Layered, Multivariate Anomaly Explanations: A First Look

Matthew Michelson MMICHELSON@FETCH.COM
Sofus A. Macskassy SOFMAC@FETCH.COM
Fetch Technologies, 841 Apollo St, Ste. 400, El Segundo, CA 90245 USA

1. Introduction

An anomaly is a data point that deviates dramatically from some set of related data points, based on some common metric. For instance, consider the flights shown in Table 1. In this table, flights 2 and 5 are anomalous because their flying time is much longer than the average even though the set of airports is similar. While there is significant research on *finding* anomalous records within data sources (e.g. (Lane & Brodley, 1999; Steinwart et al., 2005), we are unaware of research on explaining why those anomalous data points are actually anomalous (beyond their deviance from the "standard" value for the metric). Generating these explanations is the focus of this paper.

Table 1. Example flights (anomalies shown in **bold**)

Flights			
id	origin	destination	flying time
1	LAX	SFO	1h
2	**LBO**	**OAK**	**9h**
3	LAX	SFO	1.1h
4	LAX	SFO	1.2h
5	**LAX**	**OAK**	**9.3h**
6	LAX	SFO	1.2h

This paper presents our initial work on automatically generating a layered, multivariate explanation for anomalies. Given a set of anomalous records, such as Flights 2 and 5 from Table 1, an "anomaly explanation" is defined as a model that describes the variables common to the anomalous records and how they relate to each other. This model functions as a hypothesis for describing the anomalous group. For instance, using the anomalous flights of Table 1, an anomaly explanation models the notion that long flight times depend on the flight destination, which in turn depends on the flight's origin, as shown in Figure 1. As we will show, our approach leverages outside data sources to increase the richness of the features to consider in the model, and so we consider it "multivariate." Further, as we show, our approach supports drilling down into this explanation, yielding detailed explanations for specific subsets of the anomalous group. Therefore, we consider it a "layered" model. We note that while we focus on anomalous groups of data for this paper, since those groups generate useful and interesting explanations, our technique will generalize to any coherent (non-random) group of data. As long as the records are coherently grouped our technique should generate an explanation.

As mentioned above, sometimes the initial data is not rich enough to form reasonable explanation models.

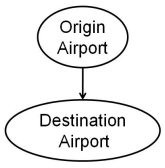

Figure 1. An example anomaly explanation model

In this case, outside knowledge can be incorporated by linking extra information to the data, yielding a "multivariate" model that incorporates features from disjoint sources. These richer, outside features help generate more detailed explanations. For example, we might link in detailed information about each flight by joining a flight registry source with the flight data, generating the new records shown in Table 2. By appending information such as the type of aircraft and the operating company, we generate a more detailed model, shown in Figure 2. This model shows that the deviant flights also depend on the type of aircraft, which is slower, and that these slow aircraft are used to fly to certain destinations.

Table 2. Example flights (anomalies shown in **bold**)

Flights					
id	origin	destination	flying time	aircraft type	top speed
1	LAX	SFO	1h	747	650kts
2	**LBO**	**OAK**	**9h**	**Prop**	**250kts**
3	LAX	SFO	1.1h	747	650kts
4	LAX	SFO	1.2h	747	675kts
5	**LAX**	**OAK**	**9.3h**	**Prop**	**225kts**
6	LAX	SFO	1.2h	747	650kts

Finally, an anomaly explanation such as Figure 2 functions at a high-level because it models the factors that determine the whole set of anomalies. However, we also want to drill into subsets of the anomalies and explain them in finer detail. For instance, although Figure 2 explains the anomalies of Table 2, other flights that are also anomalous because of their long flight time (not shown) may also fit this model. However, these flights might have slightly different features from the example Flights 2 and 5. Therefore, we further refine the explanation, grouping Flights 2 and 5 together because they share both a rare destination ("OAK") and use a rare "Prop" plane that is slow. Our approach generates these finer grained explanations for subsets of the anomalies, which is why we consider our method as "layered."

The rest of this paper is as follows. Section 2 describes our initial work on generating layered, multivariate anomaly explanations. Section 3 presents initial results, conclusions, and future directions for this work.

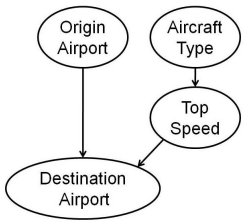

Figure 2. A richer anomaly explanation model

2. Layered, multivariate explanations

As stated in the introduction, given a set of anomalous records, $a_i \in A$, our goal is to generate a layered, multivariate explanation in the form of a model. Our approach breaks into three distinct phases. First, our algorithm incorporate as much outside knowledge as possible to generate an explanation using as rich a feature space as possible. Second, an actual model is generated from the feature space built by the first step. Finally, we cluster together the anomalies based on how well they fit various parts of the model, providing a finer grained explanation of how the anomalies fit the explanation.

Therefore, the first step is to join together the anomalous records with other outside sources of information. As stated, we start with anomalies $a \in A$, where a is a record composed of fields $f_1 \dots f_m$. We also maintain a list of outside sources S_i, each of which contains other fields $o_{i1} \dots o_{in}$, and a mechanism for joining together records in A with each outside source S_i (denoted $\bowtie$ for $f_j \in A = o_x \in S_i$). The result of this first step is a new view over the set of joined data:

$$A' = \forall_i A \bowtie S_i$$

An example is given in Table 2.

Next, our algorithm takes the richer set of data in A', and generates a Bayes Net hypothesis for the records[1]. This Bayes Net functions as our multivariate, anomaly explanation such as that shown in Figure 2. A Bayes Net model is advantageous for a number of reasons. First, it compactly models which attributes are common to the set of anomalies A', and also how the attributes relate to each other. These relations are dependencies which is particularly useful for interpreting their interplay. We do note, however, that building a Bayes net assumes that the data points in the subset are i.i.d. , which is not necessarily the case, especially when joining records across sources. We realize this is a shortcoming of our method, and hope to overcome this assumption in our future research.

A second, and important, advantage of a Bayes Net is that it provides an intuitive mathematical model of the anomaly explanation. That is, we can generate a Markov factorization for a Bayes Net, yielding a probabilistic formula for fitting anomalies in A' to the generated explanation. Given a Bayes Net with nodes X,Y,Z, the Markov factorization is:

$$P(X = x, Y = y, Z = x) = \prod P(v|parents(v))$$

The Markov factorization not only provides a fit for each anomaly to the explanation, it also provides a way in which to cluster subsets of the anomalies into groups that fit particular parts of the Bayes Net similarly. These subsets yield the finer grained explanations sought by our layered approach. In some sense, the Bayes Net functions as feature selection, narrowing the search space of similarities amongst the anomalies, while the clustering of the Markov factorization yields the actual fine grained explanations.

The factorization is a probability distribution for each part of the hypothesis explanation. Therefore, we can compare the factorizations of any two anomalies a_x and a_y by computing the Kullback-Leibler divergence (D_{KL}) between their factorized pieces. For probability distributions P and Q, D_{KL} is defined as:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Since D_{KL} is neither symmetric, nor necessarily non-negative, we instead use a modified version of the distance, known as Jensen-Shannon Divergence (D_{JS}):

$$D_{JS}(P||Q) = \frac{D_{KL}(P||M)}{2} + \frac{D_{kL}(Q||M)}{2}$$

Where $M = 1/2(P + Q)$.

To cluster our subset members using the Markov factorization and D_{JS}, we build a fully connected graph where the nodes are the records in A', and the edge weights are the D_{JS} scores between all nodes. Nodes are clustered together if their edge weight is greater than a threshold, such that resulting clusters represent sets of nodes that fit the various parts of the explanation similarly. Examining the factorized pieces of the explanation, for each cluster, generates finer grained explanations. For example, we would see that since Flights 2 and 5 of Table 2 both share a rare airport and flight type, and so they would share a low divergence score (high similarity). Therefore, we can give a finer explanation for these records based on the attributes common to their cluster.

Overall, our algorithm for discovering layered, multivariate anomaly explanations is given by Figure 3.

3. Conclusions and Future Work

Our approach is still in its initial investigation. As such, we don't yet have our full empirical results to test our approach to generating anomaly explanations. In fact, deciding how to empirically test our approach could prove to be challenging. However, we do have some encouraging anecdotal results. Our study uses an initial 48,629 flights culled from the FlightAware flight tracking service. Each initial flight record has attributes such as the origin airport and destination airport, along with numeric data such as planned flight time, actual flight time, etc. To build multivariate explanations, we join this flight data with data from

[1]We use the Bayes Net Power Constructor(Cheng et al., 1998) which allows for efficient generation and conditional independence testing

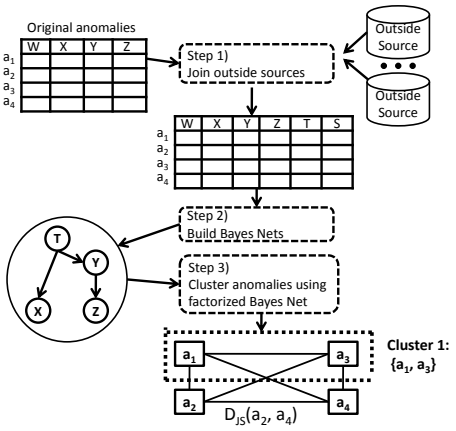

Figure 3. Generating layered, multivariate explanations

the FAA Flight Registry to augment the records with attributes such as the type of plane used, the company that operates the plane, etc.

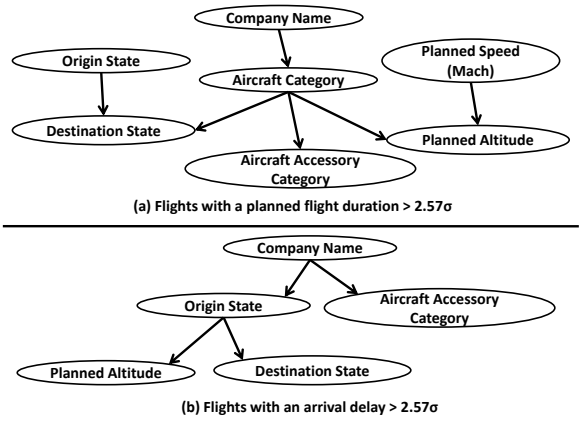

Figure 4. Generated anomaly explanations

To generate sets of anomalies to explain, we compute the z-score for each numeric attribute in the original flight data, and consider all flights whose z-score is greater than or less than 2.57 as anomalies (Assuming a normal distribution motivates our choice of 2.57). Note that we treat anomalies separately depending on whether they are above or below the z-score, since they are likely to be distinct (that is, very slow flights are different anomalies than very fast flights).

Figures 4 shows two of the generated Bayes net explanations for different sets of anomalous flights. The explanation for (a) describes flights whose planned fight time is deviant, while (b) describes flights who arrived much later than other flights. From the picture, we see that flights that planned an abnormally long flight time depended on such attributes as where they are flying to, which depends on where they are flying from, and how high they intend to fly, which depends on how fast they want to go. It is intuitive that flights that plan to take a long time will make such a decision based on where they are going and how fast they plan to get there. It is also interesting that the type of plane is so important (defined by both the aircraft category and the accessories, which further subdivide the plane types), and that the company flying the plane is such a factor. This is because certain companies will fly certain types of planes at rarer altitudes (with respect to other companies flying other types of planes).This explanation is multivariate because it uses attributes from the flight registry. The explanation for late arriving flights (shown in (b)), again suggests that the route is important, as well as how high the plane can fly. It couples this information with the companies and the types of planes, so there is some correlation between companies flying certain planes and arriving late.

We also generated the clusters for the anomalous flights of (a) and (b) above. As stated, the clusters serve as more detailed explanations for subsets of the anomalies. For the flights of (a) in the figure above, we found clusters such as flights that share a same rare company and plane (e.g. a low value for P(aircraftcategory|companyname)), while other clusters group around a commonly occurring destination and origin pair. For the flights of (b) we again see clusters such as planes that share a rare origin and destination, or those that share a common origin and destination. Although anecdotal thus far, and though space prohibits speaking about each cluster found, our approach does generate clusters that give more detailed and intuitive explanations for anomalous flights.

Although our initial results are encouraging, this is still early work. In the near future we plan to examine more advanced clustering techniques, and also perform rigorous empirical testing. Also, we must explicitly deal with the fact that our multivariate data points in the subset are not i.i.d.

Acknowledgements

This work was sponsored in part by the Air Force Office of Scientific Research under award number FA9550-09-C-0064. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the U.S. Government.

References

Cheng, J., Belland, D., & Liu, W. (1998). *Learning bayesian networks from data: An efficient approach based on information theory* (Technical Report). Department of Computer Science, University of Alberta.

Lane, T., & Brodley, C. E. (1999). Temporal sequence learning and data reduction for anomaly detection. *ACM Trans. Inf. Syst. Secur.*, *2*, 295–331.

Steinwart, I., Hush, D., & Scovel, C. (2005). A classification framework for anomaly detection. *Journal of Machine Learning Research*, *6*, 211–232.