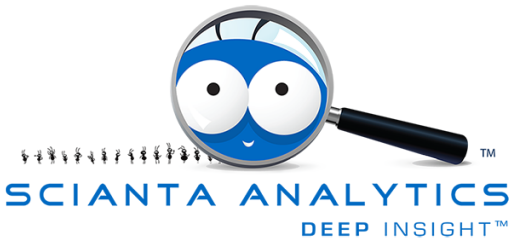
Anomalous Behavior Discovery, Quantification, and Modeling

A Non-Parametric Approach to
Detecting Provider Fraud and Abuse
in Managed Healthcare



www.sciantaanalytics.com

WP-01-001-20141101

Table of Contents

Introduction and Background

Conservative estimates of provider fraud (and abuse) – that is, fraud committed by doctors and other care givers – range between 10% and 12% of the roughly $650 billion spent annually on healthcare in the United States. Given the enormous amounts of money involved in the American healthcare industry, the shallowness of the regulatory oversight, the complexity of today's medical service protocols, and the relative ease with which abusive behaviors can be disguised or buried in the high transaction volumes processed by most insurers, it is easy to understand how abusive and ultimately fraudulent behavior can arise. Managed healthcare fraud is further complicated by the fragmentary nature of the patterns themselves – claims are dispersed across time and across many different insurance companies often renamed and defined according to different medical protocols so that no single insurer or oversight agency has a complete picture of a provider's activities.

The difficulty in detecting managed healthcare fraud is generally compounded by the insurance industry's inertia and its lack of a consistent, automated methodology. Typically, fraud detection has depended almost exclusively on claim processing administrators and auditing analysts noticing abnormal properties in claims documents. When claims were essentially paper documents manually processed by evaluators and then key punched into the insurance company's data processing system, this surveillance may have been feasible and acceptable. Today almost all claims processing is handled through on-line networks. Not only does the on-line nature of claims processing remove the source documents from the scrutiny of experienced auditors and claims processors, but the sheer volume of claims data, the complexity of the mixed private and public managed healthcare system, and the legal constraints imposed by Health Insurance Portability & Accountability Act of 1996 (HIPAA) regulations make human fraud detection a quaint bit of ancient history.

Recent attempts to bring computing power to bear on the problem of managed healthcare fraud have been ambitious but equivocal at best. Statistical analysis has proven inadequate to the over-all task (although in some narrow cases it has been quite successful). Statisticians continue to run up against a large number of roadblocks,

- ✓ Lack of clean and confident data
- ✓ Huge volume of data
- ✓ Lack of clear and workable objective functions'
- ✓ High dimensionality of the data
- ✓ Time series orientation with lead and lag relationships
- ✓ Missing and fragmentary data

On the other hand data mining, expert systems, decision trees, and neural networks, the workhorses of machine intelligence, have also generally made little headway in discovering and quantifying early stage fraud[1] and abuse. In most cases they run up against

[1] It is early stage fraud that is the most difficult but the most critical type of anomaly to catch. Intercepting providers before they become large scale abusers of the system not only saves vast amounts of money, but also returns a provider to their practice and eliminates a node in the abuse network. Late stage fraud, in any case, is relatively easy to find – an insurance company needs only sort their payment schedule in order of decreasing dollars and investigate the top N providers.

the same difficulties as the statisticians. Machine learning approaches, such as neural networks and decision tree generators, as well as subject matter expert approaches such as expert systems continually discover that finding sufficiently large numbers of fraud and abuse in historical data – necessary to either train the learning machine, or finding experts that can specify all the more subtle and hence important rules for fraud simply do not exist.

These parametric approaches have generally failed, in our view, because they rely on a definition of fraud that comes from :outside the system." That is, the definition is imposed on the model from either historical cases or from a statistical analysis of claims data. Models that relay on historical cases are too static and brittle to adapt to continuing and rapid changes in medical, regulatory, economic, demographic, delivery, and recording methodologies. Models that rely on statistical analysis alone cannot explain their reasoning, generate classification rules, handle incomplete and noisy data, and are often poorly designed to deal with very high dimensional, time-varying data.

Our approach, on the other hand, is non-parametric – it evolves from the behavior of a large population of provider peers. It flows from the observation that communities share a common behavior. A set of managed healthcare providers in a delivery organization of the same size, in the same medical specialty, and within the same geographic regions, will through shared experiences, constraints, and random interactions behave in pretty much the same way. By quantifying these characteristics, we can build a mathematical model of the provider peer population's behavior reflecting the elasticity of individual providers. The model is fashioned after Boltzmann's approach to statistical mechanics – while individual behaviors of molecules in a gas cannot be predicted, the aggregate or large scale behavior of the gas can be precisely modeled. The same is true, we assert, for individual providers in their large community of similar providers.

Detecting Anomalous Behaviors

A fundamentally intuitive approach to isolating abusive and fraudulent healthcare providers is based on the concept of non-parametric anomaly detection within the a large population. The over-all behavior a sufficiently well-contained and well-represented population, assuming the data is normally distributed[2], is governed by the Law of Large Numbers and the Central Limit Theorem[3]. In particular, the weak Law of Large Numbers say that, for a sufficiently large population, the measurement of behavior characteristics will approach the mean for the population. Thus, we can "look into the data" to find providers that have behaviors significantly at variance form their peers. At its core, the algorithm for locating anomalous behaviors is fairly straight forward and deceptively simple:

For each provider (P_i)
  Analyze and quantify P_i's behavior
  Compare P_i's behavior to their peer's behavior (S)
  If P_i is significantly different from S on important measures
    Then select P_i for Ranking
End for each

Central to this algorithm is the concept of a *peer group*. A peer group represents a collection of managed healthcare providers (P_1, P_2, P_3, ..., P_n) that work within the same type of business organization (individuals, small clinics, large clinics, small hospitals, large hospitals, etc.), have the same medical specialty, and are physically located in the same geographic area[4]. It is this backbone peer clustering that drives the anomalous behavior model. In this way the model need only know the normal behavior for the collection of peers in order to isolate those individuals who have behaviors which fall significantly outside those of their peers.

When this algorithm has been executed against all the healthcare providers the outcome is a collection of anomalous providers ranked according to their aggregate degree of difference form their peer group. The degree of difference is the weighted sum of the differences for each of the behaviors. Thus, when sorted by the this measure, the ranking produces a list of anomalous providers in order of increasingly abnormal behaviors.

[2] This is a starting assumption **not** a necessary requirement for the anomaly detection process.

[3] The most general case of the Central Limit Theorem (originally discovered by de Moivre) states that data influenced by many small and unrelated (independent) random effects can be approximated by a normal distribution.

[4] Geography is the easiest patient-centric constraint to impose on the model (the type and specialty constraints are, of course, provider-centric constraints). However, we could very well replace geography with a measure of demographics or sponsorship (for public or military agencies, as an example). In any case, however, a restricted set of patients is a crucial component of the behavior model.

Behavior Patterns

The actual properties of a healthcare provider that represent its "behavior" are often difficult to isolate. This difficulty arises from several issues.

- First, there are a large numbers of elementary data items underlying a claim (procedure codes, patient profiles, time references, costs, and other line items required by the insurance companies as well as by regulatory oversight agencies.)
- Second, many of the properties are not elemental – as an example, the number of patients per day over a particular time interval, the number of different procedures performed over a time interval, the frequency distribution of patients by gender and age, and the number of repeat patients handled over a particular time interval.
- Third, behavior patterns are seldom static – they occur over time, are seldom cleanly obvious (that is, the patterns are confused with noise), are usually incomplete, and many of the patterns are only faintly represented in the data over time.

As we will discuss shortly, the use of clustering and self-organizing maps - advanced machine intelligence and data mining methodologies - are ideally suited to handling large, multi-dimensional volumes of data, are fault and noise tolerant, and can look deeply into the data to find periodic, often transient, patterns that simultaneously span many dimensions.

Behavior Measurement Classes

In order to provide a uniform and coherent (not to mention consistent) platform for measuring behavior patterns, we organize the behaviors into a set of measurement classes[i]. Each class exposes a collection of subclasses. As Table 1 indicates, there are five major behavior categories in the current fraud model,

Category	Behavioral Measurement
FINANCIAL	The flow of money through the system
IDENTIFICATION	How providers identify themselves to the insurer
LOGISTICS	The place, time, and sequence of activities associated with each claim
MEDICAL LOGIC	Indicates whether or not a medical or diagnostic situation would normally occur given the precedence relationships of procedures
STATISTICS	Frequency and statistical properties of treatments, visits, and procedures

Table 1. The Major Behavior Categories

Within these behavior classes a variable number of analysis heuristics are used to measure individual behavior attributes. As an example, in the LOGISTICS class, THE PATIENT-

TRAVEL-DISTANCE heuristic measures the travel distance between the patient and provider's address. It is important to understand that the behavior analysis is deeply multi-dimensional. A naïve or single dimensional approach to anomaly detection would focus on a single measure – say the average distance traveled – and exposes only a single discrimination dimension – whether a provider's travel measures tend be greater than or less than their peers. However, our analysis is based on distribution analysis; hence the distance measurement reflects the degree to which the measurement is more or less spread out than the corresponding peer population (that is, the measure incorporates the standard deviation of the population in the form of the third and forth statistical moments - kurtosis and skew[ii]). Distribution based measurements allow the anomalous behavior analysis to find, as an example, providers that draw all their patients from a specific neighborhood.

Measuring Population Properties

There are several facets to measuring the characteristics of a population. Underlying the cluster of individual measurement sub-categories is a more fundamental analysis of the statistical properties of each population. Figure 3 illustrates the distributions, for a particular behavioral measurement, of the background peer and the current provider populations.

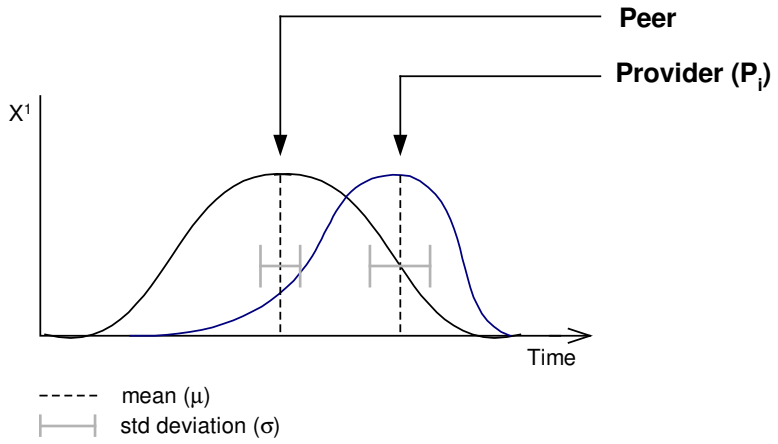


Figure 3. Peer and Provider Populations

There are, of course, conventional statistical analyses that can discriminate and measure the difference between two probability distributions (as well as determine whether two samples are drawn from the same population). These statistical measures, however, only provide clues in a search for the over-all trends and contingent differences that indicate the deep mechanics of population behaviors. In the case of anomaly detection, we are concerned with finding outliers in a very high dimensional population space that represent quantifiable abnormal behavior. To do this we apply three principal component methodologies to the model.

- First, all the underlying data is normalize to generate a symmetric scaling in the range [0,1]. To do this compute the values for the behavior measurement (if it is a ratio or other calculated field), determine their statistical properties (basically the

mean and standard deviation), and then convert the value to its z-score (or z-statistic):

$$z = \frac{x_i - \overline{x}}{\sigma}$$

This process, using the mean and standard deviation, transforms any normal random variable into a variable with a mean of [0] and a standard deviation of [1].

- Second, we measure the difference between expected and actual behavior in the high dimensional space through the use of a consistent Information Entropy Number (EN). This number measures how far out of line a provider is with their peers for each of the behavior measures. The EN is actually the sum of the Euclidean distances from the centers of all the overlapping clusters for the current behavior measurement"

$$d_k = \sum_{j=1}^{N} \left\| X_j^k - C_j^i \right\|^2$$

(Refer to the discussion of clustering, below, for a more detailed discussion of the idea behind clustering and emergent behaviors.)

- Third, although not discussed at length in this paper, we interpret all the data and boundary constraints in the model as either fuzzy numbers or quantifiers expressed as fuzzy sets. This removes the brittleness of the over-all model and allows, as one example, a data point to belong to several clusters at the same time – each with a different degree of membership. For a more detailed discussion of fuzzy sets and fuzzy models see many of the books and articles cited in the References.

With this understanding behind us, we can turn our attention to locating centers of behavior (clusters) in multi-dimensional space. This is the role of automatic cluster detection and self-organizing maps – a form of unsupervised data mining. The question naturally arises – what are we clustering? The cluster consists of the z-scores for each of the behavior measurements (which are subclasses of the measurements outlined in Table 1). Cluster analysis and detection finds patterns from large collections of data points. These clusters represent patterns of behavior that shift across time. Each pattern has a center (representing the centroid or center of gravity) for the collection of points in that cluster[iii].

Clustering and Self-Organizing Maps

In searching for anomalous behaviors the fraud detection system simultaneously examines the relationships between the behavior metrics across the complete N-dimensional space[5] and forms a set of clusters for each of the peer populations (that is, for each time invariant slice through the underlying three dimensional peer structure.) Understanding an N-dimensional space is a bit difficult. Let's begin by examining a smaller space – the number of patients seen by a provider. Figure 4 illustrates, for a particular time period, a simple clustering of patient patterns for a specific organizational size, specialty, and geographic location.

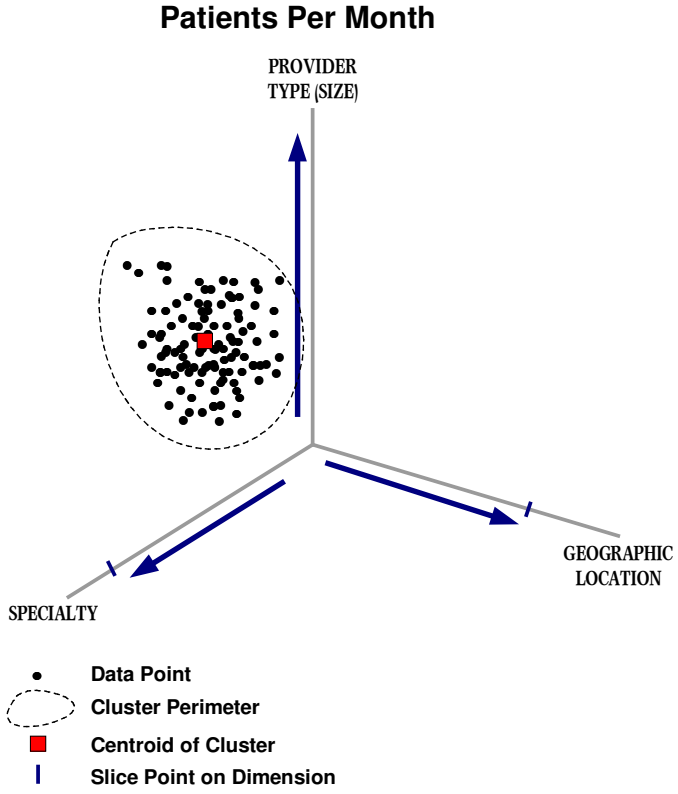


Figure 4. Example of A Cluster
Small Clinic, Osteopath, Greater St. Louis Metropolitan Area

In this significantly simplified clustering example, we can see the distribution of patient counts for small osteopathic clinics in the greater St. Louis (Missouri) area (the additional dimensions of count and month of year have been omitted for clarity.) The clustering begins to expose the natural behaviors of all the St; Louis small osteopathic clinics.

[5] The cardinality of N depends on the total number of sub-class measures in the system (see Table 1). A peer cluster space of these N-dimensions is created for each distinct point in the size-specialty-location topology so these dimensions do not contribute to the space size. Naturally, if we consider one of the topology axes as a fuzzy number (such as the organization size or the extent of the geographic (or demographic) coverage) then the topology spaces cannot be crisply separated and we must fold the fuzzy dimensions into the underlying decision space.

Since the underlying data represents all the clinics and we assume that only a small fraction of these are involved in fraud or abuse[6], the cluster represents normal behavior.

This is essentially a six dimensional space. Each point in the cluster has a value, its statistical moment parameters, and an associated date-time stamp. Although the value axis is not explicitly shown (for relative simplicity), the cluster represents a slice though the same time frame (such as the month of May 2004, implying perhaps, that each data point is aggregated from a lower level set of raw data – the provider claims for patients throughout the month). We can therefore view the peer population as a set of clusters moving through time. Figure 5 illustrates this concept by showing the ellipsoid clusters for April and May.

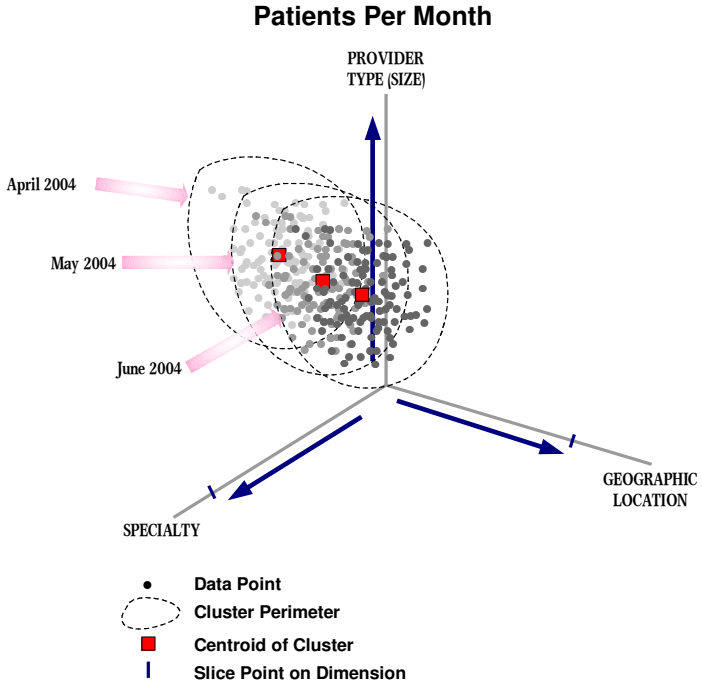


Figure 5. Patient Count Clusters Moving Over Time

If we now build collections of population clusters over time for each of the low level behavior patterns, we begin to find an emerging pattern. The underlying periodic cycles of the population are revealed. This emergent property reflects a crucial and intrinsic aspect of peer collections – that in order to understand normal behavior and in order to find anomalous behaviors we must understand how behaviors vary periodically over time. Thus, in our discussion of clustering and emergent behaviors, bear in mind that a time axis is always implicit in the higher dimensional space.

[6] Of course this might be an unwarranted assumption in some parts of the country. But, in any case, we could either (1) assume that out model will then detect the clinics that are abusive even for all the other abusive clinics or (2) drop or expand the geographic location constraint on the model thereby comparing the behavior of each clinic relative to (as a few examples): all other clinics in the state, all other clinics in the region (southeast, midwest, northeast, etc), or all other clinics in the country.

Cluster Properties

The more compact a cluster, the more homogeneous is the overall behavior among all members of that cluster. As we move out from the centroid, the behavior become less and less characteristic of the cluster. This falling off of similar behaviors is marked by a set of imprecise but never-the-less observable layers. These are called Behavior Frontiers. Figure 6 illustrates how these boundaries are arrayed in concentric layers out from the centroid[7].

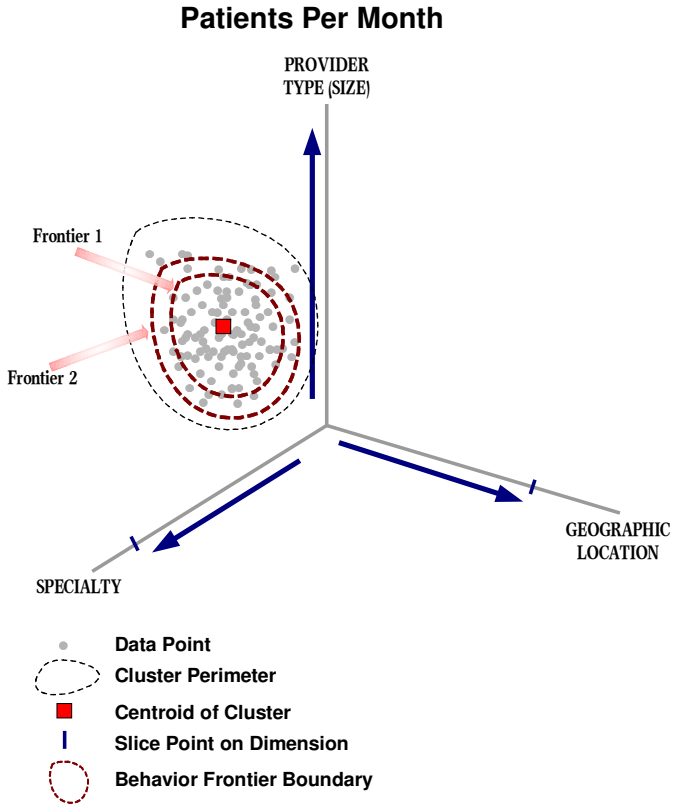


Figure 6. Behavior Frontiers in a cluster

Frontiers provide a formal way of ranking the degree of difference in the behavior *within* a cluster – that is, they are an intrinsic property of the cluster and are used to measure the degree of compactness (or, conversely, the degree of dispersal). Both peer population clusters and provider clusters have frontier layers as a measure of their compactness[8].

In practice, however, individual behavior anomalies, regarless of their lack of compactness, are generally insufficient to either rank a provider as abusive or to allocate resources to continuing an investigation. An anomaly model must construct a clear picture of the relationships between all the associated behavior measurements and determine the

[7] In actual fact the frontier boundaries are non-linear, that is, they approximate the area under the first, second, third and forth standard deviations of the Gaussian (normal) distribution. The actual boundary is, however an elastic band represented imprecise or fuzzy numbers.

[8] This would be expected, of course since a provider population is necessarily a subset of the peer population.

degree to which a provider is anomalous in some significant portion of this N-dimensional space. What does this N-dimensional space look like?

We can get a general idea by adding some more and slightly different dimensions to the model – the ratio of patient counts to total billings and the mean distance traveled by patients. Two seeming unrelated observables, but, in a behavior model, they are brought together under the relentless search for consistent, on-going patterns in the data. Figure 7 shows a greatly simplified view of how patient counts and billings emerge as a set of clusters.

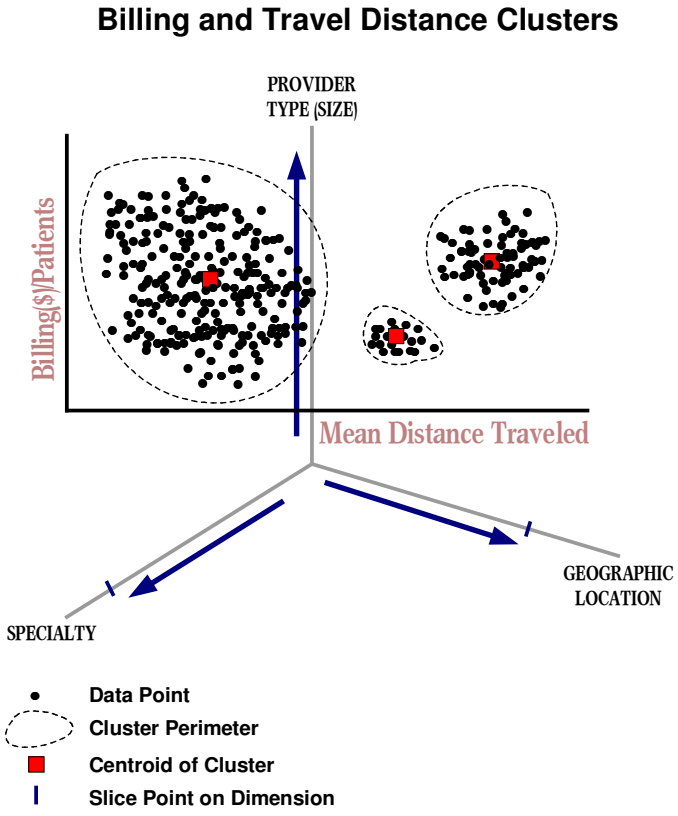


Figure 7. Higher Dimensional Clustering

An interesting aspect of higher dimensional analysis is the emergence of local clusters. These clusters tell us interesting and deep things about the behaviors rippling through our N-dimensional space. Figure 7 shows this phenomenon, when we move from looking just at patients to dollars per patient and distance traveled the data points begin to cluster in interesting ways. In this clustering we see that most of the clinic's patients come from relatively near by and the billing dollars are distributed more or less evenly through this groups. As we might expect, however, for osteopathic clinics, some group of patients come from farther away (referrals) and bring in a higher fee per patient. These kinds of behaviors emerge naturally out of the data itself.

Clusters are overlaid on vast amount of data scattered through the N-dimensional space according to various randomization patterns. As a consequence of this scattering, the boundaries of a cluster are, to a large degree, semi-permeable, that is, a data point can belong to several clusters simultaneously. This ability follows from the way data points are associated

with a cluster. A point is not "in" or "out" of a cluster, but has a degree of membership in that cluster. In this way, the actual underlying behaviors are not rigidly forced into or out of a cluster. Figure 8 illustrates a slightly more realistic (but still highly simplified) collection of emergent clusters with their overlapping regions.

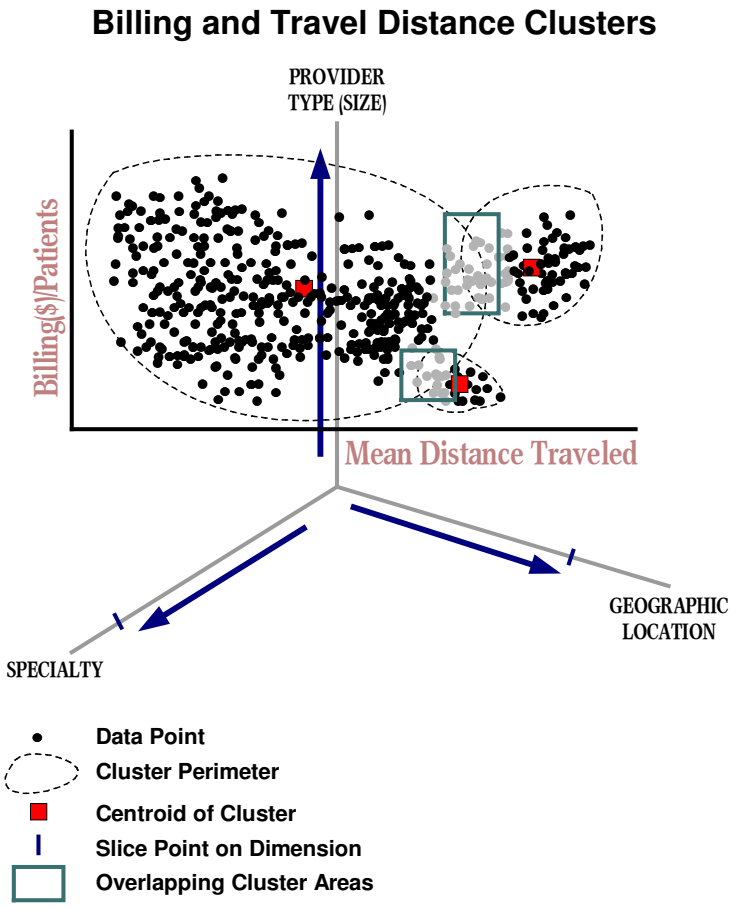


Figure 8. Overlapping Behavior Clusters

This kind of flexibility is important because behavior representation in cluster analysis has the same kind of ambiguity and on-linear separability as behaviors in the real world (which, of course, we are attempting to model). Data points on the edge of one cluster (beyond the second frontier boundary) may lie well within the inner frontier of another cluster – thus turning abnormal behavior into a conditional form[9] of normal behavior. We should expect this kind of fluidity in behaviors, even those restricted to a single time frame. The population of patients handled by a provider are not themselves homogeneous, nor is the provider's behavior entirely consistent. These natural irregularities give rise to unexpected and unpredictable segmentation in the data.

[9] Without delving deep into the underlying mathematical apparatus, data points that lie toward the edge of a cluster but are inside the perimeter of another cluster are subject to an adjacency and sufficiency algorithm that considers the size, transience, and evidence of both clusters to determine whether or not the data point is counted as an anomalous instance.

Measuring Anomalous Behavior

When the N-dimensional peer space has been constructed and populated with nascent clusters, we can measure the difference between the peer space and each provider space. The algorithm is surprisingly simple (of course all the hard work has gone into constructing the self-organizing maps (clusters) that represent the behaviors).

For each Provider (P_i)
  Create an N-Dimensional Cluster space of its Behavior Measures (B_1, B_2, …, B_k)
  For each Peer Behavior Measure (S_j) where j = 1 to N
    Find D_m = Unsupported-Difference(S_j,B_j)
    If D_m > Threshold
      Store D_m else Store 0 in Measure Vector(j) for P_i
  End for each Peer
  Measure Vector() = Measure Vector() * weight vector()
End for each Provider

The core of the algorithm is the Unsupported-Difference() method. This procedure (alluded to earlier) examines the degree of membership of B_j in all overlapping (S_j,B_j) clusters to determine whether or not it represents actual anomalous behavior or whether it belongs to the normal behavior of one of the other proximity clusters. The method returns a zero for normal behavior otherwise it returns the maximum distance from the best fit normal behavior cluster.

When this algorithm is applied to each provider, the outcome is a weighted real valued vector of N-values (D_1, D_2, D_3, ... D_N) specifying, for the j-th behavior measure, the degree of difference between the peer behavior and the provider's behavior. The larger this number (the closer it approaches the N, since this would be the Euclidean distance between two centroids at opposite sides of the configuration space) the more anomalous the behavior. Figure 9 shows how the outcome set of vectors appears after each provider has been evaluated.

provider	D_1	D_2	D_3	D_4	..	D_N
P_1	0	.3	0	0	..	.7
P_2	1	0	0	0	..	0
P_3	0	0	.2	.3	..	0
:	0	0	0	.4	..	.2
P_k	.5	0	.1	0	..	.6

Figure 9. The Outcome Measure Vector()

The use of a vector to contain all the differences for each measure also allows the selection and ranking mechanism to focus on specialized filtering methods, such as providers that have a high frequency of anomalous measures or a few measures with very large anomalous values.

The *weight vector()* couples the purely analytical processing of the cluster detection and self-organizing maps with the business policies and constraints of the enterprise. Through the set of associated weights the outcome vector emphasizes or dilutes the contribution of each behavior measure. Thus, the associated weight vector, a real array of scaling weights also dimensioned 1....N, provides expert (client) weighting of behavior measures. This allows the application of particular emphasis on certain kinds of anomalous behavior as well as the reduction or exclusion of emphasis on other kinds of behaviors. The weights are normally set to [1], thus eliminating any special emphasis.

Ranking Anomalous Behaviors

The result of the anomalous detection algorithm is a set of Measure Vectors, one for each provider. Each vector contains an array of difference measures (each vector position is either zero for normal behavior, or a value greater than zero that measures the relative degree of variance from the peer behavior). Taken together, these values represent a positioning of the provider in a large space of anomalous providers. In order to focus on the most significant of the providers, the vectors are sorted using a scalable, multi-objective, multi-constraint ranking procedure. The method, Minimum-Entropy, Ordered Weighted Aggregation (ME-OWA) effectively applies an operator aggregation or partitioning weight to each attribute (the measure value) and generates an ordered ranking of the providers. Figure 10 provides a mathematical overview of the aggregation and ranking mechanism.

$$R \leftarrow \begin{bmatrix} D_{1,1} & D_{1,2} & .. & D_{1,N} \\ D_{2,1} & D_{2,2} & .. & D_{2,N} \\ : & : & : & : \\ D_{k,1} & D_{k,2} & .. & D_{k,N} \end{bmatrix} \bullet \begin{bmatrix} w_1 & w_2 & w_N \end{bmatrix}$$

Figure 9. The ME-OWA Ranking Mechanism

The OWA weight operators adjust the way in which the measure difference values are combined to form a final ranking. The operator weight changes the AND-ness or the OR-ness of the aggregation process making it possible to alter the form of aggregation through an entire range of operator strengths. While it is beyond the scope of this paper to explore the underlying mechanics and principles of the OWA process, the weights provide a way to tighten or relax the criteria for a constrained ranking of the providers. From this final ranking process the anomaly detection system provides an ordered list of providers arranged in decreasing degree of aggregate difference between themselves and their peer group. Because the ranking vectors maintain a 1:1 coherence between the rank and the underlying measure, a complete evidence support, case support, and explanatory capability is available.

Further Reading

Cox, E.D. (1991) “Approximate Reasoning: The Use of Fuzzy Logic in Expert Systems and Decision Support”. Proceedings of the. Conf. on Expert Systems in the Insurance Industry, 24-25 April, Institute for International Research, New York.

Cox, E.D. (1991) “Company Acquisition Analysis: Formulating Queries with Imprecise Domains”. Proceedings of the First Intl. Conf. on Artificial Intelligence Applications on Wall Street, 9-11 October, IEEE Computer Society Press, Los Alamitos, CA. 194-9.

Cox, E.D. (1992) “The Great Myths of Fuzzy Logic”, AI Expert, January, 40-5.

Cox, E.D. (1992) “Solving problems with fuzzy logic”, AI Expert, March, 28-37.

Cox, E.D. (1992) “Integrating fuzzy logic into neural nets”, AI Expert, June, 43-7.

Cox, E.D. (1992) “Fuzzy fundamentals”, IEEE Spectrum, October, 58-61.

Cox, E.D. (1992) “Effectively Using Fuzzy Logic and Fuzzy Expert System Modeling — in Theory and Practice”. Proceedings of the Conf. on Advanced Technologies to Re-Engineer the Insurance Process, 17-18 September, Institute for International Research, New York.

Cox, E.D. (1992) “Fuzzy Logic and Fuzzy System Modeling”. Proceedings of the Fourth Annual IBC Conf. on Expert Systems in Insurance, 28-29 October, IBC USA Conferences, Southborough, MA.

Cox, E.D. (1992) “Applications Of Fuzzy System Models”, AI Expert, October, 34-9.

Cox, E.D. (1992) “A Close Shave With Occam's Razor: Fuzzy-Neural Hetero-Genetic Object-Oriented Knowledge-Based Nano-Synthetic Reasoning Models: Throwing The Kitchen Sink At Problem Solving”, A Workshop in the Industrial Applications of Philosophy and Epistemology to AI, Proceedings of the Third Annual Symposium of the International Association of Knowledge Engineers, 16-19 November, Software Engineering Press, Kensington, MD.

Cox, E.D. (1993) “Adaptive Fuzzy Systems”, IEEE Spectrum, February, 67-70.

Cox, E.D. (1993) “A Fuzzy Systems Approach To Detecting Anomalous Risk Behaviors In Portfolio Management Strategies”, Proceedings of the Second Intl. Conf. on Artificial Intelligence Applications on Wall Street, 19-22 April, Software Engineering Press, Gaithersburg, MD,144-8.

Cox, E.D. (1993) “Fuzzy Information Systems With Multiple Conflicting Experts”, Proceedings of the Computer Design Magazine's Fuzzy Logic'93 Conference, March, M223-1-13.

Cox, E.D. (1993) “A Model-Free Trainable Fuzzy System For The Analysis Of Financial Time-Series Data”, Proceedings of the Computer Design Magazine's Fuzzy Logic'93 Conference, March, A124-1-7.

Cox, E.D. (1994) *The Fuzzy Systems Handbook*, Academic Press Professional, Cambridge, MA.

Cox, E.D. (1999) “The New Face of Fuzzy Logic: A Close Shave With Occam's Razor”, PC/AI Magazine, May/June, Vol.13, Issue 3

Cox, E.D. (1999) "What Does Your Company Really Do? Data Fusion in the Era of Knowledge Management", PC/AI Magazine, July/August, Vol.13, Issue 4

Cox, E.D. (1999) "Striving for Imprecision Fuzzy Knowledge Bases for Business Process Modeling", PC/AI Magazine, July/August, Vol.13, Issue 4

Cox, E.D. (1999) "Rediscovering What You Do: A Data Mining and Rule Discovery Approach to Business Forecasting with Adaptive, Genetically-Tuned Fuzzy System Models", PC/AI Magazine, Sept/Oct, Vol.13, Issue 5

Cox, E.D. (1999) "Coping with the Uncertainty Principle: Predictive Project Risk Assessment and Risk Classification Using a Fuzzy Case-Based Reasoning System", PC/AI Magazine, Nov/Dec, Vol.13, Issue 6

Cox, E.D. (2000) "Distributed Intelligence in the B2B Universe: Fuzzy and Neural Connections In Web-Centric Knowledge Management", PC/AI Magazine, May/June, Vol.14, Issue 3

Cox, E.D. (2000) "Free-Form Text Data Mining: Integrating Fuzzy Systems, Self-Organizing Neural Nets and Rule-Based Knowledge Bases", PC/AI Magazine, Sept/Pct, Vol.14, Issue 5

Cox, E.D. (2001) "Building Intelligent Business Applications with Semantic Nets and Business Rules", PC/AI Magazine, Jan/Feb, Vol.15, Issue 1

Cox, E.D. (2001) "XML and Distributed Business-to-Business Intelligence: Fusing XML and Java-based Expert Applications", PC/AI Magazine, March/April, Vol.15, Issue 2

Cox, E.D. (2001) "Fuzzy Logic and the Measures of Certainty in eCommerce Expert Systems", PC/AI Magazine, May/June, Vol.15, Issue 3

Cox, E.D. (2001) "The Knowledge Navigator: An Auction Metaphor for the Brokering of Corporate Knowledge Asset", PC/AI Magazine, July/August, Vol.15, Issue 4

Cox, E.D. (2001) "Building Intelligent Models from Data Mining and Expert Knowledge: A Look at Fundamental Principles", PC/AI Magazine, Sept/Oct, Vol.15, Issue 5

Cox, E.D. (2001) "Intelligence in Terrorist Detection - Our New Secret Agents: Fuzzy Logic and Computational Intelligence in Threat Detection, Validation, and Interdiction", PC/AI Magazine, Nov/Dec, Vol.15, Issue 6

Cox, E.D. (2002) "Knowledge-Based Business Process Modeling – Complex Systems Design Through a Fusion of Computational Intelligence and Object Oriented Models", PC/AI Magazine, May/June, Vol.16, Issue 2

Dubois, D. & Prade, H. (1980) *Fuzzy Sets and Systems: Theory and Applications*. Mathematics in Science and Engineering, Vol. 144. Academic Press, San Diego, CA.

Dubois, D. & Prade, H. (1988*) Possibility Theory, An Approach to Computerized Processing of Uncertainty*. Plenum Press, New York.

Jamshidi, M., Vadiee, N. & Ross, T.J. (eds) (1993) *Fuzzy Logic and Control*, Prentice Hall, Englewood Cliffs, NJ.

Jones, P.L. & Graham, I. (1988) *Expert Systems: Knowledge, Uncertainty and Decision*, Chapman & Hall, London.

Klir, G.J. & Folger, T.A. (1988) *Fuzzy Sets, Uncertainty, and Information*, Prentice Hall, Englewood Cliffs, NJ.

Kosko, B. (1992) *Neural Networks and Fuzzy Systems, A Dynamical Systems Approach to Machine Intelligence*, Prentice Hall, Englewood Cliffs, NJ.

Kosko, B. & Isaka, S. (1993) "Fuzzy Logic", Scientific American, July, 76-81.

Masters, T. (1993) *Practical Neural Network Recipes in C++,* Academic Press, San Diego, CA.

Pedrycz, W. (1993) *Fuzzy Control and Fuzzy Systems*, 2nd edition, John Wiley, New York.

Schmucker, K.J. (1984) *Fuzzy Sets, Natural Language Computations, and Risk Analysis*, Computer Science Press, Rockville, MD.

Smets, P., Mamdani, E.H., Dubois, D. & Prade, H. (1988) *Non-Standard Logics for Automated Reasoning*, Academic Press, London.

Smithson, M. (1987) *Fuzzy Set Analysis for Behavioral and Social Sciences*, Springer-Verlag, New York.

Terrano, T.,Asai, K. & Sugeno, M. (1991) *Fuzzy Systems Theory and Its Applications*, Academic Press, San Diego, CA.

Wang, L. and Mendel, J. M. (1991) "Generating fuzzy rules from numerical data, with Applications", USC-SIPI Report No. 169, Signal and Image Processing Institute, University of Southern California, Los Angeles, CA

End Notes

[ii] Organizing behaviors into a set of comprehensive, predefined classes adds a structure to the anomalous behavior analysis that is ultimately derived from expert knowledge. A more adaptively oriented approach, and one that we follow in many risk assessment and fault analysis systems, is to use a form of genetic algorithms to explore the interplay of elementary and calculated fields to find those that seem to have a set of common relationships. This approach parallels a form of Principal Component Analysis and thus can dynamically adjust a set of behavior classes. In this high level document we focus on a set of predefined behavior classes to reduce the over-all complexity of the discussion and focus on the underlying analysis mechanisms.

[ii] Kurtosis measures the shape of the population distribution. There are three basic shapes – leptokurtic (sharply peaked), mesokurtic (the middle curve form, very similar to a normal or Poisson distribution), and platykurtic (a flat distribution). Figure E1 illustrates these various kinds of distributions.

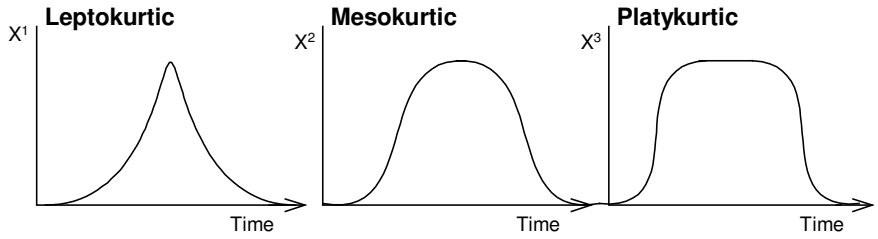


Figure E1. The Three Basic Kinds of Population Distributions

Skew, as the name implies, indicates a population that is shifted away from its mean value, thus producing either a right or left trailing tail. Figure E2 illustrates how distributions can be skewed in one of two directions.

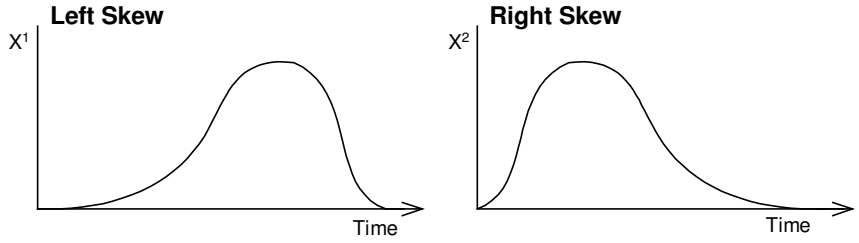


Figure E2. Left and Right Handed Skews

[iii] Membership in a cluster is determined by measuring the distance from a cluster center to a data point (see previous section, **Cluster Attributes**). Figure E3 shows the array of points around a cluster center with the distance metrics for several points.

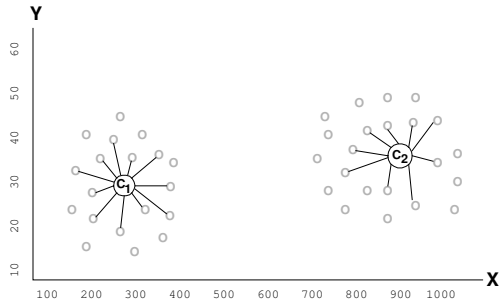


Figure E3 Data Point Distances within A Cluster

The over-all process of clustering takes the distance between a point X_m and each of the clusters $C_{1...k}$ as the simple sum of the differences between X_m and C_i (any cluster center). If there are *n* attributes in each cluster vector, we can see this in Expression E1,

$$diff_{k \to C} = \sum_{i=1}^{n} (X(x_i) - C(x_i))^2 \qquad \text{(Exp. E1)}$$

Here the Euclidean distance between the point and a center is calculated. Other distance methods can also be used, but the Euclidean calculation is fast and appears to work very well in a wide variety of real-world clustering applications. More formally, perhaps, computing the distance between a point and the cluster center is done through a process called sum of the squared difference. Expression E2 provides the mathematical summary of the approach.

$$d_k = \sum_{j=1}^{N} \left\| X_j^k - C_j^i \right\|^2 \qquad \text{(Exp. E2)}$$

Note: A distance from X_k is calculated for each of the cluster Centers ($C_{1...i}$) using the *N* attributes in each vector. For the distance (d_k) from point X_k to the center C_i we sum the squared difference between the j^{th} attribute of X_k and the j^{th} attribute of the i^{th} cluster center. From this we can see that if the X_k lies on or close to the cluster center the value will be at or near zero; on the other hand as the distance between X_k and C_i increases, the distance measure also increases rapidly.