Anomaly Detection Library

Kapil Thakkar and Reshma Kumari

27th May, 2016

Contents

Chapter 1

Window Based Correlation

1.1 Introduction

This technique is basically applied on two time-series. Let's say we have two time series as series1 and series2. So, in this method, we first find correlation at various lags between these two time series. User can specify minimum and maximum lag to consider. So, for all those values, we find correlation values.

After finding correlation values at all lags, we consider that lag at which correlation value is higher among all previously calulated correlation values at all lags. Let's say that lag be "x". So, depending upon that "x", we shift series1 or series2. If "x" is positive, we move series2 by "x" units and if it is negative than we shift series1 by $|x|$ units.

Now, we are ready to apply window correlation. Take window value,"w" as input. First window will be from 1st element to w'th elementof both the time series after aligning by lag "x". Find correlation for this window between two time-series and save it in an array. Now, slide window by "w" elements and calculate correlation value again and so on. Now, we have correlation values at multiple windows.

Now, let's say both the series should have been positively correlated. So, what we do is, we choose threshold by MAD test if not provided to us, and find all correlation values which are below that threshold and report all those windows as anomaly.

1.2 Related Functions

1.2.1 correlation(arr1, arr2, maxlag, pos, neg)

This function calculates correlation between arr1 and arr2 at all possible lags between -maxlag to +maxlag, as specified by pos and neg parameters.

- Input Parameters
    1. arr1 *(list)* : Input series 1 as a list of float values
    2. arr2 *(list)* : Input series 2 as a list of float values
    3. maxlag *(int)* : maximum (maxlag) and minimum (-maxlag) lag to consider while calculating correlation betweem arr1 and arr2

4. pos *(int, 1 or 0)* : To consider positive lag or not, i.e. 1 to maxlag
5. neg *(int, 1 or 0)* : To consider negative lag or not, i.e. -maxlag to -1

- Output *(list)* :
  Returns list of tuples of the form

(lag, correlation value at this lag)

1.2.2 getMaxCorr(arar1,positive_correlation)

This function takes list of tuples of the form (lag, correlation value at this lag) as input. Returns lag value at which correlation value is maximum if positive_correlation is True, and returns lag at which correlation value is minimum if positive_correlation is False.

Basically, if both the series are positively correlated than we will be interested in maximum positive correlation or if both series are negatively correlated than we will be interested in minimum negative correlation, which is specified by positive_correlation parameter.

- Input Parameters

1. arr1 *(list)* : list of tuples of the form
   (lag, correlation value at this lag)
   i.e. correlation values at various lags
2. positive_correlation *(boolean, "True" or "False")* :
   - True: If value of this parameter is True than it will return lag at which correlation value if maximum (positive)
   - False: If value of this parameter is False than it will return lag at which correlation value if minimum (negative)

- Output *(Tuple)* :
  returns single tuple of the form (lag,correlation value at this lag), i.e. lag at which optimum correlation value is found along with correlation value.

1.2.3 correlationAtLag(series1, series2, lag, window_size)

This function fisrt aligns two series by given lag. If lag is positive than it shifts start of series2 else start of series1. After aligning both the series according to lag, this function calculates correlation between both series at all windows.

window_size states size of the window. So, we will start with first window taking first window_size elements from each series and will calculate correlation. We will save this correlation value in list and will slide to next window. Next window will start after window_size elements. In such a way, we calculate, correlation at all windows and return the list of correlation values.

- Input Parameters

1. series1 *(list)* : Input series 1 as a list of float values

2. series2 *(list)* : Input series 2 as a list of float values
3. lag *(int)* : lag at which series needs to be adjusted as explained above
4. window_size *(int)* : window size to be considered

- Output *(list)* :
  Returns list of correlation values (of float type) for all windows calculated at given lag

1.2.4 WindowCorrelationWithConstantLag(arr1, arr2, window_size,maxlag, positive_correlation, pos, neg)

This is sort of driver function, which will call above 3 functions. This function will first get lag at which series needs to be adjusted. Than using this lag, it will calculate correlation values at all windows and will return it.

- Input Parameters

  1. arr1 *(list)* : Input series 1 as a list of float values
  2. arr2 *(list)* : Input series 2 as a list of float values
  3. window_size *(int)* : window size to be considered while calculating window correlation
  4. maxlag *(int)* : maximum (maxlag) and minimum (-maxlag) lag to consider while calculating correlation betweem arr1 and arr2, to align both the series
  5. positive_correlation *(boolean, "True" or "False")* :
     - True: This suggest that both the series are positively correlated
     - False: This suggest that both the series are negatively correlated
  6. pos *(int, 1 or 0)* : If value of this parameter is True than we will consider positive values for lag, i.e. 1 to +maxlag to align both the series initially
  7. neg *(int, 1 or 0)* : If value of this parameter is True than we will consider negative values for lag, i.e. -maxlag to -1 to align both the series initially

- Output *(list)* :
  List of tuples of the form (lag,array) Where lag is lag value for which whole series is shifted and then at that lag, we have calculated correlation for all window. Correlation value for all windows is stored in array.

1.2.5 anomaliesFromWindowCorrelationWithConstantlag(arr1, arr2, window_size=15,maxlag=15, positive_correlation=True, pos=1, neg=1, default_threshold = True, threshold = 0):

This is main function of this method. This is driver of whole method. Using previously stated methods, it will first gather correlation values at different windows. Than depending upon which type of threshold is to be used, it will filter

out anomalies. If default threshold is to be used, than it will be caulated using MAD test on the correlation values at each window, else threshold provided by user will be used.

Correlation values not satisying threshold will be reported along with the date range of that window.

- Input Parameters

  1. arr1 *(list)* : Input series 1 as a list of tuples of the from (date,value)
  2. arr2 *(list)* : Input series 2 as a list of tuples of the from (date,value)
  3. window_size *(int)* : window size to be considered while calculating window correlation
  4. maxlag *(int)* : maximum (maxlag) and minimum (-maxlag) lag to consider while calculating correlation betweem arr1 and arr2, to align both the series
  5. positive_correlation *(boolean, "True" or "False")* :
     - True: This suggest that both the series are positively correlated
     - False: This suggest that both the series are negatively correlated
  6. pos *(int, 1 or 0)* : If value of this parameter is True than we will consider positive values for lag, i.e. 1 to +maxlag to align both the series initially
  7. neg *(int, 1 or 0)* : If value of this parameter is True than we will consider negative values for lag, i.e. -maxlag to -1 to align both the series initially
  8. default_threshold *(boolean, "True" or "False")* : whether to use default threshold or not. If True, default threshold will be used using MAD test on calculated correlation values for all windows.
  9. threshold *(float)* : if default_threshold is False, than this user provided threshold will be used.

- Output *(list)* :
  This function filter out anomalies and returns them. This function returns List of tuples of the form
  (start_date,end_date,correlation_value),
  where (start_date, end_date) specifies range of the window and correlation_value if value of correlation of that window

1.3 Description

Putting all together, here is the summary:

Function "WindowCorrelationWithConstantLag", first makes use of "correlation" function, to calculate correlation values at all lags to find out at which lag it needs to be align. Result of "correlation" function is passed to "getMaxCorr" function. Which will return lag at which optimum value of correlation is present. This output will be used by "correlationAtLag" function, to caluclate correlation at all windows after aligning both series by input lag. So, in this way

"WindowCorrelationWithConstantLag" combines these three functions and returns correlation value at each window.

Function "anomaliesFromWindowCorrelationWithConstantlag" is the main driver. This function calls "WindowCorrelationWithConstantLag" and gets the correlation values at all windows and filters out anomalies (either using threshold calculated by MAD test or by user provided threshold) and returns them in the format of (start_date,end_date,correlation_value), where (start_date, end_date) specifies range of the window and correlation_value if value of correlation of that window.

Chapter 2

Linear Regression

2.1 Introduction

This technique is applied on two time series where one is independent variable and other is dependent on independent variable. Let's say independent variable is "x" is represented by series1 and "y" is dependent variable which is represented by series2, where y=f(x).

So, in this method, given values of both variables at different points, i.e. given many pairs of (x,y), which are represented here by series1 and series2, this technique tries to find relation between x and y, i.e. it tries to find best suitable function y=f(x), which can best fit given data. Note that this function can only find linear relation between two variables, i.e. it can find relation such as y = mx + c, where "m" and "c" are some variables, which are found by this method, which can best represent these two series.

After finding that function, for a given value of "x" one can predict, what should be ideal value of "y". So, this technique basically works on this principle. So, after finding that function, we again apply same function of the given series of "x" and try to predict corresponding series of "y" and see the relative difference between actual "y" series and predicted "y" series. If this relative difference is too much high or too much low or both (depending upon what user needs), we return those values as anomalies. To decide, whether value is too high or too low, we set up threshold. This threshold can be given by user or can be set automatically by using MAD test on the series generated by taking relative difference. Values beyond this threshold are reported as anomalies.

2.2 Related Functions

2.2.1 linear_regression(x_series, y_series, param = 0, default_threshold = True, threshold = 0)

This function takes two time series, x_series and y_series, as input, where x_series is series corresponding to "x" variable (independent variable) and y_series is series corresponding to "y" variable (dependent variable, dependent on "x"). Given these two series, it first finds out best linear relationship between these two variables and as described in the above section, it finds relative difference

between predicted and actual "y" series and the ones which are beyond threshold value are reported as anomaly.

As described above, threshold value may be calculated by MAD test on relative difference values by keeping "default_threshold" as "True", and it it is false, user will provide threshold value, by setting up "threshold" parameter above.

Note that i'th value in y_series should be corresponding to i'th value in the x_series.

- Input Parameters

  1. x_series *(list)* : List of float values representing "x" variable (independent variable)
  2. y_series *(list)* : List of float values representing "y" variable (independent variable)
  3. param *(int, 1 or 0 or -1)* :
     Defines what to be treated as anomaly depending on its value as follows: 0: Values going out of range, both with positive and negative error 1: Values with positive errors -1: Values with negative errors (Treat here error as relative difference crossing threshold value, positive error is relative difference which is positive and crossing positive threshold value and vice-versa).
  4. default_threshold *(boolean, True or False)* : If this is set as "True", than threshold will be calculated using MAD test, if False, than user given threshold value will be used.
  5. threshold *(float)* : Here, user can provide threshold value if, default_threshold is False. maxlag

- Output *(Tuple)* :
  returns Following tuple: (result,regression_object)

  Where, "result" is list of tuples which are anomaly according to linear regression test of following format:

  (Index_of_Data_Point,x_value,y_value,predicted_y_value,difference_between_predicted_and_actual_y_valu

  "regression_object" is an object of linear regression test, which represents y=f(x) = mx + c, which can be used to regenerate predicted values for plotting graphs afterwards or for some other task.

  Format of using: regression_object.predict(x_value), where x_value is just one value, for which we need corresponding ideal "y" value,

2.2.2 anomalies_from_linear_regression(result_of_lr, any_series)

This function basically takes result of "linear_regression" as input along with any series which is list of tuples of the form (date, value), and gives date to each anomaly.

The result returned by "linear_regression" function just provides index of data point, which is reported as anomaly. But we have time series, so we need to provide date, instead of index of data point. So, this function basically, attaches each anomaly with it s date and returns it.

- Input Parameters

  1. result_of_lr *(list)* : This is list of anomalies reported by "linear_regression" function. Note that here we are just passing list of anomalies only and not the regression object, i.e. we are passing just first element of tuple returned by "linear_regression" function.
  2. any_series *(list)* : Any list/series (x_series or y_series) of tuples in the format (Date,Value), date will be used from this series to attach each anomaly with its corresponding date.

- Output *(list)* :
  Returns list of tuples of the following form:
  (date,x_value,y_value,predicted_y_value,difference_between_predicted_and_actual_y_value)

2.2.3 linear_regressionMain(x_series, y_series, param = 0, default_threshold = True, threshold = 0)

This is main function of this anomaly detection technique. This function first calls "linear_regression" function, gets list of anomalies. After that, it calls "anomalies_from_linear_regression" function to attach date with each anomaly and than returns result.

- Input Parameters

  1. x_series *(list)* : List of tuples of the format (date,value) representing "x" variable (independent variable)
  2. y_series *(list)* : List of tuples of the format (date,value) representing "y" variable (independent variable)
  3. param *(int, 1 or 0 or -1)* :
     Defines what to be treated as anomaly depending on its value as follows: 0: Values going out of range, both with positive and negative error 1: Values with positive errors -1: Values with negative errors (Treat here error as relative difference crossing threshold value, positive error is relative difference which is positive and crossing positive threshold value and vice-versa).
  4. default_threshold *(boolean, True or False)* : If this is set as "True", than threshold will be calculated using MAD test, if False, than user given threshold value will be used.
  5. threshold *(float)* : Here, user can provide threshold value if, default_threshold is False. maxlag

- Output *(list)* :
  Returns list of tuples of the form
  (start_date,end_date,difference_between_predicted_and_actual_y_value)

Note that here, start_date is equal to end_date, as we are working day-wise in this technique, instead of any window.

2.3 Description

Putting all together, here is the summary:

"linear_regressionMain" is the main function of this technique, which calls 2 other functions and returns result. First it calls, "linear_regression" function, gets list of anomalies. After that, it calls "anomalies_from_linear_regression" function to attach date with each anomaly and than returns result.

Chapter 3

Graph Based Anomaly Detection Technique

3.1 Introduction

This technique was introduced by [1]. We have used is R implementation given by authors of this book [2]. So, here by using python script, we will be just calling R script with appropriate arguments and will be using result provided by that script.

Graph based anomaly detection technique considers each day as a node of a graph. Similar nodes are connected to each other by some weight. Similarity of nodes are calculated by making use of the values of that node i.e. value(s) of timeseries on that date. Based on this similarity, edge weights are also assigned. Than random walk algorithm is applied on this graph structure and connectivity value of each node is calculated. Graph nodes having the least connectivity values are reported as anomaly.

Note that previous techniques, like Window Correlation, Slope Based and Linear Regression techniques, can take only 2 time series as input. They also don't consider historical values, trend or seasonality. It just makes prediction on the given present data. Where as, this Graph based anomaly detection technique, can take multiple time series as input and also considers trends, seasonality as well, as explained in research paper [1].

So, here, we take multiple time series as input. Out of them, one will be dependent on rest of the others. We will call R script, it will print result in one csv file. We read that CSV file and return result. Note that here we do not have threshold value. We just give number of points with the least connectivity value and function returns them. If in future, one wants to add threshold value on connectivity than function can be modified according to that as well.

3.2 Related Functions

3.2.1 graphBasedAnomalyCall(dependentVar, numberOfVals, timeSeriesFileNames)

This function calls the R Script "graphBasedAnomaly.R". This function takes multiple time series as input, which are stored in files, whose name are stored in "timeSeriesFileNames" list. This time-series files are generated by us only. Out of these time series, one will be for dependent variable and others will be corresponding to independent variable. So variable, "dependentVar" represents which time series/variable is dependent.

this function executes R script and writes output to the file named "GraphBasedAnomalyOp.csv".

- Input Parameters
  1. dependentVar *(int)* : Index of the dependent variable, where dependentVar = function of independantVars
  2. numberOfVals *(int)* : Each CSV contains how many values? That is each time series has how many values?
  3. timeSeriesFileNames *(list)* : Name of the files to which series is stored. File should contain only series values.
- Output: This function does not generate any output. R Script will write output to CSV file as stated before.

3.2.2 generateCSVsForGraphBasedAnomaly(lists, dateIndex, seriesIndex)

In python code, we have time-series as a list. This list is list of tuples, in which first value of tuple is date and than we have more than one values in the same tuple, representing different time-series. For example, if we have testcase as onion, than for one city we have 3 time series along with date, which is rpresented as list of tuples of the the form (date, arrival, wholesale price, retail price). But, for R script, we need just time series values. So this function will take series of time series in variable "lists", where lists[i] will represent one timeseries or multiple time series for one object (like explained previously we can have multiple time series for one city).

dateIndex will say which tuple number for the list lists[i] represents date and seriesIndex represents, if lists[i] represents multiple series than which one to take out of them. This can be explained by example as follows:

Let's say, we have lists as folows:

```
[
[(1-1-2010, x1, y1, z1), (2-1-2010, x2, y2, z2), ... ],
[(1-1-2010, x1, y1, z1), (2-1-2010, x2, y2, z2), ... ],
[...], ...
];
```

So, here we have time-series corresponding to two entities, which can be accessed via lists[0] and lists[1]. Now, lists[0] gives us 3 time-series for one entity. But let's say, here we need only one corresponding to "y" time series. So, give dateIndex as 0 here and seriesIndex as 2. So, this function will create 2 CSVs, one for each entity. Each CSV will have values [y1, y2, y3, ...]. One line will contain one value in file.

Note that it is not necessary to have multiple time-series for on entity. We can have just simple structure as follows:

```
[
[(1-1-2010, x1), (2-1-2010, x2), ... ],
[(1-1-2010, y1), (2-1-2010, y2), ... ],
[...], ...
];
```

So here, we have two time-series as x and y, and we can than give dateIndex as 0 here and seriesIndex as 1. This will create two CSVs, one for "x" and other for "y".

After creating these CSVs, this function returns names of the file created.

- Input Parameters

  1. lists *(list)* : List of time-series, where lists[i] = list of tuple of the form (date, val1 [, val2, val3, ...])
     where date is in form of string and values in square brackets are optional.
  2. dateIndex *(int)* : column number of date in list of tuple (starting with 0)
  3. seriesIndex *(int)* : column number of series in list of tuple (starting with 0)

- Output *(Tuple)*:
  returns tuple of the form, (dates,fileNames),

  Where,
  fileNames: Generated multiple CSVs, corresponding to each series for the input of R script. Returns name of these files.
  dates: Separated date from the series, so that later we can combine result of the R script (anomalies) with dates.

3.2.3 getAnomalies(dates,resultFile, numOfPtsReqd)

This function does the work of combining result of R script with the date. Result generated by R will be in some file, which is passed here as resultFile parameter. This will have indices for each day. So using this we append dates to it. So now, we have connectivity value for each date. This function sorts them according to connectivity value and returns the number of points required stated by parameter numOfPtsReqd, which as low connectivity value.

- Input Parameters

1. dates *(list)* : List of dates, returned by "getAnomalies" function.
2. resultFile *(string)* : Path of file to which output of R script is written
3. numOfPtsReqd *(int)* : Number of anomalous points required

- Output *(list)*:
  reurns list of tuples of the form:
  (start_date, end_date, connectivity_value)
  Note that, here start_date will be same as end_date, as this function returns results day-wise.

3.2.4 graphBasedAnomalyMain(lists, dependentVar, numOfPtsReqd, dateIndex=0, seriesIndex=1)

This is the main function of this method. This function makes call to other functions, uses the output of one, as a input to other, combines all functions and returns generated output.

- Input Parameters

1. lists *(list)* : List of time-series, where lists[i] = list of tuple of the form (date, val1 [, val2, val3, ...])
   where date is in form of string and values in square brackets are optional.
2. dependentVar *(int)* : Index of the dependent variable, where dependentVar = function of independantVars
3. numOfPtsReqd *(int)* : Number of anomalous points required
4. dateIndex *(int)* : column number of date in list of tuple (starting with 0)
5. seriesIndex *(int)* : column number of series in list of tuple (starting with 0)

3.3 Description

Putting all together, here is the summary:

"graphBasedAnomalyMain" is the main function. First, it calls "generateCSVsForGraphBasedAnomaly", which will generate files for each series, which will be used as input for R script. It also generates list of dates. Now, this list of file is passed to function, "graphBasedAnomalyCall", which will execute R script and will generate output in predefined file. This file name, along with dates and number of anomaly points required is passed to function "getAnomalies", which will return output in required format.