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.

This is example text. The second paragraph is indented 10 points, with no extra space between lines. The text is 10 point Times New Roman. 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 text. It is 10 point Times New Roman. 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 text. It is 10 point Times New Roman. 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 text. It is 10 point Times New Roman.

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

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

1. Synthetic dataset
2. ADAMS Intrusion Detection Dataset

This dataset contains information of 5691 users of a corporate network. All activities of these users are monitored and encoded into 69 features. The activities were recorded by date i.e. for each day we have one feature vector corresponding to one user, which we can call a user day. We run the group anomaly detection on a collection of such user days for the month of April 2013. The complete dataset contains around 170K user day instances. Our goal is to identify the anomalous users i.e. the users those are actually threat. The ground truth anomalies were inserted by the Red Team Inserts, a team whose job is to insert anomalies in a realistic way to make them appear as they occurred naturally. To run the group anomaly detector we treated each user is a group where each group i.e. user contains 30 user days. Our goal is to detect users whose user day distribution appears anomalous i.e. to detect the groups those distributions are different than normal groups distributions. The point here is that the individual user days might not be anomalous but when considered as a group they might appear normal.

We assume that each user group is coming from *T* normal group types where under each group the user days belongs to one of the *K* global topic those are shared among the entire dataset.

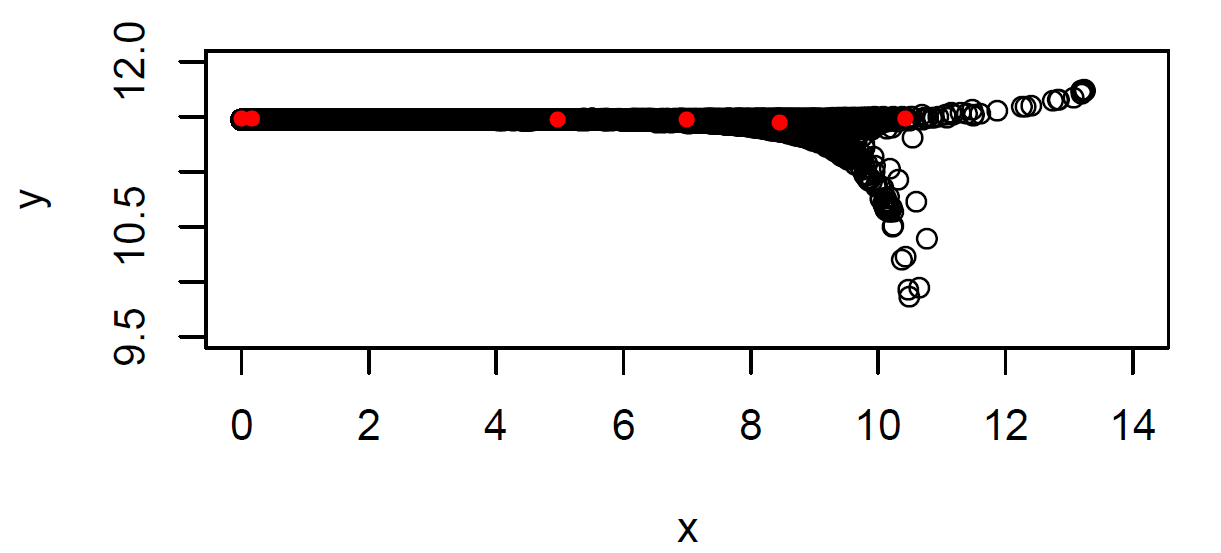
Data Preprocessing: It is common to apply some dimensionality reduction techniques to the data before applying the anomaly detection, since the dataset might contain some low variance and redundant dimensions. We used PCA to reduce our data. First, we applied PCA and retained only the principle components those represent 95% of the variance and it reduced the dataset from 69 dimension to just 2 dimensions. We also tried retaining 99% variance, which reduced the data dimension to 4. And also before applying PCA we scaled the data to have unit variance and then applied the PCA, which reduced the dimension 69 to 38. We run the group anomaly detection on these three reduced datasets and observed that they produce almost similar performance in terms of AUC. Hence, we only report and discus the results found from the reduced 2D data, since it will allow us to get more insight into the datasets.

Figure : Scatter plot of all the points along with the 6 topic centers learned (Red points).

Topics and Groups learned: We choose different values of T and K and computed the corresponding AUC. We got the best AUC of 0.78 when used T = 3 and K = 6, i.e. 3 normal group types and 6 topics. First we show the scattered plot of the reduced 2D dataset in figure 1 above along with the 6 topic centers (red dots) found after learning the parameters. It is worth to mention that we translated the points to make all the coordinates positive and then applied log on both x and y axis, hence the x and y axis in figure 1 are in logarithmic scale.

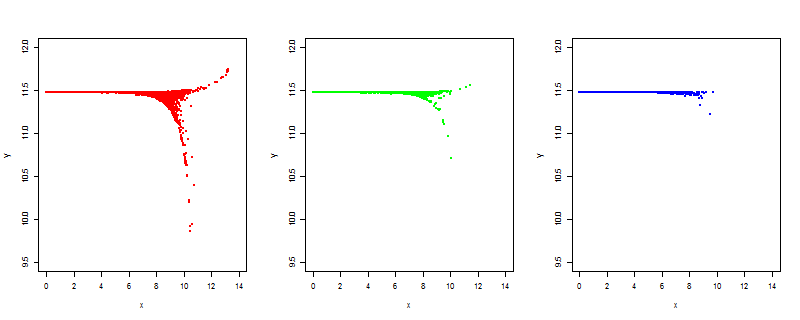
We inferred the group types for each of the users from the learned parameters. We show the data points assigned in three normal group types in figure 2 below. We observe that most of the groups are assigned to group type 1 and less number of groups are assigned to group type 3.

Figure :Scatter plot of the points assigned to group type 1 (left), group type 2 (center) and group type 3 (right).

In figure 3 above we show the distribution of topics under each group types. We see that most of the poins are generated from topic 1 and topic 6 and only a few points are generated from topic 2. Group type 3 is sparse i.e. contains most of the points from topic 1 and 6 and almost none from other four topics.

Figure : Distribution of topics under each groups

Effect of number of normal groups: We want to see how the peformance of the anomaly detector varies when we choose different number of normal group types. In figure 4 below we plot the number of normal groups in x axis and in y axis we put average AUC of different topics:

Figure : Effect of number of groups in performance of anomaly detector

We observe that in figure 3, when we increase number of groups there is a decrease in performance, i.e. AUC decreases and best performacne is achieved when using number of normal group types equals 2. The performance degrades when we increase number of normal group types is because anomalous groups will tend to fall under one of these groups which will give more false negatives.

Effect of number of topics: In figure 5 below we plot number of topics vs. average AUC in different group numbers.

Figure : Effect of number of topics in performance of anomaly detector

Choosing small number of topics does not help identifying anomalous groups. Also large number of topics does not. We observed that the best peformance is achieved when number of topics is 6. The reason is that when we have very small number of topics the group types will tend to produce similar topic distributions hence will get lots of false negatives, on the other hand if we have large number of topics the distributions will be mostly anomalous and will produce large number of false positives. In both of the cases the anomaly detector will perform badly.

1. Weather temperature data

Papers **must** be printed in two-column format on 8-1⁄2 x 11 inch paper. The margins if you use 8-1/2 x 11 inch paper must be **exactly** as follows:

• Top margin: 3⁄4 of an inch

• Left margin: 3⁄4 inch

• Right margin: 3⁄4 inch

• Bottom margin: 1-1⁄4 inches

The actual image size for your paper (minus margins) is 7

x 9 inches (17.78 x 22.86 cm). Papers that deviate from these measurements will not be published.

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)