The Netflix Prize

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

Stan Lanning
Netflix
100 Winchester Circle
Los Gatos, CA 95032
slanning@netflix.com

ABSTRACT

In October, 2006 Netflix released a dataset containing 100 million anonymous movie ratings and challenged the data mining, machine learning and computer science communities to develop systems that could beat the accuracy of its recommendation system, Cinematch. We briefly describe the challenge itself, review related work and efforts, and summarize visible progress to date. Other potential uses of the data are outlined, including its application to the KDD Cup 2007.

Categories and Subject Descriptors

I.2.6 [**Machine Learning**]: Engineering applications – *applications of techniques.*

General Terms

Experimentation, Algorithms.

Keywords

Netflix Prize, RMSE, machine learning.

1. INTRODUCTION

Recommendation systems suggest items of interest and enjoyment to people based on their preferences. They have been under development since the early 1990s [2]. These systems play an important role in many e-commerce sites, notably Amazon, MusicStrands, Pandora, Yahoo!, and Netflix.

Netflix, an on-line movie subscription rental service, allows people to rent movies for a fixed monthly fee, maintaining a prioritized list of movies they wish to view (their "queue"). Movies are mailed to them or delivered electronically over the Internet. In the case of DVDs, when they are finished watching the movie they can return it by post and the next DVD is automatically mailed, postage free.

The length of service of subscribers is related to the number of movies they watch and enjoy. If subscribers fail to find movies that interest and engage them, they tend to abandon the service. Connecting subscribers to movies that they will love is therefore critical to both the subscribers and the company.

The company encourages subscribers to "rate" the movies that they watch, expressing an opinion about how much they liked (or disliked) a film. To date, the company has collected over 1.9 billion ratings from more than 11.7 million subscribers on over 85 thousand titles since October, 1998. The company has shipped over 1 billion DVDs, shipping more than 1.5 million DVDs a day. It receives over 2 million ratings per day. The company's Cinematch recommendation system analyzes the accumulated movie ratings and uses them to make several hundreds of millions of personalized predictions to subscribers per day, each based on their particular tastes.

The Cinematch recommendation system automatically analyzes the accumulated movie ratings weekly using a variant of Pearson's correlation with all other movies to determine a list of "similar" movies that are predictive of enjoyment for the movie. Then, as the user provides ratings, an on-line, real-time portion of the system computes a multivariate regression based on these correlations to determine a unique, personalized prediction for each predictable movie based on those ratings. If no personalized prediction is available, the average rating based on all ratings for the film is used. These predictions are displayed on the website as red-stars.

The performance of Cinematch is measured in several ways. In addition to various system throughput requirements, the accuracy of the system is determined by computing the root mean squared error (RMSE) [1] of the system's prediction against the actual rating that a subscriber provides. Improvements to the algorithms and underlying system performance have resulted, to date, in a roughly 10% improvement in RMSE over simply reporting the movie average rating.

2. THE NETFLIX PRIZE

In October, 2006 Netflix released a large movie rating dataset and challenged the data mining, machine learning and computer science communities to develop systems that could beat the accuracy of Cinematch by certain amounts [3]. Winners of the various Prizes are required to document and publish their approaches publicly, enabling everyone to understand and benefit

KDDCup'07, August 12, 2007, San Jose, California, USA.
Copyright 2007 ACM 978-1-59593-834-3/07/0008…$5.00.

from the insights and techniques required to achieve the enhanced levels of predictive accuracy.

Netflix provided over 100 million ratings (and their dates) from over 480 thousand randomly-chosen, anonymous subscribers 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. It withheld over 3 million most-recent ratings from those same subscribers over the same set of movies as a competition qualifying set.

Contestants are required to make predictions for all 3 million withheld ratings in the qualifying set. The RMSE is computed immediately and automatically for a fixed but unknown half of the qualifying set (the "quiz" subset). This value is reported to the contestant and posted to the leader board, if appropriate. The RMSE for the other half of the qualifying set (the "test" subset) is not reported and is used by Netflix to identify potential winners of a Prize.

The company reported the RMSE performance of Cinematch trained on the Prize dataset against the quiz subset as 0.9514, a 9.6% improvement over simply predicting individual movie averages.

The company will award a Grand Prize to the team with a system that can improve on that accuracy by an additional 10%. In addition, Progress Prizes will be awarded on the anniversaries of the Prize to teams that make sufficient accuracy improvements.

In addition to providing the baseline Cinematch performance on the quiz subset, Netflix also identified a "probe" subset of the complete training set and the Cinematch RMSE value to permit off-line comparison with systems before submission

3. FORMATION OF THE TRAINING SET

Two separate random sampling processes were employed to compose first the entire Prize dataset and then the quiz, test, and probe subsets used to evaluate the performance of contestant systems.

The complete Prize dataset (the training set, which contains the probe subset, and the qualifying set, which comprises the quiz and test subsets) was formed by randomly selecting a subset of all users who provided at least 20 ratings between October, 1998 and December, 2005. All their ratings were retrieved. To protect some information about the Netflix subscriber base [5], a perturbation technique was then applied to the ratings in that dataset. The perturbation technique was designed to not change the overall statistics of the Prize dataset. However, the perturbation technique will not be described since that would defeat its purpose.

The qualifying set was formed by selecting, for each of the randomly selected users in the complete Prize dataset, a set of their most recent ratings. These ratings were randomly assigned, with equal probability, to three subsets: quiz, test, and probe. Selecting the most recent ratings reflects the Netflix business goal of predicting future ratings based on past ratings. The training set was created from all the remaining (past) ratings and the probe subset; the qualifying set was created from the quiz and test subsets. The training set ratings were released to contestants; the qualifying ratings were withheld and form the basis of the contest scoring system.

Based on considerations such as the average number of ratings per user and the target size of the complete Prize dataset, the user's 9 most recent ratings were selected to assign to the subsets. However, if the user had fewer than 18 ratings (because of perturbation), only the most recent one-half of their ratings were selected to assign to the subsets.

4. RELATED WORK

Previous movie rating datasets have been released, notably the 1 million rating dataset provided by MovieLens.org [4]. Attempts to release a subsequent, larger dataset were abandoned when it became clear the identity of the raters could be discovered [5] [Riedl, personal communication].

There are several other notable data mining competitions. The

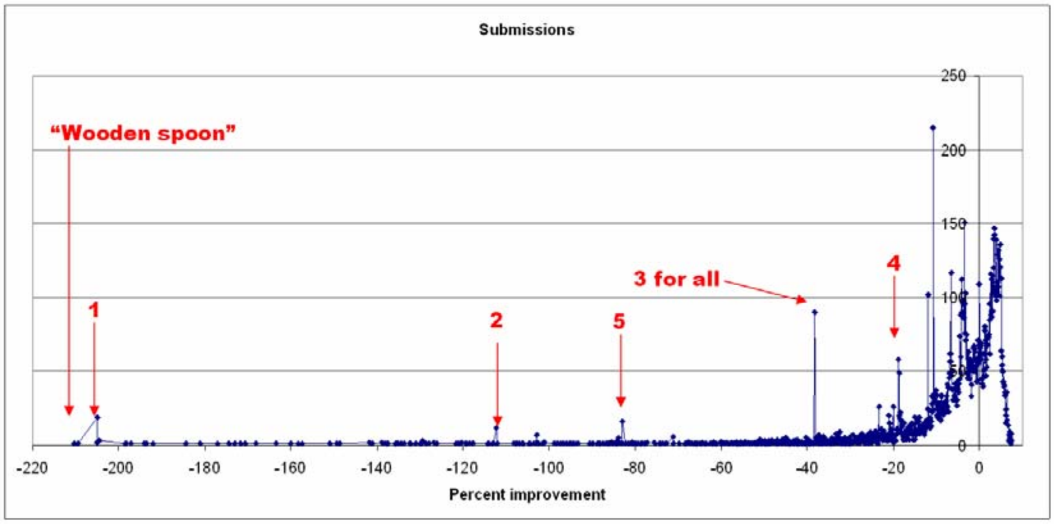

Figure 1: Distribution of submissions by percentage improvement over Cinematch.

KDD Cup [9], organized by the ACM SIGKDD, has been held every year since 1997. The Cup focuses on analyzing different datasets in search, bioinformatics, web mining, and medical diagnosis. The UCSD Data Mining Contest [10] engages American university teams in a variety of tasks, including interpreting mass spectroscopy data, time series analysis on user accounts, word and text prediction, and real property value prediction.

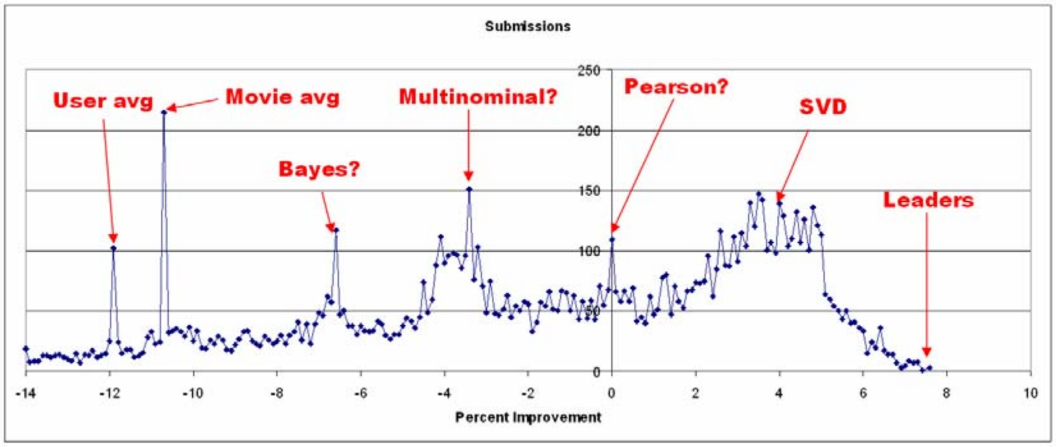

Figure 2: Detail of distribution of leading submissions indicating possible techniques

5. PRIZE STATUS, JUNE 2007

To date over 20 thousand teams have registered from over 152 countries around the world. Of those, 2 thousand have submitted prediction sets, currently over 13 thousand submissions.

The distribution of their accuracy measures with respect to Cinematch is shown in Figure 1. There have been over 650 teams that have exceeded the accuracy of Cinematch; 90 of those teams have exceeded a 5% accuracy improvement. The spikes in the distributions correspond to submissions containing a single-valued prediction for every item (e.g., all 3's).

Figure 2 shows the distribution of the leading submissions. Additional spikes with putative techniques (according to comments made on the Netflix Prize Forum) are also shown. These include the per-movie average and the per-user average rating. The current leading teams are located in Hungary, Canada, Japan, and the United States.

Figure 3 shows that after early and dramatic improvements, as of early February, progress slowed and appeared to have reached an asymptote just below 7% improvement. However, there has been a recent surge of progress as teams approach 8% improvement. A Progress Prize will likely be awarded after October, 2007.

In addition to active submissions, there has been substantial engagement between contestants on the Netflix Prize Forum [8], including sharing of code and coding ideas, additional data, insights on the patterns found in the data, even pooling of

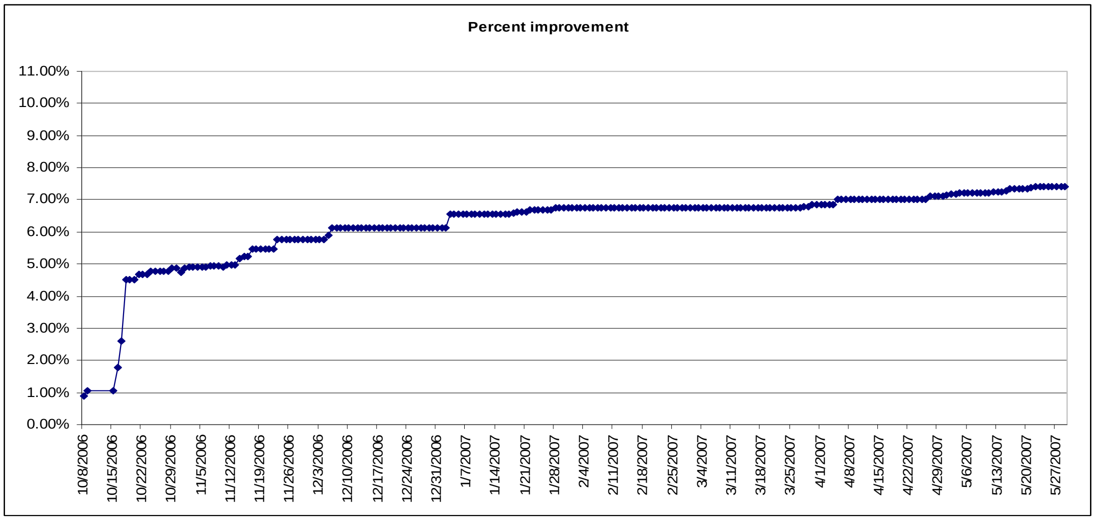

Figure 3: Aggregate improvement over Cinematch by time

submissions (and hence teams) themselves to achieve increased accuracy ("blending"). A notable event was the early lucid and detailed description of an SVD implementation by Brandyn Webb (aka "Simon Funk") and Vincent DiCarlo [6] that helped advance the pursuit of the SVD approaches. Many individuals have pursued extensions to this scheme and shared insights on what makes it work and fail. Other approaches and insights are now being reported in the literature [7].

6. FUTURE DIRECTIONS

Netflix has given permission to use the dataset for other non-commercial research purposes. In particular, the dataset is being used in another data mining contest, the KDD Cup 2007. In this contest participants may address two tasks. The first task is to predict which movies certain individuals would rate in 2006, the year after the training dataset was based. The second task is to predict the total number of ratings certain movies received in 2006 from the Netflix Prize user base.

Results from the KDD Cup 2007 competition will be available in late August, 2007.

7. ACKNOWLEDGMENTS

Our thanks to the KDD Cup committee: Charles Elkan, Bing Liu, Padhraic Smyth, and Domonkos Tikk for their feedback. And our great thanks to all the participants in both the Netflix Prize and the KDD Cup 2007 challenges.

8. REFERENCES

[1] Herlocker, J, Konstan, J., Terveen, L., and Riedl, J. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22 (2004), ACM Press, 5-53.

[2] Resnick, P. Varian, H.R., Recommender Systems, *Communications of the ACM* 40 (1997), 56-58.

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

[4] http://www.grouplens.org/node/12#attachments

[5] Frankowski, D., Crosley, D, Sen, S., Terveen, L., Riedl, J. You Are What You Say, *Proceedings of SIGIR*, 2006.

[6] Funk, Simon. Netflix Update: Try This At Home. http://sifter.org/~simon/journal/20061211.html 2006

[7] Salakhutdinov, R., Mnih, A. Hinton, G, Restricted Boltzman Machines for Collaborative Filtering, To appear in *Proceedings of the 24th International Conference on Machine Learning 2007*.

[8] http://www.netflixprize.com/community

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

[10] http://mill.ucsd.edu/index.php?page=History