4 - Support of managerial decision making processes by transductive learning

Hubertus Brandner, Universität Hamburg, Germany, hubertus.brandner@studium.uni-hamburg.de, *Stefan Lessmann*

This study analyses to which extent the promising findings of transductive approaches can be transfered to business problems of classification. Different variants of Support Vector Machines are examined to compare the established inductive learning and the transductive technique. To that end a hybrid metaheuristic is implemented to solve the mathematical programming formulations in the same way. Empirical results confirm the potential of transductive inference. Therefore it is advisable to utilize the information of unlabeled data in the context of managerial decision making and planning.

■ WA-21

Wednesday, 9:00-10:20
6.2.47

Optimization Algorithms I

Stream: Software for OR/MS
Invited session

Chair: *Simone Garatti*, Dept. of Electronics and Information, Politecnico di Milano, p.zza L. da Vinci 32, 20133, Milan, Italy, sgaratti@elet.polimi.it

1 - Solving uncertain programs via the scenario approach: the FAST algorithm

Simone Garatti, Dept. of Electronics and Information, Politecnico di Milano, p.zza L. da Vinci 32, 20133, Milan, Italy, sgaratti@elet.polimi.it, *Algo Carè*, *Marco Campi*

Uncertainty is ubiquitous in decision problems, and this leads naturally to uncertain programs (UP). Robust and chance-constrained solutions to UP can be difficult to obtain in general. In this talk, we discuss the use of the scenario approach, a handy methodology based on random sampling of constraints, to solve UP with a guaranteed degree of approximation. In particular, we introduce FAST (Fast Algorithm for the Scenario Technology), a variant of the standard scenario algorithm with reduced complexity, which improves the applicability of the scenario methodology to a high extent.

2 - Scheduling optimization in virtual enterprises based on the hybridization of a CSP with a genetic algorithm

Rabah Kassa, mathematique, Universite Bejaia algerie, universite de bejaia 06000 bejaia algerie, 06000, bejaia, Algeria, rabah_kassa2002@yahoo.fr, *Djamila Boukredera*, *Zaidi Sahnoun*

Production scheduling represents an important manufacturing function whose quality remains an essential stake for virtual enterprises. To optimize its scheduling, a virtual enterprise aims to improve its profitability while minimizing the customer's service costs and respecting manufacturing constraints. This can be formulated as a CSP. We suggest an optimization method of the CSP based on the genetic algorithm. This hybridization aim at better taking over of this kind of problem defined by a large research space and a complex constraint set and finds solutions of good quality.

3 - ParadisEO: a framework for metaheuristics

El-ghazali Talbi, University of Lille - INRIA - CNRS, Lille, El-ghazali.Talbi@lifl.fr

We present the ParadisEO white-box object-oriented framework dedicated to the reusable design of metaheuristics. It provides a broad range of features including population based metaheuristics and single-solution metaheuristics. It basedes on a conceptual separation of the solution methods from the problems they are intended to solve. The fine-grained nature of the classes allows a high flexibility. ParadisEO is of the rare frameworks providing most common parallel and distributed models; implementation is portable and models can be exploited transparently.

■ WA-23

Wednesday, 9:00-10:20
6.2.49

Model Selection in Regression Analysis

Stream: Data Mining in the Financial Sector
Invited session

Chair: *Michael Khachay*, Ural Branch of RAS, Institute of Mathematics and Mechanics, S.Kovalevskoy, 16, 620990, Ekaterinburg, Russian Federation, mkhachay@imm.uran.ru

Chair: *Vadim Strijov*, Computing Center of the Russian Academy of Sciences, Klara Zetkin 13-79A, 127299, Moscow, Russian Federation, strijov@ccas.ru

1 - Model generation and model selection in credit scoring

Vadim Strijov, Computing Center of the Russian Academy of Sciences, Klara Zetkin 13-79A, 127299, Moscow, Russian Federation, strijov@ccas.ru

The credit scorecard is the logistic regression model; it maps the feature space to the probability of default of a banking client. A classical scorecard is constructed by an analyst, who manually selects informative features and creates combinations of them. We propose a new technique for the automatic scorecard construction. To develop a scorecard, one must assign a set of primitive functions and model generation rules. The result model is an admissible superposition of the primitive functions and features. The coherent Bayesian inference is used to select features and their superpositions.

2 - Algorithms of feature selection for volatility estimation of European options

Ekaterina Krymova, Control/Management and Applied Mathematics, Moscow Institute of Physics and Technology, 9,35 b.3, Nagornaya st,,Moscow, 141981, Moscow region, Dubna, Bogolubova 33-304, 117186, Moscow, Russian Federation, ekkrym@gmail.com

The problem of multicollinearity is commonly encountered in regression analysis. This problem may lead to overfitting and result in unstable model parameters. New approach to the feature generation and feature selection was proposed. The feature generation technique is based on Kolmogorov-Gabor polynomial construction. The features are superpositions of primitive functions and free variables. The generated features require reduction of multicollinearity. For this purpose, the LARS modification is developed. Historical data of European options is used as practical example.

3 - A topological approach to formulating conditions of the uniform convergence of frequencies to probabilities

Michael Khachay, Ural Branch of RAS, Institute of Mathematics and Mechanics, S.Kovalevskoy, 16, 620990, Ekaterinburg, Russian Federation, mkhachay@imm.uran.ru

Existence of the uniform convergence of frequencies to probabilities over an appropriate events class is a well known sufficient consistency condition of the empirical risk minimization (ERM) in machine learning. The traditional approach for proving such convergence is based on a sublinear growth of entropy of the event class in question and obtaining upper VCD bounds for this class. In this paper, existence of the uniform convergence of frequencies to probabilities over an event class is related to some topological properties of the sigma-algebra, induced by this class.

4 - Benchmarking Framework for Financial Text Mining

Caslav Bozic, Institute AIFB, IME Graduate School, Karlsruhe Institute of Technology (KIT), Institute AIFB - 05.20, KIT Campus South, 76128, Karlsruhe, BW, Germany, bozic@kit.edu

Different data mining methods for financial text and various sentiment measures are described in the existing literature, without common benchmark for comparing these approaches. Implemented system (which is a part of FINDS Project) and proposed framework are based on theoretical data integration, and they facilitate combining more sources of financial data into comprehensive integral dataset. The dataset is then used to analyse the candidate measure by regressing it on different returns and other financial indicators that can be defined using the system's novel data transformation approach.

Benchmarking Framework for Financial Text Mining

Caslav Bozic
bozic@kit.edu

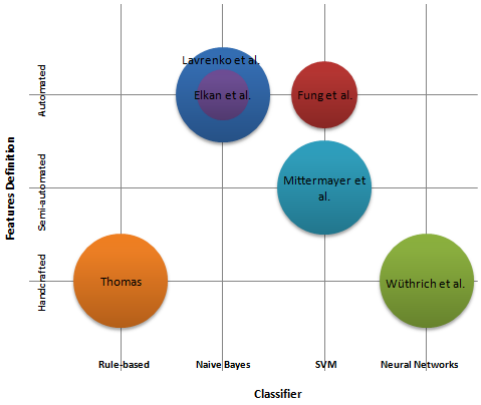
Applied Informatics and Formal Description Methods (AIFB)
Information Management and Market Engineering (IME)
Karlsruhe Institute of Technology (KIT), Germany

Agenda

FINDS Project

- Financial News and Data Service
  - Conducting innovative research on the analysis of quantitative and qualitative information from financial markets
  - Amount of financial data available (previous trades, news stories) makes it impossible for a human trader to process it in whole
  - Services to help traders by
    - filtering important news releases
    - suggesting buy-sell decisions
    - allowing making subjective connections within the data

Related Work



Agenda

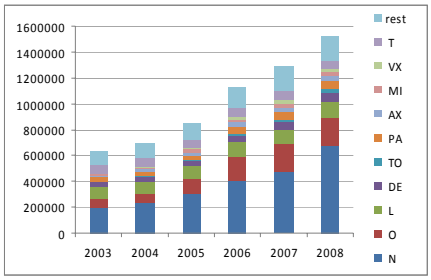
(i) Motivation
(ii) Data Description
(iii) Data Processing
(iv) First Results
(v) Summary & Future Work
(vi) References

Data Sources

- Thomson Reuters TickHistory
  - order book (some)
  - best bid and ask (most)
  - trades (all major exchanges)
  - indices values
- Reuters NewsScope Sentiment Engine
  - sentiment measure for all English-language news published through Reuters NewsScope in period 2003-2006
- Reuters Takes
  - full text of news stories for 2003

Sentiment Data

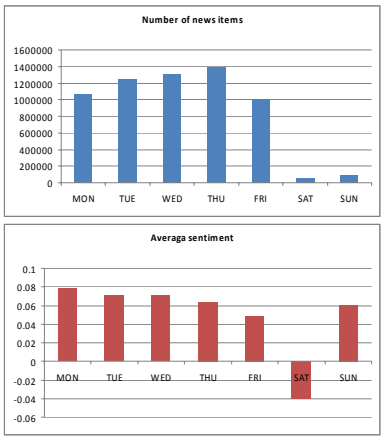
- 6 Mio records about 10,000 different companies
- 2.5 times increase in yearly volume in period 2003 – 2008
- 2 biggest US markets (NYSE & NASDAQ)
  - 40% in 2003
  - 60% in 2008



Number of records per year

Sentiment Data

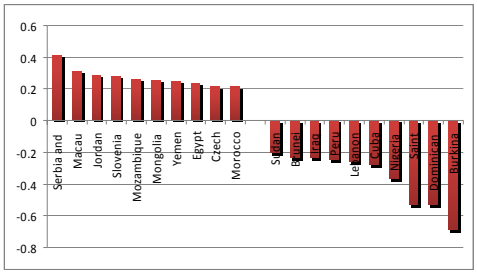
- Negative sentiment on Saturdays



Number of news items and average sentiment per days of week

Sentiment Data

- Example of aggregation: sentiment per country

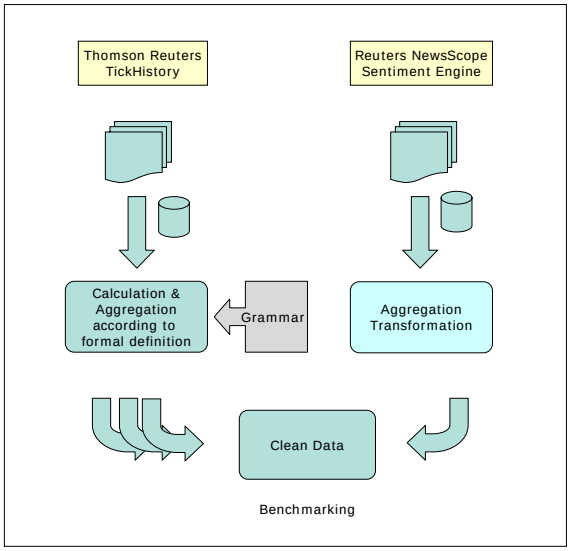


Best and worst average sentiment for countries with over 1000 mentions

Agenda

(i) Motivation
(ii) Data Description
(iii) Data Processing
(iv) First Results
(v) Summary & Future Work
(vi) References

Data Processing

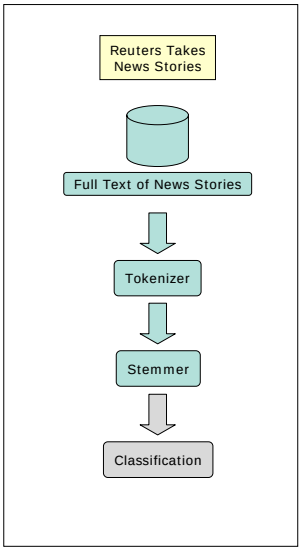


Agenda

(i) Motivation
(ii) Data Description
(iii) Data Processing
(iv) First Results
(v) Summary & Future Work
(vi) References

First Results

- 3 classifiers
  - Bayes – Fisher
  - SVM
  - Neural Network



First Results

- Comparison
  - Classifiers trained on first 9 months 2003
  - 5 big technology companies: IBM, Oracle, Microsoft, Apple, SAP
  - Tested on 3 last months 2003 – 729 news stories
  - 9 variables – 10 lagged values each
  - Statistically relevant relation could be proven for
    - Bayes-Fisher 2 values
    - SVM 1 value
    - Neural network 1 value
    - RNSE 7 values
  - Not enough data to draw certain conclusions
  - Longer period needed, more companies
  - RNSE data for NYSE and NASDAQ in period 2003 - 2008
    - Statistical relevance improved

Agenda

(i) Motivation
(ii) Data Description
(iii) Data Processing
(iv) First Results
(v) Summary & Future Work
(vi) References

Summary

- Need for benchmarking method for financial text mining
- Proposed framework and implemented system for
  - Flexible adding of new data sources
  - Formal definition of calculated fields and aggregations
- Test on 4 sentiment measures
  - 3 months on 5 companies – short period for statistical relevance
  - 6 years on 2 markets – promising results
- Future Work
  - Adding new types of data sources (Compustat for mc & earnings)
  - More sophisticated statistical analysis (heteroscedastic data)

Benchmarking Framework for Financial Text Mining

Caslav Bozic
bozic@kit.edu

Applied Informatics and Formal Description Methods (AIFB)
Information Management and Market Engineering (IME)
Karlsruhe Institute of Technology (KIT), Germany

Thank you for your attention.

Discussion

Agenda

References

- [1] Hevner, A.R., March, S.T., Park, J. & Ram, S., Design Science in Information Systems Research, MIS Quarterly, Management Information Systems Research Center, University of Minnesota, 2004, Vol. 28(1), pp. 75-105
- [2] FINDS - Integrative services, Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on, 2009, pp. 61-62
- [3] Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. & Allan, J., Mining of Concurrent Text and Time-Series, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000
- [4] Gidófalvi, G. & Elkan, C., Using news articles to predict stock price movements, Department of Computer Science and Engineering, University of California, San Diego, 2003
- [5] Pui Cheong Fung, G., Xu Yu, J. & Lam, W., Stock prediction: Integrating text mining approach using real-time news, Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on, 2003, pp. 395 - 402
- [6] Mittermayer, M.-A. & Knolmayer, G.F., NewsCATS: A News Categorization and Trading System, Data Mining, IEEE International Conference on, IEEE Computer Society, 2006, Vol. 0, pp. 1002-1007
- [7] Thomas, J., News and trading rules, 2003
- [8] Schulz, A., Spiliopoulou, M. & Winkler, K., Kursrelevanzprognose von Ad-hoc-Meldungen: Text Mining wider die Informations\überlastung im Mobile Banking, Wirtschaftsinformatik, 2003, Vol. 2, pp. 181-200
- [9] Antweiler, W. & Frank, M.Z., Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, The Journal of Finance, Blackwell Publishing for the American Finance Association, 2004, Vol. 59(3), pp. 1259-1294

References

- [10] Das, S. & Chen, M., Yahoo! for Amazon: Sentiment extraction from small talk on the web, Management Science, INFORMS, 2007, Vol. 53(9), pp. 1375-1388
- [11] Tetlock, P., Giving Content to Investor Sentiment: The Role of Media in the Stock Market, THE JOURNAL OF FINANCE, 2007, Vol. 62(3)
- [12] Tetlock, P., Saar-Tsechansky, M. & Macskassy, S., More Than Words: Quantifying Language to Measure Firms' Fundamentals, Journal of Finance, American Finance Association, 2008, Vol. 63(3), pp. 1437-1467
- [13] Pfrommer, J., Hubschneider, C. & Wenzel, S., Sentiment Analysis on Stock News using Historical Data and Machine Learning Algorithms, Term Paper, 2010
- [14] Mittermayer, M. & Knolmayer, G., Text mining systems for market response to news: A survey
- [15] Wüthrich, B., Permunetilleke, D., Leung, S., Cho, V., Zhang, J. & Lam, W., Daily prediction of major stock indices from textual www data, 1998