A Logistic Additive Approach for Relation Prediction in Multi-relational Data

Xueyan Jiang[2], Volker Tresp[1,2], and Denis Krompass[2]

[1] Siemens AG, Corporate Technology, Munich, Germany
[2] Ludwig Maximilian University of Munich, Munich, Germany

Abstract. This paper introduces a new stepwise approach for predicting one specific binary relationship in a multi-relational setting. The approach includes a phase of initializing the components of a logistic additive model by matrix factorization and a phase of further optimizing the components with an additive restriction and the Bernoulli modelling assumption. By using low-rank approximations on a set of matrices derived from various interactions of the multi-relational data, the approach achieves data efficiency and exploits sparse matrix algebra. Experiments on three multi-relational datasets are conducted to validate the logistic additive approach.

1 Introduction

Considering the problem of predicting binary relations in multi-relational data, a natural way is to build a three-dimensional tensor such that two dimensions represent the entities and a third dimension stands for different types of relations, and tensor entries represent the existence (a one) or non-existence (a zero) of a relation [7]. Relation prediction can be performed by a low rank approximation via tensor factorization. Many existing tensor factorization methods such as CP [3], Tucker [4,10], RESCAL [7] and the latent factor model [9] are optimizing a target function over all the entries in the tensor. Tensor factorization methods thus learn the multi-linear interaction among the three dimensions through the latent representation of the tensor. The prediction of the truth value of an instantiated relation then equals to the reconstruction of the corresponding entry value in the tensor. Tensor factorization methods have been shown to be efficient and become more and more popular in the areas of psychometrics, chemometrics, neuroscience, the Semantic Web and so on [6,8].

In this paper, we discuss a special case of relation prediction in multi-relational data, where users are only interested in predicting relations between entities in one specific type of relation. We argue that modelling the loss over the query relation slice and capturing different ways of linear interactions between the query entries and the rest of the tensor entries is sufficient. By using low-rank approximations of various matrices derived from the reorganized sparse tensor data, the method exploits sparse matrix algebra and achieves data efficiency.

The proposed approach consists of two steps. In the first step, components of the logistic additive model are initialized by a least-squares fit. The linear combinations of the predictions from all these components are then used as inputs for a logistic regression in the second step. Hence we further optimize the predictions from all the components by additive models using logistic regression.

While the traditional additive models described in [2] assume that each component corresponds to a regression surface describing the targets with their own inputs, we assume in our logistic additive approach that components may share the same input entries but that they are reorganized differently in order to detect different ways of interactions between the query entries and the rest in the tensor.

We provide the least-squares fit solutions for the initialization of all the components, and the update equations for the additive model using logistic regression. The main idea of this two step approach is to first use latent factorization to exploit the correlations between the targets and the rest of the tensor from different ways of interactions, then apply the robust logistic regression using a Bernoulli likelihood as a reasonable model for the binary tensor.

The remaining parts of this paper are organized as follows: Section 2 introduces the stepwise logistic additive approach in detail. Section 3 gives experimental results on Kinship, Nations and the UMLS datasets. Conclusions and possible extensions are given in Section 4.

2 A Logistic Additive Approach

Consider an adjacency tensor $\mathcal{X}$ of size $I \times J \times K$ for a multi-relational graph, where each entry represents a possible link in the graph, I and J are respectively the number of nodes that may have out edges and in edges, and K is the number of labels (i.e., relation types) for the links. Typically $\mathcal{X}$ is sparse for the large multi-relational graph. Let x_{ijk} be the random variable representing the truth of the existence of the links, where $x_{ijk} = 1$ for the links that are known to exist, otherwise $x_{ijk} = 0$.

We now consider the task of predicting the existence of a specific relationship of type q among a set of nodes, i.e. to predict the truth of $x_{ijq} = 1$ for the zero entries in relation slice q. Let the query relation slice q be the matrix $\mathcal{X}_q$. The relation prediction task is to derive a matrix with the same size of $\mathcal{X}_q$ but to replace the zeros by continuous numbers which can be interpreted from the probabilistic point of view as $P(x_{ijq} = 1|\mathcal{X})$. These continuous values can then be the basis for a further analysis to tackle the tasks of classification and ranking.

In the following we take into account different interactions between entries in $\mathcal{X}_q$ and entries in the rest of the tensor, which is denoted as $\mathcal{X}_{\bar{q}}$, $\bar{q} = [1 : q-1, q+1 : K]$. Recall that both $\mathcal{X}_q$ and $\mathcal{X}_{\bar{q}}$ represent some entries from the tensor in this paper but that the organization of the entries is not fixed with these representations. Examples of $\mathcal{X}_q$ and $\mathcal{X}_{\bar{q}}$ can refer to section 3.1;

In the rest of this section, we start with building the additive least-squares model and then extend it to our logistic additive approach.

2.1 The Additive Least-squares Model and Its Normal Equations

The standard parametric method for additive models is to predefine the form of functions f_h (e.g., to polynomial) for the components and then estimate the parameters by least-squares. Here we take a nonparametric method based on H smoothed matrices. A smoother matrix $S^{(h)} : \Re^n \to \Re^n$ is a linear mapping defined by [2]:

$$\tilde{\mathcal{X}}_q^{(h)} = S^{(h)}\mathcal{X}_q \tag{1}$$

Where h is the index of the smoothed matrix or component, $\tilde{\mathcal{X}}_q^{(h)}$ is the smoothed prediction matrix from component h. In our setting, we assume that each smoother $S^{(h)}$ depends on inputs $\mathcal{X}_{\bar{q}}$.

The assumption of our additive model is that we formulate the overall prediction as a linear combination of predictions from H components:

$$\tilde{\mathcal{X}}_q = \sum_{h=1}^{H} \tilde{\mathcal{X}}_q^{(h)} = \sum_{h=1}^{H} S^{(h)}\mathcal{X}_q \tag{2}$$

which minimize:

$$\min_{\left\{\tilde{\mathcal{X}}_q^{(h)} | h=1,\cdots,H\right\}} \quad \|\mathcal{X}_q - \sum_h \tilde{\mathcal{X}}_q^{(h)}\|_F \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

To derive a solution, we start with the solution for only one component (Equation 1) and then extend it to H components. The least-squares fit minimizes:

$$\min_{\tilde{\mathcal{X}}_q^{(h)}} \quad \|\mathcal{X}_q - S^{(h)}\mathcal{X}_q\|_F \tag{4}$$

Recall that $\mathcal{X}$ is a sparse tensor. Therefore any reorganization of $\mathcal{X}_{\bar{q}}^{(h)}$ is also sparse. Adding a regularizer to the cost function in Equation 4 and enforcing a reduced rank, the solution can be computed via the Singular Value Decomposition of $\mathcal{X}_{\bar{q}}^{(h)}$:

$$\tilde{\mathcal{X}}_{\bar{q}}^{(h)} = U_r \text{diag}\left\{\frac{d_i^3}{d_i^2 + \lambda}\right\}_{i=1}^{r} V_r$$

where r is the rank, λ is a regularizer, U_r and V_r are orthogonal matrices.

Thus here $S^{(h)}$ is a projection to the column space of $\mathcal{X}_{\bar{q}}^{(h)}$ and describes the regression space of component h:

$$S^{(h)} = U_r \text{diag}\left\{\frac{d_i^2}{d_i^2 + \lambda}\right\}_{i=1}^{r} U_r^T$$

We hence derive the predictions from component h:

$$\tilde{\mathcal{X}}_q^{(h)} = U_r \text{diag}\left\{\frac{d_i^2}{d_i^2 + \lambda}\right\}_{i=1}^{r} U_r^T X_q$$

Now we extend the solution of one component to the additive model of objective function 3. A sufficient condition for a solution is that the space of residuals $(\mathcal{X}_q - \sum_h \tilde{\mathcal{X}}_q^{(h)})$ is orthogonal to the regression space of the additive model. Since the regression space of the additive model depends on the space of each component h, we have equivalently that the regression space of component h is orthogonal to the residuals [2]:

$$S^{(h)}(\mathcal{X}_q - \sum_{h'=1}^{H} \tilde{\mathcal{X}}_q^{(h')}) = 0$$

Thus:

$$\tilde{\mathcal{X}}_q^{(h)} + \sum_{\bar{h} \neq h} S^{(h)} \tilde{\mathcal{X}}_q^{(\bar{h})} = S^{(h)} \mathcal{X}_q$$

Equivalently, the following systems of normal equations is necessary and sufficient to minimize objective function 3:

$$\begin{pmatrix} I & S^{(1)} & S^{(1)} & \cdots & S^{(1)} \\ S^{(2)} & I & S^{(2)} & \cdots & S^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S^{(H)} & S^{(H)} & S^{(H)} & \cdots & I \end{pmatrix} \begin{pmatrix} \tilde{\mathcal{X}}_q^{(1)} \\ \tilde{\mathcal{X}}_q^{(2)} \\ \vdots \\ \tilde{\mathcal{X}}_q^{(H)} \end{pmatrix} = \begin{pmatrix} S^{(1)}\mathcal{X}_q \\ S^{(2)}\mathcal{X}_q \\ \vdots \\ S^{(H)}\mathcal{X}_q \end{pmatrix}$$

The Gauss-Seidel algorithm to update each $\tilde{X}_q^{(h)}$ at a time is [2]:

$$\tilde{\mathcal{X}}_q^{(h)} new \leftarrow S^{(h)}(\mathcal{X}_q - \sum_{\bar{h}<h} \tilde{\mathcal{X}}_q^{(\bar{h})} new - \sum_{\bar{h}>h} \tilde{\mathcal{X}}_q^{(\bar{h})} old) \qquad (5)$$

2.2 Update Equations for Additive Model using Logistic Regression

A least-squares solution, as derived in the last section, implies a Gaussian noise model. Now we consider a Bernoulli noise model, more appropriate for binary data. In particular, we assume that x_{ijq} follows a Bernoulli distribution:

$$x_{ijq} \sim \text{Bernoulli}(\sigma(\tilde{x}_{ijq}))$$

where

$$\sigma(\tilde{x}_{ijq}) = \frac{1}{1 + e^{-\tilde{x}_{ijq}}} \qquad (6)$$

is the probability for $x_{ijq} = 1$ being true. We denote $\tilde{x}_{ijq}$ as the entries from $\tilde{\mathcal{X}}_q$.

We maximize the log likelihood of the distribution assumption:

$$\begin{aligned} \ell = & \sum_{i=1}^{I} \sum_{j=1}^{J} \log \quad p(x_{ijq}|\tilde{x}_{ijq}) \\ = & \sum_{i=1}^{I} \sum_{j=1}^{J} (x_{ijq} \log \sigma(\tilde{x}_{ijq}) + (1 - x_{ijq}) \log(1 - \sigma(\tilde{x}_{ijq}))) \end{aligned}$$

The stochastic gradient ascent rules for the logistic additive model are:

$$\tilde{\mathcal{X}}_q \leftarrow \sum_h \tilde{\mathcal{X}}_q^{(h)}$$

$$\tilde{\mathcal{X}}_q^{(h)} new \leftarrow \tilde{\mathcal{X}}_q^{(h)} old + \alpha(\mathbf{1} + S_h)(\mathcal{X}_q - \sigma(\tilde{\mathcal{X}}_q)) \quad for \quad h = 1, \cdots, H$$

where α is the learning rate, $\mathbf{1}$ is a matrix of ones for modelling the bias, $\sigma(\tilde{\mathcal{X}}_q)$ denotes elementwise operation to the matrix $\tilde{\mathcal{X}}_q$.

Iterate the above update equations until convergence and we get the final predictions $\sigma(\tilde{\mathcal{X}}_q)$.

2.3 Tuning of Hyperparameters

Each component of the logistic additive approach contains two hyperparameters, i.e., the regularizer λ and the rank r. We follow the method described in [1] and perform a random grid search for the best hyperparameters.

3 Experiments

In this section, we give examples about how to reorganize entries in X_q and $\mathcal{X}_{\bar{q}}$ in order to capture different interactions between the entries. We compare our approach with the additive least-squares model from section 2.1 and the state-of-the-art tensor factorization method RESCAL [7] on the Kinship, Nations and the UMLS datasets.

3.1 Exploiting Different Ways of Interactions

Following the idea in [5], we consider each entry x_{ijk} as a subject-relation-object triple in the form of (s, p, o). We take into account the following three different matrix organizations:

(1) $\mathcal{X}_q$ is a matrix with rows representing all the subject nodes, with columns representing all the object nodes, with entries representing the truth of query relation q among the subjects and objects. $\tilde{\mathcal{X}}_{\bar{q}}$ is a matrix with the same rows as $\mathcal{X}_q$, but with columns representing the object-relation pairs of all the relations except q. Here $\mathcal{X}_q$ and $\tilde{\mathcal{X}}_{\bar{q}}$ can be viewed as part of the mode-1 matrix of the tensor [6].

(2) $\mathcal{X}_q$ is the transpose of the one in case (1). $\tilde{\mathcal{X}}_{\bar{q}}$ is a matrix with the same rows as $\mathcal{X}_q$, but with columns representing the subject-relation pairs of all the relations except q. Here $\mathcal{X}_q$ and $\tilde{\mathcal{X}}_{\bar{q}}$ can be viewed as part of the mode-2 matrix of the tensor [6].

(3) $\mathcal{X}_q$ is a long vector with rows representing all the subject-object pair, with entries representing the truth of query relation q for the subject-object pairs. $\tilde{\mathcal{X}}_{\bar{q}}$ is a matrix with the same rows as $\mathcal{X}_q$, but with columns representing all the relations except q. Here $\mathcal{X}_q$ and $\tilde{\mathcal{X}}_{\bar{q}}$ can be viewed as part of the transpose of mode-3 matrix of the tensor [6].

In the experiments, these three components are taken into account for learning the components for the logistic additive approach. Thus each prediction is based on a linear combination of information related to the context of a triple's subject, object and predicate. The difference to [5] is that instead of fitting an additive least-squares model, we optimize the predictions from each components by applying logistic regressions with additive restrictions.

3.2 Datasets

We use three popular multi-relational datasets for the experiments:
Kinship This dataset describes the kinship relations in Australian tribes. It constructs a graph of 104 subjects, 104 objects and 26 kinship terms to describe the relation types among the subjects and objects.
Nations This dataset describes the relations among 14 countries. It results in a graph of 14 subjects, 14 objects and 56 interactions.
UMLS This dataset describes the causal influence among some biomedical concepts. It forms a graph of 135 subjects, 135 objects and 49 relation types.

3.3 Results

For each dataset, in turn each relation is considered as the query relation q. We ignore relations that have less than 5 nonzero entries. We perform 5-fold cross validation on Kinship and UMLS data, 3-fold cross validation on Nations. Predictions are evaluated by calculating the area under precision recall curve. The final results are presented by averaging the performance on different query relations q and different test folds. Results are shown in table 1.

Since the distribution of nonzero entries is not uniform over different relations in the tensor, we observe quite high variations on the performance for different query relations on the same dataset. The experimental results show that our logistic additive approach outperforms the additive least-squares model on all the three datasets. The logistic additive approach has performance comparable to RESCAL on the UMLS data. However it performs much worse than RESCAL on the Kinship data. But we observe nearly 20% improvement of the logistic additive approach on the Nations dataset, if compared to RESCAL.

Note that the logistic additive approach optimizes the target relations using the other relations as side information. To some extent, it is better than optimizing the whole tensor under a global model assumption. Furthermore, the way how we model the interactions between subject-object pairs over different relations favors the tensor with higher number of relation types, such as the Nations dataset but not Kinship dataset. Unfortunately, the logistic additive approach does not have collective learning ability like RESCAL and thus it could not compete with RESCAL on the Kinship dataset which requires collective learning for optimal results.

Table 1. Experiments on Kinship, Nations and UMLS

Datasets	Additive Least-squares Model	RESCAL	Our Logistic Additive Approach
Kinship	76,71 ± 27,50	**91,81** ± 15,63	81,06 ± 22,87
Nations	79,00 ± 22,13	68,68 ± 19,46	**88,23** ± 12,99
UMLS	73,66 ± 28,76	83,49 ± 26,55	**83,79** ± 24,75

4 Conclusions and Extensions

In this paper we discuss the problem of relation prediction in a multi-relational setting. Instead of optimizing all the entries under one global model assumption, we focus on the query relation slice and use the other relations as side information. A stepwise logistic additive approach is presented to show how we can exploit different interactions between the query and the known information by reorganizing entries for both query relation and the other relations. We validate our method by experiments on Kinship, Nations and UMLS datasets.

Our approach exploits sparse matrix algebra and from the scalability concern, it favours datasets with a large number of relations, e.g.: the Nations dataset. Furthermore, the logistic additive model is quite is flexible. By defining new entries to be a function of some existing entries (e.g., the product of two predicate-object pairs), one can introduce selected interactions into the additive model to exploit non-linearity behaviors.

However, there are also limitations of the approach. The proposed logistic additive approach needs to update the full prediction slices from the components which requires high memory consumption for datasets with large numbers of subjects and objects. Although the approach considers the immediate context of the triples, indirectly related context information is not yet taken into account, which would need to be done via explicit aggregation.

References

1. James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 2012.
2. Andreas Buja, Trevor Hastie, Robert Tibshirani, Andreas Buja, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, 1989.
3. Richard A. Harshman. Foundations of the parafac procedure: Models and conditions for an "explanatory " multimodal factor analysis. In *UCLA Working Papers in Phonectics*, 1970.
4. F.L. Hitchcock. *The Expression of a Tensor Or a Polyadic as a Sum of Products.* Contributions from the Department of Mathematics. sn., 1927.
5. Xueyan Jiang, Volker Tresp, Yi Huang, Maximilian Nickel, and Hans-Peter Kriegel. Link prediction in multi-relational graphs using additive models. In *Proceedings of International Workshop on Semantic Technologies meet Recommender Systems and Big Data at the ISWC*, 2012.
6. Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM REVIEW*, 51(3):455–500, 2009.
7. Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, 2011.
8. Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. Factorizing YAGO: scalable machine learning for linked data. In *WWW*, 2012.
9. A Bordes R Jenatton, N Le Roux and G Obozinski. A latent factor model for highly multi-relational data. In *NIPS*, 2012.
10. L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.