Early Detection of Insider Trading in Option Markets

Steve Donoho
Donoho Analytics, Inc.
4301 Warner Ln
Chantilly, VA 20151
steve@donohoanalytics.com

ABSTRACT

"Inside information" comes in many forms: knowledge of a corporate takeover, a terrorist attack, unexpectedly poor earnings, the FDA's acceptance of a new drug, etc. Anyone who knows some piece of soon-to-break news possesses inside information. Historically, insider trading has been detected after the news is public, but this is often too late: fraud has been perpetrated, innocent investors have been disadvantaged, or terrorist acts have been carried out.

This paper explores early detection of insider trading – detection before the news breaks. Data mining holds great promise for this emerging application, but the problem also poses significant challenges. We present the specific problem of insider trading in option markets, compare decision tree, logistic regression, and neural net results to results from an expert model, and discuss insights that knowledge discovery techniques shed upon this problem.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data mining*; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Design, Experimentation.

Keywords

Insider trading, fraud detection, behavior detection, data mining.

1. INTRODUCTION

Financial markets allow people to profit from knowledge they have without giving that knowledge away or even letting others know they have it. This fact makes financial markets tempting vehicles for illicitly profiting from "inside information" – information that affects a company's value but is not yet available to the public.

Examples of inside information include advance knowledge of:

- A company being up/downgraded in a research report.
- Unexpectedly good or poor earnings
- Revisions to earnings projections
- Settlement of a lawsuit
- FDA accepting or rejecting a new drug
- SEC announcing an investigation of a company's accounting practices
- USDA announcing first U.S. case of Mad Cow disease
- Terrorist attacks affecting a specific industry/company
- Mergers and acquisitions

A person doesn't have to be a corporate executive to have inside information. Family members, in-laws, and friends often hear corporate secrets. Psychologists have heard of pending mergers from their patients. Janitors have taken sensitive material from trash cans. Lawyers have overheard important news from colleagues. Terrorists have had advance knowledge of attacks.

When people trade on inside information, they leave faint traces of this in trading data. Trading volumes are often unusually high. Price movements are unusual. Trading in options is unusually distributed among various contract types. This is especially true in "thinly traded" securities (ones that usually have low trading volumes) because a small amount of insider trading has a larger impact on the overall trading. This is as opposed to heavily traded stocks where signals of insider trading are likely to be drowned out by the noise of normal trading. Our work focuses on option trading because options are thinly traded compared to stocks.

Most systems that find insider trading look for it after the news is public. The news is the "trigger" that begins the search for unusual trading that preceded the news. The SONAR system at NASD [3] is an excellent example of such a system. This approach works very well for most regulatory purposes where the goal is to prosecute those who have committed insider trading. For other goals such as preventing terrorist attacks or preventing being taken advantage of by a trader with inside information, a method is needed that detects problems before the news breaks.

This paper explores early detection of insider trading – detection before the news breaks. There are several uses of a system that could provide leads of where insider trading is occurring. Such leads could be used in conjunction with other intelligence to give an early warning of planned terrorist attacks. Brokerage firms could take steps to protect themselves against a sudden swing in a security's price. Corporations could know if sensitive information is being leaked and avoid brand name tarnishing that is associated with insider trading. Even if the exact timing or nature of news

KDD'04, August 22–25, 2004, Seattle, Washington, USA.

cannot be known with high certainty, leads can keep people from being caught totally off guard.

Data mining technology is a good fit for this problem because it is able to automatically pick up faint signals from noisy data. It is able to discover correlations that experts may not have been aware of. Data mining has also proven useful at finding trends in human behavior, and markets are the synthesis of the behaviors of many individuals (some with inside information, most without).

Although early detection has several important uses, and data mining is a good fit for the problem, there are several challenges to solving the problem:

- The important information about insider trading is very spread out. Between 20 and 120 different types of option contracts trade on any given day for every underlying company, and the relevant information is spread over these. The important information is also often spread across multiple trading days.
- In addition to structured trading data, much important information is also buried in unstructured text. It is crucial to weed out trading anomalies that are a reaction to news as opposed to those that are predictors of news. This involves determining if news is "market moving" news or just an unimportant press release or tangential news.
- There are relational interactions among companies. News about one company may cause abnormal trading in another. An earnings announcement by Microsoft may cause trading anomalies in Intel options. The acquisition of one bank may cause abnormal trading for other banks.

The problem affords opportunities for combinations of feature construction, text mining, and learning from relational data.

The rest of the paper is organized as follows. Section 2 describes options and option markets explaining why traders with inside information often use options. Section 3 describes the symptoms of insider trading in options and also gives two case studies. Section 4 presents the original data and describes steps taken to transform data from the option "series level" to the "company level" which was more conducive to learning. Section 5 describes experiments comparing a model manually built from expert knowledge to learned models. Also described are experiments with unsupervised algorithms to extract unknown relationships. Sections 6 and 7 discuss related work and future work.

2. OPTIONS

A "call" is an option to *buy* a specific stock at a specific price on or before a specific date. For example, an investor may buy an option to buy Microsoft stock (MSFT) at $30 on or before March 20, 2004. The stock is called the "underlying stock" (MSFT). The price is called the "strike price" ($30). The date is called the "expiration" date (March 20, 2004) because the investor's option to buy expires on that date. If the market price of MSFT is above $30 on March 20, then the owner of the option will choose to "exercise" their option, buy the stock at $30, and possibly turn around and immediately sell the stock at the higher market price. When the price of MSFT is above the strike price, the option is said to be "in the money." When the price of MSFT is below the strike price, the option is "out of the money." Options that are in-the-money on the expiration date are exercised. Options that are out-of-the-money on the expiration date simply expire worthless and are not exercised. Because it is an *option*, the owner is not obligated to buy the stock at the exercise price. An investor would buy a call when they expect the underlying stock to go up.

A "put" is an option to *sell* a specific stock at a specific price on or before a specific date. So a put is complementary to a call. A put is in-the-money when the price of the underlying stock is *below* the exercise price. An investor would buy a put when they expect the underlying stock to go down.

For any given underlying stock there will be many different options trading on any given day. There will be both puts and calls. For both there will be multiple different strike prices, some in the money and some out of the money. For all of these, there will be multiple expiration dates – some less than a month away, some up to 2 years away. Options with the same strike price, same expiration date, and same type or "class" (call or put) are said to be in the same "series." An underlying stock often has between 20 and 120 series of options trading on a given day.

Options amplify an investor's gains (and losses) and this makes them appealing to people with inside information. For example, consider someone who knows that XYZ Corp is going to be acquired in the next couple of weeks, and their stock price is likely to jump from $50 to $55. If they simply buy XYZ stock at $50, they make a 10% profit when it jumps to $55. But if they buy a call (with a strike price of $50) for $1, that call will be worth at least $5 after the jump. The trader has made a 400% profit instead of a 10% profit. The risk is that the acquisition doesn't happen as soon as expected, and the stock stays at $50. In this case the option expires worthless, and the option trader has a 100% loss as opposed to the stock trader who breaks even. But if the person is fairly confident of their information, there is much to be gained from buying calls.

There are several factors that determine the price of an option. The most obvious is how far in or out of the money the option is. A call that is $2 in the money is going to be worth more than a similar call (same underlying, same expiration) that is $3 out of the money. A second factor is the time remaining until expiration. The more time until expiration, the more an option will cost because there is more time for the price of the underlying to move around. When a person buys a call, they are paying to be able to keep their options open on whether or not they will buy at a specific price. The longer they are allowed to keep their options open, the more they will have to pay. People with inside information are likely to favor options with near-term expiration dates because the news they know about is likely to happen soon, and near-term options maximize their percent return because they are less expensive.

A third factor is the "volatility" of the underlying stock. Volatility has to do with how much the stock price moves around and how rapidly it moves. Options on a volatile stock will cost more than on a non-volatile stock because it is more likely the underlying price will end up in the money. Other less important factors include the current interest rates and whether or not the stock will issue a dividend before the exercise date. The exact relationship among these factors is described by the Black-Scholes equations

[4], and common practice is to plug them into a Black-Scholes calculator which returns a price.

Of these factors, most can be measured exactly with the exception of volatility because it is the future volatility that matters – the volatility between now and the expiration date. In lieu of knowing future volatility, recent historical volatility is often used as an estimate.

Because volatility is the only factor that cannot be measured exactly, it often becomes a catch-all input for other factors that cannot be explicitly entered into Black-Scholes. For example, supply and demand for options affect price, uncertainty about a pending court case affects price, etc. For this reason, an important concept is an option's "implied volatility." That is the volatility that is implied by the price. This is calculated by taking the option's real, market price and reverse engineering the volatility that would have been input to get that price. "Implied volatility" measures how "pricy" an option is and can be used to compare options with different underlying stocks, strike prices, and expiration dates.

3. SYMPTOMS OF INSIDER TRADING

When inside information is leaked, rumors start to swirl, and people trade on those rumors. This trading is known to manifest itself in certain ways in trading data. The first half of this section presents several "symptoms" of insider trading that one can observe in option trading data. Most of these are taken from McMillian [5] which presents a mostly-manual method of looking for insider trading. The second half of the section presents two case studies of abnormal trading that preceded material news ("material news" being news that greatly affected a company's stock price).

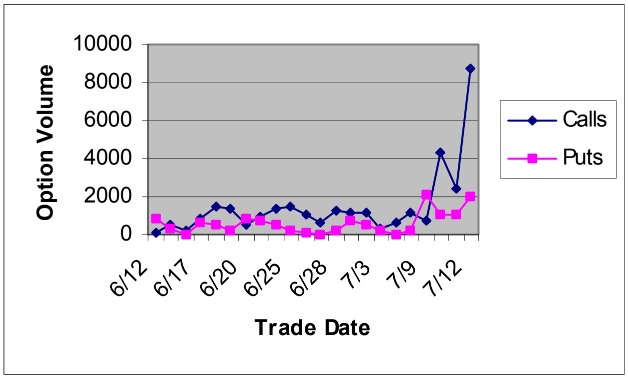

Figure 1. Example rise in option trading volume

The primary symptom is a jump in the volume of options traded. If there is a jump in the volume of calls traded, this would indicate good news (anticipating stock price increase); puts would indicate bad news (anticipating decrease). The volume of options with near-term expiration dates (within the next couple of months) is particularly important because inside knowledge is usually about events that will occur in the next couple of months, and near-term options give a higher percent return. Figure 1 shows an example rise in call volume. Historically, call volume has been around 1000 contracts per day (a contract is an option for 100 shares). Suddenly around July 12, the volume of calls rises to 4000 and then to above 8000.

A related symptom is that there is an imbalance in the volume of calls to the volume of puts. This is because inside information is usually directional – the stock price will rise or fall. A lot more calls than puts would indicate good news; more puts than calls would indicate bad news. If the volume of both rise together, this does not indicate a direction. It may mean there is big news coming, but there is disagreement among traders as to whether it will be good news or bad news. This often happens before scheduled earnings releases – traders know there will be news, but some think it will be good and others think it will be bad. In the example in Figure 1, the volume of puts increases some, but there are still four times as many calls traded.

Some stocks may trade a high volume of calls one day and then trade a high volume of puts a few days later and then calls again the next day. This lack of direction would indicate that the volume is probably not driven by people with inside knowledge.

Another symptom is that implied volatility often increases. Recall that implied volatility is a measure of how "pricy" an option is. If insiders are buying a lot of options, demand will outpace supply resulting in an increase in the relative expensiveness of the options. This will be reflected in a rising implied volatility. Implied volatility can also be an indicator of uncertainty. Abnormal trading makes traders wonder what is up and increases their uncertainty of the stock's future movements.

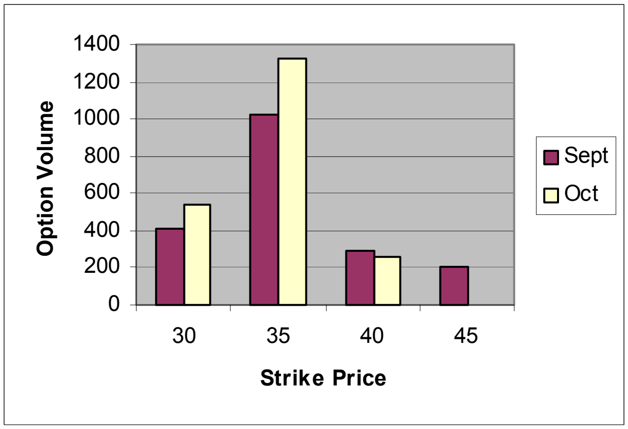

Figure 2. Volume distributed over prices and expirations

There are other reasons than inside information that cause option volume to increase. A mutual fund may have a large position in XYZ stock and sell calls because they think XYZ is near the top of its trading range. This is called writing covered calls. A trader may create a "spread" by buying a lot of calls at one strike price and selling an equal number of calls at a different strike price. These false positives can sometimes be weeded out by looking at the distribution of volume among the various series (strike prices and expiration dates). Inside information often leads to volume distributed over multiple strike prices and expiration dates as in Figure 2. This is because some inside traders will prefer slightly in-the-money options while some will prefer out-of-the-money for higher returns. Some will prefer options with more time until expiration while some will prefer shorter expirations for higher returns.

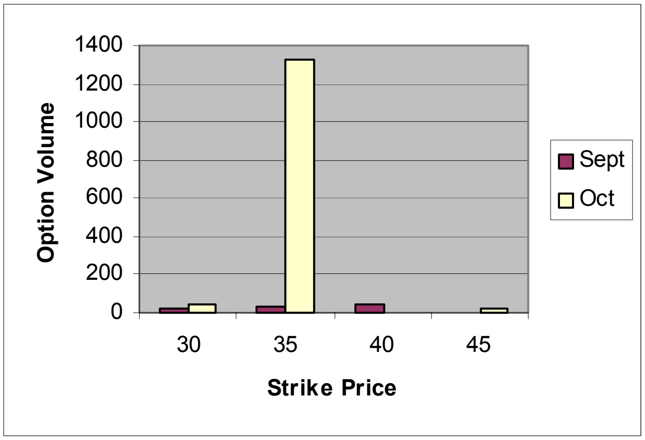

Figure 3. Volume concentrated in one price and expiration

Figure 3 shows a volume distribution pattern more indicative of covered call writing. Real volume distributions are rarely this clean and crisp, but odd volume distributions with gaps, round number of contracts, or near-identical peaks can sometimes indicate other reasons for option volume.

Some symptoms may also be evident in the market for the underlying stock. There may be increased volume, but the increases may be small compared to normal trading volume. There may also be an abnormal increase or decrease in the stock price. If there are rumors about bad news, enough people may sell or short the stock to drive the price down some.

The following two subsections give case studies of two companies that released unscheduled news, and the news was preceded by abnormal trading patterns. The first case study for EDS involves high put volume followed by bad earnings news and a price drop. The second case study for Pharmacia involves high call volume followed by an acquisition announcement and a price jump.

3.1 Case Study #1

On Sept 18, 2002 EDS stock closed at $36.46. After the market closed, EDS announced that it "expects its Q3 revenues and earnings to be lower than previous company guidance." On the following morning, EDS opened at $21.90 and fell to $17.20 by the end of the day (down 53%). A person who bought puts with a strike price of $35 expiring in Sept on Sept 18 for $0.40 could have sold them the 19th for $17.70, a 43-fold increase.

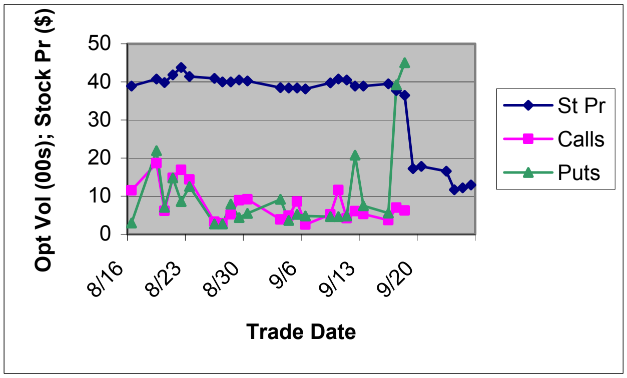

Figure 4. EDS option volume and stock price

Figure 4 plots the call and put volume in the days leading up to the news. Call and put volume were generally below 1000 contracts per day with a few exceptions in mid-August. On both Sept 17 and 18, there was unusually high put volume (over 4000 contracts) and even on Sept 12 put volume was at 2000. Meanwhile call volume remained relatively normal. Option volume after Sept 18 is not shown because the news caused volume to jump off the scale. Also shown in Figure 4 is the stock price so the drop between Sept 18 and 19 can be seen.

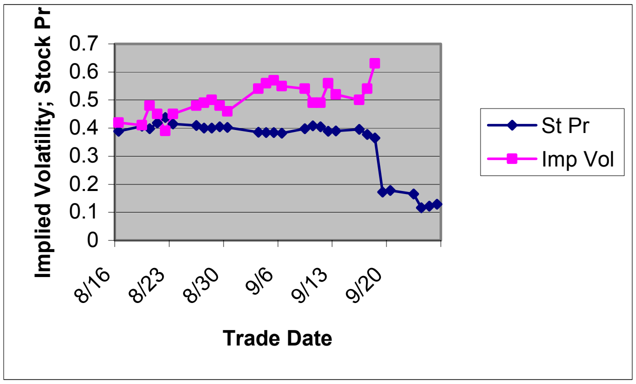

Figure 5. EDS implied volatility

Figure 5 shows EDS's implied volatility rising in general in the weeks leading up to the news and especially rising on Sept 17 and 18. Again, implied volatility after Sept 18 is not shown because the news caused it to jump off the scale. Stock price (divided by 100) is shown for comparison.

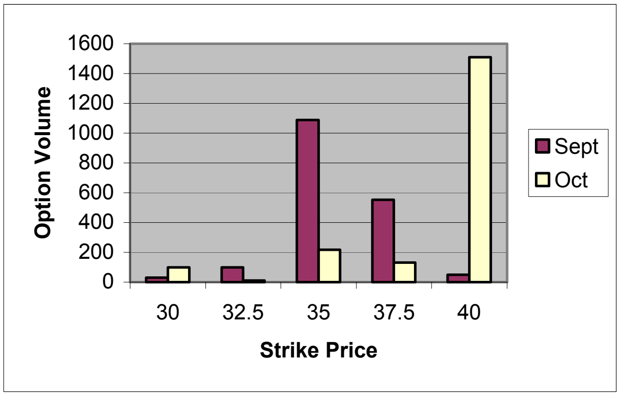

Figure 6. EDS option volume distribution on Sept 17

Figures 6 and 7 show the distribution of EDS option volume over various strike prices and expiration dates on Sept 17 and 18. On the 17th the stock price was around $38, and we see significant volume on the three strike prices closest to this. On the 18th the stock price dropped closer to $36, and there is still significant volume in the three strike prices around that. Volume on the 18th was mostly in October puts.

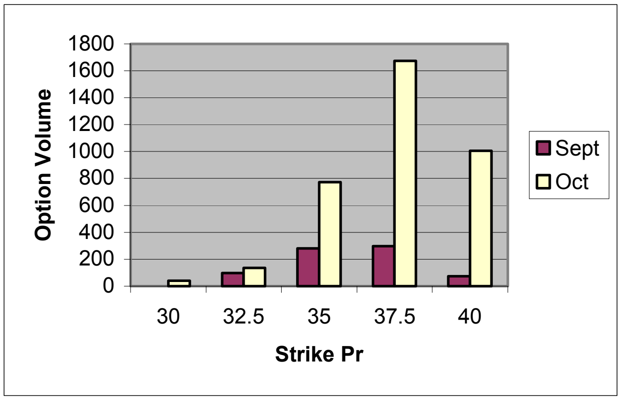

Figure 7. EDS option volume distribution on Sept 18

Regarding stock price and volume, the price of EDS stock dropped 4.5% on the 17th and 3.4% on the 18th indicating selling pressure on those two days. Additionally, the stock volume was about 20% above normal on the 17th and 18th. In summary, in the days before a large drop in EDS stock price due to the release of negative news, the following trading patterns were observed: put volume jumped over four-fold while call volume remained fairly normal, implied volatility rose, put volume was distributed over several series, and stock volume was high while stock prices fell.

3.2 Case Study #2

On Friday, July 12, 2002 Pharmacia stock (PHA) closed at $32.59. Over the weekend, there was an announcement that Pfizer was acquiring Pharmacia. On Monday, July 15, PHA opened sharply higher and closed at $39.25, a rise of over 20%. A person who bought July/$35 calls on July 12 for $0.55 could have sold them the 15th for $4.10, a 650% increase.

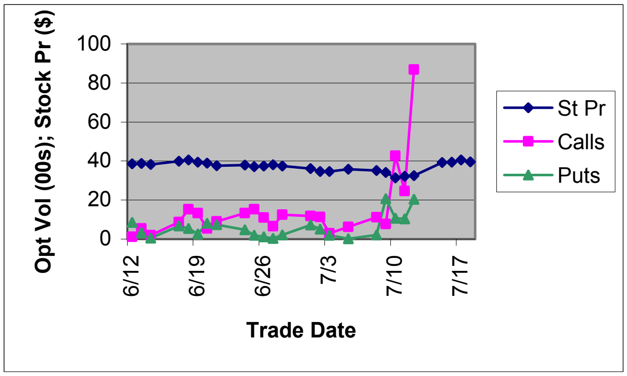

Figure 8. PHA option volume and stock price

Figure 8 plots the call and put volume in the days leading up to the acquisition. Call volume was generally below 1500 contracts per day, and put volume below 1000. On July 10th, 11th, and 12th near-term call volume rose to 4266, 2463, and 8685 respectively. Put volume rose some but stayed at or below 2000 contracts per day. Option volume after July 12th is not shown because the acquisition caused volume to jump off the scale. Figure 8 also shows the stock price. While the increase in price over the weekend does not look dramatic in the graph, it is a 20% increase.

Figure 9 shows PHA's implied volatility which jumped from about 0.4 to 0.6 three days before the acquisition and stayed above 0.6. The implied volatility had briefly jumped to 0.6 once in the previous month, but generally it had been around 0.4. Stock price (divided by 100) is shown for comparison.

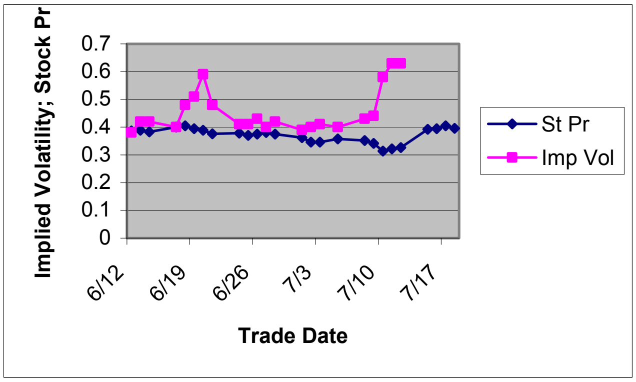

Figure 9. PHA implied volatility

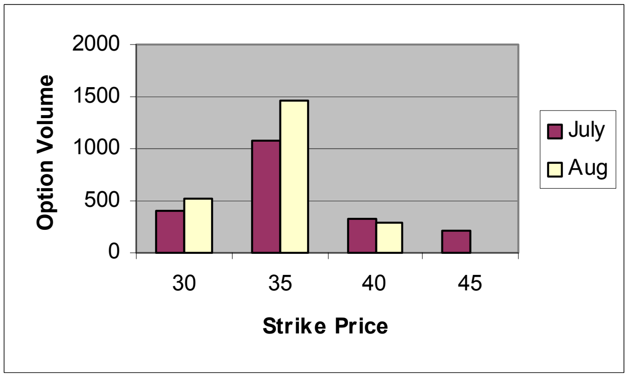

Figure 10. PHA option volume distribution on July 10

Figure 10 shows the distribution of PHA option volume over various strike prices and expiration dates on July 10. On the 10th the stock price closed close to $31, and we see significant volume on either side of this price and also up to $40 and $45 for both July and August expirations. The distribution on the 11th and 12th were similar to the 10th.

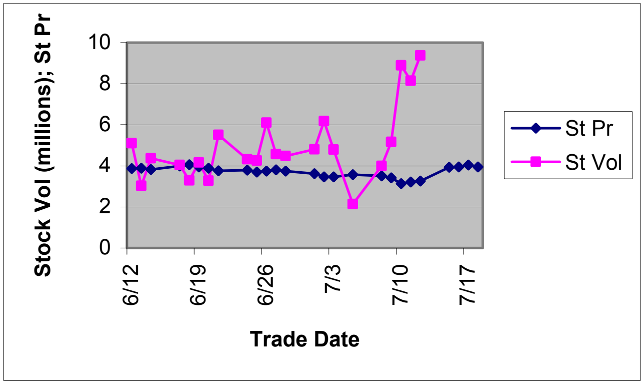

Figure 11. PHA stock volume

Regarding stock price and volume, Figure 11 shows that volume of the PHA stock rose to double the normal volume in the three days preceding the acquisition. Stock price dropped July 10 but rose modestly both the 11th and 12th.

In summary, in the days leading up to a Pharmacia acquisition announcement and a 20% rise in the stock price, the following trading patterns were observed: call volume jumped over five-fold while put volume rose only modestly, implied volatility rose, call volume was distributed over several series, and stock volume doubled while stock prices rose modestly.

4. DATA SOURCES & PREPARATION

One of the primary challenges of applying data mining techniques to early detection of insider trading is that the important information is very spread out. It is spread over multiple trading days. On any given trading day, the information is spread over multiple different option series (call/put, strike price, expiration date). It is often spread over multiple news articles.

The raw data is about individual series on individual trading dates. But people do not have inside information about a series, they have information about an underlying company. And inside information is not confined to single dates; rather it starts small on one date and spreads as time passes. Much of the problem-solving work goes into transforming the problem from the "series level" to the "company level" and from looking at individual trade dates to looking at trade dates in the context of their recent history. The first subsection explains the raw data and how much of it is at the "series level." The second subsection explains the transformation to the "company level" and how important concepts such as "near-term put/call imbalance" and "'good' distribution among strike prices and expiration dates" are represented. The third subsection addresses how information from text was integrated with trade data, and the fourth subsection tells how a labeled dataset was created.

4.1 Raw Data

Raw data came from three sources: option trading, stock trading, and news. Stock and option data were available on all U.S. companies for which options are trades (about 2160 companies). News covered these companies plus others. The date range for which all three data sources were available covered a six-month time period from March 11, 2003 to Sept 17, 2003.

Option data consisted of a daily summary for each option series. So there is an entry each day for each stock's combination of call/put, strike price, and expiration date. There were between 110,000 and 130,000 rows of data per day or about 50 to 60 series per stock on average. The available data fields were: option symbol, underlying stock symbol, trade date, class (call or put), strike price, year and month of expiration, volume (number of contracts traded), closing bid price (price at which buyers will buy), closing ask price (price at which sellers will sell), and open interest. Stock data consisted of daily summaries for each company. The available fields were: stock symbol, company name, trade date, opening price, high price, low price, closing price, and volume (number of shares traded).

News was available from several sources. Press releases came from Business Wire and PR Newswire. News stories came from Dow Jones, Reuters, and the Associated Press. Each news story consisted of a title and body in ASCII and a list of stock symbols mentioned in the story. There were often multiple stock symbols per article. 20% of press releases contained multiple stock symbols. 68% of news stories contained multiple stock symbols.

4.2 "Series Level" to "Company Level"

Our goal was to create "mid-level" features that could be used for discovery by data mining algorithms. This is as opposed to "high-level" features that completely solve the problem and "low-level" features such as were in the raw data. The desired features should measure some aspect of how a company's trading deviates from its own normal behavior [2]. Not every anomaly is interesting, though. When data mining algorithms are applied to these, the goal is to find trends in the anomalies. 27 features were constructed from the equity and option data that measured the five high-level symptoms discussed in Section 3.

4.2.1 High Option Volume

High volume indicates interest in a particular company, and 12 features were included to measure this concept. These were actually six features calculated once for calls and once for puts. The call measurements are kept separate from the puts to preserve directionality. The first was near-term call (put) volume. This was calculated simply by summing across all the call (put) series for each company on each trading date. This gives an absolute measurement of interest in the company. If only 50 calls trade, this may not be interesting even if it is abnormally high for that company. It shows whether the trading was in round amounts (e.g. exactly 500 contracts may indicate one large trade) or in non-round (e.g. 537 contracts may indicate multiple smaller trades).

Two features that measure relative volume were how many times average the volume was and how many standard deviations above average the volume was (z-Score). The average and standard deviation of near-term call (put) volume were calculated over the previous 20 trading days (because there are approximately 20 trading days in a month). These measure whether trading volume is high relative to recent history.

A separate way of measuring interest in a company is to look at the change in open interest. "Open interest" is the number of contracts that are outstanding for a particular series. If traders are opening new positions, then a high volume will result in a rise in open interest. But traders may simply be closing old positions and opening new ones. While this might result in a high volume, it does not necessarily indicate an abnormally high interest in the company. One feature was percent change in call (put) open interest from the previous day. A second feature divided today's percent change in open interest by the average change over the last 20 trading days. The last pair of features measured the number of days in the last five days that call (put) volume had exceeded two times the twenty day moving average.

4.2.2 Call-Put Imbalance

An imbalance between the volume of puts and calls indicates directionality. More calls than puts indicates optimism. More puts than calls indicates pessimism. One feature is simply the ratio of today's near-term call volume to put volume. On high volume days, this is an indicator of directionality and how strong.

A second feature compares call volume to recent high put volume. Instead of comparing to the average put volume over the last 20 days, today's call volume is compared to the average put volume of the three days out of the last 20 that had the highest put volume. This is because it is assumed that this feature will only be of interest if there is a high call volume today. Therefore it makes sense to compare today's call volume to days with high put volume. Some companies have high put volume one day and high call volume the next, and that does not necessarily indicate interest with directionality. This feature allows a measurement of that back-and-forth behavior. A similar feature compares today's put volume to recent high call volume.

4.2.3 Implied Volatility

Implied volatility is a measure of uncertainty about a company and how "pricy" their options. The implied volatility is first calculated for each series using the midpoint between the bid and ask prices and reverse engineering Black-Scholes. The near-term implied volatility is found by averaging implied volatility from series expiring in the next two months. Only series whose strike prices are close to the current equity price (closest 4 strike prices, two on either side) are used. The implied volatilities of series that are too far in-the-money or out-of-the-money have a lot of noise and can distort the measurement. To get a measure of whether implied volatility is higher or lower than usual, a second feature divides the current implied volatility by the 20-day average.

4.2.4 Distribution

Distribution is more difficult to quantify. In general a more spread out distribution is better than a concentrated one. But there are other factors such as gaps and whether the distribution is at strike prices close to the stock price. Another factor is how close the next expiration date is. If it is a couple weeks away, distribution over two expiration dates is expected. But few traders will open a position on an expiration date that is only a couple days away. This is an area where future work in automated feature construction would likely have a good payoff. For these experiments, the three features user were:

- Percent of near-term call volume in the top series (series with the highest volume).
- Percent of near-term call volume in the top two series.
- Percent of near-term call volume in the top three series.

If the percent in the top series is 95%, this indicates all the call volume is concentrated in on series. If the percent in the top three series is 75%, this indicates that volume is spread over several series. Three similar features measure put distribution.

4.2.5 Stock Behavior

Inside information may also be reflected to a lesser degree in stock trading. It is worthwhile to check stock behavior to see if it confirms or contradicts what is seen in option behavior. One feature compares stock volume to the 20 day moving average. A second feature measures the percent change in the stock price over the last two days. A moderate rise in stock price can indicate buying interest, but a large increase may mean the option volume is probably a reaction to news, not a predictor of it.

4.2.6 Features from Expert Model

Two features from an expert model described in Appendix A were the one-day score from the model and the sum of the scores from the model over the last five days. Interestingly, these features were not used very much by the learning algorithms.

4.3 Integrating Text

An additional 8 features were extracted from business news. The news analysis was used primarily to tell if trading anomalies were reactions to news or predictors of news. The biggest challenge to this is separating "material" news (important news) from the steady stream of unimportant news. To do this we relied on two basic principles:

- Material news often falls into a few categories such as earnings announcements, research reports, mergers and acquisitions, regulatory approvals or denials, and product announcements [3]. Articles of these types can often be identified by commonly used words or phrases.
- When there really is material news, it is usually reported in multiple different sources creating a spike in news volume as compared with recent history.

We used a two phase approach to quantify these characteristics. The first phase extracted features from each news article and flattened them onto a feature vector. The second phase created a daily news profile for each company. The two phases are discussed in greater detail below.

By finding common words and phrases, each article was assigned multiple scores along several dimensions such as:

- Was this good news, bad news, or neutral?
- Was this about an earnings release or prerelease?
- Was this about an upgrade or downgrade?
- Was this about a lawsuit or scandal?
- Was this about a merger or acquisition?
- Was this about FDA approval of a drug?

While this was done in a fairly crude manner (a PERL script that counts instances of key words and phrases), it proved to be fairly good at categorizing news because the same words and phrases are used over and over again in business news.

These were then rolled up into a company news profile that had a feature vector for each company for each day. If there was no news for a company on a given day, there was no vector. Three features measured the significance of the news on that day:

- Max good news score. This returned the highest "good news" score for that company that day.
- Max bad news score. This returned the highest "bad news" score for that company that day.
- Max significant news score. This summed all the "material news" scores for each article and took the highest score for that company that day.

Another five features measured the volume of news and compared it to the previous 30 days:

- One day count. Number of articles mentioning company today.
- Three day count. Number of articles mentioning company in last three days.
- Thirty day count. Number of articles mentioning company in last thirty days.
- One/Thirty ratio. One day count / Thirty day count
- Three/Thirty ratio. Three day count / Thirty day count

If there was unusual trading on a given day, the learning algorithms could use these eight features to determine if it was simply a reaction to news.

4.4 Creating Labels

A data element represented a specific company on a specific date. An element was labeled positive if within the next two weeks after the date, there was significant news that led to a sizable increase in stock price. (The experiments in this paper focus only on the task of predicting positive news and upward movements in a stock. The flip side of the problem – predicting negative news – is similar but not entirely symmetrical and thus brings out a different set of issues. In practice puts and calls are traded for different reasons. Calls are often traded by speculators while puts are often traded to hedge downward risk.)

A company was considered to have significant news if any of the following three conditions held 1) the number of articles on a company was greater on one day than 25% of the sum of the previous 30 days or 2) the company had an article with a score greater than zero in one of the "significant news" categories or 3) the company had an article with a "good news" score >= 5. A company was considered to have a sizable stock price increase if *all* of the following three conditions held: 1) the stock price increased by more than 5% in one day and 2) the company was in the top 2% of price gainers and 3) the price increased by more than $.50. These factors taken together were found to be pretty good indicators of a public announcement of market-moving news that someone could have had advance knowledge of.

While the date range of the dataset was from March 11 to Sept 17, the range of the labeled data was from April 21 to Sept 10. Thirty days were left off at the beginning of the range because the news profile needed a thirty-day history to be accurate. One week was left off the end because this gave a week after Sept 10 for a significant news event to occur so that the last elements could be labeled positive. This left 100 trade days in the dataset.

To filter out trading that was not unusual at all, data elements were excluded if the company's call volume on that day was not at least two times the 20-day moving average. The choice of two times was taken as an expert suggestion from [5] but could be parameterized in future experiments. This filtered the original 216,294 "company days" (approximately 2160 companies on 100 days) down to 5,848 data elements (approximately 2.7% of the original "company days"). Of these 5,848, there were 458 labeled positive (7.8%) and 5,390 labeled negative (92.2%).

A jump in a stock's price does not always involve insider trading. Even taken together with high volumes and other trading irregularities, it certainly looks suspicious, but it does not *prove* insider trading. The instances in this dataset labeled positive are not proven cases of insider trading but instead should be viewed as "suspicious" or "worth reviewing."

5. EXPERIMENTS

The experiments can be described as follows: Given that call volume was high on a given trade date (double the 20 day moving average), predict whether the stock will jump on material news within the next two weeks. This amounts to predicting the "label" field described in the previous section. The data was divided into a training set and a test set. Training data came from the dates April 21 to July 31 (72 trade days), and test data came from Aug 1 to Sept 10 (28 trade days). This is a realistic approximation of a real world implementation of the task where a model would be trained on a recent period and would be used to predict the subsequent time period.

Three algorithms were evaluated: C4.5 [6], backwards stepwise logistic regression, and neural networks. All models were built using Clementine 8.0. For comparison, a manually-built expert model was also tested. The model was adapted from expert indicators described in [5] and also described in Section 3 of this paper. At a high level, the model is a weighted combination of evidence where evidence for insider trading is added to a score, and evidence against is subtracted from the score. The model is described in greater detail in Appendix A.

C4.5 was run with default parameters (pruning severity of 75% and a minimum of 2 records per child branch). Misclassification costs were set at 10 for false negatives and 1 for false positives to compensate for the skewed distribution of positives and negatives. The backwards stepwise logistic regression algorithm starts with a model containing all features and iteratively removes features that don't add value (also features that were previously removed may be added back in). The default Clementine parameters were used. Neural networks were used with the prune method where the initial network has a large number of hidden layer units, and the weakest units are iteratively removed. Again, the default Clementine parameters were used.

Figures 12, 13, and 14 focus on three different portions of the same lift curve. Figure 12 gives the whole lift curve from 0% to 100% in order to show performance over the full range of data. Of the three algorithms, none clearly outperformed the other two over the entire range. If fact, their performance was roughly the same. Logistic regression outperformed in the 20% to 50% range. The neural net outperformed in the 50% to 80% range. Decision trees performed the best in the 0% to 10% range, and this will be examined further below. The expert model performed only slightly better than random for the full range but better in the lower ranges as examined below.

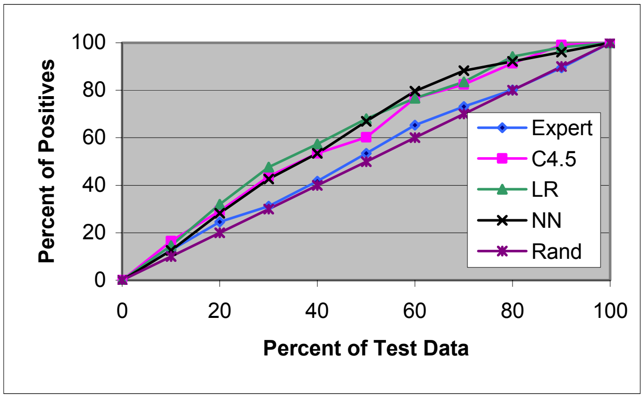

Figure 12. Lift curve 0% to 100%

Figure 13 focuses on the portion of the lift curve from 0% to 10%. This is the portion of the curve that would be most important for analysts looking for quality leads to focus on for further investigation of insider trading. In this section of the curve, C4.5 outperformed the other approaches. Neural nets and logistic regression performed slightly worse and similar to the expert model. Since the expert model was designed for analysts who would only be investigating the best alerts and ignoring the rest, it is not surprising that it performs better in the lower part of the curve and relatively poor on the upper 90% of the curve.

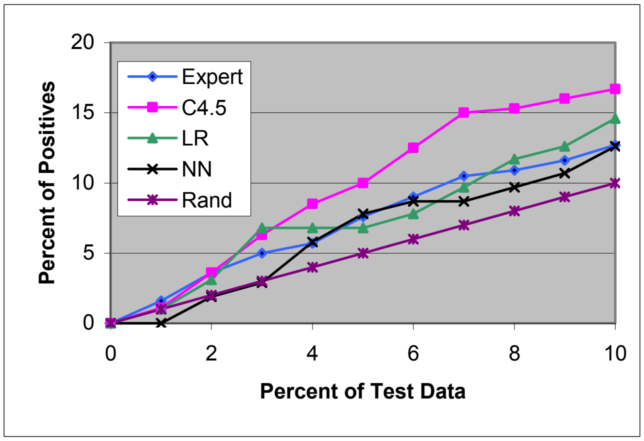

Figure 13. Lift curve 0% to 10%

Figure 14 focuses on the lowest portion of the lift curve from 0% to 2%. On this portion of the curve, the expert model has its best performance with learning algorithms performing poorly in the first 0.8% but then quickly catching up.

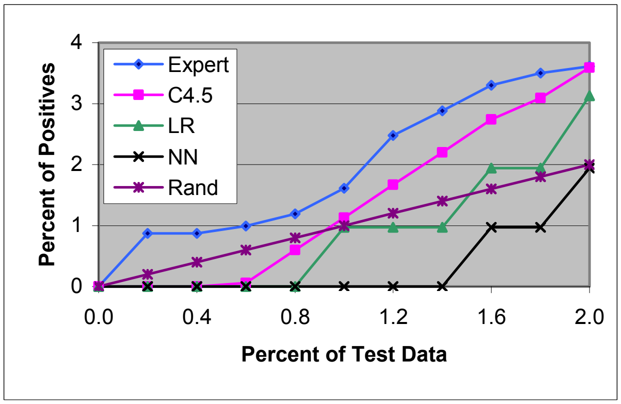

Figure 14. Lift curve 0% to 2%

Overall, while the expert model did well for the lowest part of the curve, learning algorithms were superior on the rest of the curve. Decision trees performed best on the crucial portion between 0% and 10%. If an analyst were to investigate 8% of the 5,848 possible alerts from the 100 trading days, this would amount to about 5 alerts per day, a reasonable workload. Using the C4.5 model would roughly double the hit rate over random selection and would give a hit rate 45% higher than the expert model.

5.1 Examination of Models Produced

The model produced by logistic regression used eight features: implied volatility, change in implied volatility, call volume times average, call volume standard deviations above average, call/put ratio, call/recent-put ratio, number of news articles that day, and how many of the last five days have had call volume double the moving average. Implied volatility and its change were the strongest features in the model. The model did not select companies with rising implied volatility, but instead selected companies with very high implied volatility that had remained steadily high for the last 20 days. This makes sense that companies that are known to be the most volatile might be volatile sometime in the near future. These might reflect situations where news is expected, but it is unknown whether it will be good or bad. The model was less likely to select companies with a lot of news on that day. This makes sense because the trading is more likely to be a reaction to news in those cases. The model was less likely to select companies that had multiple days in the last five where call volume was double the average. This was unexpected. High call volume on multiple days in a week would be expected to indicate interest, but in high volatility options it could also indicate news broke in the last week, and the volume and the volatility are a reaction to news. So it makes sense in the context of this model.

The decision tree algorithm produced a much larger model. Some highlights from that model include:

- Implied volatility <= 0.31 almost never were positives.
- High implied volatility (> 0.83) with volume distributed over more than two series, were likely to be positive.
- Moderate implied volatility where there had been a big jump in implied volatility were likely to be negative. A big jump in volatility could indicate a reaction to news.
- Change-in-open-interest was a complex predictor. Values near zero were labeled negative (the volume is simply people closing one call contract and opening another). Moderately high values indicated positive while very high values indicated negative (which came as a surprise but probably indicated a reaction to news). Finally, a large drops in open interest was a strong predictor of positive on a small number of examples.

5.2 Discussion of Supervised Results

All three algorithms produced lift over random and over the expert model. No algorithm clearly outperformed the others. The amount of lift was not breath-taking, but it was reasonable given the difficulty of the problem. This is especially true given that the examples that did not have high call volume (the best indicator) had already been pre-filtered from the dataset leaving only those that were most difficult to classify. Analysis of the models revealed some model logic that supported expectations. Other model logic contradicted conventional wisdom. Implied volatility was found to be a very good predictor while increase in implied volatility was less important than anticipated. When conventional wisdom was contradicted, usually an explanation was found such as indicating a reaction to news.

5.3 Clustering Results

In addition to supervised techniques, K-Means clustering was run on the positively labeled examples to see if there were any naturally occurring groupings among the positives.

- Two clusters were found with significant news on the day of the alert. One of these clusters also had high volumes of both calls and puts – this probably represents the rare case where material news was followed up within two weeks by more material news.
- One cluster had high scores by the expert model. This may indicate that the expert model captures one class of trading anomaly, but not all. Interestingly, this cluster had relatively low call volume.
- Another cluster was distinguished by high implied volatility, increasing volatility, and also high call volume compared to other clusters.

6. RELATED WORK

Our work was inspired by McMillian's hypothesis in [5] that people with inside information leave evidence in option trading data that might predict news. McMillian's method was mostly manual and required a large amount of human intuition and manual analysis. The goal of our work was to automate the analysis and discover unknown relationships. SONAR [3] is a live system used by NASD analysts to investigate insider trading. SONAR starts its analysis by finding material news and then looks backwards in time to find trading anomalies. This approach works quite well for regulators whose goal is to prosecute inside traders after the fact. The goal of our work was to find insider trading before the news becomes public. Westphal and Blaxton [7] use data mining to find cyclical price manipulation where fraudsters buy stock, drive the price of the stock up, sell their stock, wait for the price to settle down, and then repeat the cycle. They refer to this scheme as "insider trading," but they really are using the phrase differently than we are.

7. FUTURE WORK

There are many directions the work could be taken. There are relationships among companies such that news in one company affects trading in another company. Or news in one bellwether stock may affect trading in a whole industry. This information was not exploited in this paper, but it could be obtained from news [1] or publicly available industry classifications.

Our work brought in some information from previous days by comparing to twenty-day moving averages, looking at sliding three-day windows for news, and looking at volume from the previous five trading days. There is still ample room for the addition of more sophisticated time series analysis.

Our work was based on end-of-day summary data. Crucial detail is lost in the summarization process. For example, the number and size of trades during the day gives insight into whether volume is created by one large buyer or a number of small ones.

Much of our work consisted of intelligently going from low-level features to mid-level features. This makes the problem of insider trading a prime candidate for applying automated feature construction. This could be done from scratch (directly from the low-level features) or by starting with expert-initiated features and mutating them.

Our date range of data did not allow us to look for re-occurring instances of trading anomalies, but it makes sense that if a company has a "leak" there might be trading anomalies before regularly scheduled news such as earnings announcements. A history of trading anomalies (manifesting itself as seasonal changes around earnings release time) that correctly predicted news could indicate a leak and be a predictor of future news.

Other future directions include exploring put volume and downward movements, merging the expert model with learning, applying boosting, and using sophisticated text mining techniques to improve information extracted from the business news.

8. SUMMARY

Early detection of insider trading has many potential uses ranging from detecting terrorist plans to protecting investors. Because the information needed to solve the problem is spread out across many types of data (text and structured, relational and non-relational, company level and series level) it poses many challenges for data mining and allows many technologies to be brought to bear on one problem. Our work synthesized option data and news and applied data mining to predict future news. Results both supported some expectations and revealed discoveries of previously unknown relationships.

9. ACKNOWLEDGEMENTS

Thank you to Jim Hayden, Henry Goldberg, Kip Hyde, and anonymous reviewers for helpful comments on this work. Much of this work was done while the author was employed at Mantas.

10. APPENDIX A: EXPERT MODEL

To calculate the daily score for an alert, the score was initialized to zero, and the following logic was used:
If near term call volume > 300 and <= 1000, score=score+3
If near term call volume > 1000 and <= 3000, score=score+4
If near term call volume > 3000, score=score+5
If call volume # StDev above average > 2 and <=4, score=score+2
If call volume # StDev above average > 4, score=score+4
If % call volume in top series > 40 and <= 50, score=score+2
If % call volume in top series <= 40, score=score+4
If call to recent high put vol ratio >1.3 and <=2, score=score+1
If call to recent high put vol ratio >2, score=score+2
If 2 day equity price increase > -11% and <= -5%, score=score-1
If 2 day equity price increase > 0% and <= 5%, score=score+1
If 2 day equity price increase > 5% and <= 55%, score=score+2
If near term call/put volume ratio > 3 and <= 6, score=score+1
If near term call/put volume ratio > 6, score=score+2
The final score for a company was the sum of the scores from the five most recent days (but final score set to 0 if there was significant news that day or One/Thirty news ratio > 0.2)

11. REFERENCES

[1] Bernstein, A., Clearwater, S., Hill, S., Perlich, C., & Provost, F., *Discovering Knowledge from Relational Data Extracted from Business News*, SIGKDD-2002 Workshop on Multi-Relational Data Mining, July 2002, Edmonton, Alberta, CA.

[2] Fawcett, T. & Provost, F., Activity Monitoring: Noticing Interesting Changes in Behavior, *Proc. of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 1999, San Diego, CA.

[3] Goldberg, H., Kirkland, D., Lee, D., Shyr, P., & Thakker, D., The NASD Securities Observation, News Analysis & Regulation System (SONAR), *Proc. of Innovative Applications of Artificial Intelligence 2003*, August 2003 Acapulco, Mexico.

[4] Hull, J., *Options, Futures, & Other Derivatives,* Prentice Hall, 2000.

[5] McMillian, L., *McMillan on Options*, John Wiley & Sons, 1996.

[6] Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[7] Westphal, C. & Blaxton, T., *Data Mining Solutions*, John Wiley & Sons, 1998