DETECTANOMALY

The anomaly detection procedure searches for unusual cases based on deviations from the norms of their cluster groups. The procedure is designed to quickly detect unusual cases for data auditing purposes in the exploratory data analysis step, prior to any inferential data analysis. This algorithm is designed for generic anomaly detection; that is, the definition of an anomalous case is not specific to any particular application, such as detection of unusual payment patterns in the healthcare industry or money laundering detection in the finance industry in which the definition of an anomaly can be well defined.

Data Assumptions

Data. This procedure works with both continuous and categorical variables. Each row represents a distinct observation, and each column represents a distinct variable upon which the peer groups are based. A case identification variable can be available in the data file for marking output, but it will not be used in the analysis. Missing values are allowed. The SPSS weight variable, if specified, is ignored.

The detection model can be applied to a new test data file. The elements of the test data must be the same as the elements of the training data. And, depending on the algorithm settings, the missing value handling that is used to create the model may be applied to the test data file prior to scoring.

Case Order. Note that the solution may depend on the order of cases. To minimize order effects, randomly order the cases. To verify the stability of a given solution, you may want to obtain several different solutions with cases sorted in different random orders. In situations with extremely large file sizes, multiple runs can be performed, with a sample of cases sorted in different random orders.

Assumptions. The algorithm assumes that all variables are nonconstant and independent and assumes that no case has missing values for all the input variables. Further, each continuous variable is assumed to have a normal (Gaussian) distribution, and each categorical variable is assumed to have a multinomial distribution. Empirical internal testing indicates that the procedure is fairly robust to violations of both the assumption of independence and the distributional assumptions, but be aware of how well these assumptions are met.

Notation

The following notation is used throughout this chapter unless otherwise stated:

ID	The identity variable of each case in the data file.
n	The number of cases in the training data X_{train} .
X_{ok}, k = 1, …, K	The set of input variables in the training data.
M_k, k ∈ {1, …, K}	If X_{ok} is a continuous variable, M_k represents the grand mean, or average of the variable across the entire training data.

SD_k, $k \in \{1, \ldots, K\}$	If X_{ok} is a continuous variable, SD_k represents the grand standard deviation, or standard deviation of the variable across the entire training data.
X_{K+1}	A continuous variable created in the analysis. It represents the percentage of variables (k = 1, …, K) that have missing values in each case.
X_k, k = 1, …, K	The set of processed input variables after the missing value handling is applied. For more information, see “Modeling Stage ” on p. 3 .
H, or the boundaries of H: $[H_{min}, H_{max}]$	H is the pre-specified number of cluster groups to create. Alternatively, the bounds $[H_{min}, H_{max}]$ can be used to specify the minimum and maximum numbers of cluster groups.
n_h, h = 1, …, H	The number of cases in cluster h, h = 1, …, H, based on the training data.
p_h, h = 1, …, H	The proportion of cases in cluster h, h = 1, …, H, based on the training data. For each h, $p_h = n_h/n$.
M_{hk}, k = 1, …, K+1, h = 1, …, H	If X_k is a continuous variable, M_{hk} represents the cluster mean, or average of the variable in cluster h based on the training data. If X_k is a categorical variable, it represents the cluster mode, or most popular categorical value of the variable in cluster h based on the training data.
SD_{hk}, $k \in \{1, \ldots, K+1\}$, h = 1, …, H	If X_k is a continuous variable, SD_{hk} represents the cluster standard deviation, or standard deviation of the variable in cluster h based on the training data.
$\{n_{hkj}\}$, $k \in \{1, \ldots, K\}$, h = 1, …, H, j = 1, …, J_k	The frequency set $\{n_{hkj}\}$ is defined only when X_k is a categorical variable. If X_k has J_k categories, then n_{hkj} is the number of cases in cluster h that fall into category j.
m	An adjustment weight used to balance the influence between continuous and categorical variables. It is a positive value with a default of 6.
VDI_k, k = 1, …, K+1	The variable deviation index of a case is a measure of the deviation of variable value X_k from its cluster norm.
GDI	The group deviation index GDI of a case is the log-likelihood distance d(h, s), which is the sum of all the variable deviation indices $\{VDI_k, k = 1, \ldots, K+1\}$.
anomaly index	The anomaly index of a case is the ratio of the GDI to that of the average GDI for the cluster group that the case belongs.
variable contribution measure	The variable contribution measure of variable X_k for a case is the ratio of the VDI_k to the case’s corresponding GDI.
$pct_{anomaly}$ or $n_{anomaly}$	A pre-specified value $pct_{anomaly}$ determines the percentage of cases to be considered as anomalies. Alternatively a pre-specified positive integer value $n_{anomaly}$ determines the number of cases to be considered as anomalies.
$cutpoint_{anomaly}$	A pre-specified cut point; cases with anomaly index values greater than $cutpoint_{anomaly}$ are considered anomalous.
$k_{anomaly}$	A pre-specified integer threshold $1 \le k_{anomaly} \le K+1$ determines the number of variables considered as the reasons that the case is identified as an anomaly.

Algorithm Steps

This algorithm is divided into 3 stages:

Modeling. Cases are placed into cluster groups based on their similarities on a set of input variables. The clustering model used to determine the cluster group of a case and the sufficient statistics used to calculate the norms of the cluster groups are stored.

Scoring. The model is applied to each case to identify its cluster group and some indices are created for each case to measure the unusualness of the case with respect to its cluster group. All cases are sorted by the values of the anomaly indices. The top portion of the case list is identified as the set of anomalies.

Reasoning. For each anomalous case, the variables are sorted by its corresponding variable deviation indices. The top variables, their values and the corresponding norm values are presented as the reasons why a case is identified as an anomaly.

Modeling Stage

This stage performs the following tasks:

1. **Training Set Formation.** Starting with the specified variables and cases, remove any case with extremely large values (greater than 1.0E+150) on any continuous variable. If missing value handling is not in effect, also remove cases with a missing value on any variable. Remove variables with all constant nonmissing values or all missing values. The remaining cases and variables are used to create the anomaly detection model. Statistics output to pivot table by the procedure are based upon this training set, but variables saved to the dataset are computed for all cases.

2. **Missing Value Handling (Optional).** For each input variable X_{ok}, k = 1, …, K, if X_{ok} is a continuous variable, use all valid values of that variable to compute the grand mean M_k and grand standard deviation SD_k. Replace the missing values of the variable by its grand mean. If X_{ok} is a categorical variable, combine all missing values into a "missing value" category. This category is treated as a valid category. Denote the processed form of $\{X_{ok}\}$ by $\{X_k\}$.

3. **Creation of Missing Value Pct Variable (Optional).** A new continuous variable, X_{K+1}, is created that represents the percentage of variables (both continuous and categorical) with missing values in each case.

4. **Cluster Group Identification.** The processed input variables $\{X_k, k = 1, …, K+1\}$ are used to create a clustering model. The two-step clustering algorithm is used with noise handling turned on (see the TwoStep Cluster algorithm document for more information).

5. **Sufficient Statistics Storage.** The cluster model and the sufficient statistics for the variables by cluster are stored for the Scoring stage:
   - The grand mean M_k and standard deviation SD_k of each continuous variable are stored, $k \in \{1, …, K+1\}$.
   - For each cluster h = 1, …, H, store the size n_h. If X_k is a continuous variable, store the cluster mean M_{hk} and standard deviation SD_{hk} of the variable based on the cases in cluster h. If X_k is a categorical variable, store the frequency n_{hkj} of each category j of the variable based on the cases in cluster h. Also store the modal category M_{hk}. These sufficient statistics will be used in calculating the log-likelihood distance d(h, s) between a cluster h and a given case s.

Scoring Stage

This stage performs the following tasks on scoring (testing or training) data:

1. **New Valid Category Screening.** The scoring data should contain the input variables $\{X_{ok}, k = 1, …, K\}$ in the training data. Moreover, the format of the variables in the scoring data should be the same as those in the training data file during the Modeling Stage.

Cases in the scoring data are screened out that contain a categorical variable with a valid category that does not appear in the training data. For example, if *Region* is a categorical variable with categories IL, MA and CA in the training data, a case in the scoring data that has a valid category FL for *Region* will be excluded from the analysis.

2. **Missing Value Handling (Optional).** For each input variable X_{ok}, if X_{ok} is a continuous variable, use all valid values of that variable to compute the grand mean M_k and grand standard deviation SD_k. Replace the missing values of the variable by its grand mean. If X_{ok} is a categorical variable, combine all missing values and put together a missing value category. This category is treated as a valid category.

3. **Creation of Missing Value Pct Variable (Optional depending on Modeling Stage).** If X_{K+1} is created in the Modeling Stage, it is also computed for the scoring data.

4. **Assign Each Case to its Closest Non-noise Cluster.** The clustering model from the Modeling Stage is applied to the processed variables of the scoring data file to create a cluster ID for each case. Cases belonging to the noise cluster are reassigned to their closest non-noise cluster. See the TwoStep Cluster algorithm document for more information on the noise cluster.

5. **Calculate Variable Deviation Indices.** Given a case s, the closest cluster h is found. The variable deviation index VDI_k of variable X_k is defined as the contribution $d_k(h, s)$ of the variable to its log-likelihood distance d(h, s). The corresponding norm value is M_{hk}, which is the cluster sample mean of X_k if X_k is continuous, or the cluster mode of X_k if X_k is categorical.

6. **Calculate Group Deviation Index.** The group deviation index GDI of a case is the log-likelihood distance d(h, s), which is the sum of all the variable deviation indices $\{VDI_k, k = 1, \ldots, K+1\}$.

7. **Calculate Anomaly Index and Variable Contribution Measures.** Two additional indices are calculated that are easier to interpret than the group deviation index and the variable deviation index.

   The anomaly index of a case is an alternative to the GDI which is computed as the ratio of the case's GDI to the average GDI of the cluster to which the case belongs. Increasing values of this index correspond to greater deviations from the average, and indicate better anomaly candidates.

   A variable's variable contribution measure of a case is an alternative to the VDI which is computed as the ratio of the variable's VDI to the case's GDI. This is the proportional contribution of the variable to the deviation of the case. The larger the value of this measure, the greater the variable's contribution to the deviation.

Odd Situations

Zero Divided by Zero

The situation in which the GDI of a case is zero and the average GDI of the cluster that the case belongs to is also zero is possible if the cluster is a singleton or is made up of identical cases and the case in question is the same as the identical cases. Whether this case is considered as an anomaly or not depends upon whether the number of identical cases that make up the cluster is large or small. For example, suppose that there are a total of 10 cases in the training and 2 clusters are resulted in which one cluster is a singleton; that is, made up of 1 case, and the other has 9 cases. In this situation, the case in the singleton cluster should be considered as an anomaly as it

does not belong to the larger cluster. One way to calculate the anomaly index in this situation is to set it as the ratio of average cluster size to the size of the cluster h, which is:

$$\frac{n/H}{n_h}$$

Following the 10 cases example, the anomaly index for the case belonging to the singleton cluster would be (10/2)/1 = 5, which should be large enough for the algorithm to catch it as an anomaly. In this situation, the variable contribution measure is set to 1/(K+1), where (K+1) is the number of processed variables in the analysis.

Nonzero Divided by Zero

The situation in which the GDI of a case is nonzero but the average GDI of the cluster that the case belongs to is zero is possible if the corresponding cluster is a singleton or is made up of identical cases and the case in question is not the same as the identical cases. Suppose that case *i* belongs to cluster *h* which has zero average GDI; that is, average$(GDI)_h = 0$, but the GDI between case *i* and cluster *h* is nonzero, i.e., GDI(i, h) ≠ 0. One choice for the anomaly index calculation of case *i* could be to set the denominator as the weighted average GDI over all other clusters if this value is not zero, else set the calculation as the ratio of average cluster size to the size of the cluster *h*. That is,

$$\begin{cases} \dfrac{GDI(i,h)}{\frac{1}{(n-n_h)}\sum_{s=1,\neq h}^{H} n_s \cdot average(GDI)_s} & \text{if} \frac{1}{(n-n_h)}\sum_{s=1,\neq h}^{H} n_s \cdot average(GDI)_s \neq 0 \\ \dfrac{n/H}{n_h} & \text{else} \end{cases}$$

This situation triggers a warning that the case is assigned to a cluster that is made up of identical cases.

Reasoning Stage

Every case now has a group deviation index and anomaly index, and a set of variable deviation indices and variable contribution measures. The purpose of this stage is to rank the likely anomalous cases and provide the reasons to suspect them of being anomalous.

1. **Identify the Most Anomalous Cases.** Sort the cases in descending order on the values of the anomaly index. The top $pct_{anomaly}$ % (or alternatively the top $n_{anomaly}$) gives the anomaly list, subject to the restriction that cases with anomaly index less than or equal to $cutpoint_{anomaly}$ are not considered anomalous.

2. **Provide Reasons for Considering a Case Anomalous.** For each anomalous case, sort the variables by their corresponding VDI_k values in descending order. The top $k_{anomaly}$ variable names, its value (of the corresponding original variable X_{ok}), and the norm values are displayed as reasoning.

Key Formulas from Two-Step Clustering

The two-step clustering algorithm consists of: (a) a pre-cluster step that pre-clusters cases into many sub-clusters and (b) a cluster step that clusters the sub-clusters resulting from pre-cluster step into the desired number of clusters. It can also select the number of clusters automatically.

The formula for the log-likelihood distance d(j, s) between 2 clusters j and s is as follows:

$$d(j,s) = \xi_j + \xi_s - \xi_{<j,s>}$$

where

$$\xi_v = -N_v\left(\Sigma_{k=1}^{K^A}\log\left(\Delta_k + \hat{\sigma}_{vk}^2\right)/2 + \Sigma_{k=1}^{K^B}\hat{E}_{vk}\right)$$

and

$$\hat{E}_{vk} = -\Sigma_{l=1}^{L_k} N_{vkl}/N_v \log\left(N_{vkl}/N_v\right)$$

in which $\Delta_k > 0$ is a positive adjustment included in the formula to avoid the logarithm of zero in the calculation. Its value is set as:

$$\Delta_k = \frac{\hat{\sigma}_k^2}{m}$$

where m is user-specified and set to m = 6 by default, and $\hat{\sigma}_k^2$ is the sample variance of variable X_k over the entire training sample.

The log-likelihood distance can be computed as follows:

$$d(j,s) = \Sigma_{k=1}^{K^A+K^B} d_k(j,s)$$

where

$$d_k(j,s) = \begin{cases} \left\{-N_j \log\left(\Delta_k + \hat{\sigma}_{jk}^2\right) - N_s \log\left(\Delta_k + \hat{\sigma}_{sk}^2\right) + N_{<j,s>} \log\left(\Delta_k + \hat{\sigma}_{<j,s>k}^2\right)\right\}/2 \\ \left\{-N_j\hat{E}_{jk} - N_s\hat{E}_{sk} + N_{<j,s>}\hat{E}_{<j,s>k}\right\} \end{cases}$$

depending on whether the corresponding variable X_k is continuous or categorical.

See the TwoStep Cluster algorithm document for more information.