DATA MINING IN IDENTIFYING PREMIUM AND REGULAR GASOLINE USING SUPPORT VECTOR MACHINES AS NOVEL APPROACH FOR ARSON AND FUEL SPILL INVESTIGATION

Sunday O. Olatunji[#1], **Imran A. Adeleke**[#2]

[#]Faculty of Computer Science and Information Systems, Universiti Technologi Malaysia, Johor, Malaysia
[1]oluolatunji.aadam@gmail.com [2]imranade@gmail.com

Abstract

In this work, a novel data mining model based on Support Vector Machines (SVM) for the identification of gasoline types has been investigated and developed. Detection and correct identification of gasoline types during Arson and Fuel Spill Investigation are very important in forensic science. As the number of arson and spillage becomes a common place, it becomes more important to have an accurate means of detecting and classifying gasoline found at such sites of incidence. However, currently only a very few number of classification models have been explored in this germane field of forensic science, particularly as relates to gasoline identification. Thus, we have developed Support Vector Machines (SVM) based identification model for identifying gasoline types. The model was constructed using gas chromatography–mass spectrometry (GC–MS) spectral data obtained from gasoline sold in Canada over one calendar year. Prediction accuracy of the model was evaluated and compared with earlier used methods on the same datasets. Empirical results from simulation showed that SVM based model produced accurate and promising results better than the best among the other earlier implemented Artificial Neural Network and Principal Component Analysis methods on the same datasets.

Keywords: gas chromatography–mass spectrometry (GC–MS), soft margin hyper plane, Pattern recognition, Principal Component Analysis (PCA), Artificial Neural Networks (ANN).

1.0 INTRODUCTION

The importance of detection and accurate classification of gasoline for both arson and environmental spills investigation can not be over-emphasized. In this work, Support vector machines was used to classify premium and regular gasoline from gas chromatography–mass spectrometry (GC–MS) spectral data obtained from gasoline sold in Canada over one calendar year [1]. In arson, petroleum-based accelerants such as gasoline, kerosene, and paint thinners are often used to accelerate a fire. In some cases, liquid accelerant is left at the scene, which may be matched to samples that are associated with the suspect. In the environment, gasoline spills are commonplace, but identification of the source is not always straightforward.

The identification of gasoline is crucial for the successful prosecution of an offending individual and/or company. Frequently, gas chromatography is used to fingerprint fuel spills, with the gas chromatograms of the spill sample and the different candidate fuels compared visually in order to seek a best match. This method has some shortcomings. One problem with this technique is that the interpretation and classification of the data is limited by the skill and experience of the analyst. Also, visual analysis of gas chromatograms is subjective and is not always persuasive in a court of law. Pattern recognition methods offer a better approach to the problem of matching gas chromatograms of weathered fuels [2]. Pattern recognition methods involve less subjectivity in the interpretation of the data and are capable of identifying the samples correctly.

Among the pattern recognition methods that were used for this identification purpose was artificial neural network and principal component analysis (PCA) [3]. Unfortunately, accuracy of some of these earlier approaches is often limited and is sometimes bedeviled with problems like over-fitting. Recently, Support Vector Machines have been proposed as a new intelligence framework for both prediction and classification based on both structure risk minimization criterion and soft margin hyper plane. This new framework deals with kernel neuron functions instead of sigmoid-like ones, which allows projection to higher planes and solves more complex nonlinear problems. It has featured in a wide range journal, often with promising results.

In this work, we developed Support vector machines based identification model for identifying gasoline types. The model is constructed using gas

chromatography–mass spectrometry (GC–MS) spectral data obtained from gasoline sold in Canada over one calendar year [3]. Prediction accuracy of the model is evaluated and compared with earlier used methods on the same datasets. Based on the excellent performance of SVM on various identification problems surveyed in the literature coupled with the empirical results from our simulation, we found out that SVM based model produced accurate and promising results better than or at least equal to the best among the other earlier implemented methods on the same datasets. To demonstrate the usefulness of the Support Vector Machines technique as regards spill and arson investigation in particular and forensic science in general, we described both the steps and the use of Support Vector Machines as a Pattern Recognition modeling approach for identifying liquid accelerant left at the scene of arson and spill which may be matched to samples that are associated with the suspects.

A Support Vector Machines (SVM) classifier has been developed and used to identify gasoline types in arson and spill investigation. Comparative studies were also carried out to compare the performance of Support Vector Machines as a classifier with the performance of other classifiers that were already used for this same purpose, using the same datasets, such as ANN and PCA. This study also presents a comparison of SVM to PCA and ANNs for the classification of liquid premium and regular gasoline from their GC–MS chromatograms. SVM have been used in the past for many different types of pattern recognition problems, but this is the first report of applying SVM for the classification of summer and winter, premium and regular grade gasoline from GC–MS chromatograms.

2.0 LITERATURE REVIEW

SVMs are modern learning systems that deliver state-of-the-art performance in real world Pattern Recognition and data mining applications such as Text Categorization, Hand-Written Character Recognition, Image Classification, Material Identification and Bioinformatics. The SV algorithm is a nonlinear generalization of the Generalized Portrait algorithm developed in Russia in the sixties [4]. As such, it is firmly grounded in the framework of statistical learning theory, or VC theory, which has been developed over the last four decades by Vapnik et. al. [5]. In a nutshell, VC theory characterizes properties of learning machines which enable them to generalize well to unseen data.

In its present form, the SV machine was developed at AT &T Bell Laboratories by Vapnik and co-worker [6][7][8][9]. Due to this industrial context, SVM research has up to date had a sound orientation towards real-world applications. Initial work focused on OCR (optical character recognition). Within a short period of time SVM classifers became competitive with the best available systems for both OCR and object recognition tasks [10][11]. A comprehensive tutorial on SVM classiers was published by Vapnik, (1982, [5]. Also in regression and time series prediction applications, excellent performances were soon obtained [12] [13][14].

Application of SVM for both classification and regression problems continues to soar high day in day out. A good number of research works have been done recently using SVM. Emre et al [15] presented a decision support system that classifies the Doppler signals of the heart valve to two classes (normal and abnormal) by using Least-Squares Support Vector Machine (LS-SVM) classifier instead of Artificial Neural Network (ANN). The paper used a previous work where ANN was used as a classifier, as feature extractor from measured Doppler signal and concluded that LS-SVM has more advantage than ANN classifier especially in terms of training running time.

Kristof and Dirk [16] applied SVM in a newspaper subscription context. They carried out a comparison between two parameter-selection techniques needed to implement the SVM and found that SVM show good generalization performance when applied to noisy marketing data. However, they also showed that only when the optimal parameter-selection procedure is applied, SVM outperform traditional logistic regression, whereas random forests outperform both kinds of SVMs.

Kemal Polat and Salih Günes [17] conducted a breast cancer diagnosis using Least Square Support Vector Machine (LS-SVM) classifier algorithm with 98.53% accuracy. After examining the robustness of the system using classification accuracy, analysis of sensitivity and specificity, *k*-fold cross-validation method and confusion matrix, they found that LS-SVM produced a very promising result compared to the previously reported classification techniques. A similar

comparative study was carried out by [18] on breast cancer diagnosis using Linear Discriminate Analysis (LDA) and also concluded that the classificatory accuracy of the SVM-based classifier is slightly better than that of LDA.

Taboada et al [19] created a quality map of a slate deposit, using the results of an investigation based on surface geology and continuous core borehole sampling using different kinds of support vector machines (SVMs): SVM classification (multi-class one-against-all), ordinal SVM and SVM regression, and found that the SVM regression and ordinal SVM are perfectly comparable to kriging (a statistical modeling that interpolates data from a known set of sample points to a continuous surface) and possess some additional advantages in terms of interpretability and control of outliers in terms of the support vectors.

Wang et al [20] compared the performance of heuristic-based model with that of SVM-Based non-linear regression model in the prediction of surface tension data of common diversity liquid compounds. Comparing the results of the two methods, the SVM-based non-linear regression model gave a better prediction result than the heuristic method. Anthony and Goh [21] demonstrated the excellent performance of SVM in the field of geotechnical engineering by using SVM to assess a large seismic liquefaction data set, with 98% accuracy.

3.0 OPERATIONS OF SUPPORT VECTOR MACHINES

Generally, in prediction and classification problems, the purpose is to determine the relationship among the set of input and output variables of a given dataset $D = Y, X$ where $X \in R^p$ represents the n-by-p matrix of p input variables. It may be noted that $Y \in R$ for forecasting problems and $Y \subseteq$ for classification problems. Suppose $D = \{ y_i, s_{i1}, \ldots, x_{ip} \}$is a training set for all $i = 1, \ldots,$of input variables, *Xsubj* where $[X_j = (x_{1j}, \ldots x_{nj})^T]$ for $j = 1, \ldots, p$ and the output variables, $Y = (y_1 \ldots y_n)^T$. The lower case letters $X_{i1}, x_{i2}, \ldots, X_{ip}$ for all $i = 1, \ldots n$ refer to the values of each observation of the input variables, and $y = k$ to the response variable Y to refer to class A_k for all, $k = 1, 2, \ldots c$ where $c \geq 2$.

Here we briefly describe the basic ideas behind SVM for pattern recognition, especially for the two-class classification problem, and refer readers to [5][22] for a full description of the technique.

The goal is to construct a binary classifier or derive a decision function from the available samples which has a small probability of misclassifying a future sample. SVM implements the following idea: it maps the input vectors $\overrightarrow{x} \in R^d$ into a high dimensional feature space$\Phi(\overrightarrow{x}) \in H$ and constructs an Optimal Separating Hyperplane (OSH), which maximizes the margin, the distance between the hyper plane and the nearest data points of each class in the space*H*. Different mappings construct different SVMs. The mapping $\Phi(.)$ is performed by a kernel function $K(\overrightarrow{x_i}, \overrightarrow{x_j})$ which defines an inner product in the space H. The decision function implemented by SVM can be written as:

$$f(\overrightarrow{x}) = sgn\left(\sum_{i=1}^{N} y_i \alpha_i, K(\overrightarrow{x}, \overrightarrow{x_i}) + b \right) \quad \ldots (1)$$

Where the coefficients are obtained by solving the following convex Quadratic Programming (QP) problem: Maximize

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K(\overrightarrow{x_i}, \overrightarrow{x_j})$$

Subject to

$$0 \leq \alpha_1 \leq C \quad \ldots (2)$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad i = 1, 2, \ldots N$$

In the equation (2), C is a regularization parameter which controls the trade off between margin and misclassification error. These X_j are called Support Vectors only if the corresponding $\alpha, > 0$.

Several typical kernel functions are:

$$K(\overrightarrow{X_i}, \overrightarrow{X_j}) = (\overrightarrow{X_i}, \overrightarrow{X_j} + 1)^3 \quad \ldots (3)$$

$$K(\overrightarrow{X_i}, \overrightarrow{X_j}) = \exp\left(-\gamma \| \overrightarrow{X_i} - \overrightarrow{X_j} \|^d \right) \quad \ldots (4)$$

Equation (3) is the polynomial kernel function of degree d which will revert to the linear function when

$d = 1$. Equation (4) is the Radial Basic Function (RBF) kernel with one parameter γ.

Other kernel functions are:

Linear: $K(x_i, x_j) = x_i^{\gamma} x_j$

And

Sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T X_j + r)$

Here γ, r, and d are kernel parameters.

Thus, SVMs provide a new approach to the problem of pattern recognition (together with regression estimation and linear operator inversion) with clear connections to the underlying statistical learning theory. They differ radically from comparable approaches such as neural networks: SVM training always finds a global minimum, and their simple geometric interpretation provides fertile ground for further investigation [22].

4.0 IMPLEMENTATION

Here we present the description of our implementation procedures, data and software used.

4.1 Experimental data

The experimental data was originally taken from a Canadian Petroleum Products Institute Report of the composition of unleaded summer and winter gasoline in 1993 [2] and reported by [3]. In this report, 44 samples of regular gasoline (22 winter, 22 summer), and 44 samples of premium gasoline (22 winter, 22 summer) were analyzed by GC–MS. The gasolines were collected over the course of one year from different regions across the country. Forty-four compounds that may be present in automotive gasoline in concentrations of greater than 1% were reported.

The experiment was carried out using MATLAB 2007b, with support vector toolbox integrated by us.

4.2 Determining the Optimal Parameters through Cross-Validation

Percentage of correctly classified samples is employed as a criterion for the determination of the SVM parameters using test-set-cross-validation method. These criteria also provide a more accurate evaluation of the classifier.

An experimental procedure based on test-set-cross-validation was employed in our study. We used the stratified sampling approach to divide the data set into both training and testing data, such that the size of the training set is 70% of the available data and the testing is the rest. The parameters associated with the SVM were optimized through a test-set-cross-validation on the available data set. This entire process was repeated 10 times with different random splitting of the training and testing data sets using the stratified sampling approach; the final results were averaged over 10 runs.

The details of the test-set-cross-validation for optimizing the SVM parameters goes thus: For each run of generated training and testing set, the percentage of correctly classified samples for a group of parameters C (bound on the Langrangian multiplier) and (conditioning parameter for QP methods) noted. Searching through all possible values of the parameters in a given range will identify the best value of the performance measure and the corresponding values of the parameters for the fixed set of features. In our experiment, this process was repeated for every SVM kernel option available, each time with an incremental step of parameters. The optimal values of the parameters and the kernel option associated with the best performance measure were identified. A summary of the procedure is as follows.

(I) Choose the initial kernel option from the list of available kernel options.

(II) Identify the best values of the parameters C and through a test-set –cross-validation and store the corresponding performance measure.

(III) If there is no kernel option left, then go to (IV). Otherwise, add the next kernel option and go to (II).

(IV) Identify the best performance measure and its associated kernel option and the parameters' value.

(V) Use the optimized kernel option and the parameters values to train the final SVM.

(VI) Calculate the performance measure (% of correctly classified samples) for both the training and testing sets using the classifier obtained in the previous step (V).

The optimum parameters identified through the above procedures are then used to build the final SVM whose results were compared with the results of the earlier implemented models. The SVM experimentation carried out were subjected to the same conditions as

earlier implemented models, for instance the data set were divided into 50% training set and 50% testing set.

5.0 RESULTS AND DISCUSSION

We present here the results of classifying the gasoline into regular and premium in the first instance and then the classification into premium winter, regular winter, premium summer and regular summer , making four classes in this second case while the first case has just two classes.

5.1 Results of Classification into two groups of premium and regular gasoline

In this section, the results of classifying gasoline to either premium unleaded or regular unleaded are presented as follow using PCA, ANN and SVM.

From the results in table 1 and figure 1, we found that SVM and ANN performed at par with PCA performing lower than the two.

Table 1. The percentage of correctly classified regular and premium gasoline

Model	% correctly Classified (Training)	% correctly Classified (Testing)
PCA	95.46	90.91
ANN	100	100
SVM	100	100

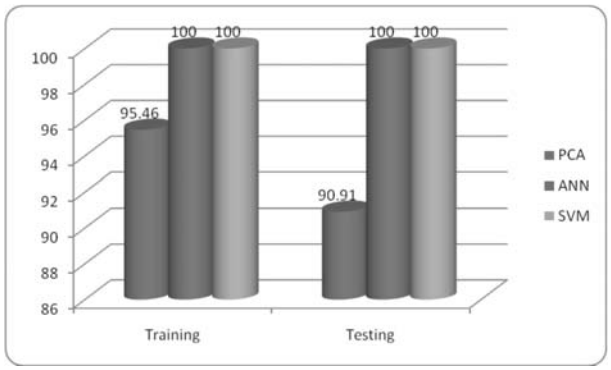

Fig. 1. Pictorial representation of the % of correctly classified regular and premium gasoline

Table 2. The percentage of correctly classified PUW, RUW, PUS, and RUS

Model	% correctly Classified (Training)	% correctly Classified (Testing)
PCA	75.0	47.72
ANN	100	84.09
SVM	100	96.47

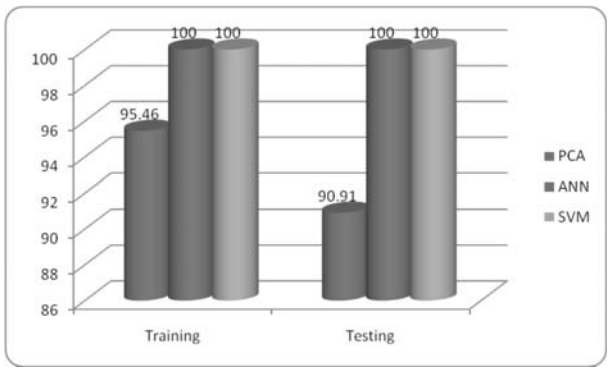

Fig. 2. pictorial representation of the % of correctly classified PUW, RUW, PUS, and RUS

5.2 Results of Classification into four groups of premium winter, premium summer, regular winter and regular summer gasoline

In this section, the results of classifying gasoline to either premium winter, regular winter, premium summer, or regular summer are presented as follow using PCA, ANN and SVM. This second classification is necessary because in a climate such as the one in Canada, where the current sample under investigation was obtained, there is a requirement for a higher fuel vapour pressure for cold engine-starting in the winter and a lower fuel vapour pressure to prevent vapour lock in the fuel line in the summer. Therefore, it was expected that the winter gasolines would tend to have more volatile compounds than the summer gasolines and that significant variations between winter and summer gasoline would be detectable.

From the reported results of the two different classification cases, contained in table 1 & 2 and figure 1 & 2, we found out that, as the number of classes increases from two to four, so also the performance of PCA and ANN decreases compared to that of SVM.

In fact PCA performed very badly as for the four-class case with just 47.72 % correct classification for the testing data set. Though ANN still performed but its testing performance is far lesser that that of SVM. Thus, SVM has again distinguished itself here as a viable tool for correctly classifying gasoline in the field of forensic science, during arson and oil spillage investigation.

6.0 CONCLUSION

A support vector classificatory model has been built in this work. It has been shown that PCA performed to some extents in the classification of gasoline to either premium or regular, though with lower accuracy compared to ANN and SVM. But PCA performed very poorly when it was used to sub-classify the gasoline samples into their respective summer/ winter grouping thereby resulting in four classes. As for ANN, it performed excellently for the two-class classification of the gasoline samples with 100% correct classification just like the SVM did. But on four-class classification of gasoline, ANN performance on the testing set reduced to 84.09% which is far lower than that of SVM that stayed at 96.47%. Thus we conclude here that SVM has again distinguished itself as a viable tool in the field of forensic science for correct classification of gasoline samples in the course of arson and oil spill investigations. Also it could also serve as powerful classificatory tool in other fields of forensic science like in the identification of broken glasses found at the scene of crime during arson investigation and the likes.

REFERENCES

[1] Lavine B.K.; Brzozowski D.; Moores A.J.; Davidson, C.E.; June 2001, Mayfield H.T.;Genetic algorithm for fuel spill identification Analytical Chemical Acta, Volume 437, Number 2, 27 pp. 233-246(14) , Elsevier

[2] 1993, Composition of Canadian Summer and Winter Gasolines, Canadian Petroleum Products Institute, Report No. 945. @REF = [3] Philip Doble, Mark Sandercocka, Eric Du Pasquier, Peter Petocz, Claude Roux, Michael Dawson, (2003) Classification of premium and regular gasoline by gas chromatography/mass spectrometry, principal component analysis and artificial neural networks. Forensic Science International 132 26–39, Elsevier.

[4] Vapnik V. and Chervonenkis A. 1964 A note on one class of perceptrons. Automation and Remote Control, 25.

[5] Vapnik V. 1995, The Nature of Statistical Learning Theory. Springer, N.Y.

[6] Boser B.E., I. Guyon M., and Vapnik V.N. A training algorithm for optimal margin classifiers. In D. Haussler, editor, 5th Annual ACM Workshop on COLT, pages 144-152, Pittsburgh, PA, ACM Press.

[7] Guyon, B. Boser, and V. Vapnik. 1993 Automatic capacity tuning of very large VC-dimension classifiers. In Stephen Jose Hanson, Jack D. Cowan, and C. Lee Giles, editors, Advances in Neural Information Processing Systems, volume 5, Pages 147 -155. Morgan Kaufmann, San Mateo, CA.

[8] Cortes and V. Vapnik. 1995, Support vector networks. M. Learning, 20:273-297.

[9] Scholkopf, C. Burges, and V. Vapnik. 1995 Extracting support data for a given task. In U. M. Fayyad and R. Uthurusamy, editors, Proceedings, First International Conference on Knowledge Discovery & Data Mining Menlo Park, AAAI Press.

[10] Scholkopf, C. Burges, and V. Vapnik_ Incorporating invariances in support vector learning machines. In C. von der Malsburg, W. von Seelen, J. C. Vorbruggen, and B. Sendho 1996, editors, Artificial Neural Networks –ICANN '96, pages 47 – 52, Berlin, Springer Lecture Notes in Computer in Computer Science, Vol. 1112.

[11] Scholkopf, P. Y. Simard, A. J. Smola, and V. N. Vapnik 1997, Prior knowledge in support vector kernels. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, Advances in Neural information processings systems, volume 10, pages 640-646, Cambridge, MA, MIT Press.

[12] Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. 1997 Support vector regression machines. In M. Mozer, M. Jordan, and T. Petsche, editors Advancesin Neural Information Processing Systems 9, pages 155 -161, Cambridge, MA, MIT Press.

[13] Stitson M., Gammerman A., Vapnik V., Vovk V., Watkins C., and J. Weston. 1999 Support vector regression with ANOVA decomposition kernels. In B. Scholkopf, C.J.C. Burges, and A.J. Smola, editors, Advances in Kernel Methods - Support Vector Learning, pages 285 – 292, Cambridge, MA, MIT Press.

[14] Mattera D. and Haykin S. 1999 Support vector machines for dynamic reconstruction of a chaotic system. In B. Scholkopf C.J.C. Burges, and A.J. Smola, editors, Advances in Kernel Methods - Support Vector Learning, pages 211-242 Cambridge, MIT Press.

[15] Emre Çomaka, Ahmet Arslana, Ibrahim Türkoglub, 2007 A Decision Support System based on Support

Vector Machines for diagnosis of the heart valve diseases, Computer in Biology and Medicine 37, pp 21-27.

[16] Kristof Coussement, Dirk Van den Poel, 2008 Churn prediction in subscription services: An application of Support Vector Machines while comparing two parameter-selection techniques, Expert Systems with Applications 34, pp 313–327, 2008 [Accepted for publication in.

[17] Kemal Polat, Salih Günes, 2007 Breast cancer diagnosis using least square support vector machine, Digital Signal Processing 17, pp 694–701, Elsevier,.

[18] Cheng-Lung Huang, Hung-Chang Liao, Mu-Chen Chen, 2008 Prediction model building and feature selection with support vector machines in breast cancer diagnosis, Expert Systems with Applications 34, pp 578–587, Elsevier,.

[19] Taboada J., Matías J.M., Ordóñez C., García P.J., 2007 Creating a quality map of a slate deposit using support vector machines, Journal of Computational and Applied Mathematics 204, pp 84 – 94, Elsevier,.

[20] Jie Wang, Hongying Du, Huanxiang Liu, Xiaojun Yao, Zhide Hu, Botao Fan, 2007 Prediction of surface tension for common compounds based on novel methods using heuristic method and support vector machine, Talanta 73, pp. 147–156, Elsevier,.

[21] Anthony T.C. Goh, S.H. Goh, 2007 Support Vector machines: Their use in geotechnical engineering as illustrated using seismic liquefaction data, Computers and Geotechnics 34, pp. 410–421, Elsevier,.

[22] Vapnik,V. (1998) Statistical Learning Theory. Wiley, New York.

S.O. Olatunji received the B.Sc. (Hons) Degree in Computer Science, Ondo State University Ado Ekiti, Nigeria (1999), M.Sc. Computer Science, University Of Ibadan, Nigeria (2003), M.S.. Degree in Information and Computer Science, King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia (2008).He is currently pursuing his Phd in Computer Science. He is a member of ACM and IEEE.