Machine learning and Facebook

By Professor Jan Willem de Graaf
Professor of Brain and Technology, Saxion University of Applied Sciences, Deventer, Netherlands

Coming back to the blog of some weeks ago, we will go further into machine learning. Behaviour of people on the social media can be measured and linked to results on (psychological) tests. One of the most famous examples is the research by Youyou et al. (2015) in which they relate the giving of likes on Facebook items with the 5 characteristics of personality traits of the Big5 personality model: extraversion, kindness, openness, accuracy and emotional stability. What is needed to allow a machine to "learn" which liked Facebook items contribute to which aspect of personality?

Firstly, items must have at least 25 likes, preferably considerably more. Secondly, we must have all the likes of many users as well as the results of a Big5 personality test. The personality test is the starting point, because the machine (the algorithm) is supervised from this data. People who are extroverted, for example "like" items with splashing action, and more static images are not (or something random). The algorithm establishes a simple equation:

Friendliness = a (item-x) + b (item-y) ... zn (item-n)

This is called the least absolute shrinkage and selection operator (Lasso) algorithm. The coefficients (a, b etc.) in the equation can have a positive or a negative value. Valuation of picture item-x, or event item-y is positive, or negative, with the property kindness. Precisely because there is a lot of data, the algorithm can estimate quite accurately which items charge positive or negative on each of the 5 alleged personal characteristics, and which properties do not really matter. The coefficient is set to 0 for properties that do not appear to do, or only a little bit. Of the properties that do matter, the coefficient is then systematically strengthened. In this way, the algorithm thus constructs, as it were, a series of items that may or may not be "like" as an alternative version of the personality test. The number of components in the Lasso equation is limited: coefficients above a certain value (criterion) are "strengthened", set below a certain value to 0 and thus removed from the equation. The criterion can be determined on the basis of the maximum number of permitted components.

Youyou et al. showed that from 70 likes the algorithm could predict the personality better than friends, from 150 likes better than family members and from 300 likes even better than the test person himself. Of course, this is only possible with items that are shared very much in the social networks and when the results of the 5 subscales of the big5 are known from a large group of people. However, there is the danger of this type of research, the fit never gets better than the model that applies as supervisor, so here is the Big5 personality test. The fact that the Big 5 was ever designed using Likertscale questions that have been arranged through a 5-factor analysis model, contributes to the distinctive character of the 5 subscales, but does not say anything about the reality of the personality structure. Precisely because the big 5 originated on the basis of factor analysis, it is not surprising that the 5 items are fairly distinctly related to other patterns of human behaviour.

However, it would be a fallacy to find a justification of the personality model in finding correspondence between the 5 personality factors and clusters found by pattern recognizers. For example, we do not know whether the properties are really introverted or extroverted, or someone can not possess both properties at the same time, but depending on the context can show 1 of the 2 properties, or the properties change over time, or even what the properties exactly mean. The machine learning algorithm does not know the person better than the colleagues, the friends or the person themselves. The machine can "only", after sufficient "supervised learning" trials (iterations), determine the correlation between the click behaviour on Facebook and the scored Big5 factors better than individuals. Prediction replaces the Big5 questionnaire. But we are not getting closer to the personality of people than with the Big5 questionnaire. Garbage in stays garbage out. In the end psychology remains a subject of the story, of the argument, even though data mining can help enormously!