Distant Supervision for Relation Extraction with Matrix Completion

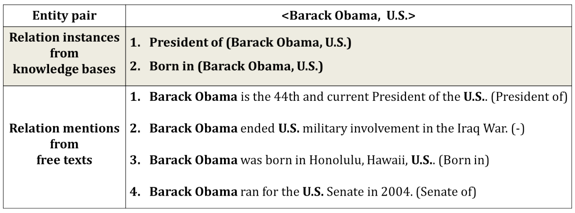
Miao Fan[†,‡,*], **Deli Zhao**[‡], **Qiang Zhou**[†], **Zhiyuan Liu**[◇,‡], **Thomas Fang Zheng**[†], **Edward Y. Chang**[‡]
[†] CSLT, Division of Technical Innovation and Development,
Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, China.
[◇] Department of Computer Science and Technology, Tsinghua University, China.
[‡] HTC Beijing Advanced Technology and Research Center, China.
[*]fanmiao.cslt.thu@gmail.com

Abstract

The essence of distantly supervised relation extraction is that it is an *incomplete* multi-label classification problem with *sparse* and *noisy* features. To tackle the sparsity and noise challenges, we propose solving the classification problem using matrix completion on factorized matrix of minimized rank. We formulate relation classification as completing the unknown labels of testing items (entity pairs) in a sparse matrix that concatenates training and testing textual features with training labels. Our algorithmic framework is based on the assumption that the rank of item-by-feature and item-by-label joint matrix is low. We apply two optimization models to recover the underlying low-rank matrix leveraging the sparsity of feature-label matrix. The matrix completion problem is then solved by the fixed point continuation (FPC) algorithm, which can find the global optimum. Experiments on two widely used datasets with different dimensions of textual features demonstrate that our low-rank matrix completion approach significantly outperforms the baseline and the state-of-the-art methods.

1 Introduction

Relation Extraction (RE) is the process of generating structured relation knowledge from unstructured natural language texts. Traditional supervised methods (Zhou et al., 2005; Bach and Badaskar, 2007) on small hand-labeled corpora, such as MUC[1] and ACE[2], can achieve high precision and recall. However, as producing hand-labeled corpora is laborius and expensive, the supervised approach can not satisfy the increasing demand of building large-scale knowledge repositories with the explosion of Web texts. To address the lacking training data issue, we consider the distant (Mintz et al., 2009) or weak (Hoffmann et al., 2011) supervision paradigm attractive, and we improve the effectiveness of the paradigm in this paper.



Entity pair	**<Barack Obama, U.S.>**
Relation instances from knowledge bases	1. **President of (Barack Obama, U.S.)** 2. **Born in (Barack Obama, U.S.)**
Relation mentions from free texts	1. **Barack Obama** is the 44th and current President of the **U.S.**. (President of) 2. **Barack Obama** ended **U.S.** military involvement in the Iraq War. (-) 3. **Barack Obama** was born in Honolulu, Hawaii, **U.S.**. (Born in) 4. **Barack Obama** ran for the **U.S.** Senate in 2004. (Senate of)

Figure 1: Training corpus generated by the basic alignment assumption of distantly supervised relation extraction. The relation instances are the triples related to President Barack Obama in the Freebase, and the relation mentions are some sentences describing him in the Wikipedia.

The intuition of the paradigm is that one can take advantage of several knowledge bases, such as WordNet[3], Freebase[4] and YAGO[5], to automatically label free texts, like Wikipedia[6] and New York Times corpora[7], based on some heuristic alignment assumptions. An example accounting for the basic but practical assumption is illustrated in Figure 1, in which we know that the two entities (<Barack Obama, U.S.>) are not only involved in the *relation instances*[8] coming from knowledge bases (President-of(Barack Obama, U.S.) and Born-in(Barack Obama, U.S.)),

[1] http://www.itl.nist.gov/iaui/894.02/related projects/muc/

[2] http://www.itl.nist.gov/iad/mig/tests/ace/

[3] http://wordnet.princeton.edu

[4] http://www.freebase.com

[5] http://www.mpi-inf.mpg.de/yago-naga/yago

[6] http://www.wikipedia.org

[7] http://catalog.ldc.upenn.edu/LDC2008T19

[8] According to convention, we regard a structured triple $r(e_i, e_j)$ as a relation instance which is composed of a pair of entities $<e_i, e_j>$and a relation name r with respect to them.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 839–849,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

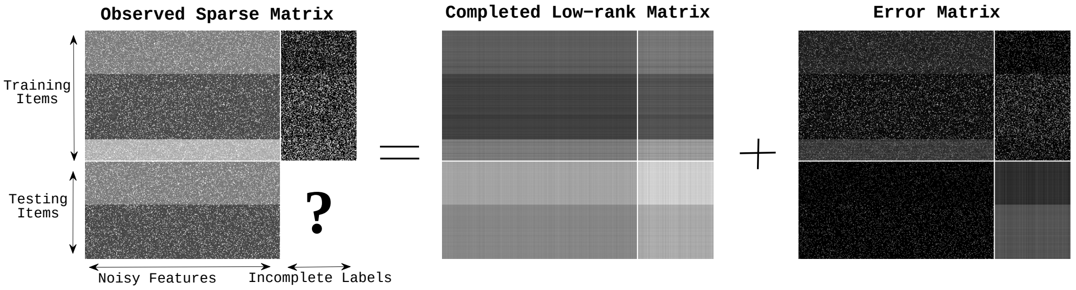


Figure 2: The procedure of noise-tolerant low-rank matrix completion. In this scenario, distantly supervised relation extraction task is transformed into completing the labels for testing items (entity pairs) in a sparse matrix that concatenates training and testing textual features with training labels. We seek to recover the underlying low-rank matrix and to complete the unknown testing labels simultaneously.

but also co-occur in several *relation mentions*[9] appearing in free texts (`Barack Obama is the 44th and current President of the U.S.` and `Barack Obama was born in Honolulu, Hawaii, U.S.`, etc.). We extract diverse textual features from all those *relation mentions* and combine them into a rich feature vector labeled by the *relation names* (`President-of` and `Born-in`) to produce a *weak* training corpus for relation classification.

This paradigm is promising to generate large-scale training corpora automatically. However, it comes up against three technical challeges:

- **Sparse features**. As we cannot tell what kinds of features are effective in advance, we have to use NLP toolkits, such as Stanford CoreNLP[10], to extract a variety of textual features, e.g., named entity tags, part-of-speech tags and lexicalized dependency paths. Unfortunately, most of them appear only once in the training corpus, and hence leading to very sparse features.

- **Noisy features**. Not all relation mentions express the corresponding relation instances. For example, the second relation mention in Figure 1 does not explicitly describe any relation instance, so features extracted from this sentence can be noisy. Such analogous cases commonly exist in feature extraction.

- **Incomplete labels**. Similar to noisy features, the generated labels can be incomplete. For example, the fourth relation mention in Figure 1 should have been labeled by the relation `Senate-of`. However, the incomplete knowledge base does not contain the corresponding relation instance (`Senate-of(Barack Obama, U.S.)`). Therefore, the distant supervision paradigm may generate incomplete labeling corpora.

In essence, distantly supervised relation extraction is an *incomplete* multi-label classification task with *sparse* and *noisy* features.

In this paper, we formulate the relation-extraction task from a novel perspective of using matrix completion with low rank criterion. To the best of our knowledge, we are the first to apply this technique on relation extraction with distant supervision. More specifically, as shown in Figure 2, we model the task with a sparse matrix whose rows present items (entity pairs) and columns contain noisy textual features and incomplete relation labels. In such a way, relation classification is transformed into a problem of completing the unknown labels for testing items in the sparse matrix that concatenates training and testing textual features with training labels, based on the assumption that the item-by-feature and item-by-label joint matrix is of low rank. The rationale of this assumption is that noisy features and incomplete labels are semantically correlated. The low-rank factorization of the sparse feature-label matrix delivers the low-dimensional representation of de-correlation for features and labels.

[9] The sentences that contain the given entity pair are called relation mentions.

[10] http://nlp.stanford.edu/downloads/corenlp.shtml

We contribute two optimization models, DRMC[11]-b and DRMC-1, aiming at exploiting the sparsity to recover the underlying low-rank matrix and to complete the unknown testing labels simultaneously. Moreover, the logistic cost function is integrated in our models to reduce the influence of noisy features and incomplete labels, due to that it is suitable for binary variables. We also modify the fixed point continuation (FPC) algorithm (Ma et al., 2011) to find the global optimum.

Experiments on two widely used datasets demonstrate that our noise-tolerant approaches outperform the baseline and the state-of-the-art methods. Furthermore, we discuss the influence of feature sparsity, and our approaches consistently achieve better performance than compared methods under different sparsity degrees.

2 Related Work

The idea of distant supervision was firstly proposed in the field of bioinformatics (Craven and Kumlien, 1999). Snow et al. (2004) used WordNet as the knowledge base to discover more hpyernym/hyponym relations between entities from news articles. However, either bioinformatic database or WordNet is maintained by a few experts, thus hardly kept up-to-date.

As we are stepping into the *big data* era, the explosion of unstructured Web texts simulates us to build more powerful models that can automatically extract relation instances from large-scale online natural language corpora without hand-labeled annotation. Mintz et al. (2009) adopted Freebase (Bollacker et al., 2008; Bollacker et al., 2007), a large-scale crowdsourcing knowledge base online which contains billions of relation instances and thousands of relation names, to *distantly supervise* Wikipedia corpus. The basic alignment assumption of this work is that if a pair of entities participate in a relation, *all sentences* that mention these entities are labeled by that relation name. Then we can extract a variety of textual features and learn a multi-class logistic regression classifier. Inspired by multi-instance learning (Maron and Lozano-Pérez, 1998), Riedel et al. (2010) relaxed the strong assumption and replaced *all sentences* with *at least one sentence*. Hoffmann et al. (2011) pointed out that many entity pairs have more than one relation. They extended the multi-instance learning framework (Riedel et al., 2010) to the multi-label circumstance. Surdeanu et al. (2012) proposed a novel approach to multi-instance multi-label learning for relation extraction, which jointly modeled all the sentences in texts and all labels in knowledge bases for a given entity pair. Other literatures (Takamatsu et al., 2012; Min et al., 2013; Zhang et al., 2013; Xu et al., 2013) addressed more specific issues, like how to construct the negative class in learning or how to adopt more information, such as name entity tags, to improve the performance.

Our work is more relevant to Riedel et al.'s (2013) which considered the task as a matrix factorization problem. Their approach is composed of several models, such as PCA (Collins et al., 2001) and collaborative filtering (Koren, 2008). However, they did not concern about the data noise brought by the basic assumption of distant supervision.

3 Model

We apply a new technique in the field of applied mathematics, i.e., low-rank matrix completion with convex optimization. The breakthrough work on this topic was made by Candès and Recht (2009) who proved that most low-rank matrices can be perfectly recovered from an incomplete set of entries. This promising theory has been successfully applied on many active research areas, such as computer vision (Cabral et al., 2011), recommender system (Rennie and Srebro, 2005) and system controlling (Fazel et al., 2001). Our models for relation extraction are based on the theoretic framework proposed by Goldberg et al. (2010), which formulated the multi-label transductive learning as a matrix completion problem. The new framework for classification enhances the robustness to data noise by penalizing different cost functions for features and labels.

3.1 Formulation

Suppose that we have built a training corpus for relation classification with n items (entity pairs), d-dimensional textual features, and t labels (relations), based on the basic alignment assumption proposed by Mintz et al. (2009). Let $X_{train} \in \mathbb{R}^{n \times d}$ and $Y_{train} \in \mathbb{R}^{n \times t}$ denote the feature matrix and the label matrix for training, respectively. The linear classifier we adopt aims to explicitly learn the weight matrix $\mathbf{W} \in \mathbb{R}^{d \times t}$ and the bias column

[11] It is the abbreviation for **D**istant supervision for **R**elation extraction with **M**atrix **C**ompletion

vector $\mathbf{b} \in \mathbb{R}^{t\times 1}$ with the constraint of minimizing the loss function l,

$$\arg\min_{\mathbf{W},\mathbf{b}} \; l(Y_{train}, \left[\begin{array}{cc} \mathbf{1} & X_{train} \end{array} \right] \left[\begin{array}{c} \mathbf{b}^T \\ \mathbf{W} \end{array} \right]), \quad (1)$$

where $\mathbf{1}$ is the all-one column vector. Then we can predict the label matrix $Y_{test} \in \mathbb{R}^{m\times t}$ of m testing items with respect to the feature matrix $X_{test} \in \mathbb{R}^{m\times d}$. Let

$$\mathbf{Z} = \left[\begin{array}{cc} X_{train} & Y_{train} \\ X_{test} & Y_{test} \end{array} \right].$$

This linear classification problem can be transformed into completing the unobservable entries in Y_{test} by means of the observable entries in X_{train}, Y_{train} and X_{test}, based on the assumption that the rank of matrix $\mathbf{Z} \in \mathbb{R}^{(n+m)\times(d+t)}$ is low. The model can be written as,

$$\begin{aligned} & \arg\min_{\mathbf{Z}\in\mathbb{R}^{(n+m)\times(d+t)}} \quad \text{rank}(\mathbf{Z}) \\ & \text{s.t. } \forall(i,j) \in \Omega_X, \;\; z_{ij} = x_{ij}, \\ & \quad (1 \le i \le n+m, \;\; 1 \le j \le d), \\ & \quad \forall(i,j) \in \Omega_Y, \;\; z_{i(j+d)} = y_{ij}, \\ & \quad (1 \le i \le n, \;\; 1 \le j \le t), \end{aligned} \quad (2)$$

where we use Ω_X to represent the index set of observable feature entries in X_{train} and X_{test}, and Ω_Y to denote the index set of observable label entries in Y_{train}.

Formula (2) is usually impractical for real problems as the entries in the matrix $\mathbf{Z}$ are corrupted by noise. We thus define

$$\mathbf{Z} = \mathbf{Z}^* + \mathbf{E},$$

where $\mathbf{Z}^*$ as the underlying low-rank matrix

$$\mathbf{Z}^* = \left[\begin{array}{cc} X^* & Y^* \end{array} \right] = \left[\begin{array}{cc} X^*_{train} & Y^*_{train} \\ X^*_{test} & Y^*_{test} \end{array} \right],$$

and $\mathbf{E}$ is the error matrix

$$\mathbf{E} = \left[\begin{array}{cc} E_{X_{train}} & E_{Y_{train}} \\ E_{X_{test}} & 0 \end{array} \right].$$

The rank function in Formula (2) is a non-convex function that is difficult to be optimized. The surrogate of the function can be the convex nuclear norm $||\mathbf{Z}||_* = \sum \sigma_k(\mathbf{Z})$ (Candès and Recht, 2009), where σ_k is the k-th largest singular value of $\mathbf{Z}$. To tolerate the noise entries in the error matrix $\mathbf{E}$, we minimize the cost functions C_x and C_y for features and labels respectively, rather than using the hard constraints in Formula (2).

According to Formula (1), $\mathbf{Z}^* \in \mathbb{R}^{(n+m)\times(d+t)}$ can be represented as $[X^*, \mathbf{W}X^*]$ instead of $[X^*, Y^*]$, by explicitly modeling the bias vector $\mathbf{b}$. Therefore, this convex optimization model is called **DRMC-b**,

$$\begin{aligned} \arg\min_{\mathbf{Z},\mathbf{b}} \;\; & \mu||\mathbf{Z}||_* + \frac{1}{|\Omega_X|} \sum_{(i,j)\in\Omega_X} C_x(z_{ij}, x_{ij}) \\ & + \frac{\lambda}{|\Omega_Y|} \sum_{(i,j)\in\Omega_Y} C_y(z_{i(j+d)} + b_j, y_{ij}), \end{aligned} \quad (3)$$

where μ and λ are the positive trade-off weights. More specifically, we minimize the nuclear norm $||\mathbf{Z}||_*$ via employing the regularization terms, i.e., the cost functions C_x and C_y for features and labels.

If we implicitly model the bias vector $\mathbf{b}$, $\mathbf{Z}^* \in \mathbb{R}^{(n+m)\times(1+d+t)}$ can be denoted by $[\mathbf{1}, X^*, \mathbf{W}'X^*]$ instead of $[X^*, Y^*]$, in which $\mathbf{W}'$ takes the role of $[\mathbf{b}^T; \mathbf{W}]$ in DRMC-b. Then we derive another optimization model called **DRMC-1**,

$$\begin{aligned} \arg\min_{\mathbf{Z}} \;\; & \mu||\mathbf{Z}||_* + \frac{1}{|\Omega_X|} \sum_{(i,j)\in\Omega_X} C_x(z_{i(j+1)}, x_{ij}) \\ & + \frac{\lambda}{|\Omega_Y|} \sum_{(i,j)\in\Omega_Y} C_y(z_{i(j+d+1)}, y_{ij}) \\ \text{s.t.} \;\; & \mathbf{Z}(:,1) = \mathbf{1}, \end{aligned} \quad (4)$$

where $\mathbf{Z}(:,1)$ denotes the first column of $\mathbf{Z}$.

For our relation classification task, both features and labels are binary. We assume that the actual entry u belonging to the underlying matrix $\mathbf{Z}^*$ is randomly generated via a sigmoid function (Jordan, 1995): $Pr(u|v) = 1/(1 + e^{-uv})$, given the observed binary entry v from the observed sparse matrix $\mathbf{Z}$. Then, we can apply the log-likelihood cost function to measure the conditional probability and derive the *logistic cost function* for C_x and C_y,

$$C(u,v) = -\log Pr(u|v) = \log(1 + e^{-uv}),$$

After completing the entries in Y_{test}, we adopt the sigmoid function to calculate the conditional probability of relation r_j, given entity pair p_i pertaining to y_{ij} in Y_{test},

$$Pr(r_j|p_i) = \frac{1}{1 + e^{-y_{ij}}}, \quad y_{ij} \in Y_{test}.$$

Finally, we can achieve Top-N predicted relation instances via ranking the values of $Pr(r_j|p_i)$.

4 Algorithm

The matrix rank minimization problem is NP-hard. Therefore, Candés and Recht (2009) suggested to use a convex relaxation, the nuclear norm minimization instead. Then, Ma et al. (2011) proposed the fixed point continuation (FPC) algorithm which is fast and robust. Moreover, Goldfrab and Ma (2011) proved the convergence of the FPC algorithm for solving the nuclear norm minimization problem. We thus adopt and modify the algorithm aiming to find the optima for our noise-tolerant models, i.e., Formulae (3) and (4).

4.1 Fixed point continuation for DRMC-b

Algorithm 1 describes the modified FPC algorithm for solving DRMC-b, which contains two steps for each iteration,

Gradient step: In this step, we infer the matrix gradient $g(\mathbf{Z})$ and bias vector gradient $g(\mathbf{b})$ as follows,

$$g(z_{ij}) = \begin{cases} \frac{1}{|\Omega_X|}\frac{-x_{ij}}{1+e^{x_{ij}z_{ij}}}, & (i,j) \in \Omega_X \\ \frac{\lambda}{|\Omega_Y|}\frac{-y_{i(j-d)}}{1+e^{y_{i(j-d)}(z_{ij}+b_j)}}, & (i,j-d) \in \Omega_Y \\ 0, & otherwise \end{cases}$$

and

$$g(b_j) = \frac{\lambda}{|\Omega_Y|}\sum_{i:(i,j)\in\Omega_Y}\frac{-y_{ij}}{1+e^{y_{ij}(z_{i(j+d)}+b_j)}}.$$

We use the gradient descents $\mathbf{A} = \mathbf{Z} - \tau_z g(\mathbf{Z})$ and $\mathbf{b} = \mathbf{b} - \tau_b g(\mathbf{b})$ to gradually find the global minima of the cost function terms in Formula (3), where τ_z and τ_b are step sizes.

Shrinkage step: The goal of this step is to minimize the nuclear norm $||\mathbf{Z}||_*$ in Formula (3). We perform the singular value decomposition (SVD) (Golub and Kahan, 1965) for $\mathbf{A}$ at first, and then cut down each singular value. During the iteration, any negative value in $\mathbf{\Sigma} - \tau_{\mathbf{z}}\mu$ is assigned by zero, so that the rank of reconstructed matrix $\mathbf{Z}$ will be reduced, where $\mathbf{Z} = \mathbf{U}max(\mathbf{\Sigma} - \tau_{\mathbf{z}}\mu, 0)\mathbf{V^T}$.

To accelerate the convergence, we use a continuation method to improve the speed. μ is initialized by a large value μ_1, thus resulting in the fast reduction of the rank at first. Then the convergence slows down as μ decreases while obeying $\mu_{k+1} = max(\mu_k\eta_\mu, \mu_F)$. μ_F is the final value of μ, and η_μ is the decay parameter.

For the stopping criteria in inner iterations, we define the *relative error* to measure the residual of matrix $\mathbf{Z}$ between two successive iterations,

$$\frac{||\mathbf{Z}^{k+1} - \mathbf{Z}^k||_F}{max(1, ||\mathbf{Z}^k||_F)} \le \varepsilon,$$

where ε is the convergence threshold.

Algorithm 1 FPC algorithm for solving DRMC-b

```
Input:
  Initial matrix $\mathbf{Z_0}$, bias $\mathbf{b_0}$; Parameters $\mu, \lambda$;
  Step sizes $\tau_z, \tau_b$.

  Set $\mathbf{Z} = \mathbf{Z_0}, \mathbf{b} = \mathbf{b_0}$.
  foreach $\mu = \mu_1 > \mu_2 > ... > \mu_F$ do
    while relative error $> \varepsilon$ do
      Gradient step:
      $\mathbf{A} = \mathbf{Z} - \tau_z g(\mathbf{Z}), \mathbf{b} = \mathbf{b} - \tau_b g(\mathbf{b})$.
      Shrinkage step:
      $\mathbf{U\Sigma V^T}$ = SVD($\mathbf{A}$),
      $\mathbf{Z} = \mathbf{U}\ max(\mathbf{\Sigma} - \tau_{\mathbf{z}}\mu, 0)\ \mathbf{V^T}$.
    end while
  end foreach

Output:   Completed Matrix Z, bias b.
```

4.2 Fixed point continuation for DRMC-1

Algorithm 2 is similar to Algorithm 1 except for two differences. First, there is no bias vector **b**. Second, a projection step is added to enforce the first column of matrix **Z** to be **1**. In addition, The matrix gradient $g(\mathbf{Z})$ for DRMC-1 is

$$g(z_{ij}) = \begin{cases} \frac{1}{|\Omega_X|}\frac{-x_{i(j-1)}}{1+e^{x_{i(j-1)}z_{ij}}}, & (i,j-1) \in \Omega_X \\ \frac{\lambda}{|\Omega_Y|}\frac{-y_{i(j-d-1)}}{1+e^{y_{i(j-d-1)}z_{ij}}}, & (i,j-d-1) \in \Omega_Y \\ 0, & otherwise \end{cases}.$$

Algorithm 2 FPC algorithm for solving DRMC-1

```
Input:
  Initial matrix $\mathbf{Z_0}$; Parameters $\mu, \lambda$;
  Step sizes $\tau_z$.

  Set $\mathbf{Z} = \mathbf{Z_0}$.
  foreach $\mu = \mu_1 > \mu_2 > ... > \mu_F$ do
    while relative error $> \varepsilon$ do
      Gradient step: $\mathbf{A} = \mathbf{Z} - \tau_z g(\mathbf{Z})$.
      Shrinkage step:
      $\mathbf{U\Sigma V^T}$ = SVD($\mathbf{A}$),
      $\mathbf{Z} = \mathbf{U}\ max(\mathbf{\Sigma} - \tau_{\mathbf{z}}\mu, 0)\ \mathbf{V^T}$.
      Projection step: $\mathbf{Z}(:, 1) = \mathbf{1}$.
    end while
  end foreach

Output:   Completed Matrix Z.
```

Dataset	# of training tuples	# of testing tuples	% with more than one label	# of features	# of relation labels
NYT'10	4,700	1,950	7.5%	244,903	51
NYT'13	8,077	3,716	0%	1,957	51

Table 1: Statistics about the two widely used datasets.

Model	NYT'10 (θ=2)	NYT'10 (θ=3)	NYT'10 (θ=4)	NYT'10 (θ=5)	NYT'13
DRMC-b	51.4 ± 8.7 (51)	45.6 ± 3.4 (46)	41.6 ± 2.5 (43)	36.2 ± 8.8(37)	84.6 ± 19.0 (85)
DRMC-1	16.0 ± 1.0 (16)	16.4 ± 1.1(17)	16 ± 1.4 (17)	16.8 ± 1.5(17)	15.8 ± 1.6 (16)

Table 2: The range of optimal ranks for DRMC-b and DRMC-1 through five-fold cross validation. The threshold θ means filtering the features that appear less than θ times. The values in brackets pertaining to DRMC-b and DRMC-1 are the exact optimal ranks that we choose for the completed matrices on testing sets.

5 Experiments

In order to conduct reliable experiments, we adjust and estimate the parameters for our approaches, DRMC-b and DRMC-1, and compare them with other four kinds of landmark methods (Mintz et al., 2009; Hoffmann et al., 2011; Surdeanu et al., 2012; Riedel et al., 2013) on two public datasets.

5.1 Dataset

The two widely used datasets that we adopt are both automatically generated by aligning Freebase to New York Times corpora. The first dataset[12], NYT'10, was developed by Riedel et al. (2010), and also used by Hoffmann et al. (2011) and Surdeanu et al. (2012). Three kinds of features, namely, lexical, syntactic and named entity tag features, were extracted from relation mentions. The second dataset[13], NYT'13, was also released by Riedel et al. (2013), in which they only regarded the lexicalized dependency path between two entities as features. Table 1 shows that the two datasets differ in some main attributes. More specifically, NYT'10 contains much higher dimensional features than NYT'13, whereas fewer training and testing items.

5.2 Parameter setting

In this part, we address the issue of setting parameters: the trade-off weights μ and λ, the step sizes τ_z and τ_b, and the decay parameter η_μ.

We set $\lambda = 1$ to make the contribution of the cost function terms for feature and label matrices equal in Formulae (3) and (4). μ is assigned by a series of values obeying $\mu_{k+1} = max(\mu_k \eta_\mu, \mu_F)$. We follow the suggestion in (Goldberg et al., 2010) that μ starts at $\sigma_1 \eta_\mu$, and σ_1 is the largest singular value of the matrix $\mathbf{Z}$. We set $\eta_\mu = 0.01$. The final value of μ, namely μ_F, is equal to 0.01. Ma et al. (2011) revealed that as long as the non-negative step sizes satisfy $\tau_z < min(\frac{4|\Omega_Y|}{\lambda}, |\Omega_X|)$ and $\tau_b < \frac{4|\Omega_Y|}{\lambda(n+m)}$, the FPC algorithm will guarantee to converge to a global optimum. Therefore, we set $\tau_z = \tau_b = 0.5$ to satisfy the above constraints on both two datasets.

5.3 Rank estimation

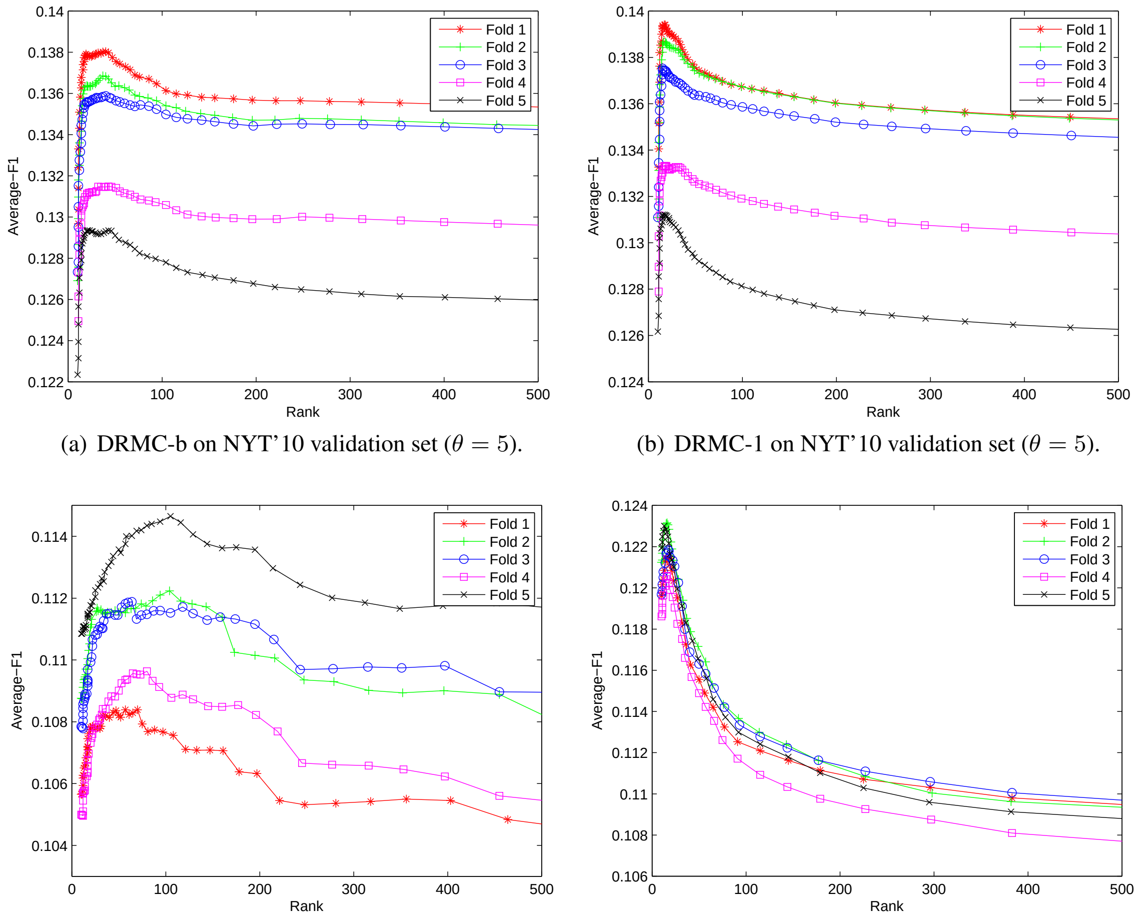
Even though the FPC algorithm converges in iterative fashion, the value of ε varying with different datasets is difficult to be decided. In practice, we record the rank of matrix $\mathbf{Z}$ at each round of iteration until it converges at a rather small threshold $\varepsilon = 10^{-4}$. The reason is that we suppose the optimal low-rank representation of the matrix $\mathbf{Z}$ conveys the truly effective information about underlying semantic correlation between the features and the corresponding labels.

We use the five-fold cross validation on the validation set and evaluate the performance on each fold with different ranks. At each round of iteration, we gain a recovered matrix and average the F1[14] scores from Top-5 to Top-all predicted relation instances to measure the performance. Figure 3 illustrates the curves of average F1 scores. After recording the rank associated with the highest F1 score on each fold, we compute the mean and the standard deviation to estimate the range of optimal rank for testing. Table 2 lists the range of optimal ranks for DRMC-b and DRMC-1 on NYT'10 and NYT'13.

[12] http://iesl.cs.umass.edu/riedel/ecml/

[13] http://iesl.cs.umass.edu/riedel/data-univSchema/

[14] $F1 = \frac{2 \times precision \times recall}{precision + recall}$



(a) DRMC-b on NYT'10 validation set ($\theta = 5$).

(b) DRMC-1 on NYT'10 validation set ($\theta = 5$).

(c) DRMC-b on NYT'13 validation set.

(d) DRMC-1 on NYT'13 validation set.

Figure 3: Five-fold cross validation for rank estimation on two datasets.

On both two datasets, we observe an identical phenomenon that the performance gradually increases as the rank of the matrix declines before reaching the optimum. However, it sharply decreases if we continue reducing the optimal rank. An intuitive explanation is that the high-rank matrix contains much noise and the model tends to be overfitting, whereas the matrix of excessively low rank is more likely to lose principal information and the model tends to be underfitting.

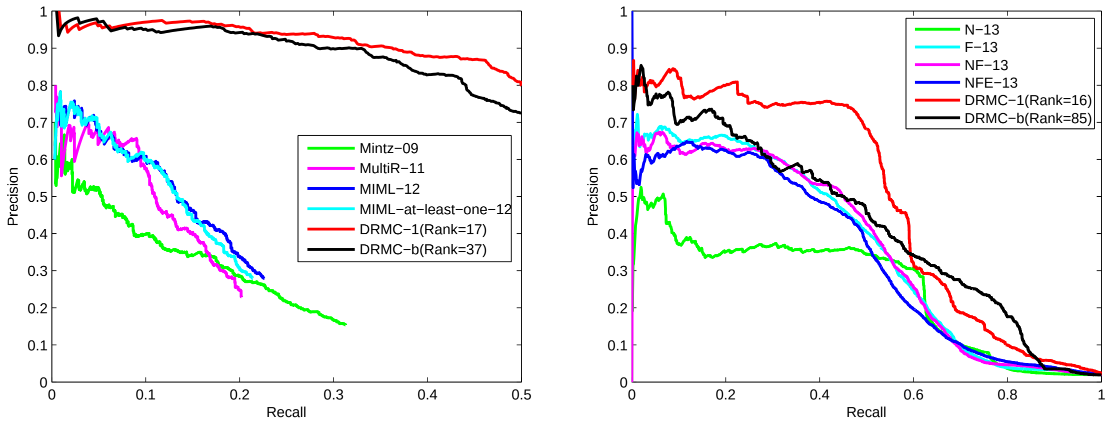
5.4 Method Comparison

Firstly, we conduct experiments to compare our approaches with Mintz-09 (Mintz et al., 2009), MultiR-11 (Hoffmann et al., 2011), MIML-12 and MIML-at-least-one-12 (Surdeanu et al., 2012) on NYT'10 dataset. Surdeanu et al. (2012) released the open source code[15] to reproduce the experimental results on those previous methods. Moreover, their programs can control the feature sparsity degree through a threshold θ which filters the features that appears less than θ times. They set $\theta = 5$ in the original code by default. Therefore, we follow their settings and adopt the same way to filter the features. In this way, we guarantee the fair comparison for all methods. Figure 4 (a) shows that our approaches achieve the significant improvement on performance.

We also perform the experiments to compare our approaches with the state-of-the-art NFE-13[16] (Riedel et al., 2013) and its sub-methods (N-13, F-13 and NF-13) on NYT'13 dataset. Figure 4 (b) illustrates that our approaches still outperform the state-of-the-art methods. In practical applications, we also concern about the precision on Top-N predicted relation instances. Therefore, We compare the precision of Top-100s, Top-200s and Top-500s for DRMC-1, DRMC-b and the state-of-the-

[15] http://nlp.stanford.edu/software/mimlre.shtml

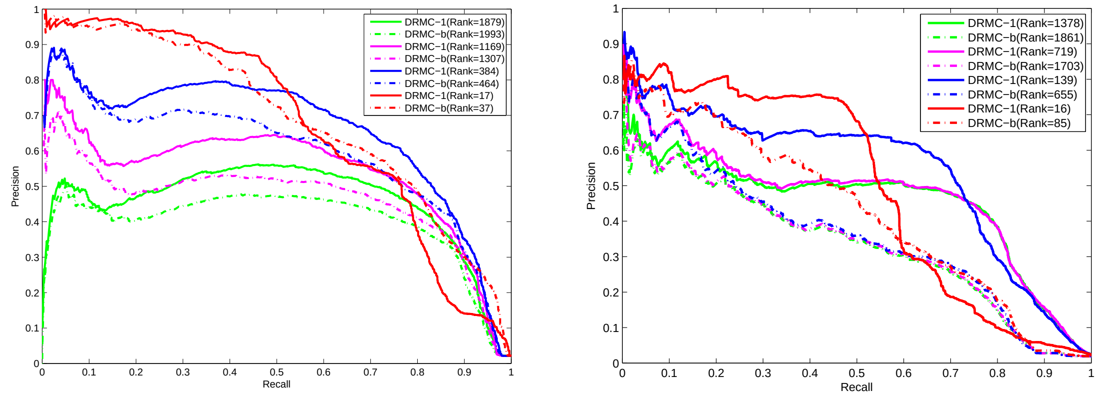
[16] Readers may refer to the website, http://www.riedelcastro.org/uschema for the details of those methods. We bypass the description due to the limitation of space.



(a) NYT'10 testing set ($\theta = 5$).

(b) NYT'13 testing set.

Figure 4: Method comparison on two testing sets.



(a) NYT'10 testing set ($\theta = 5$).

(b) NYT'13 testing set.

Figure 5: Precision-Recall curve for DRMC-b and DRMC-1 with different ranks on two testing sets.

Top-N	NFE-13	DRMC-b	DRMC-1
Top-100	62.9%	**82.0%**	80.0%
Top-200	57.1%	77.0%	**80.0%**
Top-500	37.2%	70.2%	**77.0%**
Average	52.4%	76.4%	**79.0%**

Table 3: Precision of NFE-13, DRMC-b and DRMC-1 on Top-100, Top-200 and Top-500 predicted relation instances.

art method NFE-13 (Riedel et al., 2013). Table 3 shows that DRMC-b and DRMC-1 achieve 24.0% and 26.6% precision increments on average, respectively.

6 Discussion

We have mentioned that the basic alignment assumption of distant supervision (Mintz et al., 2009) tends to generate noisy (noisy features and incomplete labels) and sparse (sparse features) data. In this section, we discuss how our approaches tackle these natural flaws.

Due to the noisy features and incomplete labels, the underlying low-rank data matrix with truly effective information tends to be corrupted and the rank of observed data matrix can be extremely high. Figure 5 demonstrates that the ranks of data matrices are approximately 2,000 for the initial optimization of DRMC-b and DRMC-1. However, those high ranks result in poor performance. As the ranks decline before approaching the optimum, the performance gradually improves, implying that our approaches filter the noise in data and keep the principal information for classification via recovering the underlying low-rank data matrix.

Furthermore, we discuss the influence of the feature sparsity for our approaches and the state-

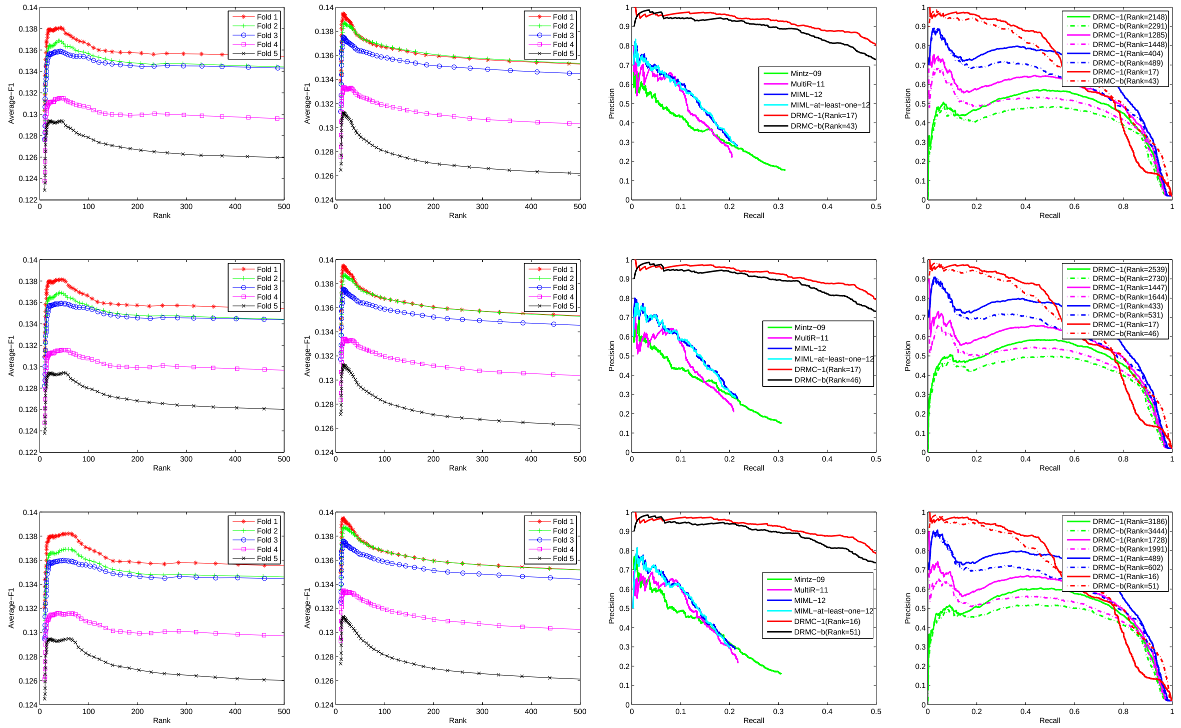


Figure 6: Feature sparsity discussion on NYT'10 testing set. Each row (from top to bottom, $\theta = 4, 3, 2$) illustrates a suite of experimental results. They are, from left to right, five-fold cross validation for rank estimation on DRMC-b and DRMC-1, method comparison and precision-recall curve with different ranks, respectively.

of-the-art methods. We relax the feature filtering threshold ($\theta = 4, 3, 2$) in Surdeanu et al.'s (2012) open source program to generate more sparse features from NYT'10 dataset. Figure 6 shows that our approaches consistently outperform the baseline and the state-of-the-art methods with diverse feature sparsity degrees. Table 2 also lists the range of optimal rank for DRMC-b and DRMC-1 with different θ. We observe that for each approach, the optimal range is relatively stable. In other words, for each approach, the amount of truly effective information about underlying semantic correlation keeps constant for the same dataset, which, to some extent, explains the reason why our approaches are robust to sparse features.

7 Conclusion and Future Work

In this paper, we contributed two noise-tolerant optimization models[17], DRMC-b and DRMC-1, for distantly supervised relation extraction task from a novel perspective. Our models are based on matrix completion with low-rank criterion. Experiments demonstrated that the low-rank representation of the feature-label matrix can exploit the underlying semantic correlated information for relation classification and is effective to overcome the difficulties incurred by sparse and noisy features and incomplete labels, so that we achieved significant improvements on performance.

Our proposed models also leave open questions for distantly supervised relation extraction task. First, they can not process new coming testing items efficiently, as we have to reconstruct the data matrix containing not only the testing items but also all the training items for relation classification, and compute in iterative fashion again. Second, the volume of the datasets we adopt are relatively small. For the future work, we plan to improve our models so that they will be capable of incremental learning on large-scale datasets (Chang, 2011).

Acknowledgments

This work is supported by National Program on Key Basic Research Project (973 Program) under Grant 2013CB329304, National Science Foundation of China (NSFC) under Grant No.61373075.

[17] The source code can be downloaded from `https://github.com/nlpgeek/DRMC/tree/master`

References

Nguyen Bach and Sameer Badaskar. 2007. A review of relation extraction. *Literature review for Language and Statistics II*.

Kurt Bollacker, Robert Cook, and Patrick Tufts. 2007. Freebase: A shared database of structured general human knowledge. In *Proceedings of the national conference on Artificial Intelligence*, volume 22, page 1962. AAAI Press.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Ricardo S Cabral, Fernando Torre, João P Costeira, and Alexandre Bernardino. 2011. Matrix completion for multi-label image classification. In *Advances in Neural Information Processing Systems*, pages 190–198.

Emmanuel J Candès and Benjamin Recht. 2009. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772.

Edward Y Chang. 2011. *Foundations of Large-Scale Multimedia Information Management and Retrieval*. Springer.

Michael Collins, Sanjoy Dasgupta, and Robert E Schapire. 2001. A generalization of principal components analysis to the exponential family. In *Advances in neural information processing systems*, pages 617–624.

Mark Craven and Johan Kumlien. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

Maryam Fazel, Haitham Hindi, and Stephen P Boyd. 2001. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734–4739. IEEE.

Andrew Goldberg, Ben Recht, Junming Xu, Robert Nowak, and Xiaojin Zhu. 2010. Transduction with matrix completion: Three birds with one stone. In *Advances in neural information processing systems*, pages 757–765.

Donald Goldfarb and Shiqian Ma. 2011. Convergence of fixed-point continuation algorithms for matrix rank minimization. *Foundations of Computational Mathematics*, 11(2):183–210.

Gene Golub and William Kahan. 1965. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial & Applied Mathematics, Series B: Numerical Analysis*, 2(2):205–224.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA, June. Association for Computational Linguistics.

Michael Jordan. 1995. Why the logistic function? a tutorial discussion on probabilities and neural networks. *Computational Cognitive Science Technical Report*.

Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434. ACM.

Shiqian Ma, Donald Goldfarb, and Lifeng Chen. 2011. Fixed point and bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1-2):321–353.

Oded Maron and Tomás Lozano-Pérez. 1998. A framework for multiple-instance learning. *Advances in neural information processing systems*, pages 570–576.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782, Atlanta, Georgia, June. Association for Computational Linguistics.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Jasson DM Rennie and Nathan Srebro. 2005. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719. ACM.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–84, Atlanta, Georgia, June. Association for Computational Linguistics.

Rion Snow, Daniel Jurafsky, and Andrew Y Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. *Advances in Neural Information Processing Systems 17*.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465. Association for Computational Linguistics.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 721–729. Association for Computational Linguistics.

Wei Xu, Raphael Hoffmann, Le Zhao, and Ralph Grishman. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 665–670, Sofia, Bulgaria, August. Association for Computational Linguistics.

Xingxing Zhang, Jianwen Zhang, Junyu Zeng, Jun Yan, Zheng Chen, and Zhifang Sui. 2013. Towards accurate distant supervision for relational facts extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 810–815, Sofia, Bulgaria, August. Association for Computational Linguistics.

Guodong Zhou, Jian Su, Jie Zhang, and Min Zhang. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 427–434. Association for Computational Linguistics.