

Mixed Statistical and Numerical Model for Sensor Array Detection and Classification

This application claims benefit to U.S. patent application no. 11/987,658, filed December 3, 2007 to Li *et al.*, which is hereby incorporated by reference in its entirety.

FIELD OF THE INVENTION

[0001] This invention is related in general to the field of biosensor and chemical sensor array detection and classification.

BACKGROUND OF THE INVENTION

[0002] Biosensor array units having sensor arrays are becoming very useful in today's society, with the threat of bioterrorism being more and more prominent. In more detail, bioterrorism and biological warfare pose both physical and psychological threats to military and civilian forces, as well as to civilian populations. Biological and chemical sensors also hold promise for environmental monitoring, indoor air quality assessment, and in food and beverage analysis.

[0003] Typically, either analytical or statistical models are utilized to generate a detection algorithm for sensor arrays. Analytical models are only applicable for quantitative data analysis while statistical models typically capture all correlations in the datasets of which some may be generated by causes that are not of interest. The sensitivity and specificity are highly dependent on operational conditions, and thus it is difficult to optimize the trade-off relationship between the detection sensitivity and specificity. In practice, it is often required to collect standard datasets to calibrate these models, and thus they are not ideal for operation under robust environments.

SUMMARY OF THE INVENTION

[0004] The present invention relates to a method and apparatus for sensor array detection and classification.

[0005] In accordance with one aspect of the invention, there is provided a method for detecting and classifying sensor data. The method includes performing a statistical analysis of the data collected from an array of sensors, to obtain statistically analyzed sensor data. Based on the statistically analyzed sensor data, it is determined whether or not there a response signal has been generated from the array of sensors. If a response signal is determined to have been generated, the method further includes performing a numerical pattern recognition on the data collected from the array of sensors, in order to categorize the data.

[0006] In accordance with another aspect of the invention, there is provided a system for detecting and classifying sensor data. The system includes a data processing unit configured to receive and perform data processing on data output from an array of sensors. The system also includes a statistical analysis unit configured to perform statistical analysis on the data processed by the data processing unit, to obtain statistically analyzed sensor data, wherein the statistical analysis unit determines, based on the statistically analyzed sensor data, whether or not there a response signal has been generated from the array of sensors. The system further includes a numerical pattern recognition unit configured to perform a numerical pattern recognition on the data collected from the array of sensors, in order to categorize the data.

[0007] In accordance with yet another aspect of the invention, there is provided a computer readable medium embodying computer program product for detecting and classifying sensor data, the computer program product, when executed by a computer, causing the computer to perform the steps of:

- fetching data collected from an array of sensors;
- performing a statistical analysis of the data collected from the array of sensors, to obtain statistically analyzed sensor data;

determining, based on the statistically analyzed sensor data, whether or not there a response signal has been generated from the array of sensors; and

performing, if a response signal is determined to have been generated, a numerical pattern recognition on the data collected from the array of sensors, in order to categorize the data.

[0008] It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the invention as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate several embodiments of the invention and, together with the description, serve to explain the principles of the invention.

[0010] Figure 1 is a block diagram showing elements involved in performing a statistical and numerical analysis of biosensor data, according to a first embodiment of the invention.

[0011] Figure 2 is a flow diagram diagramming a method of performing biosensor array detection and classification, according to the first embodiment.

DETAILED DESCRIPTION

[0012] Reference will now be made in detail to embodiments of the invention, examples of which are illustrated in the accompanying drawings. An effort has been made to use the same reference numbers throughout the drawings to refer to the same or like parts.

[0013] Unless explicitly stated otherwise, “and” can mean “or,” and “or” can mean “and.” For example, if a feature is described as having A, B, or C, the feature can have A, B, and C, or any combination of A, B, and C. Similarly, if a feature is described as having A, B, and C, the feature can have only one or two of A, B, or C.

[0014] Unless explicitly stated otherwise, “a” and “an” can mean “one or more than one.” For example, if a device is described as having a feature X, the device may have one or more of feature X.

[0015] Figure 1 shows a data analysis system that combines both statistical and numerical models have been developed as a detection and classification algorithm for a biosensor array, according to a first embodiment of the invention. The biosensor array used to provide data to the system according to the first embodiment may be a Smart Bio Sensor (SBS), and it may be a biosensor array of a handheld or portable sensor detection unit (e.g., Bioseeq or BioseeqPlus) made by Smiths Detection Inc., for example. The statistical analysis is utilized for sensor signal detection, while one or more numerical methods are applied to analyze the correlations of interest captured in the data for pattern recognition. The statistical analysis provides the detection sensitivity, and the numerical analysis enhances the biological recognition capability and thus provides the detection and classification specificity. The system, which may be implemented in Matlab and C programming language in one possible implementation of the first embodiment, can be run in real-time simultaneously with data collection. By way of example and not be way of limitation, in the first embodiment, 3 to 5 data points are utilized for detection analysis and 10 data points are utilized for classification analysis. The system according to the first embodiment is capable of carrying out self-calibration in real-time, and is thus suitable for robust operation. Other chemical/biological sensor arrays besides an SBS array may be utilized to provide sensor data to be detected and classified by way of the first embodiment.

[0016] As shown in Figure 1, the system receives data collected from an array of sensor channels and stored in a data store, into a Principle Component Analysis (PCA) statistical model, to determine if there is a response signal. For a detected sensor response, the system runs a numerical pattern recognition model that may be customized based on the sensor responses that are of interest. Information from other sensor channels, such as airborne particle counter (APC) data, may also be stored (in the data store), interpreted and utilized as parameters in a pattern recognition model

for interferent rejection. Fingerprints numerically extracted from the correlations between different biosensor channels can also be exploited for classification.

[0017] With the system according to the first embodiment, there is provided a detection and classification system for a biosensor array that is suitable for real-time and robust operation. Accordingly, no frequent model calibration needs to be done by users of a biosensor detection unit that incorporates the system and method according to the first embodiment.

[0018] In one possible implementation of the first embodiment, a statistical model Principal Component Analysis (PCA) unit is utilized for signal detection, and then a numerical model which is built upon correlations of interest captured in a biosensor dataset is run for pattern recognition to enhance the specificity of the detected sensor array data. The trade-off relationship between detection sensitivity and specificity can be optimized using a Relative Operating Characteristic (ROC) curve, for example.

[0019] Turning now to Figure 1, in a system 100 according to the first embodiment, raw biosensor data obtained by an array of biosensors (not shown) is stored in a raw data store 105. The stored raw biosensor data is processed by a data processing unit 110, and the processed data is stored in a processed data store 120. A PCA unit 130 performs signal detection of the data-processed biosensor data, based on a statistical model. If a signal detection is determined to have occurred based on analysis by the PCA unit 130, a numerical analysis of the biosensor and APC data is then performed. In more detail, an interferent rejection I unit 140, a biological match sensor data unit 150, and an interferent rejection II unit 160 are used to perform numerical analysis on the statistically analyzed biosensor and the non-statistically-analyzed APC data, and a classification sensor data unit 170 is used to classify the numerically analyzed data (if determined to be required). The units 140, 150, 160 and 170 utilize different numerical models and collectively correspond to a numerical analysis unit of the system according to the first embodiment.

[0020] Also, as shown in Figure 1, Airborne particle counter (APC) data is collected by an APC sensor (not shown), and stored in the raw data store 105. The stored raw

APC data is processed by the data processing unit 110, and the processed data is stored in the processed data store 120, whereby the APC data is also statically analyzed and numerically analyzed by the system and method according to the first embodiment. The APC data is utilized for interferent rejection as determined by the interferent rejection II unit 160.

[0021] The biosensor data obtained by an array of biosensors provides fluorescence information, while the APC data provides a total count of particles per liter of air. Thus, the biosensor data and the APC data are measures of different physical properties. Given the entirely independent nature of the data, different approaches are required to process these data. In one possible implementation of the first embodiment, the biosensor data processing is processed by a Savitzky_Golay filter to smooth the data and remove the offset, whereby this filtering corresponds to the data processing performed on raw biosensor data by the data processing unit 110. Also, in one possible implementation of the first embodiment, the data processing performed on raw APC data by the data processing unit 110 is carried out by applying a median filter followed by integrating multiple data points to smooth the data and subtract a baseline level to remove the offset. Other types of data processing on the raw biosensor data and the raw APC data by the data processing unit 110 may be envisioned by those skilled in the art, without departing from the spirit and scope of the invention.

[0022] The PCA unit 130 performs Principal Component Analysis, which is a statistical technique known in the mathematical arts used to reduce multidimensional data sets to lower dimensions for analysis. PCA involves the calculation of an Eigenvalue decomposition or a Singular value decomposition of a data set. The results of a PCA are usually discussed in terms of component scores and loadings. PCA is mostly used for making predictive models. In the first embodiment, PCA is used as a tool to reduce the multi-dimensional biosensor data to a one-dimension score vs. time. The score is of one of the principal components captured in the biosensor data, but it is typically of the first or second one. The relationship of score vs. time is then utilized to detect a correlated signal response from the biosensors.

[0023] For conventional biosensor analysis systems, a standard data set is typically used to generate a PCA model onto which unknown data is projected, in order to obtain the prediction. The PCA model of such conventional systems can be calibrated with an updated standard dataset if there is any change in operational conditions. In the first embodiment, however, no use is made of a standard dataset for calibration in its real-time model. Instead, a PCA is applied for every data point, and decision-making is performed based on the comparison between the latest change in score and the score variation in a background window. A reference model generated with the background window is updated for every data point and is considered as being calibrated in real-time. This enhances the robustness of the statistical analysis performed by the PCA model (as performed by the PCA unit 130) according to the first embodiment.

[0024] The system and method according to the first embodiment exploit the change in the value of the score to detect a non-specific response signal instead of the principal component vector for specific pattern recognition and prediction. A non-specific response signal may be caused by any analytes of interest, interferences, high level of noise, or changes in operational conditions. Only a positive result at the statistical analysis step (as performed by the PCA unit 130) will trigger the next step of numerical pattern recognition for bioalarm, as performed by the numerical analysis units 140, 150, 160 and 170.

[0025] In the first embodiment, the PCA is only performed on the biosensor data, and is not performed on the APC data, as seen by the direct path between the processed data store 120 and the Interferent Rejection II unit 160 in Figure 1. Biosensor data stored in the processed data store 120 is subject to statistical analysis as performed by the PCA unit 130, and then that statistically analyzed data is provided to the Interferent Rejection I unit 140.

[0026] The numerical analysis performed by the numerical analysis units 140, 150, 160 is essentially a pattern recognition process. It is executed only if a positive result is obtained (e.g., a response signal is generated) by the PCA unit 130. For the numerical analysis, each pattern is characterized with a set of vectors of which each is associated with a particular sensor/detector type. Each vector is assigned a logic

value of +1, 0, or -1. The assignment of the logic values according to one possible numerical analysis implementation is described below.

[0027] For biosensor data,

○ If the biosensor data meet the following condition, +1 is assigned to the relevant vector:

1^{st} derivative of (P_{t-2}, P_{t-1}, P_t) - 1^{st} derivative of $(P_{t-5}, P_{t-4}, P_{t-3}) \geq$ positive threshold
where P_t is the current data point and P_{t-1} is the point before, etc., and (P_{t-2}, P_{t-1}, P_t) defines the detection window, and $(P_{t-5}, P_{t-4}, P_{t-3})$ is considered the background window.

○ If the biosensor data meet the following condition, -1 is assigned to the relevant vector:

1^{st} derivative of (P_{t-2}, P_{t-1}, P_t) - 1^{st} derivative of $(P_{t-5}, P_{t-4}, P_{t-3}) \leq$ negative threshold

○ Otherwise, 0 is assigned to the vector.

[0028] For APC data,

○ If a net increase in APC (current APC level – baseline level) is larger than a positive threshold, +1 is assigned to the vector.

○ If a net decrease in APC (current APC level – baseline level) is less than a negative threshold, -1 is assigned to the vector.

○ Otherwise 0 is assigned to the vector.

[0029] The logical values are combined into vectors to form a pattern. The number of vectors used to express a pattern is equal to the number of sensor/detector types in the system. By way of example, there are four (4) biosensor types (S1, S2, S3, S4) and one (1) APC detector (APC1) in a sensor array. Formation of a pattern is as described below:

Vectors	Combined pattern
S1 (+1, 0, -1) S2 (+1, 0, -1) S3 (+1, 0, -1) S4 (+1, 0, -1)	$\left. \begin{array}{l} \text{S1 (+1, 0, -1)} \\ \text{S2 (+1, 0, -1)} \\ \text{S3 (+1, 0, -1)} \\ \text{S4 (+1, 0, -1)} \end{array} \right\} [\text{S1 (+1,0,-1), S2 (+1,0,-1), S3 (+1,0,-1), S4 (+1,0,-1), APC1 (+1,0,-1)}]$

[0030] Only a subset of the combined patterns is related to analytes of interest. In one possible implementation of the first embodiment, the following three conditions are used to determine whether or not a pattern is related to an analyte:

- The combined pattern does not contain generic features of interferent patterns. (interferent rejection I unit 140)
- The combined pattern captures generic features of the known analyte patterns (biological match unit 150).
- The APC vector is +1. (interferent rejection II unit 160)

[0031] The interferent and analyte generic pattern features described above are extracted from data that can be collected from in-house or from field system tests. In the first embodiment, a model is created for checking each of the three conditions listed above. The numerical models can be run sequentially or in parallel, but all of the conditions should be met to trigger a bioalarm. In the first embodiment, sequential checking is performed by the three numerical analysis units 140, 150, 160 to determine if the above three conditions have been met, to reduce redundant computing. Alternatively, the processing performed by the numerical analysis units 140, 150, 160 can be performed in parallel, depending upon the type of processor or processors used to perform the numerical analyses of the biosensor and APC data.

[0032] If the three conditions described above have all been met, a bioalarm will be issued, and this issuance will initiate the classification of the data by the classification sensor data unit 170, whereby more detailed pattern recognition is carried out. Also, other data signatures, such as a difference in response time between certain sensor types, can be analyzed by the classification sensor data unit 170 to identify the detected bioagent as belonging to one of different classes of bioagents, such as one of the following classes: bacterial spores, bacteria, viruses, and toxins.

[0033] Figure 2 shows a method of performing biosensor array detection and classification, according to the first embodiment. In a first step 210, biosensor data is collected, whereby such biosensor data may be obtained from one or more array of sensors. In a second step 220, which is typically performed at the same time as the first step 210, APC data is collected. In a third step 230, the biosensor data and the APC data are stored in a data store. In a fourth step 240, the biosensor data is

statistically analyzed by a Principle Component Analysis (PCA) statistical model, to determine if there is a response signal. In a fifth step 260, a determination is made as to whether or not a response signal exists in the biosensor data. If no response signal exists, then in a sixth step 260, no further steps are performed (no numerical analysis is performed). If a response signal is determined to exist, then in a seventh step 270, the biosensor data and the APC data is subject to one or more numerical models, in order to categorize that data as biological or non-biological detection. Then, in an eighth step 280, a determination is made as to whether or not the categorization results meet certain criteria. If Yes, a bioalarm is output in step 282, a more detailed classification of the biosensor data is made in step 295 and the classification result is output in step 297 (note that steps 282 and 295 can be initiated at the same time). If No, in step 280, no further numerical analysis is performed on the biosensor data, as shown in step 290.

[0034] The embodiments described above have been set forth herein for the purpose of illustration. This description, however, should not be deemed to be a limitation on the scope of the invention. Various modifications, adaptations, and alternatives may occur to one skilled in the art without departing from the claimed inventive concept. For example, in a second embodiment, only sensor data provided by an array of sensors or biosensors would be analyzed, and APC data would not be input and analyzed, whereby the numerical processing unit would include only the units 140, 150 and 170 in that instance. The spirit and scope of the invention are indicated by the following claims.

WHAT IS CLAIMED IS:

1. A method for detecting and classifying sensor data, comprising:
processing data collected from an array of sensors;
performing a statistical analysis of the data collected from the array of sensors,
to obtain statistically analyzed sensor data;
determining, based on the statistically analyzed sensor data, whether or not
there a response signal has been generated from the array of sensors; and
performing, if a response signal is determined to have been generated, a
numerical pattern recognition on the data collected from the array of sensors, in order
to categorize the data.
2. The method according to claim 1, further comprising:
processing airborne particle data from one or more other sensors,
wherein the step of performing a statistical analysis is performed on the data
collected from the array of sensors and not on the airborne particle data.
3. The method according to claim 1, wherein the step of performing a
statistical analysis is performed by using a principle component analysis statistical
model.
4. The method according to claim 1, wherein the step of performing a
numerical pattern comprises:
performing an interferent rejection pattern analysis on the sensor data;
performing a biological match pattern analysis on the sensor data;
performing an interferent rejection pattern analysis on the APC data; and
performing a classification pattern analysis on the sensor data.
5. The method according to claim 2, wherein the step of performing a
numerical pattern comprises:
performing a first interferent rejection pattern analysis on the sensor data;
performing a biological match pattern analysis on the sensor data;

performing a second interferent rejection pattern analysis on the airborne particle data; and

if the results of the three performing steps meet predetermined criteria, performing a classification pattern analysis on the sensor data.

6. The method according to claim 5, wherein the step of performing a first interferent rejection pattern analysis is performed on a combination of the statistically sensor data and on raw sensor data that has not been statistically analyzed.

7. A system for detecting and classifying sensor data, comprising:
a data processing unit configured to receive and perform data processing on data output from an array of sensors;
a statistical analysis unit configured to perform statistical analysis on the data collected from the array of sensors and data processed by the data processing unit, to obtain statistically analyzed sensor data, wherein the statistical analysis unit determines, based on the statistically analyzed sensor data, whether or not there a response signal has been generated from the array of sensors; and
a numerical pattern recognition unit configured to perform a numerical pattern recognition on the data collected from the array of sensors, in order to categorize the data.

8. The system according to claim 7, wherein the numerical pattern recognition unit only performs the numerical pattern recognition when the response signal is determined to have been generated.

9. The system according to claim 7, further comprising:
a second data processing unit configured to receive and perform data processing on airborne particle data output from one or more other sensors,
wherein the statistical analysis unit performs statistical analysis on the data collected from the array of sensors and not on the airborne particle data.

10. The system according to claim 7, wherein the statistical analysis unit performs statistical analysis using a principle component analysis statistical model.

11. The system according to claim 7, wherein the numerical pattern recognition unit comprises:

an interferent rejection pattern analysis unit configured to perform an interferent rejection pattern analysis on the sensor data;

a biological match pattern analysis configured to perform a biological match pattern analysis on the sensor data; and

a classification unit configured to perform a classification pattern analysis on the sensor data.

12. The system according to claim 9, wherein the step of performing a numerical pattern comprises:

a first interferent rejection pattern analysis unit configured to perform a first interferent rejection pattern analysis on the sensor data;

a biological match pattern analysis configured to perform a biological match pattern analysis on the sensor data;

a second interferent rejection pattern analysis unit configured to perform a second interferent rejection pattern analysis on the airborne particle data
and

a classification unit configured to perform a classification pattern analysis on the sensor data, when the results of the first and second interferent rejection pattern analyses and the biological match pattern analysis meet predetermined criteria.

13. The system according to claim 12, wherein the first interferent rejection pattern analysis unit performs the first interferent rejection pattern analysis on a combination of the statistically sensor data and on raw sensor data that has not been statistically analyzed.

14. A computer readable medium embodying computer program product for detecting and classifying sensor data, the computer program product, when executed by a computer, causing the computer to perform the steps of:

fetching data from an array of sensors;

performing a statistical analysis of the data collected from the array of sensors, to obtain statistically analyzed sensor data;

determining, based on the statistically analyzed sensor data, whether or not there a response signal has been generated from the array of sensors; and

performing, if a response signal is determined to have been generated, a numerical pattern recognition on the data collected from the array of sensors, in order to categorize the data.

15. The computer readable medium according to claim 14, further comprising:

fetching airborne particle data from one or more other sensors,

wherein the step of performing a statistical analysis is performed on the data collected from the array of sensors and not on the airborne particle data.

16. The computer readable medium according to claim 14, wherein the step of performing a statistical analysis is performed by using a principle component analysis statistical model.

17. The computer readable medium according to claim 14, wherein the step of performing a numerical pattern comprises:

performing an interferent rejection pattern analysis on the sensor data;

performing a biological match pattern analysis on the sensor data; and

performing a classification pattern analysis on the sensor data.

18. The computer readable medium according to claim 15, wherein the step of performing a numerical pattern comprises:

performing a first interferent rejection pattern analysis on the sensor data;

performing a biological match pattern analysis on the sensor data;
performing a second interferent rejection pattern analysis on the airborne particle data; and

if the results of the three performing steps meet predetermined criteria,
performing a classification pattern analysis on the sensor data.

19. The computer readable medium according to claim 18, wherein the step of performing a first interferent rejection pattern analysis is performed on a combination of the statistically sensor data and on raw sensor data that has not been statistically analyzed.

20. The computer readable medium according to claim 14, wherein the statistical analysis performed in the performing step is a Principle Component Analysis.