

Fault Diagnosis of Machines

Biswajit Sahoo

This story was originally written for [Augmenting Writing Skills for Articulating Research \(AWSAR\)](#) award 2018. It is written in a non-technical way so as to be accessible to as many people as possible irrespective of their educational background. The story also featured in the top 100 list of stories for the award. Full list of awardees and their stories can be found at [this link](#).

Prelude

Rising sun with its gentle light marks the arrival of morning. Birds' chirp as well as time on our clock, sometimes with a blaring alarm, confirm the arrival of morning. Each of these, among several others, is an indicator of the morning. But can we know about morning by following only one indicator? Let's deliberate. What if the sky is cloudy and we don't see the sun rising, will this mean that morning is yet to come? Of course not! Our alarm will remind us of morning irrespective of whether there is sun or not. But what if, on some occasion, our clock doesn't work. In that case, birds may chirp or sun may rise or our near and dear ones may remind us that it's morning already. So in essence, we usually don't look for only one indicator. Rather, we consider several indicators. If one indicator fails, we can check another and thus be sure. It is very unlikely that all the indicators will fail simultaneously.

So the best way to get an idea about an event, it seems, is not to rely on only one indicator. Rather, observe several indicators and depending on their collective state, arrive at some conclusion. In this way, we deliberately add redundancy in order to get reliable results. This is exactly what we do in fault diagnosis of machines. Fault diagnosis is a broad term that addresses mainly three questions. First, find out whether fault is there in the machine or not. If fault is present, next question is to find the location of the fault. Once location of the fault is found, finally, find out the type of fault and its severity. In this article, we will only limit ourselves to the last aspect. But for simplicity, we will still use the term fault diagnosis to address that particular problem.

The method

To determine the health of a machine, we collect a set of indicators that best explain the condition of the machine. In scientific jargon, we call those features. Before discussing further, let's first discuss what are those features and how they are calculated.

First, data needs to be collected from a machine whose health needs to be assessed. Data might pertain to vibration level of the machine or its temperature distribution or the sound produced by the machine or something else. Sensors are needed to collect each type of data. By analogy, a thermometer, which is used to measure body temperature of humans, is a sensor that measures temperature. Likewise, different types of sensors are available to measure different quantities of interest related to the machine. From research it has been found that vibration based data are more suitable for fault diagnosis as compared to other types of data, say, temperature or sound. So in this article, we will limit our attention to vibration based fault diagnosis. And the sensor that is most commonly used to measure the vibration of a machine is called an accelerometer. From the data collected by accelerometer(s) we calculate features like the maximum level of vibration, similarly, the minimum level and other statistical features like skewness, kurtosis, etc. It is not uncommon to collect 10-15 features.

After feature collection, the next task is to find out what type of faults are present by using those features. One way to do this is by comparing the obtained feature values to pre-existing standards. But standards are available for few specialized cases when each feature is considered in isolation. For multiple features, no concrete information can be obtained from standards. The way out of this problem is to come up with an algorithm that takes all feature values as input and produces the output related to the type of fault present.

Construction of such an algorithm requires prior faulty and non-faulty data of similar machines be fed to it. The algorithm should ideally work well on this prior data. Once fine-tuning of its parameters are done, new data are fed into the algorithm and from its output, we infer the fault type. If the algorithm is carefully constructed, error in prediction of fault type will be very small. In some cases, it is also possible to get perfect accuracy. The approach just considered is a sub-class of a broad field called pattern recognition. In pattern recognition, we try to find underlying patterns in features that correspond to different fault types. This type of pattern recognition tasks are best performed by machine learning algorithms. The simple technique just described works fine for a large class of problems. But there exist some problems for which the features previously calculated are not sufficient to identify fault. However, it is possible to modify the technique by using transformation of data as well as features. Transformations are a way of converting the original data into another type such that after transformation more insight is gained out of it. This is similar to using logarithms in mathematics to do complex calculations. While direct computation of complex multiplications and divisions is difficult, using logarithm we transform the original problem into a simpler form that can be solved easily in less time. The transformation trick along with pattern recognition methods are surprisingly effective for most fault diagnosis tasks.

Some recent advances

Up to this point, we have argued that redundancy is important. It helps us take reliable decisions. However, it requires collection of huge amounts of data. Thus, continuous monitoring of machine, also known as online monitoring, becomes infeasible. So we seek an algorithm that is capable of finding fault types using only a few measurements. One way to do this is to select a few important features that can perform fault diagnosis. Research shows that it is indeed possible. But merely finding best features is not enough. Because to calculate the features, even though small in number, we need to collect all data. Hence issues related to online monitoring will still exist. A way around this problem is not to collect all data but only a fraction of it randomly in time. And the data should be collected in such a way that all information regarding the machine can be extracted from these limited observations. An even optimistic goal is to reconstruct the original data from the limited collected data. By analogy, this is similar to reconstructing the speech of a person, who speaks, say, 3000 words, from 300 random words that you have remembered of their entire speech. The problem just described is known as compressed sensing. And no matter how much counter-intuitive it may seem, encouraging results for this problem have been obtained in signal processing and these methods are beginning to get applied to problems of fault diagnosis. The problem is still in its infancy in fault diagnosis field.

What we learned (and what we didn't!)

In summary, we have learned that to diagnose faults, we need multiple features and sometimes we have to transform the data into different domains to get better accuracy. We then observed that we can get rid of the redundancy inherent in this method by using compressed sensing methods. All these techniques come under data-driven methods. It is called data-driven because all analyses are done after we collect relevant data from the machine. These methods are quite general purpose and can be used to diagnose faults in different components, say, detecting faults in cars or in other machines.

Apart from data-driven methods there also exists another class of techniques that go by the broad name of model-based methods. In model-based methods, we formulate a full mathematical model of the machine and then try to find out how the response of the model changes if a fault is introduced and using this fact, try to find the nature of fault for a new problem. Though model-based techniques are important in their own right,

in many cases it becomes very difficult to find an accurate model of the system because of the uncertainties involved. In contrast, data-driven methods are more robust against external noise and are flexible, meaning, we can perform different analysis using the same data and obtain deeper insights. Another advantage of using data-driven methods is that the whole process of fault diagnosis can easily be automated.

In this article, we have only considered the field of fault diagnosis. In fault diagnosis, faults are already present and we wish to either detect them or segregate them depending on fault type. But there exists another branch that deals with ways to predict the time of occurrence of fault in future, given the present state. Basically, they determine the remaining useful life of the machine. This sub-branch is called fault prognosis which is also an active area of research.

Given the advancement of research and scope for automation, it may be possible, in not so distant future, to get updates on your phone about possible malfunction of a part of your car while driving your car or while enjoying a ride in a self-driving car, maybe!!