A Language and an Inference Engine for Twitter Filtering Rules

Alberto Bartoli[1] Barbara Carminati[2] Elena Ferrari[2] Eric Medvet[1]

1: Dipartimento di Ingegneria e Architettura, University of Trieste, Italy
2: Dipartimento di Scienze Teoriche e Applicate, Università dell'Insubria, Italy

October 15th, 2016

http://machinelearning.inginf.units.it

Table of Contents

Motivation

People use online social networks to share huge amount of information:

- maybe too much? $\rightarrow$ information overload
- maybe disturbing/unwanted? $\rightarrow$ trolls

Twitter particularly relevant.

Recommendation, spam, filtering

- Recommendation: select content to highlight that best fits user's interests
- Spam: select content to hide basing on content quality
- Filtering: select content to hide basing on explicit user's preferences

Recommendation, spam, filtering

- Recommendation: select content to highlight that best fits user's interests
- Spam: select content to hide basing on content quality
- Filtering: select content to hide basing on explicit user's preferences

Contributions

Filtering: select content to hide basing on explicit user's preferences

- how to specify a filtering policy?

Contributions

Filtering: select content to hide basing on explicit user's preferences

- how to specify a filtering policy? $\rightarrow$ filtering language

Contributions

Filtering: select content to hide basing on explicit user's preferences

- how to specify a filtering policy? $\rightarrow$ filtering language

Writing filtering policies may be too hard for the average Twitter user, so

- can a policy be inferred from examples?

Contributions

Filering: select content to hide basing on explicit user's preferences

- how to specify a filtering policy? $\rightarrow$ filtering language

Writing filtering policies may be too hard for the average Twitter user, so

- can a policy be inferred from examples? $\rightarrow$ policy inference from examples

Table of Contents

A simple model for the tweet

Given:

- topics $\mathcal{T} = \{$vulgarity, religion, politics, sex, work, alcohol, school, holiday, health$\}$
- post labels $\mathcal{L}_P = \{$hasMedia, hasHashtags, hasURLs$\}$
- author labels $\mathcal{L}_P = \{$isVip$\}$

A tweet p is given by $\langle T_P^p, T_A^p, L_P^p, L_A^p \rangle$:

- $T_P^p \subseteq \mathcal{T}$, topics of the tweet
- $T_A^p \subseteq \mathcal{T}$, topics of the author of the tweet
- $L_P^p \subseteq \mathcal{L}_P$, post labels of the tweet
- $L_A^p \subseteq \mathcal{L}_A$, author labels of the author of the tweet

Filtering policy

A filtering rule r is a tuple $\langle o_{T_P}, T_P^r, o_{T_A}, T_A^r, o_{L_P}, L_P^r, o_{L_A}, L_A^r \rangle$

- o_* are set operators: $\subseteq$ or $\not\subseteq$
- T_P^r, T_A^r are (empty) set of topics
- L_P^r, L_A^r are (empty) set of labels

A policy is a set of rules.

p is filtered by r if $T_P^r o_{T_P} T_P^p \wedge T_A^r o_{T_A} T_A^p \wedge L_P^r o_{L_P} L_P^p \wedge L_A^r o_{L_A} L_A^p$

- rule conditions are and-ed
- policy rules are or-ed

Example

$$r_1 = \langle \subseteq \{\text{vulgarity}\}, \subseteq \emptyset, \subseteq \emptyset, \subseteq \emptyset \rangle$$
$$r_2 = \langle \subseteq \{\text{politics}\}, \not\subseteq \{\text{politics}\}, \subseteq \emptyset, \subseteq \emptyset \rangle$$
$$r_3 = \langle \subseteq \{\text{sex}\}, \subseteq \emptyset, \subseteq \{\text{hasMedia}\}, \not\subseteq \{\text{isVIP}\} \rangle$$

Filters:

- all vulgar posts
- all the posts concerning politics not authored by users who usually tweet about politics
- all the posts concerning sex containing some media and not authored by a VIP user

Table of Contents

Problem statement

Given:

- a set P_+ of tweets to be filtered
- a set P_- of tweets not to be filtered

find the simplest consistent policy.

Solution (sketch)

An evolutionary algorithm: a set of candidate solutions is evolved by recombining and mutating fitter solutions.

- custom domain-specific individual representation (individual = rule)
- custom domain-specific genetic operators
- multi-objective fitness (minimize false rejection FRR, minimize acceptance FAR, minimize rule size)
- separate-and-conquer strategy to compose policy

Table of Contents

Aims, data, procedure

Aims:

- can the language express policies of realistic complexity?
- can the approach infer them from examples?

Data:

- from a large ($\geq 2 \cdot 10^6$) set of tweets, after cleaning...
- 1707 tweets in English with assigned topics

Procedure:

- 5 target policies (from 1 to 4 rules)
- generalization ability: policy are assessed on different sets
- 9 repetitions for each target policy

Results

#	$\lvert\rho^\star\rvert$	$\lvert P_+^0\rvert$	$\lvert P_-^0\rvert$	On P_+, P_- FRR	FAR	On $P_+^{\text{test}}, P_-^{\text{test}}$ FRR	FAR	$\lvert\rho\rvert$
1	1	110	1597	0.00	0.00	0.00	0.00	1
2	1	9	1698	0.00	0.00	0.00	0.00	1
3	2	196	1511	0.00	0.00	0.00	0.00	3
4	3	166	1541	0.00	0.00	0.00	0.00	3
5	4	32	1675	0.00	0.00	0.00	0.06	2
Avg.				0.00	0.00	0.00	0.01	

Results

#	$\lvert\rho^\star\rvert$	$\lvert P_+^0\rvert$	$\lvert P_-^0\rvert$	On P_+, P_- FRR	On P_+, P_- FAR	On $P_+^{\text{test}}, P_-^{\text{test}}$ FRR	On $P_+^{\text{test}}, P_-^{\text{test}}$ FAR	$\lvert\rho\rvert$
1	1	110	1597	0.00	0.00	0.00	0.00	1
2	1	9	1698	0.00	0.00	0.00	0.00	1
3	2	196	1511	0.00	0.00	0.00	0.00	3
4	3	166	1541	0.00	0.00	0.00	0.00	3
5	4	32	1675	0.00	0.00	0.00	0.06	2
Avg.				0.00	0.00	0.00	0.01	

- policies consistent with the examples are always found

Results

#	$\lvert\rho^\star\rvert$	$\lvert P_+^0\rvert$	$\lvert P_-^0\rvert$	On P_+, P_- FRR	On P_+, P_- FAR	On $P_+^{\text{test}}, P_-^{\text{test}}$ FRR	On $P_+^{\text{test}}, P_-^{\text{test}}$ FAR	$\lvert\rho\rvert$
1	1	110	1597	0.00	0.00	0.00	0.00	1
2	1	9	1698	0.00	0.00	0.00	0.00	1
3	2	196	1511	0.00	0.00	0.00	0.00	3
4	3	166	1541	0.00	0.00	0.00	0.00	3
5	4	32	1675	0.00	0.00	0.00	0.06	2
Avg.				0.00	0.00	0.00	0.01	

- policies consistent with the examples are always found
- good generalization ability

Results

#	$\lvert\rho^\star\rvert$	$\lvert P_+^0\rvert$	$\lvert P_-^0\rvert$	On P_+, P_- FRR	On P_+, P_- FAR	On $P_+^{\text{test}}, P_-^{\text{test}}$ FRR	On $P_+^{\text{test}}, P_-^{\text{test}}$ FAR	$\lvert\rho\rvert$
1	1	110	1597	0.00	0.00	0.00	0.00	1
2	1	9	1698	0.00	0.00	0.00	0.00	1
3	2	196	1511	0.00	0.00	0.00	0.00	3
4	3	166	1541	0.00	0.00	0.00	0.00	3
5	4	32	1675	0.00	0.00	0.00	0.06	2
Avg.				0.00	0.00	0.00	0.01	

- policies consistent with the examples are always found
- good generalization ability, some errors only with the most complex target policy

Thanks!