Risk Analysis for Intellectual Property Litigation

Mihai Surdeanu
Lex Machina, Inc.
mihai@lexmachina.com

Ramesh Nallapati
Lex Machina, Inc.
nmramesh@lexmachina.com

Christopher D. Manning
Stanford University
manning@stanford.edu

ABSTRACT

We introduce the problem of risk analysis for Intellectual Property (IP) lawsuits. More specifically, we focus on estimating the risk for participating parties using solely prior factors, i.e., historical and concurrent behavior of the entities involved in the case. This work represents a first step towards building a comprehensive legal risk assessment system for parties involved in litigation. This technology will allow parties to optimize their case parameters to minimize their own risk, or to settle disputes out of court and thereby ease the burden on the judicial system. In addition, it will also help U.S. courts detect and fix any inherent biases in the system.

We model risk estimation as a relational classification problem using conditional random fields to jointly estimate the risks of concurrent cases. We evaluate our model on data collected by the Stanford Intellectual Property Litigation Clearinghouse, which consists of over 4,200 IP lawsuits filed across 88 U.S. federal districts and ranging over 8 years, probably the largest legal data set reported in data mining research. Despite being agnostic to the merits of the case, our best model achieves a classification accuracy of 64%, 22% (relative) higher than the majority-class baseline. We also perform extensive feature analysis, which identifies the most relevant prior factors for risk estimation.

1. INTRODUCTION

Intellectual Property (IP) law handles legal property rights over creations of the mind, such as industrial, literary and other artistic works. Enforcing IP rights is crucial for many industries where the cost of replicating a product is significantly smaller than the cost of creating that good, e.g., the pharmaceutical, software, or music industries. Litigation is begun when a perceived IP violation is detected; however, many cases are also filed opportunistically or to create impediments to businesses.

IP lawsuits may have drastic consequences. For example, in a patent infringement case in the pharmaceutical industry, the damages paid by the defendant may run into millions or even billions of dollars if infringement is proved. Further, if an injunction is granted by the judge, the defendant is forced to shut down manufacturing and sales of products related to the technology of the infringed patent. On the other hand, if the infringement case is denied, the plaintiff may be forced to pay significant amounts towards the attorney fee of the defendants. More importantly, if the patent under purported infringement is found to be invalid, the plaintiff may fail to recover all the investment made in inventing the patent's technology. Hence, it is a matter of critical importance for parties involved in an IP lawsuit to continually assess their respective risks during the entire progression of the case, starting from its filing time, or even prior to filing.

Understanding the risk in a given lawsuit will help the parties decide whether to continue to fight the case in the hope of a favorable jury verdict, or to settle out of court, thus avoiding large litigation expenses and maintaining the possibility to negotiate the outcome. As a desirable side-effect, if most parties settle cases out of court using an accurate risk analysis system, it would also help ease the burden on the judicial system and make it more efficient.

The main goal of our research is to build a risk analysis model for IP litigation. Risk analysis is a complex problem that depends on a number of factors. Broadly, we classify most factors into one of two categories:

- Merits of the case: In a patent infringement case, factors in this category may include strength of the patent involved, similarity of the defendant's manufacturing technology to the patent's technology, etc.
- Prior factors: Factors in this category do not model directly the merits of the case but instead focus on past information that may influence the outcome of the current case. For example, this category includes the past win rates of parties, attorneys and law firms involved in the case, potential biases of judges estimated from past cases, etc.

While both these classes of factors are important for risk estimation, in this paper we focus on the latter category. More specifically, this work raises the following questions:

1. Are prior factors relevant in determining the risks involved in a case?
2. What are the important prior factors and what is their relative significance in estimating the overall risk of a case?

Focusing on prior factors is important for several reasons. Firstly, prior factors model ingredients that may influence the case outcome independently of the case merits, e.g., a good attorney may increase the chances of winning. Understanding what these factors are provides practical feedback for risk minimization: at the time of litigation, parties cannot change the merits of the case but can adjust prior factors to optimize chances of winning. Secondly, prior factors indirectly model the merits of the case. For example, the fact...
that a plaintiff has a high success rate in previous litigation is a likely indication that it owns strong patents. Hence, prior factors may prove useful also for modeling the merits of cases. Thirdly, risk analysis based on prior factors captures potential biases of the court decision process, e.g., some districts may favor certain parties, thus assessing the objectivity of the judicial system. And lastly, prior factors are easier to extract than the merits of the case which may require more sophisticated modeling.\footnote{See section 5 for more discussion on merit based factors.}

In this work we answer the above questions empirically using a prior risk analysis model, i.e., a model that uses only prior factors for estimation, as follows.

1. To answer the first question on the influence of prior factors for the estimation of litigation risk, we model prior risk as a binary classification task, i.e., patent-owner (plaintiff) wins or defendant wins. The risk for any party can be estimated in terms of the probability that the opposing party wins. In the prediction model we use only prior factors as features, i.e., prior performance of the entities involved in the case, which can be extracted at the time the corresponding case was filed. We train and evaluate our model using data provided by the Stanford Intellectual Property Litigation Clearinghouse (IPLC) project\footnote{http://www.law.stanford.edu/program/centers/iplc/}, which covers IP litigation from the United States from the past eight years. Our results indicate that such models have a prediction accuracy of approximately 64\%, which is significantly higher than the majority-class baseline. To our knowledge, this work is the first to show that litigation risk can be estimated using solely prior information.

2. To answer the second question on the relative importance of each feature, we performed an extensive feature analysis of our model. This analysis indicates that the prior performance of the attorneys and law firms involved in the case is a crucial feature for prediction. We also show that, overall, some districts and judges are indeed biased. Furthermore, we show that there is significant correlation between concurrent cases that share parties. We exploit this correlation using statistical relational learning to further improve the predictive power of our model.

The paper is organized as follows. Section 2 introduces our approach. Section 3 summarizes our corpus of lawsuits and the pre-processing required to construct it. Section 4 describes our experiments. Section 5 discusses current limitations of our approach and future extensions. Section 6 surveys related work and Section 7 concludes the paper.

2. APPROACH

We model the prior risk analysis problem as a supervised discriminative binary classification task whose goal is to predict the outcome of new litigation given relevant prior factors. In this paper we focus on patent-infringement cases that had a publicly known polarized outcome, i.e., cases that were not settled. We model each lawsuit as a separate datum that is assigned one of two possible labels: patent owner wins or accused defendant wins. For simplicity, in the remainder of the paper we will refer to the patent owner as the plaintiff.\footnote{This is not always the case: in declaratory judgment cases the accused party initiates a non-infringement lawsuit, hence the patent owner is legally the defendant. This happens when a party is threatened with a plaintiff-based lawsuit by the patent owner but that lawsuit is not yet filed. In this paper, we normalize these notations to their semantic interpretation, i.e., the patent owner is always the alleged prejudiced party or the plaintiff, regardless of who filed the first complaint.} We introduce next the models we built for this task.

### 2.1 Models

Each individual lawsuit is modeled as a distinct example $x_i$ in our dataset containing $n$ examples. An example $x$ is represented as a vector of features $f(x) = (f_1(x), \ldots, f_m(x))$ where $m$ is the number of distinct features. In the simplest representation, features are extracted using solely historical information, i.e., information extracted from cases that terminated before the case to be modeled was filed. Each component $f_j(x)$ could be a binary value, an integer, or a real valued fraction. To handle such arbitrary feature types, discriminative classifiers are generally considered much better than generative classifiers\footnote{\cite{Manning:2008}}. Hence, we use the $L2$ regularized logistic regression with the following objective function as the classifier:

$$P(y|x, w, \sigma^2) = \frac{\prod_{i=1}^{n} \exp(y_i w^T f(x_i))}{1 + \exp(w^T f(x_i))} \frac{\exp(-\frac{w^T w}{\sigma^2})}{\sqrt{2\pi}\sigma^2}$$

where $y \in \{1, 0\}$ represents the outcome label of the case, $w$ is a vector of feature specific weights, and $\sigma^2$ is a regularization parameter that can be tuned.

A limitation of the above model is that it cannot represent dependencies between concurrent\footnote{\cite{Manning:2008}} and correlated cases. Intuitively, concurrent cases that share information should have a correlated outcome. For example, two concurrent cases with the same company as plaintiff are likely to be on the same topic, i.e., infringement of related patents, use the same evidence and, hence, have the same outcome. This assumption is verified empirically in Table 1, which analyzes the correlation between concurrent cases in our training corpus (the corpus is detailed in the next section). The first row in the table shows the percentage of cases in the training corpus that share one or more entities in the given role with at least another concurrent case. Note that these values are generally high because there is significant chronological overlap between cases: because the average time to termination of an IP case is larger than one year, virtually every case in our corpus overlaps with other cases. The data indicates that sharing is most prevalent for law firms and least common for parties, e.g., 90\% of cases have one or more concurrent cases with the same law firm in the same role but only 36\% of cases share a party in the same role with another concurrent lawsuit. This is to be expected because law firms participate in infringements with an infringement lawsuit by the patent owner but that lawsuit is not yet filed. In this paper, we normalize these notations to their semantic interpretation, i.e., the patent owner is always the alleged prejudiced party or the plaintiff, regardless of who filed the first complaint.

### Table 1: Analysis of concurrent lawsuits

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Party</th>
<th>Attorney</th>
<th>Law firms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same outcome</td>
<td>36.1%</td>
<td>65.4%</td>
<td>90.5%</td>
</tr>
<tr>
<td>80.9%</td>
<td>69.5%</td>
<td>61.4%</td>
<td></td>
</tr>
</tbody>
</table>

We conclude this section with the description of the features that implement the prior factors in our models.

### References

\cite{Manning:2008}
The parties involved in the case, past win rates in any role, unique identifier, and attorneys have several clients, they will work on many different cases that are not necessarily correlated. The data shows that this is less likely for parties.

The above analysis suggests that there is merit in modeling the correlation between concurrent cases, especially for cases that share a party in an identical role. However, this correlation does not always hold: 20% of the concurrent cases that share parties have different outcomes. So, imposing a hard constraint at runtime that any concurrent cases must have the same outcome would be a bad idea. Instead, we prefer to learn soft constraints that estimate the strength of correlation between concurrent cases using statistical relational learning. In this paper, we use conditional random fields (CRF), implemented over the network of concurrent lawsuits. To the best of our knowledge, this work is the first to show that statistical relational learning is applicable to the problem of litigation risk analysis.

Formally, we first define a graph \( G = (V, E) \) as follows. Each vertex \( v_i \in V \) corresponds to a case \( x_i \) in the data. We also define an edge \( e_{ij} \in E \) between vertices \( v_i \) and \( v_j \) if and only if \( x_i \) and \( x_j \) are concurrent cases that share at least one party in the same role (i.e., the shared entity occurs either as plaintiff or as defendant in both cases). We now define a CRF over the graph \( G \) that jointly models the outcomes \( y = (y_1, \ldots, y_n) \) of all cases \( x = (x_1, \ldots, x_n) \) as follows:

\[
P(y|x) = \frac{1}{Z} \exp(\sum_{i=1}^{n} y_i w^T f(x_i) + \sum_{(i,j) \in E} v^T g(y_i, y_j))
\]

where \( g = (g_0, g_1) \) are our new network features, defined as \( g_0(y_i, y_j) = 1 \) if \( y_i = y_j \) and 0 otherwise, and \( g_1(y_i, y_j) = 1 - g_0(y_i, y_j) \). In other words, \( g_0 \) is active if the two cases \( x_i \) and \( x_j \) have the same outcome and \( g_1 \) is active if they have opposite outcomes. The weight vector \( v \) corresponding to these features captures the network correlation strength.

Learning

Although this model is attractive for the problem of joint modeling of concurrent cases, exact learning is intractable for an arbitrary graph such as the one we defined in our problem. Hence, in this paper we use a variant of pseudo-likelihood for training [3]. Pseudo-likelihood is known to be a consistent estimator of true likelihood and is known to work well in cases where local features are strong [10, 15]. In this method, the joint likelihood of all the variables in a model is approximated by the product of the probability of each variable, conditioned on all other variables as shown below:

\[
P(y|x) \approx \prod_i P(y_i|y_{-i}, x_i) = \prod_i P(y_i|y_{\text{N}(i)}, x_i)
\]

where the subscript \(-i\) refers to all variables not including \( y_i \), and \( \text{N}(i) \) refers to the neighbors of \( y_i \). The second step above follows from the fact that a variable is conditionally independent of all other variables given its neighbors, in an undirected graphical model.

In the case of an exponential model such as the CRF, each term in the product above would be equal to the following logistic regression function:

\[
P(y_i|x_i, y_{\text{N}(i)}) = \frac{1}{Z} \exp(y_i w^T f(x_i) + \sum_{j \in \text{N}(i)} v^T g(y_i, y_j))
\]

This is much easier to learn than the joint model because it requires no global information propagation.

Inference:

Since exact inference is computationally expensive as well, we use Gibbs sampling [1] to perform approximate inference, since it has many interesting parallels with pseudo-likelihood. Like pseudo-likelihood, Gibbs sampling deals with the same local probability \( P(y_i|x_i, y_{\text{N}(i)}) \) shown in the last equation. In this approach, we sample each variable \( y_i \) in turn from the local probability, where \( y_{\text{N}(i)} \) correspond to the latest outcome assignments of its neighbors. This iterative process, when run long enough, is guaranteed to converge to the true posterior.

Also, since we need best variable assignments rather than true posterior, we use simulated annealing with Gibbs sampling, using a linear cooling schedule, as used in [5]. According to this approach, we exponentiate the sampling distribution in the last equation by a value \( 1/C \). We initialize \( C = 1 \) at the start of Gibbs iterations, and we decrease it linearly with each iteration until \( C \rightarrow 0 \). At small values of \( C \), the sampling distribution becomes peaked at the maximizing value, thus returning the maximal assignments.

2.2 Features

We model the past behavior of all litigation entities involved in a given case. These entities are:

- The parties involved in the case, i.e., plaintiffs and defendants. Note that, in general, there are multiple plaintiffs and multiple defendants in each IP case.
- The attorneys on each side of the lawsuit, i.e., plaintiffs’ and defendants’ attorneys. Similarly to the party situation, there may be multiple attorneys on both the plaintiff and defendant side.
- The plaintiffs’ and defendants’ law firms. Again, it is common to have multiple law firms on either side of a lawsuit.
- The judges assigned to a case. Usually, there is a single judge per case, but a lawsuit may eventually have multiple judges if it is transferred to a different district.
- The districts where the case is filed. A case will commonly be filed and will terminate in a single district, but it may change its district if transferred.

It is obvious that parties, attorneys, or law firms are important parameters of a lawsuit. Judges are also important because they decide what information is presented to the jury. Hence, their own bias may indirectly influence jury decisions. The district is important because the jury is selected from the local population, which may have certain cultural biases (either pro or against IP).

We model the past behavior of the entities listed above using four different types of features:

- Unique identifier (id) – for each participating entity in a given case we generate a Boolean feature for its unique identifier concatenated with its role in the case. For example, the feature id:plaintiff-attorney:101 indicates that the entity with id 101 served as a plaintiff’s attorney in the current case. This allows the discriminative model to learn the correlation between outcome labels and entities in a given role. For example, if the previous feature appears mostly in lawsuits where the plaintiff side wins, this indicates that the corresponding attorney is usually successful when defending the plaintiff side.
- Past win rates in any role (wR) – we also model the win rates of lawsuit entities, regardless of their role in past lawsuits. We compute it explicitly for each entity as the percentage of past cases, i.e., cases that terminated before the current case and included the given entity in any role, that were won by
the side of the corresponding entity. For example, if an entity won a past case as plaintiff, another as defendant, and lost one as plaintiff, its win rate is 66%. Since there are multiple entities that have the same role in a given case, we average all win rates for the entities in a given role and use this average as the actual feature for that role. These two feature sets (id and wr) are intended to capture both the dependency between previous successes and the current outcome (e.g., a better law firm should increase chances of winning) but also to model, albeit indirectly, the merits of the case (e.g., a party with a high success rate in previous litigation is likely to own strong IP, which should be reflected in the outcome of the current case as well). For obvious reasons, these features are not computed for judges and districts.

- Judge and district bias (bias) - for judges and districts we compute a variant of the wr feature that estimates the bias of that judge or district towards one side of the litigation. This feature is computed as the ratio of cases won by the plaintiff from the set of past cases assigned to the corresponding judge or district. If this value is significantly larger than 50% (or significantly smaller than 50%) that judge or district is biased towards the plaintiff (or defendant) side. Same as above, we use as the actual feature value the average over all judges or districts assigned to the case.

- Counts of participation in past cases in any role (count) - this feature counts the presence of the corresponding entity in past lawsuits, regardless of its role and the outcome of that case. For example, the value of this feature for the entity in the above wr example is 3, because this entity participated in three past cases. This feature serves as an estimate of experience. In other words, the amount of litigation experience that an entity has is likely to be correlated with the number of cases in which it has participated. Same as the wr feature, this feature is averaged over all entities of same type. and is not computed for judges and districts.

We computed the values of all non-boolean feature types, e.g., wr, bias, and count, using only historical information, i.e., information gathered from lawsuits that terminated before the current case was filed.\(^6\) Barring the exceptions noted above, we generate all combinations of feature types and entity types. This yields a total of 22 distinct feature groups. Throughout the paper we will identify these features through a concatenation of feature type and entity type, e.g., wr:plaintiff or count:defendant-law-firm.

3. DATA

The data used in this paper was provided by the Stanford IPLC project. The corpus consists of all IP lawsuits between beginning of 2001 and end of 2008. The cases in the corpus were previously annotated with their outcomes. The annotation process followed a pipeline model: first, two IP experts generated the initial annotations; second, an IP attorney reviewed all outcomes and decided the final annotation.

The meta data available for each lawsuit stores filing and termination times, and the names of all the entities involved (parties, attorneys, law firms, districts and judges). However, all names in the corpus are just textual mentions that maintain the spelling used by the person who filed or registered the transfer of the corresponding case. To transform this data into usable information we implemented an entity resolution (ER) component that consolidates all entity mentions into a set of clusters, where each cluster contains all mentions that point to the same real-world entity. We describe the entity resolution component next and conclude this section with statistics of the processed corpus.

3.1 Entity Resolution

The ER component is a rule-based system that implements a two-step architecture: first, all names are normalized, and second, entity mentions are clustered based on the information extracted during normalization. The normalization process starts by removing common prefixes (e.g., titles for person names) and suffixes (e.g., company name suffixes such as “Ltd.”) from names. We used a total of 141 regular expressions for this step. Next, some common terms in organization names are converted to a normalized form, e.g., both “Holding” and “Holdings” are changed to “Hldg.”. We used 28 regular expressions for this conversion step. During this process we also extract hints about the type of each mention, e.g., for company names we extract the middle name (if present) to initial; for judge names we remove specific titles such as “magistrate judge”.

Finally, we cluster entity mentions using a simple graph-based algorithm. The algorithm creates the entity-mention graph with every mention as a different node and edges between compatible mentions. Each connected subgraph is then assigned a unique id, which becomes the identifier of the corresponding entity. We detect compatible mentions using two different heuristics, depending on mention type:

- For all types other than law firm, two mentions are compatible if they have the same normalized form and the two types are either identical or one is a hypernym of the other in the type taxonomy.
- For law firm mentions, we require that at least two tokens in each of the corresponding names be equal (or have significant overlap), and one of these tokens be the first token in each name. This heuristic is needed because law firms are generally partnerships with dynamic structures and names. While the first partner does not usually change in a law firm name, it is very common that newer partners are added in time or that some leave, which leads to many variations of the law firm’s name. For example, the “Quinn Emanuel, LLP” law firm has 89 different spellings in our database, e.g.: “Quinn Emanuel”, “Quinn Emanuel et al.”, “Quinn Emanuel Urquhart”, “Quinn Emanuel Urquhart Oliver & Hedges, LLP”, etc.

We evaluated the accuracy of the ER component for each of the four top nodes in the type taxonomy. As evaluation metric we used cluster accuracy coupled with a strict definition of cluster correctness: a cluster was evaluated as correct if it contained all mentions of the corresponding entity and no mentions of other legal entities.
Figure 1: Taxonomy of legal entity types. The categories in italicized font (root, party) are abstract types with no actual instances. The organization category is assigned to party names that could not be classified into one of the other known party types.

Table 2: Summary statistics for the litigation corpus. Only terminated cases with a polarized outcome are considered.

<table>
<thead>
<tr>
<th>Cases</th>
<th>Parties</th>
<th>Attorneys</th>
<th>Law firms</th>
<th>Judges</th>
<th>Districts</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,263</td>
<td>12,270</td>
<td>15,706</td>
<td>5,261</td>
<td>1276</td>
<td>88</td>
</tr>
</tbody>
</table>

We computed accuracy using 100 randomly selected clusters for each type. The accuracy values were 91% for party, 92% for law firms, and over 95% for attorneys and judges. It is important to note that, while these are good results, they are not perfect. Hence there is an amount of noise in all the features in our prediction model because they all rely on ER.

3.2 Summary Statistics

There are 20,980 annotated cases in our corpus. From this set we discarded 16,666 cases that settled, i.e., the outcome is not (publicly) in favor of one of the sides, and 51 cases that had outcomes in favor of both sides. The latter situation happens because outcomes are assigned to lawsuit claims, rather than the entire case. Since a case may have multiple claims, e.g., infringement of several patents, it is possible (but uncommon) that different claims have different outcomes. Because this work treats cases as indivisible units, we ignore these situations. The remaining 4,263 cases form the corpus used in this paper.

Table 2 shows the overall statistics for the 4,263 cases. The table shows that, even though the number of cases is relatively small, the corpus contains a significant number of entities, e.g., there are more than 12,000 distinct parties involved. Table 3 shows the relevant statistics per role (plaintiff or defendant). Note that the numbers per entity type (party, attorney, or law firm) do not add to the values reported in Table 2 because some entities appear in different roles in different cases, hence they are counted twice in this table.

Table 3 indicates that there are significantly more entities on the defendant side than on the plaintiff side. This happens because it is common that one or a small number of plaintiffs sue a large number of defendants for infringement of the same patent or patents.

Because the core of our prediction model is based on historical information we train and test our approach on cases where at least one out of the six polarized entities (i.e., party, attorney, or law firm on the plaintiff or defendant side) has participated in three or more previous lawsuits. We choose a minimal history of three previous cases because this allows the model to learn unambiguous information about past success rates. This reduces our corpus to 3,243 cases, or 76% of the original corpus.

To construct the training and testing partitions we sort all cases in chronological order of their termination date and reserve the first 70% (2,270 cases) for training and the remaining 30% (973 cases) for testing. Note that other common evaluation setups that are independent of temporal constraints, e.g., cross validation or random splitting, are not possible for this problem because in such configurations we would train on cases whose outcomes are decided in the future, relative to the testing corpus.

4. EXPERIMENTS

4.1 Overall Results

Table 4 shows overall results for three models: a baseline that predicts the majority class (plaintiff wins in our corpus), the logistic regression (LR) model that uses only historical features, i.e., Equation 1, and the CRF model that models concurrent cases jointly, i.e., Equation 2. For both models we report average results over 100 different samples of the test corpus, generated with bootstrap resampling. We compute statistical significance using two-tailed paired t-test at 99% confidence interval on these 100 samples. The table shows that both our models outperform significantly the proposed baseline. Our best model (CRF), improves the baseline with over 22 relative percentage points. These results provide empirical proof that prior factors influence the outcome of IP litigation. To our knowledge, this work is the first to show that this is possible.

Table 4 also shows that capturing the correlation between concurrent cases is beneficial. The CRF model that incorporates these correlations has statistically-significant improvements over the LR model that exploits only historical information. We hypothesize that this is caused by the repetition of similar evidence in correlated litigation. For example, a plaintiff party that participates in two concurrent cases on the infringement of the same patents, will provide the same evidence in both lawsuits. Hence, it is very likely that these lawsuits will have the same outcome. Repeated information likely appears in other similar (but not identical) cases as well, e.g., different patents owned by the same party. Such correlations are modeled implicitly in our approach, through the sharing of litigating entities between cases.

4.2 Ablation Analysis

To understand the important factors in our prediction model, we performed several post-hoc ablation experiments using our best CRF model.

4.2.1 Ablation Analysis for Entity Types

Figure 2 shows the results of an ablation experiment where feature groups are defined around entity types, e.g., we group all the features about plaintiff’s attorneys into a single set. The experiment
measures the accuracy of the CRF model as each one of these feature groups is removed in turn from the complete feature space. The figure shows the difference in accuracy scores between the system without the corresponding feature group and the best system that uses all features. Hence, negative values in the figure indicate that the corresponding feature group is useful. We draw several conclusions from this experiment:

(a) Legal entities, i.e., attorneys and law firms, are crucial. For example, removing the plaintiff party does not yield a statistically-significant change in performance, but eliminating the plaintiff’s law firm causes a drop in accuracy of almost 2 percentage points. On the defendant side, the combined contribution of the defendant’s law firm and attorney equals that of the defendant party. Three out of four of the legal entities (law firms on both sides and the defendant’s attorney) have a significant impact, with drops in accuracy ranging between 0.5 and almost 2 percentage points. We hypothesize that the strong signal of the legal entity features is caused by the fact that some legal entities have a very consistent track record because they tend to specialize in similar types of cases. On the other hand, the same cannot be said about parties, e.g., a company (especially firms that specialize on opportunistic IP enforcement) may sue on many different IP issues. The latter issue does not apply to defendant parties, which explains why features extracted from defendant parties have a significant contribution.

(b) The figure shows that features extracted from judge and district entities have a significant influence on performance. This suggests that there is a significant correlation between some judges and districts and certain lawsuit outcomes (i.e., plaintiff winning or defendant winning). More formally, the bias of these entities cannot be explained by the mean distribution for that corresponding type, as given by the training data. Using this observation, we identify these entities in Section 4.3 using Fisher’s exact test [6]. However, this correlation does not necessarily imply causation, e.g., a judge intentionally tilting the case towards a specific party. There may be other factors at play. We discuss these issues in Section 5.

(c) The ablation experiment indicates that features extracted from the plaintiff party actually have a negative performance impact. The value measured is not statistically significant, but, nevertheless, it illustrates the danger of forecasting: assuming future behavior based on past events may not always work. In our work, we address this risk by “hedgeg our bets” on eight different entities. As shown in our experiments, this strategy is successful.

4.2.2 Ablation Analysis for Feature Types

We performed a second ablation experiment, this time generating feature groups based on feature types instead of entity types. We followed the same procedure as in the previous experiment. The results are summarized in Figure 3. This experiment illustrates two additional important facts:

(d) Surprisingly, the count feature, i.e., the number of past cases that had the corresponding entity as participant, is a better predictor of future behavior than the wr or id features, which capture the success rates of entities. Our conjecture is that count is a better estimate of experience, i.e., the more cases an entity had participated in, the more experience that entity has. An additional factor in favor of the count feature is that the values of the wr feature do not account for sparsity, i.e., a legal entity with few past cases may have an inflated wr value.

(e) The id feature, which models the performance of individual entities in a specific role (plaintiff or defendant), is more important than the wr feature, which captures entity win rates regardless of their role. This experiment shows that the historical performance of an entity in a given role can be different than its overall performance, i.e., a party may be successful when litigating as a defendant but not a successful litigator overall, and its important to represent this distinction. We illustrate this situation in the next sub-section, where we show several examples of entities with different win rates for different roles.

4.3 Weight Distributions for Specific Entities

The previous subsections illustrate the relative importance of different feature types at a broad level. In this subsection, our goal is to understand the importance of each entity in the model more specifically. In order to study the relative influence of entities of a specific type independent of other types, we trained five different logistic regression models, where each model trains only on the id features of one entity type (one model each for parties, attorneys, law firms, districts and judges). We then plotted the histograms of the learned weights for each model in Figure 4. For parties, attorneys and law firms, the id features are further distinguished by the roles they play (plaintiff vs. defendant), i.e., we separated the histograms by their roles for these entity types (white vs. back bars respectively).

Although all the plots appear near bell shaped with a mean at zero (indicating neutrality on average), all the plots are characterized by a heavy tail, indicating that a few entities have significant influence in determining the outcome. The reader is cautioned that the tail may appear heavier than it is, since we used a log scale for y-axis for clarity. Nevertheless the fact remains that there are entities with highly polarized weights.
Figure 4: Histograms of the feature weights for the 1d features and all entity types (first five figures). Positive weights indicate influence towards the “plaintiff-wins” class, while negative weights indicate a bias towards “defendant-wins”. The magnitudes of the weights indicate the corresponding strength of influence. Entities in the plaintiff role are represented with white bars (histogram) and solid arrows (example entities); entities in the defendant role are represented with dark bars (histogram) and dashed arrows (example entities). Table (f) lists the number of entities in each role that fail Fisher’s exact test for two different $p$-values.

Note that a few specific districts and judges exhibit large magnitudes of weight indicating patterns of bias towards the plaintiff or defendant. We are unable to identify the specific attorneys, law firms and judges with high influence (bias) because of privacy reasons, but we provided a few interesting examples of high influence behavior in parties and districts in Figure 4. Among parties, our model indicates that 3M Technologies, Inc. wins in both plaintiff and defendant roles consistently (as indicated by its large positive weight in the plaintiff role and large negative weight in the defendant role), which suggests that it has strong IP that it successfully defends or asserts in litigation. On the other hand, Teva Pharmaceutical tends to lose both as plaintiff and defendant. This happens because Teva, a manufacturer of generic drugs, usually faces brand-name pharmaceutical companies in litigation. These companies own strong patents that they typically assert successfully against generic-drug manufacturers. Furthermore, Oakley, Inc. wins as a plaintiff but loses as a defendant (as indicated by a positive weight in both roles), which suggests that it owns valuable IP but it also infringed somebody else's patents. Among districts, the model indicates that the Western District of New York is biased towards plaintiffs while the Northern District of California leans towards defendants. These biases may not indicate an intentional bias in the district or by the judge. There could be other factors at play as discussed in Section 5.

To understand better if the biases of these entities are a result of the small sample sizes or if there is a true deviation from expected...
behavior, we performed Fisher’s exact test [6] for each specific entity in our training data. We implemented each test using contingency counts tables of plaintiff wins vs. defendant wins outcomes for that particular entity vs. all other entities. The results of these tests, displayed in Figure 4(f), indicate that a significant number of entities fail this test, or in other words, they exhibit significant biases that are unexplained by random effects due to small samples.

For example, this experiment extracts Intel and Hewlett Packard because they have a perfect record as defendants in our training data, and the Northern District of California because the outcome distribution in this district is biased towards defendants, opposite to the overall national tendency.

5. DISCUSSION

Litigation, and in particular IP litigation, is an extremely important element of the United States economy. Billions of dollars are spent each year in preparing for litigation. IP trials commonly award damages in the millions of dollars or even billions. This work essentially argues (with empirical support) that IP litigation is a problem fit for forecasting. It is true that in most real-world tasks, e.g., financial markets, forecasting is dangerous because it is virtually impossible to estimate the likelihood of (apparently) highly-improbable events, or black swans, and their impact can be devastating [14]. While we cannot claim that black-swan events do not occur in litigation, these results suggest that there is significantly more structure in litigation than other tasks. We conjecture that one of the main causes that drives the consistent behavior of litigating entities, which is essentially what our model learns, is the fact that litigation in the United States is based on legal precedent, i.e., current decisions are always based on past similar cases, if available. For example, if a company successfully defended an IP object in the past, the same outcome is likely to be assigned to any future litigation on this IP object.

Our work shows that modeling prior factors is beneficial because there is definite correlation between these factors and case outcomes. This happens due to several reasons: (a) prior factors indirectly model the merits of the case, e.g., high success rates for a party indicate that they have strong IP, (b) they capture biases in the legal system, e.g., some districts or judges are biased towards patent owners, and (c) a causation relation may exist between some of these factors and lawsuit outcomes, e.g., a better attorney may increase the probability of winning. However, we cannot claim that our study proves the latter issue. While causation probably exists between some of the modeled prior factors and case outcomes, our work proves only correlation between these factors and outcomes. For example, the correlation between good attorneys and high success rates does not necessarily entail that good attorneys caused these wins. It may simply mean that good attorneys know how to select cases likely to win in the first place. Hence, the success rates of attorneys and law firms may be another indirect measure of the case merits rather than a cause of a certain outcome. A similar explanation exists for the detected correlation between judges/districts and outcomes. While anecdotal observations circulated in the litigation world support the idea that certain districts and judges are indeed biased, e.g., towards patent-owning businesses, we cannot always claim that the judicial system is biased. For example, brand-name pharmaceutical companies, which typically own strong IP, tend to file their patent-infringement cases in the same district court. Hence, judges from this district are likely to objectively award more plaintiff wins than judges from other districts.

The above discussion drives the possible applications of this technology. For instance, it may not always be possible to choose optimal parameters as suggested by the model, especially when the parameter’s correlation with the outcome is not causative. To elaborate further, even if our analysis indicates that a certain law firm maximizes the defendant’s winning probability for a certain case, that law firm could decide not to accept that case because it evaluates the defendant side as too weak. Nevertheless, the correlation between prior factors and case outcomes is sufficient to enable risk analysis once the parameters are set, e.g., it helps parties decide if they should settle or pursue litigation. This is an important application that is likely to save money both for the parties involved in litigation and for the court system, i.e., disputes settled out of court eliminate large litigation expenses and the damages negotiated are usually smaller than the damages awarded by jury verdicts.

A limitation of our current approach is that historical information is required for a given entity in order to compute the corresponding prior factors. Hence, new entities, such as new plaintiffs, are not modeled. However, the experiments show that our heuristic, which considers a case if at least one participating entity from any of the six entity types (party, attorney, or law firms on either side of the case) has historical information, has good coverage: approximately 76% of the cases fall in this category (see Section 3.2). The remaining 24% of cases will have to be addressed using factors that model only the merits of the case.

Modeling the merits of the case is the next step in our overall strategy for litigation risk analysis. As part of our future work we will implement merit-based factors such as strength of the IP object under suit and similarity of the defendant’s technology to the plaintiff’s patents. Note however that the information necessary to model these factors may not be immediately available, e.g., the strength of a patent can be reliably estimated only after all relevant prior art is exposed in the discovery phase of the lawsuit. Also, in over 40% of the cases in our database, the pleading documents were not made available by the courts, so it is extremely difficult to model merits in such cases. Even in cases where the pleading documents are available, modeling the case merits is a complex task. For example, as a first approximation, we incorporated patent information in our model by adding features for: (a) the identifiers of the actual patents under suit, and (b) the strength of patents, estimated using the number of incoming and outgoing citations from other patents10. None of these features helped9, which indicates that, indeed, modeling the case merits is not a trivial task.

On the other hand, the information needed to estimate prior factors is readily available when cases are filed, prior factors are easy to model, and, as illustrated in this paper, there exists a correlation between prior factors and case outcomes.

6. RELATED WORK

To the best of our understanding, there is very little previous work on predicting litigation outcomes reported in artificial intelligence and data mining research communities. The only works we

---

9The intuition was that a higher number of incoming citations is correlated with stronger patents, e.g., stronger patents are cited more often by other patents. Also, more outgoing citations indicate that the patent application was carefully prepared, which is another potential sign of strength.

10Features based on patent numbers did not help because they are extremely sparse, e.g., most patents participate in a single lawsuit, where they are either invalidated or successfully asserted. Our patent strength features model only the probability that a patent is valid but they do not model the likelihood of infringement, which is the more common scenario in litigation.

---
are aware of that are related to our work are those of [2, 4]. Of the two, Arditi et al [2] is the first work to address the problem of predicting outcomes of lawsuits. In this work, they focus on litigation in the construction industry. Their data consists of a training set of 102 construction industry related cases filed in Illinois appellate courts from 1982 through 1992 and an additional test set of 12 cases filed in the same court from 1992 through 1994. They represent each case by a set of 45 domain-specific features extracted from the case pleading documents such as party type, type of contract, contract value, etc., and train a Neural Network. This model achieved 67% accuracy on the test set.

In the other work cited, Chau applied an approach based on Neural Networks [4] to litigation in the same industry. The main difference is that the data came from Hong Kong courts from 1991 to 2000. This dataset had a considerably larger number of cases (1,105 overall) but the model used only 13 unique features manually extracted from the case documents. Chau trains a Neural Network model using Particle Swarm Optimization, achieving an accuracy of 82%.

These works make an initial step towards risk analysis using the merits of the case, even though they use either shallow approximations of case merits or features that were manually coded. On the other hand, we are the first to investigate an orthogonal direction that uses only historical or concurrent information on the litigating entities to evaluate litigation risk. Additionally, all our features are automatically extracted from the training data.

With respect to modeling, conditional random fields [7] have been successfully used in many domains. For example, in the field of natural language processing, Lafferty et al. used CRFs for part-of-speech tagging [7]. Sutton et al. have used a two-layer factorial CRF to jointly model noun-phrase chunking and part-of-speech tagging [13]. In the field of robotics, Liao et al. extracted high-level activities from a sequence of GPS readings using a hierarchical CRF [8]. CRFs have been extensively used in the field of vision, for various tasks, e.g., image segmentation [12, 16] and object recognition [11]. However, to the best of our knowledge, we are the first to apply conditional random fields to the problem of litigation risk analysis.

7. CONCLUSIONS

In this work, we introduced the novel problem of assessing the risk for parties involved in IP litigation based solely on prior factors. Prior factors are attractive because they capture the merits of the case indirectly, identify potential biases in the system, and are easier to extract and model than the actual merits of the case. We modeled risk estimation by estimating the probability that the corresponding party will lose the case. We built a logistic regression classifier to capture historical features and a novel relational model using conditional random fields to jointly predict the outcomes of concurrent and related cases. Our experimental results show that the CRF-based relational classifier outperforms the baseline majority classifier by more than 22 relative percentage points in accuracy. Our extensive feature analysis unearthed the entity types that are most influential in determining the risk.

Our work has established that in the IP litigation domain risk estimation systems can be developed by modeling only the prior information of the participating entities. To the best of our knowledge, this is the first to show that this is possible. As part of future work, we would like to combine both merits of the case as well as prior factors into a single model to achieve further improvements in performance.

We believe that this work can help parties involved in an IP lawsuit make well-informed decisions in terms of settlement or continuation of a case. Having an accurate estimate of litigation risk will also reduce the number of cases that reach trial, which benefits financially all the parties involved and improves the overall efficiency of the judicial system itself.

8. REFERENCES