# Correlation-Based Incremental Learning Network for Perfume Classification

Panida Lorwongtrakool
Department of Informatin Technology
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
panida.l@email.kmutnb.ac.th

Phayung Meesad
Department of Informatin Technology Management
Faculty of Information Technology and Digital Innovation
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
pym@kmutnb.ac.th

*Abstract*— **Contamination inspection or quality inspection of raw materials or products is a very important task, especially in the perfume industry that requires an expert for inspection. However, the human nose has limitations such as fatigue, which affects the accuracy. Therefore, an electronic nose or sensor array has been developed to assist in the inspection. The signal data from electronic nose is fed into machine learning models to learn and process. Since data change over time, the input data fluctuate according to the changing environment. In addition, when there are new data with features that change from the original patterns, the classification outcome may not correct and the model will not be able to classify as effective as the original model. Therefore, to solve the problem mentioned this research proposes Correlation-Based Incremental Learning Network (CILN) combined with Sliding Window, which learns automatically by adapting to new data while maintaining the existing knowledge. The experiments were conducted on classifying perfumes. The experimental data were divided into 1 batch for the training and 2 batches for testing. The proposed algorithm was compared with other well-known classifiers. The results showed that the proposed CILN algorithm provides the highest accuracy of 100%.**

***Keywords—correlation, incremental learning, e-nose, perfume classification***

## I. Introduction

Contamination inspection or quality inspection of raw materials, foods, beverages is an important task, but different industries employ different methods, such as chemical value testing or taste testing. Those methods require direct contact with samples. Therefore, researchers have attempted to invent various inspection methods, one of which is smell inspection by using an electronic nose that is currently being used extensively for inspections, such as microbial or chemical contamination inspection which may occur during the production process, pork freshness inspection [1], fungus inspection in strawberries [2], black tea quality [3], or even medical diagnoses [4, 5].

The industry that gives the most importance to scent is the perfume industry which is currently experiencing counterfeiting products. Some perfume brands are made from quality raw materials and may be expensive, while counterfeited perfume uses less quality raw materials and is sold at a low price, hence increasing the number of industries making counterfeited perfume worldwide. Counterfeiting perfume has caused considerable damage in the economy [6] and affected consumers because counterfeited perfume uses low quality raw materials and contains harmful chemicals. When consumers use it for a long time, it can lead to health problems. Counterfeited perfume tend to be sold at a cheaper price and comes in similar packaging boxes which is difficult to verify.

A human nose can recognize the smell to inspect the quality of food, perfume, environment, chemical leaks or even weapons or drugs. Perfume inspection normally requires experts to smell the perfume, but there are limitations to the human nose in smelling certain scents. The efficiency of smelling depends on the health, and humans can get tired. Importantly, the human nose is not suitable for smelling various harmful toxins, and it is difficult to measure [7].

Basic factors that make data analysis successful is pattern recognition which use machine learning techniques. There are 2 machine learning paradigms, Batch learning refers to machine learning methods that use all the observed data at once. Incremental learning (also called online learning) refers to the machine learning methods that apply to streaming data collected over time. These methods are used to update the learned function accordingly when new data come in. Incremental learning mimics the human learning process from experiences[8].

Since, the data imported into the system will be either static, dynamic or streaming in a nonstationary environment. The process to generate these data may change over time, and input data may fluctuate as the environment changes. In addition, when there is new data with features that have changed from the original patterns the outcome may be invalid. Thus, the model will not be able to classify as effective as before and is unable to learn new information when a new sample is added, hence decreased efficiency [9]. Ultimately, all the previous data sets can no longer be used, and a new model has to be created when there is new data. The process of creating a new model leads to a phenomenon known as catastrophic forgetting [10] and, therefore, is not suitable for datasets that change according to circumstances.

Therefore, to improve the algorithm to learn new data while maintaining knowledge of old datasets. So, this paper proposed Correlation-Based Incremental Learning Network Algorithm for Perfume Classification which can learn and adapt model automatically while maintain the old knowledge. The data used in the experiment are from an array of sensors developed for measuring the perfume's response.

## II. Literature reviews

### *A. Incremental Learning*

Generally the learning algorithms not created for incremental learning. Therefore, they are sensitive to data that is constantly changing (velocity). There is a continuous transmission of data in the form of data streaming which affects the accuracy reduced. New models must be rebuild, and it is difficult to apply the existing knowledge.

In addition, in terms of data in the new situation, there are still a problem and needs to increase the learning ability of the algorithm [11]. Gradual learning of knowledge without abandoning or forgetting the knowledge gained or retrain process in the case of ANN, such as MLP happens when new input data comes in the network, and the previously learnt knowledge is forgotten. To solve this problem, incremental learning algorithms must combine the new knowledge to the previously acquired knowledge in the same manner of human learning methods in which learning is based on previous knowledge or experiences which is the main feature of incremental learning [12].

Therefore, the algorithms that can learn from new data without having to access the previous set of data and maintain prior knowledge would be a good method of classification to support both static data and data stream, especially when new data samples are added all the time.

1) Characteristics of Incremental Learning

(a) Incremental learning algorithms handle with continuous data and non-stationary distributions.

(b) It adapts to new data without forgetting the existing knowledge; it does not need to retrain.

(c) It is compatible with data streams or big data to create machine learning models faster.

(d) Use Instance windows or instance weighting mechanism without making modifications to the algorithms. New models are calculated based on time periods of using windows or weights by considering new data received.

(e) All or part of the data is used to create an initial model, check for changes in data (using the detection function), and rebuild models as needed based on new data.

(f) It can automatically modify the learning mechanism in order to increase learning. For example the weight of the artificial neural network is adjusted every time there is a new pattern in the system.

2) Types of Incremental Learning

(a) Instance-incremental learning refers to the system that receives data at Step $t$ - the input point $\mathbf{x} \in X^n$ which $X$ represents the input in $n$ dimensional space and predicts the output $y \in Y$ (Output $Y$ can be either continuous in regression or in classification.

(b) Batch incremental learning receive batches of data $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_\mathbf{p})$ and must specify the label $(y_1, y_2, \dots, y_p)$ for each input point, where $p$ is the number of data points.

Relevant studies have shown that Adding new capabilities to the artificial neural network often resulting in forgetting (Catastrophic Forgetting) [10], where the network works well in jobs that have traditional data The easiest way to solve this problem is to configure all the parameters of the network and feature extraction before training [13]. However, it was observed that the results had deteriorated considerably. Therefore, all architecture must be adjusted [14]. The research by [13] gave details of the above method. It also offered a mechanism for optimizing the entire network while maintaining the old performance.

While those methods add few parameters for new tasks; however, it does not guarantee that the model maintains the full capabilities of the old work in [15]. Neuron's learning rate will decrease. Therefore, there is research, which presents methods to maintain the integrity of the old representation while the number of parameters is slightly increased. Incremental learning based on the threshold of the classifier is a method that presents not only an incremental learning model But also helps to estimate thresholds and error in teaching The effectiveness of the proposed algorithm. For example, research evaluates the use of gender recognition. Face detection and human recognition [16].

The RBF neural network is used in incremental learning, and it is found to be quite successful. Each kernel function is related to the hidden nodes of the network linked to one cluster. Incremental learning in which hidden nodes are added by updating weights and relevant kernel parameters for real-world decision making makes the RBF Neural network more flexible than MLP architecture. Thus, RBF training is faster than other neural network architectures [9].

It can be seen that RBF is used in incremental learning [9] and has been combined with other techniques, such as Neuro-Fuzzy for Machinery Condition Health Monitoring [17], In addition to RBF, SOM is also combined with Euclidean distance for Intrusion Detection[18], and SVM with Mahalanobis for Intrusion Detection[19]. The popular distance measurements, such as Euclidean distance and Mahalanobis distance. Euclidean distance is a simple method. But there is a limitation that is sensitive to scales of variables and suitable for data that scatter in circle shape form. Mahalanobis distance consider covariance matrix , so it can solve problem in Euclidean distance. Mahalanobis distance suitable for data that scatter in oval shape form. Therefore, the selection may have to consider the scatter of data. So, Correlation is an interesting measurement for similarity measurement since it considers the correlation of variables, determine the correlation level and suitable for continuous. In this research, correlation distance has been used to measure the similarity of the data that changes over time.

### B. *Correlation Distance*

Correlation Distance is a statistical measure used to measure the independence of two values or any two vectors. Correlation values are between 0 and 1, which can be measured by the variance or standard deviation. It can be calculated according to Eq. (1).

$$d(\mathbf{p}, \mathbf{q}) = 1 - \frac{\mathrm{cov(p,q)}}{\mathrm{std(p)} * \mathrm{std(q)}} \tag{1}$$

$$\mathrm{cov(p,q)} = \sum_{\mathrm{j=1}}^{\mathrm{k}} (\mathrm{p_j} - \bar{\mathrm{p}}) * (\mathrm{q_j} - \bar{\mathrm{q}})$$

$$\mathrm{std(p)} = \sqrt{\frac{1}{\mathrm{k}} \sum_{\mathrm{j=1}}^{\mathrm{k}} (\mathrm{p_j} - \bar{\mathrm{p}})^2}$$

$$\bar{\mathrm{p}} = \frac{1}{\mathrm{k}} \sum_{\mathrm{j=1}}^{\mathrm{k}} \mathrm{p_j}$$

where $d(\mathbf{p}, \mathbf{q})$ is distance from $\mathbf{p}$ to $\mathbf{q}$; $\mathbf{p}$ and $\mathbf{q}$ are any data points in $n$ dimension space.

## III. METHODS

### A. *Samples collection*

The perfume dataset is the real world data used in the experiment which from the developed sensor array which consist of 7 sensors, including MQ-2, MQ-3, MQ-4, MQ-5, MQ-6, MQ-8, MQ-135 which these sensors have different responses to gases and we can used signals from sensor array to represent pattern or fingerprint of perfume. In order to inspect 3 scents from 3 different brands and the scents from all 3 counterfeit fragrances, so the data are composed of T_Chanel, F_Chanel, T_Chloe', F_Chloe', T_Britney spear, F_Britney spear, which is a total of 6 scents. The samples were collected by spraying the fragrances on cotton pads and keeping them in prepared glass bottles. Then, the sensor array was used to measure responses from different types of gas. The data collection process from the sensor is divided into 3 phases, which are:

1) Before the response measurement

2) During the sensor response to gas from the perfume samples

3) After taking the sensor out of the perfume samples

The time in each step is set to 20 seconds, 300 seconds, and 60 seconds, respectively. The next measurement is carried out for all samples. Therefore, each sample takes approximately 420 seconds or 7 minutes. The signal from the sensor is shown in Fig. 1. The sample collection were taken 3 times (1 time per day) as described above.

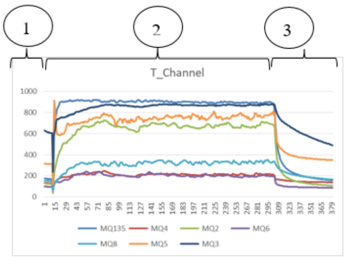

Fig. 1 Signal Received from Sensor Responses

## B. Data Preparation

The data used are in time-series dataset consisting of many variables (multivariate time-series). As shown in Figs. 2, 6 types of perfume were prepared, and noise data were eliminated. Then, the signal in the appropriate range was selected, which is in the range of the sensor signal that responds to the scent at the appropriate level as shown in number 2 in Fig. 1. The signal measurement contains 150 records /1 type. Each batch contains 900 records (150 records * 6 types), consisting of 7 attributes and which are divided into 6 classes : T_Chanel, F_Chanel, T_Chloe', F_Chloe', T_Britney spear, F_Britney spear. Since the simple collection were taken 3 times (1 collection per day), therefore the perfume dataset have 3 batches contains 2,700 records (900 records * 3 times).

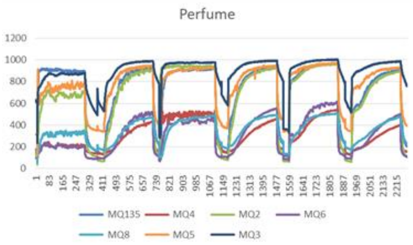

Fig. 2 Signal Obtained from Response to 6 Perfume Types

## C. Operational Structure of CILN

The main objective of this research is to develop the Correlation-based Incremental Learning Network for perfume classification by using the Correlation Distance and membership function. There are 2 steps of CILN as follows: 1) Learning phase 2) Predict phase shown in Fig. 3. The data were normalized and reframe by sliding window with the size equivalent to 5 shown in Fig. 4. To effectively verify the proposed method, take batch 1 to the training data, and take batch 2,3 to the testing data.

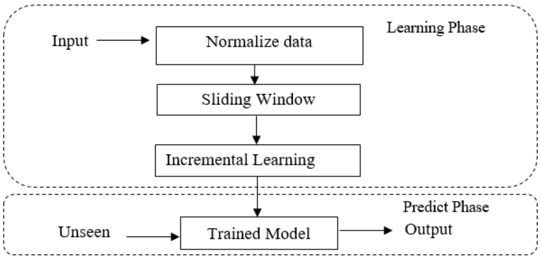

Fig. 3. Operational Structure of CILN

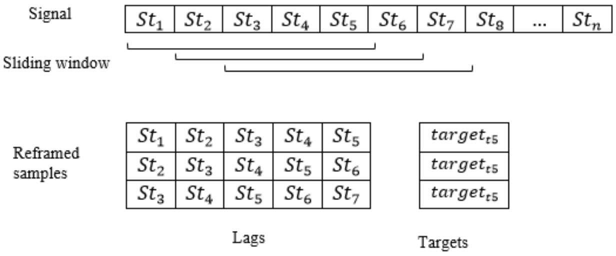
Signal
$St_1$ $St_2$ $St_3$ $St_4$ $St_5$ $St_6$ $St_7$ $St_8$ ... $St_n$
Sliding window
Reframed samples
$St_1$ $St_2$ $St_3$ $St_4$ $St_5$
$St_2$ $St_3$ $St_4$ $St_5$ $St_6$
$St_3$ $St_4$ $St_5$ $St_6$ $St_7$
$target_{t5}$
$target_{t5}$
$target_{t5}$
Lags
Targets

Fig. 4. The sliding window of training samples.

1) Learning process in this research focuses on classification with incremental learning. The algorithms used to create models and learning based on correlation distance and membership function. In addition, new data can be learned due to constant learning updates by determining the membership function and threshold of each target class. The correlation distance threshold of membership is defined as $d_{th}$, where $0 < d_{th}, <1$ and $d_c$ is correlation distance value that measured between instances. The correlation distance measurement between input (P) and centroid of cluster can be done by using correlation distance as in equation (1) and calculate the membership level with membership function as in equation (2). The node with the highest membership value is the winning node. For decision step, if $d_c < d_{th}$, the instance will be dissimilarity with cluster. On the other hand, the instance has more similarity with cluster if $d_c \geqslant d_{th}$. The system's learning model consists of weight which is adjusted according to the equation (3).

- Calculate the membership function level

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

where $\mu$ is Mean and $\sigma$ is StandardDeviation.

- Weight adjustment

$$W_{P_{J,new}} = \frac{W_{PJ,old}(C_{J,new}-1)+p}{C_{J,new}} \tag{3}$$

| | | |
|---|---|---|
| $W_{PJ,new}$ | is | new weight |
| $W_{PJ,old}$ | is | original weight |
| $p$ | is | latest input data |
| $C_{J,new}$ | is | number of members in the group |

2) Predict Phase

Read the unseen pattern to measure between the coming pattern and the prototype of each cluster and update prototype.

## IV. Experiment and Result

The algorithms NaïveBayes, BayesNet, MLP and Simple Logistics were compared by using perfume dataset. The results are as shown in Table 1.

TABLE I. COMPARISON OF PERFORMANCE AMONG EXISTING CLASSIFIERS FOR PERFUME DATA

| Algorithm | Accuracy (%) | Precision | Recall | F-Measure | Time (sec) |
|---|---|---|---|---|---|
| CILN | 100.00 | 100.00 | 100.00 | 100.00 | 115.26 |
| MLP (7-5-6) | 88.11 | 88.00 | 88.00 | 88.00 | 4.94 |
| Simple Logistics | 81.15 | 82.00 | 81.00 | 81.00 | 7.19 |
| BayesNet | 57.43 | 60.00 | 57.00 | 57.00 | 0.23 |
| NaïveBayes | 50.25 | 45.00 | 50.00 | 45.00 | 0.15 |

TABLE I. showed the comparison result of the proposed CILN algorithm with well-known algorithm with the top3 accuracy values: CILN (100%), MLP (88.11%) and Simple Logistics (81.15%) . When considering the processing time, CILN takes the most of the processing time since there are many processing steps.

## V. Conclusion

The model accuracy in data forecasting is very important. Normally, the model created will work well for some time because the input data may fluctuate according to the changing environment. As a result, the model created will not be able to classify as efficiently as before, hence reduced performance. The process of creating a new model leads to a phenomenon known as catastrophic forgetting and, therefore, is not suitable for datasets that change according to the situation. To solve this problem, this research presents the correlation-based incremental learning network by developing a sensor array consisting of 7 sensors, including MQ-2, MQ-3, MQ-4, MQ-5, M-Q6, MQ-8, MQ-135 to measure the response to 3 type of scents from 3 brands and 3 counterfeit brands, which is a total of 6 type of scents. The data were experimented by diving the data into 3 batches (batch 1 used for the training data and take batch 2,3 to the testing data). Correlation distance is used to measure similarities and membership function is used to identify membership levels. The data were patterned by Sliding Window. The findings reveal that Correlation-Based Incremental Learning Network help the system to learn the new data pattern while maintaining prior knowledge. It can increase classifier accuracy and also support time series datasets which are nonstationary distributions, and it can be used in the perfume industry, food industry, environment monitoring or others.

## References

[1] Huang, L., et al., Nondestructive measurement of total volatile basic nitrogen (TVB-N) in pork meat by integrating near infrared spectroscopy, computer vision and electronic nose techniques. Food Chemistry, 2014. 145(0): p. 228-236.

[2] Pan, L., et al., Early detection and classification of pathogenic fungal disease in post-harvest strawberry fruit by electronic nose and gas chromatography–mass spectrometry. Food Research International, 2014. 62(0): p. 162-168.

[3] Brudzewski, K., S. Osowski, and A. Dwulit, Recognition of Coffee Using Differential Electronic Nose. Instrumentation and Measurement, IEEE Transactions on, 2012. 61(6): p. 1803-1810.

[4] Schmekel, B., F. Winquist, and A. Vikström, *Analysis of breath samples for lung cancer survival.* Analytica Chimica Acta, 2014. **840**(0): p. 82-86.

[5] Westenbrink, E., et al., Development and application of a new electronic nose instrument for the detection of colorectal cancer. Biosensors and Bioelectronics, 2015. 67(0): p. 733-738.

[6] Cano, M., et al., Rapid discrimination and counterfeit detection of perfumes by an electronic olfactory system. Sensors and Actuators B: Chemical, 2011. **156**(1): p. 319-324.

[7] Loutfi, A., et al., *Electronic noses for food quality: A review.* Journal of Food Engineering, 2015. **144**(0): p. 103-111.

[8] Snell, B.Q.W.a.D., *An Introduction to Incremental Learning* Predictive Analytics and Futurism, 2016. **JULY 2016**(13).

[9] Tudu, B., et al., Electronic nose for black tea quality evaluation by an incremental RBF network. Sensors and Actuators B: Chemical, 2009. 138(1): p. 90-95.

[10] French, R.M., Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 1999. 3(4): p. 128-135.

[11] Wang, J., et al., *Deep learning for smart manufacturing: Methods and applications.* Journal of Manufacturing Systems, 2018. **48**: p. 144-156.

[12] Wan, S. and L.E. Banta, *Parameter Incremental Learning Algorithm for Neural Networks.* IEEE Transactions on Neural Networks, 2006. **17**(6): p. 1424-1438.

[13] Razavian, A.S., et al. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. in 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2014.

[14] Girshick, R., et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. in 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[15] James Kirkpatrick, R.P., Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A and K.M. Rusu, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. , Overcoming catastrophic forgetting in neural networks. 2016.

[16] Pang, Y., J. Deng, and Y. Yuan, *Incremental threshold learning for classifier selection.* Neurocomputing, 2012. **89**: p. 89-95.

[17] Yen, G.G. and P. Meesad. Constructing a fuzzy expert system using the ILFN network and the genetic algorithm. in Systems, Man, and Cybernetics, 2000 IEEE International Conference on. 2000.

[18] Tian, L. and W. Liu. Incremental intrusion detecting method based on SOM/RBF. in 2010 Girshick, R., et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. in 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[19] Myint, H.O. and P. Meesad. *Incremental Learning Algorithm based on Support Vector Machine with Mahalanobis distance (ISVMM) for intrusion prevention*. in *2009 6th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*. 2009.Yen, G.G. and P. Meesad, *An effective neuro-fuzzy paradigm for machinery condition health monitoring.* IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2001. **31**(4): p. 523-536.