CONSTRAINT DETECTION AND OUTLIER EXPLANATION

Terrible Trio

Deepti Chavan(deeptisu)| Shruti Parab(shrutide) |Sushmita Sinha(ssinha7)

Table of Contents

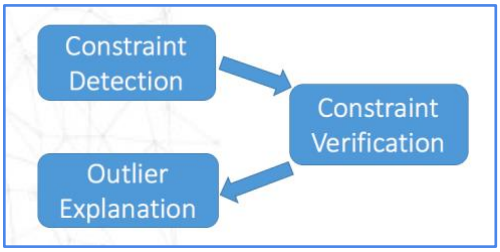
MOTIVATION

Ensuring the integrity of the database requires detection of the presence of outliers. A data row is considered to be an outlier if it violates any constraints which the database holds. These constraints are usually defined by the domain experts. However, some crucial constraints might get neglected due to the possibility of human error.

This project aims to automate the process of finding constraints over the database. Moreover, such an automated analysis can help define trends, make predictions and uncover root causes for behavior of the data.

INTRODUCTION

The high level aim is to automate the process of finding constraints in the dataset and to optimize the same. Using these constraints, find possible reasons identifying and justifying the presence of an outlier. To determine the statistically significant relationship we are using Least Dispersion Method to find constant pattern over categorical and Linear Regression for increasing and decreasing pattern over variable data.



DEFINITION

PATTERN

- Pattern (fixed, variable, aggregate_value, model, metric)
- For each group formed by fixed attribute, a pattern is determined by fitting a model on variable attributes versus aggregate values
- Metric defines the goodness of fit measure for the model

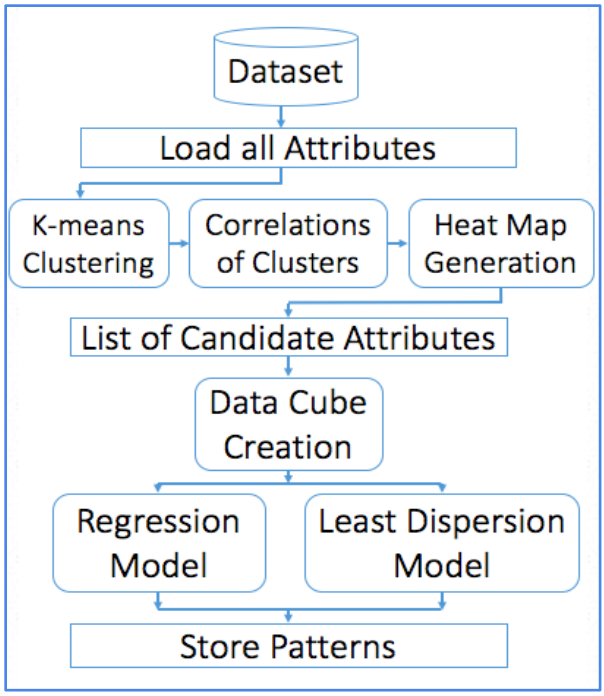
TYPE OF PATTERNS

- **Local Patterns:** Patterns discovered for the groups formed by fixed attributes
- **Global Patterns:** A summary of how many percentage of groups hold valid patterns. (Greater than 75% concludes the existence of global pattern)

MODEL

- **Linear Regression Model (LRM):**
  Fit a regression line for which the goodness of fit is determined by the score (R^2 coefficient of determination. Current threshold for valid patterns is 0.75).
  The slope determines whether the discovered pattern of aggregate values is increasing (positive slope)/ decreasing (negative slope) w.r.t. the variable attribute values.
- **Least Dispersion Model (LDM):**
  Least dispersion model finds the percentage of standard deviation w.r.t. mean for categorical data. (current threshold of 10% for valid pattern)

APPROACH



K-MEANS CLUSTERING

Using the columns that have been given by the chi-squared test, k-means clustering is done. We aim to find the correlation of the different attributes based on the representative mean of every cluster and identifying the relationship based on the heat map-based projection of the result.

DATA CUBE FORMATION

The data cube is built using GROUPING SETS and CUBE in postgres. These operations help to build the whole cube, with all possible combinations of the dimension attributes, in one single scan over the database. Once, the cube is built and stored in the database (in the form of a new table), all the results required for further analysis are picked from this cube.

LINEAR REGRESSION MODEL

This is basically, trying to fit a regression line over the given data points of variable versus aggregate values. The goodness of fit measure is the score (R^2 measure) used to determine if the regression line discovered is a significant fit. The slope of the line determines if it is an increasing or decreasing pattern.

LEAST DISPERSION MODEL

This model is used when we cannot fit a regression line over the data points. For example, we cannot fit a regression line if the variable attribute is categorical. In this case, we use a Least Dispersion Model. This model basically measures the percentage of deviation from the mean value. The inference from this model would be that for a given attribute value the value attribute is somewhat constant (since the deviation is less from the mean value)

OPTIMIZATION

CORRELATION ANALYSIS

Prune unwanted attributes which would not hold any valid pattern. This is done by finding correlation coefficient between pairs of attributes threshold to a value of 0.7. Further, explore patterns in this reduced space of attributes.

CLUSTERING

Optimize the process of finding correlation coefficient. The dataset contains millions of rows. Clustering captures the overall trend of the data points into the reduced representative means of all the clusters. Thus, finding correlation over these clustered points gives a better coefficient value and avoids the problem of pruning the significant attributes, which is highly probable in case of correlation analysis over original points.

DATA CUBE

- Fitting models over possible groups of attributes involves repeated computation of aggregates. These aggregate values are pre-computed in the form of a data cube and stored in the database.
- A data-cube is basically a model which stores the aggregates while building the model itself. We use a slice of the data-cube to compute the aggregates.

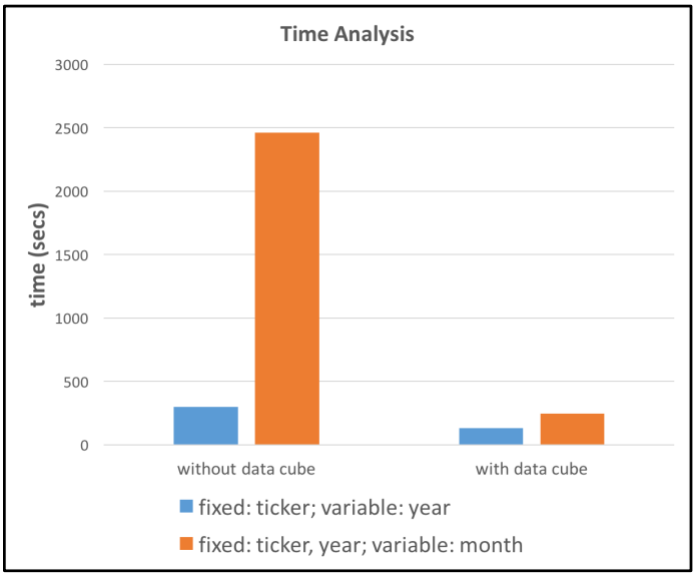
RESULTS

Following are the results for the Stock Dataset.

- Dataset: *Stock*
- Columns: *['ticker', 'date', 'open', 'high', 'low', 'volume' ,'close','ex_dividend', 'split_ratio', 'adj_open','adj_high', 'adj_low', 'adj_close', 'adj_volume']*
- Number of rows: *1877056*
- Number of value attributes = *12*
- Number of value attributes after clustering = *10*

Following is the time analysis to get the patterns in the dataset:

- For fixed attribute = 'ticker' and variable attribute = 'year', number of patterns found = 19614
- For fixed attribute = 'ticker, year' and variable attribute = month, number of patterns found = 28621



Patterns stored in the database:

id	fixed	fixedvalue	variable	aggfunction	aggvalue	pattern	metric
1123	{ticker,year}	:FB:2015	{month}	avg	open	increasing	0.90322419825370592
1124	{ticker,year}	:FB:2017	{month}	avg	open	increasing	0.98506548939663841
4600	{ticker,year}	:FB:2015	{month}	avg	high	increasing	0.90694812862416707
4601	{ticker,year}	:FB:2017	{month}	avg	high	increasing	0.98649354069204942
8066	{ticker,year}	:FB:2015	{month}	avg	low	increasing	0.90635079786186989
8067	{ticker,year}	:FB:2017	{month}	avg	low	increasing	0.98151231702915431
11554	{ticker,year}	:FB:2015	{month}	avg	volume	increasing	0.90914225368900081
11555	{ticker,year}	:FB:2017	{month}	avg	volume	increasing	0.98467031911801495
15446	{ticker,year}	:FB:2015	{month}	avg	adj_open	increasing	0.90322419825370592
15447	{ticker,year}	:FB:2017	{month}	avg	adj_open	increasing	0.98506548939663841
18996	{ticker,year}	:FB:2015	{month}	avg	adj_high	increasing	0.90694812862416707
18997	{ticker,year}	:FB:2017	{month}	avg	adj_high	increasing	0.98649354069204942
22539	{ticker,year}	:FB:2015	{month}	avg	adj_low	increasing	0.90635079786186989
22540	{ticker,year}	:FB:2017	{month}	avg	adj_low	increasing	0.98151231702915431
26112	{ticker,year}	:FB:2015	{month}	avg	adj_close	increasing	0.90914225368900081
26113	{ticker,year}	:FB:2017	{month}	avg	adj_close	increasing	0.98467031911801495

id	fixed	fixedvalue	variable	aggfunction	aggvalue	pattern	metric
609	{ticker}	FB	{year}	avg	open	increasing	0.99970841996671544
2451	{ticker}	FB	{year}	avg	high	increasing	0.99975290490964619
4286	{ticker}	FB	{year}	avg	low	increasing	0.99951940818397078
6123	{ticker}	FB	{year}	avg	volume	increasing	0.99959284736954679
7897	{ticker}	FB	{year}	avg	close	decreasing	0.85695724799976047
10899	{ticker}	FB	{year}	avg	adj_open	increasing	0.99970841996671544
12842	{ticker}	FB	{year}	avg	adj_high	increasing	0.99975290490964619
14777	{ticker}	FB	{year}	avg	adj_low	increasing	0.99951940818397078
16718	{ticker}	FB	{year}	avg	adj_close	increasing	0.99959284736954679
18545	{ticker}	FB	{year}	avg	adj_volume	decreasing	0.85695724799976047

CONCLUSION

Automated analysis of finding constraints over the database results in discovery of trends in data which are otherwise difficult to capture using manual analysis. Some of the key observation from the above project:

- Building data cube upfront with pre-computed aggregates helps to achieve at least **2X** speedup or more.
- Clustering the data helps to prune the attributes with less probability of dropping the significant ones.

The constraints discovered are then used to provide explanations to the possible outliers which the user determines. The explanation of an outlier is given in the form of top k constraints which are determined using a similarity score.