Combining Human and Artificial Intelligence for Analyzing Health Data

Erik P. Duhaime

Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02142
eduhaime@mit.edu

Abstract

Artificial intelligence (AI) systems are increasingly capable of analyzing health data such as medical images (e.g., skin lesions) and test results (e.g., ECGs). However, because it can be difficult to determine when an AI-generated diagnosis should be trusted and acted upon—especially when it conflicts with a human-generated one—many AI systems are not utilized effectively, if at all. Similarly, advances in information technology have made it possible to quickly solicit multiple diagnoses from diverse groups of people throughout the world, but these technologies are underutilized because it is difficult to determine which of multiple diagnoses should be trusted and acted upon. Here, I propose a method of soliciting and combining multiple diagnoses that will harness the collective intelligence of both human and artificial intelligence for analyzing health data.

Introduction

Many routine medical diagnoses are made by individuals or small, homogenous groups of likeminded medical practitioners, and are thus prone to systematic biases and human error. While soliciting additional opinions often improves diagnostic accuracy, it can be difficult to determine when the benefit of doing so outweighs the costs. Furthermore, once a decision is made to solicit additional opinions, it can be difficult to determine whose opinion should be solicited next, and which of the multiple diagnoses should be trusted and acted upon. At the same time, artificial intelligence systems are increasingly capable of analyzing medical images and test results, such as skin lesions and ECGs (e.g., see Monheit et al., 2011). However, for the same reason, many such systems are not utilized effectively, if at all. In other words, the collective intelligence of both human and artificial intelligence is constrained by the difficulty of aggregating multiple diagnoses. This is particularly unfortunate as the power of artificial intelligence systems continues to increase, and as advances in information technology and crowdsourcing continue to make it easier for health data to be collected and shared with people throughout the world.

Extracting the Wisdom of Crowds

Researchers have long appreciated what is known as the "wisdom of crowds" effect, which results when aggregating a group's forecasts yields predictions that are almost as good as or better than those given by any of the individuals in the group (Galton, 1907; Surowiecki, 2005), but the problem of how to best collect and aggregate individual forecasts is still very much unsolved. In practice, a simple average or median of individuals' forecasts is often used, but this will be counterproductive if some individuals have greatly superior information or abilities. For instance, while averaging the diagnoses of three doctors will normally lead to improved accuracy, averaging the diagnosis of an experienced doctor with two other clueless individuals will clearly lead to worse outcomes on average.

It has recently been shown that the wisdom of crowds effect only works to the extent that the group is diverse (Davis-Stober, Budescu, Dana, & Broomell, 2014; Page, 2008), and therefore the decision of whose opinion to solicit next should depend on whose opinion has already been solicited. For instance, at one extreme, averaging the diagnosis of three equally talented doctors who all think the same way will simply lead to the same diagnosis of just one of those doctors. Thus, soliciting additional opinions from diagnosticians whose expertise best complements the skills of the initial diagnostician will improve overall accuracy. While it can be difficult to determine how individuals' skills and expertise complement each other based solely on titles and education, an analysis of a large dataset of individuals' diagnoses on the same cases will allow for this possibility in future cases. This is especially feasible when considering how to combine human diagnoses with those of AI-generated diagnoses, since these systems can rapidly diagnose very large test sets of previous cases. Such com-

binations are also especially likely to be fruitful, since AI-generated diagnoses are less correlated with human diagnoses than human diagnoses are with each other. Either way, by considering the statistical relationships between the diagnoses of different humans with each other and with those of AI systems, statistical techniques can determine which opinion should be next solicited and at which point an aggregated diagnosis has crossed a predetermined accuracy threshold. This will not only improve accuracy on difficult cases, but it will also reduce costs when analyzing relatively easy cases.

Method

To demonstrate the promise of this approach, I offer simulations based on hypothetical diagnostic data. For instance, first consider a doctor, Doctor A, who accurately identifies that a mole is cancerous based on a picture and limited medical and demographic information 95% of the time. However, because this doctor is very cautious to avoid false negatives, she correctly identifies when a mole is not cancerous only 60% of the time. If in 50% of all cases the mole is in fact cancerous, then Doctor A will have a false negative rate of 2.5% and a false positive rate of 20%, and therefore she will incorrectly diagnose 22.5% of all cases. Now imagine that there are multiple individuals – "Type A" doctors – with similar rates of false negatives and false positives. As long as the diagnoses of Type A doctors are not perfectly correlated with one another, then combining their diagnoses through simply averaging their opinions will lead to improved accuracy. However, note that averaging the opinion of two doctors does not lead to improved accuracy because it is unclear what the decision rule should be if they disagree. This is similar to the tension that exists when doctors are provided with promising AI systems that are not as accurate as the doctors themselves.

Now consider an AI system that accurately identifies when a mole is cancerous only 65% of the time, but correctly identifies when a mole is not cancerous 85% of the time. This system will have a false negative rate of 17.5% and a false positive rate of 7.5%, and therefore it will incorrectly diagnose 25% of all cases in this example. Because the doctors are on average more accurate than the AI, such a system might not be put to use. However, once we have accumulated a large number of diagnoses, we can analyze how human diagnoses vary with those of the AI over time, which will enable us to determine when to trust the doctor and when to trust the AI. This has implications not only for how we utilize AI-generated diagnoses when they are worse, on average, than human-generated ones, but also for how we should utilize human diagnoses when they are worse than AI-generated ones. In this example, if Doctor A gives a negative diagnosis but the AI gives a positive diagnosis, it is more likely that the case is a negative. However, if Doctor A gives a positive diagnosis and the AI gives a negative diagnosis, it is slightly more likely that the case is a negative.

Importantly, the best collective diagnoses are not necessarily those generated by groups of the best individual diagnosticians. To understand, imagine that there is another group of doctors, Type B doctors. Type B doctors have the same skills that our hypothetical AI system did: a false positive rate of 7.5% and a false negative rate of 15%, and therefore lower overall accuracy than Type A doctors when 50% of all cases are positive. Even though Type A doctors are each more accurate individually, even small groups of Type B doctors are more accurate than groups of Type A doctors as long as Type B doctors are relatively less correlated with one another. This is because after we have solicited the opinions of a few Type A doctors we are almost certain that there will not be a false negative, so there are diminishing returns to soliciting more Doctor A opinions.

Furthermore, because the doctor types have differential skills – not just better or worse – a combination of both types outperforms a group of either type individually. For instance, a simple rule of soliciting three Type A doctors' opinions followed by two Type B doctors' opinions leads to higher accuracy than soliciting five opinions from either Type A or Type B. Notably, in this example, a heuristic of averaging the opinions of the three Type A doctors and continuing to solicit opinions *if and only if* the Type A doctors think the case is a positive will increase accuracy while reducing costs. With this approach, machine learning techniques can be utilized to discover many such rules and improve diagnostic accuracy even further.

Future Work

We are developing a mobile platform that will solicit quick, binary and/or categorical diagnoses for health data such as skin lesions and mammograms. We plan to use this empirical data to determine how these individuals' diagnoses should be optimally combined with those generated by AI systems in order to increase accuracy and reduce costs.

References

Engelmore, R., and Morgan, A. eds. 1986. *Blackboard Systems.* Reading, Mass.: Addison-Wesley.

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79.

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Monheit, G., Cognetta, A.B., Ferris, L., Rabinovitz, H., Gross, K., Martini, M., Grichnik, J.M., Mihm, M., Prieto, V.G., Googe, P., King, R., Toledano, A., Kabelev, N., Wojton, M., and Gutkowicz-Krusin, D. (2011). The performance of MelaFind: a

prospective multicenter study. *Archives of Dermatology*, *147*(2), 188–194.
Page, S. E. (2008). *The difference: How the power of diversity creates better groups, firms, schools, and societies*. Princeton, NJ: Princeton University Press.
Surowiecki, J. (2005). *The Wisdom of Crowds*. New York, NY: Knopf Doubleday Publishing Group.