Evaluating the Effectiveness of Margin Parameter when Learning Knowledge Embedding Representation for Domain-specific Multi-relational Categorized Data

Matthew Wai Heng Chung
HDR-UK Institute of Health Informatics
University College London
London, UK
wai.chung.18@ucl.ac.uk

Hegler Tissot
HDR-UK Institute of Health Informatics
University College London
London, UK
h.tissot@ucl.ac.uk

Abstract

Learning knowledge representation is an increasingly important technology that supports a variety of machine learning related applications. However, the choice of hyperparameters is seldom justified and usually relies on exhaustive search. Understanding the effect of hyperparameter combinations on embedding quality is crucial to avoid the inefficient process and enhance practicality of vector representation methods. We evaluate the effects of distinct values for the margin parameter (γ) focused on translational embedding representation models for multi-relational categorized data. We assess the influence of γ regarding the quality of embedding models by contrasting traditional link prediction task accuracy against a classification task. The findings provide evidence that lower values of margin are not rigorous enough to help with the learning process, whereas larger values produce much noise pushing the entities beyond to the surface of the hyperspace, thus requiring constant regularization. Finally, the correlation between link prediction and classification accuracy shows traditional validation protocol for embedding models is a weak metric to represent the quality of embedding representation.

Introduction

Information extraction aims to recover facts from heterogeneous data, and attempts to capture that information using a multi-relational representation. The problem of representing multi-relational data has gained more attention in the last decade as more knowledge graphs become available and useful as supporting resources for a variety of machine learning related applications, such as information retrieval (Büttcher, Clarke, and Cormack 2010), semantic parsing (Berant et al. 2013), question-answering (Abujabal et al. 2017), and recommender systems (Wang et al. 2019).

A knowledge graph (KG) is a multi-relational dataset composed by entities (nodes) and relations (edges) that provide a structured representation of the knowledge about the world. Freebase (Bollacker et al. 2008), Google Knowledge Graph (Dong et al. 2014), Wordnet (Fellbaum 1998), DBpedia (Lehmann et al. 2015), and YAGO (Suchanek, Kasneci, and Weikum 2007) are some well-known examples of KGs that provide reasoning ability and can be used for knowledge inference. However, the heterogeneous nature of the data sources, where facts are usually extracted from, makes the later typically inaccurate. Moreover, although containing a huge number of triplets, most of open-domain KGs are incomplete, covering only a small subset of the true knowledge domain they are supposed to represent. Thus, learning the distributed representation of multi-relational data has been used as a tool to complete knowledge bases without requiring extra knowledge input. *Knowledge base completion* or *link prediction* (LP) refers to the problem of predicting new links (or new relationships) between entities by automatically recovering missing facts based on the observed ones.

Embedding methods are able to learn and operate on the latent feature representation of the constituents and on their semantic relatedness using an algorithm that optimizes a margin-based (γ) objective function over a training set. The problem of choosing the optimal combination of hyperparameters is a common practice usually tackled by performing an exhaustive "grid-search" over several adjustable values before committing to a favorable training model. However, KGs can be very large and demand high computational costs to find the best configuration from all possible combinations of hyperparameter options.

In this work, we focus on evaluating the effectiveness of choosing adequate values for the learning margin parameter γ and how this choice is reflected on the accuracy of embedding models. There is no consensus on the optimal hyperparameters in previous work, and multiple approaches report best model accuracy with γ ranging from 0.2 to 4.0. In addition, we contrast accuracy and quality of embedding representation by comparing the results between LP and classification tasks. Results provide strong evidence that lower values for the margin parameter are not necessarily rigorous enough to help with the learning process, whereas larger values produce much noise pushing the entities beyond to the surface of the hyperspace, causing entities to require constant regularization, which has a negative effect on the validation accuracy. In addition, the correlation between LP and classification accuracy results shows traditional validation

protocol for embedding models is a weak metric to represent the quality of embedding representation. Finally, in order to advocate reproducibility and encourage the research community to test and compare novel embedding methods applied to multi-relational categorized data, we are making all the datasets used along our experiments available.

Knowledge Embedding Representation

A knowledge graph $\mathcal{G}$ is constructed with a set of facts $\mathcal{S}$ in the form of triples (h, r, t). Each triple has a pair of head h and tail t entities $e \in \mathcal{E}$, connected by a relation $r \in \mathcal{R}$. Embedding methods aim to represent entities (h and t) and relations (r) as vectors in a continuous k-dimensional vector space – though most of enhanced models represent relations as k x k matrices. Typically, embedding methods are able to learn and operate on the latent feature representation of the constituents, by defining a distinct relation-based scoring function $f_r(h, t)$ to measure the plausibility of the triplet (h, r, t), where $f_r(h, t)$ implies a transformation on the pair of entities which characterizes the relation r.

The final embedding representation is learned by optimizing a margin-based (γ) objective function (Equation 1) over a training set, while preserving the existing semantic relatedness among the triple constituents to enforce the embedding compatibility. Non-existing negative triples (h', r, t') are constructed for every observed triple in the training set by corrupting either the head or the tail entity. Both observed and corrupted triples are identically scored for comparison in the loss function, where $[x]_+ = max(0, x)$.

$$\mathcal{L} = \sum_{\substack{(h,r,t)\in S \\ (h',r,t')\in S'}} [\gamma - f_r(h,t) + f_r(h',t')]_+ \qquad (1)$$

TransE (Bordes et al. 2013) is a baseline model in translational embedding representation learning. It presents a simple and scalable method to represent KGs in lower dimensional continuous vector space, though known for its flaws on representing one-to-many, many-to-one and many-to-many relations. In TransE, entities h, t and a relation r are represented by translation vectors $\mathbf{h}, \mathbf{t}, \mathbf{r} \in \mathbb{R}^k$, chosen so that every relation r is regarded as a translation between h and t in the embedding space. The pair of embedded entities in a triple (h, r, t) can be approximately connected by $\mathbf{r}$ with low error (Equation 2), and the plausibility score for an embedded triple is calculated by a function of distance measure between $\mathbf{h} + \mathbf{r}$ and $\mathbf{t}$ (Equation 3).

$$\mathbf{h} + \mathbf{r} \approx \mathbf{t} \qquad (2)$$

$$f_r(h,t) = \|h + r - t\|_{l_{1/2}} \qquad (3)$$

Later studies addressed several weaknesses of TransE and proposed extended models enhanced with additional features, including relation-specific projections to improve modelling of different data cardinality (Wang et al. 2014; Lin et al. 2015), adapted scoring functions to allow more flexible translations (Fan et al. 2014; Xiao et al. 2015), and Gaussian embeddings to model semantic uncertainty (He et al. 2015; Xiao, Huang, and Zhu 2016). More recent approaches attempted to improve KG completion performance by exploring several learning features, including tensor factorization (Nickel, Tresp, and Kriegel 2011), compositional vector representation (Nickel, Rosasco, and Poggio 2016), complex spaces (Trouillon et al. 2016), transitive relation embeddings (Zhou et al. 2019), and neural neighborhood-aware embeddings (Kong et al. 2019). However, there are multiple issues with these models that make them difficult for further development and their adoption in other domain-specific applications, including (a) limited performance, (b) time-consuming validation processes, and (c) open-domain validation datasets that prevent embedding methods of taking advantage of more detailed and enriched metadata. Indeed, (Kadlec, Bajgar, and Kleindienst 2017) cast doubt on the performance improvement claims of more recent models whether being due to hyperparameter tuning or different training objectives instead of architectural changes, suggesting future research to re-consider the way performance of embedding models should be evaluated and reported.

HEXTRATO (Tissot 2018) is a translational embedding approach that couples TransE with a set of ontology-based constraints to learn representations for multi-relational categorized data, originally designed to embed biomedical- and clinical-related datasets. In categorical datasets, each entity e is associated with a category (or type) $c \in \mathcal{T}$. Designed to be used on embedding domain-specific data, HEXTRATO improves the translational embedding by using typed entities that are projected onto type-based independent hyperspaces, achieving great performance on the LP task on an adapted (typed) version of Freebase, even in very low k-dimensional spaces ($k < 50$), without necessarily adding complex representation structures within the model training process.

Given a training set S of categorized triples $(c_h{:}h,r,c_t{:}t)$, HEXTRATO learns embedding vectors for entities and relations, so that each categorized entity $c{:}e$ is represented by a embedding vector $e_c \in \mathbb{R}^k$, and each relation r is represented by a embedding vector $r \in \mathbb{R}^k$. A score function f_r (Equation 4) represents a L2-norm dissimilarity, such that the score $f_r(h_{c_h}, t_{c_t})$ of a plausible triple $(c_h{:}h,r,c_t{:}t)$ is smaller than the score $f_r(h'_{c_h}, t'_{c_t})$ of an implausible triple $(c_h{:}h',r,c_t{:}t')$. Then, HEXTRATO learns knowledge embedding representation by minimizing a margin-based (γ) loss function $\mathcal{L}$ (Equation 5) adapted from TransE, where γ is the margin parameter, $\mathcal{S}$ is the set of correct triples, and $\mathcal{S}'$ is the set of incorrect triples $(c_h{:}h',r,c_t{:}t) \cup (c_h{:}h,r,c_t{:}t')$.

$$f_r(h_{c_h}, t_{c_t}) = \|h_{c_h} + r - t_{c_t}\|_{l_2} \qquad (4)$$

$$\mathcal{L} = \sum_{\substack{(c_h:h,r,c_t:t)\in\mathcal{S} \\ (c_h:h',r,c_t:t')\in\mathcal{S}'}} [\gamma + f_r(h_{c_h}, t_{c_t}) - f_r(h'_{c_h}, t'_{c_t})]_+ \qquad (5)$$

The combination of these constraint strategies diminishes the impact of two problems of model training discussed in literature: (a) the "zero loss" problem (Wang, Li, and Pan 2018; Shan et al. 2018) describes the observation that at later

stages of training, corrupted triples tend to be sampled beyond the margin, which leads to a zero loss that is not useful for training – this is because the corrupted head or tail entities are selected by random sampling and we expect false entities to gradually move away during training; and (b) the "false detection" problem (Shan et al. 2018), which arises when poor-quality negative triples are constructed using entities that are often unrelated and have a different semantic type, which may give false confidence to the original triple and reduces representation accuracy. In HEXTRATO, independent vector spaces for each type, coupled with restrictions on domain and range for each relation, lessen the probability of constructing a poor-quality negative triple. The selection of corrupted triples is restricted by setting functional relations and disjoint sets instead of random sampling from the whole set of possible entities, so that training is more efficient and sped up, with reduced impact from uninformative training.

HEXTRATO is proven to be faster than other embedding models due to the set of ontology-based constraints it uses, which optimizes the selection of negative samples during training. However it is still not efficient for a grid-search optimization. For real use-case scenarios, we aim to deploy embedding models in a more effective way, requiring the prototyping models with already known good (not necessarily the best) choice of hyperparameters, which has been the biggest motivation for this work.

Methodology

Domain-specific databases provide categorized data and metadata that can be used to enrich definitions of entities and relations within a knowledge base. Thus, each resulting triple is presented in the form $(c_h{:}h,\ r,\ c_t{:}t)$, where c_h and c_t represent the types of h and t. Besides providing categorized entities, relations are also restricted by domain and range. In the following example, the relation *hasGender* is constrained by the domain *patient* and the range *gender* – in addition, using independent vector spaces to project each entity type leads to a substantial processing time improvement along the validation process.

```
(patient:P01, hasGender, gender:male)
```

We aim to evaluate the effectiveness of the margin parameter γ when learning embedding representation for domain-specific multi-relational categorized data along distinct scenarios regarding dimensionality (k) and dataset sizes and shapes (relation cardinality). Thus, we used the primary ontological constraint proposed by HEXTRATO (typed entities) in order to train a set of embedding models that allowed us to contrast the target hyperparameter performance - this approach is reported in the original paper as "H1".

We believe dataset shapes can considerably affect the performance of embedding methods, as more complex embedding representation models tend to adapt the way relations are taken into account in order to improve accuracy within the LP task. However, the entity representation is just marginally affected. Thus, this work focuses on embedding quality assessment instead of trying to improve LP performance.

Datasets

HEXTRATO was originally evaluated using the following datasets: (a) two real clinical-related datasets extracted from *InfoSaude* (Tissot and Dobson 2018) – an Electronic Health Record (EHR) system; (b) Mushroom, a publicly available dataset deposited on the UCI Machine Learning Repository that describes hypothetical samples corresponding to distinct species of mushrooms; and (c) an adapted version of FB15K dataset (FB15K-Typed), which has been simplified to a set of distinct 55 types in this work (FB55T). In addition, we also included a new dataset (BPA) that is presented as set of restrictions on how medical procedures are constrained by distinct diagnosis. An overview of each dataset[1] is presented below and statistics are depicted in Table 1.

EHR-Demographics comprises a set of 2,185 randomly selected patients from the *InfoSaude* system who had at least one admission between 2014 and 2016. Each patient is described by a set of basic demographic information, including gender, age (range in years) in the admission, marital status (unknown for about 15% of the patients), education level, and two flags indicating whether the patient is known to be either a smoker or pregnant, and the social groups assigned according to a diverse set of rules mainly based on demographic and historical clinical conditions. Demographic features are represented by *many-to-one* relations, whereas association of each patient to social groups is given by a *many-to-many* relation.

EHR-Pregnancy is a dataset used to identify correlations between pre- and post-clinical conditions on pregnant patients with abnormal pregnancy termination, comprised by a set of 2,879 randomly selected pregnant female patients from the *InfoSaude* system in which pregnancy was inadvertently and abnormally interrupted before the expected date of birth; each patient is described by age (range in years), known date of last menstrual period (LMP), whether the patient had an abortion (regardless of reason), and a list of ICD-10 (the 10th revision of the International Classification of Diseases) codes (WHO 2004) registered either before or after the LMP date. This is a dataset mostly comprising *many-to-many* relations that connects patients with corresponding diagnoses.

Mushroom uses a set of features to describe 8,124 hypothetical species of mushrooms, including shape, surface, color, bruises, odor, gill, stalk, veil, ring, spore, population, and habitat - intended to identify whether each species is edible or poisonous given its featured characteristics. All features within this dataset are given by *many-to-one* relations.

BPA (Ambulatory Production Bulletin)[2] is an outpatient care dataset that allows the service provider to be linked to the Public Health Ministry in Brazil to record the care performed at the health facility on an outpatient basis; in order to optimize the data remittance process, there are sev-

[1] https://github.com/hextrato/KER/tree/master/datasets

[2] http://datasus.saude.gov.br/sistemas-e-aplicativos/ambulatoriais/sia

eral rules for the correct completion of submitted data that must be followed strictly, including the restrictions between medical procedures and their constrained diagnoses from ICD-10. This dataset associates medical procedures with a multilevel hierarchical set of *many-to-one* relations, coupled with *many-to-many* relations that impose multiple restrictions on each procedure, mainly regarding to possible related diagnoses.

FB55T is adapted from the Freebase-Typed dataset used along HEXTRATO experiments, from which each entity was categorized based on the metadata description of their corresponding relations; in FB55T, we reduced the high cardinality of 1,248 hierarchical types to a set of 55 aggregated types, each type mostly acting as an entity domain context, though over 50% of the triples have distinct aggregated types between head and tail.

Evaluation Protocol

For each dataset, we look for the best embedding model corresponding to each possible pair of hyperparameters (γ, k), aiming to analyze the impact of distinct values for the margin parameter $\gamma \in \{0.25, 0.5, 1.0, 1.5, 2.0, 3.0, 4.0\}$ in distinct k-dimensional spaces $k \in \{32, 64, 128\}$, regarding the resulting embedding representation accuracy. Our evaluation protocol comprises two main steps: (a) we primarily assess the accuracy of the knowledge embedding representation model using LP as a reference for accuracy performance in order to compare the effects of choosing distinct values for the margin parameter γ; (b) then, the quality of the resulting embedding representation is evaluated by a classification task, designed as multi-categorical classification problem for each one of the evaluation datasets. Finally, the correlation between the accuracy in the LP and classification tasks is analyzed in order the evaluate the effectiveness of γ regarding the quality of each resulting embedding model.

The following training protocol was proposed by HEXTRATO, in which multiple learning processes are used to train independent replicas initialized with random vector representations in order to find the best configuration set:

- Given a training set S of triplets *(c_h:h, r, c_t:t)* we trained embedding models to learn vector representation for each categorized entities and relation. Entity and relation embedding vectors are initialized with the random uniform normalized initialization (Glorot and Bengio 2010). The set of golden triples is then randomly traversed multiple times along the training process, such that each training step produces a corrupted triple for each correct triple.
- The dissimilarity measure was set to the L2-norm distance based on γ, and each optimal model setup is determined by early stopping accordingly to the MRR accuracy on the validation set. Ten distinct replicas of each model are independently trained for each set of targeted hyperparameters. After traversing all the training triplets at most 1,000 epochs with a learning rate $\lambda = 0.01$, the best model is chosen by comparing the scores against a tuning set – the tuning set is used to choose the best replica.
- Final resulting scores are then calculated over the test set. FB55T has no tuning set, so that we were only able run one single replica for each target hyperparameter pair setup.
- Incorrect triples $(c_h : h', r, c_t : t) \cup (c_h : h, r, c_t : t')$ are generated by randomly corrupting either *h* or *t* in a correct triple $(c_h : h, r, c_t : t) \in \mathcal{S}$ using a uniform probability for entity replacement, in which the entity replacement is randomly chosen from the set of entities belonging to the corresponding type of each original relation domain and range. We used Stochastic Gradient Descent (SGD) (Robbins and Monro 1951) to minimize the margin-based loss functions $\mathcal{L}$.
- Within the resulting model, each entity c:e is represented by a embedding vector $e_c \in \mathbb{R}^k$, and each relation r is represented by a embedding vector $r \in \mathbb{R}^k$. Similarly to TransE, there is a score function f_r regarding each relation r (Equation 4) that represents a dissimilarity metric, where the score $f_r(h_{c_h}, t_{c_t})$ of a plausible triple *(c_h:h, r, c_t:t)* is smaller than the score $f_r(h'_{c_h}, t'_{c_t})$ of an implausible triple *(c_h:h', r, c_t:t')*.

For evaluation purposes, we designed LP as a question answering task which aims at completing a categorized triple $(c_h : h, r, c_t : t)$ with *h* or *t* missing, though the type or category c_h or c_t of the missing entity is known. The plausibility of a set of candidate entities in descending order of similarity scores records the rank of the correct missing entity according to the corresponding entity type constraint. LP results are reported by Mean Reciprocal Rank (MRR), whereas achieving higher MRR is taken as good link predictor scores. LP results are presented in Figure 1 (a,b).

Finally, each embedding model was submitted to a classification task. We selected one of the relations in each knowledge base as the target relation for each classifier – the choice was made by looking at the relation with less missing values regarding the set of head entities. Each classifier was designed with: (a) input layer size equals k, (b) hidden intermediate layer size equals $2 \times k$, and (c) output layer size equals the number of possible classes. We used logistic activation functions in all hidden and output cells. The set of relations corresponding to the target relation was split into training (80%), validation (10%) and test (10%) sets, and each classifier was submitted up to 1,000 training epochs early stopping accordingly to the classification accuracy on the validation set. After training at most 1,000 epochs with a learning rate $\lambda = 0.01$, the final classification scores are then calculated over the test set. Each result embedding model as submitted to the same classification protocol in order to evaluate effectiveness of resulting vector representation as a measure of quality. Classification results are presented in Figure 1 (c).

We believe the practical usage of this methodology is as important which was not thoroughly explored in previous publications. Instead of introducing new parameters, we aim to improve our understanding of the parameters that determine the functionality and accuracy of embedding models. Moreover, our evaluation by classification tasks propose a more realistic approach to training models for categorized datasets and challenges the current notion of a better LP score implies a better representation model.

Table 1: Statistics of domain-specific benchmark datasets, given by the number of entities, relations, types, and triples in each dataset split – training (LRN), validation (VLD), tuning (TUN) and test (TST) sets.

	Datasets				
# (number of)	**EHR Demographics**	**EHR Pregnancy**	**UCI Mushroom**	**BPA**	**Freebase FB55T**
Entities	2,237	3,088	8,487	22,874	14,951
Relations	6	5	23	23	1,345
Types	7	4	15	14	55
Triples (total)	15,345	20,768	191,088	186,177	592,213
LRN	13,875	14,588	153,057	177,727	483,142
VLD	463	1,997	9,525	2,889	50,000
TUN	475	2,093	9,564	2,729	n/a
TST	532	2,090	18,942	2,832	59,071

Results

The primary objective of this work is to assess the choice of learning margin on the accuracy of learned embedding models, particularly in the context of multi-relational categorized data. Data quality, ambiguity of category definitions and missing data erode the knowledge represented and undermine the embedding model. However, as long as we are making comparisons using the same dataset their effects would not confound our findings.

Figure 1 presents MRR, MRR for the classifier target relation (MRR$_r$), and classification accuracy for all pairs (γ, k). In general, a margin between 1.0 and 2.0 is preferred in terms of LP across Mushroom, BPA, and FB55T datasets, whereas for EHR Demographics and Pregnancy larger values between 2.0 and 4.0 tend to provide better accuracy in LP. Larger margins tend to work better in higher k-dimensional spaces when comparing results within the same dataset. A plausible reason is that higher dimensions give more room to enforce the margin onto the entities. However, in lower k-dimensional spaces, larger values of γ may create more noise, pushing entities towards the hyperspace surface. Conversely, setting a very small margin may not be adequate to improve the representation and is therefore less likely to reach the best LP accuracy. Still regarding LP, MRR scores on different dimensions are similar, so that higher dimensions do not necessarily outperform lower ones, which led us to conclude higher k-dimensional spaces may be not required to model larger datasets (e.g. FB55T), but are more related to improving performance regarding the semantic complexity of the knowledge graph (e.g. BPA).

Along the classification task, all the classifiers achieved high accuracy (Figure 1c), regardless of the range of MRR. Although linear embedding models, like TransE and HEXTRATO, cannot directly capture the complexity of non-separable problems, their resulting embedding representation is somehow able to capture the semantic representativeness of entities and their relations, directly reflecting on high accuracy when the vector representation is used as input for more complex machine learning classification tasks. Interestingly, in mostly all the scenarios, an embedding model trained with a smaller margin tends to allow classifiers to achieve higher accuracy than those using embedding models trained with a larger margin.

Finally, for each k, we calculated Pearson's and Spearman's rank-order correlation coefficients to analyze the relationship between LP and classification metrics (Table 2). In most cases, there are some correlation, though not always strong. When correlation exists, the direction is usually positive for MRR although the opposite is observed in *Demographics*, *Pregnancy* and *BPA* with $k = 64$. LP and classification metrics do not fully agree on the overall quality of embedding representation, as their correlation is inconsistent and unreliable. The correlation between MRR$_r$ and classifier accuracy follows the same direction as MRR but the strength often varies. Despite focusing on the same relation, in most cases MRR$_r$ does not have strong correlation with classifier accuracy. This implies that the two evaluation tasks are not directly related and it is inaccurate to infer one from another.

In general, we observed that a learning margin between 0.5 and 1.5 results in consistently better quality of embedding representation. Larger values push the entities beyond the surface of the hyperspace which could produce much more noise than setting the regularization to $|x| = 1$, as opposed to $|x| \leq 1$ as in TransE (Bordes et al. 2013) and some other models (Fan et al. 2014; Bordes et al. 2011). In contrast, resulting accuracy from lower margin values ($\gamma < 0.5$) is frequently inferior possibly because the margin is not rigorous enough to help with the learning process.

Based on the classification results, we have strong evidence that the LP metrics do not directly reflect the effectiveness of using the resulting vector representation in specific classification tasks. While there may be some overall correlation between the metrics, the inconsistency makes it unreliable to base our judgment entirely on MRR in these situations. While the classifiers are focused on a very specific group of triples from each KG, these experiments allowed us to reflect on the extent of extrapolating LP performance to possible subsequent tasks. Especially in specific domains with abundant categorized data such as in Electronic Health Records, embedding representation may have more applications in classifiers for clinical decision support rather than knowledge graph completion, whereas the latter has been a major research motivation in the open domain.

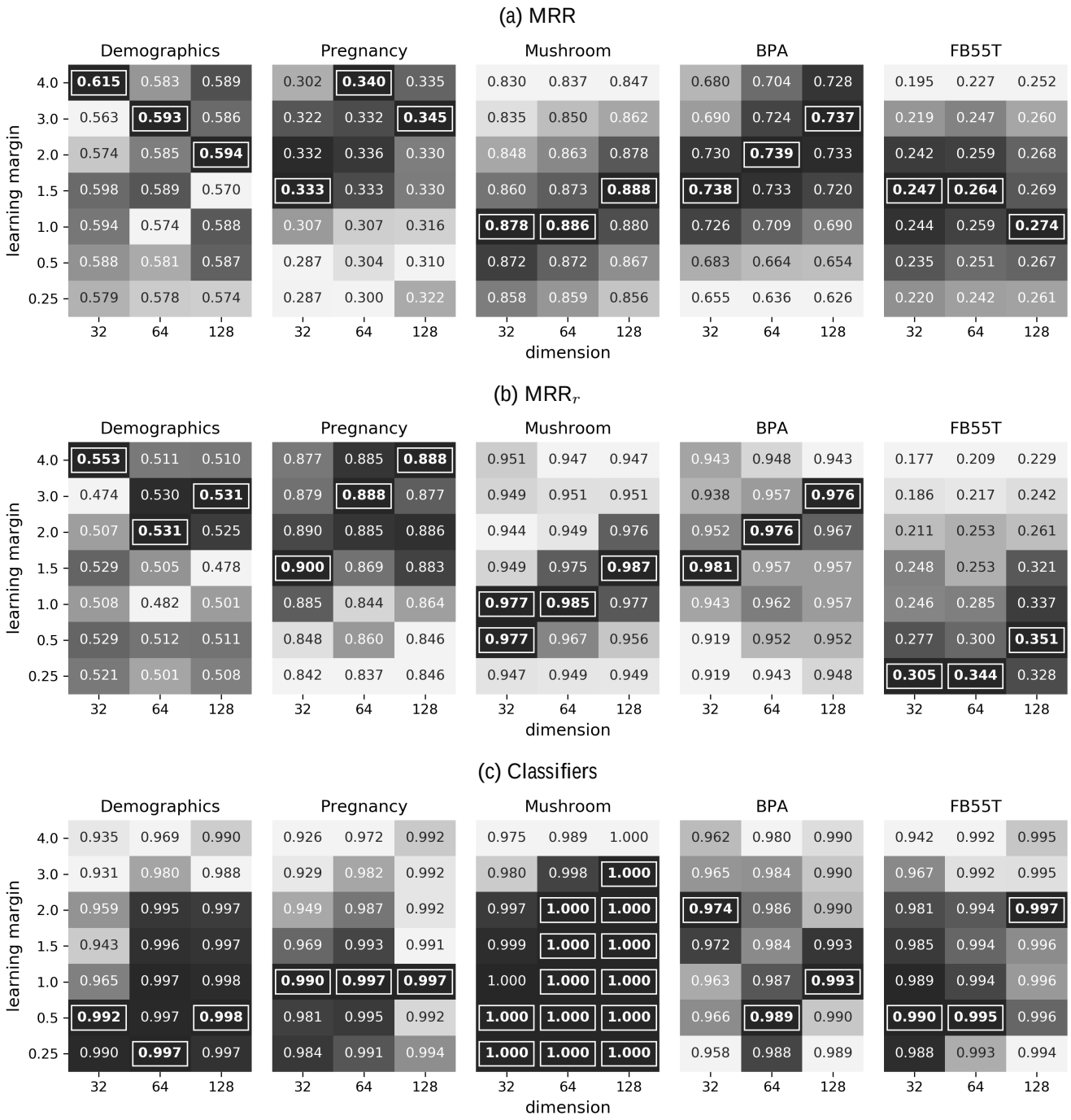

Figure 1: LP and classification metrics by learning margin and dimensionality (k) of each model. **(a)** overall MRR; **(b)** MRR for the target relation (MRR$_r$) used along the classification task; **(c)** classifier accuracy. The boxes are colored according to the scale of each column, ranging from the worst values in the lightest grey to the best values in the darkest grey. The metric score of each pair (γ, k) is annotated in the cell, and the best score is highlighted in a box.

Table 2: Pearson and Spearman correlation coefficients between LP metrics and corresponding model classifier accuracy for distinct model dimensionality $k \in \{32, 64, 128\}$ – MRR_r stands for the equivalent MRR of the classifier target relation.

Datasets	k = 32		k = 64		k = 128	
	MRR	**MRR_r**	**MRR**	**MRR_r**	**MRR**	**MRR_r**
Demographics	-0.137	0.156	-0.385	-0.344	-0.264	-0.507
Pregnancy	-0.407	-0.347	-0.695	-0.708	-0.500	-0.420
Mushroom	0.867	0.346	0.764	0.444	0.654	0.442
BPA	0.818	0.702	-0.471	0.094	0.211	0.160
FB55T	0.843	0.804	0.757	0.508	0.563	0.146

(a) Pearson correlation coefficients

Datasets	k = 32		k = 64		k = 128	
	MRR	**MRR_r**	**MRR**	**MRR_r**	**MRR**	**MRR_r**
Demographics	0.000	0.107	-0.714	-0.643	-0.143	-0.464
Pregnancy	-0.286	-0.143	-0.607	-0.721	-0.487	-0.400
Mushroom	0.793	0.174	0.802	0.539	0.612	0.612
BPA	0.821	0.564	-0.464	-0.090	0.286	0.631
FB55T	0.536	0.821	0.607	0.607	0.536	0.250

(b) Spearman correlation coefficients

Conclusions

This study focuses on the learning margin parameter as we observed several patterns of margin preference according to the dataset characteristics, model dimensionality, and evaluation task. Our findings provide preliminary evidence to understand the choice of hyperparameters in the context of learning representation for multi-relational categorized data, an area which lacks formal justification for hyperparameter optimization.

We trained and evaluated an embedding model with a range of learning margins while holding other hyperparameters constant on multiple categorized datasets with various sizes and shapes. Additionally, we analyzed multiple resulting models with different embedding vector dimensions to investigate any relationship between γ and k and their possible interactions. The evaluation is primarily focused on the standard LP task. However, subject to the intended usage of the embeddings such as for building a classifier, we question whether LP can directly reflect their effectiveness. Classifiers that predict a targeted relation in each knowledge graph are trained as a secondary evaluation task to assess the appropriateness of LP metrics.

Based on experimental results, we provide evidence that lower values for the margin parameter are not necessarily rigorous enough, whereas larger values produce much noise pushing the entities beyond to the surface of the hyperspace, leading to frequent regularization. More importantly, usual LP metrics do not necessarily represent the quality of resulting vector representation, as the correlation between LP and classification accuracy metrics tends to be weak.

Some of the ways in which this work can be extended include:

(a) Evaluating other combination of hyperparameters with respect to the data and task at hand, as such findings could be useful for working around the reliance on exhaustive grid search which is ineffective and is a huge barrier against widespread application of embedding representation for categorized multi-relational data.

(b) Gathering further evidence about the effectiveness of the learning margin regarding other ontological constraints as well as expanding the scope to other hyperparameter sets.

(c) Towards finding alternative ways of accurately evaluating an embedding representation model that corresponds to its intended use, we plan to test whether and in what extent clustering is able to replace traditional LP metrics within the embedding training and evaluation protocols.

(d) About the generalizability of our results, we may be still able to fit in a few more evaluation tasks.

Although robust conclusions cannot be drawn from these experiments performed with a single model which is an extension of TransE, we try to provide evidences that embedding models can be differently affected with the choice of hyperparameters. A more meaningful experimental setup must be considered in order to verify whether reported correlation results are a dataset shape-dependent phenomenon and lead to a more general conclusion. Now that we have a starting point in hyperparameter association and the implication of different evaluation tasks, we can scale up our experiments to include more datasets with variable sizes and shapes as well as using multiple evaluation tasks; however, we believe our current results adequately represent the trend in learning margin versus dimension and the difference between link prediction and classification metrics.

Finally, although HEXTRATO is a promising approach to learn knowledge embedding representation for multi-relational categorized data, there is still room for improvement. We plan to explore various additional modelling strategies and hyperparameter options including further ontology-based constraints, the use of projection matrices and distinct regularization constraints.

References

[Abujabal et al. 2017] Abujabal, A.; Yahya, M.; Riedewald, M.; and Weikum, G. 2017. Automated template generation for question answering over knowledge graphs. In *Proceedings of the 26th international conference on world wide web*, 1191–1200. International World Wide Web Conferences Steering Committee.

[Berant et al. 2013] Berant, J.; Chou, A.; Frostig, R.; and Liang, P. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1533–1544.

[Bollacker et al. 2008] Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, 1247–1250. New York, NY, USA: ACM.

[Bordes et al. 2011] Bordes, A.; Weston, J.; Collobert, R.; and Bengio, Y. 2011. Learning structured embeddings of knowledge bases. In *Conference on Artificial Intelligence*.

[Bordes et al. 2013] Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc. 2787–2795.

[Büttcher, Clarke, and Cormack 2010] Büttcher, S.; Clarke, C.; and Cormack, G. V. 2010. *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press.

[Dong et al. 2014] Dong, X.; Gabrilovich, E.; Heitz, G.; Horn, W.; Lao, N.; Murphy, K.; Strohmann, T.; Sun, S.; and Zhang, W. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, 601–610. New York, NY, USA: ACM.

[Fan et al. 2014] Fan, M.; Zhou, Q.; Chang, E.; and Zheng, T. F. 2014. Transition-based knowledge graph embedding with relational mapping properties. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*.

[Fellbaum 1998] Fellbaum, C., ed. 1998. *WordNet: an electronic lexical database*. MIT Press.

[Glorot and Bengio 2010] Glorot, X., and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W., and Titterington, D. M., eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS-10)*, volume 9, 249–256.

[He et al. 2015] He, S.; Liu, K.; Ji, G.; and Zhao, J. 2015. Learning to represent knowledge graphs with gaussian embedding. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, 623–632. New York, NY, USA: ACM.

[Kadlec, Bajgar, and Kleindienst 2017] Kadlec, R.; Bajgar, O.; and Kleindienst, J. 2017. Knowledge base completion: Baselines strike back. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 69–74. Vancouver, Canada: Association for Computational Linguistics.

[Kong et al. 2019] Kong, F.; Zhang, R.; Mao, Y.; and Deng, T. 2019. Lena: Locality-expanded neural embedding for knowledge base completion. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01):2895–2902.

[Lehmann et al. 2015] Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P. N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; and Bizer, C. 2015. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal* 6(2):167–195.

[Lin et al. 2015] Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, AAAI'15, 2181–2187. AAAI Press.

[Nickel, Rosasco, and Poggio 2016] Nickel, M.; Rosasco, L.; and Poggio, T. 2016. Holographic embeddings of knowledge graphs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, 1955–1961. AAAI Press.

[Nickel, Tresp, and Kriegel 2011] Nickel, M.; Tresp, V.; and Kriegel, H.-P. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, 809–816. USA: Omnipress.

[Robbins and Monro 1951] Robbins, H., and Monro, S. 1951. A stochastic approximation method. *Ann. Math. Statist.* 22(3):400–407.

[Shan et al. 2018] Shan, Y.; Bu, C.; Liu, X.; Ji, S.; and Li, L. 2018. Confidence-aware negative sampling method for noisy knowledge graph embedding. In *2018 IEEE International Conference on Big Knowledge (ICBK)*, 33–40.

[Suchanek, Kasneci, and Weikum 2007] Suchanek, F. M.; Kasneci, G.; and Weikum, G. 2007. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, 697–706. New York, NY, USA: ACM.

[Tissot and Dobson 2018] Tissot, H., and Dobson, R. 2018. Identifying misspelt names of drugs in medical records written in portuguese. *HealTAC-2018: Unlocking Evidence Contained in Healthcare Free-text*.

[Tissot 2018] Tissot, H. 2018. HEXTRATO: Using ontology-based constraints to improve accuracy on learning domain-specific entity and relationship embedding representation for knowledge resolution. In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2018, Volume 1: KDIR, Seville, Spain, September 18-20, 2018.*, 70–79.

[Trouillon et al. 2016] Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, E.; and Bouchard, G. 2016. Complex embeddings

for simple link prediction. In *Proceedings of the 34 Annual International Conference on Machine Learning (ICML)*.

[Wang et al. 2014] Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In Brodley, C. E., and Stone, P., eds., *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 1112–1119. AAAI Press.

[Wang et al. 2019] Wang, H.; Zhang, F.; Wang, J.; Zhao, M.; Li, W.; Xie, X.; and Guo, M. 2019. Exploring high-order user preference on the knowledge graph for recommender systems. *ACM Trans. Inf. Syst.* 37(3):32:1–32:26.

[Wang, Li, and Pan 2018] Wang, P.; Li, S.; and Pan, R. 2018. Incorporating GAN for negative sampling in knowledge representation learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 2005–2012.

[WHO 2004] WHO, W. H. O. 2004. *ICD-10 : international statistical classification of diseases and related health problems / World Health Organization*. World Health Organization Geneva, 10th revision, 2nd ed. edition.

[Xiao et al. 2015] Xiao, H.; Huang, M.; Hao, Y.; and Zhu, X. 2015. Transa: An adaptive approach for knowledge graph embedding. *CoRR* abs/1509.05490.

[Xiao, Huang, and Zhu 2016] Xiao, H.; Huang, M.; and Zhu, X. 2016. Transg : A generative model for knowledge graph embedding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2316–2325. Association for Computational Linguistics.

[Zhou et al. 2019] Zhou, Z.; Liu, S.; Xu, G.; and Zhang, W. 2019. On completing sparse knowledge base with transitive relation embedding. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01):3125–3132.