Development of a Complex Adaptive PNN System for the Rapid Detection of E.coli

Yinglei Li\textsuperscript{a}, William Ford\textsuperscript{a}, Kun Xiang\textsuperscript{b}, Walker Land\textsuperscript{a}, Robert Congdon\textsuperscript{b}, Omowunmi Sadik\textsuperscript{b}

\textsuperscript{a}Dept. of Bioengineering, State University of New York at Binghamton, NY, 13902, USA
\textsuperscript{b}Dept. of Chemistry, State University of New York at Binghamton, NY, 13902, USA

Abstract

The objective of this research is to develop a complex adaptive piecewise linear regression / probabilistic neural network (PNN) intelligent system for the rapid detection and classification of Escherichia coli (E.coli). The rapid detection and classification of E.coli is important because current methods require a long period of analysis before a classification can be determined. The objective of this paper is to describe the design and preliminarily evaluate an Intelligent Decision Support System (IDSS) that will validate the following hypotheses: an intelligent decision support system (IDSS) to allow the rapid collection and classification of E.coli can be designed and preliminarily evaluated, which will significantly decrease detection and classification times for E.coli bacteria, thereby addressing the food spoilage problem. The research in this paper provides a preliminary answer to: What performance improvement percentage can be realized against the 16 to 48 hours required for the conventional multistep methods of detection of microorganisms (using E.coli data as a baseline)? For the 16 hour period we have a 6.7% reduction in the time-to-detect period ($(16-15)/15 \times 100\% = 6.7\%$) and for the 48 hour period we have a 220% reduction in time ($(48-15)/15 \times 100\% = 220\%$).

Keywords: Escherichia coli; classification; probabilistic neural network (PNN)

1. Introduction

According to the Center for Disease Control, millions of people suffer from food-borne pathogens every year resulting in approximately 128,000 hospitalizations and 3,000 deaths [1,2]. Salmonella, Listeria, Staphylococcus, Norovirus and E. coli are the major food-borne pathogens [3,4]. Hence, rapid and accurate determination of these pathogens is important to prevent infection of food, water and other samples.

The conventional methods for detection of microorganisms are limited by the multi-step, time-consuming process. Completion of all phases typically requires at least 16 hours and can take as long as 48 hours. The detection is usually $10^5$-$10^6$ cells/mL without pre-enrichment. Novel pathogenic biosensors are needed to either replace the conventional, labor-intensive cell culture techniques or the process must be automated. Sadik et al. have previously reported the development of electrochemical detection with pattern recognition techniques for the detection and classification of bacteria at subspecies and strain levels [5, 6, 7]. Oxygen is a key parameter in aerobic systems: the level of oxygen consumed by the cells provides information on cell viability. The approach exploits the fact that under identical experimental conditions, various bacteria consume oxygen at different rates, which implies that electrical-current-time profiles are dependent on species and can provide distinguishable patterns.
To test this conjecture, data from two types of E.coli were collected and used as inputs to an intelligent decision support system (IDSS) designed for the quick classification of these types of bacteria. Two types of E.coli were purchased from American Type Culture Collection (ATCC, Manassas, VA, USA): E.coli type A (ATCC#25922) and E.coli type B (ATCC# 11775) to evaluate this IDSS.

2. Data collection, system IDSS overview and operational description

For each bacterial culture sample we collected voltage-domain curves hourly for 24 hours by observing the current change (delta) as voltage was swept through a range of values using a technique known as square wave voltammetry. Each curve has an E.coli peak and an oxygen peak as shown in Figure 1. By plotting the E.coli and oxygen peaks across time, the data are reduced to two time-domain curves as shown in Figures 2 and 3.

![Figure 1. Raw data of a sample of E.coli type A at 1 hour](image)

![Figure 2. Time-domain curve of E.coli peak](image)

![Figure 3. Time-domain curve of oxygen peak](image)

The analysis approach seeks to characterize the time course of the peaks by fitting these points with piecewise linear regression, and using the regression model parameters as features to distinguish different bacterial species. Each time / delta current -domain curve was partitioned into four temporal regions: \{0 – 5; 5 – 10; 10 – 15; 15 – 23 hours\}. The general form of the equation is:
\[ y = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot (t - 5) \cdot \Delta_1 + \beta_3 \cdot (t - 10) \cdot \Delta_2 + \beta_4 \cdot (t - 15) \cdot \Delta_3 \]

where \( y \) is the delta current, \( \beta_0, \beta_1, \beta_2, \beta_3 \) and \( \beta_4 \) are the coefficients, \( t \) is time

The above equation operates under the following constraints: If \( t > 5 \), \( \Delta_1 = 1 \); otherwise 0. If \( t > 10 \), \( \Delta_2 = 1 \); otherwise 0. If \( t > 15 \), \( \Delta_3 = 1 \); otherwise 0. Thus, each sample has a total of ten coefficients: \( \beta_{E0}, \beta_{E1}, \beta_{E2}, \beta_{E3}, \beta_{E4} \) for the time-domain curve representing the E.coli peak and \( \beta_{O0}, \beta_{O1}, \beta_{O2}, \beta_{O3}, \beta_{O4} \) for the time-domain curve representing the oxygen peak. Equation 1 was modified to included discontinuity, which degraded system performance as depicted in Table 1. That modification is not discussed here.

An automated hardware system was designed and implemented that is faster and less error-prone than the previous manual data collection process. The data collected by this automated system was used as an input to the piecewise linear regression process that extracted pertinent mathematical features. These features were then used to train and validate a specifically designed Bayesian PNN that was used in the E.coli classification process. The data processing flow is shown in the system overview Figure 4 below. Interested readers should consult Specht [8] for more information on the design and implementation of the PNN.

As described in previous work [5, 6, 7], the results of principal component analysis (PCA) indicated that E.coli peak and oxygen peak are the discriminating features in these cases. Therefore, the maximum delta current of the E.coli peak and the oxygen peak of each sample in time-domain were input features in the piecewise linear regression model.

The coefficients \( \beta_{E1}, \beta_{E2}, \beta_{E3} \) and \( \beta_{E4} \) correspond to the rates of change of the E.coli peak, while \( \beta_{O1}, \beta_{O2}, \beta_{O3} \) and \( \beta_{O4} \) correspond to the rates of change of the oxygen peak. Generally, the intercepts (\( \beta_{O0} \& \beta_{E0} \)) are not discriminating features and so are not used. To test discrimination at 5 hours, we used only \( \beta_{O1} \) and \( \beta_{E1} \); for 10 hour discrimination we use 4 coefficients (\( \beta_{O1}, \beta_{O2}, \beta_{E1}, \beta_{E2} \)), and so forth. These features defined the structure of the Bayesian PNN input layer. These feature sets were then normalized for ease of PNN training / validation data processing, using these two types of E. coli.

2.1 Summary of Bayesian PNN training

The Bayesian PNN was trained using the following Evolutionary Programming / Evolutionary Strategies (EP / ES) process, which is based on the original work of Fogel [9]. The evolutionary programming process as implemented in this study evolves the parameters (\( \sigma \)'s) for a population of PNN models. A generic description of this process is as follows [10, 11]:
A population of candidate solutions (PNN parameters) is randomly generated using the process shown in figure 5. Each of these candidate solutions then is copied and mutated, yielding a solution pool of twice the original size, using the expressions given below:

\[ v'_i = v_i e^{\sqrt{2n} N(0,1) + \frac{1}{\sqrt{2n}} N_i(0,1)} \]

where \( n \) is the total number of configurable parameters being evolved, \( N(0,1) \) is a standard normal random variable sampled once for all \( n \) parameters of the \( v \) vector, and \( N_i(0,1) \) is a standard normal random variable sampled for each of the \( n \) parameters in the \( v \) vector.

The second step of this mutation process comprises the updating of each configurable parameter for all elements of the evolving population. If we let the vector \( x_i \) denote these elements for each of the individual member of the population, this update process will be accomplished as follows:

\[ x'_i = x_i + C v'_i \]

All elements of this pool are scored using an objective function, for the purpose of finding the “best fit” \( n \) elements contained in the population set of \( 2n \) elements (note: \( C \) is a value sampled from a Cauchy distribution). The objective function scores are then used to order the candidate solutions from the “most fit” to the “least fit.” Better results usually are obtained from using tournament selection methodologies. With tournament selection, each candidate solution competes against a random subset of the remaining solutions. Finally, the upper 50% of the solution pool is selected to continue as (discarded) to reduce the pool to the original population size of \( n \) elements. This process is repeated for a specified number of generations, unless some other “stopping” criteria is used.

3. Data processing and results

This section addresses the data collection and analysis to support the hypotheses: an intelligent decision support system (IDSS) that will significantly decrease detection and classification times for E.coli bacteria can be designed and preliminarily evaluated. The process to evaluate these hypotheses used a limited data set size of E.coli type A and B. This dataset, because of its limited size of 24 samples, was processed using two separate cross-validation methods, which were 3-fold cross validation and 1-hold out cross validation. The K-fold (K=3) cross validation was performed to ascertain the collective validation performance of a small data set, while 1-hold out method provided some measure of the best possible performance for this system using this limited amount of information. Three-fold cross validation means that 2/3s of the data were used for training and 1/3s for validation. This process is repeated two additional times, where a different 2/3s and 1/3s of data were held out for training and validation, respectively. Finally, the results of the separate 3 validation set results are averaged for a representative estimate of system performance.
We used 1-hold out cross validation to obtain the best possible estimate for the PNN kernel density training parameters and validation accuracy, as we can show these sigma parameter (denoted by $\sigma_i$ in equation 3 above) were better (more accurately estimated) as the sample size used for training increases. One-hold out cross validation process functions as follows: each sample is individually held out and the remaining 23 samples are trained by the EP/ES training process described above. This process was repeated 23 times, once for each held out sample, and the 24 validation results averaged for an estimate of system performance. Here 24 separate PNNs were trained to provide the separate 24 validation performance results, which in theory, maximizes performance for this IDSS for this limited data set size of 24 samples.

3.1 Analysis of 3-fold and 1 hold out cross validation result

The PNN was trained using 3-fold and 1-hold out cross validation as described for 10, 15 and 24 hours. The results are depicted in Table 1 below.

Table 1. Preliminary 3-fold and 1-hold out cross validation results for recorded data

<table>
<thead>
<tr>
<th>Evaluation types</th>
<th>Average accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 hours</td>
</tr>
<tr>
<td></td>
<td>train</td>
</tr>
<tr>
<td>Using equation 1 to get coefficients</td>
<td>1-hold out</td>
</tr>
<tr>
<td></td>
<td>3-fold</td>
</tr>
<tr>
<td>Using with discontinuity variable</td>
<td>1-hold out</td>
</tr>
<tr>
<td></td>
<td>3-fold</td>
</tr>
</tbody>
</table>

Two factors are in play here: (1) the amount of information, as represented by the sample size, and (2) the accuracy with which the E.coli and oxygen peak behavior as measured by the regression curve fit coefficients. This accuracy is measured by the amount of recorded time used to ascertain the $\beta$ coefficients.

The 70.8% validation result in Table 1 for 1-hold out cross validation is smaller than the 79.2% validation result for 3-fold process for the 10 hours, which demonstrates that the recorded time and/or the sample size must be increased. Also the significant performance degradation with the discontinuity variable included (see Table 1) for the validation process demonstrates the significance of “getting” the curve fit representative accurately modeled. Consequently, this discontinuity variable was excluded from any additional analysis.

Referring again to Table 1, we observed that after 15 hours of recorded E.coli and oxygen peak data, one-hold out and 3-fold cross validation accuracy is 83.3% and 75%, respectively, with corresponding training accuracy of 89.1% and 87.5%, respectively. These preliminary results suggest that collecting the data for 15 hours (for this small size of samples) provides reasonably accurate results. Using 15 hours as the baseline, we now ask: What is the percentage of improvement in performance against the 16 to 48 hours required for the conventional methods of detection of microorganisms (using E.coli data as a baseline)?

For the 16 hour period we have a 6.7% gain in time ($(16-15)/15 \times 100\% = 6.7\%$) and for the 48 hour period we have a 220% time improvement ($(48-15)/15 \times 100\% = 220\%$). This time improvement, while encouraging, are not as
large as the 100% to 500% improvement obtained from the “A Multi-class PNN for Pathogen Classification” research study by William Ford, et.al. (this issue of CAS 2013) because (we believe): (1.) the normalization of the difference curves method provides more discriminating information than does linear regression, and (2.) that approach had a ~ 60 samples and four pathogens as compared to the 24 samples and two E.coli for this research study.

4. Conclusions

The primary objective of this paper is a preliminary evaluation of the following hypotheses: an intelligent decision support system (IDSS) to allow the rapid collection and classification of E.coli can be designed and preliminarily evaluated, which will significantly decrease detection time for E.coli bacteria. This system was designed, implemented and tested using a data set size of 24 samples. Using this data set, both K-fold and 1-hold out cross-validation processes were used to validate the IDSS performance, using this limited A and B E.coli data set. Using these cross-validation processes, the performance of the PNN classification model evaluated, which provided an answer to the following question: What is the percentage of improvement in the time-to-detect performance against the 16 to 48 hours required for the conventional methods of detection of microorganisms (using E.coli data as a baseline)? The answer is: for the 16 hour period we have a 6.7% reduction in time-to-detect ((16-15)/15 × 100% = 6.7%) and for the 48 hour period we have a 220% time improvement ((48-15)/15×100% = 220%). These percentage accuracies will be revised using an ROC analysis which will be performed when a sufficient number of E.coli samples have been collected using this new automated collection data collection system. These large differences in percentage improvement are probably due to the small sample size available for this analysis.

References

7. J. Karasinski, S. Andreescu, O. A. Sadik, B. Lavine, Multiarray Sensors with Pattern Recognition for the Detection, Classification and Differentiation of Bacteria at Species and Subspecies Levels, Analytical Chemistry, 77/24 (2005) 7941-7949