Citation-Based Document Classification

Claudia Peilich
Stern School of Business
cperich@stern.nyu.edu

Abstract

This paper presents a novel approach to document classification based on authorship and citations rather than the full text of the documents. This task requires the construction of features from relational background tables with variable numbers of authors and citations for each paper. We propose a method for aggregation relying on density estimates and distances, and highlight a number of desirable properties. We finally demonstrate on the CORA domain of scientific publications in Machine Learning the method's superior performance over state-of-the-art text classification approaches and conjecture that the outlined methodology is applicable to a wide variety of classification tasks that could profit from relational background knowledge such as customer relationship management (CRM), fraud detection, and personalization for e-Commerce applications.

Keywords: Aggregation, Density Estimation, Relational Learning, Document Classification

1. Introduction

The classification of large numbers of text documents remains a challenging problem. Although machine learning techniques such as Naïve Bayes have performed surprisingly well on this task (Friedman 97), the manual labeling of a sufficient number of documents for training as well as the storage and processing of a large number of documents remain a major challenge. Classification algorithms typically use the full text and rely on a bag of words as input representation, which can easily exceed 5,000 entries per document. However, most documents, in particular scientific texts, have additional characteristics that are commonly ignored for classification tasks, including authorship and citations of other publications. The reason for this practice is twofold: 1) across a collection of scientific documents the number of authors tends to be large, introducing a high dimensionality if they were included as categorical variables in the classification process; 2) the number of authors and citations varies across papers, which renders a feature-vector representation with a fixed number of independent attributes unsuitable. There have been a number of efforts addressing the problem of hypertext classification (e.g., Chakrabarti et al. 1998), most of which use in addition the original document the text or class label of referenced documents. More generally, domains with one-to-n relationships between entities, as in the case of paper and citations, are referred to as multi-relational, since they require multiple tables (for this domain DOCUMENT, AUTHOR, and CITATION) to represent all relevant information. This work is therefore closely related to the general problem of Relational Data Mining (Lavrac 2001), a field that has grown substantially over the last few years. However, little attention has focused on the role of complex aggregation methods beyond SQL functions like COUNT, MAX, and MIN. The potential of combining logic-based relational learning methods and logistic regression was demonstrated for the particular application of text classification with mixed success by Popescul et al. (2002).

This paper presents a methodology for relational learning for document classification using only citations and authorship rather than the text, addressing both issues of feature construction from high-dimensional categorical values and variable numbers of attributes. Compared to full-text classification, this approach minimizes the space requirements by about 3 orders of magnitude and performs very well given a comparatively small training set. The remainder of the paper is structured as follows. Section 2 provides a general overview of the steps involved in learning from multiple tables. Section 3 presents a methodology for document classification based on target-dependent aggregation using density estimation and discusses some

---

1 This argument holds also for number of words. However there are many more occurrences for every word across documents allowing the reliable estimation of class conditional priors and model parameters.
properties of the approach. Section 4 compares the classification performance of different classification and feature construction methods.

2. Learning from Citations and Author Linkage

A domain of scientific documents can be represented by three tables: the DOCUMENT table containing pairs of document identifiers and the corresponding class labels, the AUTHOR table with pairs of author names and document identifiers, and a REFERENCE table capturing the citations between pairs of documents. The document identifier links objects across the tables: for example, a list of references from a given document can be found by joining the DOCUMENT and REFERENCE tables.

The general problem of learning from multiple tables can be viewed in part as a transformation task, where the objective of the transformation is the construction of meaningful features from the available background relations (here, AUTHOR and REFERENCE). A simple example of such a transformation would be to count the number of authors for each paper. This feature could now be added to the DOCUMENT table and be used for traditional feature-vector learning. The full architecture of an automated learning system for multi-relational domains is shown in Figure 1. The transformation step has to address two subtasks: 1) the identification and exploration of related objects and 2) the compression of sets of related entities into single-valued features. The first task is easily achieved, given a typed database schema and joining on the object identifiers.

![Diagram](image)

**Figure 1: System Architecture**

De Raedt (1998) showed that for most relational databases complete transformation (without loss of information) is possible, but results in an exponentially growing feature space with potentially sparse feature vectors. Large feature spaces as well as sparse features can significantly harm the modeling performance by amplifying multiple comparison problems (Jensen and Cohen 2000) and the variance of the estimator. There have been a number of SQL-based suggestions for incomplete transformation through aggregation including count, mode, min, max, sum and average. However, the mode of categorical variables with large number of possible values introduces significant loss of information. The next section presents a novel approach to document classification based on the aggregation of multi-valued categorical variables that maintains a low dimensionality of the feature space but retains significant amounts of predictive information.

3. Density Estimation and Comparison for Automated Feature Construction

Figure 2 illustrates the two tables DOCUMENT and AUTHOR with a one-to-n mapping through the key PaperId and a classification target Class. The background table AUTHOR can only improve classification performance if the authors of papers in class 1 are somehow different from authors of papers in class 0. Assuming a theoretical framework where the entities in the background tables are drawn from two different distributions $D_{\text{Class }1}$ and $D_{\text{Class }0}$, we follow the methodology as described in detail by Perlich and Provost (2003).
The first aggregation step estimates the class-conditional distributions of authors from the training set in the DOCUMENT table. Supposing an order on the values of the categorical attributes (mapping from authors to a vector position, i: A and 2:B) the value for $D_{\text{Class } n}$ at position i is defined as the class-conditional prior:

$$D_{\text{Class } n}[i] = \frac{\text{Number of occurrences of author } i: X \text{ in authors related to document of class } n}{\text{Number of authors related to documents of class } n}$$

The resulting estimates of the class-conditional distributions for our example are given by:

$D_{\text{Class } 0} = [0.333 \ 0.666]$ and $D_{\text{Class } 1} = [0.666 \ 0.333]$

The second aggregation step is the estimation of the densities for every document:

$$D_{\text{Pn}}[i] = \frac{\text{Number of occurrences of author } i: X \text{ related to the document } P_n}{\text{Number of authors related to document } P_n}$$

The estimates for the document-specific distributions in our example are:

$D_{P1} = [0 \ 1], D_{P2} = [0.5 \ 0.5], D_{P3} = [1 \ 0], \text{ and } D_{P4} = [0.5 \ 0.5]$

The third aggregation step constructs features from the class-conditional densities and the document densities through the application of various vector distance measures including cosine, Euclidean, and Mahalanobis (variance-adjusted Euclidean distance, Mahalanobis 1936) distance. The new target table after adding the Euclidean distances (ED) between the class-conditional densities and the document densities is shown in Figure 3: A simple but effective extension is the construction of the differences between the vector distances $ED(D_{\text{Class } 1, P_n}) - ED(D_{\text{Class } 1, P_n})$ reflecting whether the case is closer to the density estimate of class 1 or class 0.

3.1 Discussion

One can observe a number of interesting properties of the outlined feature construction method:

**Dimensionality Reduction:** The use of vector distances compresses the high-dimensional space of possible categorical values into two (one for each class) dimensions per vector-distance metric. This quality allows the exploration of related entities to a much greater depth without incurring major variance errors during model estimation.

**Discriminative Information Preservation:** The loss of discriminate information is minimal. Significant differences in the class-conditional distributions will be reflected in the vector distances. If indeed the two class distributions are identical, the difference in the distances should be close to zero for all cases and the feature would be discarded during feature selection.

**Efficiency:** The total complexity of the aggregation is $O(n*k*log k)$, where $n$ is the size of the table after the join on ID and $k$ is the number of possible categorical values. The conditional class distribution can already be estimated during the join execution. One additional pass over the resulting table is required to construct the case-specific distributions and distances. The $k*log k$ factor reflects the use of hashes to store intermediate results. But even the estimation of the commonly used mode of a categorical distribution would exhibit the same overall complexity.
Domain Independence: The density estimation does not require any prior knowledge about the application domain and therefore is suitable for a variety of applications beyond text classification.

Monotonic Relationship: The use of differences of vector distances transforms the categorical attribute into a numerical feature that is monotonic in the probability of class membership. This makes logistic regression a natural choice for the model induction step.

Task-Specific Feature Construction: The advantages outlined above are possible because we use the target value during feature construction. This practice requires splitting the training set into two separate portions for 1) the class-conditional density estimation and feature construction and 2) the estimation of the classification model. Having fewer data points for model induction increases the risk of overfitting and motivates feature selection as well as a more biased model category such as logistic regression.

4. Experimental Results

We evaluate the density-based feature construction using our prototype ACORA (Automated Construction of Relational Attributes) on the CORA (McCallum et al. 2000) database. It contains 4200 publications in the field of Machine Learning that are categorized into 7 classes: Rule Learning, Reinforcement Learning, Theory, Neural Networks, Probabilistic Methods, Genetic Algorithms, and Case-Based Reasoning. The domain has 4007 unique authors with an average of 2.1 authors per paper and a total of 90,000 citations between documents. The size of the domain is 179 Kbytes, whereas the original database with the full text had a size of approximately 240 Mbytes.

4.1 Methodology

The target table DOCUMENT was split into 400 entities for training and 3800 entities for testing. All feature construction approaches that take advantage of the target variable require a further split into 200 entities for feature construction and 200 entities for feature selection and model estimation.

Feature Construction: In addition to the presented vector distances between the class-conditional densities, we constructed a number of alternative features including the number of related objects, counts for each of the 5 most common categorical values (an extension of the mode), counts for the 5 most discriminative categorical values (where the differences between entries in $D_{class1}$ and $D_{class2}$ is maximal), and vector distances to the unconditional density ($D_0$) estimated over all training examples. The last feature helps us to evaluate whether the performance of the proposed method is mostly due to the dimensionality reduction or caused by the conditioning on the class label.

Feature Selection, Model Estimation, and Bagging: Following the feature construction, we select randomly 12 features from the extended target table (12 was an arbitrary choice based on the rule of thumb that the number of observations over the number of parameters should be larger than 10; varying the number of features between 8 and 20 had little effect on the reported results), where the probability of selection is proportional to the AUC (Area under the Receiver Operating Curve, Bradley 1997) of a linear logistic classification model estimated on the particular feature. We iterate the random selection of 12 features 10 times and average the results from the 10 different models for the final prediction. We are not aware that this combined approach of feature selection and model bagging has been used before in classification and compare it to model estimation without feature selection or bagging.

Evaluation: We learn separate binary classification models for each of the 7 classes and predict the final class with the highest probability score across the 7 model predictions. We include in addition to the different feature construction methods three alternative classification approaches: a Naïve Bayes classifier using the full text learned by the Rainbow (McCallum 1996) system, a Probabilistic Relational Model (PRM, Koller and Pfeffer 98) with the results reported by Taskar et al. (2001), and a Simple Relational Classifier (SRC, Maeskassy and Provost 2003) that uses the known class labels of related documents. The latter method is closely related to Chakrabarti's work, which draws from the theoretical framework of Markov Random Fields (MRF). The Simple Relational Classifier iteratively propagates the evidence of known class labels of related papers under the assumption that documents from a particular scientific field will dominantly cite previously published papers in the same field.
<table>
<thead>
<tr>
<th>Method Description</th>
<th>Logistic Regression</th>
<th>Decision Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Feature Selection</td>
<td>All Features</td>
</tr>
<tr>
<td>1) 5 most common values</td>
<td>0.60 (0.03)</td>
<td>0.55 (0.03)</td>
</tr>
<tr>
<td>2) 5 most discriminate values</td>
<td>0.68 (0.02)</td>
<td>0.57 (0.01)</td>
</tr>
<tr>
<td>3) Unconditional densities</td>
<td>0.70 (0.05)</td>
<td>0.67 (0.04)</td>
</tr>
<tr>
<td>4) Conditional densities</td>
<td>0.78 (0.01)</td>
<td>0.61 (0.07)</td>
</tr>
<tr>
<td>5) Conditional densities &amp; 5 most discriminate values</td>
<td>0.81 (0.01)</td>
<td>0.63 (0.07)</td>
</tr>
<tr>
<td>6) SRC using related labels</td>
<td>0.68 (0.01)</td>
<td></td>
</tr>
<tr>
<td>7) Naïve Bayes on full text</td>
<td>0.74 (0.03)</td>
<td></td>
</tr>
<tr>
<td>8) PRM on text and citations</td>
<td>0.74^2 (0.01)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Classification Results

4.2 Classification Results

Table 1 shows the performance of the variants of feature construction and the 3 alternative classification algorithms in terms of classification accuracy (and the standard deviations across 10 experiments in parentheses). Row 5 combines the two target-dependent feature construction methods, 2 and 4. We additionally compare logistic regression and decision tree induction (C4.5, Quinlan 1993), as well as feature selection with bagging and no feature selection for both model classes.

**Logistic Regression with feature selection:** Focusing on our main results in the second column of Table 1, we conclude that the proposed feature construction approach using class-conditional density estimation (4 and 5) outperforms both alternative feature construction methods (1-3) and alternative text-based and relational classifiers (6-8). Using feature selection and all target-dependent construction methods (conditional densities and discriminative values) achieves the overall highest accuracy with 81%. The unconditional density distances (3) perform better than the 5 most common values (1). A similar increase in performance can be observed between the conditional densities (4) and the 5 most discriminative values (2). We conclude that the dimensionality reduction contributes significantly to the improved generalization performance. Both target-independent feature construction methods (5 most common and unconditional densities) have a lower performance than any other method in combination. This highlights once more the importance of task-specific feature construction.

**Feature Selection:** Comparing the results between columns two and three as well as four and five demonstrates the need for feature selection for both, logistic regression and tree induction. The model estimation was based on only 200 examples and limiting each model to 12 independent variables improved the generalization performance in all cases. The improvements are larger for logistic regression, consistent with the ability of trees to select a subset of relevant features.

**Model Estimation:** With feature selection, logistic regression performed never worse than the decision tree and for class-conditional feature construction, the performances are identical. The reasons for the superiority of a linear model are twofold: 1) as argued before, density distances and class probabilities should exhibit a monotonic relationship, justifying a linear bias 2) tree induction is more prone to variance errors, in particular for the feature construction methods 1-3, where the task remains noisy due to the low information content of the features. Results by Perlich et al. (2003) suggest superior performance of logistic regression over tree induction on noisy domains. This also explains the relatively high performance of tree induction over logistic regression without feature selection in experiments 4 and 5: given increasingly more discriminative features the task becomes less noisy and the relative performance of tree induction improves.

---

^2 We did not estimate the results but report the accuracies from Taskar et al. (2001). We compared the datasets and verified the number of papers and classes. However, number of authors in their case was only 1500 since they eliminated authors with fewer than two papers.
5. Conclusion and Future Work

This work applied a general purpose aggregation method for automated feature construction based on density estimates from relational background tables to a document classification task. The proposed method in combination with feature selection showed superior performance using only authorship and citations over a variety of alternative methods for feature construction, text-based, and relational classification. The outlined methodology is applicable to a wide variety of classification tasks that can profit from relational background knowledge, such as customer relationship management (CRM), fraud detection, and personalization for e-Commerce applications. Many domains collect large amounts of transaction and interaction data but so far lack a reliable and automated mechanism for feature construction to support decision making. Density estimation in combination with feature selection has the potential to fill this gap and allow the seamless integration of model estimation on top of existing relational databases relieving the analyst from the manual and time-consuming task of feature construction. Future work will extend the proposed method to regression tasks where no class label but only a continuous variable is available for task-specific feature construction.

Acknowledgements

I thank Foster Provost for countless discussions that helped to develop and clarify the ideas in this work, Sofus Macskassy for providing the comparative performance of Na"ive Bayes, and Ben Taskar and Andrew McCallum for the CORA dataset. This work was sponsored in part by the Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory, Air Force Materiel Command, USAF, under agreement number F30602-01-585.

References:


