WSDM Cup Competitions

Friday 15th, 09:00–17:00, Room 112

WSDM Cup 2019 consisted of four data mining and optimization tasks, listed below. A total of 94, 386, 136, and 92 teams participated in the four challenges, with final rankings of the teams published on the respective competition platforms[1,2,3,4]. The top-three teams for each task have been invited to present their solutions at the WSDM Cup 2019 Workshop. The four tasks and the prizes were sponsored by ByteDance, Spotify, Baidu, and Sichuan Airlines, respectively, and we thank them for their generous assistance.

Fake News Classification

ByteDance, as a leading news aggregator and content provider, has a large-scale database of known fake news articles, and any newly arrived news articles goes through a test on the truthfulness of its content before being published. A matching is run to compare the newly arrived article with the known fake news articles in the database. Articles identified as containing fake news will be withdrawn after human verification. The accuracy and efficiency of the process are crucial to make their service safe, reliable, and healthy.

In this task, participants are given the title of a known fake news article A and the title of an unknown news article B, and the aim is to classify B into one of the following three categories: (i) AGREED: B and A are about the same fake news; (ii) DISAGREED: B refutes the fake news in A; and (iii) UNRELATED: B and A are unrelated. Both training and testing datasets are given, containing 320,767 and 80,126 data items respectively. A training data item (a pair of news titles) contains an id of the item, the title of a known fake news article, the title of an unknown news article, and the true label of the news title pair. A test data item contains the same fields except that the label field is hidden from the competition participants.

Sequential Skip Prediction

A key challenge for Spotify is to recommend the right music to each user. While there is a large body of work on recommender systems, there is very little work, or data, describing how users sequentially interact with the streamed content that they are presented with. In particular within music, the question of if, and when, a user skips a track is an important implicit feedback signal.

This task focuses on session-based sequential skip prediction, that is, predicting whether users will skip tracks, given their immediately preceding interactions in their listening session. The public part of the dataset consists of roughly 130 million listening sessions with associated user interactions on the Spotify service. A further 30 million listening sessions were used for testing. These testing sessions contain all the user interaction features for the first half of the session, but

[1] https://www.kaggle.com/c/fake-news-pair-classification-challenge

[2] https://www.crowdai.org/challenges/spotify-sequential-skip-prediction-challenge

[3] https://dianshi.baidu.com/competition/24/rule

[4] https://dianshi.baidu.com/competition/25/rule

only the track id's for the second half. In total, users interacted with almost 4 million tracks during these sessions. The output of a prediction is a binary variable for each track in the second half of a test session, indicating if it was skipped or not, with a 1 indicating that the track was skipped, and a 0 indicating that the track was not skipped.

User Retention Prediction

Baidu's Hao Kan App is an aggregation service that provides users accesses to a large number of short videos. Predicting user retention, that is, whether a new user of the app will use the app again after installation, is an essential task for the app developers, especially when new features or contents are added to the app.

Given a set of new users, including their profiles (such as age, gender, education level, location, interests), app usage times, installation source, and video viewing logs, this task aims to predict whether each new user will use Baidu Hao Kan App the day after they first installed it. The training data set contains 1.3 million users, 1.2 million videos, and 97 million video viewing records of the users. The testing data set contains 0.5 million users, 0.8 million videos, and 36 million video viewing records of the users.

Flight Schedule Optimization

An increasing challenge that Sichuan Airlines faces is how to reschedule its flights when delays occur, for example, due to bad whether or aircraft maintenance issues. Given the various constraints such as the availability of replacement aircrafts, it is difficult to reschedule delayed flights to minimize the delay time and to optimism the passenger experience. This task focuses on computing an optimal flight rescheduling plan under a set of given constraints.

Assume that there is a foggy weather at Shuangliu Airport (China) from 8:00am to 10:00am, with both inbound and outbound flights delayed by 2 hours. The airport re-opens at 10:00am. Given the flight plans of Sichuan Airlines in the next four days, the aim is to reschedule the flights to optimize an objective function defined based on delay time and passenger inconvenience. The following rescheduling operations are allowed: (i) replacing the aircraft for an upcoming flight so that it does not have to wait for the aircraft of a delayed inbound flight; (ii) delaying a flight; (iii) canceling a flight; and (iv) transferring passengers to other flights. There are nine constraints, such as availability of idle aircraft for replacement, and the number of vacant seats on the transfer flights. The dataset contained a set of flights to be rescheduled (2,150 records), a set of idle aircraft (17 records), a table of aircraft turnaround times (306 records), a table of airports and their allowed types of aircrafts (168 records), a table of flight routes and their allowed types of aircrafts (498 records), a table of cross-water flight routes (134 records), a table of types of aircraft non-compatible with cross-water routes (25 records), and a table of airport opening hours (11 records).

Jianzhong Qi	The University of Melbourne
Zhifeng Bao	RMIT University
Shiwei Wu	ByteDance Inc.
Brian Brost	Spotify Research
Jianmin Wu	Baidu Inc.
Shibin Song	Sichuan Airlines