PROBABILISTIC RECORD MATCHING

ROBERT RAVIV MURCIANO-GOROFF

1. Introduction

A common problem when utilizing multiple datasets from disparate sources is linking observations about the same entity when unique identifiers are unavailable. This challenge arises when Amazon determines if two product listings are for the same item, when Google Books merges library catalog listings, or when hospitals link health records using only patient names and attributes. The process of creating such a bridge, known as record-linkage, is challenging for deterministic programs as typographical variations and data errors can make even matching records appear dissimilar.

One task for which record-linkage has been applied is finding citations to the same publication across multiple bibliometric databases [2]. Citations to articles and books often include similar information, such as author names, publication titles, and publication year. But variations in even these few fields are common. For example, names, such as Steve, Stephen, and Stephen, are often misspelled, while titles from foreign language publications can be translated or transliterated with variations. Even a numeric field such as the publication year for a journal can be recorded differently when, for example, a winter issue of a quarterly publication spans both December and January. These challenges are representative of issues that come up in many other databases, and would require intractably lengthy code enumerating all possible variations to be solved by a deterministic matching program.

Supervised classification methods from machine learning provide a means for implementing probabilistic record-linkage that can both handle these challenges and scale for use on large bibliographic databases. In this paper, I demonstrate and evaluate three classification methods for uniquely matching articles indexed in the PubMed/Medline and Web of Science (WoS) databases. These two databases contain overlapping information, such as article titles, journal names, and author listings. They also contain complimentary data: Medline contains research funding information not found in WoS, while WoS tallies forward citations unavailable in Medline. Thus, in order to analyze the costs and benefits of research work, one must link the corresponding data between these two datasets. Unfortunately, the databases do not have a single identifier for linking articles present in Medline and WoS. Logistic regression, Gaussian Discriminant Analysis, and Support Vector Machine classifiers are evaluated for their ability to predict matching versus non-matching articles. Because of stark differences in the patterns of matched and non-matching article records, logistic regression proved to be the most efficacious. Finally, the effects of missing data and heavily weighted training datasets are discussed as implementation challenges for scaling this classifier.

2. Data

Web of Science (WoS) and Medline are large bibliographic databases of scientific articles. Maintained by the National Library of Medicine at the National Institutes of Health, Medline comprises over 19 million citations for journal articles and books related to the biomedical

Field	Mutual Information Score
Title	0.5763
Volume	0.0729
Issue	0.3700
Begin Page	0.6106
End Page	0.5437

Figure 1. Mutual Information Scores for Used Features

sciences [4]. Web of Science, an online database maintained by Thomson Reuters, contains the articles of approximately 12,000 of the highest impact journals published across a wide variety of scientific fields [3].

The two databases have overlapping coverage, but also subtle differences. While many journals are indexed by both bibliometric products, Medline is not a proper subset of WoS. Thomson Reuters focuses on capturing only high impact and primarily English language journals, while Medline includes a wide variety of international, transliterated, and translated articles. Furthermore, because some articles are indexed using optical character recognition (OCR) technologies on the printed journals, even articles contained in both databases can vary in the spelling of article titles and author names. Additionally, approximately a third of articles have missing data fields in one or both databases.

In order to develop a model for linking publication citations available in both databases, I constructed a training dataset of examples of matched and unmatched articles. From research papers published in the year 2008, I extracted citations with complete article titles and page numbers as well as journal volume, issue, and International Standard Serial Number (ISSN) information. In total, the subset included 200,173 Medline articles and 1,490,842 Web of Science articles. Since the databases do not have a means of linking all matching articles, I only used records from the extract with listed Digital Object Identifier (DOI) numbers, a means of uniquely identifying documents posted online. DOI numbers on scientific publications are not very common. Only 53,655 articles, about a quarter of the Medline articles without missing fields from this publication year, have corresponding articles in the WoS with the same DOI number.

While these articles with corresponding DOI numbers provide examples of matching articles, in order to train a classifier, I also required examples of non-matching articles. By cross-joining matched article pairs appearing in the same journal, I created an additional 599,589 pairs of articles with non-matching DOI numbers. These served as the training examples of non-matched article pairs.

3. Method

The overarching strategy for linking citations was to create similarity profiles of superficially similar articles from the respective databases. Using the similarity profiles of known matches and non-matches based on the DOI numbers, a classifier could then be trained to recognize and predict whether a computed profile was the product of a matched pair of references to the same article or a non-matched pair of references to distinct articles. In particular, using a parametric modeling technique, such as logistic regression, enables this classifier to scale-up for use on large bibliometric databases since relatively few parameters need to be stored.

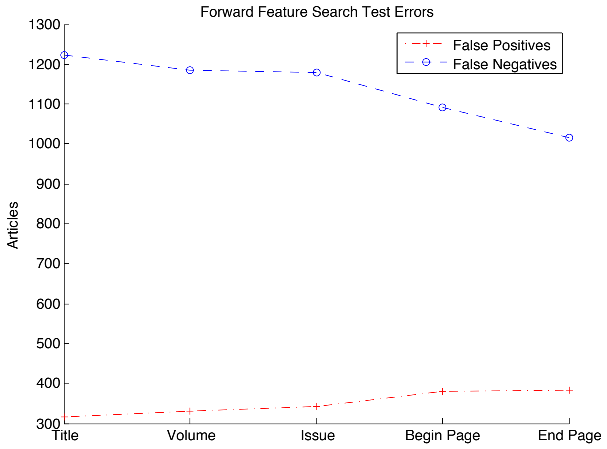

Figure 2. Forward Search

After considerable data cleaning, all available features were extracted for the 653,244 pairs of matched and non-matched articles. Because most fields useful for comparison, such as article titles, issue numbers, and author names are textual data fields, a number of string comparison methods were tested for assigning scores of text similarity. Ultimately, using the Levenshtein distance metric for shorter string fields, Jaccard comparison for longer text fields, and soundex on titles provided discernible advantages during cross-validation.

Prior to running any classifiers, features were pruned using mutual information scoring as well as forward searches. First, the correlation of each available feature with the label of training examples was assessed. Titles showed promise as strongly differentiating examples. The soundex similarity of titles has a 0.4132 correlation coefficient with the match labeling, and indeed, 99.86% of matched articles had a soundex score of 4 for their titles, while only 26.81% of non-matched articles achieved this high value. Volume and issues are a less robust field and contain many variations and typos. Their Levenshtein distances had lower correlations with the labels and provided less discerning information. The distance between the listings for the volume field had a coefficient of a mere -0.1552.

Using logistic regression, I also ran a forward feature search in order to assess the incremental value of each additional comparison metric being included in the model. Most surprising was the seemingly low value to utilizing measures comparing author names between articles. The author names data are replete with differences in spelling. Even after manipulating the author name fields such that Medline's UTF-8 could be compared with WoS's ASCII names as well as using soundex's liberal distance metrics, both the low mutual information scores and forward search cross validations indicated that this field would provide little support to the classification system.

With the features selected, I ran and evaluated three supervised classification systems: Gaussian Discriminant Analysis, logistic regression, and support vector machines (SVM).

4. Results

Because most matched and non-matched pairs have extremely contrasting features, I anticipated that Gaussian Discriminant Analysis (GDA) would provide the strongest results.

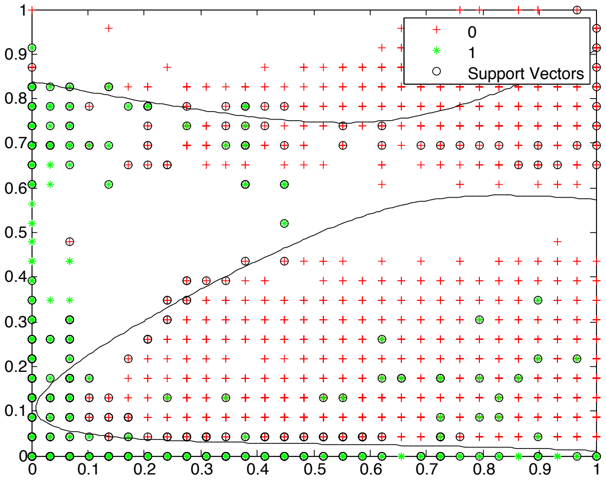

Figure 3. SVM with Polynomial Kernel for Title and End Page features

1043897	3542
155281	94750

Figure 4. Diagnostic Table for Logistic Regress

After implementing a k-fold cross-validation procedure, however, GDA proved to be the lowest performant. On its best iteration, GDA correctly binned 96.39% of all test examples and 93.86% of non-matched articles. While a considerably high proportion of the data, the need to store and provide GDA with lots of training examples makes this less attractive than logistic regression.

On the same data, logistic regression had significantly higher performance. Through k-fold analysis, on each of ten iterations the classifier achieved a rate of correctly categorizing 97.15% of the test examples. The classifier also provided even higher predictive value, 98.45%, for identifying positive matches. Given the speed, efficiency, and parametric nature of logistic regression, this classifier seems apt for applying to citation record-linkage problems.

Since the features of matches and non-matches need not be linearly separable, I also implemented and tested SVM using a quadratic kernel and SMO optimization [1]. Because both GDA and logistic regression had already shown extraordinary accuracy and recall, and because SVM is considerably more computationally expensive, I had hoped for SVM to have significant advantages over the previous two methods. While the correct categorization rates did peak at 97.50% with balanced performance on both matches and non-matches, the data is not separable and the Karush-Kuhn-Tucker (KKT) conditions had to be violated during the optimization stage. Indeed, only by allowing 3% of the examples to violate the KKT conditions would SMO even converge within 15,000 iterations. With the speed and efficaciousness of logistic regression, SVM seemed useful for batch processing of small datasets, but perhaps impractical for larger databases and more frequent matching.

5. Conclusion

Supervised learning as a tool for record-linkage problems on bibliometric datasets is highly effective. By compiling a large number of similarity profiles for both known matched and non-matched citations, even a basic classifier such as logistic regression can provide exceptional accuracy in predicting if previously unseen articles refer to the same document.

When thinking about the applicability of logistic regression and SVM for large scale usage on more years of data or even larger datasets, a number of practical challenges must be dealt with. First, the training examples are heavily biased. Because DOI numbers have been relatively rarely used in the past, the cross-join technique to construct non-matching examples will always heavily weight the training set with profiles of mismatched citation pairs. This bias can cause problems in interpreting the accuracy of a classifier, since even a classifier that defaulted to labeling all test examples as non-matches could still have unusually high performance. Thus, while the optimization techniques of these classifiers still provide excellent accuracy for both matched and non-matched test pairs, their performance must be measured against the baseline distribution of examples in the training and test sets.

Second, when applying these mechanisms to most citation databases, thoughtful consideration will have to be given for how to deal with missing data fields. My tests indicated that performance would not degrade significantly if missing fields were interpolated with the average of non-missing fields.

Finally, after examining the learning curves of these classifiers for increasingly large training set sizes, I found that they potentially suffer from bias, since despite tens of thousands of training examples, the test and training errors remained somewhat apart. An obvious fix would be to provide more or better features for the classifiers. This would likely require further manipulation and cleaning of the original data fields. For example, comments and replies to articles are often published alongside the original scholarly research. Thus, in the database, these comments appear with the same title, journal information, and even page number as the records of the original article. Thus, two records in the database may appear almost identical yet have different DOI numbers. Perhaps these more subtle and problematic cases could be differentiated by isolating a discriminating feature and giving it more weight in the classification model.

Despite these minor concerns, supervised machine learning remains an efficient, highly accurate, and scalable means of performing record-linkage across citation databases.

References

[1] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[2] Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart J. Russell, and Ilya Shpitser. Identity uncertainty and citation matching. In *NIPS*, pages 1401–1408, 2002.

[3] Thomson Reuters. Web of Science Fact Sheet, 2013. URL `http://thomsonreuters.com/products/ip-science/04_064/web-of-science-fs-en.pdf`.

[4] U.S. National Library of Medicine. Medline Fact Sheet, 2013. URL `http://www.nlm.nih.gov/pubs/factsheets/medline.html`.