Introduction

James Bennett
Netflix
100 Winchester Circle
Los Gatos, CA 95032
jbennett@netflix.com

Charles Elkan
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093-0404
elkan@cs.ucsd.edu

Bing Liu
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street
Chicago, IL 60607-7053
liub@cs.uic.edu

Padhraic Smyth
Department of Computer Science
University of California, Irvine
CA 92697-3425
smyth@ics.uci.edu

Domonkos Tikk
Department of Telecom. & Media Informatics
Budapest University of Technology and Economics
H-1117 Budapest, Magyar Tudósok krt. 2, Hungary
tikk@tmit.bme.hu

INTRODUCTION

The KDD Cup is the oldest of the many data mining competitions that are now popular [1]. It is an integral part of the annual ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). In 2007, the traditional KDD Cup competition was augmented with a workshop with a focus on the concurrently active Netflix Prize competition [2]. The KDD Cup itself in 2007 consisted of a prediction competition using Netflix movie rating data, with tasks that were different and separate from those being used in the Netflix Prize itself. At the workshop, participants in both the KDD Cup and the Netflix Prize competition presented their results and analyses, and exchanged ideas.

Categories and Subject Descriptors

H.2.8 [**Database Management]**: Database Applications – *data mining*. I.2.6 [**Artificial Intelligence**]: Learning.

General Terms

Algorithms, Experimentation.

Keywords

KDD Cup, Netflix Prize, collaborative filtering, recommendation.

1. KDD CUP 2007

This year's KDD Cup focused on predicting aspects of movie rating behavior. There were two tasks, which were developed in conjunction with Netflix and were chosen to be interesting to participants from both academia and industry.

KDDCup'07, August 12, 2007, San Jose, California, USA.

Both tasks employed the Netflix Prize training data set [2], which consists of more than 100 million ratings from over 480 thousand randomly-chosen, anonymous customers on nearly 18 thousand movie titles. The data were collected between October, 1998 and December, 2005 and reflect the distribution of all ratings received by Netflix during this period. The ratings are on a scale from 1 to 5 (integral) stars.

Task 1 (Who Rated What in 2006): The task was to predict which users rated which movies in 2006. The organizers provided a list of 100,000 (user_id, movie_id) pairs, where the users and movies were drawn from the Netflix Prize training data set. None of the pairs were rated in the training set. The task was to predict the probability that each pair was rated in 2006 (i.e., the probability that user_id rated movie_id in 2006). (The actual rating is irrelevant; each contestant only needed to predict whether the movie was rated by that user at some point in 2006. The date in 2006 when the rating was provided by the user was also irrelevant.)

Task 2 (How Many Ratings in 2006): The second task was to predict the number of additional ratings that the users from the Netflix Prize training data set gave to a subset of the movies in the training set. A list of 8863 movie_ids drawn from the Netflix Prize training set was provided. Each contestant needed to predict the number of additional ratings that *all* users in the Netflix Prize training set provided in 2006 for each of those movie titles. (Again the actual rating given by each user was irrelevant; only the number of times that the movie was rated in 2006 was required. The date in 2006 when the rating was given was also irrelevant.)

Paper Presentation: The top-ranked teams were invited to submit papers describing their algorithms. The accepted papers appeared in the workshop proceedings and were presented at the workshop.

2. NETFLIX PRIZE PAPERS

In October, 2006 Netflix released a large data set of movie-ratings and challenged the data mining, machine learning and statistical communities to develop systems that could improve the accuracy

of its recommendation system, Cinematch, by certain fixed amounts [2]. Netflix has agreed to award a Grand Prize to the team with a system that can improve the accuracy of Cinematch by 10%. In addition, Progress Prizes are to be awarded annually to teams that produce the greatest accuracy improvements over a given 12-month period. This is referred to as the *Netflix Prize* competition. Additional details about the competition can be found in Bennett and Lanning [3].

As part of this competition, many interesting data mining techniques have been (and continue to be) explored and applied to the Prize data set. The workshop was initiated with the aim to bring together competition participants, as well as other researchers interested in the Netflix Prize problem, to exchange ideas and to learn from each other in an informal setting. Netflix Prize participants were thus encouraged to submit papers describing their algorithms and experiences. After review by the program committee, selected papers were presented in the workshop.

3. INVITED PRESENTATIONS

The Netflix Prize competition has stimulated a great deal of high-quality research. With the aim of bringing some of this research to the workshop, we invited two speakers to present their recently-published work on the Netflix Prize data set. The speakers are Su-In Lee (Stanford University) and Andriy Mnih (University of Toronto), who each presented recent work based on their papers [4, 5] at the International Conference on Machine Learning (ICML-2007).

4. WORKSHOP ORGANIZERS

James Bennett, Netflix, USA

Charles Elkan, University of California, San Diego, USA

Bing Liu (Chair), University of Illinois at Chicago, USA

Padhraic Smyth, University of California, Irvine, USA

Domonkos Tikk, Budapest University of Technology and Economics, Hungary

5. PROGRAM COMMITTEE

Michael W. Berry, University of Tennessee, USA

Chris Ding, Lawrence Berkeley National Laboratory, USA

Ricci Francesco, Free University of Bozen-Bolzano, Italy

Genevieve Gorrell, University of Sheffield, UK

János Abonyi, Pannon University, Hungary

George Karypis, University of Minnesota, USA

Andras Kornai, Metacarta, USA

John Langford, Yahoo! Inc, USA

Ben Marlin, University of Toronto, Canada

Chris Meek, Microsoft Research, USA

Bamshad Mobasher, DePaul University, USA

Seung-Taek Park, Yahoo! Inc, USA

John Riedl, University of Minnesota, USA

Barry Smyth, University College Dublin, Ireland

Nathan Srebro, University of Chicago, USA

Volker Tresp, Siemens AG, Germany

Alexander Tuzhilin, New York University, USA

Lyle Ungar, University of Pennsylvania, USA

Tong Zhang, Yahoo! Inc, NYC, USA

6. ACKNOWLEDGMENTS

We would like to thank the ACM SIGKDD Chair, Gregory Piatetsky-Shapiro, for initiating the KDD Cup and Workshop 2007. We also thank the program committee members for reviewing all the Netflix Prize submissions. Stan Lanning from Netflix wrote the code used in evaluating the KDD Cup submissions. He also helped produce the training and test data for the competition tasks. We are very grateful to him.

7. REFERENCES

[1]. http://www.kdnuggets.com/datasets/kddcup.html

[2]. http://www.netflixprize.com

[3]. Bennett, J. and Lanning, S. The Netflix Prize. *Proceedings of KDD Cup and Workshop 2007*, Aug. 12, 2007.

[4]. Lee, S.-I., Chatalbashev, V., Vickrey, D., and Koller, D. Learning a Meta-Level Prior for Feature Relevance from Multiple Related Tasks. In *Proceedings of International Conference on Machine Learning* (*ICML-07*), Corvallis, OR, June 2007, pp. 489-496.

[5]. Salakhutdinov, R., Mnih, A., and Hinton, G. Restricted Boltzmann Machines for Collaborative Filtering. In *Proceedings of International Conference on Machine Learning* (*ICML-07*), Corvallis, OR, June 2007, pp. 791-798.