Model Builder for Predictive Analytics & Fair Isaac's Approach to KDD Cup 2003

Joel Carleton, Daragh Hartnett, Joseph Milana, Michinari Momma,
Joseph Sirosh, Gabriela Surpi
Fair Isaac Corporation
3661 Valley Centre Drive
San Diego, CA 92130
gabrielasurpi@fairisaac.com

ABSTRACT

Fair Isaac tackled the third task of KDD Cup 2003 using a predictive modeling approach that leveraged citation graphs, text mining, custom variable creation and linear regression. The core tools we used were embedded in our Model Builder for Predictive Analytics (MBPA) product that makes commercially available a broad set of previously proprietary methodologies used by Fair Isaac for predictive scoring systems such as credit risk and credit card fraud. This short paper reviews the KDD cup problem, our approach, and the toolset. We analyze the predictive variables in the model, the main sources of prediction errors, and the steps that could be taken to alleviate such errors in future work.

1. INTRODUCTION

The third task of the 2003 KDD Cup focused on predicting the number of downloads a paper receives within the first 60 days of submissions to the High Energy Physics - Theory (hep-th) section of the academic e-print archive (www.arxiv.org) hosted by Cornell University and supported by the NSF. As reviewed by the Cup's organizers [1], the physics archive was started in 1991 by Paul Ginsparg and quickly became the dominant mechanism used by the profession for rapidly communicating research results. Unlike formal journals, there is no review process: submitted papers are available world-wide for download the next day.

Researchers are updated of new submissions to the archive in one of two manners: either by directly hopping to the archive's Internet site or by joining the archive's email distribution list. The daily email only includes notifications of papers submitted to the archives specified by the researcher (e.g., "hep-th", "quant-ph", "cond-mat", etc.). Authors can cross-reference submitted papers, so that a paper for example submitted to "gr-qc" could also appear in the listing for "astro-ph". Papers are however uniquely assigned to a specific archive, as reflected by their reference-number. Besides a paper's title and authors, email notifications (& web-page listings) include a paper's abstract, "Comments" (that includes a paper's length as well as any additional explanations the authors wish to provide) and where appropriate, a "Journal-ref" in case the paper has already been published.

The "new" abstracts notification was provided for use for the third task of the 2003 KDD Cup where the data provided reflected the most complete, up-to-date information known by the archive at time of extract (& not the paper's original submission). The latex source of each paper in the hep-th archive was also provided as well as, most importantly, the network of references to each paper from other papers within the hep-th archive (provided as simply a double column of archive reference numbers). Such references overwhelmingly occur well after the first 60 days that follow a paper's submission.

In addition to the above data sources for all papers in the hep-th archive (29104 papers at time of extraction), the download log for the first 60 days of 6 months of papers was provided, the set comprising 2 months of papers from the Spring of the years 2000, 2001 & 2002. The third task of the 2003 KDD Cup was to predict the download volume for papers from a third, adjoining spring month from each of the aforementioned years. Download totals for the 6 tagged months ranged from 22 to an extreme value on one paper of 2927. The average download volume was 193, while the median was 142.

2. TRAINING SET & METHODOLOGY

The set of tagged data comprised a grand total of 1566 papers. Critically, the evaluation metric for the task was the L_1 norm of the error on the top 50 (~20%) downloaded papers from each of the 3 Evaluation months (150 papers in total). While a submission involved a prediction for all papers (as contestants obviously did not know which papers were in the top 20%), 80% of the predictions were ignored. Given this evaluation metric, Fair Isaac's approach was to build focused models using as training exemplars only papers from the tails of the distributions: errors on the remaining 80% were irrelevant and experimentation demonstrated that using them for training indeed biased all results to lower predictions & generated poorer performing models on the papers of relevance. The training thus involved the top 50 download papers for each month plus a few extra per month to help smooth out the borders. We also chose to exclude the aforementioned 2927 extremum (with nearly double the number of downloads of the next most popular paper), leaving a total of 323 papers. As we will see below, the remaining 1242 tagged papers as well as the entire archive of 29104 papers were nevertheless not ignored and indeed played a key role in the generated feature space of predictive characteristics.

Given the size of the training set, cross-validation on (non-overlapping) buckets of 10% was used to evaluate our models. Results quoted below are thus the average results across all 10 hold-out sets (i.e., of the 10 models on each of their respective evaluation sets). After all experimentation, the final model used to generate the submission was trained on all 323 papers.

We used as our baseline performance the “naïve” model of giving all 323 papers their average download value of 441. The average per paper of the L_1 norm of the error of this “naïve” model across the 323 papers was 161.

3. MBPA

All models were trained using Fair Isaac's Model Builder for Predictive Analytics product. Fair Isaac’s Model Builder, part of the Fair Isaac Business Science™ suite of integrated software tools, delivers a comprehensive modeling environment that includes data preparation and analysis, predictive modeling, model validation and model deployment capabilities. Model Builder delivers linear and logistic regression, as well as neural network models as standard components. Model Builder also makes available advanced techniques unique to Fair Isaac. Fair Isaac's Scorecard technology, for instance, is used to build Fair Isaac's industry-standard credit scoring models. Fair Isaac’s patented Context Vector™ technology [2] for modeling unstructured text used, as described below, for feature generation for the KDD Cup submission is scheduled for release as part of Model Builder in 2004.

4. Model Details

Table (1) contains a sample list of feature types explored as input to the predictive models. The first class involves features associated only with the Paper itself. The second class contains features involving single links between papers in the archive. Class 3 leverages the known download history of the Paper’s authors. Here and in classes 4 & 5, weightings were assigned each author by the number of previous publications found in the archive, thereby attempting to distinguish between “established” researchers and their “students”. Classes 4 & 5 involve significantly more complicated network links. Class 4 examines the fame of the paper’s authors using the average number of citations found in the Latex bibliography. A similar metric was used to evaluate the most famous author citing a Paper. Class 5 involves a Context Vector build of the abstracts of all papers in the hep-th archive to classify their content. Nearest neighbor evaluations of the Paper to other papers in the tagged training set are then used to generate download predictions for a Paper. We note that the canonical “authority-hubs” measure [3] of a paper was also explored but found less predictive than the most significant class 2 features.

1	Year submitted; Weekday submitted; Publication Journal; Authors’s email; Number of Pages; Number of Authors; Title_keyword ; Citations to Other archive sites
2	Number of authors the Paper cites; Number hep-th citations received (no self-citations); Number of hep-th authors who cite Paper ; Number hep-th citations to any publication by Authors ; Number hep-th publications by Authors before Paper
3	Weighted download of other papers by Authors; Largest download by any Author
4	Fame of Authors of the Paper ; Fame of most famous author that cites Paper.
5	Similarity prediction of Paper’s Context Vector ; Ranking prediction of Paper’s Context Vector.

Table 1 : Sample list of feature types explored.

A significantly longer list of features was generated using various transformations and metrics on each type (e.g. citations received in 240 days, etc.). The total was reduced through backwards selection using the linear regression module of MBPA. The final submission involved scores from two linear regression models distinguished by whether the Authors had other papers in the training set (class 3 above).

Table 2 provides performance figures for the models on both the training and test sets used for model development, as well as the submitted solution’s results on the 3 months of Evaluation data. The fourth column excludes one extreme paper in the Evaluation set that had 7160 downloads (providing thereby a fairer comparison with columns one & two as it mirrors the exclusion in the development set discussed previously). For the latter set, the average download for the 149 papers was 410.

	Train	Test	Eval	Eval*
$< L_1 >$	85	97	146	104

Table 2: Model results for the development & Evaluation sets

5. Conclusions

Although Table 2 indicates a fairly robust model with fairly consistent results across all data sets, better performance could have undoubtedly been obtained if additional information had been provided. First, the size of the data set, as winnowed by the evaluation metric was ultimately quite small. Over one order of magnitude of relevant data could have been easily extracted using download histories over many years of the archive. Second, the citation data, limited to only other papers within the hep-th archive itself, was severely incomplete. It did not capture the full dynamics of the e-print archive wherein papers thru cross-listing can address a significantly larger community than that of hep-th itself. The last variable of Class 1 of Table 1, “Citations to Other archive sites” (extracted from the paper’s Latex source) was an attempt to proxy this effect but it was clearly not a replacement for the citation-list that could have been generated from the entire archive.

6. ACKNOWLEDGMENTS

We thank the Cup’s organizers for both this invitation as well as for all their efforts designing and assembling this year’s competition.

7. REFERENCES

[1] Ginsparg, P., Gehrke, J. & Kleinberg J., see: http://www.cs.cornell.edu/projects/kddcup/download/KDDCup-Overview.pdf.

[2] Caid, W. & Qing P., “System & Method of Context Vector Generation and Retrieval”, U.S. Patent 5,619,709 (1997).

[3] Kleinberg J., “Authoritative sources in a hyperlinked environment”, Journal of ACM (JACM), 46, 1999.