Enhancing white-box machine learning processes by incorporating semantic background knowledge

Gilles Vandewiele

Department of Information Technology
Ghent University - imec, IDLab
gilles.vandewiele@intec.ugent.be

Abstract. Currently, most of white-box machine learning techniques are purely data-driven and ignore prior background and expert knowledge. A lot of this knowledge has already been captured in domain models, i.e. ontologies, using Semantic Web technologies. The goal of this research proposal is to enhance the predictive performance and required training time of white-box models by incorporating the vast amount of available knowledge in the pre-processing, feature extraction and selection phase of a machine learning process.

Keywords: White-box machine learning, knowledge incorporation, semantic knowledge bases

1 Introduction

Most machine learning techniques are data-driven and thus ignore most of the vast amounts of existing available knowledge already captured in domain models [1], such as SNOMED [2], SSN [3] and UMLS [4]. The advantage of applying a purely data-driven approach is that the model is robust to outliers and noise. The disadvantage is that a computationally expensive training phase needs to be executed and that valuable prior knowledge is not taken into account. In many critical domains such as electronic health care and law enforcement, wherein wrong decisions made can have significant repercussions, knowledge-based systems such as expert systems were long preferred [5] as they can easily give a comprehensible corresponding explanation with their predictions. Moreover, they can be deployed without requiring a lot of data, which was rather hard to collect prior to the big data era. The main disadvantage of a purely knowledge-based approach is that the performance is completely biased to the content of the knowledge base [6], which can take a lot of time to construct and maintain, and that it is not able to learn new patterns or insights. Moreover, this approach is often not robust, e.g. in the case of conflicting rules or samples that do not comply to any of the defined rules.

Within the data-driven approaches, two large families of techniques can be distinguished. First, there are black-box techniques, such as artificial neural networks, which are often able to learn features automatically, thus not requiring a feature extraction and selection phase, and tend to achieve high predictive performances [7]. However, they cannot provide an explanation for their predictions, making them impractical in applications where decision support, instead

of decision making, is crucial. Secondly, white-box techniques, such as decision tree induction and classification rule mining, construct an easily comprehensible predictive model from the data. While the predictive performance of these technique tends to be lower than their counterpart, they are able to give a corresponding explanation, therefore being ideally suited to provide decision support for experts within critical domains.

Given the advantages of both data-driven and knowledge-based approaches, advancements within the machine learning domain, the growth of data within all domains [8] and the vast amount of prior knowledge already available on the Semantic Web, a hybrid approach seems to be ideal. In such an approach, a white-box predictive model, such as a decision tree or an ordered rule list, is constructed from the given data with incorporation of prior knowledge in each of its steps. Ideally, the advantages of both approaches would be retained, i.e. robustness to outliers and noise, ability to give a corresponding explanation, a less expensive and more performant training phase and the ability to deduce new insights and knowledge.

The remainder of this paper is as follows. A use case which will be used as a running example throughout the rest of this paper is presented in Section 2, followed by a discussion of the related work in Section 3. A problem statement with corresponding hypotheses and research questions are presented in Section 4. A methodology to provide an answer on these research questions in proposed in Section 5. Then, we discuss how our future research will be evaluated in Section 6 and finally, a conclusion is given in Section 7.

2 Use case: primary headache diagnosis

Primary headaches [9] are an increasingly common health issue in modern society, having a large prejudicial impact. In Europe, it has a prevalence of more than 50% and according to the World Health Organization (WHO), severe headache attacks are one of the top 10 most disabling conditions [10]. Currently, it costs a lot of time to diagnose a patient correctly because a lot of different aspects need to be taken into account and because many different types of primary headache exist. Furthermore, a lot of research by medical experts has already been done in the headache domain, resulting in a vast amount of available prior domain and expert knowledge [11]. Therefore, the automatic diagnosis of primary headaches seems an ideal use case to combine both the data-driven and knowledge-driven approach which can have a very positive impact. For my master dissertation, a mobile headache journal[1] was developed that allows headache patients to register their headache attacks and medicine consumptions. The semantically annotated data generated by this mobile application, in combination with background knowledge [4], can be used to generate a decision tree in order to support an expert in making a correct diagnosis. An overview of this workflow can be found in Figure 1. This use case will be used as a running example throughout this paper and, in addition to well-known benchmark datasets, to evaluate the different proposed techniques.

[1] https://play.google.com/store/apps/details?id=be.ugent.chronicals&hl=en

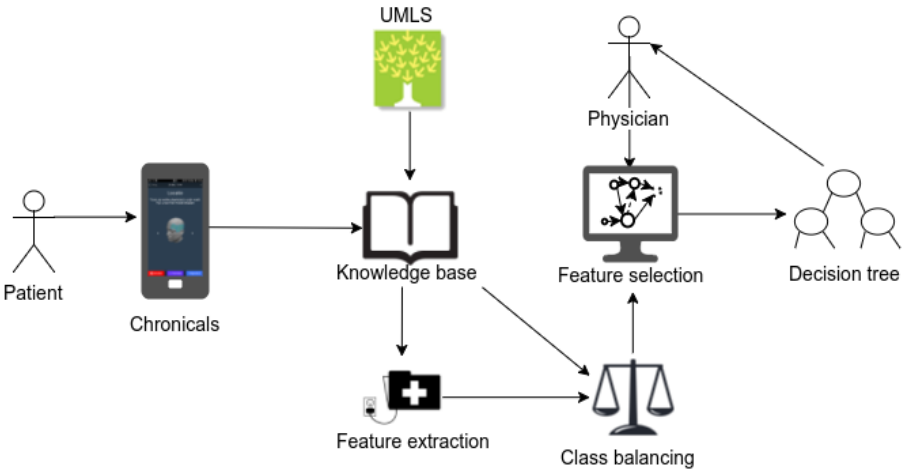

Fig. 1: Schematic overview of the machine learning work-flow, with incorporation of prior knowledge into the different phases, to diagnose a patient with primary headache.

3 Related work

Combining the advantages of knowledge-driven and data-driven approaches, sometimes referred to as semantic meta-mining, has been investigated before. Two very thorough and recent surveys can be found in [12, 13]. A traditional white-box data-driven approach consists of several main steps, which can be identified in Figure 2. In a first step, numerical features that have a high discriminative power are extracted from the raw data, which is optionally pre-processed first. Pre-processing examples include applying transformations to the data or generating and removing samples to balance the dataset. When all features are extracted, a selection phase is applied in order to discard the uninformative features, which allows for better generalization. Finally, a white-box model is constructed from the selected features. In the following subsections, related work for each these phases is presented.

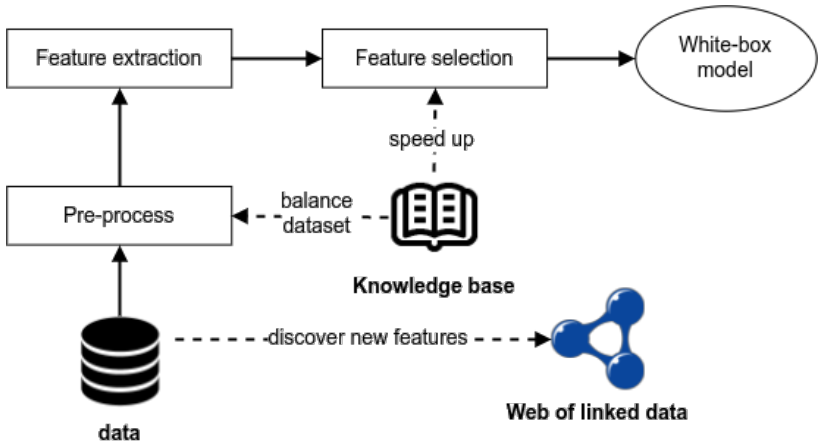

Fig. 2: The different steps of a white-box machine learning approach and how prior knowledge can be incorporated.

3.1 Automatic feature discovery

In a typical machine learning work-flow, a very large amount of time is spent on data cleaning and feature extraction. Generic features, which can be applied in a large number of problems, are available, but often, the most efficient features require some prior knowledge about the task to solve. Facilitating this feature extraction process by exploiting the concept of linked data to automatically discover new informative features could therefore significantly reduce the time required to create a predictive model. In order to do this, entities in the training set are mapped to a URI which corresponds to a node in the graph of linked data. From here, we can traverse edges to discover new features [14–16]. While this is a very interesting approach, there are many possible optimizations left, such as automatic measurements of feature importance, heuristics to decide when to stop traversing the immensely large graph and pruning parts of the graph in order to reduce the gigantic search space.

3.2 Class balancing

In the classification domain, a dataset is called imbalanced when the distribution of the classes in the training set is skewed. An imbalanced dataset is very common in the financial and medical domain, e.g. fraud and epilepsy detection respectively. Class imbalance gives rise to a few potential problems. First, the classifier will be biased towards the largest populated class as this has the highest impact on the objective function it is trying to optimize, while this is often the class of least importance to the expert. Second, general metrics, such as accuracy, to evaluate the model give a wrong representation of the predictive performance [17, 18]. Two large approaches to tackle with data imbalance can be identified. On the one hand sampling techniques can remove or create new samples in order to make the distribution of the classes more uniform [19]. On the other hand, the classification algorithm can be modified (e.g. adapting the objective function) to pay more attention to samples in the minor class [20, 21]. Sampling techniques are very interesting, as they can be applied as a pre-processing step of the machine learning work-flow, and can therefore be seen as model-agnostic. Sampling techniques can be divided in either oversampling, where the number of samples in the minor class in increased, or undersampling, where the number of samples in the major class is decreased. In current state-of-the-art oversampling algorithms, such as SMOTE [22] and its adaptations and ADASYN [23], virtual samples of the minority class are generated by using the small amount of data available and thus no prior knowledge is used. On the other hand, researchers have already attempted to generate 'virtual' samples solely based on the prior knowledge available [24–28]. While the latter research attempts were not done in the context of imbalanced dataset but more in the context of data augmentation, a hybrid approach, which combines the positive characteristics of both approaches, can be very interesting.

3.3 Feature selection

When all of the possible features are extracted from the raw data, a selection phase can optionally be applied in order to remove uninformative features. This

can mitigate the curse of dimensionality and thus possibly increases the generalization capability of the model while reducing the amount of training time required. Research for incorporating prior knowledge into the selection phase is a very young and pre-mature research field. In Ringsquandl et al. [29], the Semantic Sensor Network (SSN) ontology [3] is adapted to allow for automatic feature selection. Here, features are selected based on dependency relations defined by an expert between predictor variables or between a predictor variable and the target variable. This technique has a lower computational complexity than current feature selection techniques, as it is dependent only on the number of features and not on the number of data samples, which can become very large in many cases. Moreover, in contrast to dimensionality reduction techniques such as t-SNE [30] and PCA [31], interpretability of the features is maintained and the selection phase only has to be re-applied when new features are added to the model, instead of when a certain amount of new samples is added. Unfortunately, this technique is still rather simplistic and is equivalent to manual feature selection.

4 Problem statement

By analysis of the state of the art, one open problem can be identified:

P1 Current white-box machine learning techniques learn from scratch and often only use a limited amount of information (i.e. the training set) as they do not make full use of the vast amount of prior background and expert knowledge available in ontologies and on the web of linked data [32].

From this, the following hypotheses can be deducted:

H1 The automatic discovery of new features by exploiting the concept of linked data can lead to a reduction in the labor needed for feature extraction while resulting in an increase in the predictive performance of the model.
H2 Balancing the dataset using both knowledge and the limited amount of samples in the minority class will result in a better predictive performance for the minority class than sampling methods that are based only on this limited amount of samples.
H3 Applying feature selection based on a ranked list of features, generated by applying a ranking algorithm on a graph of features defined by an expert, will require less time than current feature selection techniques and result in a better generalization capability. Moreover, it allows for experts to have more control of the algorithm, which can increase their will to adopt such a system.

To deliver proof for the given hypotheses, the following research questions will be resolved:

Q1 Can we improve existing or develop new techniques that map the entities in the dataset to a URI identifiable on the web of linked data in order to traverse the graph of data to extract new relevant, discriminant features for the task to solve.
Q2 Can we develop a hybrid technique that uses both the limited amount of samples in the minority class and the knowledge about the minority class in order to generate new samples to balance the dataset? Moreover, how does this hybrid technique compare to the techniques where only one of the two is used?

Q3 Is it possible to improve the feature selection phase by creating a new algorithm that ranks the different features based on their relations defined by an expert?

5 Methodology

5.1 Automatic feature discovery

In order to augment the data with information from the web of linked data, a mapping phase must first be applied. Here, the entities in the initial dataset are mapped on a URI identifiable on the web of linked data or on a semantically annotated electronic health record in the medical domain. This mapping has to occur with minimal user interaction. When each of the samples are mapped on a URI, we can try to find new features by doing a breadth-first search in the graph of linked data. The reason for a breadth-first strategy is because of the almost infinite depth of the graph. In order for a new candidate feature to be informative, not too many missing values may occur and there must be correlation with the target variable (or must improve the cluster quality in the unsupervised case). Since counting the number of missing values and calculating correlations between a new candidate predictor and the target variable for a large dataset can take a significant amount of time, a subset of the initial dataset can be used to provide an approximation. Moreover, to decide heuristically which feature-threshold combination results in the most optimal split of data from all possible candidates, the Hoeffding bound [33, 34] can be applied. Since the graph we are traversing has an immense size, we need to define conditions when to stop the search, e.g. stop when we traversed k levels deeper in the graph without finding a new usable feature. Finally, pruning of the graph can optionally be applied by calculating semantic concept relatedness [35, 36] between a new subject and the target concept. When there is almost no semantic relation between a new concept and the target concept, that part of the graph can already be pruned. Many different metrics exist to calculate this relatedness [37, 38]. I will perform a clear evaluation of different metrics in order to find the most suited one for this task.

For the headache use case, a user profile in the mobile headache journal needs to be mapped to the patient's corresponding semantically annotated electronic health record. This can be done by joining on unique identifiable information such as the combination of name and email. As the electronic health record is semantically annotated, it can be seen as a graph, which can be traversed to discover new informative features that help in formulating a correct diagnosis for a primary headache patient. Moreover, datasets ideally suited for evaluation of this technique exist. Examples include the zoo dataset from UCI [39] and the datasets curated by the University of Mannheim [40]. The property of these datasets is that they contain a limited amount of information about rich concepts (such as cities or animals), and therefore rely on automatic feature discovery to obtain reasonable results.

5.2 Class balancing

In order to balance the classes, oversampling as a pre-processing step will be investigated, enabling a model-agnostic approach. I will create a hybrid approach

that combines the positive characteristics of data-based sampling algorithms, such as SMOTE [22] and ADASYN [23] and knowledge-based sampling algorithms, where samples are generated that comply to a pre-defined knowledge base. First, consistency of the knowledge base or the given data needs to be checked by evaluating whether the small amount of samples in the minority class complies to this knowledge. If this is not the case, there is either an anomaly in the data or a inconsistency/fault in the knowledge base that needs to be resolved. When we find that a certain fraction of the samples in the minority class do not comply to one specific rule in the knowledge base, chance are high that the rule is inconsistent with the ground truth and we can remove the rule. Else, the sample is probably an anomaly and can therefore be removed. Alternatively, both the rule and the sample can be removed. An evaluation is required to determine which technique (and threshold on the fractions of samples) is most suited for a dataset with certain properties. After this phase, data can be generated based on the knowledge base and on the small amount of samples in our dataset. For each dimension (i.e. feature) for which knowledge is available, these dimensions of a new virtual sample are set to values that comply to this defined knowledge (e.g. the value must be in a certain range). Of course, it is infeasible to have complete information about each dimension. For these dimensions, the values of samples in our dataset can be used as follows: we find the two nearest neighbors to our new virtual sample in the feature space defined by the features of which knowledge is available; then we can generate a random point on the link between these two neighbors.

One of the most severe primary headache types is cluster headache. It has been discovered quite recently and is a rather and rare condition, with a prevalence of 1 out of 1000 [41] as opposed to 1 out of 7 for migraine [42], making it very hard to diagnose. This, in combination with the fact that a lot of domain knowledge is available [11], makes it an ideal use case to evaluate the new technique on.

5.3 Feature selection

I will design a method that allows to represent the knowledge base as a graph, where each feature defined in the knowledge base or dataset corresponds to a node, and each relation between two features (such as dependsOn or independentOf) corresponds to an edge between their two corresponding nodes. I will then rehone a ranking algorithm, similar to e.g. Google PageRank [43], to calculate a weight for each of the nodes (or features) in the graph [44–46]. Finally, we can sort the features on their rank and return the top k features [47]. An example is given in Figure 3.

For the headache use case, the newly discovered features (see Subsection 5.1) and their corresponding descriptions, in combination with the features obtained from the semantically annotated information produced by the mobile headache journal, can be visualized for a neurologist in a GUI. The neurologist can then define relations between these features, analogue to Figure 3. Finally, the ranking algorithm can be applied to create a list of features, ordered by their importance. This technique can easily be compared to other feature ranking techniques by taking the k top ranked features of both approaches and measuring the predictive performance of the model, trained on these features.

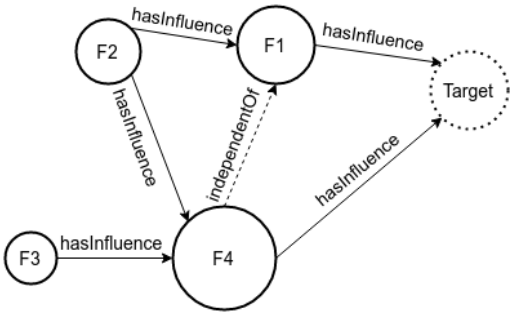

Fig. 3: Feature selection by applying a technique similar to PageRank to the knowledge graph of features.

6 Evaluation

To evaluate the impact of prior knowledge incorporation in each of the phases, a comparison will be done between the process with and without incorporation regarding the following criteria (sorted by decreasing priority):

- predictive performance of the model: by calculating the accuracy, balanced accuracy, precision, recall, AUC, F-measure, etc.
- predictive model complexity: by visual inspection and counting the maximal depth, number of nodes or leaves in the resulting decision tree
- computational time: by timing the execution of each of the phases in the machine learning process

The evaluation will be done for both incorporation in each phase separately and incorporation in all (possible subsets) of the phases. To take the no-free-lunch theorem [48] into account, the evaluation will be done on multiple benchmark datasets with varying characteristics.

7 Conclusion

In this research proposal, related work and methodologies are presented to incorporate prior background and expert knowledge, represented using Semantic Web technologies, into the different phases of a white-box machine learning approach. We are convinced that the incorporation of prior knowledge into these phases will allow for higher predictive performances and reduced training times. An evaluation regarding computational time, model complexity and predictive performance will be done by comparing the process with and without incorporation on multiple benchmark datasets and a real-world use cases.

Acknowledgements

I would like to thank my promoters prof. Filip De Turck and dr. Femke Ongenae from Ghent University for their support and valuable input in the realization of this work. This research is funded by a PhD SB fellow scholarship of FWO (1S31417N).

References

1. Tony Jan and John Debenham. Incorporating Prior Domain Knowledge Into Inductive Machine Learning. *Journal of Machine Learning*, pages 1–42, 2007.

2. Stefan Schulz et al. Snomed reaching its adolescence: Ontologists and logicians health check. *International journal of medical informatics*, 78:S86–S94, 2009.
3. Michael Compton et al. The ssn ontology of the w3c semantic sensor network incubator group. *Web Semantics: Science, Services and Agents on WWW*, 2012.
4. Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.
5. MW Kattan. Expert systems in medicine. 2001.
6. Volker Tresp et al. Towards machine learning on the semantic web. In *Uncertainty reasoning for the Semantic Web I*, pages 282–314. Springer, 2008.
7. Tjen Sien Lim et al. Comparison of prediction accuracy, complexity, and training time of thirty-three classification algorithms. *Machine Learning*, 2000.
8. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1):97–107, 2014.
9. J Caemaert and EJA Baert. In *Neurologie*, pages 191–208. Springer, 2003.
10. Lars Jacob Stovner, J-A Zwart, Knut Hagen, GM Terwindt, and J Pascual. Epidemiology of headache in europe. *European journal of neurology*, 13(4):333–345, 2006.
11. Morris Levin. The international classification of headache disorders. *Headache: The Journal of Head and Face Pain*, 53(8):1383–1395, 2013.
12. Dejing Dou, Hao Wang, and Haishan Liu. Semantic Data Mining: A Survey of Ontology-based Approaches. *2015 Ieee 9Th International Conference on Semantic Computing (Icsc)*, pages 244–251, 2015.
13. P Ristoski and H Paulheim. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on*, 2016.
14. Maximilian Nickel et al. A Review of Relational Machine Learning for Knowledge Graphs From Multi-Relational Link Prediction to Automated Knowledge Graph Construction. *Proceedings of the IEEE*, pages 1–18, 2015.
15. Heiko Paulheim, Petar Ristoski, Evgeny Mitichkin, and Christian Bizer. Data Mining with Background Knowledge from the Web. *RapidMiner World*, 2014.
16. Petar Ristoski. Towards Linked Open Data Enabled Data Mining. *The Semantic Web. Latest Advances and New Domains*, pages 772–782, 2015.
17. Rushi Longadge and Snehalata Dongre. Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*, 2013.
18. Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.
19. Nitesh V Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 853–867. Springer, 2005.
20. Vaishali Ganganwar. An overview of classification algorithms for imbalanced datasets. *IJETAE*, 2(4):42–47, 2012.
21. Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2009.
22. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
23. Haibo He et al. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *IJCNN*, pages 1322–1328. IEEE, 2008.
24. P Niyogi, F Girosi, and T Poggio. Incorporating Prior Information in Machine Learning by Creating Virtual Examples. *Proceedings of the IEEE.*, 86(11):2196–2209, 1998.
25. Ridwan Al Iqbal. A Generalized Method for Integrating Rule-based Knowledge into Inductive Methods Through Virtual Sample Creation. *arXiv:1101.4924*, 2011.
26. Jing Yang et al. A novel virtual sample generation method based on gaussian distribution. *Know.-Based Syst.*, 24(6):740–748, August 2011.

27. Liang-Sian Lin et al. Improving virtual sample generation for small sample learning with dependent attributes. *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 715–718, 2016.
28. Der-Chiang Li and I-Hsiang Wen. A genetic algorithm-based virtual sample generation technique to improve small data set learning. *Neurocomput.*, 143:222–230, November 2014.
29. Martin Ringsquandl, Steffen Lamparter, and Sebastian Brandt. Semantic-Guided Feature Selection For Industrial Automation Systems. 2015.
30. Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.
31. Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.
32. Çağlar Gülçehre and Yoshua Bengio. Knowledge matters: Importance of prior information for optimization. *Journal of Machine Learning Research*, 17(8):1–32, 2016.
33. Pedro Domingos and Geoff Hulten. Mining high-speed data streams. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 71–80. ACM, 2000.
34. Yordan Terziev. Feature Generation using Ontologies during Induction of Decision Trees on Linked Data. *ISWC PhD Symposium*, 2016.
35. Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJcAI*, volume 7, pages 1606–1611, 2007.
36. Pieter Bonte, Femke Ongenae, and Filip De Turck. Learning semantic rules for intelligent transport scheduling in hospitals. *CEUR Workshop Proceedings*, 1586:1–6, 2016.
37. Samer Hassan and Rada Mihalcea. Semantic relatedness using salient semantic analysis. In *AAAI*, 2011.
38. Iryna Gurevych. Using the structure of a conceptual network in computing semantic relatedness. In *ICON*, pages 767–778. Springer, 2005.
39. M. Lichman. UCI machine learning repository, 2013.
40. Petar Ristoski, Gerben Klaas Dirk de Vries, and Heiko Paulheim. A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In *International Semantic Web Conference*, pages 186–194. Springer, 2016.
41. M Fischera et al. The incidence and prevalence of cluster headache: a meta-analysis of population-based studies. *Cephalalgia*, 28(6):614–618, 2008.
42. Rebecca C Burch, Stephen Loder, Elizabeth Loder, and Todd A Smitherman. The prevalence and burden of migraine and severe headache in the united states: updated statistics from government health surveillance studies. *Headache: The Journal of Head and Face Pain*, 55(1):21–34, 2015.
43. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999.
44. Andreas Thalhammer and Achim Rettinger. Pagerank on wikipedia: Towards general importance scores for entities. *LNCS*, 9989 LNCS:227–240, 2016.
45. Alex D Wade et al. Wsdm cup 2016: Entity ranking challenge. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 593–594. ACM, 2016.
46. Sangkeun Lee et al. Random walk based entity ranking on graph for multidimensional recommendation. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 93–100, New York, NY, USA, 2011. ACM.
47. Dino Ienco, Rosa Meo, and Marco Botta. Using pagerank in feature selection. In *SEBD*, pages 93–100, 2008.
48. David H Wolpert and William G Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1(1):67–82, 1997.