

Statistical Pattern Recognition Using Enterprise Miner and Self-Organizing Maps

Keith A. Jones

There has been significant increasing interest within the last several years in the use of Pattern Recognition (PR) and other Artificial Intelligence (AI) techniques for machine automation software, including applications in speech recognition, robot vision, biometric security systems, and other intelligent software automation techniques. One of the more recent innovations in Pattern Recognition and AI involves Statistical Pattern Recognition (SPR), which has particular significance to SAS users because most of the tools for SPR are very similar to the techniques used in SAS Customer Relationship Management as provided by SAS Enterprise Miner for data mining using Neural Network Node and Decision Tree Node. However, although data mining capabilities are available in many statistical software products, only SAS Enterprise Miner comes equipped with one of the most powerful SPR techniques, Kohonen Self-Organizing Maps (SOMs) in the SOM Node. Although SAS Enterprise Miner is an extremely valuable data mining platform that can support several powerful SPR techniques, the use of the SOM Node can provide the basis for building many advanced automated recognition applications and software agents that can be implemented using SAS/AF and SAS/C for use in web clients and Java pages, as well as traditional systems automation.

The first and very simplest pattern recognition techniques involve various methods for matching signals or image features to marker files with some type of data signature or data array mapping pattern. This is First Generation SPR method.

These First Generation pattern recognition techniques can involve either direct matching of exact values or string patterns against a look-up table, or various methods involving probabilities of a match based on a specific sub-set of key features, or a hierarchy of decision rules that weight together key features. Usually there is some combination of the two approaches, with an increasing trend away from simple matching of raw histogram or template look-up values, which is still used primarily only in discrete voice command speech recognition and many conventional biometric ID security systems. More often, conventional pattern recognition systems use increasingly complex SPR methods involving probability vectors and complex decision tree rules, such as in large vocabulary natural language translation and autonomous robot active vision.

Notably, this increased emphasis on statistical and probabilistic approaches to pattern recognition are behind the increasing trend for integrating SAS statistical programming objects into conventional SPR systems, due to the dominance of SAS as the most powerful statistical programming language and its capabilities for MDDb and web integration. As previously noted, the Base SAS programming language is also very well suited for the handling of both implicit and explicit subscript arrays for use in probability vectors as well as various types of look-up tables, thus SAS is the best of all world for developing both simple and complex histogram and template matching pattern recognition problem domain solutions..

It should also be easily understood by both SAS data step programmers as well as SAS applied statistical procedure programmers not only how but also why both SAS unidimensional and multidimensional arrays are such handy tools for Statistical Pattern Recognition. Histograms are widely used by all SAS programmers and statisticians in almost every exploratory data analysis that is either used to profile a populations with respect to key variables or dimensions, so it should not be hard to understand why it is a valuable technique to stratify or band numeric continuous data into discrete layers or ranges.

Although the template matching may involve element by element scans to compare two or more arrays across an entire vertical task domain matrix, or only to "zero in" on a small number of critical or key features within the constraints of a more limited vector subspace, in all cases the template is simply a multidimensional equivalent of unidimensional histograms. Thus, in both the histogram and template matching used in First Generation pattern recognition, often a simple linear model using PROC GLM or logistic regression using PROC LOGISTIC may identify the key bands or histogram ranges that can be defined as dummy variables (0,1) or binary variable inputs that are most critical to fast solution for a response model, key signal bands and image features may yield fast statistically reliable predictors.

Thus, regardless of the nature of the First Generation pattern recognition problem domain and application, a most critical aspect of the task domain and solution approach will always involve the representation of the signal or imaging data. There are generally 3 levels of representation of the data,

which applies to the specification of file formats for unidimensional as well as multidimensional SAS arrays, for both signal histogram and image template processing.

The first level is pure binary data, and this corresponds to high-low or plus-minus time base waveform data in the case of signals, such as audio speech data, as well as pure Black-White images such as are used in newspaper print media. The second level is what is called "Gray-Scale" or halftone data, which is in effect data that is either time-sliced as in the case of digitally filtered or preprocessed analog audio signal data, or images that are converted or transformed into halftone "mosaics". These are usually very small row and column cells that are binary encoded to be either black or white based on some algorithm that averages across a threshold or baseline, or assigns binary states based on some underlying feature of the surrounding cells. Clearly, the sampling level or granularity of the halftone image is critical to the recognition capability of the image, and this is a traditional problem of print photography as well as digitizing photographs for electronic transmission. Obviously, it should be normally possible to improve the quality as well as quantity of information that can be extracted or recognized by altering both the size of the cells in the image as well as the sampling rate that is used to scan or compare images if the problem domain is to detect changes. This is in fact basis for all modern Information Theory, and the basis for the concept of entropy. Furthermore, it should also be obvious when considering the problem of halftone and Gray-scale filtering by assigning high-low binary values to cells how this same method can not only improve digital information levels from an image, but the same process can be used to assign binary probability vectors to a signal other than images, such as in audio speech recognition, or any related sound problem that involves a need for intelligent SPR, and this is in fact the case.

The third level involves color scale data, which is not only the method that is most commonly used to represent Red-Blue-Green (RGB) in both color video as well as all color graphic file formats, but it is also used to refer to all methods that involve not only a multidimensional matrix or vector of binary values that are externally referenced to another file or factor or coefficient that can be used to transform encoding or decoding for purposes of even greater resolution that would ever be possible within the constraints of the binary halftone or Gray-Scale data. Although in some cases a color graphics file format may contain real number values for each of the 3 primary RGB combinative color spectrum bands, all of the most advanced and emerging graphics formats not only have triplets of RGB values for each x,y coordinate pixel point in an image, but also contain increasingly complex header and footer information in the file that can be used to externally derive much greater resolution in the image or signal encoding using standardized methods of transformation and compression that enable order of magnitude increases in the level of information that can be potentially stored in and analyzed from the file.

In fact, this is why such formats are all known as "External Data Formats" and include such commonly used file extensions as .GIF, .JPG, .BMP, .TIF, .PIC, and several hundred others. The most important point with regard to the topic of External Data Formats is to simply understand that any graphics images within any of these formats that involve either compression or encryption can make them virtually worthless for most Statistical Pattern Recognition applications. Therefore, it is usually necessary to convert such External Data Format images into an appropriate raw data image format using SAS, using the most advanced suitable data transformation and cleaning methods that are normally used by SAS data modeling statisticians and programmers, but the results will always contain less usable information than the original image. So as will be explained in more detail later, it is always best if at all possible to capture and retain raw signal and image formats as data that can be analyzed using appropriate SAS procedures to get the best results from SPR.

Therefore it should be clear that there are many levels and formats of data that are needed for Statistical Pattern Recognition, but in fact they are not that much different than the same methods that SAS programmers and statisticians every day to build models and analyze samples and manage decision enabling data bases. In addition, it should be clear that all of the First Generation methods that are based on histograms and template matching are also very similar to the same methods that SAS programmers and statisticians use every day to build models and analyze samples and manage their data, using exactly the same basic SAS programming skills.

There are however, several advanced techniques and topics that relate to exactly how these various levels of First Generation pattern recognition methods can be expanded using SAS programming language and procedures in order to build the Next Generation of Statistical Pattern Recognition application systems and technologies. Among these are histogram pattern recognition enhancement SAS based techniques, such as Discriminant Analysis Thresholding, Entropy Thresholding, Multi-Thresholding, Local and Global Quantum Equalization, Color Quantization, Histogram Back-Propagation and Template Indexing.

In addition, it is important to understand that the same statistical principles that are involved in statistical techniques to enhance upon traditional histogram and template matching are also critical to effectively applying and improving upon one of the most basic traditional pattern recognition techniques, Edge Detection. There are a great many traditional Edge Detection methods that can be easily applied, and integrated into complex SPR systems using SAS programming language and procedures.

Some of the most critical methods involve Edge Thinning, Edge Texture Convolution, Shadow Line Detection, Circle Detection, Chain Detection, Closing Gaps, Regionalization, and various other methods for separating or combining SPR object features using SAS. As can be seen in the attached figure, the first and most basic application of these methods can be applied starting with a simple scan of a color photograph into an RGB graphic file in an External Data Format and then transformed back to a binary Black-White, Gray-Scale, and transposed SAS dataset, which can be most easily preprocessed and stored for future analysis as raw binary encoded character image file, transformed Gray-Scale image file using various edge detection methods, and stored in a raw data format as a SAS dataset with suitable ranges encoded using PROC FORMAT.

The Second Generation of pattern recognition involved much more sophisticated techniques than pure simple matching, and involve the selection of features that are critical to deriving pattern recognition classifiers. Although the terms are slightly different, for all practical purposes, a classifier is a behavior model and a feature is a variable in a response model to predict accurate recognition.

In practical reality, the automated pattern classifier method that is the most common basis for Second Generation pattern recognition is most simply represented by the same problem domain architecture that is used by all SAS programmers and statisticians to build and apply all SAS predictive models. This method involves what is commonly referred to as a generic supervised prediction problem, that consists of a SAS dataset of some sample of a universe of cases, observations, examples, instances or measurements, and each case is associated with a vector of model key input independent variables, which are also called predictors in conventional response modeling, or features in Statistical Pattern Recognition.

The target or dependent variable of interest is usually the response or outcome variable in predictive modeling, and in Statistical Pattern Recognition, the goal is normally very simply to predict the accuracy or reliability of recognition that is possible using a particular classifier or model based on key variables or features that can be estimated and calculated with coefficients for each key variable or feature as part of a general equation. Although they are still not exactly the same by strict definitions, for all practical purposes, the nature of features and classifiers is approximately the same as the independent and dependent variables that are input variable and fit variables of any response model that can be built by SAS in PROC GLM or PROC LOGISTIC, such as in the following:

OBS RESPONSE VARS (i.e., "FEATURES") TARGET VARS (i.e., "CLASSIFIERS")

1	(x1 x2 x3 x4 x5 x6 x7.. xn)	(y1 y2 .. yn)
2	(x1 x2 x3 x4 x5 x6 x7.. xn)	(y1 y2 .. yn)
3	(x1 x2 x3 x4 x5 x6 x7.. xn)	(y1 y2 .. yn)
..	..	..
N	(x1 x2 x3 x4 x5 x6 x7.. xn)	(y1 y2 .. yn)

It should also be understandable that the defined distribution assumptions and characteristics constitute possible features of all possible patterns, both real and imaginary, and part of the objective of all pattern recognition is to separate the actual features from the features that may be artifacts or some distortion of the data itself. This can best be addressed by probabilities the same as in many other applied statistical techniques. It is also important to apply all of the same statistical principles that are based on the foundations of statistical assumptions, such as sampling theory, information theory and entropy.

For example, assuming that an input audio signal recording of speech can be divided into intervals of appropriate fixed size, such as around 10 kHz with frame of about 12 ms, the beginning and end of typical phonemes will be reflected in adjacent intervals, since research indicates each spoken phoneme is of average duration of between 10 to 20 ms. This is a duration sizing window that is suitable and easily handled by SAS processing speeds on all platforms, and it is a simple matter to combine adjacent intervals to a single phoneme. Given that the first half of the phoneme has a ramp up in both amplitude and

frequency that can be detected and used to pair phonemes as they are initiated, and conversely to split each known half-phoneme similar to the same methods used in halftone transforms of Gray-Scale images, it is useful to either convert to binary or index by a scaling algorithm.

Thus, in each signal frame, additional statistical transforms or pre-process algorithms or rules are usually appropriate prior to loading the data to best represent a key feature of interest, or Region Of Interest (ROI) as this is most commonly called in image understanding problem domains. This includes the need to standardize the ROI into sub-image arrays, such as 16x16 or 32x32 pixels based on the resolution needs of the application and the available hardware and storage.

So when the requirement is to measure a random event or detect the pick-off point as well as duration of the event, which can be extremely critical in realtime target acquisition and tactical solutions that are commonplace in so many military pattern recognition systems, the random measure of X_j for the random variable X_j contains information that can be derived from the outcome response that will be directly dependent on its vector probability. So if the probability to observe the random variable X_j is small, a large degree of information must be assumed and predictable with greater risk, since the occurrence of the random variable is rare.

In contrast, if the random variable X_j has a large probability of being observed in reality then there will most likely be a small amount of information that will be observable. Thus, very simply, what happens easily and often and fast, such as the population of all cars that will pass you on the street, and you are looking for any red car, this problem domain will be easily solved as any red car is a problem domain that is easily recognized, and captured, many times even if you only look up a few times in an hour.

But what happens rarely, such as to personally see a particular make of an exotic and expensive very fast sports car in the color of red, and only if a convertible, may require higher sampling levels to look up every few minutes, and must likely sample a much larger universe. This might be handled by walking around the city continuously for many months, or years, in order to see the same red convertible sports car in real life that you first saw in a magazine and wanted to see for real. But if you do not have months or years, and just can not wait to see this particular sports car, then it might be better to use a series of statistical procedures to first find the best location to see sports cars, such as on a street corner near a sports car club, or the location where that type of car is sold, and then to begin your process to sample and watch for more suitable and economical use of your time and resources, and have better probability to succeed.

At a recent seminar sponsored by a leading university to discuss Statistical Pattern Recognition, when this example was given, a theoretical mathematician in the audience protested, saying: "This is simply cutting corners, and it defies all of the rules for empirical models, because if we allow solutions to problems that do not follow the entire mathematical process from start to finish then we will not really understand what or why we have a solution". This motivated one statistician in the audience to respond: "No that is not really cutting corners, that is simply Statistical Pattern Recognition -- and the difference is fundamentally the same as between all theoretical sciences and engineering, because the objective of SPR is not to try to understand all the known universe to explain all variables involved in reality -- which are all involved in trying to understand what a car is, what the color red is, what makes people drive cars here or there, or pick a particular street corner to stand on, or all of the other hundreds of things that someone might build more empirical models to describe. Yet the most critical point is that the person who wants to find that beautiful red car does not really care about any of that, the purpose is to recognize car -- and that is what differs between the Next Generation of pattern recognition, which is Statistical Pattern Recognition, and the Last Generation, or Second Generation pattern recognition based strictly on mathematically rigorous models may be universal but take forever to formulate and prove or solve just one, meanwhile many new applied SPR systems can be developed".

So it should be clearly understood that greatest value of SAS programming language to the design, development and integration of advanced Statistical Pattern Recognition programming objects into advanced Artificial Intelligence applications normally has its goal to create an improved device or process according to the most common objectives of engineering as opposed to theoretical science, although both SPR and SAS can be used for these purposes also. But the greatest strength and benefit of using SAS for SPR is related to its rapid applied statistics capability.

Yet this is not to say that SAS cannot provide the basis for mathematical rigor as well as applied statistics and practical engineering benefits. With regard to the previously defined problem domain, it is

possible to formalize these conditions with the observation that random variables with a large probability of being observed have a very small degree of information. Thus, as is one of the most basic tenants of Information Theory, the amount of information that can be contained in any data source or derived from it can be defined as:

$$-x_i = -\log p(x_i)$$

This is commonly known as the most important property of any randomized information source, which is also known as "entropy" and can also be more formally defined according to the actual average of the real information, such that:

$$H(S) = - \sum p(x_i) \log p(x_i)$$

The important part is that entropy is the measure of the amount of the information that can be derived and used beyond the simple derivation of that classifier or model and applies it based on the information that is coming in from the best understanding of the ongoing or realtime analysis of the data. This is the critical basis for all improvements upon all the Second Generation pattern classifier methods that are most similar to response models, and what is the basis for the Third Generation of pattern recognition, Statistical Pattern Recognition, which is easily applied and implemented using SAS programming language. However, it is important to understand that all the previously discussed tenants of pattern recognition, such as Signal-To-Noise ratios, and other entropy related components that influence the strength and information of signals, all still apply under SPR.

It is also very critical to understand at least 3 major aspects of traditional classifier pattern recognition technology in order to apply them using SAS.

First is the nature of comparative spatial versus spectral features, which are very similar to the previously discussed considerations to compare analog and digital signals to raw images and templates. Second involves methods of filtering and smoothing pattern recognition data. Just as smoothing is a widely understood concept to behavior response modeling statisticians and SAS programmers, it is just as important for SPR programmers and statisticians using SAS for pattern recognition to understand both basic filter method and advanced wavelet transforms. Third is traditional Numerical Pattern Classification, since this is normally a method that can either augment or circumvent transposition from unidimensional to multidimensional processing, and several methods integrate both steps directly from raw numerical data, and can be implemented by using PROC GLM if the data is linear versus using PROC LOGISTIC or PROC NLIM if there is detectable nonlinearity.

The spatial and spectral nature of speech and image features is first of all critical to the nature of SPR input data and how it is both represented and stored, but it is also important to be aware of several critical aspects of spatial versus spectral features themselves in order to effectively apply them within SPR applications and problem domains. It should first of all be understood that SPR techniques require that statistical assumptions are adhered to, which obviously requires that any feature must be analyzed within the confines of a distribution of sufficiently large size and sampling to determine its normality characteristics.

This clearly requires the need for collecting repeated observations to measure particular features within regions of particular interest, which is normally in the form of time-slices or frames or windows, as in the case of signals, and blocks or grids or cells in the case of images. Spatial features are best characterized by sampling without transforming the signal across or into another frequency band or time domain. Most efficient SPR algorithms for spatial feature detection must work on raw quantized signals, which may be analyzed within frameworks specific by types of spatial features. On the other hand, spectral features not only are more possible to sustain through transformation of the data, but many SPR techniques assume particular transformations, such as the Fourier transform for example, which can be a critical part of preprocessing of the data.

In both spatial and spectral feature measurements and computations, many of the same statistical properties should be examined, such as zero-crossing rate, autocorrelation, sequence and average slope for zero-crossings, and the time base energy and its rate of ramp-up. Some pattern recognition problem domains and application domains, however, such as speech recognition for example, may require additional feature dimensions that map back or provide link functions from corresponding unidimensional signal features to image or other multidimensional mappings.

In the case of speech recognition, a very specialized spectral feature category is the so-called "cepstrum" which defines cross-linkages between the frequency and duration features for speaking of particular phonemes to the anatomical dimensions of a speaker's mouth, such that the data from any incoming speech signal or data stream can be quickly recognized as fitting within the normal distribution of

audio speech values for a small child, or large man, for example. After this first discrete recognition is made it is possible to enable faster continuous speech recognition by weighting or adjusting values from the audio input and evaluating using a smaller vector space table that are most appropriate for the identified class of speakers. Similar SPR techniques are used in image as well as speech problems, which can all be developed using SAS data step programming and SAS statistical procedures.

It is also important to understand statistical foundations for filtering and preprocessing signals, which can all be implemented using SAS logic conditions. This will include basic linear filters and moving average smoothing methods, as well as rank order operations, conditional averages and reconstruction.

The speech classifier linkage method previously described is only one of many techniques that are the basis for Second Generation pattern recognition, which involves Automated Classification (AC) of features by using vectors to link classes. One most common is Numerical Pattern Classification (NPC).

Numerical Pattern Classification is a broad topic with both theoretical framework and practical opportunities for implementation using SAS. Many methods of NPC are supported by SAS and fall into the domain of categorical analysis, whereby features or variables can be assigned according to a set of formalized constraints into classes which are defined as sets of discrete categorical variables. This approach is most useful not only for the type of speech recognition of audio signal inputs as previously described, but is also very useful for very complex multiple-channel applications such as used in many systems to concurrently monitor multiple dimensions of patient life sign signals from body attached sensors. As in all these methods, the signals can be highly degraded by ambient noise in the data, as a function of probabilistic environmental and behavior factors, and thus are most effective when statistical procedures are used in order to estimate and weight to remove as much noise as possible before starting the process of Numerical Pattern Classification to assign features to classes.

The design of automated classifiers are fundamental to all the most advanced Statistical Pattern Recognition technology, and can involve such statistical procedures as multiple linear discriminants with class constraints, polynomial classifiers, Bayesian classifiers, geometric classifiers, nearest neighbor classifiers, and many other decision rule threshold and vector space methods. Just as the First Generation of pattern recognition could be understood with a framework of traditional Information Theory, these automated classifiers as are used in the Second Generation of pattern recognition are easily understood within a framework of what has come to be formalized as conventional Decision Theory.

The Third Generation of pattern recognition, which is also regarded as most current state-of-the-art in pattern recognition, is often generally referred to as Statistical Pattern Recognition (SPR). Several other emerging technologies within the Third Generation of pattern recognition are also fairly widely known, such as Image Understanding (IU), Bio-Metric Identification (BMI) and Fuzzy Logic (FL) are all considered state-of-the-art pattern recognition methodologies, and although their problem domains may be more or less specific, these methodologies all share a large number of common statistically-based foundations, and all are just as easily designed and applied using SAS programming and SAS statistical procedures as SPR.

All of this Third Generation of pattern recognition, including SPR, are each characterized by at least 3 very important commonalities. One, each of these advanced methods have "autonomous understanding" as a primary goal. This does not necessarily mean that SPR and other Third Generation pattern recognition has a pure goal of totally understanding real world behaviors and acting on its own, however these may be part of problem domains that may be best supported by SPR and other Third Generation pattern recognition methods. The goals of such autonomous understanding systems as can be designed using SPR methods may also include pure prediction as a primary goal, as in previous generation automated classifier systems, and indeed as well as predictive response or behavior models.

The meaning of "autonomous understanding" here is that desirable outputs of problem domains suitable to SPR methods may benefit if a training file architecture is involved which can enable autonomous derivations of recommendations or solutions based on previous iterations of previous processing of SPR solutions. Thus a critical consideration of autonomous understanding is it involves "training".

Two, the Third Generation pattern recognition tools are based on both methods and computational architectures that often involve some combination of Artificial Intelligence (AI) as well as advanced systems engineering architectures such as parallel processing, distributed storage arrays, and association networks that often approximate and emulate corresponding biological or psychological systems such as the human brain or perceptual motor processing. This is not to say that all Third Generation in general, and SPR pattern recognition in particular, have a primary goal to mimic the human brain or any sensorimotor system, although the methods that are used in SPR may in fact be based upon statistical and mathematical

frameworks that are actually based upon empirical study and analysis of the brain or human perceptual processes, such as is the case with Kohonen formalizations which led to various SOM (Self Organizing Map) methods available as SAS/EM nodes. In fact, this common perception that Statistical Pattern Recognition and many of its underlying techniques such as SOM and Artificial Neural Networks (ANN) has often led to unreasonable expectations among many technologists as well as the public that these methods actually involve capabilities to replace skilled humans. But this is not the case, and it should always be understood SPR is just a device.

Three, it is most important to point out that all of the Third Generation of pattern recognition, including SPR, are in the strictest sense still "supervised" rather than unsupervised predictive problem solving techniques. Such supervised prediction models are not different in concept and problem domains than automated classifier models that were discussed previously, which consist of analyzing a number of observations with a vector of input variables, predictors, or features, and fitting a model to a particular target response variable. Such supervised prediction models can handle most types of pattern recognition problem domains very effectively, and the introduction of training set or iterative training processes to fit the model provides for very reliable classifier design for most recognition.

The number of such supervised prediction statistical procedures that are provided with SAS/Base and SAS/Stat software, as well as SAS/Enterprise Miner data mining tools, combined with its powerful and flexible SAS/Multi Dimensional Data Base, SAS/Interactive Matrix Language and its many visualization and Java/C++ integration platforms for web localization and distributed delivery systems, make SAS the premier programming language of choice for both small and large scale advanced Statistical Pattern Recognition applications. The most important of these SAS statistical procedures and tools include supervised prediction methods, as well as advanced associative network architectures supporting many unsupervised methods.

There are a wide variety of supervised prediction procedures in the SAS/Stat arsenal that can be used to build state-of-the-art Statistical Pattern Recognition applications to address every category of pattern recognition problem domain. Many of these involve such methods as: Univariate Function Estimation; Non Linear Regression; Non Parametric Regression; Kernel Regression; Spline Smoothing; Multivariate Function Estimation; Posterior Probabilities; and Multiple Regression, to name but a few.

There are in fact many practical difficulties in effectively applying network architecture methods such as Artificial Neural Networks (ANN's) and Self Organizing Maps (SOM's), as well as decision trees and perceptrons, to approach and implement solutions to Statistical Pattern Recognition domains. The most critical of these techniques, which are solidly within the arsenal of proven SPR techniques, can also be implemented using SAS programming language and SAS procedures. Some of these include: Multiple Linear Regression, Logistic Regression, Sigmoidal Functions, Link Functions, Hidden Layers, Feed Forward Neural Networks (FFNN), Multiple Layer Perceptrons (MLP), Ordinary Radial Basis Functions (ORBF), Normalized Radial Basis Functions (NRBF), Flexible Discrimination Analysis, Nearest Neighbor Analysis, K-Clustering Analysis, and Logistic Discrimination.

In addition, several of the most advanced and cutting edge "unsupervised" prediction and automated classification techniques are included in the SAS arsenal of statistical procedures for SPR and other Third Generation pattern recognition techniques. These include: Principal Components Analysis (PCA); Projection Pursuit methods; Interest Index measure; Binocular Cortical Model (BCM); Polar Nearest Neighbor (PNN); Non Linear Feature Extraction (NLFE); Multi-Dimensional Scaling (MDS); Dissimilarity Coefficients; Principal Co-Ordinate Analysis (PCOA); Sammon Mapping; K-Means Clustering; Adaptive Resonance Theory (ART); Fuzzy Clustering (FC); Tree Vector Quantization (TVQ); and Self-Organizing Maps (SOM's), all of which can be implemented in SAS/EM or SAS/IML.

It should be noted at this point that these Statistical Pattern Recognition techniques based on Unsupervised Prediction or Unsupervised Classification are in fact not Third Generation techniques, these are all the emerging state-of-the-art "Next Generation" in SPR and pattern recognition technology, which will probably soon be formalized and regarded as Fourth Generation pattern recognition. However, it is important to note that these methods are still being refined and are not yet as reliable or controllable as the previous generations of more conventional pattern recognition techniques, and are demonstrated to the extent that these methods are considered reliable and available as SAS procedures or SAS/EM nodes. Yet, because each of these and other emerging new SPR technologies can be fairly easily implemented using SAS/IML and SAS data step programming, and will probably all be available as SAS/EM nodes and SAS/Stat procedures in the future, they are techniques that every SAS user should begin to be aware of.

Author Information:

Keith A. Jones, Ph.D.
2913 Myrtlewood
Plano, TX 75074

Telephone: (972) 424-8487
E-Mail: drkajones@aol.com