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 Marketing” 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 Transactions 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-Stubble
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
Éric 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
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution and Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stijn Viaene</td>
<td>KU Leuven, Belgium</td>
</tr>
<tr>
<td>Rob A. Vingerhoeds</td>
<td>Ecole Nationale d'Ingénieurs de Tarbes, France</td>
</tr>
<tr>
<td>Claus Weihs</td>
<td>University of Dortmund, Germany</td>
</tr>
<tr>
<td>Terry Windeatt</td>
<td>University of Surrey, UK</td>
</tr>
</tbody>
</table>
# Table of Contents

## Invited Talk
Distances in Classification .................................................. 1  
* Claus Weihs and Gero Szepannek*

## Data Mining in Medicine and Agriculture
Electronic Nose Ovarian Carcinoma Diagnosis Based on Machine Learning Algorithms ................................................. 13  
* José Chilo, György Horvath, Thomas Lindblad, and Roland Olsson*

Data Mining of Agricultural Yield Data: A Comparison of Regression Models ........................................................... 24  
* Georg Ruß*

Study of Principal Components on Classification of Problematic Wine Fermentations ......................................................... 38  
* Alejandra Urtubia U. and J. Ricardo Pérez-Correa*

A Data Mining Method for Finding Hidden Relationship in Blood and Urine Examination Items for Health Check ......................... 44  
* Kazuhiko Shinozawa, Norihiro Hagita, Michiko Furutani, and Rumiko Matsuoka*

Application of Classification Association Rule Mining for Mammalian Mesenchymal Stem Cell Differentiation .......................... 51  
* Weiqi Wang, Yanbo J. Wang, René Bañares-Alcántara, Zhanfeng Cui, and Frans Coenen*

Computer-Aided Diagnosis in Brain Computed Tomography Screening ................................................................. 62  
* Hugo Peixoto and Victor Alves*

Knowledge Representation in Difficult Medical Diagnosis ............. 73  
* Ana Aguilera and Alberto Subero*

## Data Mining in Marketing, Finance and Telecommunication
Forecasting Product Life Cycle Phase Transition Points with Modular Neural Networks Based System ........................................... 88  
* Serge Parshutin, Ludmila Aleksejeva, and Arkady Borisov*
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visualizing the Competitive Structure of Online Auctions</td>
<td>103</td>
</tr>
<tr>
<td>Stephen France and Douglas Carroll</td>
<td></td>
</tr>
<tr>
<td>Credit Risk Handling in Telecommunication Sector</td>
<td>117</td>
</tr>
<tr>
<td>Monika Szczerba and Andrzej Ciemski</td>
<td></td>
</tr>
<tr>
<td>Sales Intelligence Using Web Mining</td>
<td>131</td>
</tr>
<tr>
<td>Viara Popova, Robert John, and David Stockton</td>
<td></td>
</tr>
<tr>
<td>A Sales Forecast Model for the German Automobile Market Based on</td>
<td>146</td>
</tr>
<tr>
<td>Time Series Analysis and Data Mining Methods</td>
<td></td>
</tr>
<tr>
<td>Bernhard Brühl, Marco Hülsmann, Detlef Borscheid, Christoph M. Friedrich, and Dirk Reith</td>
<td></td>
</tr>
<tr>
<td><strong>Data Mining in Process Control, Industry and Society</strong></td>
<td></td>
</tr>
<tr>
<td>Screening Paper Runnability in a Web-Offset Pressroom by Data Mining</td>
<td>161</td>
</tr>
<tr>
<td>A. Alzghoul, A. Verikas, M. Hälander, M. Bacauskiene, and A. Gelzinis</td>
<td></td>
</tr>
<tr>
<td>An Application of Different Data-Mining Techniques</td>
<td></td>
</tr>
<tr>
<td>Fabio Tango and Marco Botta</td>
<td></td>
</tr>
<tr>
<td>SO_MAD: SensOr Mining for Anomaly Detection in Railway Data</td>
<td>191</td>
</tr>
<tr>
<td>Julien Rabatel, Sandra Bringay, and Pascal Poncelet</td>
<td></td>
</tr>
<tr>
<td>Online Mass Flow Prediction in CFB Boilers</td>
<td>206</td>
</tr>
<tr>
<td>Andriy Ivannikov, Mykola Pechenizkiy, Jorn Bakker, Timo Leino,</td>
<td></td>
</tr>
<tr>
<td>Mikko Jegoroff, Tommi Kärkkäinen, and Sami Ayrämö</td>
<td></td>
</tr>
<tr>
<td>Integrating Data Mining and Agent Based Modeling and Simulation</td>
<td>220</td>
</tr>
<tr>
<td>Omar Baqueiro, Yanbo J. Wang, Peter McBurney, and Frans Coenen</td>
<td></td>
</tr>
<tr>
<td>Combining Multidimensional Scaling and Computational Intelligence</td>
<td>232</td>
</tr>
<tr>
<td>for Industrial Monitoring</td>
<td></td>
</tr>
<tr>
<td>António Dourado, Sara Silva, Lara Aires, and João Araújo</td>
<td></td>
</tr>
<tr>
<td>A Case of Using Formal Concept Analysis in Combination with</td>
<td>247</td>
</tr>
<tr>
<td>Emergent Self Organizing Maps for Detecting Domestic Violence</td>
<td></td>
</tr>
<tr>
<td>Jonas Poelmans, Paul Elzinga, Stijn Viaene, and Guido Dedene</td>
<td></td>
</tr>
<tr>
<td><strong>Data Mining on Multimedia Data</strong></td>
<td></td>
</tr>
<tr>
<td>Ordinal Evaluation: A New Perspective on Country Images</td>
<td>261</td>
</tr>
<tr>
<td>Marko Robnik-Šikonja, Kris Brijs, and Koen Vanhoof</td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>Evaluation of Fusion for Similarity Searching in Online Handwritten Documents</td>
<td>276</td>
</tr>
<tr>
<td>Sascha Schimke, Maik Schott, Claus Vielhauer, and Jana Dittmann</td>
<td></td>
</tr>
<tr>
<td>Self-training Strategies for Handwriting Word Recognition</td>
<td>291</td>
</tr>
<tr>
<td>Volkmar Frinken and Horst Bunke</td>
<td></td>
</tr>
<tr>
<td>On a New Similarity Analysis in Frequency Domain for Mining Faces within a Complex Background</td>
<td>301</td>
</tr>
<tr>
<td>D.A. Karras</td>
<td></td>
</tr>
</tbody>
</table>

**Theoretical Aspects of Data Mining**

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering with Domain Value Dissimilarity for Categorical Data</td>
<td>310</td>
</tr>
<tr>
<td>Jeonghoon Lee, Yoon-Joon Lee, and Minho Park</td>
<td></td>
</tr>
<tr>
<td>The Normalized Compression Distance as a Distance Measure in Entity Identification</td>
<td>325</td>
</tr>
<tr>
<td>Sebastian Klenk, Dennis Thom, and Gunther Heidemann</td>
<td></td>
</tr>
<tr>
<td>Attribute Constrained Rules for Partially Labeled Sequence Completion</td>
<td>338</td>
</tr>
<tr>
<td>Chad A. Williams, Peter C. Nelson, and Abolfazl (Kouros) Mohammadian</td>
<td></td>
</tr>
<tr>
<td>Mining Determining Sets for Partially Defined Functions</td>
<td>353</td>
</tr>
<tr>
<td>Dan A. Simovici, Dan Pletea, and Rosanne Vetro</td>
<td></td>
</tr>
<tr>
<td>On the Integration of Neural Classifiers through Similarity Analysis of Higher Order Features</td>
<td>361</td>
</tr>
<tr>
<td>D.A. Karras and B.G. Mertzios</td>
<td></td>
</tr>
<tr>
<td>On Cellular Network Channels Data Mining and Decision Making through Ant Colony Optimization and Multi Agent Systems Strategies</td>
<td>372</td>
</tr>
<tr>
<td>P.M. Papazoglou, D.A. Karras, and R.C. Papademetriou</td>
<td></td>
</tr>
<tr>
<td>Responsible Data Releases</td>
<td>388</td>
</tr>
<tr>
<td>Sanguthevar Rajasekaran, Ofer Harel, Michael Zuba, Greg Matthews, and Robert Aseltine</td>
<td></td>
</tr>
</tbody>
</table>

**Author Index** | 401 |
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
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 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. 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, 2003). Is this clustering ready for interpretation? How well are the clusters connected? Ignoring the white regions which were not clustered, Table II 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 \[\text{---} \] - cluster as indicated in Figure 1. Obviously the \[\text{---} \] - cluster is very much connected, whereas the other clusters are much more dispersed.

![Fig. 1. Clusters of districts of the City of Dortmund (Germany)](image-url)
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>
<thead>
<tr>
<th>cluster</th>
<th>$d_2$</th>
<th>$d_0$</th>
<th>$p_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.1</td>
<td>4</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>5.9</td>
<td>14</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>6.5</td>
<td>18</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>7.8</td>
<td>13</td>
<td>0.90</td>
</tr>
</tbody>
</table>

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 $N = \text{number of observations}$, $k = \text{number of clusters}$, $u_{ij} = \text{probability that observation } i \text{ belongs to cluster } j$. 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

$$D_c = \text{percentage of districts in the maximum connected parts of all clusters}.$$

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 13partials, 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)
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 re-defined 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...
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.

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

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


Electronic Nose Ovarian Carcinoma Diagnosis Based on Machine Learning Algorithms

José Chilo¹, György Horvath², Thomas Lindblad³, and Roland Olsson⁴

¹ Center for RF Measurement Technology, University of Gävle, S-801 76 Gävle, Sweden
² Department of Oncology, Sahlgrenska University Hosp, Gothenburg, Sweden
³ Department of Physics, Royal Institute of Technology, S-106 91 Stockholm, Sweden
⁴ Department of Computer Science, Ostfold University College, N-1757 Halden, Norway

Abstract. Ovarian carcinoma is one of the most deadly diseases, especially in the case of late diagnosis. This paper describes the result of a pilot study on an early detection method that could be inexpensive and simple based on data processing and machine learning algorithms in an electronic nose system. Experimental analysis using real ovarian carcinoma samples is presented in this study. The electronic nose used in this pilot test is very much the same as a nose used to detect and identify explosives. However, even if the apparatus used is the same, it is shown that the use of proper algorithms for analysis of the multisensor data from the electronic nose yielded surprisingly good results with more than 77% classification rate. These results are suggestive for further extensive experiments and development of the hardware as well as the software.

Keywords: Machine learning algorithms, odor classification, ovarian carcinoma, medicine.

1 Introduction

Although most people would agree on the fact that there is no “artificial nose” [1] with the same sensitivity as that of a dog. It is claimed that a dog can detect less than 100 molecules per cubic meter. A sophisticated gas chromatograph with the proper injection system can maybe detect a 1000 molecules. However, it is obvious that there are advantages with a small and simple electronic device, even if its performance is not as good. A few years ago the authors presented an electronic nose for detecting explosives. This nose could in no way compete with trained “bomb dogs” to sense the presence of explosives, but it could distinguish between various types of chemicals (alcohols) and explosives [2–4]. The nose is relatively small and is shown in Fig. 1. The square tube holds four types of sensors, each operating at four different temperatures (ranging from 20 – 750 deg C). The sensors are mounted on the sides of the tube with pertinent electronics and support system directly on a printed circuit board. The
The electronic nose with its holder (circle to the left), the square tube with four sensors and electronics on each side and the exhaust fan to the right

fan and exhaust tube are shown to the right in Fig. 1. A small modification has been made to carry out this study: it is a combined holder and inlet (made of teflon an easily replaceable) shown to the left with a circle.

Many cancers are detected in late stages with consequential high mortality rates. For example, the majority of human ovarian carcinomas are diagnosed in stage III or IV, and 70% of these patients will die within five years. Thus, it is essential to develop inexpensive and simple methods for early diagnosis.

It has been proposed, and also to some extent demonstrated, that electronic noses could be used in medical diagnostics [1]. Most recently, it has also been shown that a dog can be trained to distinguish between histopathological types and grades of ovarian carcinomas, including borderline tumors, from healthy control samples [5]. This study clearly demonstrated that human ovarian carcinomas may be characterised by a specific odor. To detect this odor, a trained dog was used. However, dogs may be influenced by several factors before and during their work, leading to changes in the accuracy rates. Use of an artificial nose with adequate sensitivity should limit the possible diagnostic errors and might be developed for early diagnosis. We will investigate here whether our electronic nose could carry out such a task.

2 Experimental Details

The electronic nose consists of four different types of gas-sensors from Figaro [3]. The four types of sensors are: TGS2600 different “air pollutions”, TGS2610 combustible gases, TGS2611 combustible gases and TGS2620 alcohols of different kinds. The gas-sensors work by the changing resistance induced by gas molecules that bind on two electrodes that are separated a small distance. The response of each individual sensor is strongly influenced by temperature, so the Nose is actually an assembly of four individuals of the four above sensor types, each held at a different internal temperature, producing in total 16 independent analog signals, each varying with time and smell [4]. The gas-sensors, which change conductivity upon exposure to the odor under test, are used in a voltage divider circuit to provide and output voltage which increases with gas concentration. The voltage is sampled at 10 Hz and the raw data
file is compacted to a data matrix consisting of 16 columns by 500 rows. Hence, what we measure is simply the slowly varying DC voltage. The measurement starts a few seconds before the sample is being analyzed, since (as shown in the case of the bomb nose) the rise time of the signals from the sensors holds much information.

In the first study pieces of human seropapillary adenocarcinomas were used, these were clearly smaller (a few millimeter) than in the case of explosives (a few centimeter). Tumors were taken from three different individuals and sized to 3x3x3 mm. Two similar sizes of human tuba obtained from two different healthy individuals were also used as control samples.

In the second study, tumors from 6 different individuals were divided into 21 pieces and analyzed by 92 runs. Myometrium control samples from 3 different individuals were divided into 23 pieces and analyzed by 70 runs.

![Fig. 2. Signals from the 16 sensors of the electronic nose. The horizontal axis is the time scale and the vertical scale is the DC voltage. Cases are shown for healthy tissue (up) and for tissue diagnosed as cancer (bottom).](image)
Both ovarian cancer samples and controls were kept at -80°C in our tumor bank (Ethical Committee license number: S-154-02). Samples were thawed to room temperature for 15-30 minutes before being used.

As mentioned, the primary data are the 16 signals from the various sensors (4 different sensors operating at 4 different temperatures each) as digitized and stored for each sample. An example of the 16 sensor outputs when the array is exposed to healthy tissue and to tissue diagnosed as cancer are shown in Fig. 2.

3 Evaluation of the Primary Signals with WEKA

In this work we use the Waikato Environment for Knowledge Analysis (WEKA) [6]. This is an open source data mining toolbox (written in Java) developed by Ian Witten’s group at the University of Waikato. It provides tools for all the tasks usually performed in data mining, including numerous algorithms for pre-processing, classification, regression and clustering.

Here we utilized 19 different classification algorithms in WEKA. We used their default parameters unless otherwise stated. These algorithms are grouped into five groups in WEKA according to the models they create. Below we give a brief summary of these and some pertinent references.

Bayes includes algorithms where learning results in Bayesian models. In our study we use BayesNet and NaiveBayes algorithms. NaiveBayes is an implementation of the standard naïve Bayes algorithm, where normal distribution is for numerical features. BayesNet creates a Bayesian Network with the ability to represent the same model as NaiveBayes or other more complex models where the independence between features is not assumed.

Lazy is comprised of algorithms that delay construction of classifiers until classification time. The IB1, IBK, Kstar and LWL algorithms were used in this work. IB1 is a nearest-neighbor algorithm that classifies an instance according to the nearest neighbor identified by the Euclidean distance as defined in [7]. IBK is similar to IB1 except that the K- nearest neighbors is used instead of only one. We determined the appropriate number of neighbors using leave-one-out cross-validation. The Kstar algorithm uses entropic distance measure, based on the probability of transforming one instance into another by randomly choosing between all possible transformations [8] and turns out to be much better than Euclidean distance for classification. The LWL (Locally weighted learning) algorithm differs from the other three algorithms in using only a nearest-neighbor algorithm to weight the instances in the training set before applying another classification algorithm to them.

Rules contains methods which creates classification rules. We use NNge, JRip, Ridor and PART algorithms. NNge is a nearest-neighbor algorithm which learns rules based on the hyper rectangles that it divides the instance space into [9]. JRip is an implementation of Cohen’s RIPPER. Ridor creates first a default rule and then recursively develops exceptions to it and PART constructs rules based on partial decision trees.
Functions are algorithms that can be represented mathematically. In this work we use MultilayerPerceptron, RBFNetwork, SimpleLogistic and SMO algorithms. RBFNetwork is an implementation of radial basis functions, and SimpleLogistic constructs linear logistic regression models. SMO is a sequential minimum optimization algorithm for building Support Vector Machine (SVM) [10]. We used a polynomial kernel, which is default in WEKA.

Trees includes algorithms that creates trees as models. The ADTree and J48 algorithms were used in this study. The ADTree is similar to options trees and the J48 is an implementation of the popular C4.5 [11].

Miscellaneous contains simply the rest of algorithms that do not fit into any of the other groups. VFI, which was used in our work, finds intervals for each feature, and attributes each class according to number of instances with the class in the training set for the specific interval. Voting is used to select the final class for an instance.

In this paper, the following features were used to form a feature vector that in total has 48 components as inputs to the classifiers: transient slope (TS), saturation slope (SS) and maximum slope (MS) when the sample is closed of each sensor.

In Table 1 we give the results for 24 runs (15 cancer tissues and 9 healthy tissues) from the first experiment. In Table 2 we give the results for 162 runs (92 cancer tissues and 70 healthy tissues) from the second experiment. We used ten-fold cross validation in our experiments, which means that each dataset was divided into ten equal sized folds and ten independent runs of each algorithm were conducted for each dataset. For the ith run, the ith fold was designated as the test set and the patterns in the remaining nine folds were used for training. At the end of training the classifier’s generalization was measured on the test set.

| Table 1. Classification results |

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Cancer tissues</th>
<th>Healthy tissues</th>
<th>Total (%) Correctly Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes Network</td>
<td>11/15</td>
<td>2/9</td>
<td>54</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>15/15</td>
<td>6/9</td>
<td>88</td>
</tr>
<tr>
<td>Multilayer Perceptron</td>
<td>9/15</td>
<td>4/9</td>
<td>54</td>
</tr>
<tr>
<td>RBF Network</td>
<td>15/15</td>
<td>5/9</td>
<td>83</td>
</tr>
<tr>
<td>SimpleLogistic</td>
<td>11/15</td>
<td>2/9</td>
<td>54</td>
</tr>
<tr>
<td>SMO</td>
<td>9/15</td>
<td>5/9</td>
<td>58</td>
</tr>
<tr>
<td>IB1</td>
<td>11/15</td>
<td>8/9</td>
<td>79</td>
</tr>
<tr>
<td>KNN</td>
<td>11/15</td>
<td>8/9</td>
<td>79</td>
</tr>
<tr>
<td>KStar</td>
<td>13/15</td>
<td>7/9</td>
<td>83</td>
</tr>
<tr>
<td>LWL</td>
<td>14/15</td>
<td>6/9</td>
<td>83</td>
</tr>
<tr>
<td>ClassificationVia Regression</td>
<td>13/15</td>
<td>7/9</td>
<td>83</td>
</tr>
<tr>
<td>ThresholdSelector</td>
<td>12/15</td>
<td>7/9</td>
<td>89</td>
</tr>
<tr>
<td>VFI</td>
<td>11/15</td>
<td>8/9</td>
<td>79</td>
</tr>
<tr>
<td>ADTree</td>
<td>13/15</td>
<td>7/9</td>
<td>83</td>
</tr>
<tr>
<td>J48</td>
<td>15/15</td>
<td>5/9</td>
<td>83</td>
</tr>
<tr>
<td>JRip</td>
<td>15/15</td>
<td>8/9</td>
<td>95</td>
</tr>
<tr>
<td>NNge</td>
<td>14/15</td>
<td>3/9</td>
<td>71</td>
</tr>
<tr>
<td>PART</td>
<td>15/15</td>
<td>5/9</td>
<td>83</td>
</tr>
<tr>
<td>Ridor</td>
<td>12/15</td>
<td>5/9</td>
<td>71</td>
</tr>
</tbody>
</table>
Table 2. Classification results

<table>
<thead>
<tr>
<th>Model</th>
<th>Correctly Classified (% of Cancer tissues)</th>
<th>Correctly Classified (% of Healthy tissues)</th>
<th>Total (%) Correctly Classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes Network</td>
<td>83/92</td>
<td>66/70</td>
<td>92</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>84/92</td>
<td>63/70</td>
<td>91</td>
</tr>
<tr>
<td>Multilayer Perceptron</td>
<td>82/92</td>
<td>61/70</td>
<td>88</td>
</tr>
<tr>
<td>RBF Network</td>
<td>81/92</td>
<td>57/70</td>
<td>85</td>
</tr>
<tr>
<td>Simple Logistic</td>
<td>86/92</td>
<td>64/70</td>
<td>93</td>
</tr>
<tr>
<td>SMO</td>
<td>84/92</td>
<td>62/70</td>
<td>91</td>
</tr>
<tr>
<td>IB1</td>
<td>75/92</td>
<td>50/70</td>
<td>77</td>
</tr>
<tr>
<td>KNN</td>
<td>75/92</td>
<td>50/70</td>
<td>77</td>
</tr>
<tr>
<td>KStar</td>
<td>81/92</td>
<td>61/70</td>
<td>88</td>
</tr>
<tr>
<td>LWL</td>
<td>82/92</td>
<td>58/70</td>
<td>86</td>
</tr>
<tr>
<td>ClassificationVia Regression</td>
<td>83/92</td>
<td>59/70</td>
<td>88</td>
</tr>
<tr>
<td>ThresholdSelector</td>
<td>77/92</td>
<td>57/70</td>
<td>83</td>
</tr>
<tr>
<td>VFI</td>
<td>72/92</td>
<td>61/70</td>
<td>82</td>
</tr>
<tr>
<td>ADTree</td>
<td>84/92</td>
<td>66/70</td>
<td>93</td>
</tr>
<tr>
<td>J48</td>
<td>79/92</td>
<td>53/70</td>
<td>81</td>
</tr>
<tr>
<td>JRip</td>
<td>78/92</td>
<td>62/70</td>
<td>86</td>
</tr>
<tr>
<td>NNge</td>
<td>83/92</td>
<td>61/70</td>
<td>89</td>
</tr>
<tr>
<td>PART</td>
<td>79/92</td>
<td>65/70</td>
<td>89</td>
</tr>
<tr>
<td>Ridor</td>
<td>79/92</td>
<td>59/70</td>
<td>85</td>
</tr>
</tbody>
</table>

4 Evaluation of the Data with the ADATE Code

4.1 A Brief Introduction to ADATE

Automatic Design of Algorithms through Evolution (ADATE) [12] is a system for general automatic programming in a first order, purely functional subset of Standard ML. ADATE can synthesize recursive programs for standard algorithm design problems such as sorting, searching, string processing and many others. It has also successfully been used to generate programs for more advanced tasks such as segmentation of noisy images [13] and driving a robot car.

However, ADATE is also well suited to a more traditional machine learning problem such as analyzing data from an electronic nose to diagnose cancer and offers several advantages in comparison with the standard methods in the WEKA toolbox, such as a better ability to find compact and yet descriptive models.

The models generated by ADATE are formulated in a general programming language which is more expressive than any of the various formalisms used in WEKA discussed above. This means that programs generated by ADATE may be more compact than any of the WEKA models. Compact and still accurate models are important both to avoid overfitting, to enhance readability and above all to give clues for further optimization and redesign of the electronic nose so that it becomes better at cancer detection.

ADATE maintains a hierarchically structured so-called “kingdom of programs”. The most basic principle used to organize the kingdom is that each program must be better than all smaller ones found so far. Thus, ADATE generates a progression of
gradually bigger and more accurate programs, where each program is optimized over and over again to be the best for its size on the training data.

Program transformations in varying combinations are employed to produce new programs that become candidates for insertion into the kingdom. The search for program transformations is mostly systematic and does not rely on randomization for purposes other than introducing new floating point constants.

ADATE has the ability to define new auxiliary functions “on-the-fly”. However, the effectiveness of its program synthesis may strongly depend on the set of predefined functions that it is allowed to use. For the experiments reported in this paper, we included addition, multiplication, subtraction and division of floating point numbers in this set and also the hyperbolic tangent function $\tanh$ that is commonly used in neural networks.

Since ADATE is able to effectively introduce and optimize floating-point constants on its own, there was no need to include any special, predefined constants.

The above set of predefined functions is a superset of what is needed to implement standard feed-forward neural networks with any number of hidden layers, which can express quite good approximations to any non-linear function [14]. Therefore, the same result holds for our evolved programs.

In practice, however, the limiting factor for most neural and evolutionary computation techniques is not the theoretical expressiveness of the languages that they employ but their ability to avoid entrapment in local optima in the search space. Another key limiting factor is overfitting. We believe that ADATE excels at both reducing overfitting and avoiding local optima, but we do not have space here for a discussion of the many mechanisms employed to do so [12].

4.2 ADATE Experiments for Analyzing Data from the Electronic Nose

Given that the e-nose employs four types of sensors and that each sensor is measured at four different temperatures as discussed above, we have a total of 16 time series for each sample. Each time series was converted to three parameters as described above, giving a total of 48 floating point inputs to a program to be generated by ADATE. These inputs are called $TS_0$, $TS_1$, ..., $TS_{15}$, $SS_{16}$, $SS_{17}$, ..., $SS_{31}$, $MS_{32}$, $MS_{33}$, ..., $MS_{47}$.

We first conducted a simple ADATE run where half of the patients, randomly chosen, were used for training and the other half for testing. Thus, we had 81 patients in the training set and 81 in the test set and obtained overfitting characteristics as shown in Fig. 3. The horizontal axis in that figure shows the size of the programs, measured in bits, whereas the vertical axis shows the number of correctly classified instances for training and test data respectively.

When just saying that all patients in the training set have cancer, we obtain 46 correct classifications, but when instead classifying according to the following simple rule, we suddenly obtain 71 correctly classified patients.

\[
\text{if } SS_{23} < SS_{16} \text{ then healthy else cancer}
\]
If a small increase in model complexity gives a big increase in classification accuracy, we typically have a change in the model without any overfitting. In other words, if a small amount of extra theory can explain many more observations, that extra theory is believed to be generally valid. As can be seen in Fig. 3, there is an equally big jump in accuracy for both training and testing data when moving to the simple rule above, which has a size of about 34 bits according to ADATE’s built in syntactic complexity measure.

The rule above correctly classifies 71 out of 81 training cases and 70 out of 81 test cases, giving an accuracy of 86.4% on the test data and a 95% confidence interval between 77% and 94%. Note that WEKA was run using ten-fold cross validation, which means that 90% of the data were used for training instead of only 50% as in the ADATE experiments. But even if ADATE was given much less training data, it still created results comparable with those of WEKA given in Table 2 and additionally a very simple model that is easy to understand and use for optimization of the Enose.

5 A Pilot Gas Chromatography Experiment

To show that there really is a difference between the healthy and the cancer sample, an extended gas chromatography plus mass spectroscopy has been initiated [15]. The preliminary results of this study will be published soon and here we only present two spectra to show that there is a significant difference between the samples (Fig. 4).
Fig. 4. It is clearly seen from the chromatogram obtained from the healthy sample tissue (upper chromatogram) and one from the cancer tissue that there are differences in the occurrence of certain peaks as well as in their intensities. This indicates that there is a reason for the electronic nose to react differently to the different tissues.

6 Summary, Conclusions and Future Work

The hardware in the present investigation is the same as in the case of a bomb nose. However, the feature extraction and analysis is different. In the first case we simply used rise times and saturation points and the PCA approach to define the regions of interest. In the present case we have tested several algorithms, e.g. the WEKA ones. Hence, from Table 1 we may possibly conclude that the “best” algorithm is JRip. This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER). As we can see 68% of the used machine learning algorithms classifies correctly at least 79%. When we extend the study to include 162 tests, the “best” algorithms are SimpleLogistic and ADTree with 93% correctly classified.

The results from the ADATE test are interesting and suggestive. They tell us which sensors operated at which temperature are important. Hence some sensors of the original bomb nose may be changed and a more redundant and efficient system could be designed.

The results show that the proposed method, although simple and inexpensive, is probably a rather efficient ovarian carcinoma identification system. It should be stressed again that the sensors are probably not optimal for the present samples. This means that we need to study the results further. We need to test on several more