Detection of Explosive Markers Using Zeolite Modified Gas Sensors

Electronic Supporting Information

William J. Peveler,[a] **Russell Binions,**[b]**, Stephen M.V. Hailes,**[c] **and Ivan P. Parkin**[*d]

1 Background

Support Vector Machines (SVMs) are a form of maximum margin classification, first proposed by Boser *et al.*[1]. A training data set is used to derive an algorithm that separates classes of data. A basic SVM is a binary classifier that *linearly* separates data into two groups. As not all data sets are linearly separable in their basic (*input space*) form, a function maps the data into higher dimensional space (*feature space*), and then a linear separating hyperplane is determined. Once the hyperplane has been determined, the SVM can be used to classify data. The distance between the separating hyperplane and the nearest data point of each class is calculated to be as large as possible (hence *maximum margin* classifier), which ensures that the classifier is robust against new data points that lie slightly outside the observed class boundaries.

In practice, although some data are linearly separable in this way, many are not; measurement noise in observed points that are on the margins of a class can have a disproportionate effect on the hyperplane chosen and may even mean that it is not possible to separate classes without error. Consequently, in the *soft margin* approach to classification, a hyperplane is chosen in such a way as to minimise a cost function that balances the impact of training errors against the size of the margin for correctly classified data, using a parameter C that is tuned through cross validation.

[0a] *Dept. of Security and Crime Science, University College London, 35 Tavistock Sq., London, UK, WC1H 9EZ*

[0b] *School of Engineering and Materials Science, Queen Mary University of London, Mile End Road, London, UK, E1 4NS*

[0c] *Dept. of Computer Science, University College London, Gower St., London, UK, WC1E 6BT*

[0d] *Dept. of Chemistry, University College London, 20 Gordon St., London, UK, WC1H 0AJ Fax: +44 (0)20 7679 7463; Tel: +44 (0)20 7679 4669; E-mail: i.p.parkin@ucl.ac.uk*

As described above, SVMs are binary classifiers, distinguishing between two classes; however, in many practical cases it is necessary to classify data into multiple classes. There are two approaches to the multiclass case: (i) a classifier is trained per class - each classifier defines the hyperplane that separates the examples of a particular class from all the remaining data points (the *"one-against-all technique"*); (ii) a classifier is trained for every pair of classes and, when classifying a point, a number of votes is allocated to each class based on the pairwise comparisons. The assigned class is selected as that with the greatest number of votes (the *"one-against-one technique"*).

There are a number of SVM-based tools in common use; that selected for use in this study was libSVM, designed and maintained by Chang and Lin[2]. This applied a *"one-against-one technique"* to extend the binary classifier to a multiclass problem, as the most effective implementation of a multiclass SVM.[3,4] The implementation used was run in the data mining package, WEKA.[5]

A C-SVC classification SVM was used and the Gaussian radial-basis-function kernel was selected, as recommended by the literature. Training this form of classifier requires two parameters to be set: the cost parameter C, discussed above, which is a measure of how strictly a point in the training set must be classified accurately; and γ, which is an inverse width parameter for the Gaussian kernel (related to the curvature of the decision boundary between classes). Before the classifier can be used, values for both parameters must be determined, and this is typically achieved by searching over a grid of plausible values using cross validation to determine classification accuracy for a given parameter setting. In this case, the grid was formed using values of 2^x, with x between -6 and 10 for C and -10 and 5 for γ, in 0.5 step intervals. Grid extension was allowed where necessary, so numbers outside of this range were accessible to the algorithm. This search was carried out with 2-fold cross-validation to find the best approximate values quickly; next, sequential 10-fold cross validations were performed, using adjacent parameter pairs to refine the values. These optimised C and γ parameters were then used to build the SVM model and test data.

2 Building the SVM

The data used for training and testing the SVM models consisted of 948 data vectors containing the attributes: $|S|$, Direction of Response (1 for

Ro/R, 0 for R/Ro), Sensor Material (7 types), Pulse Length (600 s), Humidity (0%) and Sensor Temperature (in °C). Concentration (in ppm) was initially included, but then later removed in further experiments. Each vector was also labelled with the class (NO_2, $MeNO_2$, NH_3 or EtOH).

The data were randomly split into a training set of 664 vectors and a testing set of 284 vectors. The model was trained, using the grid search to optimise C and γ, on the training data and, once trained, was tested with the test data. This was run using the original split data and a normalised form, to see if this offered any improvement. The original data was well classified with 94% accuracy, however the SVM built using normalised data appeared to mis-classify much of the NH_3 data, leading to a lower accuracy of 81%. The confusion matrices are given in Table 1.

Table 1 Confusion matrices for training with 664 data vectors and testing on 284 with optimised C and γ. True class is defined vertically and output classification horizontally.(a) Data was not normalised and accuracy was 94.37%. (b) Data was normalised and accuracy was 80.63%

(a)

	Classification			
Class	NO_2	$MeNO_2$	NH_3	EtOH
NO_2	**106**	0	0	0
$MeNO_2$	0	**67**	0	0
NH_3	1	0	**46**	3
EtOH	0	3	9	**49**

(b)

	Classification			
Class	NO_2	$MeNO_2$	NH_3	EtOH
NO_2	**106**	0	0	0
$MeNO_2$	0	**67**	0	0
NH_3	0	0	**8**	42
EtOH	0	0	13	**48**

Next, concentration information was removed from the training and test data, as this information would not necessarily be available in a sensing situation. The training and testing was run using non-normalised and normalised data, with optimised parameters. As expected, these

SVMs were not as successful as the non-normalised model with concentration information; however in this instance the normalisation of the data slightly improved classification, from 85.6% to 85.9% (Table 2).

Table 2 Confusion matrices for training with 664 data vectors and testing on 284 with optimised C and γ. True class is defined vertically with the output classification across the horizontal. Concentration values were omitted for both train and test data and data were not normalised for (a), but were subsequently normalised in (b). Classification accuracies were (a) 85.56% and (b) 85.92%

(a)

	Classification			
Class	NO_2	$MeNO_2$	NH_3	EtOH
NO_2	**105**	1	0	0
$MeNO_2$	0	**61**	0	6
NH_3	0	1	**39**	10
EtOH	1	13	9	**38**

(b)

	Classification			
Class	NO_2	$MeNO_2$	NH_3	EtOH
NO_2	**106**	0	0	0
$MeNO_2$	0	**59**	5	3
NH_3	0	1	**44**	5
EtOH	0	15	11	**35**

Finally the SVM was trained using all 948 data vectors (all with 600 s gas pulse and 0% humidity) and tested first against a further 100 data vectors collected at 50% humidity, and then 303 vectors collected in dry air but with 150, 300 and 900 s gas pulses. The SVM was trained as before, using normalised data without the concentration attribute in either training or testing data, and the confusion matrices are given in Table 3.

The use of humidity reduced classification accuracy very slightly from the model in Table 2b by only 0.92%, to 85.00%. Variable pulse length data, was actually more successfully classified, with an SVM accuracy of 89.77%.

Table 3 Confusion matrices for training with 948 data vectors and testing on (a) 100 vectors at 50% humidity and (b) 303 vectors at 150, 300 and 900 s. C and γ were optimised. True class is defined vertically with the output classification across the horizontal. Concentration values were omitted for both train and test data in each case and the data were normalised. Classification accuracy was 85.00% for (a) and 89.77% for (b)

(a)

	Classification			
Class	NO_2	$MeNO_2$	NH_3	EtOH
NO_2	**13**	0	0	0
$MeNO_2$	0	**38**	0	2
NH_3	1	0	**13**	7
EtOH	0	4	2	**21**

(b)

	Classification			
Class	NO_2	$MeNO_2$	NH_3	EtOH
NO_2	**39**	0	0	0
$MeNO_2$	0	**117**	0	3
NH_3	0	3	**54**	6
EtOH	0	9	10	**62**

References

[1] B. E. Boser, I. M. Guyon and V. Vapnik, Proceedings of the fifth annual workshop on computational learning theory, 1992.

[2] C.-C. Chang and C.-J. Lin, *ACM Transactions on Intelligent Systems and Technology*, 2011, **2**, 27:1–27:27.

[3] C.-W. Hsu and C.-J. Lin, *Neural Networks, IEEE Transactions on*, 2002, **13**, 415–425.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006, ch. 7, pp. 325–344.

[5] Y. El-Manzalawy, *WLSVM*, 2005, http://www.cs.iastate.edu/∼yasser/wlsvm/.