The Linked Data Mining Challenge 2016

Petar Ristoski[1], Heiko Paulheim[1], Vojtěch Svátek[2], and Václav Zeman[2]

[1] University of Mannheim, Germany
Research Group Data and Web Science
{petar.ristoski,heiko}@informatik.uni-mannheim.de
[2] University of Economics, Prague, Czech Republic
Department of Information and Knowledge Engineering
{svatek,vaclav.zeman}@vse.cz

Abstract. The 2016 edition of the Linked Data Mining Challenge, conducted in conjunction with Know@LOD 2016, has been the fourth edition of this challenge. This year's dataset collected music album ratings, where the task was to classify well and badly rated music albums. The best solution submitted reached an accuracy of almost 92.5%, which is a clear advancement over the baseline of 69.38%.

1 The Linked Data Mining Challenge Overview

Linked Open Data [9] has been recognized as a valuable source of background knowledge in many data mining tasks and knowledge discovery in general [7]. Augmenting a dataset with features taken from Linked Open Data can, in many cases, improve the results of a data mining problem at hand, while externalizing the cost of maintaining that background knowledge [4]. Hence, the primary goal of the Linked Data Mining Challenge 2016 is to show how Linked Open Data and Semantic Web technologies could be used in a real-world data mining task.

This year, the Linked Data Mining Challenge was held for the fourth time, following past editions co-located with *DMoLD* (at ECML/PKDD) [11], Know@-LOD 2014 [10] and Know@LOD 2015 [8]. The challenge consists of one task, which is the prediction of the review class of music albums. The dataset is generated from real-world observations, linked to a LOD dataset and it is used for a two-class classification problem.

The rest of this paper is structured as follows. Section 2 discusses the dataset construction and the task to be solved. In section 3, we discuss the entrants to the challenge and their results. We conclude with a short summary and an outlook on future work.

2 Task and Dataset

The 2016 edition of the challenge used a dataset built from music albums recommendations, turned into a two-class classification problem.

2.1 Dataset

The task concerns the prediction of a review of music albums, i.e.,"good" and "bad". The initial dataset is retrieved from Metacritic.com[3], which offers an average rating of all time reviews for a list of music albums[4]. Each album is linked to DBpedia [3] using the album's title and the album's artist. The initial dataset contained around 10,000 music albums, from which we selected 800 albums from the top of the list, and 800 albums from the bottom of the list. The ratings were used to divide the albums into classes, i.e., albums with score above 79 are regarded as "good" albums, while albums with score less than 63 are regarded as "bad" albums. For each album we provide the corresponding DBpedia URI. The mappings can be used to extract semantic features from DBpedia or other LOD repositories to be exploited in the learning approaches proposed in the challenge.

The dataset was split into training and test set using random stratified split 80/20 rule, i.e., the training dataset contains 1,280 instances, and the test dataset contains 320 instances. The training dataset, which contains the target variable, was provided to the participants to train predictive models. The test dataset, from which the target label is removed, is used for evaluating the built predictive models.

2.2 Task

The task concerns the prediction of a review of albums, i.e.,"good" and "bad", as a classification task. The performance of the approaches is evaluated with respect to accuracy, calcuated as:

$$Accuracy = \frac{\#true\ positives + \#true\ negatives}{\#true\ positives + \#false\ positives + \#false\ negatives + \#true\ negatives} \quad (1)$$

2.3 Submission

The participants were asked to submit the predicted labels for the instances in the test dataset. The submissions were performed through an online submission system. The users could upload their prediction and get the results instantly. Furthermore, the results of all participants were made completely transparent by publishing them on an online real-time leader board (Figure 1). The number of submissions per user was not constrained.

In order to advance the increase of Linked Open Data [9] available as a side-effect of the challenge, we allowed users to also exploit non-LOD data sources, given that they transform the datasets they use to RDF, and provide them publicly. Since the Metacritic dataset is publicly available, the participants were asked not to use the Metacritic music albums' rating score to tune the predictor for the albums in the test set.

[3] http://www.metacritic.com/
[4] http://www.metacritic.com/browse/albums/score/metascore/all

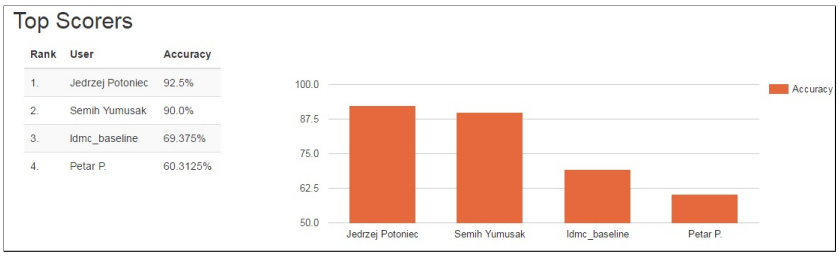

Fig. 1: Participants Results

3 The Linked Data Mining Challenge results

In total, three parties participated in the challenge. We compare those results against a baseline approach.

3.1 Baseline Models

We provide a simple classification model that serves as a baseline. In this baseline approach, we use the albums' DBpedia URI to extract the direct types and categories of each album. On the resulting dataset, we built a k-NN classifier (k=3), and applied it on the test set, scoring an accuracy of 69.38%. The model is implemented in the RapidMiner platform[5], using the Linked Open Data extension [7], and it was publicly available for the participants.

3.2 Participants' Approaches

During the submission period, three teams completed the challenge by submitting a solution to the online evaluation system and describing the used approach in a paper. In the following, we describe and compare the final participant approaches. A summary of all approaches is given in Table 1.

Jedrzej Potoniec. Not-So-Linked Solution to the Linked Data Mining Challenge 2016 [6]

By Jedrzej Potoniec

In this approach, the authors extract features from several non-LOD datasets, which are then used to build a Logistic Regression model for classification of albums. To extract the features, the authors start by scraping the *Wikipedia* pages for the given albums using the *Scrapy* tool[6]. From the collected data, the

[5] http://www.rapidminer.com/
[6] http://scrapy.org/

authors focus on the album reviews and ratings. Furthermore, reviews and ratings are collected from *Amazon*[7] and *Discogs*[8], while *MusicBrainz*[9] is used to obtain the number of users owning an album and its average score. The final dataset contains 94 numerical attributes in total.

To train the classification model, the authors use logistic regression, using the RapidMiner platform. Before training the model, a Z-transformation is performed on all attributes, so all attributes have an average of 0 and standard deviation 1. The authors perform 10-fold cross-validation on the training dataset, achieving accuracy of 91.7%. This value is consistent with 92.5% on the test set reported by the challenge submission system.

Furthermore, the authors provide some insights on the relevance of the features for the classification task, based on the learned logistic regression coefficients for each attribute. For example, the results show that Metacritic ratings highly correlate with ratings from other sources, like *Pitchfork*[10], *AllMusic*[11], *Stylus*[12], and others.

The code and the data can be found online[13].

Semih Yumusak. A Hybrid Method for Rating Prediction Using Linked Data Features and Text Reviews [12]

By Semih Yumusak, Emir Muñoz, Pasquale Minervini, Erdogan Dogdu, and Halife Kodaz

In this approach, the authors use Linked Open Data features in combination with album reviews to build seven different classification models. DBpedia is used as a main source for LOD features. More precisely, the authors manually select predicates that might be relevant for the given classification task. Along the direct predicate values, aggregate count features are used as well. Besides the LOD features, the authors also use albums' reviews retrieved from Metacritic as textual features. The reviews are first preprocessed, i.e., lower-case transformation, non-alphanumeric normalizations, stopwords removal and stemming, then standard Bag-of-Words is used to represent each review.

Furthermore, the authors identify that discretizing some of the features leads to better representation of the data, e.g., the award feature of an artist could be marked as "high" if the number of awards is more than one, and "low" otherwise.

In the next step, the authors experiment with seven different classification models, i.e., linear SVM, KNN, RBF SVM, Decision Trees, Random Forest, AdaBoost, and Naive Bayes. The hyper parameters for each model are determined

[7] http://www.amazon.com/
[8] https://www.discogs.com/
[9] https://musicbrainz.org/doc/MusicBrainz_Database/Download
[10] http://pitchfork.com/
[11] http://www.allmusic.com/
[12] http://www.stylusmagazine.com/
[13] https://github.com/jpotoniec/LDMC2016

manually via incremental tests, and results extracted from the training set. They evaluate each model on the training dataset using 10-fold cross-validation. The experiments were performed using the sckit-learn library[14]. The best performance on the training dataset is achieved using Linear SVM with an accuracy of 87.81%. Applying the same model on the test set scores an accuracy of 90.00%, which confirms that the model is not overfitted on the training dataset. The authors evaluate the relevance of different features group separately, showing that most of the models perform the best when using both LOD and text-based features.

Furthermore, the authors provide some interesting observations about the task. For example, "Bands are more successful than single artists", "Shorter albums are likely to be worse", "The genre of the album indicates if the album is good or bad", and others.

The source files, the crawler code and the reviews, the enriched knowledge base in RDF, and the intermediate files are published as an open-source repository[15].

Petar P. Can you judge a music album by its cover? [5]

By Petar Petrovski and Anna Lisa Gentile

In this approach, the authors use an unconventional method for the task of music album classification. They explore the potential role of music album cover arts for the task of predicting the overall rating of music albums and investigate if one can judge a music album by its cover alone. The proposed approach for album classification consists of three main steps. Given a collection of music albums, the authors first obtain the image of their cover art using DBpedia. Then, using off-the-shelf tools obtain a feature vector representation of the images. In the final step, a classifier is trained to label each album, only exploiting the feature space obtained from its cover art.

To extract the image features, the authors use the Caffe deep learning framework [2], which also provides a collection of reference models, which can be used retrieving image feature vectors. More precisely, the authors use the bvlc model[16], which consists of five convolutional layers, and three fully-connected layers, and it is trained on 1.2 million labeled images from the ILSVRC2012 challenge[17]. To obtain features for each image, the output vectors of the second fully-connected layer of the model are used. Such features capture different characteristics of images, e.g., colors, shapes, edges etc.

To build a classification model, the authors use linear SVM model. The model is evaluated on the training set using 10-fold cross-validation, achieving accuracy of 58.30%. The accuracy of the model on the test set is 60.31%. Hence, the results

[14] http://scikit-learn.org/
[15] https://github.com/semihyumusak/KNOW2016
[16] bvlc_reference_caffenet from caffe.berkeleyvision.org
[17] http://image-net.org/challenges/LSVRC/2012/

Table 1: Comparison of the participants approaches.

Approach	Classification methods	Knowledge Source	Tools	Score	Rank
Jedrzej Potoniec	Logistic Regression	Wikipedia, Amazon, Discogs, MusicBrainz	RapidMiner, Scrapy	92.50%	1
Semih Yumusak	SVM, KNN, Decision Trees, Random Forest, AdaBoost, Naive Bayes	DBpedia, Metacritic	sckit-learn lib	90.00%	2
Petar P.	SVM	DBpedia	RapidMiner, LOD extension, Caffe	60.31%	3

show that using only features extracted from the album cover arts is not sufficient for the given classification task.

The dataset is available online[18], along with the extracted feature vectors and used processes.

3.3 Meta Learner

We made a few more experiments in order to analyze the agreement of the three submissions, as well as the headroom for improvement.

For the agreement of the three submissions, we computed the Fleiss' kappa score [1], which is 0.373. This means that there is not a good agreement of the three approaches about what makes good and bad albums. We also calculated the Fleiss' kappa score for the top two approaches, which is 0.687. This means that there is a good, although not perfect agreement of the top two approaches.

To exploit advantages of the three approaches, and mitigate the disadvantages, we analyzed how a majority vote of the three submissions would perform. The accuracy totals at 90.00%, which is lower than the best solution submitted. This shows that the majority vote is highly influenced by the low scoring submission using the image features, which does not outperform the baseline. We also perform a weighted majority vote, using the achieved accuracy of each approach as the weight. The accuracy totals at "92.50", which is same as the best solution submitted.

4 Conclusion

In this paper, we have discussed the task, dataset, and results of the Linked Data Mining Challenge 2016. The submissions show that Linked Open Data is a

[18] https://github.com/petrovskip/know-lod2016

useful source of information for data mining, and that it can help building good predictive models.

One problem to address in future editions is the presence of false predictors. The dataset at hand, originating from Metacritic, averages several ratings on albums into a final score. Some of the LOD datasets used by the competitors contained a few of those original ratings, which means that they implicitly used parts of the ground truth in their predictive models (which, to a certain extent, explains the high accuracy values). Since all of the participants had access to that information, a fair comparison of approaches is still possible; but in a real-life setting, the predictive model would perform sub-optimally, e.g., when trying to forecast the rating of a *new* music album.

In summary, this year's edition of the Linked Data Mining challenge showed some interesting cutting-edge approaches for using Linked Open Data in data mining. As the dataset is publicly available, it can be used for benchmarking future approaches as well.

Acknowledgements

We thank all participants for their interest in the challenge and their submissions. The preparation of the Linked Data Mining Challenge and of this paper has been partially supported by the by the German Research Foundation (DFG) under grant number PA 2373/1-1 (Mine@LOD), and by long-term institutional support of research activities by the Faculty of Informatics and Statistics, University of Economics, Prague.

References

1. Joseph L Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intra-class correlation coefficient as measures of reliability. *Educational and psychological measurement*, 1973.
2. Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.
3. Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, 2013.
4. Heiko Paulheim. Exploiting linked open data as background knowledge in data mining. In *Workshop on Data Mining on Linked Open Data*, 2013.
5. Petar Petrovski and Anna Lisa Gentile. Can you judge a music album by its cover? In *5th Workshop on Knowledge Discovery and Data Mining meets Linked Open Data (Know@LOD)*, 2016.
6. Jedrzej Potoniec. Not-so-linked solution to the linked data mining challenge 2016. In *5th Workshop on Knowledge Discovery and Data Mining meets Linked Open Data (Know@LOD)*, 2016.

7. Petar Ristoski, Christian Bizer, and Heiko Paulheim. Mining the web of linked data with rapidminer. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:142–151, 2015.
8. Petar Ristoski, Heiko Paulheim, Vojtech Svátek, and Vaclav Zeman. The linked data mining challenge 2015. In *5th Workshop on Knowledge Discovery and Data Mining meets Linked Open Data (Know@LOD)*, 2015.
9. Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In *The Semantic Web–ISWC 2014*, pages 245–260. Springer, 2014.
10. Vojtěch Svátek, Jindřich Mynarz, and Heiko Paulheim. The linked data mining challenge 2014: Results and experiences. In *3rd International Workshop on Knowledge Discovery and Data Mining meets Linked Open Data*, 2014.
11. Vojtěch Svátek, Jindřich Mynarz, and Petr Berka. Linked Data Mining Challenge (LDMC) 2013 Summary. In *International Workshop on Data Mining on Linked Data (DMoLD 2013)*, 2013.
12. Semih Yumusak, Emir Muñoz, Pasquale Minervini, Erdogan Dogdu, and Halife Kodaz. A hybrid method for rating prediction using linked data features and text reviews. In *5th Workshop on Knowledge Discovery and Data Mining meets Linked Open Data (Know@LOD)*, 2016.