Relational Learning for Securities Market Regulation

Position Paper for the IJCAI 2003 Workshop

Henry G. Goldberg*

NASD (National Association of Securities Dealers), 9513 Key West Avenue, Rockville, MD 20850
GoldberH@NASD.com

Introduction and Background

Over the past six years, NASD's Market Regulation department has built and operates two major "break detection" systems [Senator 2002] – the Advanced Detection System (ADS) and Securities Observation, News Analysis, and Regulation (SONAR) – for surveillance of the Nasdaq and several other markets. These systems rely for their effectiveness on the detection of instances of scenarios of regulatory interest – episodes in market activity where some violation may have occurred – many of which comprise relationships among transactions, market participants, securities, issuers, and other subject entities. I will discuss these systems in a bit more detail and then describe some of the kinds of scenarios for which statistical learning would be most beneficial.

Advanced Detection System (ADS)

ADS monitors trades, quotations, and orders in the Nasdaq, Over the Counter (OTC), and Nasdaq-Liffe (futures) stock markets to identify patterns and practices of behavior of potential regulatory interest. [Kirkland 1999] ADS has been in operational use at NASD since summer 1997 by several groups of analysts, processing roughly 25 million transactions per day, generating several thousand breaks per day. More important, it has greatly expanded surveillance coverage to new areas of the market and to many new types of behavior of regulatory concern. It's technology has been expanded to surveillance of the corporate and municipal bond markets and to NASD's new Alternative Display Facility. ADS combines detection and investigative components in a single system which supports multiple regulatory domains and which share the same market data. ADS makes use of a variety of AI techniques, including visualization, pattern recognition, and data mining, in support of the activities of regulatory analysis, alert and pattern detection, and knowledge discovery. ADS relies on a rule pattern matcher and a time-sequence pattern matcher. Data and market visualizations allow analysts to see the market context of breaks and temporal relationships of events in large amounts of data.

Temporal/Sequence Relationships in ADS

ADS relies heavily upon heuristic, manually coded patterns describing temporal sequences of market transactions. These patterns are input to a sequence matcher which finds instances of the patterns in databases of market transactions. The sequence matcher algorithm is similar to a regular expression matcher. It maintains a list of potential match states. At each step, a row is fetched and a new state is started for each pattern. Existing states are advanced if they match data constraints on the current transaction. When a state reaches the end of a pattern, it is a match. The sequence matcher may be in increasing or decreasing time order depending on whether the triggering event for the sequence occurs before or after the other necessary conditions. In a single pass, multiple tables may be scanned for several patterns concurrently. The sequence pattern language uses a syntax and precedence similar to the C programming language.

The sequence match has several problems. It is extremely brittle, in the sense that patterns and data constraints must be very carefully drawn not to inadvertently exclude a potentially valued match. A single failed match kills the entire chain. As a result, break detection errors are usually allowed to run heavily towards the false positives. Pattern discovery is limited to a semi-automated, iterative process in which patterns are carefully refined in an attempt to achieve the desired results and error rates.[Senator 2000] However, this refinement is, of necessity, haphazard and incomplete in its ability to model the variability in the data. Finally, there is a critical need to detect what market analysts call a "pattern and practice" – a set of similar or related matches from which one may infer intention violation of rules.

It is likely that statistical modeling can help to address all three problems. Models which produce a likelihood that an episode belongs to the modeled population are less brittle. Pattern refinement through statistical method would be more consistent and a comprehensive in dealing with data variability. And a model which describes a population of sequence episodes is a promising step towards defining "pattern and practice" detection.

*The author of this paper is an employee of NASD. The views expressed herein are those of the author and do not represent an official policy statement of NASD.

Securities Observation, News Analysis, and Regulation (SONAR)

SONAR was developed by NASD to monitor the Nasdaq, Over the Counter (OTC), and Nasdaq-Liffe (futures) stock markets for potential insider trading and fraud through misrepresentation. [Goldberg 2003] SONAR has been in operational use at NASD since December 2001, processing approximately 10,000 news wires stories and SEC filings, evaluating price/volume models for 25,000 securities, and generating 40-50 alerts (or "breaks") per day for review by several groups of regulatory analysts and investigators. SONAR makes use of several AI and statistical techniques, including NLP text mining, statistical regression, rule-based inference, uncertainty, and fuzzy matching. Sonar combines these enabling technologies in a system designed to deliver a steady stream of high-quality breaks to the analysts for further investigation. Additional components including visualization, text search agents, and flexible displays add to the system's utility.

Entities, Relationships, and Events

SONAR mines news wire stories and SEC filings for entities such as companies which issue securities, company officers, brokers, the securities themselves, regulatory bodies such as the FDA which have an impact on stock values, and others. It also finds material events: product announcements, earnings reports, mergers and acquisitions, etc. Finally, SONAR mines for relationships both explicit and implicit among the entities and events. The results of the text mining stage are contained in the top-level predicates output by a linguistic rule-base used by SONAR NLP component (from ClearForest). These entities, relationships, and events form particular episodes or scenarios, with specific identifiers and values which may be incompletely mined. Learning statistical models of these episodes would improve detection, especiallyin dealing with stories where the "components" of a scenarios are not all present.

News Stream Segmentation

Insider Trading is defined as trading upon inside information of a "material" nature – information which a reasonable investor would take as a reason to buy or sell a security. Thus, two crucial events in an insider trading break are the appearance of material news and a movement in the market in response to it. It is critical that SONAR is able, therefore, to determine when a news item is material, but also when it is first made public. The drawing of relationships among entities and events mined from several news stories is currently performed by a fairly simple template match. But, clearly, news is re-written, expended upon, and interpreted. Any failure of this match will "create" a new trigger for an insider trading break. Membership in the same model, drawn from a broad population of multiple story events, seems to be better way to detect truly "new" news.

Misrepresentation Fraud

Fraud by misrepresentation is another critical target activity of SONAR. While we currently mine for several dozen "flags", likely indicators of stocks which are being falsely touted, much more could be done with the ability to draw comparisons across stories and sources (e.g. compare an announcement of $50M dollars in contracts with an SEC filing indicating the company has a staff of 2 with no assets.) Linking such evidence across text sources and learning statistical models of misrepresentation seems to be a promising approach.

Break Detection and Fraud

Break Detection Systems are powerful tools for detecting errors, violations, or other anomalous conditions and activities. [Senator 2002] However, they are limited to the immediate activities which they find in the input data stream. Background knowledge, aggregation of detection over a priori identifiers (brokers, issuers, etc.) can start to draw a picture of an underlying intentional pattern and practice. Without powerful but tractable models of populations of breaks, we are limited to counts and percentages as a decision tool for investigators. As target activities become more complex and varied, and as the cost of regulation continues to rise, NASD feels increased need for such models to cull and derive the greater benefit from its break detection systems.

References

Goldberg, Henry G, Kirkland, James D., Lee, Dennis, Shyr, Ping, and Thakker, Dipak, "The NASD Securities Observation, News Analysis & Regulation System (SONAR)," *presented at IAAI-2003, Acapulco, Mexico.*

Kirkland, James D., Senator, Ted E., Hayden, James J., Dybala, Tomasz, Goldberg, Henry G., and Shyr, Ping, "The NASD Regulation Advanced Detection System (ADS)," *AI Magazine* 20(1):55-67, 1999.

Senator, Ted E., "Ongoing Management and Application of Discovered Knowledge in a Large Regulatory Organization: A Case Study of the Use and Impact of NASD Regulation's Advanced Detection System (ADS)," in *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-00)*, pp. 44-53, ACM, 2000.

Senator, Ted E. and Goldberg, Henry G., "Break Detection Systems." in W. Kloesgen and J. Zytkow (eds.), *Handbook of Knowledge Discovery and Data Mining*, pp. 863-873, Oxford University Press, 2002.