View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by MURAL - Maynooth University Research Archive Library

Measurement and Control

http://tim.sagepub.com/

Predicting organic acid concentration from UV/vis spectrometry measurements - A comparison of machine learning techniques

Christian Wolf, Daniel Gaida, André Stuhlsatz, Thomas Ludwig, Seán McLoone and Michael Bongards

Transactions of the Institute of Measurement and Control published online 19 September 2011
DOI: 10.1177/0142331211403797

The online version of this article can be found at:
http://tim.sagepub.com/content/early/2011/09/14/0142331211403797

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Institute of Measurement and Control
The Institute of Measurement and Control

Additional services and information for *Transactions of the Institute of Measurement and Control* can be found at:

Email Alerts: http://tim.sagepub.com/cgi/alerts

Subscriptions: http://tim.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

>> OnlineFirst Version of Record - Sep 19, 2011

What is This?

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

Transactions of the Institute of
Measurement and Control
0(0) 1–11
© The Author(s) 2011
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0142331211403797
tim.sagepub.com

Predicting organic acid concentration from UV/vis spectrometry measurements – A comparison of machine learning techniques

Christian Wolf[1], Daniel Gaida[2], André Stuhlsatz[3],
Thomas Ludwig[2], Seán McLoone[1] and Michael Bongards[2]

Abstract

The concentration of organic acids in anaerobic digesters is one of the most critical parameters for monitoring and advanced control of anaerobic digestion processes. Thus, a reliable online-measurement system is absolutely necessary. A novel approach to obtaining these measurements indirectly and online using UV/vis spectroscopic probes, in conjunction with powerful pattern recognition methods, is presented in this paper. An UV/vis spectroscopic probe from S::CAN is used in combination with a custom-built dilution system to monitor the absorption of fully fermented sludge at a spectrum from 200 to 750 nm. Advanced pattern recognition methods are then used to map the non-linear relationship between measured absorption spectra to laboratory measurements of organic acid concentrations. Linear discriminant analysis, generalized discriminant analysis (GerDA), support vector machines (SVM), relevance vector machines, random forest and neural networks are investigated for this purpose and their performance compared. To validate the approach, online measurements have been taken at a full-scale 1.3-MW industrial biogas plant. Results show that whereas some of the methods considered do not yield satisfactory results, accurate prediction of organic acid concentration ranges can be obtained with both GerDA and SVM-based classifiers, with classification rates in excess of 87% achieved on test data.

Keywords

Anaerobic digestion, classification, feature extraction, GerDA, LDA, neural networks, online measurement, organic acids, random forest, RVM, SVM, UV/vis spectroscopy

1. Introduction

The use of powerful computational intelligence and data analysis methods in conjunction with new and existing online-measurement and advanced control systems allows the development of highly sophisticated and robust systems for efficient process monitoring and optimization. There is a vast range of applications, for example artificial neural networks for modelling and prediction purposes, fuzzy control to include expert knowledge in plant operation, genetic algorithms for the optimization of complex processes, and machine learning methods to detect critical operation states or to further process online-measured information in so-called soft sensors (Ozkaya et al., 2007; Puñal et al., 2002; Steyer et al., 2002; Strik et al., 2005). The application of feature extraction and classification methods to predict organic acid concentrations in anaerobic digestion processes using UV/vis spectroscopic probes is an example of such a hybrid system.

There are many industrial applications for organic acid measurements. In addition to being one of the most important parameters in anaerobic digestion processes, it is commonly used to monitor the quality of beer and wine and inhibition in composting processes (Batista et al., 2010; Cheung et al. 2010; Rodrigues et al., 2010). In particular, monitoring and control of anaerobic digestion in biogas plants has proven to be extremely difficult because of a lack of robust and feasible online-measurement systems and the high non-linearity of anaerobic digestion processes. Nevertheless, it becomes more important than ever to offer solutions for advanced process control of biogas plants, because efficient plant operation is the major issue when it comes to feasible long-term operation of such plants. The monitoring of organic acid concentrations is one of the key

[1] Department of Electronic Engineering, National University of Ireland Maynooth, Co. Kildare, Maynooth, Ireland
[2] Institute of Automation and Industrial IT, Cologne University of Applied Sciences, Steinmüllerallee, Gummersbach, Germany
[3] Institute for Information Technology, Department of Mechanical and Process Engineering, University of Applied Sciences Düsseldorf, Josef-Gockeln-Str., Düsseldorf, Germany

Corresponding author:
Christian Wolf, Institute of Automation and Industrial IT, Cologne University of Applied Sciences, Steinmüllerallee 1, 51643 Gummersbach, Germany
Email: christian.wolf@fh-koeln.de

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

parameters for process stability in biogas plants, as high concentrations leads to acidification of the anaerobic biology and consequently a collapse of biogas production. Its robust measurement is required for long-term monitoring to recognize critical process states in time. Furthermore, a detailed process monitoring regime is an excellent basis for the development and testing of innovative optimization and control strategies for anaerobic digestion processes.

There are several online-measurement systems for organic acid concentrations available on the market, such as online capillary gas chromatography and online spectro-fluorimetric systems, which are widely used by the chemical industry. Nevertheless, these systems are too complex and sensitive to disturbances to be used efficiently at agricultural biogas plants (Diamantis et al., 2010; Palacio-Barco et al., 2010). Calibration and maintenance of such systems require a high degree of expert knowledge, which operators normally do not have. Furthermore, because of their high purchase, maintenance and repair costs, these systems are not feasible for most agricultural biogas plants.

The availability of UV/vis spectroscopic probes offers a new approach to measuring organic acid concentrations indirectly and online. By employing powerful feature extraction and classification methods, organic acid concentrations can be predicted from the absorption spectra measurements taken from diluted fermentation sludge. As the use of UV/vis spectroscopic probes is well established in the wastewater sector for the online measurement of the chemical oxygen demand (COD) in sewage systems and wastewater treatment plants, these probes have proven to be extremely robust, requiring less maintenance than the alternatives mentioned above (Bongards et al., 2007).

Two factors support the use of pattern recognition methods for this type of application: 1) organic acid concentrations can be divided into different concentration ranges/classes, which correspond to different process conditions; 2) a high precision measurement of organic acid concentrations is not necessary, as the determination of concentration ranges is sufficient for plant operation. Furthermore, the chosen concentration ranges/classes can be easily applied to the development of fuzzy control systems.

In this paper we consider the well-known linear discriminant analysis (LDA) and the generalized discriminant analysis (GerDA), which is a novel and powerful extension of the classical LDA algorithm (Stuhlsatz et al., 2010a), to extract features automatically from the raw UV/vis spectrogram measurements. In addition to these feature extractors, we use linear classifiers to classify the extracted features into different concentration ranges. For comparison, we also investigate the use of random forest (RF), neural networks (multilayer perceptron, MLP), support vector machines (SVM) and relevance vector machines (RVM), which have proven to be very efficient methods for multi-class classification (Balabin et al., 2011; Guo et al. 2011; Wang et al., 2010; Yogameena et al., 2010). RF is used for feature selection and classification, whereas MLPs are used for the classification of reduced feature spaces created by applying partial least squares regression (PLS), forward selection regression (FSR) and GerDA to the raw measurements. Finally, SVMs and RVMs are investigated for direct classification of the spectrogram as well as for classification of the GerDA and RF features.

Table 1 Definition of the class labels and the number of samples in each class ϑ for the complete (N_ϑ), training ($N_{T,\vartheta}$) and validation dataset, $N_{V,\vartheta}$

Class $\vartheta \in \Theta$	Organic acid concentration $c_a[g/l]$	N_ϑ	$N_{T,\vartheta}$	$N_{V,\vartheta}$
1 (low)	1.1,. . ., 1.4	228	171	57
2 (low-normal)	1.5,. . ., 1.8	1528	1146	382
3 (normal)	1.9,. . ., 2.2	1880	1410	470
4 (normal-high)	2.3,. . ., 2.6	731	549	182
5 (high)	2.7,. . ., 3.0	70	52	18

The remainder of the paper is organized as follows. Section 2 first formulates the classification task and the data set, and gives a brief introduction to the practical background and application. Section 3 provides a short introduction to the feature extraction and classification methods investigated, namely LDA, GerDA, RF, MLP, SVM and RVM for classification. Section 4 presents the classification results and provides a comparison of the performance of the different methods. A final evaluation of the pattern recognition methods considered, as well as a discussion of future work to improve the complete system, are provided in Section 5.

2. Case study and test methodology

2.1 Description of the data set

The spectrometric measurement device provides a characteristic absorption curve, called a fingerprint, over $p \in \mathbf{N}$ wavelengths. The values are given in $[Au/m]$ and stored as a column vector, where the ith one is denoted by $\mathbf{x}_i \in \mathbf{X}$, with the feature space $\mathbf{X} \subseteq \mathbf{R}^p$. In total, we have $N \in \mathbf{N}$ such fingerprints, i.e. $i = 1, \ldots, N$. Associated with each such vector $\mathbf{x}_i$ is the ith organic acid sample with unit $[g/l]$, denoted by $c_{a,i} \in \mathbf{R}$. To formulate the mapping from $\mathbf{x}_i$ to $c_{a,i}$ as a classification problem, the measurements $c_{a,i}$ are clustered into $C = 5$ classes, which account for the whole range of given c_a values. The class $\vartheta \in \Theta$ to which the ith organic acid measurement belongs is given by $\vartheta_i \in \Theta$, where $\Theta = \{1,2,3,4,5\}$ are the class labels as defined in Table 1. These classes correspond to low, low-normal, normal, normal-high and high concentrations of organic acid, respectively. A total of $N = 4437$ samples were obtained from the biogas plant and these were used to generate training and validation data sets with $N_T = 3326$ and $N_V = 1109$ samples respectively. The distribution of the samples across classes is illustrated in Table 1.

From an initial investigation of the data set for the full spectrum spanning 200–750 nm using LDA, it was determined that better results could be obtained by omitting the longer wavelengths; hence, as a final pre-processing step, wavelengths above 640 nm were removed leaving a $p = 176$ dimensional feature vector $\mathbf{x}_i$ for analysis. This cut-off point was determined by optimizing the LDA classification results with respect to p. This is simply a reflection of the fact that better generalization can be obtained if irrelevant feature vectors are

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

discarded. The absorption beyond 300 nm is very low, which leads to the conclusion that absorption characteristics at higher wavelengths do not have a high impact on the concentration of organic acids. Nevertheless, the inclusion of wavelengths above 300 nm is justified to take account of substrate colouring and the changing matrix of mains water in the measurement. Therefore, the appropriate cut-off point needs to be determined.

2.2 Practical application

The measurement of organic acid concentrations in biogas plants is essential for monitoring anaerobic digestion processes and to assure stable and efficient plant operation. The online measurement of this key parameter is important to detect and solve problems in plant operation quickly. High organic acid concentrations decrease the pH level in the bioreactor and cause high stress to methane-producing bacteria, which are no longer able to process the available substrate. Such a change in environmental conditions may easily lead to a complete collapse of the anaerobic digestion process as shown in Figure 1.

The state-of-the-art way to measure and monitor organic acid concentration on agricultural biogas plants unfortunately is still to perform laboratory analysis of the fermentation sludge and substrate feed on a regular basis. This thorough analysis allows efficient process operating conditions to be determined and indicates whether a process is stable or in danger. However, performing the analysis and interpreting the results requires detailed knowledge about the fermentation process, and access to such expertise is generally only cost effective for the largest biogas production facilities. Furthermore, laboratory analyses are difficult, expensive and time-consuming, which makes effective process monitoring and control impractical.

Based on the current situation in the field of online-measurement systems for biogas plants, the need for new developments in sensor technology and engineering solutions in this area is huge. Existing technologies for online measurements of organic acid concentrations in biogas plants, such as gas-phase chromatographs or automatic titrators, are only available to a small group of biogas plant operating companies, as they are expensive and high-maintenance products.

The new approach discussed in this paper is to use UV/vis spectroscopy, which uses ultraviolet light (200–750 nm) to determine the concentration of a certain substance in a liquid sample. The main problem for the application on biogas plants is the high concentration of organic acids in the substrate and also the relatively high concentration of solids. Thus, an automated sample preparation and dilution system has been developed, which addresses these issues and which is installed on an industrial biogas plant near Gummersbach, Germany. This industrial biogas plant with an electrical power output of 1.3 MW uses biological municipal waste for fermentation. In particular, high amounts of leftovers, which rapidly increase organic acid production, may compromise plant operation and stability. Because of these operating conditions, the plant operator has a high interest in testing and validating new promising measurement systems.

Laboratory tests conducted with the S::CAN spectro::lyser show that organic acid concentrations can be detected by analysing the absorption over several wavelengths as shown in Figure 2 (Schmidt and Rehorek, 2008). Different organic acids (acetic acids, propionic acid, lactic acid) were measured in different concentrations to determine the effect on the measured absorption intensity. It is obvious that with higher concentrations the maximum absorption shifts towards longer wavelengths and that for all three acids the absorption maximum is at 230 nm, which makes it very difficult to distinguish between different organic acids. This indicates that organic acid concentrations cannot be measured separately but as a composite parameter, which makes UV/vis spectroscopy well suited for organic acid measurement on biogas plants.

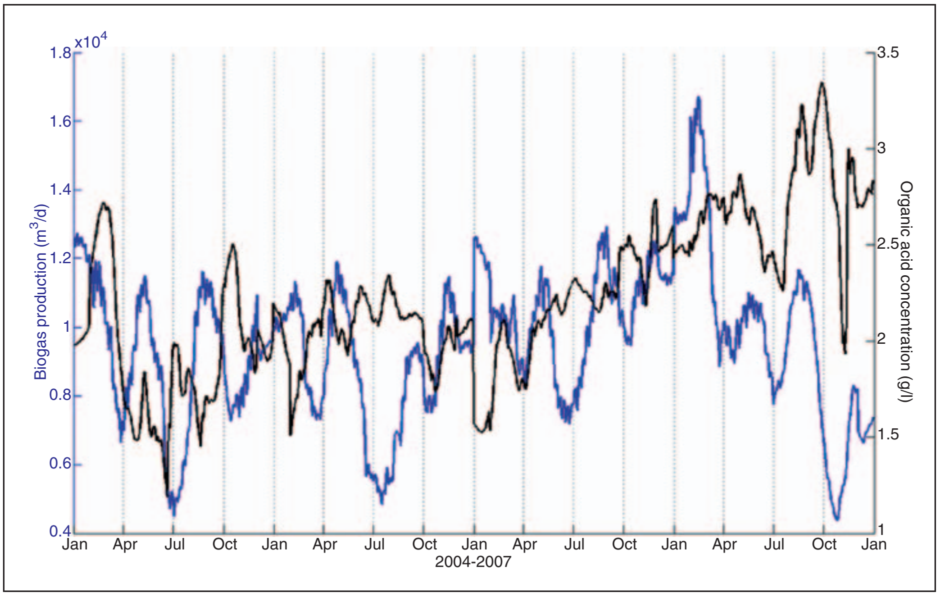

Figure 1 Collapse of biogas production because of rising organic acid concentration at an industrial biogas plant.

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

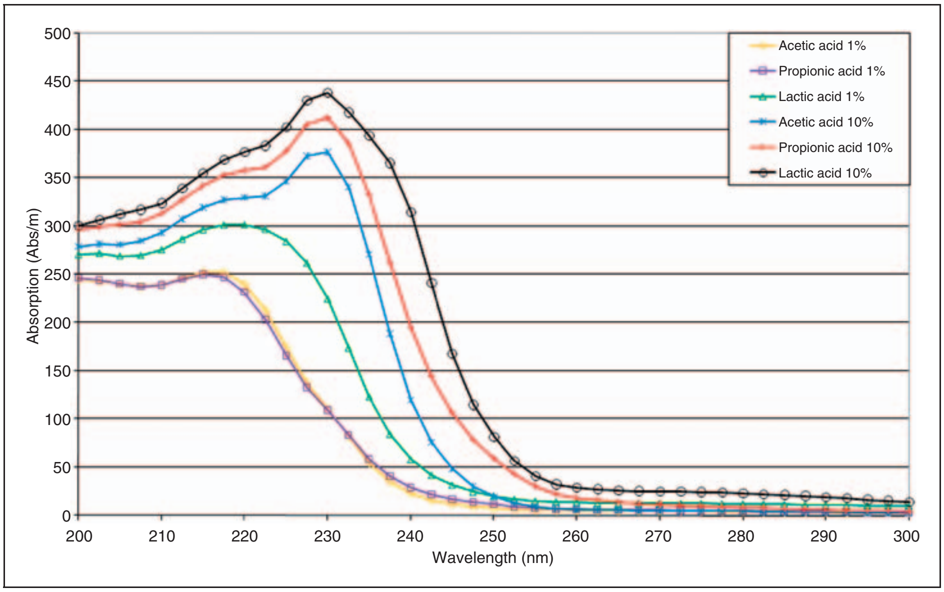

Figure 2 Laboratory measurements of different acids and concentrations using an S::CAN UV/vis probe.

A very popular method often used to deduce chemical parameters from UV/vis measurements is PLS (Langergraber et al., 2003). However, an initial investigation of PLS for this problem yielded very poor results with error rates of about 50%. This led us to consider advanced pattern recognition methods in preference to more traditional linear regression tools.

2.2.1 Online-measurement apparatus. Because of the fact that total solids (TS) concentration in the digester is up to 20%, a direct measurement of the absorption of the substrate at different wavelengths is not feasible, as the 1-mm gap width of the UV/vis probe (S::CAN spectro::lyser) is easily soiled. For this reason, it is necessary to build up a special dilution system for the fermentation sludge. In this case, water from fermentation sludge dewatering is used for online measurements, as organic acids are mainly present in the liquid phase of the sludge.

Laboratory tests have shown that the optimal ratio between water and sample is 1:80 to obtain a clear spectrum. To reach this dilution degree for an accurate measurement with the S::CAN spectro::lyser probe, the dilution unit is filled with 41 of water every 30 min (batch process). A flexible-tube pump is used to administer a defined amount of the fermentation press water (50 ml). Figure 3 shows the layout of the measurement and dilution system.

3. Machine learning techniques

3.1 Linear discriminant analysis (LDA)

LDA searches for a linear transformation $\mathbf{A} \in \mathbf{R}^{m \times p}$, $m \leq p$, such that the transformed data $\mathbf{Y} = \mathbf{A} \times \mathbf{X}$, $\mathbf{Y} := (\mathbf{y}_1, \ldots \mathbf{y}_{N_T}) \in \mathbf{R}^{m \times N_T}$, can be linearly separated better than the original feature vectors $\mathbf{X} := (\mathbf{x}_1, \ldots \mathbf{x}_{N_T})$. The linear transformation $\mathbf{A}$ is determined by solving an optimization problem that corresponds to maximizing the well-known Fisher discriminant criterion:

$$trace(\mathbf{S}_T^{-1} \cdot \mathbf{S}_B) \qquad (1)$$

where $\mathbf{S}_T$ and $\mathbf{S}_B$ are the total scatter-matrix and between-class scatter-matrix, respectively as defined in Duda et al. (2001).

The LDA and a subsequent linear classifier are both implemented in MATLAB® (Moore, 2009). An LDA transformation into a feature space of $m = C - 1 = 4$ dimensions yields the best subsequent linear classification results.

3.2 Generalized discriminant analysis (GerDA)

LDA is a popular pre-processing and visualization tool used in different pattern recognition applications. Unfortunately, LDA and subsequent linear classification procedures produce high error rates on many real world datasets, because a linear mapping cannot transform arbitrarily distributed features into independently Gaussian distributed ones. A natural generalization of the classical LDA is still to rely on having intrinsic features $\mathbf{h} = f(\mathbf{x})$ with the same statistical properties as assumed for LDA features. Unlike LDA a function space F of non-linear transformations $f = \mathbf{R}^p \rightarrow \mathbf{R}^m$ is used. The idea is that a sufficiently large space F potentially contains a non-linear feature extractor $f^* \in F$ that can increase the discriminant criterion (1) compared with what can be achieved with the optimum linear extractor $\mathbf{A}$.

GerDA defines a large space F using the topology of a deep neural network (DNN), and consequently the non-linear feature extractor $f^* \in F$ is given by the DNN, which is trained with measurements of the data space such that the objective function (1) is maximized. Unfortunately, training a DNN with standard methods, like back-propagation, is

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

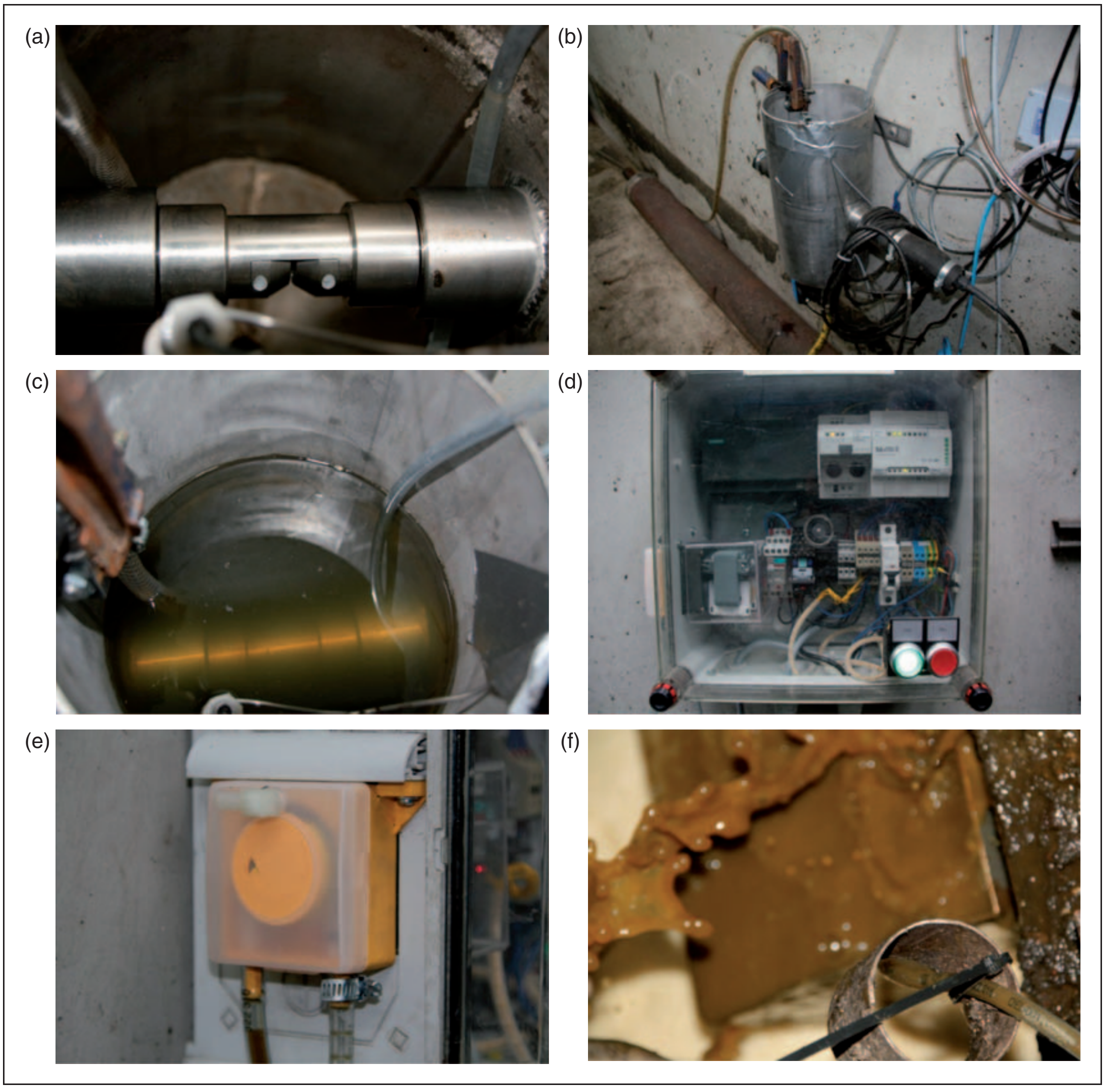

Figure 3 (a) UV/vis-probe with 1 mm gap width; (b) complete layout of the measurement system; (c) online-measurement in progress; (d) control cabinet for the measurement system; (e) flexible-tube pump for exact dosing of the fermentation sludge; (f) collection container for the press water of the fermentation sludge.

known to be challenging because of many local optima in the objective function considered. Therefore, randomly initializing the network parameters and restarting until thrown near to a good solution is ineffective for optimizing DNNs.

To train a large DNN efficiently with respect to (1), Stuhlsatz et al. (2010a, 2010b) have developed a stochastic pre-optimization based on greedily layer-wise trained restricted Boltzmann machines (RBM) (Hinton et al., 2006). In order to appropriately initialize a full GerDA-DNN, a stack of trained RBMs is used (Figure 4). Each RBM is trained with the inputs clamped to the output states of its predecessor RBM via minimizing the difference of two Kullback–Leibler distances, d,

$$CD_n(\Theta) := \mathrm{d}\big(P^0 \| P^\infty; \mathbf{\Lambda}\big) - \mathrm{d}(P^n \| P^\infty; \mathbf{\Lambda}) \tag{2}$$

with respect to the network parameters $\mathbf{\Lambda}$.

The RBM's states are assumed to be Boltzmann distributed according to the distributions P^0, P^n and P^∞. Minimizing (2) can be performed using a very efficient training method for RBMs called contrastive divergence (CD) (Hinton, 2002). In (Stuhlsatz et al., 2010a), the CD heuristic is adapted for learning input–output associations by an output RBM (Figure 5). Training of all RBMs in a stack is unsupervised, with the exception of the output RBM, which requires supervised training through minimization of the mean squared error (MSE) between specific target codes $\vartheta_i \in \Theta := \{\mathbf{t}_1, \ldots, \mathbf{t}_C\}$, $\mathbf{t}_j := \big[t_{j,1}, \ldots t_{j,C}\big]^T \in \mathbf{R}^C$, and the RBM's predictions $\mathbf{v}^{out}(\mathbf{x}_i) \in \mathbf{R}^C$. Minimizing the MSE with respect to the coding

$$t_{jk} := \begin{cases} \sqrt{N_T/N_k} & \text{if } j = k \\ 0 & \text{otherwise} \end{cases} \quad j, k = 1, \ldots, C \tag{3}$$

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

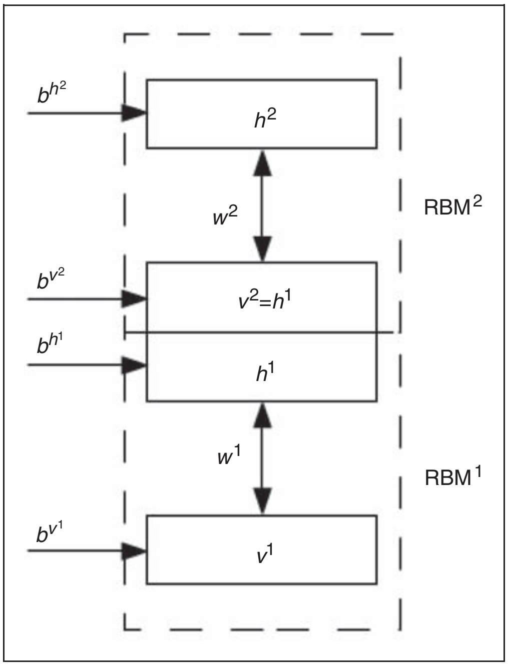

Figure 4 A simple stack of two restricted Boltzmann machines.

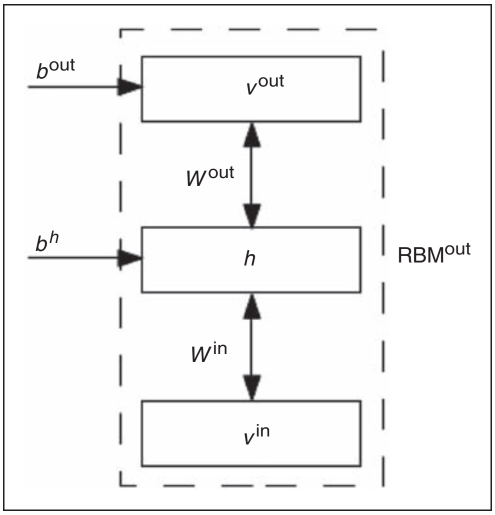

Figure 5 An input–output associative restricted Boltzmann machine.

where N_k is the number of examples of class k, can be shown to asymptotically maximize the discriminant criterion (1) at the hidden units $\mathbf{h} \in \mathbf{R}^m$ (Osman and Fahmy, 1994). A layer-wise training, with all weights **W** and biases **b** up to the last hidden layer **h**, i.e. the output layer of the output RBM is discarded, is used to initialize a GerDA-DNN. Nevertheless, pre-optimization is suboptimal in maximizing (1), thus a subsequent fine-tuning of the GerDA-DNN is performed using a modified back-propagation of the gradients of (1) with respect to the network parameters. In Stuhlsatz et al. (2010a, 2010b), it is shown that stochastic pre-optimization and subsequent fine-tuning yields very good discriminative features and training time is substantially reduced compared with random initialization of large GerDA-DNNs.

For the extraction of intrinsic features from the raw measurements, we used GerDA with a p–250–50–25–m topology, i.e. a five-layer DNN consisting of one input layer with p units, three hidden layers with 250, 50 and 25 units respectively, and one output layer with m units resulting in more than 265 million free parameters. To avoid the effect of over-fitting of the training data, we terminated the fine-tuning after the pre-training stage using an early-stopping criterion dependent on the training error. The topology of GerDA as well as the early-stopping criterion was evaluated on the training data via fivefold cross-validation. Additionally, the best intrinsic dimensionality $m \geq C - 1$ was cross-validated too. The GerDA-framework is implemented in MATLAB®.

The results presented in Section 4.2 were obtained by using a DNN with the topology p–250–50–25–m, with $p = 176$ and $m = 4$. A topology with $m = 5$ was also examined, but classification performance obtained was slightly inferior.

3.3 Random forest (RF)

RF is an efficient algorithm for solving complex classification and regression problems, introduced by Breiman (2001). The RF-algorithm used here is the R-based random forest package for classification and regression presented by Liaw and Wiener in 2002. R is a free software environment for statistical computing and graphics R (R Development Core Team, 2010).

The algorithm is an ensemble of unpruned decision trees. Hence, the classification consists of an ensemble of classification trees, where each tree is trained on a bootstrapped sample of the original training data set (also called in-Bag), and at each new branch the candidate set of variables is a random subset of all variables. For the investigation in this paper, the number of input variables was set to 30 and the number of trees in the forest set to 800.

One third of the training data set is not present in the in-Bag. This left over data is known as out-of-bag (oob) data and is used to obtain a running unbiased estimate of the classification error, as trees are added to the forest as well as to obtain estimates of variable importance. The average misclassification over all trees is known as the oob-error estimate. In this case, the algorithm estimates the importance of all variables by looking at how much the oob-error increases for one variable, while all other variables are not considered. This important information can be used to minimize the number of variables in the dataset in order to minimize computation time and costs. The output of the classifier is determined by a majority vote of the trees.

3.4 Neural networks (MLP)

The MLP is a feedforward artificial neural network, which consists of multiple layers of neurons that are fully connected from one layer to the next. Being an advancement of the standard linear perceptron, MLPs can distinguish data that is not linearly separable, which makes them perfectly suited for learning highly complex and non-linear mappings

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

(Cybenko, 1989). Furthermore, MLPs have several desirable properties like universal function approximation capabilities, good generalization properties and the availability of robust efficient training algorithms (Haykin, 1999). For the classification problem at hand, a single hidden-layer MLP is used to map the non-linear relationship between the intrinsic GerDA features $\mathbf{h}_i^* := f^*(\mathbf{x}_i)$ and the corresponding class labels $\vartheta_i, i = 1, \ldots, N_T$. Applying PLS and FSR on the original data set, two further feature sets are generated, which are also mapped to the class labels by a second and third MLP. This feature extraction significantly accelerates MLP design optimization. MLP training is performed using a BFGS training algorithm with stopped minimization used to prevent over-fitting (McLoone et al., 1998). The optimum number of neurons in the hidden layer and the optimum number of input features were determined for each model by cross-validation on the test data set. For PLS, the optimal MLP design was p = *30*, $n_h = 40$ and for FSR p = 25, $n_h = 60$. Here p is the number of input features and n_h is the number of neurons in the hidden layer.

3.5 Support vector machines (SVM)

SVM offer a computationally efficient method for multi-class classification problems by finding hyperplanes, which separate data sets into classes in a high dimensional feature space. For the classification problem under consideration, a C-support vector classification is used with soft margin optimization and a radial basis function kernel (RBF kernel) (Cortes and Vapnik, 1995) using the SVM implementation LIBSVM (Chang and Lin, 2001).

For classification of the spectral data set, a Gaussian RBF kernel is used, because of several advantages. The RBF kernel is perfectly suited for a non-linear relation between class labels and attributes, and the linear kernel is a special case of the RBF kernel, as proved by Keerthi and Lin (2003). Furthermore, the number of parameters that have to be optimized is limited to two parameters c and λ, which makes model selection easier and faster, when compared with polynomial kernels. c is a trade-off parameter between margin and error, and λ is a standard parameter of the Gaussian RBF kernel. A grid search is performed to determine the best parameters c and λ for the RBF kernel function according to the training data using the misclassification rate (MCR) defined in Equation (4). Training is performed with a fivefold cross-validation procedure (one against one), and different pairs of c and λ values are tested. Finally, the one that yields the best cross-validation accuracy is picked. As suggested by Hsu et al. (2003), in a first pass, exponentially growing sequences of c and λ are evaluated to identify interesting regions for a detailed grid search.

3.6 Relevance vector machine (RVM)

The RVM is a Bayesian formulation of the classification problem with priors selected to encourage sparse representations. They are structurally similar to SVMs and have been shown to provide comparable performance while offering a number of additional benefits. RVM predictions are probabilistic, facilitating the estimation of the uncertainty in predictions, and typically the number of relevance vectors can be reduced significantly compared with SVMs, leading to more robust and computationally more efficient predictions. The RVM was introduced by Tipping (2000, 2001) as part of a general sparse Bayesian learning framework in which sparsity is achieved by assigning parameterized priors to the model weights that encourage sparsity. As a result, predictions for new data are made by estimating the marginal likelihood over the parameters of the priors (referred to as hyperparameters). For the classification at hand, the RVM is used with a Gaussian RBF kernel so that it is directly comparable with the SVM implementation employed. The toolbox used for RVM classification is Version 2 of the sparse Bayesian modelling toolbox developed by Tipping (2009). Because of the fact that RVM training for data sets of high dimensionality has proven to be very slow, RVMs are applied on the intrinsic 4D-GerDA features and the reduced 30D-RF features (Silva and Ribeiro, 2010) only. As the toolbox only supports two-class RVMs, a separate RVM is trained and optimized for each class using a one-versus-all methodology and the overall prediction is determined by selecting the RVM with the highest class probability. The width parameter of the RBF kernels, λ, used with each RVM is determined by cross-validation on the test data set. The optimization is performed using particle swarm optimization (Clerc, 2006) for λ in the range 1×10^{-5} to 1×10^{-1}.

4. Results and analysis

4.1 Performance measures

To validate and compare the classification performances of different methods, the mean misclassification rate (MCR) in per cent can be used,

$$\begin{aligned} \mathrm{MCR} &:= 100 \cdot \left(1 - \frac{1}{N_V} \cdot \sum_{i=1}^{N_V} 1(\mathbf{x}_i)\right), \\ 1(\mathbf{x}_i) &:= \begin{cases} 1 & \text{if } f_{classifier}(\mathbf{x}_i) = \vartheta_i \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

where $f_{classifier} : X \rightarrow \Theta$ is the mapping function. Since in our experiments the number of samples per class is not uniformly distributed, the large classes, such as classes 2 and 3, may dominate the mean MCR too optimistically.

Because our goal is to identify each class with equal certainty, performance measure (4) is not a good choice. Given the confusion matrix $\mathbf{K} := (k_{jl}) \in \mathbf{R}^{C \times C}$, with $\sum_{l=1}^{C} k_{j,l} = 100$, $j = 1, \ldots, C$, for each classifier, an alternative measure of performance, called the normalized MCR (NMCR), which gives equal weighting to each class, is given by:

$$\mathrm{NMCR} := 100 - \frac{1}{C} \cdot \sum_{j=1}^{C} k_{j,j} \tag{5}$$

We decided to use the latter measure for validation, because it is an unweighted measure of performance independent of the number of samples N_ϑ.

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

Table 2 Confusion matrices for different feature extraction and classification methods applied to the UV/vis spectrum data set

	Predicted					
Given	[%]	1	2	3	4	5
(a) LDA						
	1	**68.4**	14.0	8.8	8.8	0.0
	2	7.1	**64.9**	20.2	6.0	1.8
	3	1.9	17.0	**71.1**	8.7	1.3
	4	1.6	17.0	30.8	**42.3**	8.2
	5	0.0	5.6	5.6	5.6	**83.3**
(b) GerDA						
	1	**98.3**	0.0	0.0	0.0	1.8
	2	3.1	**91.6**	4.2	0.8	0.3
	3	0.0	4.5	**88.7**	4.0	2.8
	4	1.1	3.3	12.1	**68.7**	14.8
	5	0.0	0.0	11.1	0.0	**88.9**
(c) RF						
	1	**82.1**	10.7	3.6	3.6	0.0
	2	3.4	**87.4**	6.0	2.4	0.8
	3	0.0	7.0	**82.1**	8.9	1.9
	4	1.1	4.4	12.6	**75.8**	6.0
	5	0.0	5.6	0.0	5.6	**88.9**
(d) RF	(GerDA features)					
	1	**91.1**	7.1	0.0	0.0	1.8
	2	2.4	**91.9**	4.5	1.0	0.3
	3	0.0	4.5	**89.1**	4.0	2.3
	4	1.1	3.8	8.8	**73.1**	13.2
	5	0.0	0.0	5.6	0.0	**94.4**
(e) MLP	(FSR features)					
	1	**86.0**	12.3	1.7	0.0	0.0
	2	5.0	**90.6**	4.4	0.0	0.0
	3	0.0	7.4	**89.8**	2.8	0.0
	4	0.0	0.0	39.6	**58.8**	1.6
	5	0.0	0.0	11.1	66.7	**22.2**
(f) MLP	(GerDA features)					
	1	**86**	12.3	0.0	1.7	0.0
	2	2.6	**91.4**	5.7	0.3	0.0
	3	0.0	3.8	**92.1**	4.1	0.0
	4	0.0	3.3	19.8	**76.9**	0.0
	5	0.0	0.0	16.7	83.3	**0.0**
(g) SVM	(RF features)					
	1	**94.6**	3.6	0.0	1.8	0.0
	2	3.1	**90.6**	6.0	0.3	0.0
	3	0.0	8.7	**88.8**	1.5	1.0
	4	3.3	4.9	13.3	**76.9**	1.6
	5	0.0	5.6	16.7	38.8	**38.9**
(h) SVM						
	1	**93**	5.3	1.7	0.0	0.0
	2	2.0	**92.4**	5	0.6	1.8
	3	0.2	6.0	**89.4**	3.6	0.8
	4	2.7	4.4	6.0	**85.2**	1.7
	5	0.0	5.5	16.7	33.3	**44.5**
(i) RVM	(RVM, GerDA features)					
	1	**84.2**	12.3	0.0	3.5	0
	2	2.1	**92.9**	3.5	1	0.5
	3	0.0	3.8	**90.6**	3.7	1.9
	4	1.1	2.2	9.9	**83.0**	3.8
	5	5.6	0.0	11.1	33.3	**50.0**
(j) RVM	(RVM, RF features)					
	1	**89.3**	5.3	3.6	1.8	0
	2	3.4	**88.5**	5.7	2.4	0
	3	0.6	5.8	**87.9**	5.1	0.6
	4	1.6	3.3	22.0	**71.5**	1.6
	5	0.0	5.6	11.1	50.0	**33.3**

(continued)

Table 2 Continued

(a) LDA used as a feature extractor to get a 4D feature space followed by linear classification; (b) GerDA used as feature extractor to get a 4D feature space followed by linear classification; (c) RF used for feature selection and classification on a 30D feature space; (d) RF used for classification of the 4D GerDA features; (e) MLP used for classification on the 30D FSR features; (f) MLP used for classification on the 4D GerDA features; (g) SVM used for classification on the 30D RF features; (h) SVM used for direct classification on the raw dataset, (i) RVM used for classification on the 4D GerDA features and (j) RVM used for classification on the 30D RF features. LDA, linear discriminant analysis; GerDA, generalized discriminant analysis; RF, random forest; FSR, forward selection regression; MLP, multilayer perceptron; SVM, support vector machine; RVM, relevance vector machine.

4.2 Comparison of classification results

In order to compare the machine learning methods introduced in this paper, confusion matrices and MCR and NMCR performance measures were computed for the test data set. These are presented in Tables 2 and 4, respectively. The NMCR results show that some machine learning methods substantially outperform others for this classification problem. Significantly, some of the methods that underperform are competitive in terms of their MCR, suggesting that the highly unbalanced data set may be a contributing factor. It is clear that GerDA and RF are both capable of achieving very high accuracy for all classes despite the heavily unbalanced data set. In particular, the combination of the GerDA features and RF classification is very effective and yields one of the best classification results with an NMCR of 12.1%. In contrast, the MLP, SVM and RVM classifiers have serious problems classifying class 5 correctly because of the small number of samples in this class, whereas their recognition of the remaining classes is even better than obtained with GerDA and RF. The MLP results on both the weighted and unweighted PLS features were comparable but slightly inferior to the MLP results on the FSR features, hence only the MLP-FSR results are included as representative of both feature sets in Tables 2, 3 and 4.

It can be concluded that distinguishing between classes 4 and 5 seems to be very challenging and is made even more difficult because of the uneven distribution of training examples. One approach to correct for the unbalanced data set is to apply a weighting to class 5 during the training process. In the case of the SVM implementation, LIBSVM, this weighting can be introduced directly as a parameter in the optimization process. For the MLP and RVM, the weighting can be achieved by replicating the samples in class 5 and adding them to the dataset. Using this approach, the SVM, MLP and RVM classifiers were retrained with a 10-fold weighting

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

Table 3 Confusion matrices for different feature extraction and classification methods with weighting introduced to class 5 during training to compensate for the uneven distribution of samples in the training data

Given	Predicted					
	[%]	1	2	3	4	5
(a) W-MLP	(FSR features)					
	1	**64.9**	31.6	1.8	1.7	0.0
	2	4.7	**89.8**	5.2	0.3	0.0
	3	0.2	11.0	**84.3**	3.4	1.1
	4	0.0	1.7	51.1	**41.2**	6.0
	5	0.0	0.0	0.0	33.3	**66.7**
(b) W-MLP	(GerDA features)					
	1	**87.7**	10.5	0	1.8	0.0
	2	1.8	**91.6**	6.3	0.3	0.0
	3	0.0	3.4	**90.9**	4.7	1.0
	4	0.0	2.7	22	**65.9**	9.4
	5	0.0	0.0	0.0	22.2	**77.8**
(c) W-SVM	(weighted SVM, RF features)					
	1	**94.6**	3.6	0.0	0.0	1.8
	2	3.4	**89.8**	6	0.5	0.3
	3	0.0	8.1	**86.0**	4	1.9
	4	2.7	4.4	9.9	**72.5**	10.4
	5	0.0	0.0	0.0	5.6	**94.4**
(d) W-SVM	(weighted SVM)					
	1	**94.7**	3.5	1.8	0.0	0.0
	2	2.4	**91.4**	5.2	1	0.0
	3	0.0	6.2	**89.1**	3.2	1.5
	4	2.2	3.3	9.9	**75.8**	8.8
	5	0.0	5.5	0.0	5.6	**88.9**
(e) W-SVM	(weighted SVM, GerDA features)					
	1	**86.0**	10.5	0.0	1.8	1.7
	2	2.3	**92.9**	3.7	0.8	0.3
	3	0.0	4.5	**90.6**	3.0	1.9
	4	1.6	2.7	11.6	**72.0**	12.1
	5	0.0	0.0	0.0	5.6	**94.4**
(f) W-RVM	(weighted RVM, GerDA features)					
	1	**82.5**	12.3	0.0	3.5	1.7
	2	2.9	**92.7**	3.1	1	0.3
	3	0.6	4.3	**89.6**	3.6	1.9
	4	1.6	2.7	9.9	**75.8**	9.9
	5	5.6	0.0	0.0	0.0	**94.4**

(a) MLP classification of the 30D FSR features using a weighted training set; (b) MLP classification of the 4D GerDA features using a weighted training set; (c) SVM classification on the 30D RF features using a weighted SVM optimization; (d) SVM classification from the raw dataset using a weighted SVM optimization; (e) SVM classification on the 4D GerDA features using a weighted training set; and (f) RVM classification on the 4D GerDA features using a weighted training set. W-MLP, weighted multilayer perceptron; FSR, forward selection regression; GerDA, generalized discriminant analysis; RF, random forest; SVM, support vector machine; RVM, relevance vector machine.

Table 4 Overall results with normalized misclassification rate (NMCR) and misclassification rate (MCR)

Feature extractor	Classifier	NMCR [%]	MCR [%]
LDA	Linear	34.0	35.7
GerDA	Linear	12.8	13.1
RF	RF	16.7	17.0
GerDA	**RF**	**12.1**	**12.4**
none	SVM	19.1	10.8
RF	SVM	19.1	10.8
RF	RVM	25.9	15.4
GerDA	RVM	19.8	10.8
FSR	MLP	30.5	16.3
GerDA	MLP	30.7	12.4
None	**W-SVM**	**12.0**	**12.0**
RF	W-SVM	12.5	14.3
GerDA	W-SVM	12.8	11.8
GerDA	W-RVM	13.0	11.0
GerDA	W-MLP	17.2	13.3
FSR	W-MLP	30.6	22.1

LDA, linear discriminant analysis; GerDA, generalized discriminant analysis; RF, random forest; FSR, forward selection regression; SVM, support vector machine; RVM, relevance vector machine; MLP, multilayer perceptron.

applied to class 5. The resulting confusion matrices are given in Table 3 and the corresponding MCR and NCMR values are recorded in Table 4. As can be seen, the NMCR performances of SVM, MLP and RVM classifiers have improved significantly. The weighted SVM results yield the best overall performance for the NMCR (12.0%). A comparison of the weighted SVM and weighted RVM results on the 4D GerDA data set shows that both methods provide very good and comparable results, with only 0.8% difference in the MCR and 0.2% in the NMCR. However, the number of support vectors used is significantly lower for the RVM, which uses 60 support vectors instead of 398 for the SVM.

5. Discussion and conclusion

This paper demonstrates a new approach for online estimation of organic acid concentrations using UV/vis spectrometric measurements, which offers new possibilities for advanced plant operation and control. The close monitoring of anaerobic digestion processes and the development of control strategies for optimal organic acid concentrations will substantially increase process efficiency and stability. However, results show that this online measurement is far from trivial, such that advanced pattern recognition methods are needed to achieve good results. A comparison of the different feature extraction, selection and classification methods shows that the unbalanced data set available for training is a major problem, when it comes to achieving low NMCR results with some classifiers. However, application of appropriate class weightings during the training process can effectively counter the effect of the very small set of samples available for class 5.

The optimum results were obtained using SVM and a novel method named GerDA in combination with RF classification, both of which yielded an NMCR of 12%. This is sufficiently accurate to be of value for the online measurement of organic acids.

The relatively poor MLP results suggest that the complexity of the classification space cannot be captured adequately with a single hidden-layer network. Further tests with multi-hidden-layer network designs may lead to better results and is the subject of future work.

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

Of the methods considered, the combination of RF and GerDA yields the best error rate for the unweighted data set (12.1%) and furthermore has many desirable properties. The GerDA-framework is self-contained and easy to use with learning performed in a partly unsupervised and partly supervised manner. It can be used as a pre-processing dimension reduction step for different classification methods. Furthermore, the extracted features are very low-dimensional and particularly suitable for simple linear classification (Stuhlsatz, 2010a) and data visualization. Consequently, classification can be performed very quickly and the method is naturally applicable to multi-class problems. Regarding the weighted data set, SVM achieves the best overall results with an NMCR of 12% without requiring any feature selection or extraction methods. The comparison of the weighted SVM and weighted RVM on the GerDA features reveals that both methods perform equally well on the test data set, but that RVMs are more robust and provide more efficient predictive performance because of the significantly lower number of support vectors. This makes RVM well suited for applications where fast classification is the highest priority. As the measurement of organic acids in anaerobic digestion processes is not time-critical, classification time is not an issue for this application.

As already noted, the non-uniform distribution of class sizes biases the training of the pattern recognition methods in favour of the larger class sizes. To detect this effect, it is important to use an unbiased performance measure such as the NMCR for validation and also for the determination and optimization of classifier hyperparameters.

To address the problems posed by having a biased dataset, in future work, sampling with replacement methods will be considered for the generation of balanced datasets for training. Future work will also look at introducing additional meta-classes, which are constructed by the unification of different classes. Since the organic acid concentration is a continuous quantity, this may facilitate more effective time series analysis.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Balabin RM, Safieva RZ and Lomakina EI (2011) Near-infrared (NIR) spectroscopy for motor oil classification: from discriminant analysis to support vector machines. *Microchemical Journal* 98 (1):121–128.

Batista L, Monteiro S, Loureiro VB, Teixeira AR and Ferreira RB (2010) Protein haze formation in wines revisited. The stabilising effect of organic acids. *Food Chemistry* 122: 1067–75.

Bongards M, Hilmer T and Kern P (2007) Online-Konzentrationsmessung in Kanalnetzen – Technik und Betriebsergebnisse. *Forschungsbericht der Fachhochschule Köln (Research Report Cologne University of Applied Sciences)* 173–6.

Breiman L (2001) Random forests. *Machine Learning* 45(1): 5–32.

Chang CC and Lin CJ (2001) *LIBSVM: A Library for Support Vector Machines*. Taiwan.

Cheung HNB, Huang GH and Yu H (2010) Microbial-growth inhibition during composting of food waste: effects of organic acids. *Bioresource Technology* 101: 5925–34.

Clerc M (2006) *Particle Swarm Optimization*. London: ISTE.

Cortes C and Vapnik V (1995) Support-vector network. *Machine Learning* 20: 273–97.

Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems* 2(4): 303–14.

Diamantis V, Melidis P and Aivasidis A (2010) Continuous determination of volatile products in anaerobic fermenters by on-line capillary gas chromatography. *Analytica Chimica Acta* 573–574: 189–94.

Duda RO, Hart PE and Stork DG (2001) *Pattern Classification*, 2nd ed. New York: Wiley.

Guo L, Chehata N, Mallet C and Boukir S (2011) Relevance of airborne lidar and multispectral image data for urban scene classification using random forests. *ISPRS Journal of Photogrammetry and Remote Sensing* 66(1): 56–66.

Haykin SS (1999) *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ: Prentice Hall.

Hinton GE (2002) Training products of experts by minimizing contrastive divergence. *Neural Computation* 14(8): 1771–800.

Hinton G, Osindero S and Teh YW (2006) A fast learning algorithm for deep belief nets. *Neural Computation* 18(7): 1527–54.

Hsu CW, Chang CC and Lin CJ (2003) *A Practical Guide to Support Vector Classification*. National Taiwan University.

Keerthi SS and Lin CJ (2003) Asymptotic behaviours of support vector machines with Gaussian kernel. *Neural Computation* 15(7): 1667–89.

Langergraber G, Fleischmann N and Hofstaedter F (2003) A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. *Water Science Technology* 47(2): 63–71.

Liaw A and Wiener M (2002) Classification and regression by random forest. *R News* 2(3): 18–22.

McLoone S, Brown M, Irwin G and Lightbody G (1998) A hybrid linear/nonlinear training algorithm for feedforward neural networks. *IEEE Transactions on Neural Networks* 9: 669–84.

Moore H (2009) *MATLAB for Engineers*, 2nd ed. New York: Pearson Education International.

Osman H and Fahmy MH (1994) On the discriminative power of adaptive feed-forward layered networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16: 837–42.

Ozkaya B, Demir A and Bilgili MS (2007) Neural network prediction model for the methane fraction in biogas from field-scale landfill bioreactors. *Environmental Modeling and Software* 22(6): 815–22.

Palacio-Barco E, Robert-Peillard F, Boudenne J, L and Coulomb B (2010) On-line analysis of volatile fatty acids in anaerobic treatment processes. *Analytica Chimica Acta* 668: 74–9.

Puñal A, Palazzotto L, Bouvier JC, Conte T and Steyer JP (2003) Automatic control of VFA in anaerobic digestion using a fuzzy logic based approach. *Water Science Technology* 48(6): 103–10.

R Development Core Team (2010) R: A language and environment for statistical computing. Vienna.

Rodrigues JEA, Erny GL, Barros AS, Esteves VI, Brandão T, Ferreira AA, Cabrita E and Gil AM (2010) Quantification of organic acids in beer by nuclear magnetic resonance (NMR)-based methods. *Analytica Chimica Acta* 674: 166–75.

Schmidt H and Rehorek A (2008) New online-measurement methods for biogas plants. Diploma thesis, Cologne University of Applied Sciences, Cologne.

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012

Silva C and Ribeiro B (2010) Inductive Inference for Large Scale Text Classification. Kernel Approaches and Technique*s*. (Studies in Computational Intelligence; 255). Berlin: Springer.

Steyer JP, Bouvier JC, Conte T, Gras P, Harmand J and Delgenes JP (2002) On-line measurements of COD, TOC, VFA, total and partial alkalinity in anaerobic digestion processes using infra-red spectrometry. *Anaerobic Digestion IX* 45(10): 133–8.

Strik D, Domnanovich AM, Zani L, Braun R and Holubar P (2005) Prediction of trace compounds in biogas from anaerobic digestion using the MATLAB Neural Network Toolbox. *Environmental Modeling and Software* 20(6): 803–10.

Stuhlsatz A, Lippel J and Zielke Th (2010a) Feature extraction for simple classification. *Proceedings of the International Conference on Pattern Recognition (ICPR)*, Istanbul.

Stuhlsatz A, Lippel J and Zielke Th (2010b) Discriminative feature extraction with deep neural networks. *Proceedings of the 2010 International Joint Conference on Neural Networks (IJCNN)*, Barcelona.

Tipping ME (2000) The relevance vector machine. *Advances in Neural Information Processing Systems 12*. Cambridge, MA: MIT Press, 652–8.

Tipping ME (2001) Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research* 1: 211–44.

Tipping ME (2009) *SparseBayes. 2.0. An Efficient Matlab Implementation of the Sparse Bayesian Modelling Algorithm [Software]*. Vector Anomaly Ltd, Suffolk.

Wang D, Leung H, Kurian AP, Kim H, J and Yoon H (2010) A deconvolutive neural network for speech classification with applications to home service robot. *IEEE Transactions on Instrumentation and Measurement* 59(12): 3237–43.

Yogameena B, Veera Lakshmi S, Archana M and Raju Abhaikumar S (2010) Human behavior classification using multi-class relevance vector machine. *Journal of Computer Science* 6(9): 1021–6.

Downloaded from tim.sagepub.com at National Univ. of Ireland on September 4, 2012