CONSTRAINT DETECTION AND OUTLIER EXPLANATION

Terrible Trio

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

TABLE OF CONTENTS

PROBLEM

The high level aim is to find the trends and set constraints governing the dataset and try to optimize the process of finding correlations in the data. Using these constraints to find possible reasons justifying or identifying the presence of an outlier.

HIGH LEVEL APPROACH

FIND THE TRENDS AND DEFINE CONSTRAINTS

Functional dependency is a constraint between two sets of attributes in a relation from a database; hence defining the relationship between attributes. It governs how the data should look like, hence becomes an integral part in data cleaning, integration and repair. We can understand the different outliers present in the data and find which constraints did it violate. A better domain knowledge can help us identify the possible reasons for the outlier behavior or if it was just a data error.

We intend to start with a brute force approach that will try to find the correlation coefficient between different attributes. Optimizing this process is crucial and will work on using different techniques like Pearson's chi squared test, identifying the dimensional columns, random sampling instead of aggregation over the entire dataset etc.

FIND AND EXPLAIN THE OUTLIER

The constraints act as a benchmark for ideal data and once the constraints are defined we will progress towards identifying the data points that do not satisfy the constraints and mark them as outliers. Statistical distortion gives us the measure of how much the data is deviated from ideal data and helps us identify the outliers. The presence of outliers give rise to the problem of explanation finding an explanation of their presence. It is important to know where did the outlier come from and why does it not fit your model. The point can be in violation of multiple constraints and hence narrowing down to which constraint was most violated. Outliers can contain important information, identifying special events, hence outlier analysis is essential.

DATASET

We are using stock dataset comprising of 3090 ticker symbols, distributed daily over period of 3 years. The columnar values comprise of fields like ticker, date, open, high, low, close, volume, ex-dividend, split_ratio, adj_open, adj_high, adj_low, adj_close, adj_volume.

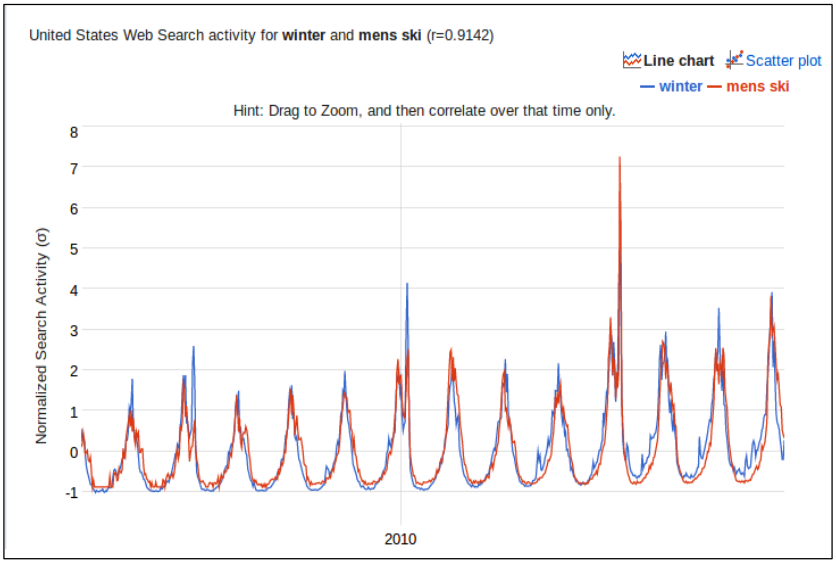
RELATED WORK

There have been different studies in the area of finding correlations in the database. Here are some of the examples:

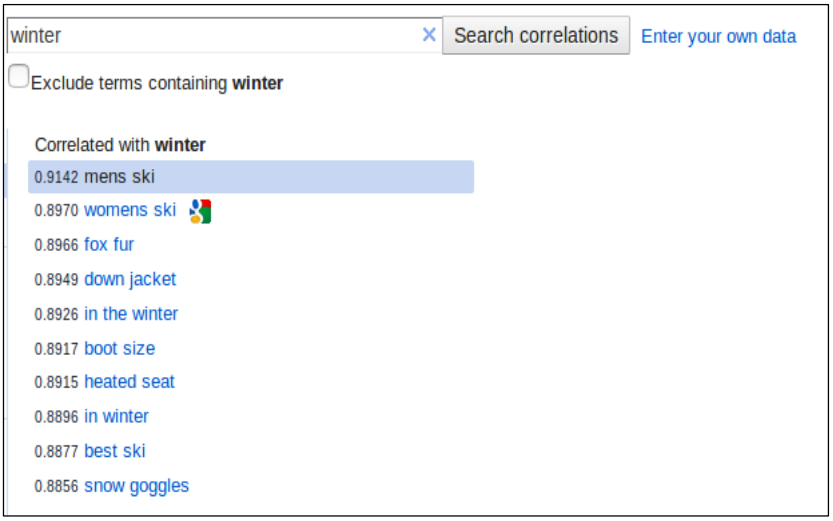
GOOGLE CORRELATE

Google Correlate basically finds out correlations between queries. It calculates the Pearson's correlation coefficient and returns those queries which have a high value for this coefficient indicating a strong correlation.

For example, if I search for the query "winter" you can see that google returns "mens ski" as the most closely correlated query. The graph looks something like this where you can see the pattern of both the queries is quite similar.



It also returns the top correlations along with their Pearson's correlation coefficients.



TOP-K Φ CORRELATION COMPUTATION

- This paper presents an algorithm to find out top-k correlated pairs in the database. The strategy described in the paper focuses on market basket datasets containing binary data.
- This paper also defines ϕ correlation coefficient which is a form of Pearson's correlation coefficient for binary data. The paper shows a comparison between ϕ correlation coefficient and chi square statistics and reasons why the ϕ correlation coefficient is better.
- It further describes an algorithm which avoids the brute force approach of finding all correlated pairs and selecting top k out of them. Instead, it uses pruning to find out the most correlated top *k* pairs. https://www.researchgate.net/publication/220668966_Top-k_Correlation_Computation

CORDS: AUTOMATIC DISCOVERY OF CORRELATIONS AND SOFT FUNCTIONAL DEPENDENCIES

This paper aims at finding correlations and soft functional dependencies in the datasets. It starts with finding out candidate columns which can be potentially correlated. It also makes use of sampling to find a subset of the dataset. Further, it uses chi-squared test and a threshold to find the correlations in the dataset.
https://cs.uwaterloo.ca/~ilyas/papers/cords.pdf

DATA QUALITY ON TEMPORAL DATA

This paper describes how to find the outliers and glitches in the data streams from a given set of constraints. This can prove as an application of the constraints derived from the data. The paper starts off with explaining how to find glitches in two correlated data streams. Further, it illustrates the study with the example of NYSE stock data.
http://web2.research.att.com/export/sites/att_labs/techdocs/TD_101818.pdf

HOW OUR APPROACH IS DIFFERENT THAN THE EXISTING WORK

The above mentioned strategies make use of correlation coefficients to rank the relations between two attributes. However, it only gives the information that the two attributes are related; it does not say much on how they are related.

Our objective is basically to find out interesting patterns from the database. Examples include relations like “total volume of stocks may vary on day to day basis but is consistent over a month” ; “the average difference between the starting and the ending price for the stock of a company is consistent over a month”.

PERFORMANCE METRICS

The performance of our project will be measured based on the accuracy and complexity in finding out the trends/constraints and explaining why outliers exist.

ACCURACY

There exist tests for correlation coefficients, viz. Pearson’s test, Chi-squared test, etc. that give us the measure of strength and direction of the linear relationship between two variables. We will cross-verify our results against these types of tests.

Another approach in testing the accuracy of our outcomes include corroborating them with known results. A synthetic dataset, where all the correlations are known to us beforehand can be used to validate the results returned by our method.

COMPLEXITY

We intend to find an improved methodology compared to the traditional Brute-Force approach in finding out the constraints.

Techniques like identifying the redundant work and storing the intermediate results, working on candidate columns instead of all the attributes, will help to reduce the time complexity.