Web Scale Data Mining homework 1

Karol Kubicki

24.11.2014

1.1) **When machine learning algorithms are trained on larger data sets, they perform better. However, there is often a trade-off involved. Read the paper by Banko and Brill [2] and answer the question: what is the downside of training their algorithms on larger data sets? Think of and describe a scenario in which you would decide against using a large data set for training.**

The downside of training algorithms on larger datasets is the increasing cost of getting the data. According to figure 1 of the paper by Banko and Brill the learning curves of different algorithms were more or less linear with respect to increasing magnitudes of the amounts of data used. This means that to get linear growth in performance we need exponential growth in cost, assuming fixed price per unit of labeled data. There are way of addressing this issue like active learning or unsupervised learning.

The best example of a scenario where we don't want to use large data set for training is developing machine learning algorithm which could decide on a treatment for patient based on several factors describing the patient's health. In this case, to gain training data we would need to try different treatments on a large number of patients. We can't allow to risk somebody's health, so there is no easy way of obtaining rich training dataset.

1.2) **Google Flu Trends predict flu outbreaks in the U.S. based on locations of search queries associated with flu symptoms. A group of MIT researchers used data from smartphones' accelerometres to identify road surface defects (potholes). Imagine that you have access to all the reasonable data sources: search queries, satellite imagery, social media, open government data. You have a potential to convince a group of users to install your app on their devices. Think of and describe an innovative project that would rely on big data analysis to make the world a slightly better place. Describe the types of data that you would use for that purpose, and how you would use them.**

Lets call my app MoveIT. The basic functionality would be:

- a start questionnaire that would ask every user to rate their health condition, give estimate of how often they are ill, their weight, height, etc,
- regular (once a week) questionnaire about their current health condition - maximum 5 minutes to fill in,
- constant tracking of the amount of movement (based on accelerometer, gps, etc) as well as time of the movement - based on that we can estimate daily walking distance of users and the distribution of that walking throughout the day,
- after every meal the user would tell the app 'I've eaten' stating the size (small, medium, big). The app would automatically record the time of the meal,
- user would also notify the app just before going to sleep (the estimate of the time of waking up would be the time that mobile starts moving judged by accelerometer),
- optionally (best solution) user could store his/hers trainings like in endomondo.

Based on gathered data we could discover healthy living patterns based on the amount and distribution of movement, the amount and distribution of sleep, the frequency and size of meals, and so on. It would be interesting to see if the healthy patterns are different in different parts of the world and how many of those patterns exist. The basic advantage for user would be that our big data analysis could cluster users into groups with common healthy life pattern. Then app would suggest slight changes in daily behavior to make particular user healthier. The advantage of such approach is that we could find healthy pattern tailored for particular user. Thanks to that we hope that the user wouldn't have to make drastic changes in their life to feel better.

Above I described how I would use the data. Below is the list of kinds of data used:

- numbers - all the questions in questionnaires would ask for ratings (how often are you ill per year? how do you feel in scale of 1 to 10? How old are you? How much do you weight?)
- movement patterns - time and length of a walk, the velocity, etc,
- geospatial - location of person, his/hers walks, the altitude,
- eating and sleeping habits - time and size of every meal, time and length of every sleep,
- weather conditions - amount of sun, temperature, etc. We assume that together with other patterns it has influence of health.