Group Anomaly Detection using Hierarchal Models

Author Names Go Here Using the Author Name Style

Affiliations and Addresses Go Here Using the Affiliation and Address Style  
publications15@aaai.org

Abstract

This is an example abstract. It goes here in 9 point Times New Roman type.

Introduction*[[1]](#footnote-1)*

Group anomaly detection refers to the problem of finding anomalies pattern of a collective groups.

This is example text. It is 10 point Times New Roman. The first sentence after the heading begins without a paragraph indent.

Related Works

The text below a second-level heading begins without indentation. This is example text. It is 10 point Times New Roman.

This is the second paragraph. It in formatted with the Text-indent style. This is example text. This is example text. It is 10 point Times New Roman. This is example text. It is 10 point Times New Roman.

This is a Subsubsection Heading

Figure 1 AUC comparison of GMM with MGMM

This is example text. It is 10 point Times New Roman. This is example text. It is 10 point Times New Roman.

This is example indented text. It is 10 point Times New Roman. This is example indented text. It is 10 point Times New Roman. This is example indented text. It is 10 point Times New Roman. This is example indented text. It is 10 point Times New Roman..

This is an example of an extract or quotation. Note the indent on both sides. Quotation marks are not neces­sary if you offset the text in a block like this, and properly identify and cite the quotation in the text.

This Is an Example of a Figure Caption.

Experiments

In this section we provide the experimental result produced by the group anomaly detection on three domain sets.

1. Synthetic dataset

Gaussian Mixture Model (GMM) using synthetic dataset. We generated the synthetic dataset using the generative described on the in [section 1]. For our experiments we generated set of synthetic dataset with 2, 5 and 10 dimension of 50 groups and 5000 points. Then we inject anomalies point both point and group type anomalies in the generated synthetic dataset.

To compare the accuracy of the point wise anomaly detector GMM with the MGMM, we aggregated the score of individual points in the groups. We took log sum of the score of each points to get the group wise likelihood for the GMM. We compare the score with the group likelihood generated from the MGMM of the 50 groups. The overall performance comparison is shown below.

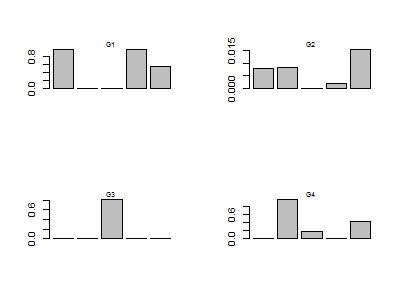
The result shows, the group anomaly detection achieved better result in all of the dimension dataset compared to the point wise Gaussian mixture Model.

1. ADAMS dataset
2. Weather temperature data

In the experiment, we tested the performance of the group anomaly detection in air temperature weather data from Oklahoma stations. We select 12 weather station with distance of below 40 miles from ‘SPEN’ weather station. The assumption is, nearest weather stations will have highly correlated readings in specific time instance. We collected the air temperature the 12 stations where each station records 288 readings per day. So, each station have 288 dimension for the 365 days of the whole year. But, the some of the station exhibit fault, which result in anomaly readings. Here we plan to detect these anomalies by grouping each day readings from all stations, because in each day the readings will be highly correlated from all stations. Then we group the readings based on the day’s readings, so that we have 365 groups throughout the whole year readings. To process the data we reduced the dimension of the data from 288 to 4 using PCA by retaining 95% variance.

We tested our proposed hierarchal model on this domain to find the likelihood score of each groups using different topic and genre (group type) proportion. Our assumption is if there is a faulty readings from one station in particular day, this will make the group more anomalies. For the experiment we used different parameter set of Topic and genre, T=2, 3, 5, 10, 20 and 100 and K= 3 to 7. The performance is measured using area under the ROC curve (AUC) of finding the anomalies from the test set. For the ground truth we have label for each day readings of a particular station. To generate ground truth for the groups, we assumed if there is one or more faulty stations in that day the group is flagged as faulty.

The result shows, the overall detection performance is not really good. There reason could be because in the dataset the number of group members are small. In this case in each group we have around 12 points. But, in comparison the algorithm achieves better for T=5 and K=4. In other hand, there result in the weather data shows, there no improvement as we increase number of topics or group type.

Then from the AUC result we took the better result for group type K=4 and topics T=5 which is around 0.71. Using this configuration we compute the topic proportion of each group type, which is shown in the figure below.

The result figure shows, how each group type are different in their topic composition. For example, Group type 1 are mostly from topic 1, 4 and 5, and Group type 3 are mostly composed of topic 3.

from the MGMM of the 50 groups. The overall performance comparison is shown below.

Conclusion

All fonts must be embedded for your paper to be published. No exceptions.

References

Engelmore, R., and Morgan, A. eds. 1986. *Blackboard Sys­tems.* Reading, Mass.: Addison-Wesley.

*(If you use EndNote and it complains "object has been deleted" when you try to generate references, temporarily delete the heading with the copyright notice attached as a footnote, and reinsert it after EndNote finishes.)*

1. Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved. [↑](#footnote-ref-1)