ON MEASURING ABNORMALITY
The effect of time of observation in detecting anomalies or frauds

Introduction

I have been working on many anomaly and fraud detection systems for five years and what I have found in these years shows that:

1. We can't use a single machine learning or statistical model to detect any abnormal behavior in a system, like an only deep neural network - which people are usually very fond of it! - or a single recurrent network or an autoencoder, etc. Instead, we have to use many classifiers or clustering processes or even statistical models and wire them together to get the result. So basically, it means we cannot do it on a shell scripting environment unless we are facing a simple problem. We need to design and build a real application.
2. Regardless of what algorithm we use to detect the abnormalities, if the algorithm can't give acceptable reasoning to the customer, they are not happy.

Here we try to describe a simplified version of an algorithm that can help you to address this problem.

Common sense vs Science

Meriam Webster defines common sense as: "sound and prudent judgment based on a simple perception of the situation or facts," while science is based on observation and experiments and processes of logical deduction and reasoning. But again, when we use a scientific algorithm to give the result if people can't get satisfied with their common sense, they just assume that result as just a result! The best way to overcome this problem is using common sense as the fusing part of the algorithm and use scientific methods to do the rest. To get familiar with the topic let us bring some examples:

Example 1: Your friend comes to the office, and you see something is different with him. You have never seen him wearing glasses.

Example 2: Your friend comes to the office, and you see something is different with him. You have never seen him in casual dress at work, but you already knew he always dresses casually on weekends.

These two simple examples show our judgment works based on our previous observations, and this is common sense, but the problem is when you are dealing with a system having many dimensions or a long history of observations, our brain doesn't pay attention to all of these details and chooses some other methods. We already know that some people's common sense oversimplifies the problems and gets some weird results, like believing in ghosts, etc. Just do not forget that the simplification process is nothing but the process of building a model, their problem is that their model is too much simple and remember that models should not be either complex and over fit or too simple and under fit.

The importance of observation time

Even if the problem you are working on, has nothing to do with time, the time of observation is essential to be considered, like the below examples:

Example 3: It is 1985, and you go to the best buy and see they are selling 3.5" floppy disk.

Example 4: It is 2018, and you go to the best buy and see they are selling 3.5" floppy disk!

But sometimes the problems are directly related to time like:

Example 5: You use your credit card at 2:00 pm to buy a $2,000 computer.

Example 6: You use your credit card at 2:00 am to buy a $2,000 computer!

So, these examples show you can't get rid of time even if the system you are dealing with its measures, doesn't have anything to do with time.

Applying time factor

It is simple, to model the effect of time on how much we should consider the result of an observation important, we need to consider these assumptions:

1. The past observations are less critical.
2. The recent observations are less critical.

At first glance, you might think these two are denying each other, but they are not. System's behavior changes through the time, for example, you used to smoke twenty years ago but never since then, now if you start smoking it is not normal. Or hundred years ago the traffic rush hour used to be around 7:00 am, but nowadays it is around 8:30 am. So, what it means is that the samples you have should not have equal value, so you can't use simple frequency counting and calculation the probability to find out how much the recent outcome of your observation is abnormal, you need to consider the first assumption and the second too.

The reason to consider the second assumption is that something gets normal or at least not abnormal if it happens for a certain number of times. If you see your old friend smoking after 20 years, you are not going to call him a smoker. But if you continue observing him smoking, after a while you can tag him as a smoker.

The very same example is right for credit card usage. If you used to use your credit card like $10,000 a month, but in the past ten years your usage has been around $1,000 a month, and now records show it has been used for $10,000 in this month, it could be a sign of fraud!

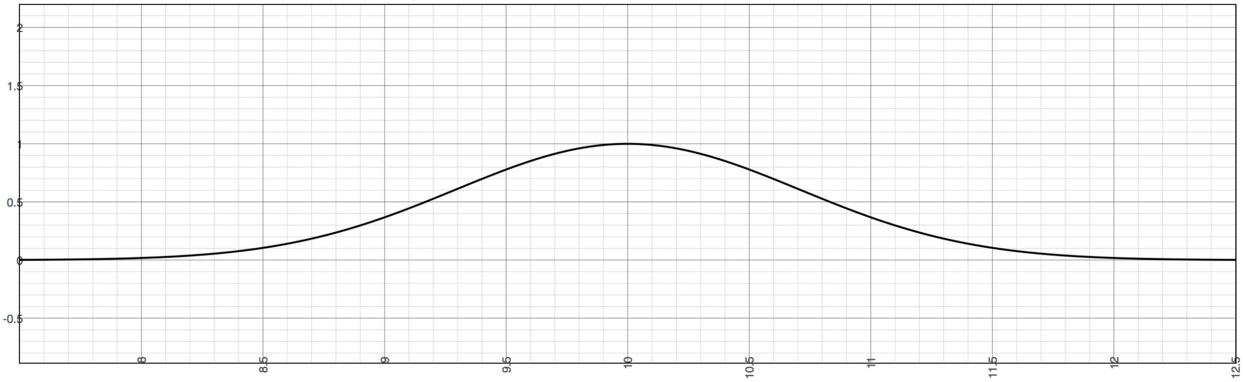
A measure for weight of observation time

We can use any form of the model to describe the time effect we talked about on frequency counting, but since most of the phenomena in nature somehow follow the exponential saturation model, we use a general form as below:

$$observationWeight(t) = e^{-(\frac{t-t_0}{df})^2}$$

(1) General form of calculating *observationWeight* of an observation outcome.

Let's name this factor in (1) as *observationWeight* a measure that shows how much you can count on the observation that has happened at time *t*. Here in (2) you can see a sample *observationWeight* curve. And by the way, *df* in (1) helps you to change the width of the bell curve.



(2) Sample observation weight model

So now the least what we can do when counting the frequencies, is using formula (1) as the normalized score or weight instead of assuming an equal score of one for every single outcome regardless of the time they happened.

Fuzzy outcomes

The problem with the real world is that most of the time the outcome of an observation is not as clear as we like. We can't tell the result belongs to class A or B or C, even most of the time we can't do any classification and have to use clustering methods. So regardless of what we do to build a model for clusters or classes the result of applying these methods to our training sets gives us some groups of outcomes, like G_1, G_2, G_3, ... Note that in any of these groups, there are many observation outcomes happened in different time with different *observationWeight*. So, to calculate the probability of how much each group is normal we have to use the *observationWeight* as in (3).

$$abnormality(G_i) = 1 - \frac{\sum_{observation \in G_i} observationWeight(t)}{\sum observationWeight(t)}$$

(3) A simple measure for abnormality

Formula (3) mainly says even though Earth has experienced living of millions of dinosaurs for millions of years while modern human beings have been around just for about hundred thousand of years, still observing a human is normal while observing a dinosaur is not! But formula (3) is still simple; we can do better.

Consider a credit card transaction, most of the time we can't tell a transaction is 100% similar to one of the classified or clustered groups. The best way to calculate *abnormality* of a transaction is either using fuzzy classification or clustering and then calculating the *abnormality* of a transaction based on weighted sum of the group *abnormality* or using instance-based models.

Using instance-based learning

Many people believe that instance-based learning methods are not a learning method at all! While I think it is one of the critical ways we humans use to earn knowledge. Let's get back to credit card example, how a bank can identify if a transaction is a fraud or not? We can use the combination of observation weight we talked about, and the idea of a transaction is normal if it already has happened many times before. How many times? It depends on the business, for a bank transaction we might say if we see ten similar transactions at least in a month, then it is normal, if not it is abnormal.

Now by using the idea of observation weight, that defined number of times can give its place to a score index. So, for a transaction, we should look back and calculate the sum of the *observationWeight*s of the similar ones, if they pass the minimum score in that period, it is OK. We can also define a similarity function between two transactions and instead of just using *observationWeight* use the formula shown in (4) to calculate the score of a transaction *T* at time *t*, consider we already have a set of instances like *Ti.*

$$score(T, t) = \sum similarity(T, T_i) \times observationWeights\,(T_i\,, t_i)$$

(4) An instance-based model for calculating score of a transaction *T* at time *t* to see if it is normal or not.

The formula in (4) might be a bit deceptive; you might think it ends up to a summation of many terms, but it does not. Because if you check the graph (2) you see transactions that have happened in just a portion of time have enough weight to alter the score, the others have weights very close to zero, so it limits the number of instances[1].

[1] And most of the time you also have other constraints that limit your instances, like the day of the week, time of day, geo location, ...

The similarity function which is a pseudometric function[2] also returns one, when two instances are equal and zero, when they don't have minimum required similarity. So, based on our definition in a linear model it might return zero or something between like 0.6 and 1.0. We can also map the [0.6, 1.0] interval to [0, 1.0] using any linear or nonlinear transformation. We should not forget that the similarity function could be as complex as the model we are discussing.

Conclusion

Anomalies and Frauds, in particular, are not simple phenomena to catch, especially when you are dealing with complex systems with changing dynamics. Unless you are working with single or double dimensions data, you can't use simple clustering or classification for detecting abnormal data or if you do your system faces many false positive or false negative. To use machine learning technologies in this field, you'd better use them as some micro pattern recognition units and then connect these groups using business knowledge engineering or statistical models. That gives you more control and understanding over what happens in the detection process and enables you to describe the reasoning to the customer.

There are many other ways we can use to find abnormal behaviors, like using hierarchical clustering or building belief networks or even extending the similar method we discussed to other dimensions. The only thing we tried to describe in this white paper was that we have to consider the observation or sampling time as a critical factor in the detection process.

[2] Function f is pseudometric if:

- $f(x, y) \geq 0$
- $f(x, x) = 0$
- $f(x, y) = f(y, x)$
- $f(x, y) + f(y, z) \geq f(x, z)$