Human breath-print identification by E-nose, using information-theoretic feature selection prior to classification

X. Rosalind Wang a,*, Joseph T. Lizier a, Amalia Z. Berna b, Florence G. Bravo b, Stephen C. Trowell b

a CSIRO Computational Informatics, P.O. Box 76, Epping, NSW 1710, Australia
b CSIRO Ecosystem Sciences and Food Futures Flagship, GPO Box 1700, Canberra, ACT 2601, Australia

A R T I C L E   I N F O

Article history:
Available online 6 November 2014

Keywords:
Classification
Electronic nose
Feature selection
Machine learning
Metabolic phenotype
Mutual information

A B S T R A C T

The composition of bodily fluids reflects many aspects of health status of a patient. Breath is another sample that may be useful for diagnosis of infectious and other diseases. Analysis of breath has the advantage of being less invasive than analysis of other fluids such as blood and bronchial biopsy. Two recent studies, using either mass spectrometry or electronic nose (E-nose) technologies, showed there are definite “breath-prints” that characterised individuals despite temporal variation in internal metabolism and environment. In this study we demonstrate that by employing an information-theoretic feature selection method that is specific to the problem together with machine learning techniques, we can dramatically improve (cross-validated) identification of individuals through their breath using a very small selected subset of E-nose measurement features. Indeed, we demonstrate here that we can identify the 10 individuals in this study with perfect accuracy using fewer than 10 features.

Crown Copyright © 2014 Published by Elsevier B.V. This is an open access article under the CC BY-NC-SA license (http://creativecommons.org/licenses/by-nc-sa/3.0/).

1. Introduction

Metabolomics is the study of metabolism at the global level and involves collecting data that represents a broad range of metabolites, the end products of cellular processes, from cells, tissues, organs or organisms [1]. Unlike genetic information, metabolomic data is influenced by environmental factors [2], individuals’ lifestyle and diet [3], etc, and thus varies in time. This poses the question of whether individuals’ metabolic phenotypes can be stable over time, which would be important in achieving personalised healthcare through metabolomics.

The study and analysis of exhaled breath is an attractive and promising area of metabolomics. Firstly, breath collection is totally non-invasive and safe, making it much easier to collect than bio-fluids such as blood and urine. Secondly, breath carries a large number of volatile metabolites [4,5], potentially providing researchers with relevant biochemical information. There are only a few reports in the literature about the variability of healthy human breath. One study using gas chromatography and mass spectroscopy (GC–MS) observed a total of 3481 different volatile organic compounds (VOCs) in the breath of 50 healthy humans, with only 27 of these VOCs observed in all the subjects [5]. Another study of 40 healthy individuals using GC-MS showed 618 compounds detected from all subjects, with only 35 of these found in all individuals [6]. These results demonstrated that there are large inter-individual variations. Additionally, a recent study [7] with 11 subjects using a quadrupole time-of-flight mass spectrometer (Q-TOF) showed there is a definite ‘breath-print’ for each individual over a period of 9 working days (on average 18 samples were collected from each subject).

While all the work discussed above showed promise for using breath as a means for achieving personalised medicine, the data were all analysed with mass spectroscopy instruments. Mass spectrometers (MS) allow detailed analysis of the composition of breath samples, but they are generally bulky, expensive and require the users to have expert knowledge. This makes them unsuitable for personal healthcare, especially in situations where immediate diagnosis is required. It is generally impractical to collect and send away breath samples for analysis by MS, and also where patients require personal portable machines for continuous monitoring. Electronic nose (E-nose) technology has been used to typify exhaled breath for research purposes [8], and its compactness and ease of use make it a viable alternative to MS instruments for point of care diagnosis through breath. In the past, E-noses have been tested for diagnosing lung cancer [9,10], asthma [11], urinary tract infections [12], tuberculosis [13], etc.
Nevertheless, there remains an important requirement in e-nose systems to determine the degree of variability in measuring exhaled metabolites from different subjects. Such information is essential as it will help support interpretation of whether an observed difference in repeated measurements is likely to represent a true change in clinical status.

In this paper, we show, by applying a feature selection method that is specific to the problem of classification, that we can select a very small subset of E-nose measurement features to achieve perfect classification performance.

Feature selection is becoming increasingly important with the rapid increases in speed and dimensionality of data acquisition in many fields. Feature selection is the act of reducing the dimensionality of data – be it from a sensor that collects a vast number of data points per sample, or data collected in parallel with many different sensors at the same time – to find the most useful subset of data for tasks such as classification, fault detection, or diagnostics. Previous work in applying feature selection to E-nose data [14,15] showed the best classification results can be achieved with a small subset of the full feature sets.

We use mutual information (MI), an information-theoretic approach to select features for classification. Mutual information is a measure of the amount of information one random variable contains about another [16]. The MI between two variables tells us the reduction in uncertainty of one due to the knowledge of the other. In the case of feature selection, this approach maximises the value of the information obtained for reliable and accurately accurate classification, rather than simply maximising the information collected. We recently showed that the method can select feature sets which achieve near-optimal classification performances [15] as compared with exhaustive search [14] but at a much reduced computational cost. Here, the selected E-nose measurement features were used as input for a variety of common classifiers: support vector machine (SVM), k-nearest neighbours (KNN), Bayesian networks (BN), neural networks (NN) and maximum likelihood (ML). Principal component analysis (PCA) was also applied to demonstrate the effectiveness of the feature selection method.

2. Materials and methods

In this section, we will first describe the procedure of collecting the breath samples and the instrument used to measure them. We will then describe the data processing steps from pre-processing, to selecting features from all the data collected, and then how classification of the identities of individuals was achieved. Finally, we will discuss in brief the hypothesis test we used for comparing classification results.

2.1. Experimental setup and breath collection

Our experiment analysed breath samples of 10 volunteers (5 males and 5 females) from the staff of CSIRO Ecosystem Sciences at Black Mountain (Canberra, Australia), all of whom were between the age of 18 and 45, healthy and are non-smokers. The research was approved by the CSIRO Animal, Food and Health Sciences Human Research Ethics Committee (LR04/2012). All subjects signed an informed consent form to participate in the research.

Volunteers breathed tidally through a mouthpiece connected to a three-way sliding valve (Hans Rudolph 2870K Series, Hans Rudolf, Kansas city, MO, USA). Neither a nose clip nor VOC filter were used. After a few seconds of breathing, volunteers were asked to perform a single slow exhalation and to open the slide valve so that only the last portion of the breath (alveolar air) was trapped in a 3-L impermeable gas bag (Shanghai Sunrise Instrument Co., China) connected to the sampling port of the valve.

Samples were collected between 08:00 and 10:00 and at least 1 h after breakfast. Three separate alveolar samples were taken per volunteer on each visit with at least 1 L of breath collected in each bag. A 2-min break was allowed between collections. Breath samples were collected from each subject over 4 visits, with at least one week between each subsequent visit. Therefore, a total of 120 breath samples were collected for the experiment.

Breath collection was performed on a daily basis with a maximum of two participants per day. All samples were stored at room temperature and analysed on the next day after collection.

2.2. Electronic nose measurement

All breath samples were analysed by the DiagNose (The eNose Company, The Netherlands) electronic nose instrument. The instrument contains 12 individual gas sensors consisting of six types of doped tin dioxide (SnO2) sensors and one type of tungsten trioxide (WO3) sensor. Table 1 lists the sensor types used in the E-nose and their sensor ID numbers.

<table>
<thead>
<tr>
<th>Sensor type</th>
<th>Doped</th>
<th>No. of sensors</th>
<th>Sensor ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>SnO2</td>
<td>Ag</td>
<td>1</td>
<td>S1</td>
</tr>
<tr>
<td>SnO2</td>
<td>Pd</td>
<td>1</td>
<td>S2</td>
</tr>
<tr>
<td>SnO2</td>
<td>Cu</td>
<td>1</td>
<td>S8</td>
</tr>
<tr>
<td>SnO2</td>
<td>PtAg</td>
<td>1</td>
<td>S12</td>
</tr>
</tbody>
</table>

The E-nose sampled the expired breath directly from the bags. For each sample, 200 ml of breath was analysed at a flow rate of 40 ml/min (each analysis lasted 5 min). The DiagNose instrument measures the sample every 20 s, henceforth called “one time index”. During each 20 s period, the sensor surface temperature was modulated with a 32 step sinusoidal signal between temperatures of approximately 260 and 340 °C, starting at ~305 °C where the temperature increases to maximum then decreases to minimum before finally increase again to ~300 °C. Fig. 1(a) shows the responses of the different sensors as temperature modulates over the 20 s period, the approximate temperature at each modulation is shown on the top x-axis. Fig. 1(b) shows the sensor responses against the actual temperatures.

After each analysis of a breath sample, the instrument was purged with instrument air (hydrocarbon free air: 20.9% oxygen, 78.1% nitrogen and 0.9% argon) at a flow rate of 300 ml/min for 35 min. This step allows the instrument to return to baseline for the next breath analysis. Fig. 2 shows sample raw data from one temperature measurement of one sensor (i.e. one of the 32 measurements per time period for the one sensor, which is a single data point in Fig. 1) for approximately 35 min. In this figure, the breath analysis occurred between 11:35 and 11:40 h, and from 11:40 onwards is the baseline recovery phase as the readings of the instrument return to baseline.

2.3. Data pre-processing

We pre-process the raw DiagNose data to correct for two inconsistencies: Firstly, we found the sensor measurements sometimes do not occur at the prescribed 20 s interval. Thus, instead of extracting n samples for analysis, we extract the data according to their timestamps in relation to the peak time. The peak time is the end of the 5 min analysis time and is recorded separately. We extract 300 s of data before and including the peak plus another 100 s after the peak. We then interpolate the possible measurement at exactly 20 s interval. Therefore, we extract data for 20 time indices per sample.
The second step in data pre-processing is to correct for a day of analysis effect. That is, raw data returned from analysis by the E-nose within the same day across different subjects have significant commonalities, and can be seen as clustered together in a PCA analysis. To correct for this, the data analysed on the same day is first normalised by their mean and standard deviation (i.e. $z$-scored): \( (x - \mu) / \sigma \), where \( x \) is a data sample, \( \mu \) is the mean of the data, and \( \sigma \) is the standard deviation. We $z$-scored the data by each sensor-temperature, that is, we take all the readings measured by one sensor at one temperature point within the same day as a separate set of data for normalisation.\(^1\)

2.4. Feature selection

The optimal feature selection task is to select a subset of features, \( v \subseteq \{1, 2, \ldots, m\} \), where \( m \) is the total number of features available, such that the resulting subset of features gives the best classification performance for the given size constraint \( n \) on the number of features in \( v \).

A common approach is to select the subset of features by maximising the mutual information \( I(Z^v, C) \) [16] between the selected features \( Z^v = (Z^v_1, \ldots, Z^v_n) \) and class \( C \):

\[
I(Z^v, C) = \sum_{Z^v, C} p(Z^v, C) \log \frac{p(C | Z^v)}{p(C)}. \tag{1}
\]

This approach was suggested by Battiti [17], since it minimises the uncertainty (entropy) \( H(C | Z^v) \) about the class given the features.

Challenges in evaluating Eq. (1) here for a given size constraint \( n \) include: (i) estimating the multivariate joint and conditional density functions for a small data set (\( N = 119 \) for each training set); and (ii) selecting \( n \) from large number of features to choose from (12 sensors $\times$ 32 temperature measurements $\times$ 20 time indices $= 7680$ features, giving \( \binom{7680}{n} \) possible combinations).

We take a simple approach to address these issues. To select \( n \) features, we take those with the highest individual mutual information to the class [18]. This means that we make only \( n \) mutual information calculations for single features, thereby avoiding multivariate joint and conditional density function calculations (for which measurement uncertainty increases with a small data set) and bypassing the combinatorial explosion of possible feature sets as \( n \) increases. This approach can be viewed as a simplification of Battiti’s approach [17] by effectively assuming that the information provided by each feature about the class is independent, i.e.:

\[
I(Z^\circ, C) = \sum_{j=1}^{n} I(Z^\circ_j, C). \tag{2}
\]

Of course, such an approach does not account for redundant information about the class held by multiple features, nor does it capture synergies held jointly in multiple features about the class (see formal definitions of redundancy and synergy in mutual information by Williams and Beer [19]). One can of course compute MI using Eq. (1) for the candidate sets of \( n \) variables (e.g. [15]), or use pairwise approaches coupled with forward feature selection (e.g. by Battiti [17], and the mRMR [20] and QPFS [21] methods). In general

\[^1\] There are many ways to $z$-score this data, for example, $z$-score within each sample, $z$-score of all data collected on the same day, $z$-score of all data collected on the same day and within one sensor, $z$-score of all data collected on the same day and within one temperature measurement of one sensor, etc. We found the classification performance of the last two $z$-scoring approaches perform the best and report here the result from the last type of $z$-scoring only.
these will be more accurate, but are more computationally expensive. As we shall see in the Results section, our simple approach works very well already on this data set, though for more complicated applications it may be pertinent to use better approximations to the joint mutual information as discussed above.

Finally, we employed the Kraskov–Grassberger technique [22] (estimator 2) for estimating the mutual information. This technique uses dynamic kernel widths and bias correction, which provide robustness against small data size. The code used for the estimation is the publicly available Java Information Dynamics Toolkit [23].

2.5. Classification

All results reported in this paper were obtained from one-against-all (leave-one-out) cross-validation. That is, the full data sets were partitioned into 120 pairs of a training set (of 119 data samples) and a test set (of the remaining 1 data sample). This process is repeated 120 times to cover all data samples in the data set. Feature selection is performed on each training set to select the features to train a given classifier, then the classifier is applied to the test set. The performance of a given classifier is the average correct classification rate it obtains over these 120 training-test pairs.

We used several common classifiers to evaluate the effectiveness of the feature selection.

A support vector machine (SVM) [25] constructs a hyperplane to separate training data in different classes. We used the libsvm library [26] to perform the classification using C-SVC (SVM classification with cost parameter of C). Nine C values, starting with 1 and increasing by a factor of 4 for each subsequent value, were tested (i.e., $C = \{1, 4, 16, 64, 256, 1024, 4096, 16384, 65536\}$) to capture the classifier’s performance over large range of C values. Two types of SVM were used: linear and a nonlinear SVM using Gaussian radial basis kernel function. The Gaussian radial basis kernel function is, $k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, where $x_i$ is the vector of the data for sample $i$ and $\gamma$ is the kernel width. We selected $\gamma$ using the inverse of number of features (the default $\gamma$ value in libsvm). As previous studies showed this gives the best classification performance [15].

We found in both types of SVM, the performances do not improve for $C > 256$, and there is no significant difference in performance between the two types of SVM. Therefore, we show here classification performance of linear SVM with $C = 256$ in comparison with the other classifiers.

The $k$ nearest neighbour (kNN) algorithm compares the input data with an existing set of training data by computing a distance metric [27]. The neighbours of the input data are the $k$ data points with the smallest distance metric. The input data’s class is determined to be that with the most data points in the neighbourhood. We tested the neighbourhood size for $k = \{3, 5, 7, 9, 11\}$. The maximum value of $k$ is set at 11 since there are 12 samples of each class in the data set, so for any test data (with our leave-one-out cross-validation) there will be at most 11 samples of the same class in the training data that can appear in the neighbourhood. We found there are no differences in classification performance between any of the $k$ values, therefore, we show here results from $k = 7$.

In Bayesian networks (BN) [28], each node represents a random variable (these could be observed data or latent variables), $X$, with associated probability functions, $p(X)$. The nodes in a BN are linked by directed edges which indicate the conditional dependency between the variables, for example, if the nodes $X_1$ and $X_2$ in a BN are linked by a directed edge from $X_1$ to $X_2$, then $X_2$ is dependent on $X_1$ and $X_1$ is the parent node to $X_2$. Lastly, the network as a whole is acyclic. We used the BNT toolbox [29] to implement the network and perform the classifications. We implemented a Naive Bayes Network, the simplest form of BN for classification, since the features were not selected for their dependencies. In a Naive Bayes Network, a node representing the class is parent to all other nodes each of which represents a feature, and there are no edges between the child nodes.

An artificial neural network (NN) [30] consists of a group of nodes, or neurons, that are interconnected. We used here a feedforward NN called multilayer perceptron (MLP) to map the feature data (input neurons) onto the classes (output neurons) through two layers of hidden neurons. The MLP was implemented using the Netlab toolbox [31] with maximum of 100 iterations using three optimisation algorithms (quasi-Newton, conjugate gradients, and scaled conjugate gradients) and 6 or 12 hidden nodes. We found the MLP with 12 hidden neurons using quasi-Newton optimisation performs the best. We further tested MLP using quasi-Newton optimisation for 6, 8, 10, 12, 14, and 16 hidden nodes and found the classification performance using 12, 14 or 16 hidden nodes do not differ significantly and are better than the rest. Therefore, we report here results from MLP with 12 hidden neurons using quasi-Newton optimisation.

Mutual information maximum likelihood (MI ML) makes a classification as the class c which maximises the local or pointwise mutual information $I(Z^* = z^*; C = c)$ with the observed (selected) features, where the pointwise mutual information is $I(Z^* = z^*; C = c) = \log p(z^* | C = c) / p(z^*)$ [32]. This approach is equivalent to selecting the class which maximises the probability $p(z^* | C = c)$ of the observation given the class, and is also equivalent to a classic maximum likelihood classifier (maximising the posterior $p(C | z^*)$) under the assumption of the marginal distribution for each class $p(C)$ being equiprobable [33]. The mutual information here was evaluated using Kraskov-Grassberger estimation [22], using the Java Information Dynamics Toolkit [23].

2.6. Hypothesis test

The two-sided two-sample binomial test is used to test the hypothesis that the classification performances of any two classifiers (using the features selected by the MI criterion) are equal. If the hypothesis is true, then the classification performances of the two classifiers being tested, $p_1$ and $p_2$, are equal with a common value of $p$. Since $p$ is unknown, we estimate its value by

$$\hat{p} = (n_1 + n_2) / (n_1 + n_2)$$

where $n_1$ and $n_2$ are the number of data samples, and $X_1$ and $X_2$ are the number of successful classifications by the classifiers. The $p$-value is approximately (for two-sided test) [34]:

$$P = 2 - 2 \Phi \left( \frac{|p_1 - p_2|}{\sqrt{\hat{p}(1 - \hat{p}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \right),$$

where $\Phi(y)$ is the cumulative distribution function of $\Lambda(0, 1)$ evaluated at $y$. We reject the null hypothesis that the classifiers’ performances are equal if $P < 0.05$.\footnote{To ensure we did not overfit [24], we have also performed cross-validation where each test set contained all 3 samples collected from a subject on the same day (whilst the corresponding training set contained all other 117 samples). We found there was no significant difference between the classification results using this cross-validation, therefore, we only report here the results of one-against-all cross-validation.}
3. Results and discussion

3.1. Mutual information analysis

We first calculate the mutual information between individual features and the classes (i.e. individual subjects) using all 120 breath samples, to observe where (in terms of time indices and sensors) any information about the class is situated. Fig. 3 shows the mutual information results as a heat map for all features and all sensor responses to the breath samples for the first 21 time indices (7 min after breath sample is first introduced to the instrument).

From Fig. 3 we can see a couple of interesting aspects: Firstly, most of the highest MI between features and class occurred at earlier time indices \(T = 1 \) and \(2 \) instead of around maximal response of the sensors \(T = 15 \). Further, we observe significant information about the class in the sensors’ responses at later time indices \(T = 19 \) and \(20 \), which occurs during baseline recovery phase of analysis. These MI results are contrary to both our intuition and the common practice in E-nose literature of selecting features near maximal sensor response \([35]\). Secondly, while sensors of the same type (that is, sensors \(\{3, 4, 5\}, \{9, 10, 11\}, \) and \(\{6, 7\}\) have similar results in Fig. 3, there are still some differences in the amount of information they hold about the class. This difference is especially obvious for the 2 WO3 sensors \(S6 \) and \(S7\). This could be due to the differences in these sensors when they were manufactured, though a more likely reason is that the physical location of these sensors in the instrument (they are not placed next to each other) plays a role here. Lastly, Sensor 8 (SnO2 doped with Cu) shows no information about the classes for the second half of the temperature cycle (when sensor surface temperature is approximately 260–300 °C), which indicates this sensor do not respond differently to the individuals’ breath samples at low temperature.

3.2. Feature selection for training sets

Since the mutual information calculations from the previous section (Fig. 3) used all 120 data samples of data, the features with the highest MI there are not necessarily those which will be selected for the training sets in our cross-validation. In this section, we examine the features selected from the training sets (each containing 119 out of the 120 samples, 120 training sets in total).

Fig. 4 shows the features selected from all the training sets for six different size constraints: 5, 10, 20, 40, 100 and 200. The features selected are shown as coloured cells according to their percentage of selection in all the training sets. A quick glance of the figures shows that almost all the features selected were from the first, and to a lesser extent second, time index. This agrees with the MI map in Fig. 3, where almost all the information about the class was found in the first two time indices.

Fig. 4(a) shows the features selected with a size constraint of 5. Here, nine different features were chosen from all the different training sets (see Table 2 for the break down), as each training set is subtly different from the others thus different features could be selected. From the figure, we can see that these features were chosen from three different sensors: S2 (SnO2 with +Pd doping, four features selected), S12 (SnO2 with PtAg doping, four features selected), and S8 (SnO2 with Cu doping, one selected feature). From Table 2, we can see that out of the nine features, only one was selected from all the training sets (S12, temperature point 27 at time index 1). In the remaining eight features, four were selected across more than 80% of the training sets. Further, the four less popular selected features are from neighbouring temperature points of the five more commonly selected features. This is not unexpected, as sensors have similar responses when operated at similar temperatures.

Fig. 4(b) shows the features selected with size constraint of 10, doubling the number of selected features from Fig. 4(a). In this new...
Fig. 4. Features selected by maximal MI using one-against-all cross validation. Each sub-figure shows the selected features for certain size constraints (5, 10, 20, 40, 100 and 200). The selected features are coloured according to the number of times they were selected for in the cross-validation, expressed as a ratio over the full size of the data set (120 samples). That is, a feature coloured in darkest red was selected from all 120 training sets, whereas a feature coloured in dark blue was only selected from less than 10% of the training sets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Features selected for size constraint of 5. Each row represent a selected feature: its sensor number, temperature index (and approximate temperature) and time index, and the number of training sets that selected the feature. This information is represented graphically in Fig. 4(a).

<table>
<thead>
<tr>
<th>Sensor No.</th>
<th>Temperature index</th>
<th>Approximate temperature</th>
<th>Time index</th>
<th>% in Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>305</td>
<td>1</td>
<td>82.50</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>310</td>
<td>1</td>
<td>6.67</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>325</td>
<td>1</td>
<td>15.83</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>330</td>
<td>1</td>
<td>89.17</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>275</td>
<td>1</td>
<td>16.67</td>
</tr>
<tr>
<td>12</td>
<td>22</td>
<td>270</td>
<td>1</td>
<td>4.17</td>
</tr>
<tr>
<td>12</td>
<td>27</td>
<td>275</td>
<td>1</td>
<td>100.00</td>
</tr>
<tr>
<td>12</td>
<td>29</td>
<td>285</td>
<td>1</td>
<td>85.83</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>320</td>
<td>2</td>
<td>99.17</td>
</tr>
</tbody>
</table>
size constraint, six new features have been added to the set. Comparing the two feature sets with size constraints of 5 and 10, we find almost all of the features in the smaller set are now selected in all training sets for the larger feature set. When the size constraint is doubled to 20 (Fig. 4(c)), the features selected in size constraint of 10 are now selected in all training sets. This trend continues as the size constraint increases. This shows that while there are subtle differences between the training sets resulting in different feature sets, the features selected are still some of the ones with the most information about the classes.

An interesting aspect of the first three subfigures in Fig. 4 is that the selected features are concentrated in only three sensors. This is despite the fact that a total of 26 different features were chosen in size constraint of 20, the only exception is a feature in sensor S7. As shown in the next section, we only needed a small set of features to give excellent classification results for this data set. Having only three sensors chosen in the feature set is advantageous if a portable instrument was to be manufactured for cost reduction.

Fig. 4(e) and (f) shows the selected features for size constraints of 100 and 200, confirming a few observations from the MI plot for the full data set in Fig. 3. Firstly, we see from these figures that even at large size constraints, the features are almost chosen all from the first two time indices (apart from several features in the later time indices chosen in sensor S6). Secondly, while there are triplicate and duplicates of some sensor types, the features selected from the individual sensors (e.g. S3, S4 and S5 are all Pd doped SnO2) are slightly different. This latter observation could be due to subtle difference in the manufacturing of the sensors, or the placement of the sensors within the instrument. While these selected features do not make any difference to the overall classification results in this data set (see below), we believe it should be subjected to future research.

3.3. Classification

Fig. 5 shows the classification results of five different classifiers using the features selected by MI method for size constraints up to 60. For each value on the x-axis (number of selected features) in the figure, we:

1. select feature sets according to the size constraint for each training set,
2. we then train a model using only the data from the selected features for each training set,
3. finally, we classify the test data using the models learned from the corresponding training set, and report the average classification accuracy from all 120 training-test pairs.

The first observation we can make from Fig. 5 is that the classification results become very good very quickly – four out of the five classifiers achieved, or almost achieved, 100% accuracy with fewer than 10 features. Further, from the feature analysis in the previous section, we know that these features are concentrated in three, at most four, sensors and all taken at the first 2 time indices. This result is surprising and counter-intuitive, as we might have expected the selected features to be around the maximal response, and that a good classification result would only be achieved using a large number of features from many of the sensors.

Out of the classifiers tested, we see that four of them (SVM, kNN, MI ML, and BN) have very similar classification performances. Indeed, performing a two-tailed binomial test between the results (see Section 2.6), we found there is no statistical significance between these classification results. Of the classifiers tested, NN has the worst performance, but using the same hypothesis test

---

**Fig. 5.** Cross-validated results for classifying the identity of an individual using the features selected by maximal mutual information for size constraint up to 60. Five classifiers were tested: support vector machine (SVM), k-nearest neighbours (kNN), mutual information maximum likelihood (MI ML), neural networks (NN), and Bayesian networks (BN). Notice the y-axis shows the performance in a highly non-linear, logarithmic scale.

---
we found this difference in performance is also not statistically significant when corrected for multiple comparisons. Further we compared the average performances between the classifiers and found no statistical significance in the difference in performance. In our previous work [15], where we applied a similar feature selection method and the same classifiers to a different set of data measured by a different E-nose instrument, the BN and SVM classifiers performed better than the rest, while NN was not the worst classifier tested. While this is only a very small sample set (2 different data sets), comparing the results here with our previous work, we can tentatively conclude that classifiers’ performance is dependent on data sets and it would be advisable to test a variety of classifiers for each new type of data set before selecting the one with the best performance.

3.4. Principal component analysis

Fig. 6 shows the first three principle components of the data set for all data and for the first 10 features selected via maximal mutual information. The principle components were calculated based on the covariance matrix of the data. We use the first 10 features selected using all the data (i.e. Fig. 3) as we see in Fig. 4 the features selected through training sets do agree with each other as the size constraint increases. We chose the size constraint to be 10, as we see in Fig. 5 that at 10 features, the classifiers have reached their optimal performance.

From Fig. 6 we can see how important it is to select a subset of the features for classification: In the PCA plot for all data, we cannot see distinct clusterings of samples into each separate class, however, in the PCA plot using only the top 10 features, the samples very clearly separate into clusters for each individual subject. There remains however, some variance in these clusters – while the samples from subjects 1, 2, 5 and 8 are almost on top of each other in the PCA plot, the other subjects do have samples outside their main cluster. The latter might also explain why the classifiers did not achieve 100% accuracy on cross-validation at some size constraints. By examining the mis-classification at various feature size constraints, we found the most common subject to be mis-classified
is subject 10, followed by subjects 6, 7 and 9. From Fig. 6b, we can see that these are all clusters with one or two samples away from the rest of the class.

4. Conclusion

We applied an information-theoretic method of feature selection to a set 120 samples of E-nose data collected from 10 different individuals, and then attempted to classify the identity of these individuals using the selected features. Our cross-validated classification result of 100% accuracy with fewer than 10 features shows that the E-nose can distinguish between different individuals. Our result is better (using the same hypothesis test and P < 0.05) than a previous study using mass spectrometry on 193 samples collected from eleven individuals where the authors achieved an over 76% classification accuracy [7] using a combination of Kruskal–Wallis/PCA/CA for feature selection and kNN for classification. Further, our result is also much superior than using the Kruskal–Wallis/PCA/CA method, where the error rate was over 90% on the same data set [24]. There are several reasons for the difference in results here: (1) the Kruskal–Wallis test is a linear approximation of our method and it does not use the values of the data in its calculations; (2) the mutual information method in this work is a direct measurement of the information contained in the features about the class; further, (3) the Kraskov–Glassberger technique for estimating the probability distribution of the data is robust against small data sets and outliers in data sets.

We found by selecting features that contain the most information about the classes (individual subjects’ IDs), we were able to rapidly improve the classification result with very few features (fewer than 10) concentrated into three out of the twelve sensors used in the instrument. Therefore, we demonstrated in this work that the E-nose instrument, a simpler and cheaper tool than the mass spectrometer, can be used to investigate and identify breath prints from different individuals. By reducing the number of features to a very small subset, our work here shows the possibility that ultimately a device might be built to allow individual patients to easily monitor their health at home and contribute to personalised healthcare.

Acknowledgements

We would like to thank CSIRO Advanced Scientific Computing (http://www.hpsc.csiro.au) for the use of their supercomputer clusters in performing the experiments for this paper.

We would also like to thank Thomas Nowotny of Sussex University, UK for providing the software to extract and pre-process raw DiagNose data; Nadja Wiziack and Benjamin Padovan for their contribution to the breath collection method.

This research was supported by the Science and Industry Endowment Fund (http://www.sief.org.au). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.snb.2014.09.115.

References


Biographies

X. Rosalind Wang studied Bachelors of Engineering and Science at the University of Sydney, Australia, where she also received a PhD in engineering for Mechatronics Engineering in 2008. Since joining CSIRO as a Research Scientist, she has applied her skills in machine learning to a diverse field of projects ranging from energy management, anomaly detection in sensor networks, medical diagnostics, swarm intelligence and artificial life. Rosalind’s main research interest is machine learning for complex systems.

Joseph T. Lizier is currently a Research Scientist at CSIRO in Sydney, Australia (since 2012). Previously he was a Postdoctoral Researcher at the Max Planck Institute for Mathematics in the Sciences in Leipzig, Germany (2010-12), and a Senior Research Engineer at Telstra Research Laboratories (2001-06) and Seeker Wireless (2006-10) in Sydney. He received a PhD in Computer Science (2010), and Bachelor degrees in Engineering (2001, with University Medal) and Science (1999), all from the University of Sydney. His research interests include information-theoretic approaches to complex systems, complex networks and computational neuroscience.

Amalia Z. Berna received her engineering degree in food technology from the Agrarian National University, Peru and PhD in bioscience engineering from Catholic University of Leuven, Belgium. Amalia is a Senior Research Scientist at CSIRO. Amalia leads the volatiles and biomarker discovery component of the Innovative Bioproducts group, her research focuses on the detection of low abundance volatiles release above food, plants and in human’s breath, with the aim of providing faster tools for predicting quality and health. She has experience in the application of multivariate statistical methods to sensors and sensor arrays. Amalia is author of over 25 refereed international journal papers with 460 life citations and inventor on one patent family.

Florence G. Bravo obtained her engineering degree in materials sciences from the National Institute of Applied Sciences in France. Florence has worked for 8 years in pharmaceutical, petroleum, agribusiness and plastic industries where she gained extensive experience in the formulation and processing of polymers. Florence joined the Innovative Bioproducts group at CSIRO in 2012 as Research Projects Officer. As part of Amalia Z. Berna’s team, Florence has been involved in the collection and chemical analysis of volatile compounds in human breath samples.

Stephen C. Trowell studied biochemistry at the University of Cambridge and, in 1987, completed his Ph.D. on the biochemistry of visual transduction at the Australian National University. In 1989, Stephen joined CSIRO, Australia’s leading scientific research organisation. His current role is Group Leader in the Food and Nutrition Flagship where he leads a team developing biosensors to measure the quality of foods and beverages. Stephen has over 100 publications and patents. One of the most rewarding aspects of his role is bringing scientists together from different disciplines to focus on a common goal.