Binary Vector based Propositionalization Strategy for Multivalued Relations in Linked Data

Florian Jakobs, Yordan Terziev and Volker Gruhn

University of Duisburg-Essen, Essen, Germany
florian.jakobs@stud.uni-due.de,
{yordan.terziev,volker.gruhn}@paluno.uni-due.de

Abstract. Machine learning on linked data is strongly dependent on the selection of high quality data features to achieve good results and build reusable and generalizable models. In this work, we explore the problem of representing multivalued relations in a suitable form for machine learning while keeping the human comprehensibility of the resulting model. Specifically, we propose the use of a binary vector representation and compare it to two state of the art approaches. Our evaluation shows that the binary vector representation achieves mostly higher accuracy in comparison to standard propositionalization techniques. It also achieves comparable accuracy to a recently presented graph embeddings approach, while retaining the human comprehensibility.

Keywords: Propositionalization, Linked open data, Data mining

1 Introduction

Linked Open Data (LOD) has become one of the main approaches for combining structured and unstructured data from different domains. It provides context and relations between entities and connects them to other domains. In this way, LOD can be used as source for querying all kind of properties for entities of interest. The queried properties can be used as features for training machine learning (ML) models that subsequently can be applied to multiple real-world problems like classifying the entities or validating their relations to other entities.

However, a typical problem in ML on LOD is the generation and selection of features. To extract features from a LOD graph different transformations have to be performed in a process called propositionalization [1]. The different transformations may produce features for every outgoing/ingoing relation, binary features that represent the existence of a given relation or numerical features that count the number of relations of a certain type [2, 3]. Furthermore it is also possible to generate features using graph embedding techniques and graph sub-structures [2, 4]

In this work, we complement the proposed approaches for generating features with a previously unexplored technique for representing multivalued properties as binary vector. We evaluate this technique and compare it to the approaches published in [2].

2 Related Work

There are multiple ways to perform a propositionalization of multivalued features. On the one hand Ristoski and Paulheim [2] use the Weisfeiler-Lehman algorithm (WL) [4]in order to generate sub structures in the RDF graph. On the other hand, they take into consideration neural language modelling techniques [5, 6] based on random graph walks for latent feature generation (e.g. *W2V CBOW 500* in [2]).

While in these approaches the problem of representing multivalued features is non-existent, the generated features (latent or compounded) are not comprehensible for humans. Therefore, the resulting models cannot be used in domains where one needs to understand how the model makes its decisions.

Another approach for generating features is presented in [7], described in [3] and most recently evaluated in [2]. In this approach known as *rel-vals out* [2], the features are derived from generic relations-values for outgoing relations of an entity including the value of the relation. The multivalued relations on the other hand, are handled by counting every same named relation and using the sum as the attribute's value [7].

3 Approach

We propose a binary vector representation for propositionalization of multivalued relations that keeps the connected values. The implementation can be found on our GitHub account [8].These type of relations have been only indirectly covered by the counting approach in [3, 7]. The main goal of our approach is, to generate features that are comprehensible for humans and decisions made by the generated models are retraceable. This is crucial for critical domains like medicine, where ethical and legal concerns are at stake. To present our approach, we use the example RDF Graph in **Fig. 1** and the resulting representation in **Table 1**.

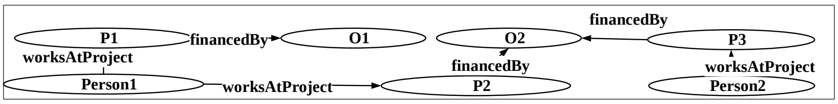
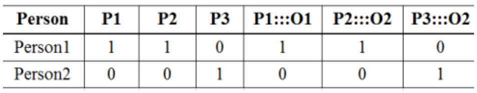


Fig. 1. Example RDF Graph

Let's say the feature generation starts at *Person1*. The approach performs a Breadth First Search (BFS) with a predefined depth given by the user. For the first iteration, we generate a feature for every entity and property reached by the BFS. We consider only outgoing relations and their target's value. Because of the relations presented in **Fig. 1**, we generate a feature for the projects a person is working at. This feature is multivalued, because a person can participate in multiple projects. For each subsequent iteration, we take the values of the previously generated features and perform a BFS on them until the predefined depth is reached. This results in new features like the financing organization (O1 or O2) of a project shown in **Fig. 1** that indirectly finances the person. After the final step of generating features, we transform the data into binary vectors. The example from **Fig. 1** results in the binary vectors (one per row) shown in **Table 1**.

Our final vectors consist of nominal values for single valued relations and of the values 0 and 1, which result from the transformation of each value of every multivalued relation. This means we produce a feature that for example links a specific project with its financier (e.g. P3:::O2), as shown in **Table 1**

Table 1. The resulting binary set of the data shown in Figure 1



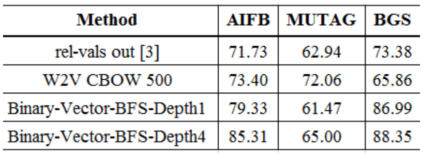
Person	P1	P2	P3	P1:::O1	P2:::O2	P3:::O2
Person1	1	1	0	1	1	0
Person2	0	0	1	0	0	1

4 Evaluation

We compare the here presented approach to *rel-vals out* and *W2V CBOW 500* [2] on the datasets presented in [9]. We used the same small datasets and the DBPedia datasets as presented in [2] and also recreated the presented evaluation environment. Currently we measured the performance accuracy using only C4.5 ML algorithm, because decision trees are closest to the human understanding of decision making thus keeping the model comprehensible. For building the decision trees we used Weka Version 3.6.13.

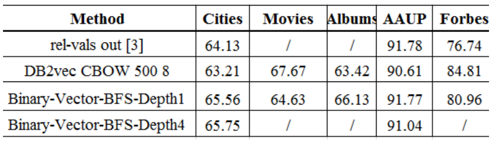
Table 2 and **Table 3** show the measured accuracy of the binary vector representation in comparison to the results published in [2]. The results show that our approach executed with BFS depth 1 outperforms rel-vals out in all datasets, except MUTAG. For the Albums and Movies dataset Ristoski et. al. state that their approach [2] does not terminate within 10 days or has run out of memory. We compare these two approaches, because both only take the directly connected values in consideration.

Table 2. Classification accuracy in percent on the small datasets.



Method	AIFB	MUTAG	BGS
rel-vals out [3]	71.73	62.94	73.38
W2V CBOW 500	73.40	72.06	65.86
Binary-Vector-BFS-Depth1	79.33	61.47	86.99
Binary-Vector-BFS-Depth4	85.31	65.00	88.35

Table 3. Classification accuracy in percent on the large datasets.



Method	Cities	Movies	Albums	AAUP	Forbes
rel-vals out [3]	64.13	/	/	91.78	76.74
DB2vec CBOW 500 8	63.21	67.67	63.42	90.61	84.81
Binary-Vector-BFS-Depth1	65.56	64.63	66.13	91.77	80.96
Binary-Vector-BFS-Depth4	65.75	/	/	91.04	/

W2V CBOW 500 and *DB2vec CBOW 500 8* on the other hand, use paths with length 8 (node hop distance of 4) to generate sentences. Thus, we compare these techniques to *Binary-Vector-BFS-Depth4*. On the small datasets, our approach outperforms the *W2V CBOW 500* except on MUTAG. On the large datasets similarly to [2] our approach failed to terminate after 3 days on the Albums, Movies and Forbes datasets. On the other two datasets (Cities and AAUP) our approach achieved slightly better accuracy.

5 Conclusion

Having the goal of comprehensibility in mind the proposed representation as binary vectors provides a better understanding of the resulting model in comparison to latent features generated by RDF2Vec [2]. The approach outperforms the standard propositionalization strategy on almost all datasets. Comparing the binary vector representation to the word embeddings approach showed that on the small datasets, it achieved mostly higher accuracy. On the large datasets the accuracy was comparable, however, our approach didn't terminate within 3 days on some of them. Our assumption is that for the smaller datasets the neural network couldn't be trained sufficiently, and therefore performed with lower accuracy. However, this assumption has to be further evaluated in our future work. Furthermore, we would like to experiment with other machine learning models, improve accuracy by restricting the relations to semantic correct ones and experiment with Depth First Search as graph traversal strategy.

Acknowledgments. The presented work has been partly funded by the German Federal Ministry of Education and Research (BMBF, FKZ 16SV7120K) SIM4BGM

References

1. Džeroski, S., Lavrač, N. (eds.): Relational Data Mining. Springer Berlin Heidelberg, Berlin, Heidelberg (2001)
2. Ristoski, P., Paulheim, H.: Rdf2vec. Rdf graph embeddings for data mining. In: International Semantic Web Conference, pp. 498–514 (2016)
3. Ristoski, P., Paulheim, H.: A comparison of propositionalization strategies for creating features from linked open data. Linked Data for Knowledge Discovery 6 (2014)
4. Shervashidze, N., Schweitzer, P., van Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-lehman graph kernels. Journal of Machine Learning Research 12, 2539–2561 (2011)
5. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
6. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp. 3111–3119 (2013)
7. Paulheim, H., Fümkranz, J.: Unsupervised generation of data mining features from linked open data. In: Proceedings of the 2nd international conference on web intelligence, mining and semantics, p. 31 (2012)
8. Jakobs, F.: ba_edtEvaluation, https://github.com/Floishy/ba_edtEvaluation
9. Ristoski, P., Vries, G.K.D. de, Paulheim, H.: A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In: International Semantic Web Conference, pp. 186–194 (2016)