Proceedings of Machine Learning Research 77:20–32, 2017 KDD 2017: Workshop on Anomaly Detection in Finance

Analytical Techniques for Anomaly Detection Through Features, Signal-Noise Separation and Partial-Value Association

Nong Ye NONGYE@ASU.EDU
Arizona State University, Tempe, Arizona (USA)

Abstract

This paper presents three analytical techniques for anomaly detection which can play an important role for anomaly detection in finance: the feature extraction technique, the signal-noise separation technique, and the Partial-Value Association Discovery (PVAD) algorithm. The feature extraction technique emphasizes the importance of extracting various data features which may be better at separating anomalies from norms than using raw data. The signal-noise separation technique considers an anomaly as the signal to detect and the norm as the noise and employs both anomaly models and norm models to detect anomalies accurately. The PVAD algorithm enables learning from data to build anomaly patterns and norm patterns which capture both partial-value and full-value variable relations as well as interactive, concurrent effects of multiple variables.

Keywords: feature extraction, signal-noise separation, partial-value variable association

1. Introduction

Anomaly detection is a critical task in many fields including fraud detection in finance, cyber attack detection, airport security, and so on. It is challenging to perform anomaly detection with a high level of accuracy (low false alarms and few misses). This paper introduces three data analytic techniques which can play important roles in achieving anomaly detection with the high level of accuracy: the feature extraction technique, the signal-noise separation technique, and the PVAD technique. The feature extraction technique and the signal-noise separation technique are originally developed and employed for cyber attack detection, as described in (Ye, 2008). These two techniques are introduced and presented in this paper in Section 2 and Section 3. Section 4 describes the PVAD technique which is developed to overcome a major shortcoming of existing data analytic techniques in having the same one model of variable relations over full ranges of variable values and to enable learning both partial- and full-value variable associations and both individual and interactive effects of multiple variables.

2. The Feature Extraction Technique

A feature is a measure of a property which exists in a sample of data observations for one or multiple data variables. Univariate features are features of a data sample from one data variable. Four univariate mathematical/statistical features are employed for cyber attack detection (Ye, 2008): statistical mean, probability distribution, autocorrelation, and wavelet-based signal strength. Mathematical/statistical methods of extracting the mean, probability distribution, autocorrelation, and wavelet features from data are described in (Ye, 2008).

Among the four features, the distribution feature gives a more comprehensive picture of a data sample than the mean feature. It is observed (Ye, 2008) that computer and network data fall in the following five types of data patterns:

- Spike (including upward and downward spikes),
- Random fluctuation,
- Step change (including step increase and step decrease),
- Steady change (including steady increase and steady decrease), and
- Sine-cosine wave with noise.

There is a link between the above data patterns and the following probability distributions (Ye, 2008):

- Left skewed distribution,
- Right skewed distribution,
- Normal distribution,
- Multimodal distribution,
- Uniform distribution.

A spike pattern produces a skewed probability distribution of data. Specifically, an upward spike pattern produces to a right skewed distribution with a right tail, meaning that most data observations have values falling into the lower part of the data range, and a few data observations spread over a part of larger values than those in the lower part. A downward spike pattern produces a left skewed distribution with a left tail, meaning that most data observations have values falling into the upper part of the data range, and a few data observations spread over a part of smaller values than those in the upper part. A random fluctuation pattern produces a symmetric, normal distribution. A step change pattern produces a multimodal distribution. A step change with two dominant levels of values produces a bimodal distribution. A step change involving several distinctive levels of values, such as the step change with one dominant level and a few other levels, produces a multimodal distribution with more than two modes. A steady change pattern produces a uniform distribution. A sine-cosine wave with noise pattern may lead to a normal distribution if there is much noise, or a uniform distribution if there is little noise and the sine-cosine wave is well-formed.

Both the wavelet feature and the autocorrelation feature reveal relations of data observations over time. Data is autocorrelated if data observations are correlated over time. The autocorrelation feature focuses on the general autocorrelation aspect of time series data, whereas the wavelet feature focuses on special forms of time-frequency data patterns.

Many objects have periodic behavior and emit special signals at a certain frequency. The frequency of signal in data over time has long been used for signal detection or object identification due to the fact that many objects have their special time-frequency signal characteristics. Fourier analysis has traditionally been used to analyze and represent signal frequencies. However, Fourier analysis does not reveal the specific time location of a given frequency characteristic. Wavelet analysis allows the analysis and representation of time-frequency signal in both frequency characteristics and their time locations (Ye, 2008, 2013). The Paul wavelet, the Derivative of Gaussian (DoG) or Mexican Hat wavelet, the Haar wavelet, the Daubechies D4 wavelet, and the Morlet wavelet, are used to extract and approximate the spike, random fluctuation, step change, steady change, and sine-cosine with noise data patterns observed in computer and network data (Ye, 2008).

If an activity (anomaly or norm) causes a significant change in a specific feature of a data variable, this change is considered a data characteristic of the activity. If a specific data characteristic appears during a given anomaly but not during other anomalies or norms, this data characteristic is considered a unique data characteristic of that anomaly and can be used to uniquely detect and identify this anomaly. Note that an activity may manifest through more than one data characteristic (e.g., more than one data variable or more than one feature of a data variable). The identified anomaly characteristics in the mean, distribution, autocorrelation and wavelet features can be used to uncover similarities and differences of the anomalies.

The application of the feature extraction technique to anomaly detection in finance lets us consider using not just raw finance data but also features, e.g., transaction frequency, which may better capture characteristics of normal transactions versus frauds/anomalies, for anomaly detection.

3. The Signal-Noise Separation Technique

The signal-noise separation technique considers anomalies as signals to detect and normal activities as noise. Because an anomaly occurs when normal activities are going on, data collected and used to detect the anomaly contains data effects of both the anomaly and normal activities. The mixture of both signal and noise data effects buries data characteristics of both signal and noise and thus makes it difficult to detect the signal using signal data characteristics directly or large deviations from noise data characteristics as anomalies. The signal-noise separation technique is developed (Ye, 2008) to handle the data mixture of signal and noise to achieve the high level of detection accuracy in the following steps:

1) Data modeling: define the signal data model and the noise data model to represent the signal data characteristic and the noise data characteristic,

2) Noise cancellation: use the noise data model to cancel the data effect of noise that is present in the data mixture of signal and noise, and

3) Signal detection: use the signal data model to detect and identify the presence of signal in the residual data from Step 2 after canceling the data effect of noise.

Steps 2 and 3 are designed to handle the mixed data effects of signal and noise. The signal data model and the noise data model defined in Step 1 are required in Steps 2 and 3.

In other words, a thorough understanding and an accurate modeling of both signal data and noise data are necessary to handle the mixed effects of the signal data and the noise data. In addition, the knowledge of how the signal data and the noise data are mixed together is necessary to enable Step 2. There are many ways in which the signal data and the noise data can be mixed together, e.g., in an additive manner, a multiplicative manner, and so on.

Many signal processing techniques exist to perform noise cancellation and signal detection. The cumulative score (cuscore) is used to carry out the signal-noise separation technique (Ye, 2008). The mathematical foundation and details of the cuscore are provided in (Box and Luceno, 1997; Box and Ramrez, 1991; Ye, 2008). If the signal data and the noise data are mixed in an additive manner as follows:

$$y_t = f(x_t) + \theta g(x_t) + \epsilon_t \tag{1}$$

where y_t is a data observation at time t, $f(x_t)$ is the noise data model, $g(x_t)$ is the signal data model, ϵ_t represents the white noise, and $\theta = \theta_0 = 0$ when no signal is present, the cuscore is:

$$Q_0 = \sum_{t=1}^{n} \left[y_t - f(x_t) \right] g(x_t) \tag{2}$$

When $\theta = 0$, Q_0 should fluctuate around zero. When $\theta \neq 0$, $[y_t - f(x_t)]$ in Q_0 has the element of $g(x_t)$ which is then correlated with $g(x_t)$ in Q_0 in Formula 2, making the Q_0 values to move upward or downward consistently, depending on the positive or negative sign of $[y_t - f(x_t)]\, g(x_t)$.

Note that $[y_t - f(x_t)]$ in Q_0 acts like canceling the effect of the noise data in the observed data, y_t, which has the effect of the noise data only when there is no signal and becomes the mixed signal and noise data when a signal is present. The residual from $[y_t - f(x_t)]$ is then correlated with the signal data model through multiplication to detect the presence of the signal defined by the given signal model. Hence, unlike conventional anomaly detection techniques which detect large deviations from a given normal use data model, the cuscore detects a specific anomaly defined in the signal data model under a given normal activity condition defined in the noise data model with a high level of detection accuracy.

If signal data and noise data are mixed in a multiplicative manner as follows:

$$y_t = \theta f(x_t) g(x_t) + \epsilon_t \tag{3}$$

where $\theta = \theta_0 = \frac{1}{g(x_t)}$ when no signal is present, the cuscore takes the following form:

$$Q_0 = \sum_{t=1}^{n} \left[y_t - f(x_t) \right] f(x_t) g(x_t) \tag{4}$$

When $\theta = 0$, Q_0 should fluctuate around zero; otherwise, the Q_0 values move upward or downward consistently, depending on the positive or negative sign of $[y_t - f(x_t)]\ f(x_t)g(x_t)$.

If the signal data and the noise data are mixed in other ways, the statistical model of cuscore provided in (Ye, 2008) can be used to drive the cuscore equation for other types of signal and noise mixtures.

Because the detection of an anomaly signal requires observing the cuscore values on a number of data observations moving upward or downward consistently, a single data observation is not sufficient to detect an anomaly signal using the signal noise separation technique. After a signal is detected by observing the consistently upward or downward moving of teh cuscore values for a number of data observations, the cuscore value needs to be reset back to zero to start a new cycle. If there are different types of anomaly signals that need to be detected, multiple cuscore equations should be developed for detecting multiple signal types, and multiple streams of cuscore values based on those equations should be employed to monitor the same data. If the specific form of the anomaly signal along with its model $g(x_t)$ is not known, the signal noise separation technique cannot be used to detect the unknown anomaly signal.

The signal-noise separation technique for cyber attack detection is illustrated in (Ye, 2008) to separate effects of the attack data and the normal use data in their mixture to enhance the performance of cyber attack detection. The application of the signal-noise separation technique to anomaly detection in finance lets us consider that normal transactions and frauds/anomalies are expected to occur in the same period, creating the signal-noise data mixture of frauds/anomalies and normal transactions. The use of the signal-noise technique for anomaly detection in finance will allow us to filter out data effects of normal transactions from the data mixture of normal transactions and frauds/anomalies to examine characteristics of frauds/anomalies and detect frauds/anomalies accurately.

4. The PVAD Algorithm

A significant shortcoming of existing data analytic techniques for learning from data to build models of variable relations is having the same one model of variable relations which covers all values of variables, that is, the full ranges of all variable values. In many fields, a relation may not exist among all values of variables but among certain values of variables, or different relations exist for different values of variables. For example, many physical systems behave differently under normal temperatures, extremely low temperatures, and extremely high temperatures so that system variables have different relations under different conditions. For another example, Fishers Iris data set (http://archive.ics.uci.edu.ml) provides four predictor variables (sepal length, sepal width, petal length, and petal width) of Iris plants and one target variable for the type of plants with three values of Iris Setosa, Iris Versicolour, and Iris Virginica. Only one value (Iris Setosa) of the target variable can be identified using the four predictor variables, whereas the two other values (Iris Versicolour, and Iris Virginica) of the target variable cannot be separated using the four predictor variables. Hence, for Fishers Iris data, we can define a relation of only one value of the target variable with certain values of the four predictor variables, and this relation is not applicable to other values, in other words, all values of the target variable. If variables have different relations for different ranges of values or relations exist only for certain values but not all values, the same one model of variable relations for all values of variables produced by existing data analytic techniques will not provide a good fit to data, that is, giving a poor representation of data and/or failing to learn and build adequate, accurate models of variable relations.

The PVAD algorithm is developed to overcome this shortcoming and learn from data to construct associative networks which are multiple-layer models of variable associations, capture both individual and interactive effects of multiple variables, and cover both partial-value and full-value relations of variables (Ye, 2017a,b). Each node of an associative network represents certain value(s) of certain variable(s). Each directed link from a conditional node I to an associative node J, in an associative network represents an association from node I to node J and indicates that certain value(s) of certain variable(s) in node I affects certain value(s) of certain variable(s) in node J. A node in an associative network can have one variable and its value or multiple variables and their values to capture interactive effects of these variables and their values on certain values of some other variables. An association can represent a causal relation such as I causing J or some other kind of associative effects.

The PVAD algorithm consists of three steps:

Step 1. Identify value intervals/categorical values of variables,

Step 2. Discover partial-value associations of variable values,

Step 3. Consolidate and generalize partial-value associations which are used to construct an associative network of variable associations.

In Step 1, we consider that there are two kinds of variables: categorical variables and numeric variables. A categorical variable already comes with its categorical values which can be used directly in Step 2 of the PVAD algorithm. For a numeric variable, we need to transform the numeric variable into a categorical variable. One method of transforming a numeric variable into a categorical variable is to plot the sorted values of the numeric variable, identify data clusters, and use the non-overlapping intervals of data clusters to define the categorical values of the numeric variable. Figure 1 shows an example of using this method to determine the data clusters and the categorical values of the numeric variable, Contact, from the chemical data set (http://www.stat.columbia.edu/˜gelman/book/data/). The data plot in Figure 1 shows 16 values of this variable in the sorted order of increasing values in three data clusters with the three non-overlapping intervals which are used to define three categorical values of c1, c2, and c3: [0.0115, 0.0135] = c1, [0.0260, 0.0410] = c2, and [0.0840, 0.0980] = c3. These three data clusters are identified by visually inspecting distances of consecutive data points in the data plot so that we have distances of data points within a data cluster smaller than distances of data points in different data clusters. We can also use elbow points in the data plot that start line segments with different slopes to identify data clusters since changes in line segment slopes indicates big changes in distances of consecutive data points. Clustering techniques such as hierarchical clustering (Ye, 2013; Ye, 2003) can also be used to produce the same data clusters.

Step 2 of the PVAD algorithm discovers partial-value associations of variable values, and each partial-value association is in the form of $X = A \rightarrow Y = B$ where X and Y are the vectors of one or more variables, A and B are the values of X and Y, respectively. X is called conditional variable(s), and Y is called associative variable(s), $X = A$ are called conditional variables' values (CV), and $Y = B$ are called associative variables' values (AV). Each data record in the data set is called an instance with instance #. Step 2 of the PVAD algorithm consists of the following steps.

- Step 2.1: discover 1-to-1 partial-value associations, $x = a \rightarrow y = b$, where the association involves only one CV and only one AV. For each value a of each variable x, each value b of each variable y, and the candidate association, $x = a \rightarrow y = b$, we carry out the following steps:

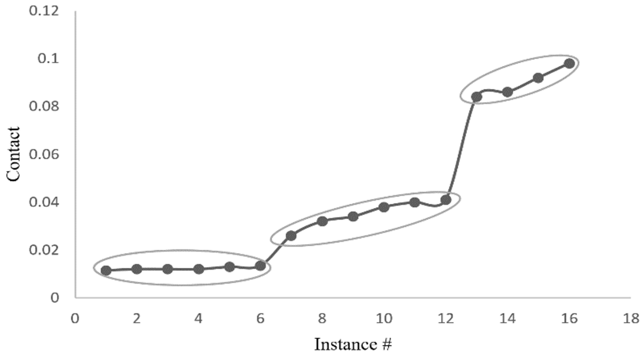

Figure 1: The plot of the sorted values of Contact from the chemical data set.

- Step 2.1.1: Compute the co-occurrence ratio (cr) of each candidate association, $x = a \to y = b$ as follows:

  $$\mathrm{cr}\,(x = a \to y = b) = \frac{N_{x=a,\ y=b}}{N_{x=a}} \tag{5}$$

  where $N_{x=a,\ y=b}$ is the number of instances containing both $x = a$ and $y = b$, and $N_{x=a}$ is the number of instances containing $x = a$.
- Step 2.1.2: store each candidate 1-to-1 association, including its cr value, N_{CV}, and the set of instances supporting this association with those instances containing $x = a$ and $y = b$, i.e., $N_{\mathrm{CV\ and\ AV}}$ and instance #'s, where N_{CV} is the total number of instances containing CV ($x = a$), $N_{\mathrm{CV\ and\ AV}}$ is the total number of instances containing CV and AV ($x = a$ and $y = b$). Any 1-to-1 association with $cr = 0$ or ∞, that is, $N_{x\ =\ a\ \mathrm{and}\ y\ =\ b} = 0$ or $N_{x\ =\ a} = 0$, is removed from the set of stored candidate 1-to-1 associations.
- Step 2.1.3: establish the 1-to-1 partial-value association, $x = a \to y = b$, if $\mathrm{cr}(x = a \to y = b) \geq \alpha$, where α is set to a value in the range of (0, 1] and is close or equal to 1. The associations with $cr \geq \alpha$ are called the established associations. Hence, an established association has cr $\geq \alpha$, and a candidate association may have any cr value in (0, 1].

- Step 2.2: discover p-to-q partial-value associations, $X = A \to Y = B$, where the association involves either multiple conditional variables or multiple associative variables, using the methods of YFM1 and YFM2 and the procedure shown in Figure 2 to generate all the established p-to-q partial-value associations. These methods and the procedure of using them in Step 2.2 are described later.

- Step 2.3: Among all the established associations from Step 2.1 and Step 2.2, use β to remove the associations whose supporting set of instances has fewer than the β number of instances, and use γ to remove the associations whose CV or AV exists in more than or equal to γ of all the instances in the data set, where β denotes the

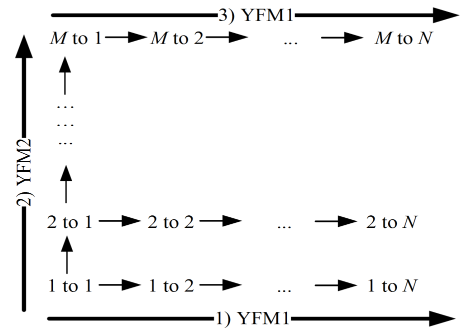

Figure 2: The procedure of using YFM1 and YFM2 to establish partial-value associations.

number of instances and can be set to a positive integer equal to or greater than 1, and γ is the percentage of instances in the data set and can be set to a value in (0%, 100%]. β is used to remove any association which is not supported by a sufficient number of instances in the data set. γ is used to remove any 1-to-1 association with its CV or AV being present in too many instances of the data set (too common in the data set). An association with such common CV or AV gives little meaningful, useful information of variable associations since the association is established solely due to the common presence of CV or AV in the data set.

In the method of YFM1 in Step 2.2, for each group of the established k-to-l associations (e.g., the established 1-to-1 associations from Step 2.1) with one of the following satisfied:

1) The associations in the group have the same or inclusive set of supporting instances and

   (a) the same CV: we establish k-to-q associations, $\{\text{CV}\} \to \{\text{AV}_q\}$, where $q > 1$ and $\{\text{AV}_q\}$ is any combination of AVs from the associations in the group, due to the following:

$$\begin{aligned} \text{cr}\left(CV \to \{\text{AV}_q\}\right) &= \frac{N_{\text{CV and }\{\text{AV}_q\}}}{N_{\text{CV}}} \\ &= \frac{min\{N_{\text{CV and AV}_i},\mid i=1,\ \ldots,\ q\}}{N_{\text{CV}}} \\ &= \min\left\{\text{cr}\left(CV \to \text{AV}_1\right), \ldots, cr\left(CV \to \text{AV}_q\right)\right\} \geq \alpha \quad (6) \end{aligned}$$

   (b) the same AV, we establish p-to-l associations, $\{\text{CV}_p\} \to \{\text{AV}\}$, where $p > k$ and $\{\text{CV}_p\}$ is any combination of CVs from the associations in the group, due to the following:

$$\begin{aligned}\mathrm{cr}\left(\{\mathrm{CV}_p\} \rightarrow AV\right) &= \frac{N_{\{\mathrm{CV}_p\}\text{ and AV}}}{N_{\{CV_p\}}} \\ &\geq max\{cr\left(\mathrm{CV}_1 \rightarrow AV\right), \ldots, cr\left(\mathrm{CV}_p \rightarrow AV\right)\} \geq \alpha \quad (7)\end{aligned}$$

where $N_{\{\mathrm{CV}p\}} \leq N_{\mathrm{CV}i}$, $i = 1, \ldots, p$, and $N_{\{\mathrm{CV}p\}}$ and AV = min $\{N_{\mathrm{CV}i\text{ and AV}} \mid i = 1, \ldots, p\}$.

2) The associations in the group have the same CV, and their common subset of supporting instances satisfies the following condition:

$$\frac{N_{\mathrm{CommonSubset}}}{N_{\mathrm{CV}}} \geq \alpha \tag{8}$$

where the common subset contains instances which are in all sets of supporting instances for associations in the group, $N_{\mathrm{CommonSubset}}$ is the number of instances in the common subset, we establish k-to-q associations, $\{\mathrm{CV}\} \rightarrow \{\mathrm{AV}_q\}$, where $q > 1$ and $\{\mathrm{AV}_q\}$ is any combination of AVs from the associations in the group, due to the following:

$$\mathrm{cr}\left(CV \rightarrow \{\mathrm{AV}_q\}\right) = \frac{N_{\mathrm{CommonSubset}}}{N_{\mathrm{CV}}} \geq \alpha \tag{9}$$

For example, suppose that we have the group of the established 1-to-1 associations from Step 2.1, $x_1 = a_1 \rightarrow y_1 = b_1$ with the supporting set of instances {1, 2, 3, 4, 5} stored in Step 2.1.2 and $x_1 = a_1 \rightarrow y_2 = b_2$ with the supporting set of instances {1, 2, 3, 4} stored in Step 2.1.2, where 1, 2, 3, 4, and 5 are instance #. {1, 2, 3, 4} is an inclusive set to{1, 2, 3, 4, 5} as {1, 2, 3, 4, 5} includes {1, 2, 3, 4}, in other words, the inclusive set{1, 2, 3, 4} is a subset of {1, 2, 3, 4, 5}. We establish the 1-to-2 association, $x_1 = a_1 \rightarrow y_1 = b_1$, $y_2 = b_2$, using YFM1's part 1a).

In the method of YFM2 in Step 2.2, we establish p-to-1 associations from the candidate (p-1)-to-1 associations in the following steps:

1) Sort the (p-1)-to-1 associations by their CV

2) For each group of associations with the same CV

a) For each association t_i in the group, determine the minimum number of instances required for the common subset, m_i, as follows:

$$m_i = \lceil n_i \times \alpha \rceil \tag{10}$$

where n_i, is the total number of instances in the set of instances supporting the association t_i.

b) For every other association t_j in the group, whose AV is not same as AV in t_i, if $n_{\text{CommonSubset}}$ m_i, where $n_{\text{CommonSubset}}$ is the number of instances in the common subset of the two instance sets supporting t_i, and t_j, we establish a new p-to-1 association with CV = {CV and AV in t_i} and AV = AV in t_j, because this new association has the cr value $\geq \alpha$:

$$\text{cr}\left(\{CV\ and\ AV\ in\ t_i\} \rightarrow AV\ in\ t_j\right) = \frac{n_{\text{CommonSubset}}}{n_i} \geq \frac{m_i}{n_i} \geq \alpha \tag{11}$$

.

For example, suppose that we have $\alpha = 0.5$ and a group of two 2-to-1 associations with the same CV as follows:

$x_1 = a_1$, $x_2 = a_2 \rightarrow x_3 = a_3$, with the supporting set of instances {1, 2}, thus $n_1 = 2$

$x_1 = a_1$, $x_2 = a_2 \rightarrow x_4 = a_4$, with the supporting set of instances {1, 3}, thus $n_2 = 2$.

For the first association, t_1, we have: $n_1 = 2$,

$$m_1 = \lceil n_1 \times \alpha \rceil = \lceil 2 \times 0.5 \rceil = 1 \tag{12}$$

The second association, t_2, has AV which is not same as AV in the first association. The common subset of the two instance sets supporting the first and second associations is {1}. The number of instances in the common subset is $1 \geq m_1$. Thus we establish a new 3-to-1 association:

$x_1 = a_1$, $x_2 = a_2$, $x_3 = a_3 \rightarrow x_4 = a_4$.

In the procedure of using YFM1 and YFM2 in Step 2.2 as shown in Figure 2, at first we use YFM1 to establish 1-to-2, . . . , 1-to-M associations from the established 1-to-1 associations, where M is the total number of variables in the data set. Secondly, we use YFM2 to establish 2-to-1 associations from the candidate 1-to-1 associations, 3-to-1 associations from the candidate 2-to-1 associations, . . . , and M-to-1 associations from the candidate (M-1)-to-1 associations. At last, we use YFM1 to establish 1-to-2, . . . , 1-to-M associations from the established 1-to-1 associations, . . . , M-to-2, . . . , M-to-M associations from the established M-to-1 associations. However, this procedure of Step 2.2 can be cut short (that is, stopped earlier) if an application needs p-to-q associations up to the given values of p and q, where $p < M$ and $q < M$.

In Step 3 of the PVAD algorithm, partial-value associations from Step 2 are consolidated and generalized, and a multi-layer model of partial/full-value associations is constructed. Step 3 of the PVAD algorithm consists of the following steps.

- Step 3.1: consolidate and generalize the partial-value associations from Step 2 as follows. If we have one association, $x = a_1 \rightarrow y = b$, and another association, $x = a_2 \rightarrow y = b$, and
  - If a_1 and a_2 are two different but non-consecutive values of x, we replace $x = a_1 \rightarrow y = b$ and $x = a_2 \rightarrow y = b$ by a consolidated/generalized association, $x = a_1/a_2 \rightarrow y = b$, where the operator / of two terms represents either of two terms but not both terms;

  - If a_1 and a_2 are two consecutive values of x, we replace $x = a_1 \rightarrow y = b$ and $x = a_2 \rightarrow y = b$ by a consolidated/generalized association, $x = a \rightarrow y = b$, where a is a new categorical value of x including a_1 and a_2.

  If all values of a variable have the same association, this step will consolidate and generalize the same associations for various values of the variable into one association for all values of the variable. Hence, the PVAD algorithm can identify both partial-value and full-value associations of variables. We use, $x = *$, to represent all values of x.

- Step 3.2: use partial/full-value associations from Step 3.1 to construct an associative network which is a multi-layer model of partial/full-value associations from Step 3.1. A node is added to the associative network to represent each unique CV or AV of associations from Step 3.1. A directed link, called an associative link, is drawn in the associative network to represent each association.

- Step 3.3: remove a direct link between one CV node and one AV node if there are multiple paths from CV node to the AV node since the direct link can be derived from a path of associative links from this CV node to this AV node. For example, if we have two paths going to node $x_3 = a_3$: $x_1 = a_1 \rightarrow x_3 = a_3$ and $x_1 = a_1 \rightarrow x_2 = a_2 \rightarrow x_3 = a_3$, we remove the direct link $x_1 = a_1 \rightarrow x_3 = a_3$ because the direct link $x_1 = a_1 \rightarrow x_3 = a_3$ can be derived from $x_1 = a_1 \rightarrow x_2 = a_2 \rightarrow x_3 = a_3$.

There are three parameters, α, β, and γ, which need to have their value set when we apply the PVAD algorithm to a data set. The parameter α can be set to the value of 1 or some value close to 1 (e.g., 0.95) to start generating associations. Established associations can be examined to see if they are meaningful and useful to the application. If no meaningful and useful associations are found using a high value of α, the α can be lowered to run tbe PVAD algorithm on the data set again and generate associations. This process can be repeated with several values of α being used to generate association results until meaningful and useful associations are discovered.

The value of β depends on how many supporting instances are considered sufficient to establish an association for an application. One application may need a large number of supporting instances to establish an association, another application may needs just two supporting instances to establish an association.

The value of γ can be set for a data set in two steps. In the first step, the frequency of each value of each variable in the data set is computed and used to compute the percentage of each variable value in the data set. The percentage of a variable value is the ratio of its frequency in the data set to the total number of instances in the data set. For an application, there are usually some special values of certain variables which we are interested in and what to see associations involving those variables and their special values. In the second step, the γ value can be set to be greater than the frequency of those variable values of interest so that associations with those variable values will not be removed from established associations.

The application of the PVAD algorithm to anomaly detection in finance will enable learning from finance data to discover patterns of both normal transactions and frauds/anomalies in the form of both partial-value and full-value variable associations or associative networks involving interactive, concurrent effects of multiple transaction attributes. For example, transactions may be characterized using multiple variables such as:

- Transaction location, e.g., home town vs not home town,
- Expense/vendor type, e.g., gas/fuel,
- Dollar amount, and
- Transaction frequency.

The following is an example association of multiple transaction attributes for normal transactions:

$$\textit{expense type} = \textit{gas/fuel, dollar amount} < \textit{\$100} \rightarrow \textit{norm}$$

This association for normal transactions can be obtained by applying the PVAD algorithm to financial transaction data with a number of normal transaction records supporting these associations. An transaction record with the following attribute values:

expense type $=$ *gas/fuel, dollar amount* $\geq$ *$100*

can be detected as an anomaly since the multivariate combination of *expense type* $=$ *gas/fuel, dollar amount* $\geq$ *$100* is different from the combination values of the same two attributes in the association for normal transactions.

Hence, the association of multiple transaction attributes will enable capturing characteristics of normal transactions and frauds/anomalies more accurately than univariate characteristics (e.g., *dollar amount* $\geq$ *$100* in many cases for anomaly detection in finance to achieve the high level of detection accuracy.

5. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Number 1561496. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The author would like to thank Archana Anandakrishnan for time and help to edit this paper and the reviewers for helpful comments.

References

G. Box and A. Luceno. *Statistical Control by Monitoring and Feedback Adjustment.* New York, NY: John Wiley & Sons, 1997.

G. Box and J. Ramrez. *Cumulative Score Charts. Report No. 58, The Center for Quality and Productivity IMprovement.* Madison, Wisconsin: University of Wisconsin, 1991.

N. Ye. *Secure Computer and Network Systems: Modeling, Analysis and Design.* London, UK: John Wiley & Sons, 2008.

N. Ye. *Data Mining: Theories, Algorithms, and Examples.* Boca Raton, Florida: CRC Press, 2013.

N. Ye. A reverse engineering algorithm for mining a causal system model from system data. *International Journal of Production Research*, 55(3):828–844, 2017a.

N. Ye. The partial-valeu association discovery algorithm to learn multi-layer structural system models from system data. *IEEE Transactions on Systems, Many, and Cybernetics: Systems*, 47(12), 2017b.