smallest and the biggest involved frequency.

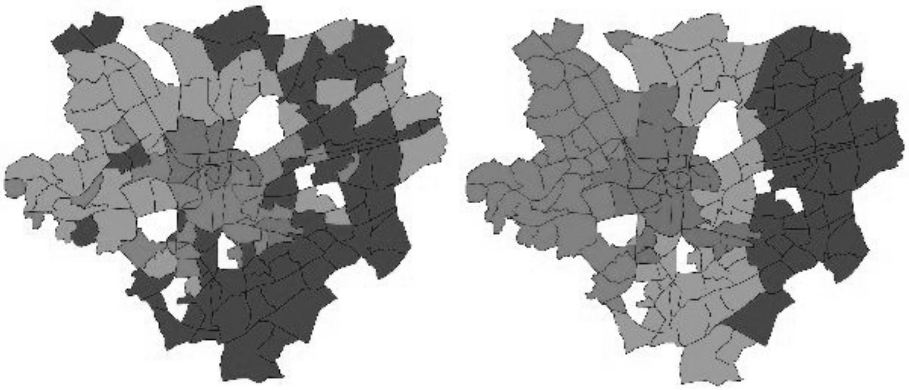

Fig. 2. Clusters with 20%- (left) and 200%- (right) weighting of district centres

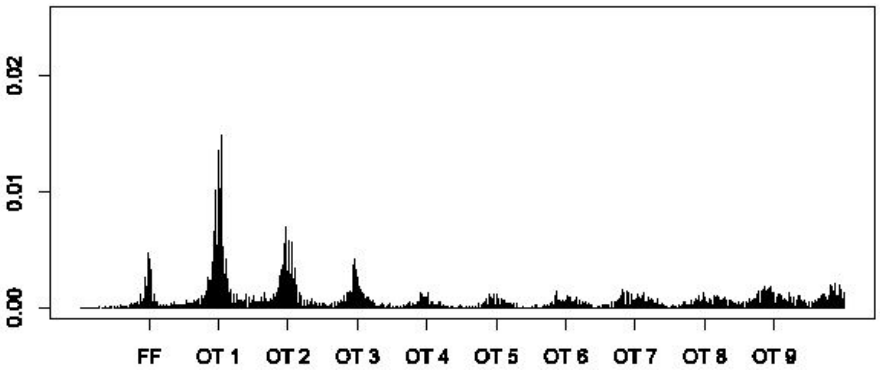

Fig. 3. Pitch independent periodogram (professional bass singer)

These 26 characteristics were determined for each individual tone, as well as averaged characteristics over all involved tones leading to only one value for each characteristic per singer or instrument. LDA based on these characteristics results in an astonishingly good prediction of register (classes low / high) (Weihs et al., 2005). The register of individual tones are predicted correctly in more than 90% of the cases for sung tones, and classification is only somewhat worse if instruments are included in the analysis. Even better, if the characteristics are averaged over all involved tones, then voice type (high or low) can be predicted without any error.
However, this classification appeared, in a way, to be too good so that it was suspected that mass and/or width might somewhat reflect frequency and thus register though the pitch independent periodogram was used. And indeed, simulations showed that width is frequency dependent because it is measured in number of half tones (s. Figure 4). However, if the absolute width in number of involved Fourier-Frequencies is used instead, then this dependency is dropped leading, though, to poorer classification quality. This example distinctly demonstrates an effect of choosing a wrong transformation, and thus a wrong distance measure.

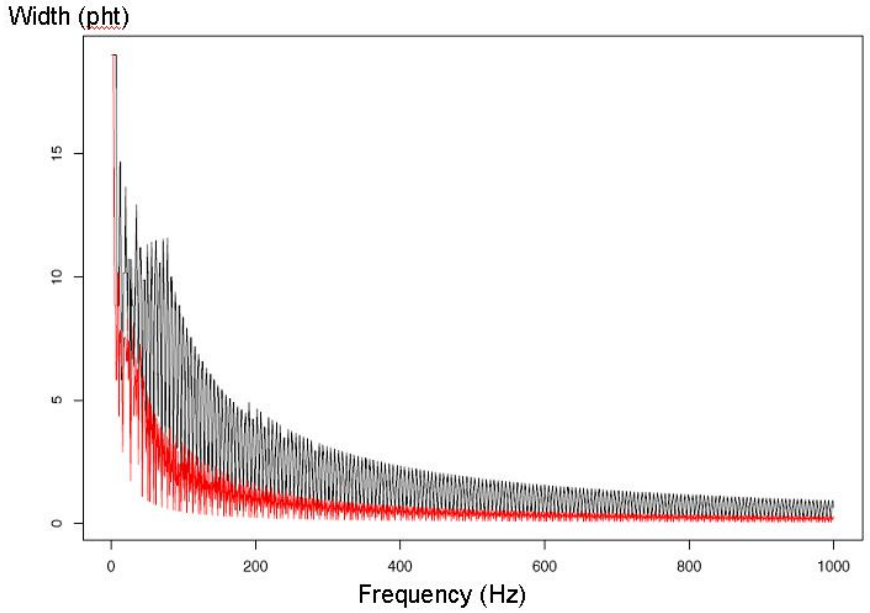

Fig. 4. Width measured in parts of half-tone (pht) dependent on frequency (upper line = fundamental frequency, lower line = first overtone)

In subsequent analyses (Weihs et al., 2006, Szepannek et al., 2008) this redefined width characteristics was applied to a data set consisting of 432 tones (= observations) played / sung by 9 different instruments / voices. In order to admit different behaviour for different instruments, so-called **local modelling** was applied building local classification rules for each instrument separately. For this, we consider the population to be the union of subpopulations across the classes high / low. Then, a mixture distribution is assumed for each class. The problem to be solved consists in register prediction for a new observation if the instrument (and thus the choice of the local model) is not known. This task can be formulated as some globalization of local classification rules. A possible solution is to identify first the local model, and further work only with the parts of the mixtures in the classes corresponding to this model.
Imagine all local (subpopulation-) classifiers return local class posterior probabilities $P(k|l,x)$, where $k = 1,\ldots,K$ denotes the class, x is the actual observation and $l = 1,\ldots,L$ is the index of the local model, i.e. the instrument in our case. The following **Bayes Rule**

$$\hat{k} \;=\; \arg\max_k \sum_l P(k|l,x)P(l|x)$$

showed best performance for the musical register classification problem. To implement this, an additional classifier has to be built to predict the presence of each local model l for a given new observation x. Using LDA for both classification models, the local models and the global decision between the local models, leads to the best error rate of 0.263 on the data set. Note that - since only posterior probabilities are used to build the classification rule - all models can be built on different subsets of variables, i.e. subpopulation individual variable selection can be performed. This may lead to individual distance measures for the different localities (voices, instruments) and for the global decision.

2.3 Common scale

In production economics often many quality criteria have to be taken into account with very different scaling. In order to find a compromise optimum, a pre-transformation, called desirability, onto the same scale may be used.
In a specific clustering problem in production economics product variants should be clustered to so-called product families so that production interruptions caused by switching between variants (so-called machine set-up times) are minimal (Neumann, 2007). Three different distance measures (Jaccard, simple-matching, and Euclidean) and many different clustering methods partly based on these distance measures are compared by means of four competitive criteria characterizing the goodness of cluster partitions, namely the similarity of the product variants in the product families, the number of product families, the uniformity of the dispersion of the product variants over the product families, and the number of product families with very few product variants. Therefore, partition quality is measured by $d = 4$ criteria. Overall, the problem is therefore to identify the

cluster method and the corresponding distance measure, as well as the number of clusters, i.e. the number of product families, optimal to all four criteria. In order to rely on only one compromise criterion a so-called desirability index is derived.
In order to transform all these criteria to a common scale, the four criteria are first transformed to so-called **desirabilities** w_i, a value in the interval $[0, 1]$, where 1 stands for best and 0 for worst, unacceptable quality. In order to join the criteria to one objective function, a so-called **desirability index** W (Harrington, 1965) is defined

$$W : \{w_1, w_2, \ldots, w_d\} \rightarrow [0, 1].$$

Harrington (1965) suggests the geometric mean for W:

$$W(w_1, \ldots, w_d) = \sqrt[d]{\prod_{i=1}^{d} w_i}.$$

This choice has the advantage that $W = 0$ already if one desirability $w_i = 0$, and $W = 1$ only if all $w_i = 1$. Another reasonable index choice would be $\min(w_1, \ldots, w_d)$ with the same properties. The geometric mean will be used here.
In order to minimize the average machine set-up time the following desirability is defined:

$$w_1(C^{(k)}) = 1 - \sum_{i=1}^{k} \sum_{X_j, X_l \in C_i, j \neq l} d_{Jac}(X_j, X_l),$$

where $C^{(k)}$ is a partition with k clusters, and $d_{Jac}(X_j, X_l)$ is the Jaccard distance between product variants X_j and X_l characterizing the machine set-up time between these products.
In this application, for the number of product families a certain range is assumed to be optimal. This lead to the desirability function w_2 indicated in Figure 5, where the number of product families with desirability = 1 are considered optimal.

For application roughly equal sized clusters are of advantage. This leads to a criterion based on the number n_w of within cluster distances of a partition, i.e. the number of distances between objects in the same cluster. When $\min C^{(k)}(n_w)$ is the minimal number of distances over all possible partitions of size k with n objects, and $\max C^{(k)}(n_w)$ the corresponding maximum, this leads, e.g., to the following criterion to measure how imbalanced the cluster sizes are:

$$w_3(C^{(k)}) = 1 - \frac{n_w - \min C^{(k)}(n_w)}{\max C^{(k)}(n_w) - \min C^{(k)}(n_w)}.$$

Product families with less than five product variants are not desirable. This leads, e.g., to the criterion:

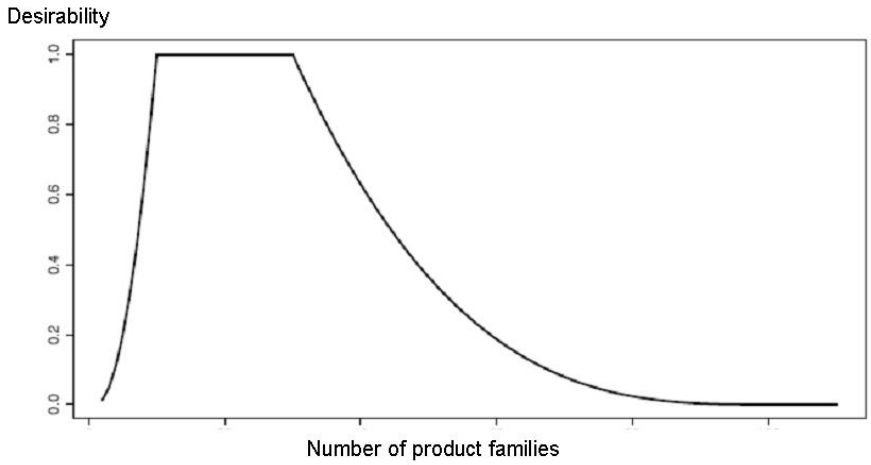

Fig. 5. Desirability function w_2

$$w_4(C^{(k)}) \;=\; 2^{-a}$$

with a = number of product families with less or equal five variants.

All these desirability functions are to be maximized, and thus also the desirability index. No attempt has been undertaken up to now to find an adequate distance measure between partitions $C^{(k)}$ based on desirability indices. That such a task might not be trivial might be derived from an attempt to find the distribution of desirability index (Trautmann and Weihs, 2006). Therefore, standard cluster methods are applied to the problem.

Some results of different cluster methods (for each method based on the most appropriate distance measure) evaluated with the desirability index of the four desirability criteria are shown in Figure 6. Obviously, Ward clustering (Ward, 1963) appears to be best, and for about the intended number of product families the index is maximal.

3 Conclusion

In section 2 it is demonstrated by means of examples from very different application areas that various transformations might be necessary to be able to use an adequate distance measure for unsupervised and supervised classification. In section 2.1 additional variables were added with tentative weights, in section 2.2 the original variables were transformed before application of standard methods and local measures appeared adequate, in section 2.3 original criteria were transformed to a common scale and combined to one criterion used for optimal clustering. All these examples showed that application of standard methods to originally observed variables might not be adequate for problem solution.

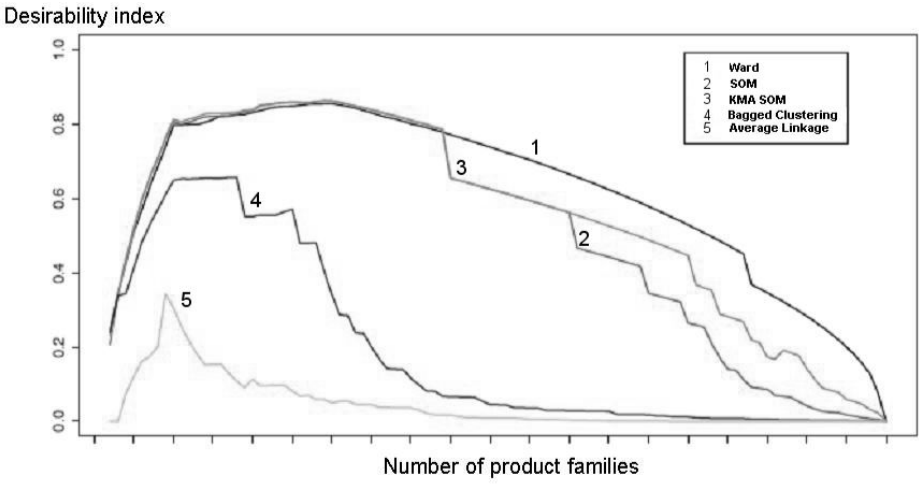

Fig. 6. Desirability index for different cluster methods

Acknowledgments

The authors thank cand. Stat. O. Mersmann for conducting the cluster analyses including district centres. Also, financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

References

Anderberg, M.R.: Cluster Analysis for Applications. Acadamic Press, New York (1973)

Breiman, L., Friedman, J. H., Olshen, R. A., Stone, C. J.: CART: Classification and Regression Trees. Wadsworth: Belmont, CA (1983)

Dash, M., and Liu, H.: Feature selection for classification. Intelligent Data Analysis 1, 131-156 (1997)

Gnanadesikan, R.: Methods for Statistical Data Analysis of Multivariate Observations. Wiley, New York (1977)

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning - Data Mining, Inference and Prediction. Springer, New York (2001)

Lopez De Mantaras, R.: A Distance-Based Attribute Selection Measure for Decision Tree Induction Machine Learning 6, 81–92 (1991)

Neumann, C.: Einsatz von Clusterverfahren zur Produktfamilienbildung. Diploma Thesis, Department of Statistics, TU Dortmund (2007)

Perner, P.: Case-based reasoning and the statistical challenges. Journal Quality and Reliability Engineering International 24 (6), 705–720 (2008)

Perner, P.: Data Mining on Multimedia Data. Lecture Notes in Artificial Intelligence, Vol. 2558. Springer, Berlin (2002)

Roever, C. and Szepannek, G.: Application of a Genetic Algorithm to Variable Selection in Fuzzy Clustering. In: C. Weihs, W. Gaul (eds.): Classification - the Ubiquitous Challenge, pp. 674 - 681. Springer, Heidelberg (2005)

Szepannek, G., Schiffner, J., Wilson, J. and Weihs, C.: Local Modelling in Classification. In: Perner, P. (ed.): Advances in Data Mining: Medical Applications, E-Commerce, Marketing, and Theoretical Aspects; Lecture Notes in Computer Science, vol. 5077, pp. 153 – 164. Springer, New York (2008)

Tan, P.-N., Steinbach, M. and Kumar, V.: Introduction to Data Mining. Addison-Wesley (2005)

Trautmann, H. and Weihs, C.: On the Distribution of the Desirability Index Using Harrington's Desirability Function. Metrika 63, 207–213 (2006)

Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association 58, 236–244 (1963)

Weihs, C., Ligges, U., Mörchen, F. and Müllensiefen, D.: Classification in Music Research. Advances in Data Analysis and Classification (ADAC) 1(3), 255–291 (2007)

Weihs, C., Szepannek, G., Ligges, U., Lübke, K., and Raabe, N.: Local models in register classification by timbre. In: V. Batagelj, H.-H- Bock, A. Ferligoj, A. Ziberna: Data Science and Classification; Springer; 315–332 (2006)

Weihs, C., Reuter, C. and Ligges, U. : Register Classification by Timbre. In: C. Weihs, W. Gaul (Eds), Classification - The Ubiquitous Challenge, Springer-Verlag, Berlin, 624–631 (2005)

Weihs, C. and Ligges, U.: Voice Prints as a Tool for Automatic Classification of Vocal Performance. In: Kopiez, R., Lehmann, A. C., Wolther, I. and Wolf, C. (eds.), Proceedings of the 5th Triennial ESCOM Conference, Hanover University of Music and Drama, Germany, September 8-13, 332–335 (2003)