Learning symbolic features for rule induction in computer aided diagnosis

Sebastijan Dumančić, Antoine Adam and Hendrik Blockeel

Department of Computer Science, Katholieke Universiteit Leuven, 3001 Heverlee, Belgium
{sebastijan.dumancic,antoine.adam,hendrik.blockeel}@cs.kuleuven.be

Abstract. In computer aided medical diagnosis (CAD), interpretability of learned models is an important concern. Unfortunately, the raw data used to train a model are often in sub-symbolic form (for instance, images), which makes the application of symbolic learning methods difficult. One way to alleviate this problem is to construct symbolic features that describe images, and learn to extract those features from raw images. The sub-symbolic part of the model is then limited to the lowest layer, making the model as a whole more interpretable. This paper presents a case study of how simple rule-based learners can be used to learn interpretable models from visual data by including a symbolic feature extraction step, in the domain of CAD. The symbolic representation is supported by literature and learned in the supervised way by means of deep learning. It turns out that the learned models are equally accurate as the black-box models that constitute the current state of the art.

Keywords: computer aided diagnostics, inductive logic programming, deep learning, symbolic feature learning

1 Introduction

Computational systems assisting humans in decision making have become very common lately, covering a wide large of applications. One notable example are recommender systems that allow massive online retailers to help their customers browse large amount of available items. Other examples include search engines rank the information by its relevance [1] , while computer vision techniques are used in biology for tracking cells (or other objects) and analyse them [2, 3].

One domain that can hugely benefit from computational assistance systems is medicine. The possibilities there are numerous; such systems can double-check physicians' decisions, pre-select potential infected patients from a large pool of test specimen and many more. One case that attracted a lot of attention recently is an anti-nuclear antibody (ANA) test for auto-immune diseases. The workflow of this test is fairly straightforward - starting with an image containing many cells, a physician is required to identify a staining pattern those cells exhibit. Examples of such patterns are shown in figure 1. The test is based purely on a

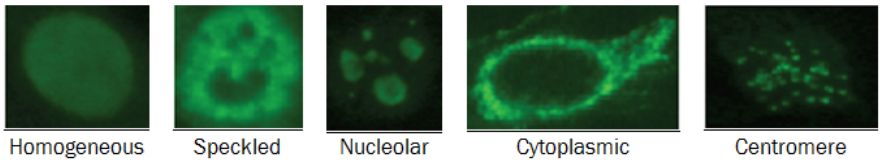

Fig. 1. Examples of HEp-2 staining patterns

visual assessment of different staining patterns. Each pattern further maps to a specific disease. This test is known to be subjective [4]; it depends heavily on the expertise of a physician, and on the varieties of reading systems and optics. The subjectiveness of the test might be significantly reduced by an intelligent system helping doctors make their decisions. In this work, we focus on this specific use case.

In the last couple of years, a number of solutions to this problem have been proposed [5, 6]. However, when dealing with image data, machine learning solutions typically provide a black-box solution. Although such solution might be very accurate, in many cases a black-box non-interpretable solution in not a desirable solution. In a critical domain such as medicine, it is very important that a physician can interpret a solution provided by a computer system. Even more, it is important that a physician can understand why a program made certain decision. Knowing precisely **why** a system made certain decision may greatly help in a situation when a physician is uncertain about his/her decision. If a system used to double-check physician's decision makes a conflicting decision, having a black-box solution can not really resolve a conflict. However, if a system could explain its decision, it would be easy to compare reasoning steps and see where they differ. Having an image data, this is rarely possible at the moment. This motivates our approach to this problem.

In this paper, we want to break open the black box. We propose to learn interpretable models from raw image data by introducing a feature construction step that extracts symbolic features using sub-symbolic learning. We achieve this by first extracting interesting features from medical text and further employing a deep learning methods to learn those features. This pre-defined set of features is learned in the supervised way - we know which features are interesting, but lack a way of specifying it formally. Having these interpretable features, we employ simple rule induction algorithms to learn rules describing the staining patterns. We focus on simple rule-based models because of their simplicity and interpretability. Additionally, we demonstrate how these simple models can be very helpful in this particular situation by introducing a collective classification settings. We elaborate this later on.

The rest of this paper is structured as follows. Section 2 discusses some background and related work. Section 3 provides more information about the data set used for this case study, and outlines our approach to this problem. It also focuses on the *feature extraction* step: it describes how models were learned that automatically extract the symbolic feature values from images, using the manually annotated images as training examples, and it evaluates the quality of these models. In section 5, several techniques are compared for learning to classify cells based on their own symbolic description, or on the description of other cells occurring in the same image. Section 6, finally, presents our conclusions.

2 Related work

This use case has been presented as a contest at the International Conference on Pattern Recognition 2012. The summary of the results is provided in [5]. For details about the approaches we refer to the paper, however, for this work it is important to state that all approaches employ high-dimensional pixel-based feature representations and complex classifiers such as Support vector machines [14]. To our knowledge, the best performance so far was reported by Xu at al [7]. The author have used a Linear Local Distance Coding method to extract the features, which were further fed into a linear Support vector machine. The approach achieved an accuracy of 95.59 %.

2.1 Deep learning and Deep belief networks

When learning a interpretable symbolic features from raw images, we focus our work on methods from deep learning[11], namely deep belief network[12]. Deep learning is a relatively new approach to machine learning which is often referred as Representation learning. It is built upon artificial neural networks and imitates the human brain in representing data. The main idea behind deep learning is to re-represent the data with many intermediate layers that represent a gradual abstraction of input data. The motivation for learning representations is quite clear the form in which data is represented is important. The success of our classifier depends on the quality of data used for training.

The deep belief network can be seen as a multi-layer generative model where each layer consists of multiple nodes, similar to a neural network. The first layer, often referred as the *visible layer*, represents the raw input data, while every higher-level layer is referred to as the *hidden layer*. It is trained in two steps - first unsupervised then supervised.

For the unsupervised phase, *Restricted Boltzmann machines* [13] are used. The Restricted Boltzmann machine is a generative energy-based model that shares the parametrization with the neural network. The Restricted Boltzmann machines are trained by maximizing the probability of data:

$$\arg\max_{W} \prod_{v \in V} P(\mathbf{v}) \tag{1}$$

where $\mathbf{v}$ represent a raw data instance, or visible layer of pixels when trained on images, while probability is represented as an energy

$$P(\mathbf{v}, \mathbf{h}) = \frac{e^{-\mathtt{E}(\mathbf{v},\mathbf{h})}}{\mathtt{Z}} \tag{2}$$

$$\mathtt{E}(\mathbf{v}, \mathbf{h}) = -\mathbf{b}^T\mathbf{v} - \mathbf{c}^T\mathbf{h} - \mathbf{h}^T\mathbf{W}\mathbf{v} \tag{3}$$

where $\mathbf{v}$ represents raw data instance, $\mathbf{h}$ response of hidden units, $\mathbf{b}_i$ and $\mathbf{c}_i$ are the offsets associated with a single element from $\mathbf{x}$ or $\mathbf{h}$ and W_{ij} are weights associated with each pair of units from different layers. $\mathtt{Z}$ is a normalization factor. After unsupervised training, the deep belief network is *fine tuned* by back-propagation [16].

3 Data and approach outline

3.1 Data

As said in the introduction, this case study focuses on a cell classification problem considered as part of a contest at the International Conference on Pattern Recognition in 2012. The original dataset considered is a set of 28 images, where each image contains a number of cells. These images were manually segmented into separate cells by human experts, leading to a second dataset containing 1456 images of individual cells. For these individual cells, it is known which original image they were extracted from; that is, we have information about which cells were originally on the same image. This is important information we plan to utilize, while none of the previous approach uses that information. Although this information is trivial to extract, none of the previous approaches uses it, as they don't have a way to integrate this information.

The images in this dataset contain also a different kind of cells, namely mitotic cells. Mitotic cells are cells that have already started dividing at the moment an image is taken. For this particular use case, they are considered very important - depending on a pattern type, mitotic cells can take different forms. This information typically helps physicians in making decisions. However, none of the previous approaches use this information. The main problem with mitotic cells is that they might not appear on every image. For this particular dataset, there are approximately 70 mitotic cells compared to 1456 regular cells, while 3 out of 28 images does not contain any mitotic cells. In that sense, the information about mitotic cells is often missing. Although not used by other methods, the dataset provides the information about the mitotic cells. We later show how methods from Inductive Logic programming (ILP)[17], a rule induction set of methods that rely on first-order logic for data representation, allow us to elegantly incorporate this information while bypassing the problem of missing data.

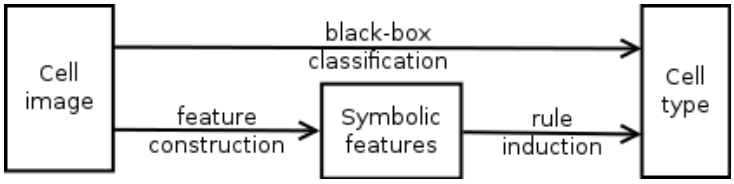

Fig. 2. System scheme

3.2 Our approach

Our approach is illustrated in figure 2. Compared to a black-box model, our approach proceeds in two steps:

1. it first assigns to each image a set of symbolic features
2. based on these symbolic features, it assigns to each image the pattern class.

The extraction of symbolic information from raw images is the key component of our system. This is done by the deep learning methods explained earlier. The set of symbolic features is pre-defined and extracted from medical texts. The goal of this case study is to investigate to what extent the relationship between raw images and their classification can be made more interpretable by building models in which the sub-symbolic component is isolated from the symbolic, interpretable component.

Extracting features As our goal is to work with features that make sense to human experts, we have searched the medical literature [8, 9] for features used by humans when classifying this type of cells. As the ANA test is based on visual interpretation, we restrict ourselves to features that describe visual properties of the cells. This led to the following list of features and their possible values:

- **shape**: circular, irregular
- **fluorescence intensity level**: positive, intermediate
- **structure**: homogeneous, speckled
- **organelle type**: dark, bright, neutral
- **organelle number**: none, few, lots
- **texture**: smooth, sparkly, blob

These six feature describe purely visual properties of a cell and can easily be labelled. We have manually annotated all cells with the value for these features. These features now serve as labels for a classifier mapping a raw image to the predefined set of symbolic features.

As the importance of mitotic cells was previously discussed, we include this information in our model. This feature takes as value the type of the mitotic cells that were present on the same image as the cell of interest. All mitotic cells

on the same image are of the same type. The mitotic cells were also manually segmented by the human experts. However, it is important to note a difference between the six visual feature described above and mitotic cell information - it cannot be derived from a cell image in isolation, information about other cells in the same original image is needed.

How these features are learned is explained in section 4. Having each image now described with these feature we can run any rule induction algorithm to learn how to detect the target patterns.

Utilizing collective classification As it was previously mentioned, the original images consist of many cells that are later manually segmented to the individual cells. The fact that all cells from the same image have to be of the same type might be further utilized to gain performance. This scenario significantly resembles collective classification [10]. In collective classification, related instances (or objects to be classified) are classified not just based on their own set of attribute values but also based on the attribute values and class labels of the related instances. In this specific use case, this means that each cell is classified not only by its attributes, but by looking at the attributes of other cells on the same image too.

One may argue that in this case, classifying each cell individually and taking the majority vote as a final class for each cell is enough. While that is true for images of high quality containing a lot of cells, it is not true for the case of low quality images containing only a couple of cells (which is more often the case). When an image is of low quality, a classifier will most likely make many mistakes. If there is a small number of cells on the same image, it might be very difficult to find a majority vote on one class confidently. On the contrary, in the collective classification settings when all cells are classified as a whole, a classifier's decision will be mostly influenced by the cells that can confidently be predicated as a certain class.

To see how exactly collective classification might help, imagine you are given an image containing a number of cells. Assume you are about to classify a cell that given its attributes cannot be confidently assigned to a particular class by a model learned from data. For the sake of illustration, assume also that a model gives as a probability distribution over classes as an output. If a model cannot make a confident prediction about a given cell, its output can be seen as an approximately uniform distribution over classes. However, with collective classification, we can put a restriction that every cell on the same image has to belong to the same class. In that case, if a system figures out that there is a certain cell on the same image that can be confidently classified as the particular class **b**, the system will use that information to increase the probability of the uncertain cell being the class **b**. We omit a lot of details here due to the space restrictions and refer a reader to [21].

4 Learning symbolic features

Having defined the features used in the intermediate layer, we need to build models that extract the values of these features from cell images. None of these models are straightforward: the definitions of the features are to some extent subjective. Therefore, the models are learned from the dataset. This can be seen as a supervised feature construction - we know which features we want (from medical literature), but are unable to specify a model for them. A different model is learned for each feature, in a supervised manner, using the manual annotations as examples.

For all features except Shape and Fluorescence Intensity Level (briefly, Intensity) a deep belief network [12] was trained, as these are known to work well for identifying visual properties of images. A separate network with Bernoulli units was trained for each feature. Shape and Intensity are learned in the following way:

Shape: Visual shape classification is a well-studied topic in computer vision and methods suitable for our goal already exist. We adopt the following method, motivated by Belongie et al [15]. Each individual cell image is divided into 4-by-4 blocks, and for each block, the proportion of pixels inside the extracted segment is calculated. A support vector machine [14] with radial basis function (RBF) kernel is next trained, using these 16 proportions as input features.

Intensity: This describes the *clarity* of the cells in an image. Determining the fluorescence intensity level is a separate task in the ANA workflow. The medical literature does not provide a precise definition for it, only a provisional ranking of four possibilities [4], described in terms of how easy is to distinguish cells from the background. All approaches mentioned in Section 2 suggest to recognize only two classes - *positive* when cells are clearly distinguishable from the background, and *intermediate* when it is difficult to distinguish cells from the background. Our method to estimate fluorescence intensity level works as follows. Our methods starts with an observation that, although cell express different intensities across image, the background is always constant and darker compared to cells. The major assumption taken here is that each image histogram (a distribution of grayscale colors or intensities across an image) can be segmented in two distinct parts - one representing the background and the second one representing the cells. Following this intuition, we approximate every image histogram with 2 Gaussian distribution. In images with positive intensity, the two components should be well separated from each other, while in images with intermediate intensity they should be relatively close. To classify cells as having positive or intermediate fluorescence intensity level, an SVM with RBF kernel is trained that uses the mean and variance of both fitted Gaussians as inputs.

5 Results

As we already said, our goal in this work is to map raw images to the set of predefined symbolic features that would allow usage of an interpretable models to learn the domain. Although any rule-based induction algorithm can be used, here we have focused on the methods from Inductive logic programming (ILP) [17] and its probabilistic extension. The main reason why we focus on these models is that they allow us (1) to easily incorporate missing information (which is necessary for the mitotic cells) and (2) make use of collective classification. We have chosen to compare `FOIL` [18] and `Aleph` [19] as ILP methods, and their probabilistic extension in Markov logic networks [20]. We leave out the details how to train such models here and point the reader to the references, but emphasize here that these models use first-order logic as a knowledge representation, which makes them interpretable.

We focus on answering the following questions:

1. how well our model with interpretable features compare to black-box models from prior work?
2. how well our model performs when information about mitotic cells is added, compared to the base in *1)*?
3. how well our model performs when collective classification is performed, compared to the base in *1)*?

Important thing to notice here is that questions **2)** and **3)** do not allow us to perform any comparison with prior work, as to the best of our knowledge none of the previously used methods uses this particular information (mitotic cells and image location information). However, our goal is to test how much this information can help in this prediction task, together with interpretable features we learn. We first test our approach using ground truth features - assigned by human, to test the usefulness of selected features. Finally, we test our approach in full settings - we first use deep belief networks to learn the features, and then use those learned features to classify cells.

5.1 Experimental settings

For our experiments, we have used the dataset from the ICPR 2012 contest [1]. The original dataset considered is a set of 28 images, where each image contains a number of cells. These images were manually segmented into separate cells by human experts, leading to a second dataset containing 1456 images of individual cells. The correct symbolic feature values described in section 4 are manually assigned to each cell. As it is mentioned before, the features are designed to represent simple visual shapes so that the expert knowledge about the domain is not necessary.

[1] http://mivia.unisa.it/datasets/biomedical-image-datasets/hep2-image-dataset/

In each of our experiments, we have used 10-fold cross validation to evaluate our approach. To fully utilize the strengths of relational learners, the folds are created on image level - folds represent non-overlapping partitions of a set of original images (containing a number of cells). This ensures that individual cells from the same image do not appear in both training and test sets. This slightly differs from cross-validation settings usually employed, but it is crucial for properly testing relational learning methods. We report the accuracy of the classifiers for each experiment. The dataset with learned features was created in the same way - we use 9 folds to learn the model parameters, as proposed in section 4, and fill in the values in the remaining fold. The predictions on the leave-out folds are then aggregated to a new dataset with features learned by the system.

5.2 Tests with the ground truth data

We first evaluate our model using only manually assigned features that describe the visual properties of each cell. We first exclude the mitotic cells from the feature set and classify cells using only their visual properties. The results are summarized in table 1. For each setting, due to the space limitations, we present only accuracy of each model. The test with the ground truth data corresponds to the first row of the table.

The results show that the state-of-the-art solution proposed by Xu et al. [7] performs significantly better than logic-based approaches chosen for our work. The accuracy of the state-of-the-art solution is 95.59 %. This is somehow an expected result as we have to sacrifice expressiveness to gain interpretability. As we try to use features understandable by humans, the performance is bounded by their expressibility. It is worth noting that `FOIL` performs as well as the human expert on the same dataset [5]. Sophisticated image analysis methods might find enough information to separate difficult cases even without mitotic cells, but it would be extremely difficult to tailor understandable features to express those differences.

This experiment answers the first question. Although our model performs worse than the state-of-the-art solution, we believe it makes a step forward in making these models interpretable to human experts.

To answer the second question, we include the information about the mitotic cells in the dataset used to train the model. The results are shown in the second row of the table 1. It is immediately clear from the results that mitotic cells play an important role in the diagnostic procedure, as the difference in the accuracies is substantial. In this case, the results are comparable to the state-of-the-art. These results demonstrate that the interpretable features defined, together with the mitotic cells, are sufficient for the task, and in they sacrifice the performance only slightly. Note again that this is an unfair comparison with prior work as

Table 1. Performance of the classifiers in different settings

Settings	Xu at al	`MLN`	`FOIL`	`Aleph`
visual features	95.59	74.05	81.45	40.41
mitotic cells included	–	93.05	93.30	84.06
complete information	–	94.88	98.00	89.35
learned features	–	89	97.32	89.28

mitotic cells were not used there, but our aim is to show how this, obviously important, information can be easily integrated in a model using the simple ILP techniques.

Finally, we have tested our model collective classification settings. To achieve collective classification, we had to add a logical predicate `SameImage(x,y)` that evaluates to true when cells `x` and `y` are located on the same image (we leave out the details how collective classification is performed in the system). The mitotic cells were also included in this experiment. The results are presented in the third row of the table 1. Not surprisingly, collective classification clearly helps and increases the performance of the system. By combining both the information about mitotic cells and collective classification, `FOIL` even outperforms the state-of-the-art approach. This is a very pleasing result for the following two reasons:

1. it outperforms the state-of-the-art approach while maintaining an interpretable representation that sacrifices a lot of expressivity
2. it mimics the setup of the test in practice, and at the same time makes use of relational information other systems cannot easily incorporate.

5.3 Tests with the features learned by the system

The previous section aimed at demonstrating the capability of the predefined set of interpretable features for this task. In this section, we evaluate our system in full. We first train the deep belief network to assign symbolic features to a given cell, as described in section 4. Then, we use those features to predict the class of each cell. As some symbolic features will be mislabelled, this evaluates the robustness of our approach given imprecise data. Mitotic cells are included for this experiments, as well as the collective classification setting as they lead to the most successful results.

Table 1, final row, lists classification performance when learning from the dataset when the features are learned. Compared to the dataset containing the true features, the performance drops slightly, but not dramatically. This shows that even with the noisy information that is inherent to automatic feature extraction, quite accurate classification can be obtained. More importantly, using the collective classification seems to provide more stable results as the misclassified features affect the classification accuracy only slightly.

6 Conclusion

Learning symbolic interpretable representations from images is a very difficult task, but necessary in many domains. An example of such a domain are medical diagnostic procedures based on visual interpretation of images. In this paper we presented a case study of detecting antibodies patterns from images demonstrating the benefit of using ILP methods for the task. The outcomes of the paper are three-fold. First, we have proposed a method that constructs interpretable features for this application domain. Construction of such interpretable models from sub-symbolic data is a non-trivial task. In our approach we first identify and define symbolic features that are interpretable to humans and demonstrate how these features can be learned automatically by means of deep belief networks. Second, we have demonstrated a benefit of using the information about mitotic cells for the task. Related approaches ignore this information at the moment, mainly because mitotic cells do not appear on every image and raise the question of how to represent missing information. However, ILP methods provide us with an elegant way to include this information. Finally, we have demonstrated the benefits of using collective classification for the task.

Experiments show that, on the domain considered, this interpretable model can achieve accuracy comparable to black-box models, and even outperform them. This is a positive result as the final goal of the work is to build an interpretable model, that sacrifices the performance as little as possible. Deep learning show as a promising approach for this direction.

Within this particular application, other possible future work includes automatic mitotic cell detection and artefact removal, so as broader experimentation with deep learning approaches and automatic segmentation of individual cells.

Acknowledgements

This work is funded by the KU Leuven Research Fund (project IDO/10/012).

References

1. Radlinski, F. and Joachims, T. : Query Chains: Learning to Rank from Implicit Feedback International Conference On Knowledge Discovery and Data Mining, ACM SIGKDD, 239–248 (2005)
2. Harder,N. et al. : Automatic analysis of dividing cells in live cell movies to detect mitotic delays and correlate phenotypes in time Genome Research, vol. 19(11), 21132124 (2009)
3. Godinez, W. et al : Deterministic and probabilistic approaches for tracking virus particles in time-lapse fluorescence microscopy image sequences Medical Image Analysis, vol. 13, 325-342 (2009)

4. Rigon, A. and Soda, P. and Zennaro, D. and Iannello, G. and Afeltra, A. : Indirect immunofluorescence in autoimmune diseases: Assessment of digital images for diagnostic purpose Cytometry Part B-clinical Cytometry, vol. 72B, 472-477 (2007)
5. Foggia, P. et al.: Benchmarking HEp-2 Cells Classification Methods. IEEE Trans. Med. Imaging, vol. 32, 1878-1889 (2013)
6. Agrawal,P., Vatsa, M., Singh, R.: HEp-2 Cell Image Classification: A Comparative Analysis. Lecture Notes in Computer Science, Machine Learning in Medical Imaging, vol. 8184, Springer International Publishing, 195-202, (2013)
7. Xu, X. et al: Linear Local Distance coding for classification of HEp-2 staining patterns Winter Conference of Application of Computer Vision, 393-400 (2014)
8. Wiik, A.S. and Hier-Madsen, M. and Forslid, J. and Charles, P. and Meyrowitsch, J.: Antinuclear antibodies: A contemporary nomenclature using HEp-2 cells Journal of Autoimmunity , vol. 35 (3), 276-290 (2010)
9. Bolon, P.: Cellular and Molecular Mechanisms of Autoimmune Disease Toxicologic Pathology, vol. 40 (2), 216-229 (2012)
10. Sen, P. and Namata, G. and Bilgic,M. and Getoor, L. and Gallagher, B. and Eliassi-Ra, T. : Collective Classification in Network Data AI Magazine, vol. 93, 93 – 106 (2008)
11. Bengio, Y.: Learning Deep Architectures for AI. Found. Trends Mach. Learn., vol. 2, Now Publishers Inc., 1–127 (2009)
12. Hinton, G, E., Osindero, S. and Teh, Y.: A Fast Learning Algorithm for Deep Belief Nets. Neural Comput., vol. 18, MIT Press, pp. 1527–1554 (2006)
13. Larochelle, H., Bengio,Y.: Classification using discriminative restricted boltzmann machines. 25th international conference on Machine learning, ACM (2008)
14. Vapnik, V., Cortes, C.: Support-Vector Networks. Machine learning, vol. 20, Kluwer Academic Publishers, 273–297 (1995)
15. Belongie, S., Malik, J. and Puzicha, J. Matching Shapes. International Conference on Computer Vision, vol. 1, 454–461 (2001)
16. Murphy, K. : Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series). MIT Press (2012)
17. Lavrač, N. and Džeroski,S. : Inductive Logic Programming: Techniques and Applications. Routledge (1993)
18. Quinlan, J. R. earning logical definitions from relations. Machine Learning, vol. 5, 239–266 (1990)
19. Muggleton, S. Inverse Entailment and Progol. New Generation Computing, vol 13, 245–286 (1995)
20. Richardson, M. and Domingos, P. : Markov Logic Networks. Machine learning, volume 62 (1-2), 107-136 (2006)
21. Crane,R. and Mcdowell, L.: Investigating markov logic networks for collective classification. In ICAART (2012)