July 2019

Scientific and Pedagogical Plan

Cláudia Antunes
INSTITUTO SUPERIOR TÉCNICO - UNIVERSIDADE DE LISBOA

In the last years, the explosion of deep learning have diverted the attention of the research community to neural networks, making more difficult to address old problems in knowledge discovery from data, such as the exploitation of explicit domain knowledge or the use of known temporal dependencies to inform the discovery process. Nevertheless, these have been recognized as core transversal issues for research all along the years (Yang & Wu, 2006).

Those problems have been addressed in a scattered way, with a few approaches deeply depending on user intervention. The exception to this panorama is the use of constraints in pattern mining (Silva & Antunes, 2016), which have been proven to adequately focus the discovery process in accordance to user expectations. However, the specification of those constraints usually require the use of byzantine specification languages, making impossible its wide use. Moreover, in the area of classification, sporadic approaches were proposed (see for example the work of Pedro Domingos over statistical relational learning (Richardson & Domingos, 2006)), and when it happened they lacked scalability at a huge level.

In the last years, deep learning brought the ability to explore hidden patterns and collect them through unsupervised ways, then making use of them for classification purposes. However, we continue to not be able to incorporate domain knowledge in the learning process in a direct way, and the lack of interpretability of the resulting models rise old questions.

Cláudia's current research plan aims for answering the following questions:

1. Is it possible to incorporate domain knowledge in the classification process, directly from existing domain knowledge?
2. Is it possible to discover patterns through unsupervised techniques and use them as features in any classification process, without using deep learning?
3. Is it possible to automate the discovery process, in particular the data preparation step with the automatic creation of the right features (*feature engineering*)?

The central approach to answer those questions will be the exploration of data models, along with available metadata, in particular relational, star schemas and constellations. Definitely, those models, plenty available behind databases and data warehouses, already represent significant domain knowledge, that may be useful for guiding the discovery process.

Along with the definition of a framework that maps the features in the datasets to the elements in the data model, Cláudia expects to develop a set of domain driven classification algorithms and methodologies, validating them in the context of a couple of case studies from different domains. These case studies would cover problems on the fields of healthcare, education and environment.

Following a brief summary of the work about using domain knowledge in classification, the main goals to achieve are identified, and the approach to follow is described. A description of each task to develop is presented, finishing with a set of expected results.

1 Literature Review

The explosion of interest on data science, in general, and the usage of deep learning, in particular, made clear the need of exploring new paths to learn self-explanatory models. In the last two decades, the field of machine learning and data analysis proved to be effective when applied to almost all domains and a large range of kinds of data, however, the lack of ability to make use of previously existing domain knowledge and the lack of interpretability and simplicity of the models discovered, keep hounding such advances.

Historically there have been two major approaches to research in machine learning: one based on logic representations, such as ILP (Blockeel 1998), and the other focused on numerical ones, like Bayesian methods and neural networks. While the first ones tend to focus on handling the complexity of the real world, the counterpart tackles the uncertainty existing in the generality of fields of application (Domingos

1997), producing yield good levels of accuracy for unseen data, when the training set was properly balanced and sized

Along years these approaches have evolved separately, but the recognition that the inability to exploit domain knowledge in the process is a plausible cause for the lack of simple and interpretable models, has contributed to change the research efforts along the years.

One of the first proposals for this, in the last fifteen years, was the one designated as D3M (Domain Driven Data Mining) (Cao2006), which for some time, tried to create methods for extracting actionable knowledge, and proposing a paradigm shift from data-centered to domain-driven actionable knowledge discovery. However, this research was centered on methods dedicated to specific domains, with a special emphasis on the actionability of the results, and their specificity difficult their generalized application.

Along with the efforts in D3M, the use of existing domain knowledge to enrich the learning process has also been explored under the umbrella of Semantic Aspects of Data Mining, by trying to add semantics to data under analysis, in particular through the use of sound knowledge representation formalisms, to guide the mining algorithms.

Among these attempts, the work of Pedro Domingos stood out, presenting an interesting approach to inform statistical approaches through the use of existing knowledge, represented as formulas in logic. A Markov logic network (MLN) (Richardson 2006) is one of the attempts to combine first order logic and probabilistic models in a single representation. It consists of a first-order knowledge base with a weight attached to each clause, trying to bring together the ability of probabilistic models to efficiently handle uncertainty and the expressive power of first order logic. However, the size of the network depends exponentially on the number of constants in the domain of discourse and the inference in this context remains to be improved, tending to be too slow in real problems.

With the same goal, domain ontologies were also proposed to represent existing knowledge, and have been shown to be efficiently used in the context of pattern mining (Antunes 2009). In particular, the D2PM framework (Antunes 2012) provides a new formulation of the pattern mining process in the presence of domain knowledge, mapping the elements from the mining problem to the elements in the domain ontology. Along with this formulation, the framework affords a set of constraints and efficient constrained algorithms to perform the task.

In the context of this framework, attempts to apply the same principles to classification were pursued, by adapting the training of decision trees and naïve Bayes algorithms to use taxonomies to inform the mining process – HDTL and HNBL algorithms (Vieira 2014), or other axioms as in ODTL (Vieira MSc thesis). The idea behind those algorithms was to test the discriminative power of the attributes that describe the dataset as usual, but also testing their aggregations at any abstraction level, in accordance to given taxonomies for each attribute, represented through is-a axioms in a domain ontology. By computing the frequencies as in the creation of multidimensional data hypercubes, algorithms' efficiency levels are kept, but their accuracy improves, mostly in the presence of partially observed data.

Despite the incorporation of explicit domain knowledge in the discovery process, still remains unsolved, the accuracy of the generality of classification approaches have been reaching astonishing levels. These results however, are obtained after a thorough preparation of the data, by selecting and creating new features, wasting more then 80% of the total time of the discovery process.

Indeed, the selection and creation of features are the most common use of domain knowledge in the discovery process. However, this kind of work has depended exclusively on the ability of the data scientist to understand the domain and be able to translate that knowledge into new features or removing useless ones.

In order to address this issue, *autoML* has emerged in the last few years as a very promising field. With the aim of reducing and/or avoiding the requirement of human intervention in the discovery process, the new area appears as a new pathway to enhance our results. Workshops for addressing *autoML* are present in the major conferences in the area, such as KDD and ICML[1].

While the majority of the attempts are centered on choosing the best parameters for the learning algorithms (like (Wang, et al., 2018), (Davis & Giraud-Carrier, 2018), (Weisz, et al., 2018) and (Thomas, et al., 2018)), others try to choose the best architecture for neural networks ((Kamath, et al., 2018) and (Zela, et al., 2018)).

[1] https://sites.google.com/view/automl2019-workshop
https://sites.google.com/site/automl2018icml/home

Another field that has been deserving some attention is the automatic creation of new features from domain knowledge, called *feature engineering* (Nargesian, et al., 2017). In particular, natural language processing and text mining have been the biggest beneficiaries of this approach. See for example the work by (Brazdil, 2018).

Nevertheless, the promising results, there is plenty of work to do.

2 Knowledge Driven Classification

Given the need to overcome those issues, Cláudia proposes to develop a framework for formalizing the classification problem in the presence of domain knowledge and known temporal dependencies, and to propose a set of new algorithms and methodologies to train classifiers in that environment.

Framework Definition

The framework to define will be inspired on the D2PM framework previously proposed by Cláudia (Antunes 2012), and similarly it will map the features in the datasets to the elements in the data model. However, the main idea here is to explore the data models behind the datasets to mine, along with available metadata, in particular relational, star schemas and constellations. Definitely, those models, plenty available behind databases and data warehouses, already represent significant domain knowledge, that may be useful for guiding the discovery process.

Cláudia's argument is that the use of this knowledge, in particular the taxonomies described as hierarchies in the different dimensions, may contribute significantly to enrich the mining process, producing simpler models, less prone to overfitting, more generalizable and easier to apply to partially observed data.

The goal for the first task to accomplish is to define the new framework (the D2Class framework), covering tabular data, potentially with temporal dependencies. In particular, the framework shall be able to deal with data at different granularities, either for time or any other dimension, corresponding to the different levels of the hierarchies present in each dimension.

An important issue is the ability to represent and connect the different occurrences of the same variable (attribute) in each record. For example, a record representing the history of medical visits for a patient, contains a set of snapshots described by a fixed set of variables. The framework has to be able to identify the set of variables, and to manipulate the different set of features for each variable.

In this manner, the framework will enclose the basic definitions of all the terms considered in the classification process of tabular data, including the ones needed to include temporal information. The goal is to provide a mean where all the concepts in the data to mine, and the relations among those concepts are precisely defined in the context of a classification task, and their usage clearly established.

The first attempt is to define the new framework based on the D2PM framework (Antunes 2012) proposed before, where domain knowledge is represented through a domain ontology. In this new context, a dataset could be seen as just a set of records described by a set of attributes, with each record corresponding to a set of instances. Note that an instance is an element in the A-box (the set of known individuals in a domain ontology), which instantiate a given concept (an element in the T-box representing the definition for a given kind of individuals). In order to be able to consider the domain knowledge, all the attributes describing the dataset should correspond to concepts in the ontology, since this enables the existence of a taxonomy among those attributes. In this definition of a dataset, nothing about the structure of the records is required, both comprising tabular records and sequences. Moreover, a sequence may just be seen as a record with multiple instances that instantiate repeatedly the same set of concepts, usually ordered along time, which encompasses the snapshots referred above.

In terms of temporal dependencies exploitation, the first goal in the plan is to explicitly tackle the temporal information of each record, first by considering time as an additional attribute (the timestamp) and second by making use of it to aggregate all the remaining data elements to the best temporal granularity. Actually, the idea is to mimic the data aggregation performed in OLAP analysis, where data may be considered at any level of abstraction, following the different aggregations resulting in the different data hypercubes. Furthermore, the idea is to automatically choose the best temporal aggregation and abstraction level for each attribute in terms of its class discriminative power.

Domain Driven Classifiers

The goal of such a framework is to provide a generic context for exploring domain knowledge and temporal dependencies in classification, through any classification technique, but also support other kinds of domain knowledge exploration.

The first approach to follow is to adapt training algorithms, such as C4.5 or naïve Bayes, to work in the new environment. The new algorithms will be based on the ones already proposed in the D2PM project (PTDC/EIA-EIA/110074/2009), the HDTL and HBNL (Vieira 2014). After this, the adaptation of other Bayesian learners and classifiers, such as Bayesian networks, will be explored, to be guided by taxonomies. The work previously done on hierarchy based naïve Bayes learner (HNBL) and classifier (HNBC) will be extended in order to make use of the network structure, considering it designed at the different abstraction levels. Consider for example a network with two input nodes (say A and B), an unobservable variable C, conditionally dependent on both A and B, and a class node X, conditionally dependent on A and C. Instead of just using this network, other networks should be contemplated, one for each combination of the different abstraction levels of each attribute / node. For example, a network would cover Pa(A), Pa(B), C and X, with Pa(x) the next abstraction level for attribute x. The idea is, as in HNBL, to explore the different combinations of granularities as in the data hypercubes, usual in OLAP data exploration. The great challenge is to avoid the train of all the possible combinations and choose the best one achieved, but be able to identify the most promising ones and just consider them.

The second pathway to explore is to address the incorporation of previously discovered patterns as new features in the classification task. These patterns will be discovered by pattern mining methods, and selected in order to increase the accuracy of the trained classifiers, as thoroughly described in (Antunes, 2018), but the method has to be redefined in the light of the new framework.

The third way to explore is to study new methods for feature engineering, either from the exploration of the knowledge represented through the framework, or the patterns discovered as described above. In particular, Cláudia expects to explore the known relations among the different features for the same variable to create important insights to the training process.

Along with these new approaches, the extension of some classification methodologies should be proposed to work in the framework context. In particular, the methodologies that specifically deal with temporal dependencies previously proposed, such as asap classifiers (Antunes 2010) and the methodologies for prognosis (Cardoso 2014). The challenge comprises both choosing the best level of abstraction for each attribute, but also to consider each one at the best time aggregations.

In order to accomplish such goal, the first approach will comprehend the exploration of time stamped data, which encompasses two challenges: to deal with the temporal attribute *per se* and to deal with the remaining attributes at the best temporal aggregation. The strategy will be to approach time as any other attribute, with well-known and studied granularities, represented through taxonomies as other attributes. In this manner, the challenge will be focused on transforming the remaining (time stamped) data to the most adequate aggregation, not in terms of its abstraction but their aggregation along time. A simple approach will be choosing the best time granularity and aggregate the remaining attributes to that granularity.

Validation

This validation of the framework and methods proposed will be accomplished through running the methods and algorithms in a set of publicly available datasets, usually used for algorithms evaluation. This will be useful for comparing our methods with others comparable.

Against this kind of evaluation, the methods will face a second assessment, through two case studies in different domains. Their choice depends on the method to validate, and the available domain knowledge. Nevertheless, and since the methods will mainly cover taxonomies, their collection won't be a problem, since this kind of knowledge is spread available.

The domains of election for the case studies are healthcare, education and environment, since they are the domains that currently deserve more attention, being the ones where it is easier to collect domain knowledge, either through the collaboration with experts from the field or through bibliographic research.

As usual in data mining and classification, the methods will be evaluated using quantity measures, like accuracy, precision, recall and others. However, and since one of the goals is to achieve models less prone to overfitting, the methods will also be assessed in terms of size / complexity of the models learnt.

Expected Results

Through following this plan, we expect to be able to create a generic framework for classification under domain knowledge and temporal data, with a wide applicability due to the independency of the domain area.

Through the proposed approach, the methods will be entirely independent of the domain at hands, and solely on the elements of the domain, warranting the independence of the approach.

Still, by climbing along the different taxonomies, the models learnt would potentially be defined at higher levels of abstraction, which *per se* reduces the risk of overfitting. Previous results have shown that the learnt models were simpler, by exploiting higher levels of abstraction without loosing accuracy (Vieira 2014).

3 Pedagogical Benefits and Plan

The explosion of the interest on data science presents several challenges to Técnico and to the Department of Computer Science and Engineering. The continuous demand of know-how in this area and the number of students aiming for study these topics, bring a set of opportunities along with a series of additional difficulties.

The research plan proposed above is one of the ways to answer that demand, through several dimensions.

First, it presents a wide variety of topics, framed by a well-defined problem. Each of these topics may be explored and first approached through students pursuing their master graduation. Indeed, in the next year, Cláudia is going to supervise three master students: Hélio Domingos is going to approach feature engineering techniques, Carlos Branco is going to explore the temporality and granularities for classification, and Miguel Simões is going to profile students driven by domain knowledge.

Secondly, this research plan encompasses two FCT projects (PTDC/CCI-CIF/28939/2017 and PTDC/CCI-CIF/30754/2017), where Cláudia leads Técnico's teams, which offers a funded environment to fulfill the defined goals.

Along with the research in this area, the proposed plan provides experience on one of the bigger current challenges in knowledge discovery – the automation of the process. Beside the advantages of pursuing such dare task, the plan encompasses dealing with the data preparation step, which just a few Professors in Técnico usual deal with.

This and the experience on working with real data, gives to Cláudia enough experience to be able to address classes in this area with much more confidence and quality. In particular, courses like Data Science or Business Process Management (component of Process Mining) in the Masters Degree of Computer Science and Engineering, will be the first beneficiaries of such experience.

Beside these courses, the new programs for advanced qualification being proposed by Técnico +, and for which Cláudia is one of the coordinators, also benefit from Cláudia's experience

References

Ansdell, M et. all., 2018. Scientific Domain Knowledge Improves Exoplanet Transit Classification with Deep Learning.

Altınel, B. & Ganiz, M. C., 2018. Semantic text classification: A survey of past and recent advances. Information Processing & Management, pp. 1129-1153.

Antunes, C., 2009. Mining Patterns in the Presence of Domain Knowledge. s.l., INSTICC, pp. 188-193.

Antunes, C., 2010. Anticipating student's failure as soon as possible. In: C. Romero, S. Ventura, M. Pechenizkiy & R. Baker, eds. Handbook for Educational Data Mining. New York: CRC Press, pp. 353-363.

Antunes, C., 2018. Is Pattern Mining Dead? - The Role of Pattern Mining in the New Era.

Antunes, C. & Bebiano, T., 2012. Mining Patterns with Domain Knowledge: a case study on multi-language data. In. Shanghai, China, IEEE Press, pp. 167-172.

Brazdil, M. J. F. a. P., 2018. Workflow Recommendation for Text Classification with Active Testing Method. Stockholm, Sweden, s.n.

Brust, C.-A., Denzler, J. & Stifel, M., 2018. Integrating domain knowledge: using hierarchies to improve deep classifiers.

Cao, L. & Zhang, C., 2006. Domain-driven data mining: A practical methodology. Int. Journal Data Warehousing and Mining, 2(4), p. 49–65.

Cardoso, D. & Antunes, C., 2014. Computer-Aided Prognosis based on Temporal Dependencies. New York, USA, s.n., pp. 549-550.

Davis, C. & Giraud-Carrier, C., 2018. Annotative Experts for Hyperparameter Selection. Stockholm, Sweden, s.n.

Domingos, P. & Pazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, Volume 29, p. 103–137.

Kamath, P., Singh, A. & Dutta, D., 2018. AMLA: an AutoML frAmework for Neural Network Design. Stockholm, Sweden, s.n.

Nargesian, F. et al., 2017. Learning Feature Engineering for Classification. s.l., s.n., pp. 2529-2535.

Richardson, M. & Domingos, P., 2006. Markov Logic Networks. Machine Learning, 62(1-2), p. 107–136.

Silva, A. & Antunes, C., 2016. Constrained Pattern Mining in the New Era. International Journal on Knowledge and Information Systems.

Song, Y. & Roth, D., 2017. Machine Learning with World Knowledge: The Position and Survey.

Thomas, J., Coors, S. & Bischl, B., 2018. Automatic Gradient Boosting. Stockholm, Sweden, s.n.

Vieira, J. & Antunes, C., 2014. Decision tree learner in the presence of domain knowledge. Wuhan, China, s.n.

Wang, Z. et al., 2018. Automatic Hyperparameter Tuning of Machine Learning Models under Time Constraints. Seattle, WA, USA, s.n.

Weisz, G., Gyorgy, A. & Szepesvari, C., 2018. CapsAndRuns: An Improved Method for Approximately Optimal Algorithm Configuration. Stockholm, Sweden, s.n.

Yang, Q. & Wu, X., 2006. 10 Challenging Problems in Data Mining Research. Int'l Journal of Information Technology & Decision Making, 5(4), p. 597–604.

Zela, A., Klein, A., Falkner, S. & Hutter, F., 2018. Towards Automated Deep Learning: Efficient Joint Neural Architecture and Hyperparameter Search. Stockholm, Sweden