A Statistical Framework for Cluster Health Assessment and Its Application in Anti-Money-Laundering Systems

By using cluster analysis to continuously assess the health of peer groups used by anti-money-laundering systems, banks can better understand the reasons for cluster deterioration over time.

Executive Summary

The art and science of clustering is used in many fields, from ubiquitous customer segmentation for gauging marketing effectiveness, to predicting default patterns among credit card holders, to its recent application by financial institutions to segment investment customers to enhance liquidity management. In the area of anti-money-laundering, the use of peer grouping, or segmenting, is even more prevalent as it provides a tailor-made solution for detecting unusual transactional activities.

The premise: Customers are expected to exhibit the transactional behavior of the peer group in which they fall; any deviation is deemed unusual. Increased sophistication in statistical methodologies and advancement in IT solutions have ensured that peer grouping becomes the foundation of various anti-money-laundering (AML) solutions. Despite this, however, there hasn’t been much development around approaches that could ensure that predefined peer group, or clusters, remain healthy (i.e., reflective of precise transactional behaviors); peer group validation remains overlooked. A peer group is called healthy when a majority of its constituents exhibit similar characteristics and dissimilar characteristics to constituents of other peer groups.

This white paper describes the ways in which the health of a cluster or peer group may be erroneous or bad due either to poor choice of segmentation variables or to the movement of entities between clusters over time. Importantly, it proposes a generic statistical methodology to provide an objective assessment of peer group health. In other words, this paper provides a methodology to create a quantitative indicator of the extent to which a peer grouping system under consideration conforms to the fundamental traits of a good peer group.

Healthy Clusters: A Definitional Foundation

Clustering is increasingly used across various fields in innovative ways and has proved to be extremely helpful in predicting customer behavior and identifying outlier patterns. Nevertheless, most techniques employ primitive methodologies to update or maintain clusters. This paper proposes a generic framework that can
be used to assess cluster health and improve the predictive capabilities of such a clustering system by locating the probable causes of cluster health deterioration.

There are several reasons why clusters can be said to have deteriorated over time. The most notable factors include:

- Clusters are often created on the basis of expert judgment, which is liable to go awry when markets turn dynamic.
- Segmentation variables, which were selected to create clusters, may not be the most appropriate ones and do not differentiate constituents of clusters enough to produce clear segments.
- Legitimate changes in clusters due to actual behavioral change of customer groups over time.
- Poor-quality data or lack of data while forming clusters, necessitating a relook given the availability of new data.
- Additional information gained over time about cluster constituents may demand a relook at existing clusters.

We begin by explaining the traits of a healthy peer group, followed by the methodology to assess health and identify the reasons for health deterioration – notably concerning the top two bullet points above. The assessment methodology adopted, though largely generic, can be used only if the problem conforms to a particular structure. We explain the methodology based on our analysis on a set of real customer data. We also briefly describe various statistical measures of cluster health and when they can be used.

**Traits of Healthy Clustering**

From a business perspective, the degree of presence of each of the factors explained further down (e.g., identifiability, compactness, etc.) will define how good or bad the cluster system is. It is notable to mention here that there are statistical measures that correspond to particular parameters: entropy, for example, measures both the homogeneity and separation of clusters.

There are important caveats to all of this. The parameters mentioned below are not entirely independent of each other. A high degree of homogeneity and compactness is likely to be observed. But this is not necessarily true in cases where clusters are highly dispersed and highly separated (i.e., within one customer segment), behaviors are not too similar (big, dispersed clusters), but there are reasonable differences from other groups (large differences among clusters).

- **Identifiability/homogeneity (entropy, purity):**
  - Can we see clear differences between segments?
  - Is the transaction behavior of one peer group sufficiently different from that of other peer groups?
- **Compactness (variance ratio, additive margin):**
  - Are the data points of each cluster as close to each other as possible? A common measure of compactness is variance, which should be minimized.
- **Separation (L-separatability, entropy):**
  - The clusters should themselves be widely spaced.
  - Measured by distance between two clusters: single linkage, complete linkage, comparison of centroids.
- **Substantiality:**
  - Are the segments large enough to warrant separate groups and expected transactional differences?
  - Does the peer grouping need to change if there are very few data points in one peer group?
- **Stability:**
  - Do the peer groups remain stable over a certain period of time?
  - Can we implement dynamic profiling if over a period of time customer behavior changes?
- **Scalability:**
  - The peer groups should be able to accommodate and/or transform in the case of a huge number of varied data points.

**A Brief Methodology**

The following methodology can be used to assess the health of customer groups formed on the basis of a set of business variables. The terms “clusters” and “peer groups” may be used interchangeably for all practical purposes.

- **Variables definition:** At the outset, let us define some terms that are going to be used later in the paper.

A set of variables that the organization uses to create/form clusters is referred to as initial/
input/system variables. Demographic variables are a typical example. These are essentially different from the variables used to validate the clustering actually formed, which are referred to as observed/output variables. Observable transactional behavior variables such as value or volume of transaction are typical examples of such kinds of variables.

The methodology contains the following steps:

1. The initial assumption is that the input/initial variables used to create customer peer groups by the organization will correctly predict the customer behaviors in terms of the observed variables.

2. Using various tools, the organizations create clusters based on the initial variables.

   - **Clustering on initial variables:** Generally, organizations create clusters based on a set of initial variables (different from observed variables) that are available while forming clusters. In anti-money-laundering, for example, the initial set of variables is annual income, age, living area type (city/village/town) and product types. The nonavailability of data on output or observed variables is the driver. In the context of AML solutions, output variables depict transactional profiles such as value and volume.

   - **Analysis of observed variables:** For any general business problem, the health of clustering can be judged by looking at the groups of observable variables that define the constituents of that cluster. For example, in the case of an anti-money-laundering peer group, the clusters of customers formed should be “good” based on their transaction profiles. That means transaction profiles of all constituents of a peer group or cluster should be somehow similar. Transaction profiles are represented by the transaction volume, transaction value and transaction types. Specifically, this means that the clusters, which were formed at the time of system configuration and used for detection of unusual transactions, should be “good” when assessed using the observed variables — transaction values, volumes and types.

   - **Quick Take**

     **Statistical Measures**

The mathematical details of each of the measures mentioned here are explained in the glossary.

- **Entropy:** Entropy is a measure of the homogeneity of objects with a single class label (here, types of products). If the resulting clusters are not healthy based on entropy or purity, it is assumed that the clustering is bad and needs to be redone. If the entropy calculations yield satisfactory results within a predefined confidence interval, then further means of cluster analysis like variance ratio, additive margin and L-separatability can be applied.

- **L-separatability:** This can be simply described as the ratio of the distance of each point in the entire population with the population centroid to the distance of the combined two clusters from their average centroid. A lower value indicates a better separatability from the adjoining cluster.

- **Additive margin:** Simply put, this is the ratio of the average difference between the distance of points of a cluster to its centroid within the same cluster and the centroid of the nearest cluster to the average within cluster distance. A higher value indicates better quality.
A representation of the methodology discussed in Steps 2 & 3 can be elucidated by Figures 1 and 2. The figures plot the transactional behaviors (transactional value, transactional volume) of a set of ~100,000 records of customer data for anti-money-laundering systems of a leading U.S. brokerage firm.

Figure 1 represents the initial expectations of the customer transactional behaviors in terms of observed variables during the peer group configuration phase. Figure 2 represents the actual observed transactional behaviors, showing clusters corrupted due to one or both of the following reasons over a period of time:

- Specifically in anti-money-laundering transactions it may be argued that the customers that were grouped together on the basis of some parameters may have moved/changed over time to other groups due to a legitimate change in their characteristics. Typically, organizations do not regularly check the peer groups formed, and hence the discrepancy.
- It is possible that the initial variables expected to correctly predict the customer behaviors were wrongly chosen. For example, the initial set of variables used to create clusters might have included “gender,” which does not necessarily reflect customer behavior in the long term. It is also quite possible to have missed important input variables such as “income” in the initial variable set, resulting in poor grouping.

These reasons, among other scenario-specific reasons, provide an insight into the deterioration of the health of clusters over time.

Assumed Transactional Behavior of Peer Groups Per Premise

![Figure 1](image1.png)

Actual Transactional Behavior of Peer Groups Over Time

![Figure 2](image2.png)
4. If the health of the clusters is “bad,” the organization should take steps to:
   - Reconsider and redefine the initial variables taken.
   - Check if the clustering has deteriorated, not because of wrong initial variables chosen but due to time-dependency.

5. The organization should remedy the problems identified, and repeat the process again for validation.

Calculating Cluster Health

A perfect clustering would mean that groups assessed on the basis of observed variables are healthy and thus the clusters formed using initial variables remain good in terms of observed variables as well. Different business requirements may lead to a different selection of statistical measures (e.g., entropy, L-separatability, additive margins, etc.). In the beginning, business decisions must be made to determine the allowed value and variation of the measure being used.

However, calculation of clusters’ health using a complete set of observed parameters may not be possible. Not all parameters and their effects can be quantified. For example, a creditissuing company developing parameters to form customer segments may focus on salary segments (in dollars) and types of products (loans, credit cards, etc.), among other criteria. While it may be easy to plot and cluster the customers using salary figures and to analyze clusters, it is difficult to visualize the type of product, which is a non-ordinal variable that can’t be plotted.

This difficulty can be eliminated by using the statistical measure of entropy to see if clusters formed are homogeneous in nature.

Step Sequence for Calculating the Health Index for Typical AML Systems

This section depicts the complete sequence of steps or the framework used to calculate the health of peer groups using techniques identified in earlier sections of this paper. This generic framework can be used, with relevant scenario-specific modifications, to approach the cluster health problem.

Figure 3 represents the sequence of steps leading to the final statistical measure of cluster (peer group) health. It is assumed that the “health” of clusters on the basis of non-ordinal measures such as transaction types can be calculated with the help of entropy.

A System Architecture to Analyze Cluster Health

A perfect clustering would mean that groups assessed on the basis of observed variables are healthy and thus the clusters formed using initial variables remain good in terms of observed variables as well.
Applying the Framework to Different Clustering/Peer Group Systems

To apply our framework to calculate peer group health, the peer group or clustering system should conform to some basic constructs. For instance:

- Clusters should have been formed to put constituents having similar profiles together.
- These profiles have to be represented by two or more dimensions. At least two of these dimensions should be representable either quantitatively or in an ordinal manner. All of these dimensions should be equally important in business decision-making; moreover, these dimensions should be orthogonal to each other.
- While creating these clusters, data on these dimensions should not be available. Hence, these clusters should have been formed using some other “predictor” variables, referred to as initial variables in this paper.
- These profiles, as mentioned in the first bullet point above, should be an important consideration while making business decisions. For example, in case of AML systems, the transaction profile of a customer determined whether that customer should be declared suspicious.

Although this construct seems very specific, it is found in most scenarios where clustering is used. However, careful consideration is required to fit a given problem in the above construct, so that a peer group health index framework can be applied in the most appropriate way.

Looking Forward: Additional Applications

While this paper demonstrates the use of this framework in the operational risk area of finance, the generic nature of this framework makes it extremely versatile and pliable for applications in a wide range of subject areas that span financial services, consumer marketing and behavioral analytics. As such, all that is needed for such a peer group health assessment is proper understanding of the subject area and an intelligent analysis of initial and observed variables.

The benefits of this approach have already been seen in the transaction monitoring area for anti-money-laundering systems. This methodology helped a brokerage firm identify issues with its peer groups, which led to corrective measures and eventually to a reduction in the number of false alerts.

The recent emergence of big data technologies supporting high density data, velocity and other parameters can enable faster and easier implementation of this framework. We mention some common fields where the cluster health index can be applied:

- Transaction analysis in anti-money-laundering systems for customer groups.
- Customer segmentation for credit-card-issuing organizations.
- Marketing effort validation for marketing campaigns for targeted customers.
- Mutual fund rebalancing for segments of stocks grouped by stock price movement characteristics.
Glossary

The mathematical details of the statistical measures used in this white paper include the following:

- **Entropy:** To calculate the entropy of a set of peer groups, we first compute the class distribution of the objects in each peer group – i.e., for each cluster $j$ we compute $p_{ij}$, the probability that a member of cluster $j$ belongs to class $i$. Given this class distribution, the entropy of cluster $j$ is calculated as

$$E_j = - \sum_{i} p_{ij} \log(p_{ij})$$

(1)

taken over all classes. The total entropy for a set of clusters is computed as

$$E = \sum_{j=1}^{k} \frac{n_j}{n} E_j$$

(2)

the weighted sum of the entropies of all clusters, as shown in (2), where $n_j$ is the size of cluster $j$, $k$ is the number of clusters, and $n$ is the total number of data points.

- **L-separatability:** Measures like L-separatability help normalize the loss functions to obtain scale invariance.

$$L_{\text{Sepmax}}(C, X, d) = \frac{\max_{ij} d(c_i, c_j)}{\max_{i \in C} \{ j : d(c_i, c_j) \}}$$

is sensitive to maximal separation between clusters.

Here, $C$ is $\{C_1, C_2, \ldots, C_k\}$ is some k-clustering. $C_i$ is a clustering identical to $C$ except with clusters $C_i, C_j$ merged.

- **Additive margin:** If instead of looking at ratios we want to evaluate quality using differences, we use additive margin. The additive margin of a point $x$ is $C-\text{AM}_x = d(x, c_j) - d(x, c_i)$ where $c_i$ is the closest center to $x$ and $c_j$ is the second closest center to $x$ and $C$ is a center based clustering over $(X, d)$.

$$AM_x(C) = \frac{\sum_{i \in C} (d(x, c_i) - d(x, c_j))}{\sum_{i \in C} d(x, c_i)}$$

The range is $[0, \infty]$.

References

About Cognizant

Cognizant (NASDAQ: CTSH) is a leading provider of information technology, consulting, and business process outsourcing services, dedicated to helping the world’s leading companies build stronger businesses. Headquartered in Teaneck, New Jersey (U.S.), Cognizant combines a passion for client satisfaction, technology innovation, deep industry and business process expertise, and a global, collaborative workforce that embodies the future of work. With over 50 delivery centers worldwide and approximately 171,400 employees as of December 31, 2013, Cognizant is a member of the NASDAQ-100, the S&P 500, the Forbes Global 2000, and the Fortune 500 and is ranked among the top performing and fastest growing companies in the world. Visit us online at www.cognizant.com or follow us on Twitter: Cognizant.

About the Authors

Anshuman Sharma is a Financial Risk Consultant in the Governance Regulatory Compliance Practice within Cognizant Business Consulting. He has three-plus years of experience in the finance sector, focused on credit and market risk implementation, governance changes for Basel III/Dodd-Frank and regulatory reporting. He previously worked for the hedge fund D.E. Shaw & Co. and holds an M.B.A. in finance from XLRI Jamshedpur, India and an engineering degree from Indian Institute of Information Technology Allahabad, India. Anshuman can be reached at Anshuman.Sharma2@cognizant.com.

Raghvendra Kushwah is a Consulting Manager within Cognizant Business Consulting, heading the Operational Risk Division, and has deep domain experience in fraud and anti-money-laundering processes and implementation. His areas of expertise include analytics for behavioral finance, where he has led numerous operational risk and analytics projects across several geographies. He has eight years of experience in the IT and finance industries and holds an engineering degree from Indian Institute of Technology, Delhi and an M.B.A. from Indian Institute of Management Lucknow. He can be reached at Raghvendra.Kushwah@cognizant.com.

Anshuman Choudhary is a Director and heads the Governance Regulatory Compliance Practice within Cognizant Business Consulting. His areas of expertise include consulting in risk management and regulatory reporting. He has 14 years of business technology consulting and domain experience and is a qualified GARP financial risk manager. Anshuman has an M.B.A. in finance from Indian Institute of Social Welfare and Business Management and a bachelor’s degree in metallurgical engineering from REC Durgapur, India. He can be reached at Anshuman.Choudhary@cognizant.com.