A comparative analysis of various machine learning techniques for trend prediction on social media

Application oriented project

CSI5155

Diana Lucaci - group 8

University of Ottawa

23rd April 2019

Motivation

What?

- Online content - rich source of information
- Detect changes in behaviours, trends, predict popularity

Why?

- Inform people, business, and authorities
- Help individuals prepare and act accordingly
  (high sales growth, major events, strikes)

Objectives

- Draw insights from data
- Find the most promising features
- Research ML techniques and best practices
- Binary classification of Twitter instances

How?

- Analyze the dataset
- Preprocessing techniques
- Perform experiments using ML algorithms from 4 categories:
  - Linear models
  - Tree-based
  - Distance-based
  - Rule-based
  - Ensemble

Dataset

- 11 attributes measured over a 7 days period (total of 77 attributes)
- Binary relative labeled data (increment by 500 of popularity level before and after the observed time frame)
- Class ratio:
  - 97.5% negative instances
  - 2.5 % positive instances (BUZZ)

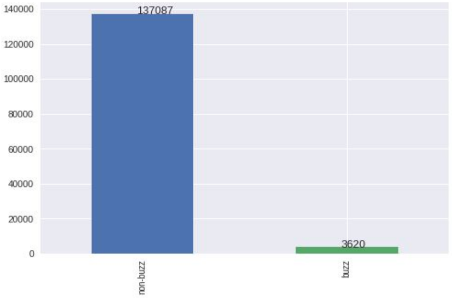


Figure 1: Dataset imbalance

Attributes

1. Number of Created Discussions (NCD)
2. Author increase (AI)
3. Number of atomic containers (NAC)
4. Number of Authors (NA)
5. Attention Level
    - AS(NA) - based on the number of authors (users)
    - AS(NAC) - measure with number of contributions
6. Burstiness Level (BL)
7. Contribution Sparseness (CS)
8. Author Interaction (AT)
9. Average Discussions Length (ADL)
10. Number of Active Discussions (NAD)

Data manipulation

1. **Linear scaling to unit variance - Standardization**

Data manipulation

1. **Linear scaling to unit variance - Standardization**
    - Transformation to a new distribution with mean 0 and standard deviation 1
    - Obtained the best results for distance-based and linear models

Data manipulation

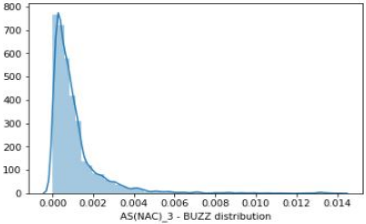
1. **Linear scaling to unit variance - Standardization**
    - Transformation to a new distribution with mean 0 and standard deviation 1
    - Obtained the best results for distance-based and linear models
2. **Min-Max scaling**
    - preserves the initial distribution
    - the larger the range, the better the performance

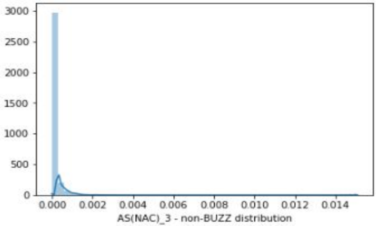
Data manipulation

1. **Linear scaling to unit variance - Standardization**
   - Transformation to a new distribution with mean 0 and standard deviation 1
   - Obtained the best results for distance-based and linear models
2. **Min-Max scaling**
   - preserves the initial distribution
   - the larger the range, the better the performance
3. **Robust scaling**
   - It uses interquartile range and it is robust to outliers.
   - Non-buzz instances contain outliers => higher precision

Data manipulation

1. **Linear scaling to unit variance - Standardization**
   - Transformation to a new distribution with mean 0 and standard deviation 1
   - Obtained the best results for distance-based and linear models
2. **Min-Max scaling**
   - preserves the initial distribution
   - the larger the range, the better the performance
3. **Robust scaling**
   - It uses interquartile range and it is robust to outliers.
   - Non-buzz instances contain outliers => higher precision





Data manipulation

1. **Linear scaling to unit variance - Standardization**
    - Transformation to a new distribution with mean 0 and standard deviation 1
    - Obtained the best results for distance-based and linear models
2. **Min-Max scaling**
    - preserves the initial distribution
    - the larger the range, the better the performance
3. **Robust scaling**
    - It uses interquartile range and it is robust to outliers.
    - Non-buzz instances contain outliers => higher precision
4. **Rowwise normalization to unit norm**
    - preserves the distribution rowwise
    - helps normalizing the values of features with very different ranges (eg. [0,1] and [0,12000])
    - L2 normalization improves the F1 score

Feature selection

Principal Component Analysis

- 95% of the data is explained by 17 features
- 35 of the features have the most prominent influence on the generated principal components
- assumptions: data is zero centered $=>$ standardization needed

Evaluation metrics

- **F1 score**
- AUC weighted by support (the number of true instances for each label)
- **TPR (recall)**
- **Precision**
- TNR

Results

	F1	AUC	**TPR (recall)**	**Precision**	TNR
DT_IG	0.456540	0.7210	0.456233		0.9857
KNN_st_sc _norm_L2	**0.553923**	0.7317	0.469306	0.676864	0.9941
NC_mahalanobis	0.488551	0.7755	0.572034	0.432421	0.9791
LinSVC_minmax (0,37505)	0.473080	0.6761	0.356028	0.731774	0.9962
LinSVC_minmax (-300,300)	**0.338376**	**0.9015**	0.892426	0.208842	0.9106
LogReg_unitvar	0.493299	0.6848	0.373457	0.735563	0.9963
SVM_SGD _L2_norm_unitvar	0.516521	0.7605	0.535105	0.514327	0.9859
RF_st	**0.570278**	0.7239	0.451403	0.776747	0.9965
BAG+RF	0.548917	0.7087	0.420316	0.794784	0.9971

Comparisons

Table 1: T-Test - F1

Algotrithm pair	p-value	Statistical difference (95% conf)
BAG_DT - BAG_KNN	0.05712	yes
BAG_DT - BAG+KNN_stsc_norm_L2	1.17e-05	yes
BAG_KNN - BAG+KNN_stsc_norm_L2	5.12e-06	yes
KNN_stsc_L2 - BAG_DT	1.55e-05	yes
KNN_stsc_L2 - BAG+KNN_stsc_norm_L2	1.93e-05	yes
KNN_stsc_L2 - BAG+KNN_stsc_norm_L2	0.2565	no

Key learnings

1. Evaluation metrics importance
2. Testing the generalization power
3. Comparing multiple algorithms
4. The class-imbalance problem needs to be addressed using techniques suitable for the data set on discussion
5. Important features
6. Dependent features (NAC vs Attention Level)

Future work

- Neural networks and deep learning
- Addressing the class imbalance problem with different techniques
  - Learn the minority class itself instead of by comparing it with the non-BUZZ class
  - Oversampling and undersampling
  - Improve the models learning on each mini-dataset
- Other methods of feature selection (e.g. univariate selection)
- Inhance the dataset with the text of the posts
- Regression task and comparisons with other datasets
- Use only the most important features extracted from the PCA into the weak classifiers ensemble

References

- Dataset: http://archive.ics.uci.edu/ml/datasets/Buzz+in+social+media+#


THANK YOU FOR YOUR
ATTENTION
ANY QUESTIONS?