Decision Support Systems 52 (2011) 189–198

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A text-based decision support system for financial sequence prediction ☆

Samuel W.K. Chan [a,*], James Franklin [b]

[a] Department of Decision Sciences The Chinese University of Hong Kong Shatin, Hong Kong
[b] School of Mathematics & Statistics University of New South Wales Sydney, Australia

ARTICLE INFO

Article history:
Received 4 September 2009
Received in revised form 6 June 2011
Accepted 19 July 2011
Available online 23 July 2011

Keywords:
Textual analysis
Decision support systems
Classifier based machine learning

ABSTRACT

Although most quantitative financial data are analyzed using traditional statistical, artificial intelligence or data mining techniques, the abundance of online electronic financial news articles has opened up new possibilities for intelligent systems that can extract and organize relevant knowledge automatically in a usable format. Most information extraction systems require a hand-built dictionary of templates and thus need continual modification to accommodate new patterns that are observed in the text. In this research, we propose a novel text-based decision support system (DSS) that (i) extracts event sequences from shallow text patterns, and (ii) predicts the likelihood of the occurrence of events using a classifier-based inference engine. The prediction relies on two major, but complementary, feature sets: adjacent events and a set of information-theoretic functions. In contrast to other approaches, the proposed text-based DSS gives explanatory hypotheses about its predictions from a coalition of intimations learned from the inference engine, while preserving robustness and without indulging in formalism. We investigate more than 2000 financial reports with 28,000 sentences. Experiments show that the prediction accuracy of our model outperforms similar statistical models by 7% for the seen data while significantly improving the prediction accuracy for the unseen data. Further comparisons substantiate the experimental findings.

1. Introduction

There has long been a strong interest in applying computational intelligence to the analysis of financial data. Such analysis has traditionally concerned forecasting based on past price data. One area of limited success in financial prediction comes from textual data [39]. Textual data contain more information than numeric data because the former not only allow us to predict financial trends but also provide us with justification of the predictions. For example, a news article on a company containing words and phrases such as "shortfall", "risk of default", "resignation" gives reason to expect a fall in the company's stock price, even if the company's reported financial figures appear sound. The current availability of huge volumes of financial electronic text has created a pressing need for better knowledge discovery and the construction of applications for managing the knowledge that is extracted. Most of the existing research into financial text mining or knowledge discovery from text (KDT) relies on the identification of a predefined set of keywords. In this approach, a text is usually scanned for a specific type of event template, such as corporate acquisitions. The main goal is to fill in the values for sets of handcrafted and predefined template slots. Consequently, the construction of event extraction templates is a fairly laborious activity. It is difficult to design templates that anticipate all of the possible combinations of events or objects of interest that can be described, as well as to cover trivial redescriptions. Given the weakness of the approach and the demand for high-level representations, that is, not just keywords, to take advantage of linguistic knowledge, there has been considerable interest in the development of an automatic means of learning shallow event patterns from text, without indulging in linguistic formalism.

In this article, we propose a novel inference engine for financial text sequence prediction that brings together the benefits of shallow text processing and classifier-based inferences to produce effective knowledge discovery. Our approach aims to extract key underlying event sequences from financial texts and then hypothesize and assess incoming, even new and unseen, event sequences in the prediction. Unlike similar approaches, an inferential mechanism is developed in the engine that can extract event sequences from a collection of relevant texts, and collate the sequences in such a manner that both explicit and implicit information can be tailored to the needs of users. The task we are addressing and the problem of predicting the financial event sequence can be stated as follows. Given a corpus of financial documents that demonstrate event sequences, we explain how to extract all of the event sequences from the texts and predict the interesting and unseen relationships between them. The rest of this article is organized as follows. Section 2 provides an overview of the research that applies linguistic style information to enhance KDT. Section 3 gives an overview of our system and issues regarding text preprocessing, shallow parsing, textual information generalization,

☆ DSS ID: DSS#09-10-2749R(2).

* Corresponding author.
E-mail addresses: swkchan@cuhk.edu.hk (S.W.K. Chan), j.franklin@unsw.edu.au (J. Franklin).

0167-9236/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2011.07.003

and event sequence extraction. Section 4 describes the design of the inference engine for event sequence prediction. A practical boosting algorithm is introduced into the engine to produce a set of prediction rules. Numerous features that characterize the sequences and their latent inter-event relations are captured in the engine. The system prototype is implemented and we conduct a series of experiments to evaluate and compare the engine with the hidden Markov model. Section 5 provides an overview of our experimental design. We also quantify the outcome and give a detailed analysis of the results in our evaluation. Finally, the conclusions and further research directions are presented in Section 6.

2. Related work

Knowledge discovery from text (KDT) is not an isolated activity, as it is influenced by, and in turn influences, other decision support activities such as business intelligence, text classification, and information extraction. Schumaker and Chen [34] employ a predictive machine learning approach for financial news article analysis that uses several different textual representations. They find that their model containing both article terms and stock price performs best in stock price prediction. They also find that among the textual representations, proper nouns outperform noun phrases which in turn outperform the de facto standard, the bag of words (BOW). The success of noun phrases is attributed to their having less term noise. Coussement and van den Poel [8] introduce a screening mechanism to improve an e-mail complaint-handling strategy. Using a set of linguistic style information as a new type of textual information, the mechanism can differentiate customer complaint e-mails from non-complaint ones. The differences in the linguistic style between the two types of e-mail are also investigated. It is found that special linguistic features, such as the number of words or articles, presence of time words, or use of different verb tenses, are all indicative of a customer complaint. They also demonstrate that incorporating linguistic style features into a conventional e-mail classification model results in an increase in its predictive performance. Lavrenko et al. [23] devise a language model that can characterize some financial trends. Associating news stories with forthcoming trends, the model learns that words such as *loss*, *shortfall*, and *bankruptcy* are most likely to precede a downward trend in the stock price, whereas *merger*, *acquisition*, and *alliance* are likely to be followed by an upward trend. This model produces noticeably better predictions for all trend types than the popular vector-space model. Similarly, Fuller et al. [15] suggest a language model for deception detection in verbal communication. They incorporate different linguistic clues into their text-based decision support system (DSS) for deception detection tasks. Armed with more than 31 different linguistic clues that involve sentence length, content words, temporal ratio, lexical diversity, and verb and word quantities, their DSS prototype has an accuracy level approaching 74%, while professional lie catchers, such as police officers or customs officers, usually demonstrate the overall accuracy ranging from 49 to 64% [41]. Their experiments also show that an automated text-based DSS can be built and provide value-added results to decision makers. However, the search for a more parsimonious, but appropriate, set of linguistic clues is both vital and desirable for all of the tasks in KDT [1,21].

Recently, many researchers advocate document warehousing to capture complete business intelligence, in response to the hard fact that textual linguistic clues provide an extra dimension in decision making. Document warehouses, unlike traditional document management systems, include extensive semantic information about documents, cross-document feature relations, and document grouping or clustering to provide more accurate and efficient access to text-oriented business intelligence [40]. However, few systems have applied machine learning to the task. One of the earliest systems for the acquisition of extraction patterns is the AutoSlog [31]. AutoSlog is a knowledge acquisition tool that uses a training corpus to generate proposed extraction patterns. It learns extraction patterns in the form of domain-specific semantic case frames that contain a maximum of one slot. The creation of the semantic frames relies on the existence of a partial parser, small lexicon, and small set of general linguistic rules. Similarly, the PALKA system applies a conceptual hierarchy, which is a set of predefined keywords that can be used to trigger each pattern, and a semantic class lexicon [22]. To learn extraction patterns, PALKA looks for sentences that contain case frame slots using semantic class information. The conceptual hierarchy is used to control the generalization or specialization of the target case frame slots. Likewise, WHISK requires the syntactic preprocessing of the text and learns extraction patterns in the form of semantic case frames [38]. The triggers for the patterns in WHISK comprise a detailed specification of linguistic context that includes the subject, verb, or object of any of the surrounding constituents. A set of inductive learning techniques, which successively generalize input examples, are derived to learn extraction patterns and their relatively complicated triggering constraints. Given a preprocessed input, WHISK can also be extended to handle semi-structured text.

One of the customary steps in extracting information from texts is to use a knowledge base to support the text inferences. Harabagiu and Moldovan [17] address this issue by using WordNet [27] as a commonsense knowledge base and designing relation-driven inference mechanisms that look for common semantic paths from which to draw conclusions. Hearst [19] proposes a domain-independent method for the automatic discovery of semantic relations in unrestricted text collections by searching for corresponding syntactical patterns. Once the basic relations, such as hyponyms and hypernyms, are triggered in WordNet between the texts, common links are built, and the existing patterns in the text are then shared among the others. One of the main advantages of this method is its low cost and the simplicity of the relations. However, it relies heavily on a perfect ontology in which general and specific concepts are all included and the semantic relations are well defined. Although much progress has been made in KDT, many research issues still need to be resolved. First, although some of the abovementioned models take advantage of different linguistic clues in their DSS, most KDT tasks are forced to depend on impoverished text models such as BOW, when the decisions that they are making ought to be based on the meanings of those words in context. The lack of linguistic clues inevitably induces heavy noise and causes misinterpretation of meaning that results in failure to extract the knowledge [8]. Second, although existing approaches work well when a large amount of pre-annotated information is made available, it is debatable whether the same methods would work for domains with limited annotation, or for a user who is not experienced enough to tag the information. How does one effectively extract knowledge from a text based on minimum training? Last but certainly not least, most of the methods described above address the constituents to be extracted that appear explicitly in the text, without any attempt to invoke the prior knowledge with which they have been associated. For example, the `oil-price-surges` pattern is associated with the `stock-market-dives` pattern in financial news articles. The identification of these implicit contexts as a cause-consequence pair is important in every KDT system. In this article, we attempt to circumvent the handcrafted and massive annotation of financial text by using shallow parsing. We also demonstrate an inference engine that can extract knowledge from a collection of relevant financial texts, and collate the knowledge in such a manner that both explicit and implicit information can be tailored to the text-based DSS using a classifier-based inference engine.

3. System overview

In this section, we first present an overview of our system and the relevant methodology used throughout this study. Our text-based DSS includes four major components: text preprocessing, textual information generalization, event sequence extraction, and a classifier-based inference engine. The system architecture of the DSS is shown