Indian Tea Discriminator: SVM Approach

Princee Gupta
M.tech Scholar
SIRT,Bhopal

Rajesh K Shukla
HOD, CSE
SIRT, Bhopal

ABSTRACT
Artificial Organoleptic Systems are being used today for a variety of detection tasks from quality control of food products to medical diagnosis. The optimization of sample preparation, signal processing, feature extraction, classifier are as important as choice of sensors within the array in enhancing the performance of the organoleptic system. It is difficult to determine if all features considered are necessary for the classifier while classifying megavariate data. The presence of irrelevant features increases the dimensionality of the search space, which can potentially deluge the accuracy of the Pattern Recognition (PARC) techniques. Hence, a systematic method is required to reduce the number of features in order to optimize the performance of PARC.

Tea in present time is the most popular beverages having huge global marketing. It is a very complex chemical compound graded by various testers' score, which led to many human errors and may vary from person to person. This problem can be solved by using an instrument called "Electronic Tongue (i-tongue)" that gives fast, reliable and repeatable results. This system analyses liquid including an array of non-specific chemical sensors with partial specificity for different component in liquid samples and appropriate pattern recognition capable of recognizing the qualitative and quantitative composition of sample and complex solutions. In this project we use "Principal Component Analysis (PCA)" to reduce the dimension of features and "Support Vector Machine (SVM)" to classify different tea samples including an array of non-specific chemical sensors.

Keywords
PCA, SVM, PARC, Electronic Tongue, i-tongue, Hyperplane, Pattern Recognition.

1. INTRODUCTION
The i-Tongue was capable of discriminating between substances with different taste modalities on the basis of impedance reading being collected at particular frequency interval and could also distinguish different substances eliciting the same basic taste [1]. The i-Tongue employs some nonspecific multi-electrode electrochemical impedance spectroscopy for classification of Indian tea. The impedance response at logarithmic frequency interval in the predefined range is used by one of the feature extraction method like PCA which reduces the dimensions but before this features should be selected. The features are being selected by one of the feature selection method to select the best feature in the dataset provided. Feature selection techniques [2] are based on the design of a criterion function and the selection of a search strategy. In this project we have used Ant Colony Optimization as search strategy. The i-tongue is used in various applications viz. food industry, medicine, safety, environmental pollution monitoring, chemical industry, Legal protection of inventions-digital fingerprints of taste and odors.

SVM is a set of supervised learning methods used for classification and regression that constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space. There are many hyperplanes that might classify the data, the best hyperplane is the one that represents the largest separation, or margin, between the two classes. Hence, we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. Such type of a hyperplane known as maximum-margin hyperplane and the linear classifier it defines is known to be a maximum margin classifier, since in general the larger the margin the lower the generalization error of the classifier.

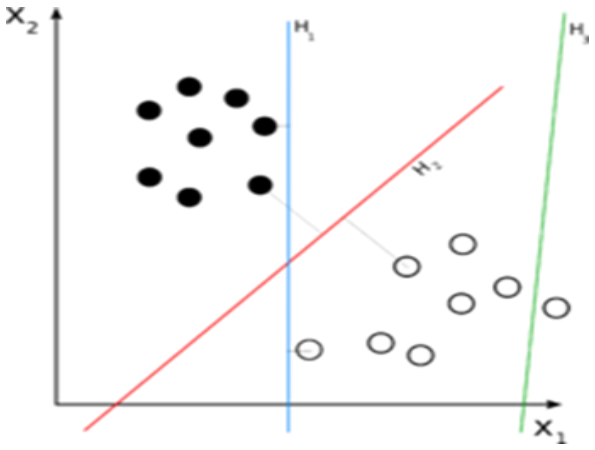

Figure 1: Maximum Margin Hyperplane

In Figure 1 H3 (green) doesn't separate the 2 classes. H1 (blue) does, with a small margin and H2 (red) with the maximum margin. Currently, SVM is widely used in object detection & recognition, content-based image retrieval, text recognition, biometrics, bioinformatics (Protein classification, Cancer classification), speech recognition, etc.

There are some advantages and disadvantages of support vector machines:

Advantages:

- It gives optimized result in high dimensional spaces.
- It is memory efficient because it uses a subset of training data points in the decision function (called support vectors).
- It is versatile because different Kernel functions can be specified for the decision function. Common kernels are provided, but it is preferred to specify custom kernels.

Disadvantages:

- If the number of features of data set is much greater than the number of samples, the method is likely to give poor performances.
- SVM does not directly provide probability estimates.

There are many benefits to this method. The solution to the optimization problem is a global minimum, whereas other machine learning methods, such as neural networks, can often terminate in local minima [3], thereby modeling the training data inaccurately. The SVM solution is an expansion on a subset of the original training data, resulting in a sparser model and less computation time required for subsequent classification. Finally, an SVM minimizes the expected generalization error, rather than just the empirical error, on the training data set.

2. RELATED WORK

Tea is one of the widely consumed beverage in the world and India being the second largest producer, has its tremendous agro-commercial importance. A number of efforts have been made to classify different tea using sensor array and electrochemical techniques such as Cyclic Voltammetry, Potentiometry and Conductivity. However, in comparison to potentiometry, especially with voltammetry, the impedance measurements are advantageous because of the potential experimental simplicity and the reduction of the response times [4].

Instrumental evaluation of black tea is quite complex because of presence of many compounds and therefore it is being distinguished by tea tasters on their scores [5, 6]. Impedance tongue are sensor array for qualitative and quantitative analysis and it is used to differentiate basic standard taste. The classification models can be created by using supervised and unsupervised techniques and artificial neural network [7, 8].

Ekachai Phaisangittisagul [9] proposed an approach of sensor subset selection in machine olfaction in which each sensor should provide different selectivity profiles over the range of target odor application so that a unique odor pattern is produced from each sensor in the array. He employed, a state-of-the-art classification algorithm, Support Vector Machine (SVM), by selecting the first few seed sensors based on maximum margin criterion among different odor classes. From the experimental results on the soda data set, the numbers of selected sensors were not only significantly reduced but the classification performance was also increased.

Cheng Tan et al [10] used Support Vector Machines (SVM) technique to identify fuel types. Flame oscillation signal were captured by a three-cell flame monitor. Thirty flame features were extracted from each flame signal. Then Principal Component Analysis (PCA) was used to choose the principal components of each features vector. An SVM was deployed to map the principal components, size-reduced flame features, to an individual type of fuel. The data of eight different types of coal obtained from a combustion test facility demonstrated that the SVM technique was effective for identifying the fuel types

Xiaodong Wang et al [11] presented a new intelligent method for signals recognition of electronic nose, based on support vector machine (SVM) classification. The SVM operates on the principle of structure risk minimization hence a better generalization ability is guaranteed. The experiments of the recognition of three different gases, ethanol, gasoline and acetone, have been presented and the method achieves higher recognition rate at reasonably small size of training sample set and overcomes disadvantages of the artificial neural networks.

Lihong Zheng et al [12] review some pattern recognition schemes published in recent years. After giving the general processing steps of pattern recognition, they discuss several methods used for steps of pattern recognition such as Principal Component Analysis (PCA) in feature extraction, Support Vector Machines (SVM) in classification, and so forth. Different kinds of merits are presented and their applications on pattern precognition are given.

3. PATTERN RECOGNITION

Pattern recognition aims to classify data (patterns) based either on a priori knowledge or on statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements or observations, defining points in an appropriate multidimensional space. Its ultimate goal is to optimally extract patterns based on certain conditions and to separate one class from the others.

Types of classification: In supervised classification we have a set of data samples with associated labels, the class types. These are used as exemplars in the classifier design. In unsupervised classification, the data are not labeled and we seek to find groups in the data and the features that distinguish one group from another. The fig 2 shows how patterns will be classified.

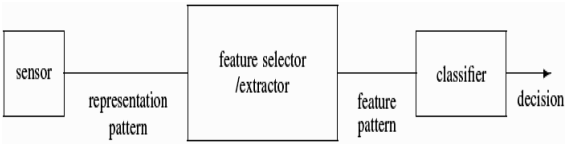

Fig 2: Pattern Classifier

4. METHODOLOGY OF I-TONGUE

Figure 3 shows the block diagram of a novel impedance-Tongue (i-Tongue). This approach uses nonspecific multi-electrode electrochemical impedance spectroscopy for classification of samples of beverages in terms of sensory scores, manufacturing parameters and qualitative evaluations etc. In this approach the impedance of test solution is measured for a range of sinusoidal frequencies. The frequency specific impedance response of the electrodes is then used for classification of the samples. As a result of which a number of features come in limelight having large dimensions, so to reduce the dimension and to get optimized result feature extraction has been done by using feature extraction technique such as Principal Component Analysis (PCA). These extracted features are then subjected to a pattern recognition engine (PARC) which classifies the sample based on its learning parameters. In this project we prefer SVM as PARC.

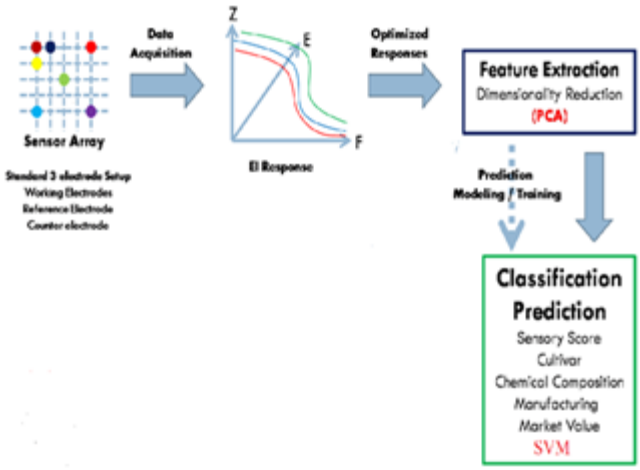

Fig 3: Generic Architecture of a novel impedance-Tongue

4.1 Sample Preparation

Take tea leaves of different categories. Grind the leaves in a grinder. Now, take one gram of grinded tea leaves and boil them with 100ml of distilled water. Mix the mixture for 10 minutes with magnetic stirrer. Then filter the solution using Whatman Filter No. 42. Take three samples of 20 ml each of this tea solution. Then we analyze the impedance of this solution using impedance analyzer.

4.2 Data Acquisition

Impedance values, including real and imaginary values with its phase degree, are exported from analyzer for a frequency range of 20 Hz to 150 MHz at three different voltages viz. 1v, 10mv, 100mv. Hence we have a 120x401 matrix which is fed into SVM for classification by using LIBSVM software package [14] where 120 are tea samples (8 samples with 15 readings) and 401 are frequency points within the previous specified frequency range. To reduce the dimensionality of data set we applied PCA to the matrix by preserving information as much as possible of the variation presented in the data set. As a result of which we got reduced matrix having 300 features.

The Nyquist Plot of impedance values at different voltages is shown below in Fig (4, 5, 6).

-Img(z)

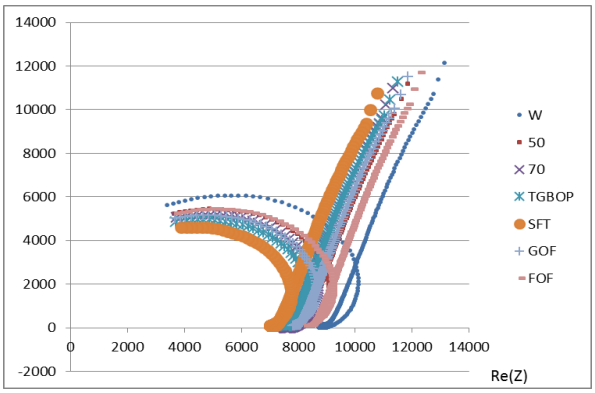

Fig 4: Nyquist Plot for Impedance at 1 V

-Img(z)

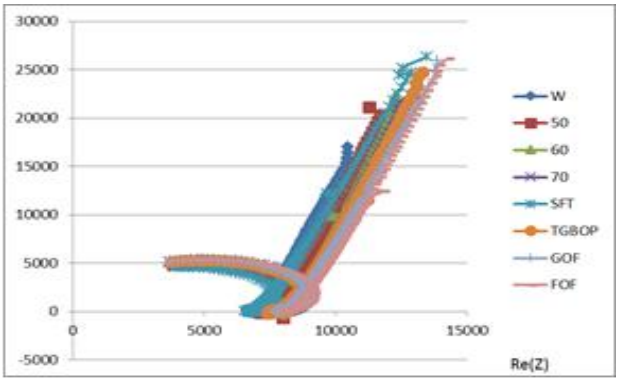

Fig 5: Nyquist Plot for Impedance at 10 mv

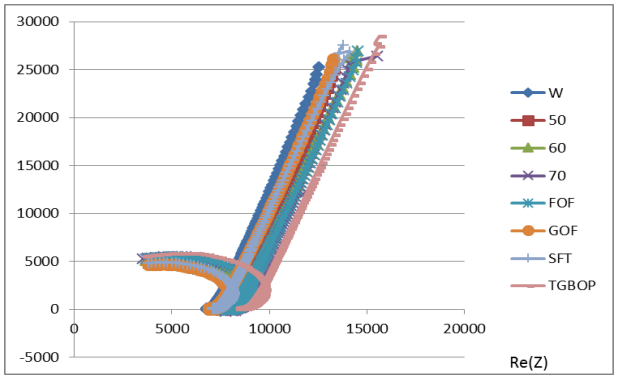

Fig 6: Nyquist Plot for Impedance at 100 mv

5. SVM CLASSIFICATION

We got tea data set in vector form, these tea data vectors were randomly divided into two sets. One was a training set of SVM that possesses 80% of the data. The other was a test set that possesses the residual 20% of the data. The training and testing process of SVM was repeated for some number of trials. The average success rates for 5 trials at 1V, 10 mV, and 100 mV are 98.32%, 99.16% and 100% with 300 features respectively. A confusion matrix shows the classification confusion grid. Figure below shows the confusion matrix with eight classes.

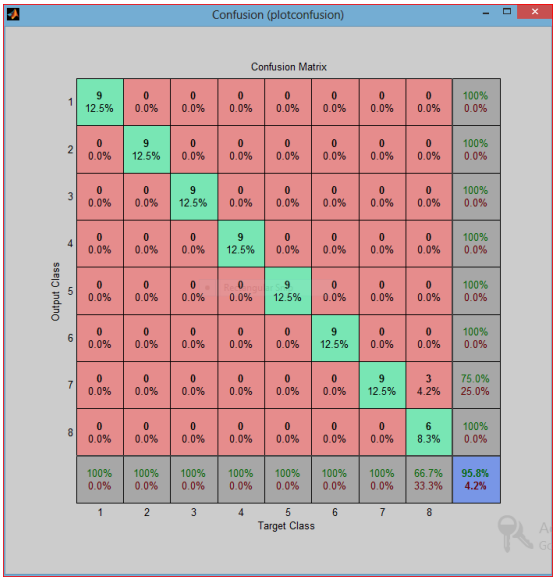

Figure 7: The Confusion Matrix showing recognition rate at 1 V for 300 features

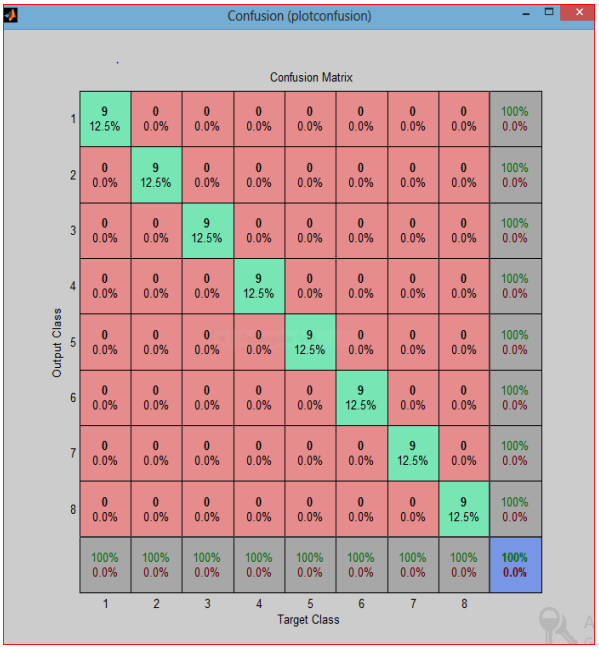

Figure 8: The Confusion Matrix showing recognition rate at 10 mv for 300 features

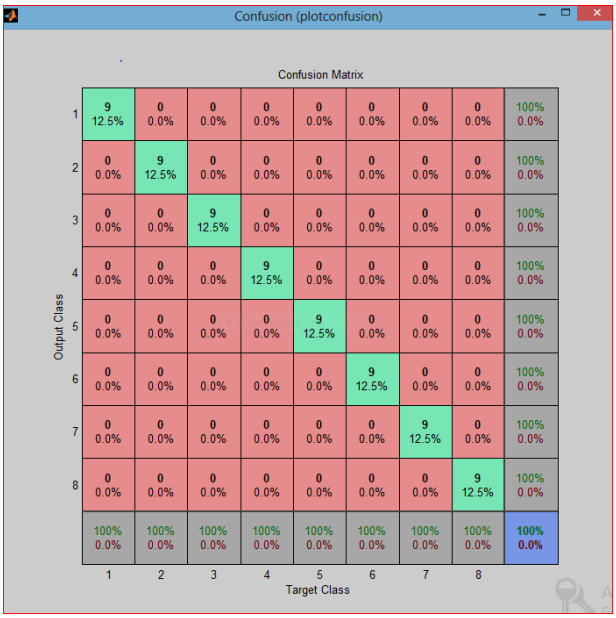

Figure 9.The Confusion Matrix showing highest recognition rate at 100 mv for 300 features

6. CONCLUSION

We set the LIBSVM system parameters to the followingvalues.

1. Distance-

s=0 (Type of SVM, C-SVC)

t=2 (Type of kernel function, linear)

c=5 (cost of misclassified samples)

No of Features=300

The average success rates for 5 trials at 1V, 10 mV, and 100 mV are 98.32%, 99.6.5% and 100% respectively.

On increasing the features to 400, the average success rates for 10 trials at 1v, 10 mV, and 100 mV are 94.16%, 96.66% and 96.66% respectively. Thus we saw decrease in the recognition rates on increasing the no of features being fed to the SVM classifier.

2. Distance-

s=0 (Type of SVM, C-SVC)

t=2 (Type of kernel function, RBF)

c=5 (cost of misclassified samples)

γ=2 (width of RBF coefficient, gamma)

No of Features=300

The average success rates for 10 trials at 1V, 10 mV, and 100 mV are 96.7%, 96.2% and 96.2% respectively. The SVM classifier exhibited good generalization performance for the testing samples on using the Radial Basis Function kernel.

With increase in the features to 400, the average success rates for 10 trials at 1V, 10 mV, and 100 mV are 97.1%, 97.1% and 98.3% respectively which clearly shows an increase in the classification accuracy on increasing the no of features.

3. Distance-

s=0 (Type of SVM, C-SVC)

t=2 (Type of kernel function, RBF)

c=1 (cost of misclassified samples)

No of Features=400

Table 1: Recognition results of eight classes of tea using SVM with different Kernel parameters.

Kernel parameter	Average Recognition Rate for 20 trials of tea samples (%)		
	1V	10 mV	100 mV
γ = 0.1	78.6	98.4	95.8
γ = 1	95.3	98.5	97.9
γ = 10	97.1	97.9	96.9
γ = 100	96.9	97.9	97.5
γ = 1000	97.5	96.9	97.1

Table 1 presents recognition results using the SVMs, with different values of γ and C=1, for the tea samples. As shown in Table 1, the SVM classifier generated good recognition results. The SVM classifier exhibited good generalization performance for the testing samples. The experiments show that the average recognition rate is high for the testing samples at voltage 100 mv.

7. REFERENCES

[1] Andrey Legin, Alisa Rudnitskaya, David Clapham, Boris Seleznev, Kevin Lord and Yuri Vlasov "Electronic tongue for pharmaceutical analytics — quantification of tastes and masking effects" J. Bioanalytical Chemistry, 2004, V. 380, pp. 36-45.

[2] P. Devijver and J. Kittler, Pattern Recognition: A Statistical Approach, Prentice Hall, 1982.

[3] Bishop C M. Neural Networks for Pattern Recognition. Oxford University Press. 1995.

[4] A. Riul, H.C. de Sousa, R.R. Malmegrim, D.S. dos Santos, A.C.P.L.F. Carvalho, F.J. Fonseca, O.N. Oliveira, L.H.C. Mattoso, Wine classification by taste sensors made from ultra-thin films and using neural networks, Sensors and Actuators B: Chemical 98 (2004) 77–82.

[5] B. Tudu, A. Jana, A. Metla, D. Ghosh, N. Bhattacharyya, R. Bandyopadhyay, Elec- tronic nose for black tea quality evaluation by an incremental RBF network, Sensors and Actuators B: Chemical 138 (2009) 90–95.

[6] N. Bhattacharyya, R. Bandyopadhyay, M. Bhuyan, A. Ghosh, R.K. Mudi, Correlation of multi-sensor array data with "Tasters" panel evaluation for objective assessment of black tea flavour, in: Int. Proc. ISOEN-2005, Barcelona, Spain, April 13–15, 2005.

[7] Yu. Vlasov, A. Legin, A. Rudnitskaya, C.DiNatale, A. D'Amico,Nonspecific sensor arrays ("electronic tongue") for chemical analysis of liquids (IUPAC Technical Report), Pure and Applied Chemistry 77 (2005) 1965–1983.

[8] K. Toko, Electronic sensing of tastes, Electroanalysis 10 (1998) 657–669.

[9] E. Phaisangittisagul, H.T. Nagle, Sensor Selection for Machine Olfaction Based on Transient Feature Extraction, Instrumentation and Measurement, IEEE Transactions on, 57 (2008) 369-378.

[10] Cheng Tan, Lijun Xu, Zhang Cao," On-Line Fuel Identification Using Optical Sensing and Support Vector Machines Technique", I2MTC 2009 - International Instrumentation and Measurement Technology Conference Singapore, 5-7 May 2009, 2009 IEEE.

[11] Xiao-Dong Wang, Hao-Ran Zhang, Chang-Jiang Zhang," Signals Recognition Of Electronic Nose Based On Support Vector Machines", Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005, 2005 IEEE.

[12] Lihong Zheng and Xiangjian He," Classification Techniques in Pattern Recognition", Proceedings ISBN 80-903100-8-7 WSCG'2005, January 31-February 4, 2005.

[13] Spector, A. Z. 1989. Achieving application requirements. In Distributed Systems, S. Mullender.

[14] Chih-Chung Chang and Chih-Jen Lin. LIBSVM -- A library for Support Vector Machines. From: http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html [Online].