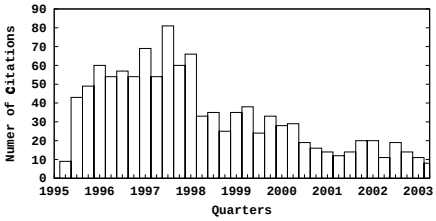
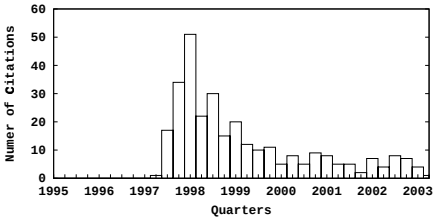
Predicting citation rates for physics papers: Constructing features for an ordered probit model

Claudia Perlich
New York University
Stern School of Business
New York, NY 10012
cperlich@stern.nyu.edu

Foster Provost
New York University
Stern School of Business
New York, NY 10012
fprovost@stern.nyu.edu

Sofus Macskassy
New York University
Stern School of Business
New York, NY 10012
smacskas@stern.nyu.edu





1. INTRODUCTION

Gehrke et al. introduce the citation prediction task in their paper "Overview of the KDD Cup 2003" (in this issue). The objective was to predict the *change* in the number of citations a paper will receive–not the absolute number of citations. There are obvious factors affecting the number of citations including the quality and topic of the paper, and the reputation of the authors. However it is not clear which factors might influence the change in citations between quarters, rendering the construction of predictive features a challenging task. A high quality and timely paper will be cited more often than a lower quality paper, but that does not suggest the change in citation counts. The selection of training data was critical, as the evaluation would only be on papers that received more than 5 citations in the quarter following the submission of results. After considering several modeling approaches, we used a modified version of an ordered probit model[1]. We describe each of these steps in turn.

2. FEATURE CONSTRUCTION

Investigating the citations of a number of papers over time shows some common properties. The two figures show examples of such time series. Both papers show a hump-shaped curve that after a sharp rise shortly after publication decreases and eventually levels off. This suggests time-series analysis as the first focus for our feature construction.

Time-Series Features

In order to capture the temporal shape of the curve we constructed features from lagged citation counts for the last 6 quarters prior to the quarter of interest. Given that the shape is a function of the age (time since publication) of the paper we also included the quarter of publication. One important difference across papers is the scale. The maximum number of citations ever received by a paper was 2400, whereas the average was only 15. Also, not only did the paper in the first graph receive more citations, it also has a different shape. Very popular papers have a longer first period of very active citing and do not show as early or as steep a decrease. This suggests two things: the total number of citations should be included and it might be of use to normalize the raw citation counts. The two examples also show that predicting a particular change between two quarters is a very noisy problem and "pooling" information across papers through normalization might help to reduce the model variance. Other factors beyond the temporal shape of each particular paper are potential structural patterns in the process of document publication. One might for instance suspect the existence of "publishing seasons" (times where more or fewer documents are published) based on the productivity of the authors over the academic year or conference schedules. We therefore also constructed quarterly dummies. This raises concerns with regard to the stationarity of such a seasonal effect over the entire period of 10 years and the timeliness of certain topics. The seasons may differ for different subfields of physics. It might also be the case that particular topics dominate at a particular point in time and draw more publications and thereby more citations of relevant work. It could be useful to induce subfields or topics from either the text or the citation graph.

Topic-based Features

Rather than clustering the full text we extracted a small set of keywords. Given those keywords we constructed for each paper the total change in citations over the most recent quarter over all papers that shared the same keywords, hoping to capture which topics are currently "hot" or "cool".

Relation-based Features

It is likely that an author will publish dominantly in a particular subfield or topic. In addition, the reputation of an author has a strong impact on the number of citations. This becomes important for very recent papers where few (or no) lagged values of citation counts are available. New papers

Year	Q2-Q1	Q3-Q2	Q4-Q3	Q1-Q4
2002	-3	-4	-2	-4
2001	-2	-4	-1	-3
2000	-2	-1	-2	-2

Table 1: Median change in citations between quarters

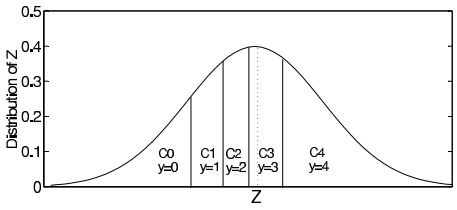
by authors with many well-cited papers can be expected to receive many citations directly after publication, since they are more likely to be read by other scientists. We therefore included the total number of publications of the author, the total number of citations of all papers by the author, the average number of citations per paper by that author, and the typical change in citations of papers by that author after the same number of quarters since publication. That is, if the prediction is to be made for a paper that was published 2 quarters ago we would take the average change over all papers by the same author(s) between the second and third quarter after their publication.

3. SELECTING THE TRAINING DATA

A number of factors affected our selection of training data. Since the evaluation was done on papers with more than 5 citations during the first quarter of 2003 we also limited our training to periods where number of citations in the last quarter was larger than 5. This limits our sample to 3117 unique papers and a total of about 10000 paper-periods (6 quarters) from more than 20000 documents. To decide whether to include all 10 years or to limit the training to only recent periods we analyzed the median change between quarters over the last 3 years yielding the mixed results in table 1. We observe a strong seasonality over the last 2 years; however the year before did not show the same pattern. We used cross-validation to test the effect of curtailing the training periods and concluded to keep all years despite the unstable seasonality.

4. MODEL SELECTION AND ESTIMATION

The task of citation prediction is a specific form of regression where the independent variable is constrained to integer values. We identified three further issues that influenced our choice of model: L1 loss as the evaluation metric, the high degree of noise in this particular task, and the limited training size. We considered as possible model classes linear regression, neural networks, and an ordered probit model that under certain simplifications learns a piecewise linear model. We evaluated all three models using cross-validation. The best generalization performance was achieved by our simplified ordered probit model, closely followed by linear regression. The neural network was consistently outperformed by a constant model that always predicts the median. Both linear regression and ordered probit have a lower variance and are therefore better suited to this noisy task. Disadvantages of the neural networks and the linear model are the minimization of the L2 norm (which results in predicting the expected value rather than the expected median that is optimal under L1) and the prediction of continuous rather than integer values. An ordered probit model [1] on the other hand assumes that the integer values of y reflect ordered categories that correspond to an unobservable continuous variable z. y is assumed to take on a particular category



Ci if the corresponding z was between the two cutoff points Ci_{min} and Ci_{max}. It is furthermore assumed that z is a linear function of the independent variables $X = (x_1, ..., x_n)$. The estimation procedure uses maximum likelihood to estimate the model parameters and the cutoff points under the assumption that the error term is normally distributed. Ordinary ordered probit will predict the most likely category y given X and the estimate of z. However, in the case of a very noisy task the most likely category tends to be one of the outer two, $C0$ or Cn, due to the large variance of the estimate of z. This leads to a large probability mass in the two extreme categories as shown in the figure. The most likely category will therefore not be the one in which the expected value z fell ($C3$ in the figure) but one of the two outermost (e.g., $C4$) due to the larger probability mass when integrating from the last cutoff value to ∞.

To address this problem we decided to predict the category into which the expected value fell. To estimate the model reliably we had to limit the number of categories to 10, assigning all papers with a change of citations of less than -7 to the lowest group and all papers gaining more than 2 into the highest group. Even for the linear and the ordered probit we still observed overfitting. Curtailing the final predictions between -4 and 0 (assigning -4 to all predictions smaller than -4 and 0 to all prediction larger than 0) improved the performance consistently. This effect might also be caused by the discrepancy of predicting the mean rather than the median. The mean tends to be larger in absolute terms in the outer categories than the median, but the median is the optimal prediction under L1 loss.

5. RESULTS

Our final ordered probit model with curtailed predictions scored 1360 on the evaluation set. Without curtailing, the score was 1366. The performance of the linear model would have been 1372 with curtailing and 1380 without. The optimal constant prediction for the task was -2 and would have reached a score of 1403.

6. ACKNOWLEDGMENTS

We would like to thank William Greene for providing us with software for the ordered probit model and valuable advice, and Jeffrey Simonoff for the extensive discussion of the task. This work is sponsored in part by the Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory, Air Force Materiel Command, USAF, under agreement number F30602-01-2-585.

7. REFERENCES

[1] R. Zavoina and W. McElvey. A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, Summer:103–120, 1975.