ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Metric Learning with Relational Data

Jiajun Pan and Hoel Le Capitaine

University of Nantes, LS2N UMR CNRS 6004, Nantes, France

Abstract. The vast majority of metric learning approaches are meant to be applied on data described by feature vectors, with some notable exceptions such as times series, trees or graphs. The objective of this paper is to propose metric learning algorithms that consider multi-relational data. More specifically, we present a metric learning approach taking into account the features of the observations, as well as the relationships between observations. Experiments and comparisons of the two settings for a collective classification task on real-world datasets show that our method i) presents a better performance than other approaches in both settings, and ii) scales well with the volume of the data.

1 Introduction

A lot of real-world datasets present aspects of multi-relation between the observations. For instance, social service networks, Wikipedia network, molecular biology classification and so forth. Given that studying relations between entities is a rich domain for knowledge discovery, a number of important science and technology domains could benefit from advances in relational learning. One of the most important tasks in relational learning is collective classification. The core idea of collective classification is that relational data provides useful information for classification of related entities that may share identical classes [1].

It is also well known that metric learning offers great support on classification for machine learning algorithms, which all need to use a distance function or a metric that reflect reasonably well the relationships existing between the different dimensions of the data. In this paper, we want to formalize or learn, a metric that specifically takes into account multi-relational data. More precisely, our purpose is that the proposed metric provides an embedding space for classification, or visualization, that incorporates both relational and label constraints.

2 Preliminaries

Metric learning is a branch of machine learning whose objective is to find a good representation of entities through mapping spaces [2]. It offers a metric, adapted to the data distribution, that is subsequently used in machine learning methods. This representation is based on a good description of the differences or similarities between entities. The general objective of metric learning is to find a distance $d_M(x_i, x_j) = \left((x_i - x_j)^T M (x_i - x_j)\right)^{1/2}$ where M is a linear projection matrix [3]. Given supervised information (labels, ranks, preferences) and unsupervised information (features, graph links, relations), metric learning algorithms make use of the constraints on similarity/dissimilarity, or relative similarity. In the

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

feasible set with these constraints, we generally consider the generic loss function written as $L(M) = \sum_{(i,j,k)\in\mathcal{C}} \ell_M(i,j,k) + \lambda r(M)$, where $\ell_M(i,j,k)$ is the encoded loss from every triple (i,j,k) in the selected constraints set $\mathcal{C}$, and $r(M)$ is a regularization term on the matrix M (e.g Frobenius norm, trace-norm).

There are a lot of linear metric learning algorithms. LMNN (Large-Margin Nearest Neighbors) [4] whose objective is to learn a metric keeping k-nearest neighbors in the same class while giving examples from different classes a large margin to the given point, by pulling or pushing examples in the embedding space. ITML (Information-Theoretic Metric Learning[5]) is based on the of LogDet divergence, which helps to formulate the distance function as a divergence between Gaussian distributions.

Relational learning deals with learning models for which data consists in a generally complex relational structure. Such models are learned to perform specific relational tasks, such as collective classification, or link prediction.

There are many relational learning algorithms, most of them are based on relational models, such as Probabilistic Relational Models (PRMs) [6] and Markov Logic Networks (MLNs) [7]. A PRM is based on a specific object representation through a Bayesian network, with relational schema describing classes in domains. An MLN is a combination of Markov networks and first-order logic. In addition to PRM and MLN, lots of relational learning algorithm are based on graphical models, such as Relational Dependency Networks [8], Relational Markov Networks [9] or Bayesian Logic Programs [10]. All these approaches are probabilistic graphical models, directed or undirected. A relational tensor is a tensor which stores the relationships of relational data with the characteristic function φ_R. For dyadic relational data modeling, we use a labeled directed graph, in which entities are nodes and relationships are labeled directed edges pointing from the subject to the object. The relational tensor then consists in the union of the characteristic function of the relations.

A relational tensor with n entities and n_r different relations can be written as $T \in \mathbb{R}^{n\times n\times n_r}$:

$$t_{ijr} = \begin{cases} 1 & \text{if } R_r(i,j) = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Several approaches using relational tensors have been proposed for relational learning. In [11], focusing on existing link prediction models, the authors extend matrix factorization to use the side information and overcome imbalance. Tucker Decomposition (TD) model is used on user-tag-item relational tensor to provide high-quality tag recommendations. In [12] and [13], they propose RESCAL factorization. RESCAL decomposes the relational tensor to a core tensor R and a matrix A. The matrix A quantifies the similarity of the relationships between entities and can be seen as a new latent feature space. In this paper, we apply the usual metric learning algorithms in the latent space produced by RESCAL for a fair comparison. Few works exist for complex, and potentially non-iid, data such as graphs, trees, strings, sequences, see [14] for examples. In this paper, we

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

are particularly interested in a very popular data structure, namely relational data. In [15], the authors propose a metric learning algorithm, SPML, based on the adjacency matrix A of a network for learning a Mahalanobis distance metric defined by matrix M, which is more effective on the inherent connectivity structure of the network. w

In the next section, we propose to generalize this approach to multi-relational data: multiple entity tables with multiple relationships. Furthermore, we also propose to actually use the target labels of the entities, as opposed to SPML, that only uses links.

3 Relational Metric Learning

There are many metric learning methods related to graph data, however, most of them focus on the relational pattern or graph structure information, and do not consider the label information or ignore the information from the feature of entities.

To the best of our knowledge, there are currently no metric learning approaches specifically dealing with multi-relational data. Some related works are only considering graph links, where all graph nodes are the same kind of observations (e.g. [15]). Other approaches to metric learning are focusing on multi-modal data. For instance, in [16] they consider several modalities for each observation (an image) and learn a metric for each of the modalities. In this paper, our objective is to propose a new approach of metric learning considering all of the three types of information: features, links and labels. Formally, we propose to use the following objective function

$$L(M) = \sum_{(i,j,k)\in C} \ell_M(i,j,k) + \lambda r(M) \tag{2}$$

where $C = \{C_R \cup C_L\}$, i.e. the union of constraints obtained from relational information, C_R, and constraints obtained from labels, C_L. The function ℓ_M is a triplet-wise loss function detailed in the next subsection. Contrary to SPML, we propose to use fully relational data, and not only a graph in which nodes correspond to one type of entity. Furthermore, we make use of supervised information of the node labels, as opposed to SPML. Let us consider relative constraints, expressed by $d_M^2(x_i, x_j) + \gamma < d_M^2(x_i, x_k)$, $\forall (i,j,k) \in C_L$. The set C_L contains (i,j,k) triples of data, where the (x_i, x_j) share the same label and (x_i, x_k) have different labels, and γ is a margin. The relative constraints make sure the entities in different classes be farther with the margin than the entities with the same labels. Based on common usage [4], we choose $\gamma = 1$. We can decompose the loss ℓ_m of Eq. (2) into two distinct losses that take into account label constraints and relational constraints. Using a hinge-loss, we obtain

$$\ell_L = \frac{1}{|C_L|} \sum_{(i,j,k)\in C_L} \max(d_M^2(x_i, x_j) - d_M^2(x_i, x_k) + 1, 0) \tag{3}$$

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

for label constraints. For the relational constraints, we propose to use a multi-relationship tensor in place of the adjacency matrix. More precisely, every slice of the relational tensor is seen as an adjacency matrix. Therefore, we consider the following constraint for each slice r of X, $d^2_M(x_i, x_j) > (1 - R_r(i,j)) \max_l(R_r(i,l) d^2_M(x_i, x_l)), \forall(i,j)$. Summing up over all slices, and using a hinge loss, gives

$$\ell_R = \frac{1}{n_r} \sum_{z=1}^{n_r} \frac{1}{|C_z|} \sum_{(i,j,k) \in C_z} \max(d^2_M(x_i, x_j) - d^2_M(x_i, x_k) + 1, 0), \tag{4}$$

where C_z is the set of constraints obtained through the z-th relation of the tensor cube, more precisely $C_z = \{(i,j,k) | R_z(i,j) = 1, R_z(i,k) = 0\}$. Note also that $\bigcup_{z=1}^{n_r} C_z = C_R$. Combining label constraints and relational constraints finally gives $L(M) = \alpha \ell_R + (1-\alpha)\ell_L + \lambda r(M)$, where the parameter α is introduced to control the trade-off between relational constraints and label constraints. Setting α to 0 makes the approach consider only label constraints, while setting it to 1 only uses relational constraints. The sub-gradient of the objective function is :

$$\begin{aligned} \nabla L(M) = \lambda M + \tfrac{1-\alpha}{|C_L|} \textstyle\sum_{(i,j,k) \in C_L^+} X C^{(i,j,k)} X^T \\ + \tfrac{\alpha}{n_r} \textstyle\sum_{z=1}^{n_r} \tfrac{1}{|C_z|} \sum_{(i,j,k) \in C_z^+} X C^{(i,j,k),z} X^T, \end{aligned} \tag{5}$$

where C_L^+ and C_z^+ are subset of C_L and C_z, respectively, for which $d^2_M(x_i, x_j) - d^2_M(x_i, x_k) + 1 > 0$. $C^{(i,j,k)}$ is a sparse matrix storing the parameters $C^{(i,j,k)}_{jj} = 1$, $C^{(i,j,k)}_{ik} = 1$, $C^{(i,j,k)}_{ki} = 1$, $C^{(i,j,k)}_{kk} = -1$, $C^{(i,j,k)}_{ij} = -1$ and $C^{(i,j,k)}_{ji} = -1$. Otherwise $C^{(i,j,k)} = 0$. Then, we use a stochastic sub-gradient descent with mini-batches to optimize the loss function. We call the corresponding proposed approach Multi-Relational Metric learning (MRML) in the sequel.

4 Experiments and Results

To conduct this study, we use 5 benchmark real-world databases that are traditionally used in relational learning. In Table 1, properties of the datasets we used are given, where n is the number of instances, n_r is the number of types of relations and m is the number of features. For the Elite dataset [17], the target label distinguishes if the elite is top200 or not. In the Mondial dataset [18], labels are defined by the class of the entities. In the movie one [19], labels are movie types. For the UW dataset [20], the phase is used as target label. Finally, the Mutagenesis [21] dataset, in which labels are defined by the types of atom.

As mentioned in the introduction, usual metric learning approaches solely rely on the use of features, and do not make use of the relations between observations. In order to fairly compare metrics, we propose to first embed the data into a space that reflects the relations within the data. To this aim, we propose to use a powerful and recent multi-relational tensor factorization called RESCAL [13]

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Dataset	n	n_r	m	Classes
Elite	4747	41	7	2
Mondial	185	23	4	2
Movie	1804	26	5	18
UW	278	4	3	2
Mutagenesis	4893	6	2	3

Table 1: Dataset characteristics.

[12] as a baseline for encoding the relationships between entities, and then use a metric learning algorithm in the corresponding latent space.

In all experiments, we use the K-nearest neighbors (KNN) algorithm with $K = 5$ and measure the effect of every metric with the cross-validation accuracy score (other values for K were tested without significant changes). The value of λ in the MRML algorithm is set to 0.5 (other values were tested, without significant changes). To be fair, for all learning algorithms, we set the same maximum number of constraints n_c, and after testing several different sets the ranks are similar. In [22] they choose 100 for n_c and in [4, 5] they set maximum number of iteration, the result we show in this paper with n_c equal to 200 for meeting the requirements of each algorithms. Results are given in Table 2. As

	Accuracy				Running time (s)					
Dataset	ITML	LSML	LFDA	MRML	RESCAL	ITML	LSML	LFDA	SPML	MRML
Elite	89.3±0.3	90.8±0.8	90.2±0.5	**91.2±1.3**	2864	206.6	4.13	3.02	24.19	**25.30**
Mondial	61.3±6.6	56.5±11.3	59.5±7.1	**71.2±6.3**	1.17	1.89	0.09	**0.01**	20.33	21.52
Movie	39.2±2.8	38.5±1.8	39.9±1.5	**40.8±1.3**	832	434.1	0.96	2.19	14.93	**22.15**
UW	70.2±1.5	55.9±7.0	**90.6±5.3**	88.5±2.7	0.96	44.4	0.06	**0.01**	17.46	16.80
Muta	79.7±1.5	70.6±1.6	72.0±1.3	**86.2±1.3**	96.14	92.61	3.19	1.37	34.70	**21.12**

Table 2: Cross-validation accuracy of KNN with different metric learning methods.

can be observed, MRML performs better than other approaches, except on the UW dataset, for which LFDA is better (although closely followed by MRML). Total running time of each of these algorithms is given by adding the RESCAL projection time and their individual running times.

5 Conclusion and perspectives

In this paper, we propose a new metric learning method, MRML, based on both features and relationships of entities in multi-relational data. We extend and generalize the SPML approach [15] using the adjacency matrix of a network to a relational tensor, with feature vectors as entities. Then, we present a stochastic subgradient descent algorithm to learn this metric. A parameter, α, allows controlling the amount of supervision of the algorithm, ranging from no use of labels to no use of relations. As perspectives, let us mention the extension of the

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

approach to non-binary relations (e.g. the relation between a user and a movie can be a rating, which is valued), or vector-valued relations.

References

[1] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93, 2008.

[2] Eric P Xing, Andrew Y Ng, Michael I Jordan, and Stuart Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, pages 505–512, 2003.

[3] Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012.

[4] Kilian Q Weinberger, John Blitzer, and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, pages 1473–1480, 2006.

[5] Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *ICML*, pages 209–216. ACM, 2007.

[6] Daphne Koller. Probabilistic relational models. In *International Conference on Inductive Logic Programming*, pages 3–13. Springer, 1999.

[7] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1):107–136, 2006.

[8] Jennifer Neville and David Jensen. Relational dependency networks. *Journal of Machine Learning Research*, 8(Mar):653–692, 2007.

[9] Ben Taskar, Pieter Abbeel, Ming-Fai Wong, and Daphne Koller. Relational markov networks. *Introduction to statistical relational learning*, pages 175–200, 2007.

[10] Kristian Kersting and Luc De Raedt. 1 bayesian logic programming: Theory and tool. *Statistical Relational Learning*, page 291, 2007.

[11] Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 437–452. Springer, 2011.

[12] Maximilian Nickel. *Tensor factorization for relational learning*. PhD thesis, lmu, 2013.

[13] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, volume 11, pages 809–816, 2011.

[14] Aurélien Bellet, Amaury Habrard, and Marc Sebban. Metric learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9(1):1–151, 2015.

[15] Blake Shaw, Bert Huang, and Tony Jebara. Learning a distance metric from a network. In *NIPS*, pages 1899–1907, 2011.

[16] Pengcheng Wu, Steven CH Hoi, Peilin Zhao, Chunyan Miao, and Zhi-Yong Liu. Online multi-modal distance metric learning with application to image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 28(2):454–467, 2016.

[17] Wilco Dekker and Ben van Raaij. *De elite*. De Volkskrant, 2008.

[18] Wolfgang May. Information extraction and integration with FLORID: The MONDIAL case study. Technical Report 131, Universität Freiburg, Institut für Informatik, 1999. Available from `http://dbis.informatik.uni-goettingen.de/Mondial`.

[19] M. Lichman. UCI machine learning repository, 2013.

[20] Hassan Khosravi, Oliver Schulte, Jianfeng Hu, and Tianxiang Gao. Learning compact Markov logic networks with decision trees. *Machine Learning*, 89(3):257–277, 2012.

[21] A. K. Debnath, R. L. Lopez de Compadre, G. Debnath, A. J. Shusterman, and C. Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. Correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

[22] Eric Yi Liu, Zhishan Guo, Xiang Zhang, Vladimir Jojic, and Wei Wang. Metric learning from relative comparisons by minimizing squared residual. In *ICDM*, pages 978–983. IEEE, 2012.