Petra Perner (Ed.)

LNAI 5633

Advances in Data Mining

Applications and Theoretical Aspects

9th Industrial Conference, ICDM 2009
Leipzig, Germany, July 2009
Proceedings

Lecture Notes in Artificial Intelligence 5633

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Petra Perner (Ed.)

Advances in Data Mining

Applications and Theoretical Aspects

9th Industrial Conference, ICDM 2009
Leipzig, Germany, July 20-22, 2009
Proceedings

Springer

Volume Editor

Petra Perner
Institute of Computer Vision
and Applied Computer Sciences, IBaI
Kohlenstr. 2
04107 Leipzig, Germany
E-mail: pperner@ibai-institut.de

Library of Congress Control Number: Applied for

CR Subject Classification (1998): I.2.6, I.2, H.2.8, K.4.4, J.3, I.4, J.1

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-642-03066-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-03066-6 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12718375 06/3180 5 4 3 2 1 0

Preface

This volume comprises the proceedings of the Industrial Conference on Data Mining (ICDM 2009) held in Leipzig (www.data-mining-forum.de).

For this edition the Program Committee received 130 submissions. After the peer-review process, we accepted 32 high-quality papers for oral presentation that are included in this book. The topics range from theoretical aspects of data mining to applications of data mining, such as on multimedia data, in marketing, finance and telecommunication, in medicine and agriculture, and in process control, industry and society.

Ten papers were selected for poster presentations that are published in the ICDM Poster Proceedings Volume by *ibai-publishing* (www.ibai-publishing.org).

In conjunction with ICDM two workshops were run focusing on special hot application-oriented topics in data mining. The workshop Data Mining in Marketing DMM 2009 was run for the second time. The papers are published in a separate workshop book "Advances in Data Mining on Markting" by *ibai-publishing* (www.ibai-publishing.org). The Workshop on Case-Based Reasoning for Multimedia Data CBR-MD ran for the second year. The papers are published in a special issue of the *International Journal of Transactios on Case-Based Reasoning* (www.ibai-publishing.org/journal/cbr).

We are pleased to announce that we gave out the best paper award for ICDM fourth time. More details are mentioned at www.data-mining-forum.de. The final decision was made by the Best Paper Award Committee based on the presentation by the authors and the discussion with the auditorium. The ceremony took place at the end of the conference. This prize is sponsored by ibai solutions (www.ibai-solutions.de) one of the leading data mining companies in data mining for marketing, Web mining and E-commerce.

The conference was rounded up by a session on new challenging topics in data mining before the Best Paper Award Ceremony.

We also thank the members of the Institute of Applied Computer Sciences, Leipzig, Germany (www.ibai-institut.de) who handled the conference as secretariat. We appreciate the help and understanding of the editorial staff at Springer, and in particular Alfred Hofmann, who supported the publication of these proceedings in the LNAI series.

Last, but not least, we wish to thank all the speakers and participants who contributed to the success of the conference. The next ICDM will take place in Berlin in 2010.

July 2009 Petra Perner

Industrial Conference on Data Mining, ICDM 2009

Chair

Petra Perner	IBaI Leipzig, Germany

Committee

Klaus-Peter Adlassnig	Medical University of Vienna, Austria
Andrea Ahlemeyer-Stubbe	ENBIS, Amsterdam
Klaus-Dieter Althoff	University of Hildesheim, Germany
Chid Apte	IBM Yorktown Heights, USA
Eva Armengol	IIA CSIC, Spain
Bart Baesens	KU Leuven, Belgium
Isabelle Bichindaritz	University of Washington, USA
Leon Bobrowski	Bialystok Technical University, Poland
Marc Boullé	France Télécom, France
Henning Christiansen	Roskilde University, Denmark
Shirley Coleman	University of Newcastle, UK
Juan M. Corchado	Universidad de Salamanca, Spain
Da Deng	University of Otago, New Zealand
Antonio Dourado	University of Coimbra, Portugal
Peter Funk	Mälardalen University, Sweden
Brent Gordon	NASA Goddard Space Flight Center, USA
Gary F. Holness	Quantum Leap Innovations Inc., USA
Eyke Hüllermeier	University of Marburg, Germany
Piotr Jedrzejowicz	Gdynia Maritime University, Poland
Janusz Kacprzyk	Polish Academy of Sciences, Poland
Mehmed Kantardzic	University of Louisville, USA
Ron Kenett	KPA Ltd., Israel
Mineichi Kudo	Hokkaido University, Japan
Eduardo F. Morales	INAOE, Ciencias Computacionales, Mexico
Stefania Montani	Università del Piemonte Orientale, Italy
Jerry Oglesby	SAS Institute Inc., USA
Eric Pauwels	CWI Utrecht, The Netherlands
Mykola Pechenizkiy	Eindhoven University of Technology, The Netherlands
Ashwin Ram	Georgia Institute of Technology, USA
Tim Rey	Dow Chemical Company, USA
Rainer Schmidt	University of Rostock, Germany
Yuval Shahar	Ben Gurion University, Israel
David Taniar	Monash University, Australia

Stijn Viaene	KU Leuven, Belgium
Rob A. Vingerhoeds	Ecole Nationale d'Ingénieurs de Tarbes, France
Claus Weihs	University of Dortmund, Germany
Terry Windeatt	University of Surrey, UK

Table of Contents

Data Mining in Process Control, Industry and Society

Data Mining on Multimedia Data

Theoretical Aspects of Data Mining

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 of 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 with very irregular borders. Then adequate spatial distances 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 of 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 point came from each class; the

P. Perner (Ed.): ICDM 2009, LNAI 5633, pp. 1–12, 2009.

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 valid if the data for each class is similarly distributed, however, nothing prevents from using it if they are not. 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, e.g., like in **Support Vector Machines** (SVM) (Hastie et al., 2001, p. 378) standard transformations were explicitly introduced in classification methods, in order to transform the data so that the images can be separated linearly as with LDA.

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 lead to 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 can then be used for defining the distance between groups of data. Examples are minimum distance between the elements of the groups (**single linkage**), maximum distance (**complete linkage**), and 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).

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.

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 would also be 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 these 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 redefinition 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. Among others, there are the following possibilities to measure the distance of two unconnected cluster parts:

- One could rely on the minimum Euclidean distance between two regions in different parts of the cluster defined by the minimum distance of points $||.||_2$ in the regions (single-linkage distance), or
- one could use the minimum distance measured by the number of borders $||.||_b$ between such regions (cp. Sturtz, 2007).

The former distance $||.||_2$ reflects the idea that region borders could be mainly ignored in the distance definition, or that regions mainly have the form of balls in space. The latter distance reflects the assumption that the regions are thoughtfully fixed and can have all forms not necessarily approximately similar to a ball.

In Figure 1 you can find a typical partition of a city into districts (please ignore the colours and the numbering for the moment). Obviously, the districts have all kinds of forms, not all similar to balls.

In order to define the **dispersion of one cluster** (say d_2 or d_b relying on $||.||_2$ and $||.||_b$, respectively) first define sequences of most neighboured connected parts of the cluster, and then sum up the distances between all sequential pairs. The dispersion may be defined as the minimum such sum over all possible sequences.

Consider the partition in Figure 1 obtained by a clustering algorithm based on social properties of the city districts of Dortmund (cp. Roever and Szepannek, 2005). Is this clustering ready for interpretation? How well are the clusters connected? Ignoring the white regions which were not clustered, Table 1 gives the dispersions d_2 and d_b of the four clusters. As an example, please consider the ordering of the connected parts of the [redacted] - cluster as indicated in Figure 1. Obviously the [redacted] - cluster is very much connected, whereas the other clusters are much more dispersed.

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

Another possible, but simpler, dispersion measure would be the percentage pc of districts in the maximum connected part of a cluster. With this measure, the - cluster is the least dispersed (see Table 1).

Table 1. Dispersion of clusters

cluster	d_2	d_b	p_c
	1.1	4	0.88
	5.9	14	0.83
	6.5	18	0.79
	7.8	13	0.90

In Roever and Szepannek, 2005, dispersion was not utilized for clustering. However, there would be the option to use dispersion as a penalty (or complexity) term for clustering. Roever and Szepannek 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. However, this fitness-function could have been supplemented with a dispersion measure to force non-dispersed clusters. For this, the dispersions for the individual clusters should be combined to one measure D. For ease, one could use the simple measure

D_c = percentage of districts in the maximum connected parts of all clusters.

A possible combination of fitness functions is then $CE - c \cdot D_c$, where $c > 0$ is a constant to be fixed, e.g., so that the two parts of the fitness function are well-balanced.

Another option would be to base optimization on two different fitness functions, CE and D_c, where CE is to be minimized, and D_c to be maximized, and combine them, e.g., by means of a desirability index (cp. section 2.3).

However, for this paper we have tried to take into account the cluster dispersion in a different way. We introduced new variables representing the x- und y-coordinates of the district centres. By this, 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).

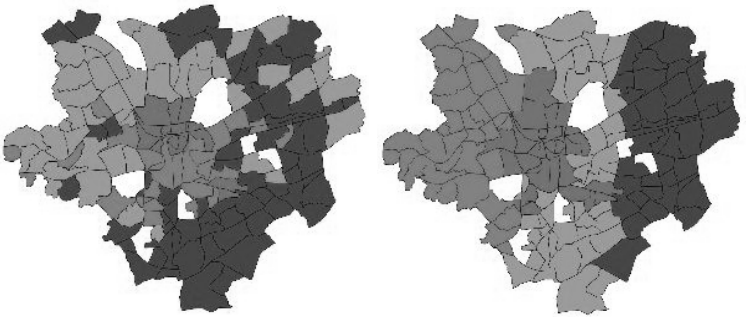

Fig. 2. Clusters with 20%- (left) and 200%- (right) weighting of district centres

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.

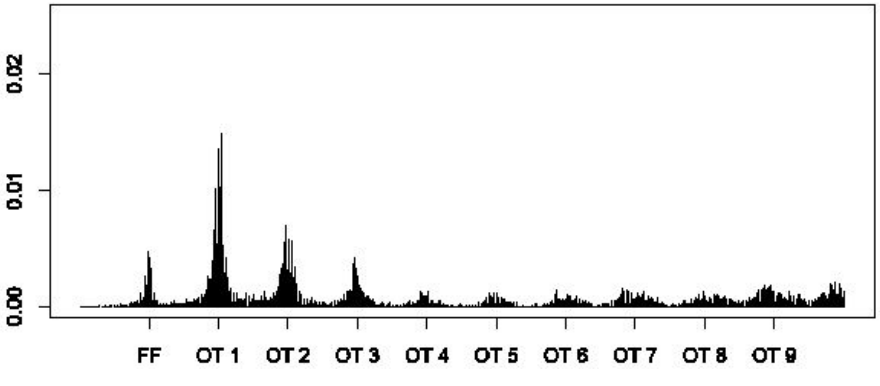

Fig. 3. Pitch independent periodogram (professional bass singer)

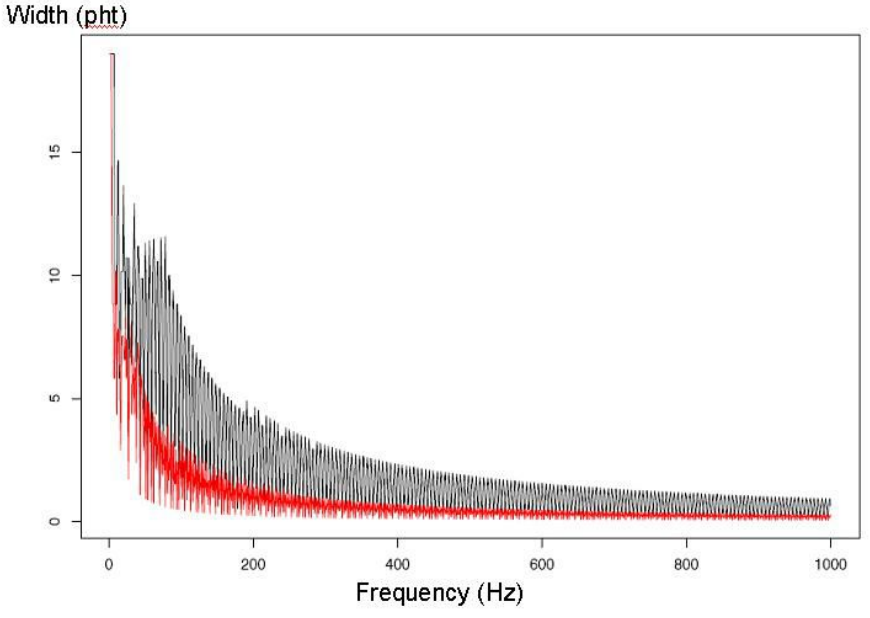

Fig. 4. Width measured in parts of half-tone (pht) dependent on frequency (upper line = fundamental frequency, lower line = first overtone)

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.

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 economic 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:

$$w_1(C^{(k)}) = 1 - \sum_{i=1}^{k} \sum_{X_j, X_l \in C_k} d(X_j, X_l),$$

where $C^{(k)}$ is a partition with k clusters, and $d(X_j, X_l)$ is the machine set-up time between product variants X_j and X_l measured by one of the above distance measures (Jaccard, simple-matching, Euclidean).

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.

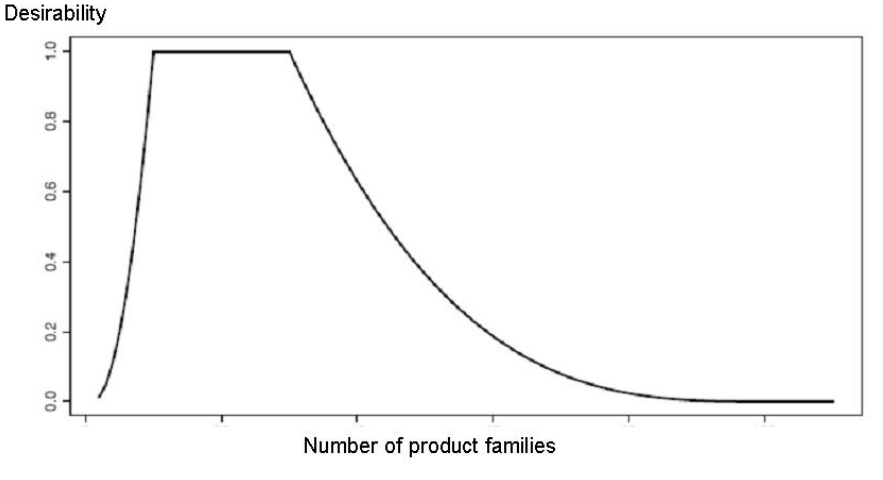

Fig. 5. Desirability function w_2

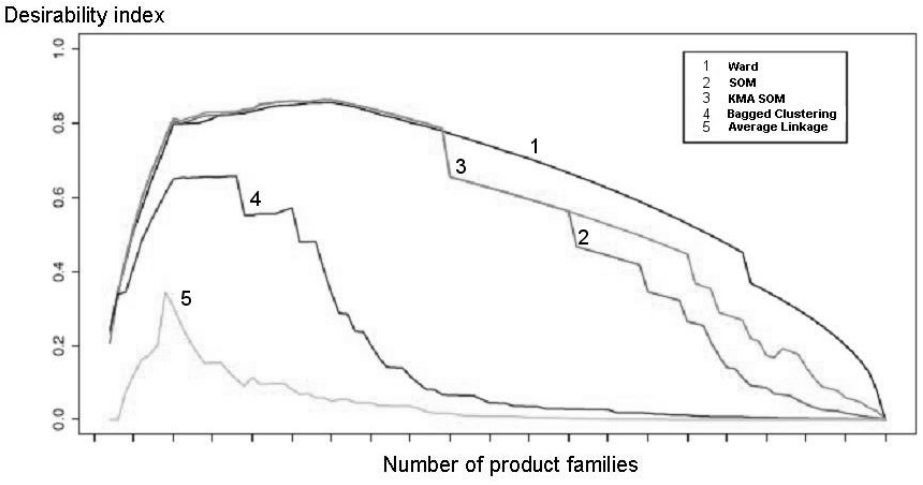

Fig. 6. Desirability index for different cluster methods

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:

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

with

$$a \;=\; \text{number of product families with less or equal five variants.}$$

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