A Study of Feature Construction for Text-based Forecasting of Time Series Variables

Yiren Wang*, Dominic Seyler*, Shubhra Kanti Karmaker Santu, ChengXiang Zhai

{yiren, dseyler2, karmake2, czhai} @illinois.edu

Motivation

- How to leverage text data that is related to a time series variable for time series forecasting?
  - Present a general formulation of the problem
  - Present a general strategy for constructing data sets
  - Study general strategies for constructing and combining word and topical features

Problem Formulation

Problem Definition

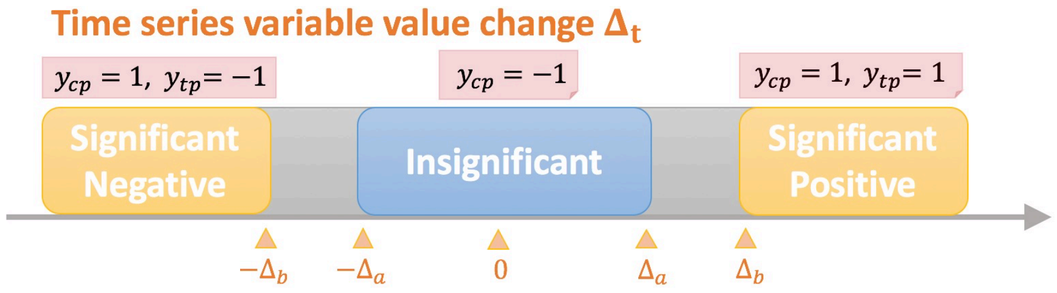
• **Change Prediction Problem (CP):** y_{cp} is the label that marks whether there is significant change in time series variable.

$$y_{cp}^{(t)} = \begin{cases} 1, & \Delta_t \in (-\infty, -\Delta_b] \cup [\Delta_b, +\infty) \text{ (significant)} \\ -1, & \Delta_t \in [-\Delta_a, \Delta_a] \text{ (insignificant)} \end{cases} \quad (1)$$

• **Trend Prediction Problem (TP):** y_{tp} marks whether the direction of change is positive or negative.

$$y_{tp}^{(t)} = \begin{cases} 1, & \Delta_t \in [\Delta_b, +\infty) \text{ (positive)} \\ -1, & \Delta_t \in (-\infty, -\Delta_b] \text{ (negative)} \end{cases} \quad (2)$$

Time series variable value change Δ_t



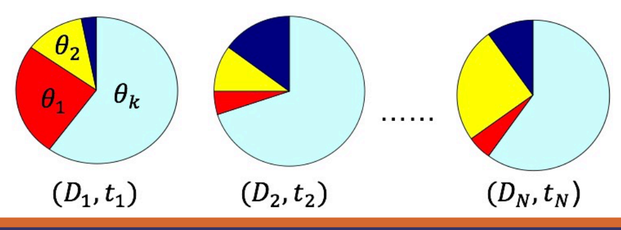
- Relative improvement: $\Delta_t = (x_{t+1} - x_t)/x_t$
- Thresholds: $0 \leq \Delta_a \leq \Delta_b$

Input Data:

- Time series $X = \{(x_1, t_1), \ldots, (x_N, t_N)\}$
- Text document $D = \{(d_1, t_1), \ldots, (d_N, t_N)\}$
- Label: $y_{cp}, y_{tp} \in \{-1, 1\}$

Feature Construction:

- **Word-based features:**
  - Raw count: $f_c(n) = count(w(n), d)$
  - TFIDF score: $f_{tfidf}(n) = f_c(n) \times \log\left(\frac{1+N_d}{1+df(d,w(n))}\right)$
- **Topic-based features:**
  - Topic Distribution (TD): θ_t
  - Topic Change (Chg): $\theta_t^{\Delta} = \theta_t - \theta_{t-1}$
  - Topic History ($Hist$): $\theta_t^{\text{hist}}(\alpha, L) = \Sigma_{i=1}^{L} \alpha^i \theta_{t-i}$



Experiments

Dataset

- Stock prices of 53 companies over 7 years
- Corpus of Reuters and Bloomberg news articles
- Map companies to documents using Lucene index
- Concatenate documents per company, per day
- Available for download: http://bit.ly/2w12Ybp

Machine Learning Framework

- **Classifier:** Logistic-regression (LR), 3-fold-cross validation
- **Feature Selection**: Top-k according to chi-square (x^2) and mutual-information (MI)

Models

- **Vector Auto Regression (VAR)**: econometric model that captures the linear inter-dependencies among multiple TS
- **Time Series-based Features**: percentage of change in opening prices as features for logistic regression
- **Text-based Features**: n-gram features with n = 1, 2, feature selection and topic modeling (Latent Dirichlet Allocation)

Models	Performance
VAR	50.95
Price Change	50.71
n-grams (n = 1)	56.26
n-grams (n = 2)	56.20
n-grams (n = 1) + Price Change	56.39
n-grams (n = 2) + Price Change	56.23
Topic-based	52.73
Topics + n-grams (n = 1)	56.02
Topics + n-grams (n = 2)	56.00
Topics + n-grams + Price change	**56.44**

Table 1: Comparison of different models (TP)

Unigram		Bigrams	
X^2	MI	X^2	MI
fell	percent	fell percent	share fell
myspace	apple	rose percent	rose percent
yahoo	company	drop percent	fell percent
rose	google	gain percent	gain point
burbank	has	fell point	new york

Table 2: Top-5 selected n-grams using $X^2$and MI (TP)

k	TD	Chg	TD + Chg	Hist.add	Hist.cont	TD + Chg + Hist.add	TD + Chg + Hist.cont
10	52.92 / 51.91	52.25 / 51.88	52.78 / 52.06	51.99 / 51.71	**53.38** / 52.10	52.28 / 52.58	53.15 / ***52.73***
15	**54.40** / 51.13	52.58 / **51.80**	53.96 / 51.49	54.15 / 51.28	54.33 / 51.62	54.10 / 51.60	54.07 /51.67
20	53.96 / 51.47	53.55 / 52.43	54.07 / 51.02	54.04 / 50.87	54.29 / 52.17	54.45 / 51.56	**54.49** / **51.73**
25	54.01 / 50.87	53.70 / 50.97	53.95 / 51.06	55.13 / 50.58	**54.80** / **52.36**	54.78 / 51.39	54.60 / 51.41
30	**55.21** / 51.34	53.06 / 51.19	54.09 / 51.52	54.98 / 51.80	54.98 / 51.86	54.53 / 51.62	54.65 / **52.04**
35	54.64 / 50.84	53.71 / 50.32	54.53 / 51.32	54.66 / 51.28	**54.83** / 51.86	54.48 / 51.36	54.43 / **52.06**
40	54.72 / 50.76	54.09 / 51.52	53.33 / 51.71	54.83 / 51.13	54.53 / 51.21	**54.57** / **51.97**	54.41 / 51.49
45	53.52 / 50.61	53.42 / 50.84	53.79 / 50.61	53.84 / 50.35	53.87 / 51.36	**54.26** / 51.06	53.95 / **52.32**
50	53.39 / 51.23	53.52 / 51.26	52.63 / 50.37	53.70 / 51.13	53.79 / **52.14**	**54.30** / 51.13	53.93 / 51.88

Table 3: Performances of topic-based features (CP / TP)

Conclusions and Future Work

Experiment Conclusions

- Text-based features are more effective than time-series features
- Topic-based features can be combined with the word-based features to further improve accuracy
- Best performance is achieved when word-based, topic-based and time series-based features are used in combination.

Future Work

- Exploration of our features on different datasets (e.g., political news for election forecasting)
- Derive other topic-based features and explore how deep learning can be leveraged for this task test classifiers in addition to logistic regression
- Show how our features can be utilized in concrete applications, e.g., decision support for stock trading

DAIS The Data and Information Systems Laboratories at The University of Illinois at Urbana-Champaign
Large Scale Information Management and Mining

TIMAN