AN ONLINE ADAPTIVE CLASSIFICATION OF GOOGLE TRENDS DATA ANOMALIES FOR INVESTOR SENTIMENT ANALYSIS

Duygu Dere¹, Mert Ergeneci^{2*}, Kaan Gokcesu³

¹Project Group International, Ankara, Turkey

²Bilkent University Nanotechnology Research Center (Nanotam), Ankara, Turkey

³Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

*Corresponding Author Email: mertergeneci@bilkent.edu.tr

This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ARTICLE DETAILS

Article History:

Received 12 March 2018

Accepted 12 April 2018

Available online 15 May 2018

ABSTRACT

Google Trends data has gained increasing popularity in the applications of behavioral finance, decision science and risk management. Because of Google's wide range of use, the Trends statistics provide significant information about the investor sentiment and intention, which can be used as decisive factors for corporate and risk management fields. However, an anomaly, a significant increase or decrease, in a certain query cannot be detected by the state of the art applications of computation due to the random baseline noise of the Trends dataset. Since through time, the baseline noise power shows a gradual change an adaptive threshold method is required to track and learn the baseline noise for a correct classification. To this end, we introduce an online method to classify meaningful deviations in Google Trends data. Through extensive experiments, we demonstrate that our method can successfully classify various anomalies for plenty of different data.

KEYWORDS

Adaptive Data Processing, Behavioral Finance, Convex Optimization, Online Learning.

1. INTRODUCTION

The use of Internet has increased largely among people. Individuals search everything that they seek to learn immediately via the use of Google and similar web-related tools due to their limited time and concentration constraint [1]. Through the statistical information of people's search density that is accumulated in such web engines, it is possible to make certain financial analysis, which is mostly related to the investor sentiment analysis [2]. The changes in the search queries reflect the emotions and trends among the vast range of people [3]. Thus, such information can be used for further corporate and risk management analysis [4]. However, the state of the art approaches cannot detect the significant changes in the Google Trends data flow due the instability of the baseline offset of the data [2]. In order to overcome this problem, we present an online adaptive threshold mathematical method, which sequentially learns the data

according to a classifier and automatically makes a decision real-time.

The remainder of the paper is organized as follows. In Section 2, we define the method that solves the anomaly detection problem. Next, in Section 3, the application of the proposed method over several Google Trends datasets is shown. Finally, some concluding remarks are given in Section 4.

2. ANOMALY DETECTION METHOD

In this section, the adaptive anomaly detection algorithm is detailed in three subsections, which are classifier, threshold and decision parts. Prior to the classifier subsection, it is useful to identify the data types and variable names.

Figure 1: Block diagram for the anomaly detection method steps

Suppose $X[n]$ is the Google Trends data of size N , then each index of $X[n]$ function gives the corresponding magnitude of the query rate out of 100.

2.1 Binary Classifier

In this subsection, we use a mathematical transformation to classify anomalies from normal data. To do that, we emphasize the anomalies and suppress relatively small changes, the discrete derivative of the data, $X[n]$, is calculated. The discrete derivative series of the X , i.e., D , is estimated in order to track the amount of changes for each index. To acquire the D

series, the X is subtracted from its one index shifted version as the following:

$$D[n-1] = X[n] - X[n-1], \quad \forall n \rightarrow N, \quad (1)$$

where $X[n-1]$ is the one index shifted version of $X[n]$. Thus, D gives the slope, change, between each X bin. After acquiring the instant slope for the discrete frequency bins, we group the D values into windows of size T and take its root mean square (RMS) as the following:

$$Y[n] = \sqrt{\sum_{t=0}^{T-1} \frac{D[t+n]^2}{T}}, \quad (2)$$

As seen in equation (2), grouping and calculating the RMS value of the discrete derivative series approximate the data to a binary classifier via amplifying the differences between small and fast changes. However, increasing the window size T has a significant tradeoff, since it causes a loss of resolution, i.e., detail of time values as week, month, and year. To this end, in order to benefit from both of the advantages, we set T as 3.

2.2 Adaptive Threshold

After obtaining a classifier metric, a threshold is applied to detect the anomalies. Instead of using a fixed value, we use an adaptive update to minimize decision errors due to long-term slow changes in data magnitude. Thus, the threshold value is updated in an online manner via the use of Online Gradient Descent (OGD) algorithm. The loss function for the OGD model is selected as the squared-loss against the binary classifier results. The threshold calculation with the selected specifications becomes as the following:

$$\tau[n+1] = \tau[n] + \frac{(Y[n] - \tau[n])}{n^\alpha}, \quad (3)$$

where $\tau[n]$ is the current threshold value, $Y[n]$ is the classifier metric for the n^{th} index and α is the learning rate. The resulting value of this equation, $\tau[n+1]$ is the estimated future threshold value. The calculation of the classifier metric, $Y[n]$, is described in Section 2.1. In order to create a soft maximum evaluation for better threshold offset, the positive exponential transformation of $Y[n]$ value is taken prior to $\tau[n+1]$ estimation. Then the update algorithm can be written as the following:

$$\tau[n+1] = \tau[n] + \frac{(e^{Y[n]} - \tau[n])}{n^\alpha}, \quad (4)$$

2.3 Decision

In the last subsection for the anomaly detection method, a decision is realized according to the comparison of calculated binary classifier metric and threshold value for each corresponding data frame. Since in the Equation 4, a positive exponential transformation is applied to the binary classifier metric $Y[n]$, the inverse of this exponential, namely $\ln(\cdot)$ conversion should be made before a decision. If the classifier $Y[n]$ is larger than the $\ln(\tau[n])$, then it can be said that an anomaly situation is present for the n^{th} index of the certain Google Trends data. In the opposite situation, where the threshold is higher than the classifier, the index value is considered as a normal condition.

3. EXPERIMENTS

In this section, the application of the proposed method on two different datasets, which are iPhone and Garmin, from Google Trends is demonstrated. The longest duration of statistical information that can be acquired from Google is 14 years, starting from 2004 to the present. First, the acquired datasets are plotted with respect to the corresponding date. Then, the anomaly detection method is applied in MatLab R2017b and the resulting classifier along with threshold values are visualized on a plot. In Figure 2. and Figure 3. the results for the data of iPhone and Garmin are displayed [5]. As seen in the figures, the classifier model successfully separates slow and fast changes. Also, the threshold learns the characteristics of the metric sequentially and updates itself accordingly. The learning rate of threshold depends on time; thus as the time passes the update rate slows down accordingly. In both figures the system gives false positive errors, that is to say, anomaly decision is made under normal non-anomaly conditions [6]. The reason for that is the system starts learning in the beginning and calibrates itself to the data. During this period, the couple of false positive errors are encountered but later all the anomaly detections are correct and there are neither false positive nor false negative errors.

Figure 2: The graph of raw Google Trends dataset for Garmin (top), and the graphs for classifier and threshold values (bottom).

Figure 3: The graph of raw Google Trends dataset for iPhone (top), and the graphs for classifier and threshold values (bottom).

4. CONCLUSION

In this paper, we have proposed an adaptive mathematical model to detect the Google Trend data anomalies. The method is applied to two different datasets that are acquired from Google and the results are displayed in the experiments section. As expected, the anomalies are detected correctly. The benefit of such adaptive approach is that the system learns the characteristics of the dataset and calibrates itself to the flow gradually. By this way, instead of converging to a fixed value, our threshold moves dynamically in accordance with the incoming data and thus result much more accurate decisions.

REFERENCES

- [1] Barber, Brad, M., Odean, T. 2008. All That Glitters, The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors. *Review of Financial Studies*, 21, 785–818.
- [2] Zhi, D., Engelberg, J., Gao P. (2011). In Search of Attention. *The Journal of Finance*, 66, 1461– 1499.
- [3] Vozlyublennaya, N. (2014). Investor Attention, Index Performance, and Return Predictability. *Journal of Banking and Finance*, 41, 17–35.
- [4] Barberis, Nicholas, Shleifer, A., Vishny, R. (1998). A Model of Investor Sentiment. *Journal of Financial Economics*, 49, 307–343.
- [5] Kewei, H., Peng, L., Xiong, W. (2008). A Tale of Two Anomalies : The Implications of Investor Attention for Price and Earnings Momentum.” SSRN. Retrieved from <http://ssrn.com/abstractD976394>.
- [6] Morgan, H. (2013). Twitter : The Carnival Barker of Investing. Retrieved from <http://www.fool.com/investing/general/2013/04/05/twitter-the-carnival-barker-of-investing.aspx>.