Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)

Statistical Relational Learning Towards Modelling Social Media Users

Golnoosh Farnadi
Dept. of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
Dept. of Computer Science, Katholieke Universiteit Leuven, Belgium
Golnoosh.Farnadi@ugent.be

1 Motivation

Nowadays web users actively generate content on different social media platforms. The large number of users requiring personalized services creates a unique opportunity for researchers to explore user modelling. To distinguish users, recognizing their attributes such as personality, age and gender is essential. To this end, substantial research has been done by utilizing user generated content to recognize user attributes by applying different classification or regression techniques. Among other things, we have inferred the personality traits of Facebook users based on their status updates using three different classifiers [Farnadi *et al.*, 2013]. But as we concluded in [Farnadi *et al.*, 2014b], user attributes are not isolated, as emotions expressed in Facebook status updates are related to age, gender and personality of the authors. Using multivariate regression or multi-target classification techniques is one approach to leverage these dependencies in the user attribute learning process. For example, to employ the dependencies between different personality traits, we applied five multivariate regression techniques to infer the personality of YouTube vloggers [Farnadi *et al.*, 2014c].

The above mentioned techniques are powerful types of machine learning approaches, however they only partially model social media users. Users communicate directly as friends, or indirectly, by liking others content. These types of interaction between users are a valuable source but modelling them with the traditional machine learning approaches is challenging. Furthermore, user generated content in social media comes in different modalities such as visual and textual content, whereas different pieces of evidence might contradict each other. Moreover, in extracting features from each source, a reasonable amount of noise is expected. Hence, using the extracted features as a single feature space without considering features' noise, dependencies or conflicts, reduces the quality of the learning process. To overcome these limitations, we introduce a new statistical relational learning (SRL) framework [Getoor and Taskar, 2007] suitable for modelling social media users, which we call PSL^Q [Farnadi *et al.*, 2014a].

2 SRL with Soft quantifiers

PSL^Q is the first SRL framework that supports reasoning with soft quantifiers, such as "most" and "a few". We start with probabilistic soft logic (PSL), an available SRL framework which defines templates for hinge-loss Markov random fields [Bach *et al.*, 2013] and extend it to a new framework with soft quantifiers. Unlike other SRL frameworks whose atoms are Boolean, atoms in PSL can take continuous values in the interval $[0, 1]$, which facilitates analysis of continuous domains such as user behavior in social media. Indeed, in practice user behavior is not always black-and-white. For example, under interpretation I, $I(Friend(Bob, Alice)) = 1$ and $I(Friend(Bob, Chris)) = 0.2$, denote that Alice is a close friend of Bob, while Chris is a distant friend. In models for social media it is common to assume that friends are influenced by each other's behavior, beliefs, and preferences. PSL, similar to other SRL frameworks, uses the existential ($\exists$) and universal ($\forall$) quantifiers from first-order logic to express this dependency. An often cited example in SRL contexts describing smoking behavior among friends is $\forall X \forall Y Friend(X, Y) \rightarrow (Smokes(X) \leftrightarrow Smokes(Y))$ [Richardson and Domingos, 2006]. This formula states that if two people are friends, then either both of them smoke or neither of them. In this case, the probability that a person smokes scales smoothly with the number of friends that smoke. However, many traits of interest might not behave this way, but instead, having a trait only becomes more probable once *most* or *some* of one's friends have that trait as with smoking. Expressing this dependency requires a soft quantifier, which none of the available SRL frameworks allow.

Syntactically, a quantifier expression in PSL^Q is of the form: $Q(V, F_1[V], F_2[V])$, where Q is a soft quantifier, and $F_1[V]$ and $F_2[V]$ are formulas containing a variable V. A formula can be an atom as well as a negation, a conjunction or a disjunction of formulas. Formulas are interpreted in Łukasiewicz logic, i.e., for x and y in $[0, 1]$ (the $\tilde{}$ indicates the relaxation over Boolean values): $x \tilde{\wedge} y = \max(0, x + y - 1)$,

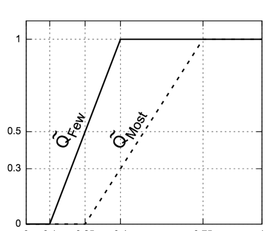

Figure 1: Example of "most" and "a few" mappings

$x \tilde{\vee} y = \min(x + y, 1)$ and $\tilde{\neg} x = 1 - x$. Similar to friendship, smoking behavior is represented by varying degrees: for example, Chris might be a heavy smoker, while Alice might be only a light smoker. All these degrees can and should be taken into account when computing the truth degree of statements such as "a few friends of Bob smoke" and "most friends of Bob smoke". To this end, we define the semantics of a quantifier expression based on the approach of [Zadeh, 1983]. Thus, we define the truth value of $Q(V, F_1[V], F_2[V])$ as:

$$\tilde{Q}\left(\frac{\sum_{x \in D_V} I(F_1(x)) \tilde{\wedge} I(F_2(x))}{\sum_{x \in D_V} I(F_1(x))}\right) \quad (1)$$

where $\tilde{Q}$ is a $[0, 1] \rightarrow [0, 1]$ mapping representing the meaning of Q and D_V is the domain for variable V. Figure 1 depicts possible quantifier mappings for the soft quantifiers "a few" and "most". Using these mappings, the statement "a few friends of Bob smoke" is true to degree 1 as soon as 40% of Bob's friends are smokers, while 75% of Bob's friends are required to be smokers for the statement "most friends of Bob smoke" to be fully true.

A PSL^Q model consists of a collection of PSL^Q rules. A PSL^Q rule r is an expression of the form: $\lambda_r : \underbrace{B_1 \wedge B_2 \wedge \ldots \wedge B_k}_{r_{body}} \rightarrow \underbrace{H_1 \vee H_2 \vee \ldots \vee H_l}_{r_{head}}$, where $B_1, B_2, \ldots, B_k, H_1, H_2, \ldots, H_l$ are atoms, negated atoms, quantifier expressions or negated quantifier expressions and $\lambda_r \in \mathbb{R}^+ \cup \{\infty\}$ is the weight of the rule r. The distance to satisfaction of a rule r under an interpretation I is defined as: $d_r(I) = \max\{0, I(r_{body}) - I(r_{head})\}$. Note that a PSL^Q model without any quantifier expressions is a PSL model.

Inference in PSL^Q is based on the most probable explanation (MPE), where the goal of optimization is to minimize the weighted sum of the distances to satisfaction of all rules. MPE inference for a PSL program is casted as a convex optimization problem. However, the main challenge in introducing soft quantifiers in the language of PSL is their non-linear formulation (Equation 1). We propose a new iterative MPE inference algorithm, which solves a PSL^Q program through solving a sequence of convex optimization problems. The optimization problem at each iteration is determined by the output of the previous iteration using a PSL MPE-solver.

Structural balance theory [Heider, 1958] implies that users are more prone to trust their neighbors in the network rather than unknown other users. Using this theory, Bach et. al modelled social trust with PSL rules such as $Trusts(A, B) \wedge Trusts(B, C) \rightarrow Trusts(A, C)$, which indicates that A trusts C to a degree that B (as a trustee of A) trusts C. To investigate whether we can improve the accuracy of the predictions by introducing rules with soft quantifier expressions, we constructed PSL^Q rules based on a triad relation over a *set* of users instead of *individual* third parties. Using the soft quantifier "a few", we extended the above rule as: $Few(X, Trusts(A, X), Trusts(X, C)) \rightarrow Trusts(A, C)$, which indicates that A trusts C to a degree that a few trustees of A trust C. Experimental results show that using soft quantifiers not only expands the expressivity of the model, but also increases the accuracy of the inferred results (Table 1).

Model	PR+	PR-	AUC
PSL [Bach *et al.*, 2013]	0.9770	0.4457	0.8118
PSL^Q	**0.9789**	**0.4670**	**0.8247**

Table 1: Results using 8-fold cross-validation using the Epinions sample with 7,974 trust vs. 701 distrust relations. Values in bold are statistically significant with a rejection threshold of 0.05 using a paired t-test w.r.t. the PSL model.

3 Future directions

The work that we presented here can be extended in the following three directions: (1) Besides social trust, many other AI applications could benefit from the use of soft quantifiers. Modelling users in social media and inferring their attributes with a PSL^Q model is a promising direction for our future work. (2) We defined the semantics of a quantifier expression using the approach of Zadeh. Studying other approaches for quantifiers and their complexity of integrating them into SRL frameworks is a direction for our future work. (3) Designing a suitable PSL^Q model is often time consuming, thus we would like to extend the capability of PSL^Q to learn the structure of data to automatically generate rules when no or little background knowledge is available. This would also include an automatic way of learning the best quantifier mapping for each quantifier expression in a PSL^Q model.

References

[Bach *et al.*, 2013] Stephen Bach, Bert Huang, Ben London, and Lise Getoor. Hinge-loss Markov random fields: Convex inference for structured prediction. In *Uncertainty in Artificial Intelligence (UAI)*, 2013.

[Farnadi *et al.*, 2013] Golnoosh Farnadi, Susana Zoghbi, Marie-Francine Moens, and Martine De Cock. Recognising Personality Traits Using Facebook Status Updates. In *Proc. of WCPR13 workshop at ICWSM13*, 2013.

[Farnadi *et al.*, 2014a] Golnoosh Farnadi, Stephen Bach, Marie-Francine Moens, Lise Getoor, and Martine De Cock. Extending PSL with Fuzzy Quantifiers. In *Proc. of StarAI2014, Workshop at Statistical Relational AI at AAAI2014*, 2014.

[Farnadi *et al.*, 2014b] Golnoosh Farnadi, Geetha Sitaraman, Mehrdad Rohani, Michal Kosinski, David Stillwell, Marie-Francine Moens, Sergio Davalos, and Martine De Cock. How are you doing? Emotions and Personality in Facebook. In *Proc. of EMPIRE2014, workshop at UMAP2014*, 2014.

[Farnadi *et al.*, 2014c] Golnoosh Farnadi, Shanu Sushmita, Geetha Sitaraman, Nhat Thon, Martine De Cock, and Sergio Davalos. A Multivariate Regression Approach to Personality Impression Recognition of Vloggers. In *Proc. of WCPR14 workshop at ACMMM2014*, 2014.

[Getoor and Taskar, 2007] Lise Getoor and Ben Taskar. *Introduction to Statistical Relational Learning*. MIT press, 2007.

[Heider, 1958] Fritz Heider. The Psychology of Interpersonal Relations. *New York: Wiley*, 1958.

[Richardson and Domingos, 2006] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62:107–136, 2006.

[Zadeh, 1983] Lotfi A Zadeh. A Computational Approach to Fuzzy Quantifiers in Natural Languages. *Computers & Mathematics with Applications*, 9(1):149–184, 1983.