ITMO UNIVERSITY

On some humble tries to predict some properties of some population of some cities

Ivan Samborskii[1,2], Leonid Startsev[2,3] and **Andrey Filchenkov**[2,3]

[1]School of Computing, National University of Singapore

[2]Computer Technology Chair, ITMO University

[3]Computer Technologies Lab, ITMO University

AINL-ISMW FRUCT'15. Nov 12, 2015, St. Petersburg, Russia

Team

- **Ivan Samborskii**
  - junior researcher and Master student
  - **role**: Java coder, insider

- **Leonid Stratsev**
  - senior lab assistant and Bachelor student
  - **role**: Python coder, enthusiast

- **Andrey Filchenkov**
  - PhD, researcher, associate professor
  - **role**: PowerPoint coder, demagogue

Core steps

- Task selection
  - Too many data so we need to define what to do
- Data preprocessing
  - We need to prepare selected data
- Feature engineering
  - Applying feature selection and feature extraction
- Classification
  - Classifying results

Task selection (by Andrey)

- Gender prediction – OK
- Age + education prediction:
  - (In school) OR (Age $\leq$17)
  - (Student) OR (19 $\leq$ Age $\leq$ 22)
  - (Age $\geq$23)
- Relationship prediction
  - Has someone (relationship; future wedding; marriage)
  - Alone
- Occupation prediction – SORRY, NO (too many labels, no fast ways to build macrolabels)

Future work

- Extremely imbalanced classes: merging classes in a more intellectual ways.
- Facebook ground truth may be ok for gender and sometimes for age, but not for occupation and education (LinkedIn seems to be much better in these terms) and not for relationship (surveys only for tests, Facebook statuses may be used as noisy labels).
- For precise selection of labels: estate or income level are more useful, than occupation or education.

Data Preprocessing (by Leonid and Ivan)

- Merging feature sets
- Adding proper class labels
- Throwing away unlabeled objects for each task

Future work

- Noise (object-based and label-based) cleaning

Feature selection (Ivan)

- We used the following feature sets:

1. 4S_o6
2. 4S_LDA
3. Twi_(LIWC+Custom)
4. Twi_LDA
5. Insta
6. All the basic
7. FS_1: *a filter* for (All the basic)
8. FS_2: PCA for (4S_L + Twi_(LIWS+Custom))
9. FS_3: *a filter* for (Insta_L)
10. FS_4: *a filter* for (FS_2 + FS_3)

ITMO UNIVERSITY

Filter selection (sponsored commercial)

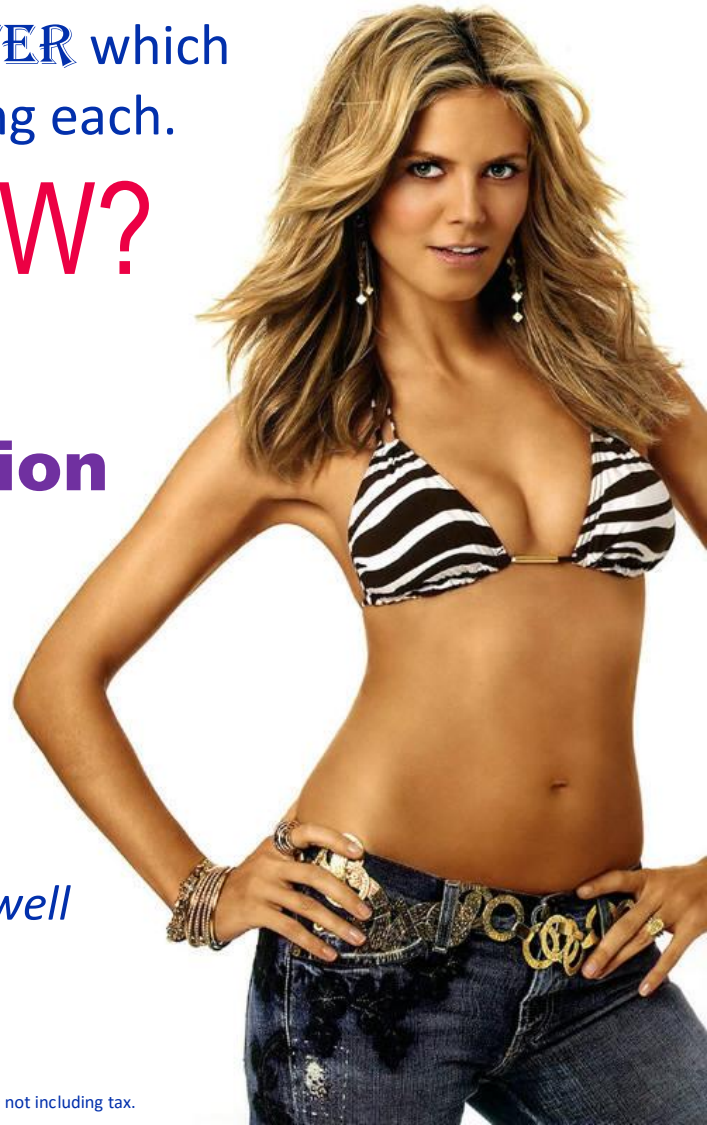
On the previous step *a filter* is the BEST FILTER which was chosen AUTOMATICALLY, without trying each.

WANNA KNOW, HOW?

Only on **Saturday**, **Section 7**, at **18:50**

Dataset Meta-Feature Description for Recommending Feature Selection Algorithm

by Andrey Filchenkov and Arseniy Pendryak



*

"I really enjoyed to read this paper. It is well presented and very understandable."

Reviewer #3

*The proposal is valid only for AINL-ISMW FRUCT participants and CT students. For any other attender the price is $79.99 not including tax.

Datasets for each labels

sex_binary	age_education	relationship_binary
4S_o6		
4S_L		
Twi_(LIWC+Custom)		
Twi_LDA		
Insta		
FS_1_s_b	FS_1_a_e	FS_1_r_b
FS_2		
FS_3_s_b	FS_3_a_e	FS_3_r_b
FS_4_s_b	FS_4_a_e	FS_4_r_b
All the original		

Future work

- Apply LDA for 4S + twi words frequencies
- Apply LDA for 4S + insta + twi words frequencies
- Put more afford on how to choose features
- Group 4S in 2, 4 and 6 months into features representing behavior stability

Classification (Leonid, and Ivan a bit)

- We used the mostly “vanilla” classifiers:

1. NB : GaussianNB() in python
2. kNN : KNeighborsClassifier() in python
3. AdaBoost : AdaBoostClassifier() in python
4. RF : RandomForestClassifier() in python
5. SVM : SVC() in python
6. BayNet: BayesNet() in WEKA
7. Stacking: Stacking() in WEKA

Future work (1/2)

- Since there are many unlabeled data, it's better to use semi-supervised learning: we build constrained clustering on all the data, such that
  - each cluster has a labelled object
  - labels with different labels are not allowed to appear in a cluster
  - after cluster making with labels of labelled objects in that, the distribution of classes should be similar to the labels distribution in population.
- It can be softed with choosing proper quality measure

Future work (2/2)

- Age should be predicted as ranking, not category
- Algorithm parameters fine-tuning
- More complex structure of classifiers. Use more strong models such as DNNs

Results: Singapore age + education

		AdaBoost	SVM	Naïve Bayes	Random Forest	kNN	Bayes Network
1	Foursquare (LDA)	0,441	0,688	0,350	0,664	0,664	0,356
2	Foursquare (venues)	0,706	0,694	0,301	0,709	0,647	0,410
3	Instagram (image concepts)	0,697	0,708	0,245	0,689	0,647	0,276
4	Twitter (LIWC + heuristic text)	0,488	0,690	0,448	0,703	0,634	0,436
5	Twitter (LDA)	0,502	0,687	0,275	0,713	0,691	0,444
6	Multi-Source	0,595	0,701	0,361	0,714	0,635	0,443
7	Signific (Multi-Source)	0,607	0,716	0,275	**0,732**	0,686	0,438
8	PCA (2 + 4)	0,408	0,692	0,619	0,692	0,667	0,272
9	CFS-GS (3)	0,706	0,707	0,254	0,686	0,655	0,276

Results: Singapore gender

		AdaBoost	SVM	Naïve Bayes	Random Forest	kNN	Bayes Network
1	Foursquare (LDA)	0,604	0,598	0,580	0,589	0,570	0,384
2	Foursquare (venues)	0,656	0,619	0,511	0,658	0,573	0,414
3	Instagram (image concepts)	0,584	0,688	0,524	0,703	0,654	0,369
4	Twitter (LIWC + heuristic text)	0,708	0,542	0,589	0,722	0,536	0,403
5	Twitter (LDA)	0,765	0,636	0,650	0,772	***0,690***	0,395
6	Multi-Source	0,794	0,519	0,550	0,767	0,530	0,369
7	Cons-RS (Multi-Source)	***0,798***	0,699	0,603	***0,800***	0,616	0,431
8	PCA (2 + 4)	0,714	***0,760***	0,560	0,730	0,676	0,457
9	Cons-SWS (3)	0,568	0,713	0,526	0,701	0,673	0,371
10	Cons-LS (8 + 9)	0,719	0,740	***0,690***	0,727	0,687	***0,468***

Results: Singapore relationship

		AdaBoost	SVM	Naïve Bayes	Random Forest	kNN	Baysian Network
1	Foursquare (LDA)	0,664	0,688	0,666	0,653	0,638	0,272
2	Foursquare (venues)	0,651	0,685	0,397	0,675	0,619	0,272
3	Instagram (image concepts)	0,685	0,696	0,396	0,662	0,625	0,291
4	Twitter (LIWC + heuristic text)	0,659	0,687	0,653	0,701	0,644	0,397
5	Twitter (LDA)	0,648	0,688	0,662	0,694	0,681	0,383
6	Multi-Source	0,660	***0,704***	0,406	0,699	0,652	0,392
7	Cons-SFS (Multi-Source)	0,672	0,703	***0,723***	***0,727***	0,665	***0,404***
8	PCA (2 + 4)	0,646	0,700	0,660	0,685	***0,686***	0,370
9	Cons-SWS (3)	***0,694***	0,695	0,695	0,608	0,645	0,274
10	CFS-TS (8 + 9)	0,674	0,699	0,685	0,674	0,639	0,347

Results: London age + education

Results: London gender

		AdaBoost	Random Forest	SVM	kNN	Naïve Bayes	Bayes Network
1	Foursquare (LDA)	0,694	0,671	0,702	0,638	0,677	0,372
2	Foursquare (venues)	0,702	0,698	0,702	0,644	0,354	0,275
3	Instagram (image concepts)	0,635	0,632	0,655	0,582	0,572	0,393
4	Twitter (LIWC + heuristic text)	?	?	?	?	?	0,335
5	Twitter (LDA)	0,740	**0,746**	0,702	0,695	0,580	0,391
6	Multi-Source	?	?	?	?	?	0,437
7	Cons-RS (Multi-Source)	0,723	0,714	0,660	0,618	0,680	0,417
8	PCA (2 + 4)	0,686	0,700	0,703	0,680	0,575	0,389
9	Cons-SWS (3)	0,645	0,612	0,656	0,610	0,639	0,343
10	Cons-LS (8 + 9)	0,647	0,648	0,656	0,626	0,607	0,372

Results: London relationship

		AdaBoost	Random Forest	SVM	kNN	Naïve Bayes	Bayes Network
1	Foursquare (LDA)	0,576	0,595	0,586	0,578	0,591	0,315
2	Foursquare (venues)	0,584	0,603	0,557	0,519	0,492	0,241
3	Instagram (image concepts)	0,505	0,484	0,442	0,484	0,516	0,393
4	Twitter (LIWC + heuristic text)	?	?	?	?	?	0,229
5	Twitter (LDA)	0,610	0,629	0,568	0,640	0,596	0,406
6	Multi-Source	?	?	?	?	?	0,414
7	Cons-SFS (Multi-Source)	0,656	**0,684**	0,627	0,634	0,582	0,397
8	PCA (2 + 4)	0,597	0,645	0,603	0,550	0,515	0,417
9	Cons-SWS (3)	0,530	0,558	0,445	0,513	0,516	0,230
10	CFS-TS (8 + 9)	0,660	0,603	0,610	0,589	0,482	0,415

Results summary

- Age should be predicted as ranking, not category
- Algorithm parameters fine-tuning
- More complex structure of classifiers. Use more strong models such as DNNs