Biometric Analytics Cost Estimating

Sean McKenna, Ph.D.
Lead Scientist

Joseph Sarage
Senior Economist

Booz Allen Hamilton Inc.

International Cost Estimating and Analysis Association (ICEAA) Conference
San Diego, CA
June 10, 2015

Abstract

This paper examines the interplay between biometric technologies and advanced analytics, referred to as biometric analytics, as a way to detect fraudulent entries in a biometric system. We follow a systematic approach, based on cost estimating standards, to ascertain whether deploying such a capability is worthwhile. A simple case study is presented that illustrates the key aspects of our analysis. In addition, there are a series of cost elements that quantify the impact biometric vulnerabilities have on individuals, companies, and countries.

Introduction

As biometric technologies and systems find their way into more and more facets of daily life, the need for such systems to be reliable and secure has become that much more important. Concurrently, fields such as data science are growing at a dynamic rate and are providing organizations new and powerful ways in which to leverage large quantities of data to drive improvements and transformation. This paper aims to explore the confluence of these two fields, what we refer to as *biometric analytics*, and begin to estimate the benefits and costs of implementing a biometric analytics capability that looks to address the problem of fraudulent records in a biometric data store.

This paper is organized as follows. We begin by presenting a focused background on biometrics and advanced analytics as a way to orient the reader for what is to come and introduce the motivation behind our study. A brief summary of our estimation approach follows. With our approach established, we present an example case study that illustrates the key aspects of the problem. We conclude with some salient points that emerged during our analysis and thoughts for future work.

Background

In this section, we provide the reader with some relevant background information on the topics of biometric and advanced analytics.

Biometrics

Biometric technologies measure and analyze human physiological and behavioral characteristics. Identifying an individual's physiological characteristics is based on direct measurement of a part of the body, e.g., fingertips, hands, face, and eye retinas and irises. Identifying behavioral characteristics is based on information derived from actions, such as speech and how one signs his/her name. Because the characteristics they measure are thought to be distinct to each person, biometrics can be very effective personal identifiers. Unlike more traditional identification methods that rely on something one has, such as an identification card for building access, or something one knows, such as a PIN to access an ATM, biometrics are integral to something about the individual. Being inherently linked to the individual, they are more reliable, cannot be forgotten, and are less easily lost, stolen, or spoofed.

While biometric technologies vary in complexity, capabilities, and performance, all share several elements in common. At a fundamental level, all biometric identification systems reduce to pattern recognition systems. They use sensors such as cameras and scanning devices to capture images, recordings, or measurements of an individual's characteristics along with computer hardware and software to extract, encode, store, and compare these characteristics. Because

the process is almost always automated, biometric decision-making is typically very fast, and in some cases, real-time.

Depending on the application, biometric systems can be used in one of two modes: verification or identification. Verification, or authentication, is used to *verify* a person's identity; i.e., to authenticate that an individual's reported identity is their true identity. Identification, on the other hand, is used to *establish* a person's identity; i.e., to determine who a person is. Although biometric technologies measure different characteristics in substantially different ways, all biometric systems involve similar processes that can be divided into two distinct stages: (1) enrollment and (2) verification or identification.

In enrollment, a biometric system is populated with the information needed to identify a specific person. The person first provides an identifier, such as an identification card. He or she then presents the biometric (e.g., fingertips, hand, iris) to a suitable acquisition device, the distinctive features are located, and one or more samples are extracted, encoded, and stored as a reference template for future comparisons. Finally, this biometric is linked to the identity specified on the identifier.

In verification systems, the objective is to verify that a person is who he or she claims to be (i.e., the person who enrolled). After the individual provides the identifier that was used during enrollment, the specific biometric is presented. The system captures the biometric and generates a trial template. The system then compares the trial biometric template with this person's reference template to determine whether the individual's trial and stored templates match. Verification is often referred to as 1:1 (one-to-one) matching. Verification systems can contain databases ranging from dozens to millions of enrolled templates but are always predicated on matching an individual's presented biometric against his or her reference template.

In identification systems, the objective is to identify who a person is. Unlike verification systems, an identifier is not necessary. To find a match, instead of locating and comparing the person's reference template against his or her presented biometric, the trial template is compared against the stored reference templates of all individuals enrolled in the system. Identification systems are referred to as 1:N (one-to-N, or one-to-many) matching because an individual's biometric is compared against multiple biometric templates in the system's database.

Advanced Analytics

The field of analytics is as broad as that of biometrics, arguably broader. At its core, analytics is the discovery and communication of meaningful patterns in data, relying on the simultaneous application of statistics and mathematics, computer programming, and data manipulation to extract valuable knowledge from data. While analytics can be as austere as fitting a line to a set

of data points, it can also be as complex as developing an artificial neural network to perform speech recognition.

In the context of our analysis, we will focus on more complex analytics, rooted in the field of machine learning. Machine learning, as defined rather formally by Mitchell [1], is a framework where, "a computer program is said to learn from experience *E* with respect to some task *T* and some performance measure *P*, if its performance on *T*, as measured by *P*, improves with experience *E*." More intuitively, machine learning is most commonly used to mean the application of induction algorithms and other algorithms that can be said to "learn." An algorithm is said to be inductive if it takes as input specific instances and produces a model that generalizes beyond these instances. The learning aspect is typically realized through a process called supervised learning, wherein the algorithm is presented with a training data set from which to learn. This training data set consists of example inputs and their desired outputs or "labels." For instance, the inputs could be physical characteristics of a person, such as height, weight, hair color, and so on; the corresponding labels could then be male or female. The algorithm would use these inputs and the corresponding labels to "learn" a model that mapped inputs to output. In cases where a training data set is not available, one turns to unsupervised learning. Here, no labels are given to the algorithm, leaving it on its own to find structure in the input data. Clustering, or grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other group, is a classical unsupervised machine learning task.

Motivation

Biometric systems have long been used in law enforcement and forensics, and more recently, has gained prominence as a reliable, cost-effective means of personal authentication. Government and commercial applications include immigration, border control, airport security, physical access control, ATM authentication, and mobile device security [2]. To quickly perform verification of a subject or to perform identification against a watch list, government agencies and other users of these systems maintain large databases of digital biometric records. For example, today, the FBI's Integrated Automated fingerprint Identification System (IAFIS) contains fingerprint records for over 64 million individuals [3].

Biometric systems are vulnerable to attacks at various stages in the biometric recognition process, including attacks on the database in which enrolled entries are stored. Large biometric databases pose challenges to testing and protecting the integrity of collected data. For example, fingerprint databases may be vulnerable to cyberattacks aimed at impersonating or concealing an individual's identity through the use of synthetically generated fingerprint images (spoofs). These images could allow the attacker to replace their own fingerprints in the database so that the attacker is not recognized during subsequent identification attempts. Additionally, an

attacker can perform a “masquerade attack,” in which they impersonate another individual by injecting a synthetic image that has been reconstructed from the desired individual’s feature set.

A number of advanced analytics techniques (e.g., machine learning approaches) have been proposed to address the problem of spoofed biometric detection [4, 5]. The belief is that with the rapid growth of fields such as data science and our ability to mine and exploit massive data stores (such as those associated with biometric records), identification of spoofed records in large databases should now be possible. Furthermore, identification of fraudulent biometric authentications, in near-real-time, should also be practical. In order to transform these possibilities into reality, a biometric analytics framework is needed. The exact details of such a framework will depend on the application. For instance, detecting spoofed biometric records in a database would likely required a form of unsupervised clustering wherein bona fide records were assigned to one group and spoofs, to another. In contrast, detecting a fraudulent authentication could be accomplished using a supervised algorithm that has been trained to recognize legitimate biometric features as different from spoofed features.

In any event, the goal of implementing a biometric analytics capability would be to reduce, in an automated fashion, the instances of fraud within a system. Such a reduction would translate to a cost savings, whether it be a readily quantifiable savings (e.g., reducing welfare abuse) or a less tangible cost reduction such as reducing occurrences of illegal entry or access (e.g., illegal entry into the Unites States by someone on a watch list). Clearly, these sorts of benefits do not come without a cost. In this case, it is the cost of developing, implementing, and maintaining a biometric analytics capability. Determining whether or not adopting such a capability is ultimately worthwhile is an important decision that requires a systematic analysis. Our approach for estimating these costs and benefits is outlined in the following section.

Approach

We follow a systematic approach, based on cost estimating standards, to ascertain whether deploying biometric capabilities are a worthwhile investment. To quantify the costs of biometric vulnerabilities, our approach was to assess the impact at the individual, company, and country level using publically available data.

For the cost estimates, we used a Booz Allen Hamilton, Inc. Monte Carlo Simulation software tool called Argo™ fitting the data with a triangular distribution and random variable generation to for impact value estimation. For each of the cost element structure items, we selected a “low,” “mode,” and “high” value to bound our variable set and we ran a 5,000 trial Monte Carlo Simulation analysis. The results are shown in the Case Study section of this paper.

In addition, we also estimated the cost of implementing a biometric capability at a company to reduce the likelihood of a vulnerability. Given the potential cost impact of a fraudulent

biometric entry, the benefits of implementing a robust biometric analytics capability outweighs the cost for a company or government entity. There are various development, production, and sustainment costs associated with implementing a robust biometric analytics capability however these investment costs directly contribute to the reduction of fraud and illegal activity from interlopers.

Case Study

For our case study, we examined a large data store of biometrics data to include fingerprints. The data store of prints has anomalous entries in it (spoofs, for example). Depending on how this data store is used (for example, as a match against persons entering the US), we quantified what the cost is as a result of there being a chance of a person exploiting the spoof and gaining entry illegally.

For cost estimating purposes, a triangular distribution was used to estimate the cost of harm to a person if a biometric feature was compromised.

The table below shows the impact in U.S. dollars.

		Risk Distribution Parameters					
Biometric CES	Name	Distribution Type	Low	Mode	High	Random Var.	Impact
1	Cost of harm to a person	Triangular	$ 52,800	$ 71,500	$ 92,400	0.19	$61,469

For cost estimating purposes, a triangular distribution was used to estimate the cost of harm to a company if a biometric feature was compromised.

The table below shows the impact in U.S. dollars.

		Risk Distribution Parameters					
Biometric CES	Name	Distribution Type	Low	Mode	High	Random Var.	Impact
2	Cost of harm to a company	Triangular	$ 134,400,000	$ 210,000,000	$ 336,000,000	0.07	$208,839,429

For cost estimating purposes, a triangular distribution was used to estimate the cost of harm to a country if a biometric feature was compromised.

The table below shows the impact in U.S. dollars.

		Risk Distribution Parameters					
Biometric CES	Name	Distribution Type	Low	Mode	High	Random Var.	Impact
3	Cost of harm to a country	Triangular	$ 27,700,000,000	$41,550,000,000	$58,170,000,000	0.35	$36,400,709,555

For cost estimating purposes, a triangular distribution was used to estimate the cost to implement a biometric analytics capability.

The table below shows the impact in U.S. dollars.

		Risk Distribution Parameters					
Biometric CES	Name	Distribution Type	Low	Mode	High	Random Var.	Impact
4	Cost of implementing a biometric analytics capability	Triangular	$ 33,600,000	$70,000,000	$168,000,000	0.24	$87,453,660

In essence, from a cost standpoint, a fingerprint recognition algorithm is required for analysis of fingerprint images of different levels of the quality to produce a matching score.

For cost, assign a cost factor by each of the “illegally-gained entry” points. For example, if a fingerprint is tied to a social security number (SSN), estimate the cost on a per unit basis of what this would ultimately cost the victim. I can derive a cost estimating relationship that will change given the type of illegally-gained entry point. We can come up with a list of “illegally-gained entry” points and depict how the costs change / how impacts change as well.

Conclusion

Biometric systems are vulnerable to attacks at various stages in the biometric recognition process, including attacks on the database in which enrolled entries are stored. Through our research and analysis, it is evident there are numerous costs that impact individuals, companies, and countries if biometric data is compromised. One of the key challenges is determining whether or not adopting such a capability is ultimately worthwhile and this is an important decision that requires a systematic analysis.

There is significant opportunity for future work in this field to include deriving ways to reduce the overall cost of implementing a biometric capability at the individual, company, and country level. As technologies improve and become more accessible, costs of implementing are reduced and these capabilities will be accessible to a broader population.

References

[1] T. Mitchell, *Machine Learning*, McGraw Hill, 1997, p. 2.

[2] D. Maltoni and R. Cappelli, "*Fingerprint Recognition*," in Handbook of Biometrics, 2008, pp. 23–42.

[3] K. R. Moses, P. Higgins, M. McCabe, S. Prabhakar, and S. Swann, "*Automated Fingerprint Identification Systems (AFIS)*," in National Institute of Justice/NCJRS, 2010, pp. 6:1–33.

[4] R. Derakhshani, "*Spoof-Proofing Fingerprint Systems Using Evolutionary Time-Delay Neural Networks*," in Proceedings of the 2005 IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety, April, 2005.

[5] S. Nikam and S. Agarwal, "*Fingerprint anti-spoofing using ridgelet transform*," in Biometrics: Theory, Applications and Systems, 2008, BTAS 2008, 2nd IEEE International Conference on, September 2008, pp. 1–6.

[6] Bureau of Justice Statistics (BJS).

[7] Bureau of Labor Statistics (BLS).