Petra Perner (Ed.)

Advances in Data Mining

Applications and Theoretical Aspects

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

LNAI 5633

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

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

Electronic Nose Ovarian Carcinoma Diagnosis Based on Machine Learning Algorithms

José Chilo[1], György Horvath[2], Thomas Lindblad[3], and Roland Olsson[4]

[1] Center for RF Measurement Technology, University of Gävle, S-801 76 Gävle, Sweden
jco@hig.se
[2] Department of Oncology, Sahlgrenska University Hosp. Gothenburg, Sweden
gyorgy.horvath@oncology.gu.se
[3] Department of Physics, Royal Institute of Technology, S-106 91 Stockholm, Sweden
lindblad@particle.kth.se
[4] Department of Computer Science, Ostfold University College, N-1757 Halden, Norway
Roland.Olsson@hiof.no

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

Both ovarian cancer samples and controls were kept at -80^{0}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 *i*th run, the *i*th 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

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

Table 2. Classification results

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

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 hyberbolic 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.

if $SS_{23} < SS_{16}$ *then healthy else cancer*

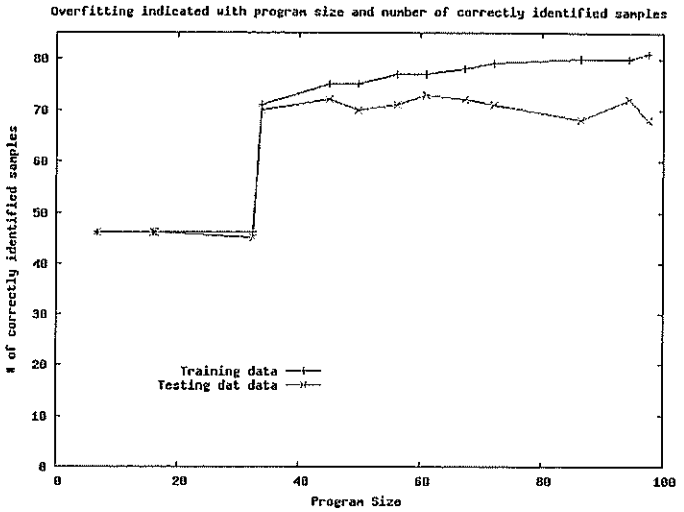

Fig. 3. Overfitting shown with program size and number of correctly identified samples in the training and testing data set

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).

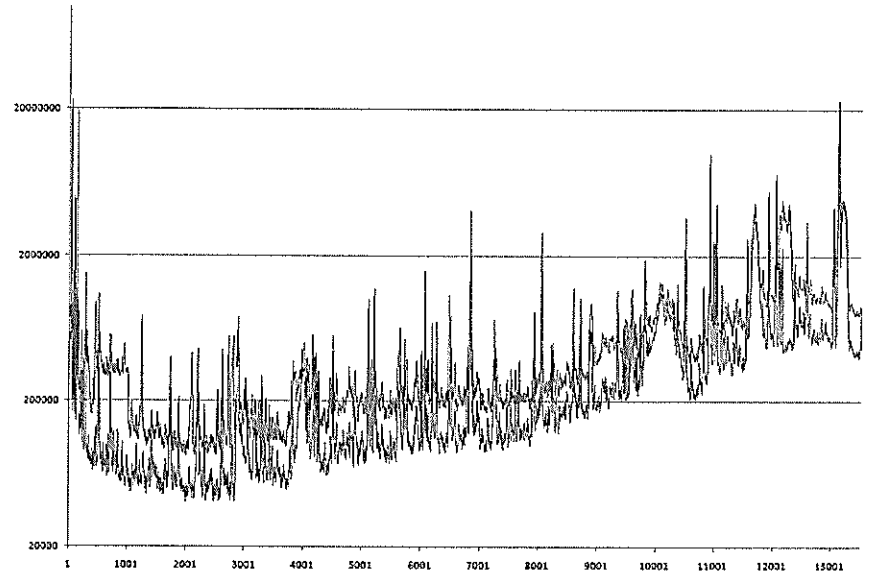

Fig. 4. It is clearly seen from the chromatogram obtained from the healthy sample tissue (upper chromatogram) and one from the cancer tissue that there are differences in the occurrence of certain peaks as well as in their intensities. This indicates that there is a reason for the electronic nose to 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

cases with tissues of the same character to see if there is a difference between healthy tissue-samples. Improved multivariate analysis combined a sophisticated gas chromatography and mass spectroscopy test is a logical next step. Although such spectra will most likely show hundreds of lines, it may give hints on which sensors to use. We further need to see in detail if there are any systematic effects. Again, it would be desirable to study the effects on the confusion matrices and to reduce the errors on the "healthy tissues" even if the "cancer tissues" will yield a larger uncertainty. We tried reducing the dimensionality of the inputs using PCA, with varying numbers of principal components, but this did not yield any increase in the classification accuracy obtained with WEKA. However, it would be quite interesting to try some non-linear dimensionality reduction algorithms such as Isomap or auto-encoders optimized with ADATE. Even if one needs to elaborate on modifications of the present e-nose system, one should, indeed, recall that it was originally designed to detect various explosives. From this point of view the "nose is doing quite well".

References

1. Wang, P., Chen, X., Xu, F.: Development of Electronic Nose for Diagnosis of Lung Cancer at Early Stage. In: 5th Conference on Information Technology and Application in Biomedicine, Shenzhen, China, pp. 588–591 (2008)
2. Roppel, T., Dunman, K., Padgett, M., Wilson, D., Lindblad, Th.: Feature-level signal processing for odor sensor array. In: Proc. Ind. Electronics Conf. IECON 1997, IEEE catalog No 97CH36066, pp. 212–221 (1997)
3. Waldemark, J., Roppel, T., Padgett, M., Wilson, D., Lindblad, Th.: Neural Network and PCA for Determining Region of Interest in Sensory Data Pre-processing. In: Virtual Intelligence/Dynamic Neural Networks Workshop 1998 SPIE, vol. 3728, pp. 396–405 (1998)
4. Kermit, M., Eide, Å.J., Lindblad, Th., Agehed, K.: Intelligent Machine Olfaction, IASTED. In: Int. Conf. on Artificial Intelligent Machine Olfaction and Computational Intelligence (ACI 2002), Tokio, Japan, pp. 25–27 (2002)
5. Horvath, G., Järverud, G.K.: Human ovarian carcinomas detected by specific odor. Integr. Cancer Ther. 7(2), 76–80 (2008)
6. Witten, I.H., Frank, E.: Data mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann Publishers, San Mateo (2005)
7. Aha, D.W.: Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms. International Journal of Man-Machine Studies 36(2), 267–287 (1992)
8. Cleary, J.G., Trigg, L.E.: K*: and instance-based learner using an entropic distance measure. In: Proceedings of the 12th International Conference on Machine Learning, pp. 108–114 (1995)
9. Cohen, W.W.: Fast Effective Rule Induction. In: Proceedings of the 12th International Conference on Machine Learning, pp. 115–123 (1995)
10. Platt, J.C.: Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. In: NIPS conference, pp. 557–563 (1999)
11. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo (1993)

12. Olsson, R.: Inductive functional programming using incremental program transformation. Artificial Intelligence 1, 55–83 (1995)
13. Berg, H., Olsson, R., Lindblad, Th.: Automatic Design of Pulse Coupled Neurons for Image Segmentation. Neurocomputing - Special Issue for Vision Research (2008)
14. Mitchell, T.: Machine Learning. McGraw-Hill Companies, Inc., New York (1997)
15. Chilo, J., Horvath, G., Lindblad, Th., Olsson, R., Redeby, J., Roeraade, J.: A Flexible Electronic Nose for Ovarian Carcinoma Diagnosis in Real Time. Accepted to IEEE NPSS Real Time Conference, Beijing, China (2009)