Artificial Intelligence Techniques for Biological Detectors

James Clark, Paul Smith and Sue Neal

BIRAL
PO Box 2, Portishead, Bristol BS20 7JB, UK

Abstract

Detectors for biological warfare agents generate very large data sets from the analysis of airborne particles. The data consists of particle characteristics that are measurable in airborne suspension such as size, shape and fluorescence. These characteristics are used to differentiate agent material from the wide range of particulate material, both natural and pollutant, found in the atmosphere. The mathematical tools that can be used to achieve good differentiation and trigger reliable alarms are reviewed.

Keywords: Biological detection, artificial intelligence, algorithms

1 Introduction

Biral have been involved for a number of years in the development of techniques for the recognition of biological warfare agents in the atmosphere under the very challenging conditions that may occur in the battlefield. The techniques developed for real-time detection employ physical characterisation techniques measured in the aerosol phase. These are secondary characteristics, some of which are shared by the vast array of natural and pollutant materials found in the atmosphere. Data are generated at rates of up to 10,000 particles per second and so sophisticated algorithms, based on artificial intelligence (AI) techniques, are required to process the data and generate alarms. One such technique has been used to process the data generated by the Biral Aerosol Size and Shape (ASAS) technology in the active military deployment of the sensor. This has been shown to be highly reliable and effective. More recently new techniques have been investigated for use with the increased data set generated by the Biral VeroTect technology that measures aerosol fluorescence characteristics as well as particle size and shape.

2 Candidate Techniques

2.1 Character Matching

The simplest option for differentiating classes of particles would be straightforward character matching. This would be possible if, when a particle had a particular size, shape and fluorescence characteristic, it could be unequivocally classed as, say, a bacterium. This is unfortunately never the case as the secondary characteristics that can be measured are not unique to any particular class. However the technique can be used to select only appropriate data for further analysis. For example, by limiting analysis only to particles in the respirable size range.

2.2 Classical Statistical Techniques

An option that is widely used for this, and for many other related types of classification, is statistical matching and discrimination of populations. These techniques may use libraries containing the characteristics, such as the medians, means and variance of the measured parameters and compares data sets from unknown samples with the contents of the library. A range of statistical tests, such as the χ^2 test, can be used to obtain the best match and quantify the probability that the unknown data come from the same population as the best match in the library.

This technique can work well in some circumstances but also has considerable limitations. Firstly, it assumes that a library can be compiled that contains all the elements that a detection and characterisation system may encounter. As most applications involve making measurements against the almost infinitely variable atmospheric background, that can never be truly the case. In most potential applications it is likely that the target of the detection will also have highly variable characteristics.

It also assumes that the populations are statistically "well behaved". Parameters such as standard deviation only have true meaning when they relate to a distribution that can be described mathematically. They may still have a value in describing the characteristics of imperfect distributions but when they are used in formal statistical tests the outcome may be unreliable.

2.3 Knowledge-based Techniques

This is a composite technique that can make use of all classes of information, by constructing an empirical rule base and testing how well the data fit the rules. The knowledge may be analytical, statistical or heuristic. Where this wins over the straightforward statistical approach is in the flexibility that can be applied to the development of the rule base. For example: when statistics, such as means and standard deviations, are used these can be applied much more readily to poorly defined distributions as the calculated parameters can be used as indicators rather as inputs to formal statistical tests. The ability to use a range of data types within the rule-base enables a much more flexible analysis technique that can be adapted to a wide range of varying circumstances.

2.4 Artificial Neural Networks (ANN)

Neural networks were the first class of techniques to create "intelligent" computers that were capable of learning from experience. Algorithms were developed that changed the weighting associated with tracks through the system in response to learning data where the end class was assigned. So if, say, it was required to design a system that sorted cups from saucers it would then be necessary to make characterising measurements on a wide range of examples of both. Data files on each would be passed through the system with it configured so that the system always directed it to the correct answer and the algorithm would modify the weightings appropriately. Once this "supervised learning" process had been completed then unknown, mixed data could be fed through the system and it would correctly classify each item as either a cup or a saucer.

Many variations on the early neural network techniques have been developed and these are used for a very wide range of applications. They are particularly appropriate for classification applications, where large quantities of representative data are available but where no direct analytical or statistical relationship between the descriptive parameters and the result can be established. This makes aerosol particle characterisation an ideal candidate for this type of analysis, as most direct measurement techniques are capable of generating data on many thousands of particles per second.

2.5 Principal Component Analysis (PCA)

When analysing large multiparameter data sets, it is often desirable to reduce their dimensionality. Principal component analysis (PCA) is one technique for doing this. It replaces the original, measured variables by a smaller number of derived variables, the principal components, which are linear combinations of the original variables. Often, it is possible to retain most of the information in the original variables with a very much smaller number and this both simplifies and clarifies the analysis.

PCA is a popular technique for the analysis and display of multi-variate aerosol data. It is particularly valuable as a tool that illustrates how readily classes may be separated, or if they are separable, using the data available. However, in its most usual form, it is limited to linear transformations of the data, which may not give the optimum separation of the classes

2.6 Support Vector Machines (SVM)

Support Vector Machines (SVMs) are learning techniques that use a particular class of mathematical algorithms known as kernels. They are beginning to replace ANNs in a wide range of difficult classification and pattern recognition applications. They have the potential to overcome many of the problems that have been encountered in the application of Neural Networks, including parameter weighting in the input vector, high dimensionality and a problem, often associated with complex data sets, called overfitting.

In relatively simple terms, the technique uses linear learning machines (linear discriminators) combined in a "kernel" function to separate classes, in high dimensional feature space, by linear hyperplanes. This concept is easier to understand than it sounds. It may help to consider a simple linear regression, where the objective is to draw the most representative line through noisy data. This is generally achieved by computing the minimum of the sum of the squared deviations from the regression line (least squares). If this principle is extended to as many dimensions as the data have descriptive features (f) then data can be separated in f-dimensional space by hyperplanes with dimensions of f-1.

A great advantage of kernel based learning machines is that there is no requirement to directly define the relationship between characterisation variables. The kernel can be designed so that this becomes implicit in what is effectively a non-linear version of PCA. This overcomes what is known as "the curse of dimensionality" which is often encountered when the number of characterisation parameters in ANN analysis is increased.

3 Systems Applied to Biological Detection

An advanced form of neural network, known as Learning Vector Quantisation (LVQ) has been used very successfully to analyse the data generated by Biral's Aerosol Size and Shape (ASAS) technology. The system uses feature maps developed by training the system with known materials. The maps may include potential interferents as well as detection targets. LVQ is then used match unknown data with the map that gives the best fit and so each particle is then assigned to a class. A great advantage over other types of ANN is that maps can be added or removed without the need to retrain the whole system making the system highly adaptable. This technique has now been employed for many years in complex military bio-detection systems for the analysis of the ASAS

data. Its success has been such that it has been carried over, largely unmodified, into radically updated equipment.

When particle fluorescence measurements were combined with the ASAS technology in the Biral VeroTect biological detector ANN techniques were found to be much less appropriate. This was largely due to the fact that the data was of different types but the higher dimensionality would also have been likely to cause greater uncertainty. The first technique, developed for trials, used statistical matching of unknown data with a library of data from different targets and backgrounds. This proved to be highly effective when operating in predictable environments with a limited range of targets. However, it was less successful when operating with unpredictable backgrounds against a wider range of targets, generated by different techniques.

For the second generation, operational software, the company has developed a knowledge-based analysis technique suitable for use in the widest possible range of environments. The rule base is entirely numeric and uses a number of weighted measures and statistics to differentiate targets from, potentially, highly variable backgrounds and interferents. It can be tailored to meet different sensitivity and false alarm requirements. The system can also learn and adapt to new background conditions and so enhance alarm performance.

This software has now been tested under a wide range of conditions and has proved to be accurate, robust and versatile in detecting and classifying biological agent simulants. It has also proved successful in differentiating simulants from potential interferent materials, with similar characteristics, that may also be found in the atmospheric aerosol.

4 Future Directions

Advanced techniques, such as SVM, are showing considerable promise but there are a number of hurdles to overcome before they can offer superior performance, in all respects, to the more established techniques. However, advances in analysis techniques are generating ever greater numbers of characterisation parameters. These will challenge the more established techniques and it is likely that SVM, perhaps with a PCA kernel, will offer significant advantages. Biral intend to continue work on developing the application of this technique, with advice and assistance from leaders in the field.