Benchmarking the "Now"

Aleksander Kolcz
Twitter, Inc

Abstract

Unstructured and semi-structured text has contributed substantially to the rapid proliferation of data. Large quantities of automatically, editorially and user created documents are generated in large quantities daily, covering web pages, blog posts, text messages, tweets, and more. They are accompanied by large quantities of auxiliary text data, corresponding to various kinds of logs (search, activity, etc). These data can be (and are) analyzed from different angles, but it is often of interest to investigate how large quantities of data help in improving the quality of certain important tasks, such as finding relevant information or document categorization. Of particular interest is to apply algorithms and methods to past data so as to investigate avenues of improvement and provide better performance in the future. There are problems with such approaches, however, whereby the existing data are biased by algorithms and methods facilitating the data delivery and creation at the data collection time. Secondly, concepts such as relevance are time-sensitive and user-sensitive. Finally, there are many issues related to user privacy. While it would be ideal to be able to share all signals indicative of the level of interest of a user in a piece of information, this poses serious danger of releasing information that users deem private and which users would be unwilling to share.

Benchmarking The Personalized "Now" in Data Streams

Much of the "Big Data" arrives in the form of streams. In the case of Twitter for example, users' tweets form a natural global stream, and each user individually has their personal stream represented by a "timeline". The goal of many algorithms operating on streams is to take into consideration the past stream data up to the current moment in time and use this information to optimize for the present time (and the future). This optimization may involve category assignments, ranking of items returned in response to search queries and so forth. To the extent that effectiveness of these algorithms can be objectively verified, evaluation of competing approaches is possible. For some tasks (e.g., content categorization) it is relatively straightforward (by inspection) to obtain the ground truth at any point in time, and thus it is possible to assess the performance of methods developed in the future (i.e., after the data have been received).

A problem arises, however, in the context of personalization and time-sensitive recommendation. It is difficult to judge if an alternative way of

recommending content would have been preferable to any particular user unless the user was presented with that option at that time. Industry tends to handle these types of questions with population split AB testing. Here the user population is divided at random with the assumption that distributions of user types and preferences in the different user groups are identical. With such a split, competing approaches can be assessed by presenting different approaches to different populations. These types of tests cannot be done retrospectively, however, and it is hard to predict how users would have responded if presented with different types of information. This is due to the fact that for large datasets, most of the data is unlabeled from the viewpoint of any single user.

Thus, to the extent that massive amounts of user data are being collected, they document the reaction of users to algorithms operating on the data at that time. It is not clear however how best to use this information to estimate the effectiveness of alternative approaches. The challenge is to use this information to facilitate experimentation with new methods and techniques in a meaningful way, so that one can have confidence that if a new AB test was to be performed, the results would be reasonably aligned with the results obtained via the benchmark

Benchmarking Adversarial Problems

A special problem of temporal sensitivity relates to various forms of spam. Spam is pervasive in many forms of data (web, email, social networks) and is adversarial in nature. This means that even with large quantities of past data it is difficult to assess how a new algorithm would fare against the spammers unless spammers are actually forced to fight against that algorithm in the first place. This makes it difficult to benchmark new approaches to fight spam even if large quantities of labeled data are available (and of course the data is mostly unlabeled). One can certainly perform temporal splits and use only past data for learning detectors to be applied to future events. Nevertheless, this does not guarantee time pollution since the experimenter may already know certain characteristics of the "future" data. This also does not equate facing a real adversary since the change in future spam distribution, as evidenced in the data collected, happened only in response to the methods that were applied to combat it and not to the methods being tried experimentally.

Benchmarking with Real User Data

Both in the case of personalization and spam detection, protecting user privacy is important. But, at the same time it is rather unclear how to obtain ground truth data that is informative and yet cannot be tied to individuals. This is particularly important in creating benchmark sets that are to be shared between researchers and organizations, where enforcing confidentiality of the data is difficult. Unless the privacy issues are meaningfully addressed, the types of data that can be shared in benchmark dataset will be limited.