Predicting order direction using support vector machines

Enrique Areyan Viqueira [1] Esfandiar Haghverdi [2]

[1]Department of Computer Science, Brown University

[2]School of Informatics and Computing, Indiana University

October 31, 2015

Outline

1. Prediction of Order Direction
2. Trade Classification Algorithms
3. Support Vector Machines
4. Data
5. Results
6. Conclusions & Future Research

Prediction of Order Direction

- The problem: Given an order, predict its direction (i.e. buy or sell).

Prediction of Order Direction

- The problem: Given an order, predict its direction (i.e. buy or sell).
- Order types: limit buy order, a partial cancellation, a deletion, etc.

Prediction of Order Direction

- The problem: Given an order, predict its direction (i.e. buy or sell).
- Order types: limit buy order, a partial cancellation, a deletion, etc.
- Contribution: A novel classification (prediction) method for order directionality.

Prediction of Order Direction

- The problem: Given an order, predict its direction (i.e. buy or sell).
- Order types: limit buy order, a partial cancellation, a deletion, etc.
- Contribution: A novel classification (prediction) method for order directionality.
- A more general problem than trade classification problem
- Use Support Vector Machines (SVM) for one stock of the NASDAQ market

Trade Classification Algorithms

P_T: execution price of a trade T.
T': the trade right before T
T'': the previous trade closest to T with $P_T \neq P_{T''}$.

Tick Rule
If $P_T > P_{T'}$, then $T =$ Buy.
If $P_T < P_{T'}$, then $T =$ Sell.
If $P_T = P_{T'}$, then (if $P_T > P_{T''}$ then $T =$ Buy, else $T =$ Sell).

Note that this algorithm is inconclusive in case there is no previous trade T'' such that $P_T \neq P_{T''}$.

Trade Classification Algorithms (cont.)

Let *Bid* and *Ask* be the best bid and ask quotes at time t

Quote Rule

A trade is a Buy (Sell) if it is executed at a price that is higher (lower) that the quote midpoint.

If $P_T > \frac{Bid+Ask}{2}$, then $T =$ Buy.
If $P_T < \frac{Bid+Ask}{2}$, then $T =$ Sell.
If $P_T = \frac{Bid+Ask}{2}$, then *inconclusive.*

The biggest disadvantage of this algorithm: it cannot determine the direction of the trade if the execution price is the same as the quote midpoint.

Trade Classification Algorithms (cont.)

- **LR** (Lee and Ready)
- If $P_T = \frac{Bid+Ask}{2}$, use **Tick Rule**, else use **Quote Rule**.

Trade Classification Algorithms (cont.)

- **LR** (Lee and Ready)
- If $P_T = \frac{Bid+Ask}{2}$, use **Tick Rule**, else use **Quote Rule**.

- **EMO** (Ellis et al.)
- If ($P_T = Bid$ or $P_T = Ask$), then use **Quote Rule**, else use **Tick Rule**.

Trade Classification Algorithms (cont.)

Decile Rule (Chakrabarty et al.)
The bid-ask spread is divided into deciles (10% increments).
Let s denote the spread: $s = Ask - Bid$ and $Mid = \frac{Ask+Bid}{2}$

If $(P_T > Ask$ or $P_T < Bid$ or $Mid - 0.2s \leq P_T \leq Mid + 0.2s)$
then use **Tick Rule**.

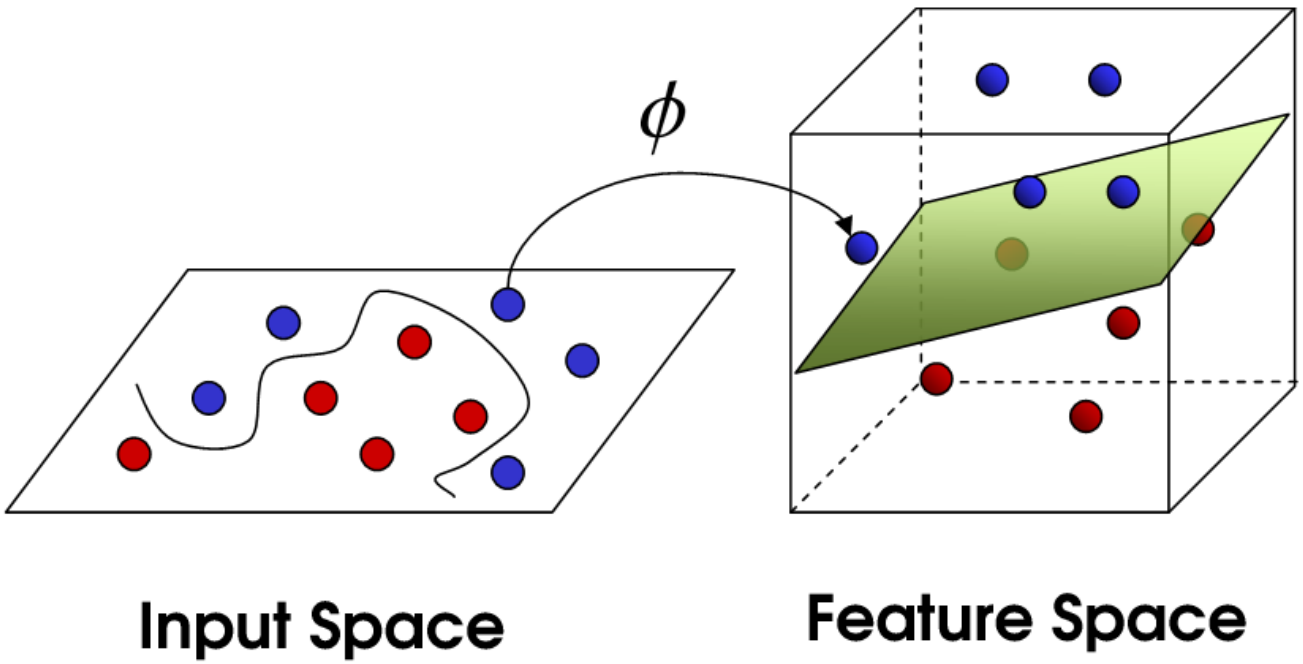
If $(Mid + 0.2s < P_T \leq Ask$ or $Bid \leq P_T < Mid - 0.2s)$
then use **Quote Rule**.

Support Vector Machines

$$f(x) = h(x)^T\beta + \beta_0 = \sum_{i=1}^{n} \alpha_i y_i \langle h(x), h(x_i)\rangle + \beta_0.$$

We use:

$$K(x, x') = exp(-\gamma||x - x'||^2).$$



[1] Image from: http://www.inf.unitru.edu.pe/revistas/2014/13.pdf

Data

1. *Time*: seconds after midnight with decimal precision of at least milliseconds and up to nanoseconds
2. *Type*: this is a categorical feature with 6 possible values:
   - 1: submission of a new limit order.
   - 2: partial cancellation of a limit order
   - 3: total deletion of a limit order
   - 4: execution of a visible limit order
   - 5: execution of a hidden limit order
   - 7: trading halt indicator
3. *Order ID*: unique order reference number
4. *Size*: number of shares
5. *Price*: dollar price
6. *Trade Direction*:
   - -1: Sell limit order
   - 1: Buy limit order

Feature Selection

- Fundamental set of features $\mathcal{F} = \{Size, Price\}$.

Feature Selection

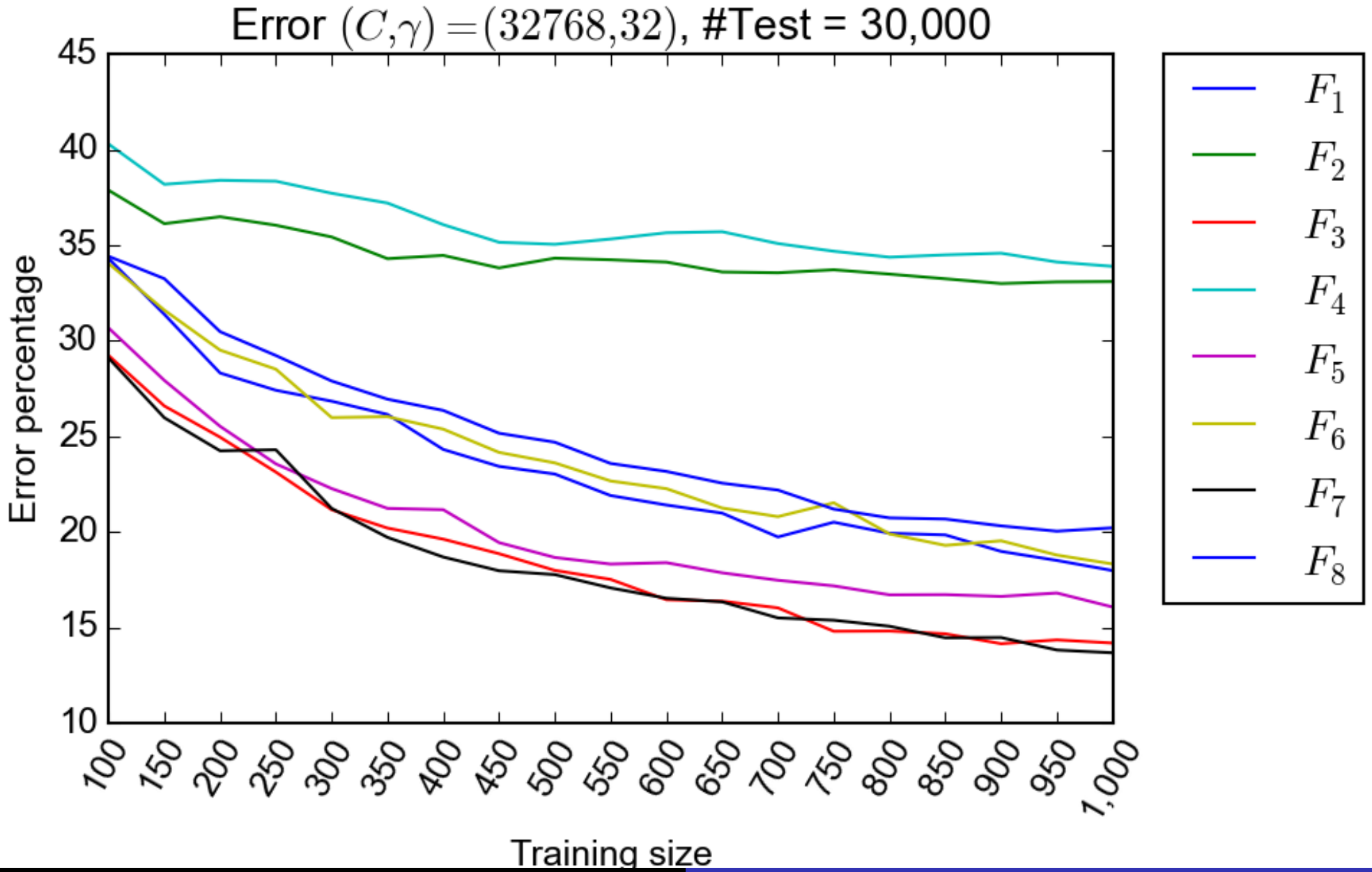
- Fundamental set of features $\mathcal{F} = \{Size, Price\}$.
- Secondary features $S = \{Time, Type, OrderId\}$

Feature Selection

- Fundamental set of features $\mathcal{F} = \{Size, Price\}$.
- Secondary features $S = \{Time, Type, OrderId\}$

$$
\begin{aligned}
\mathcal{F}_1 &= \{Size, Price, Time, Type, OrderId\} \\
\mathcal{F}_2 &= \{Size, Price\} \\
\mathcal{F}_3 &= \{Size, Price, Time\} \\
\mathcal{F}_4 &= \{Size, Price, Type\} \\
\mathcal{F}_5 &= \{Size, Price, OrderId\} \\
\mathcal{F}_6 &= \{Size, Price, Time, Type\} \\
\mathcal{F}_7 &= \{Size, Price, Time, OrderId\} \\
\mathcal{F}_8 &= \{Size, Price, Type, OrderId\}
\end{aligned}
$$

Feature Selection (cont.)



Parameter Optimization

- Need to choose two parameters: (C, γ).

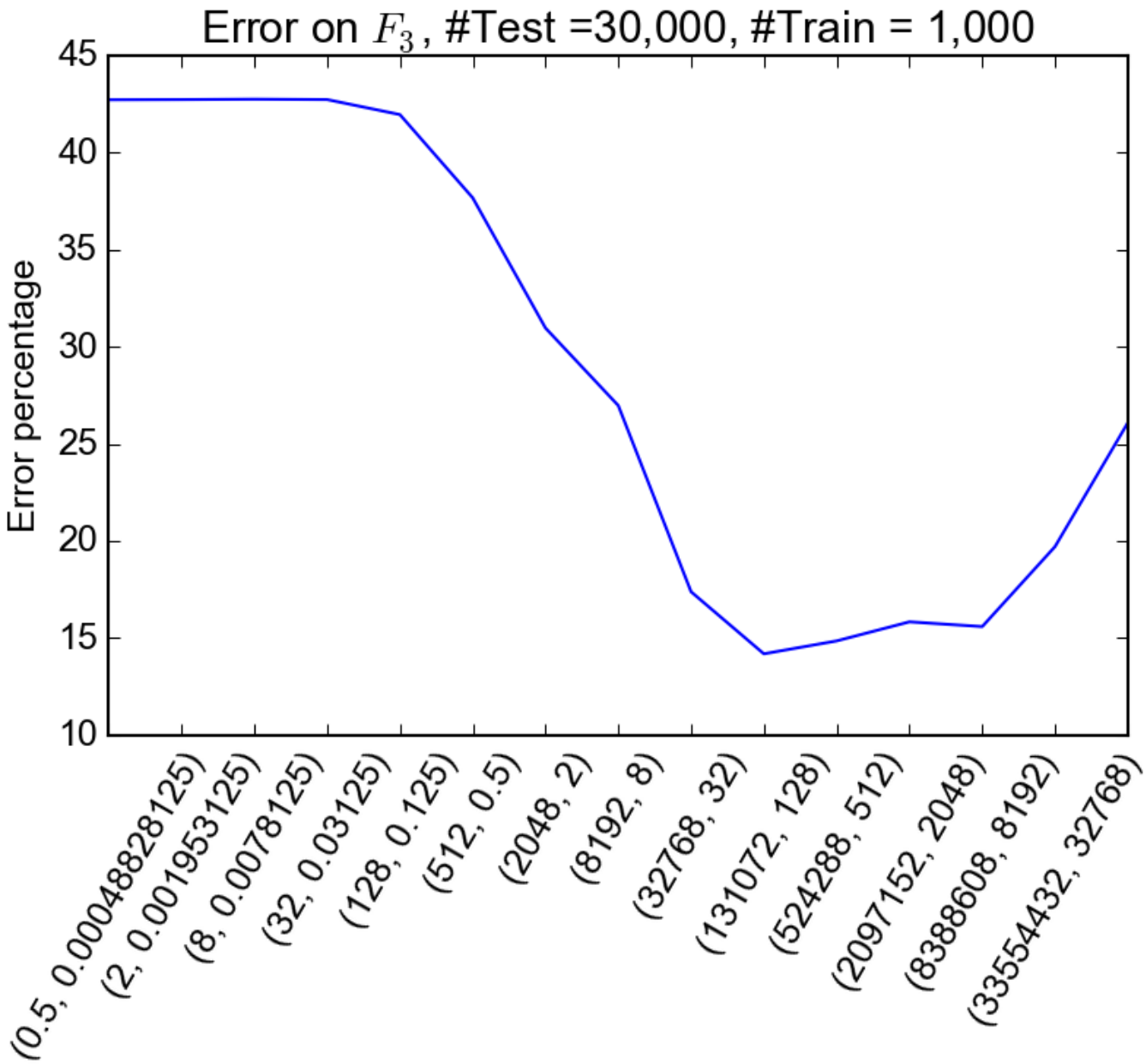
Parameter Optimization

- Need to choose two parameters: (C, γ).
- No a priori knowledge about what values of C and γ will work

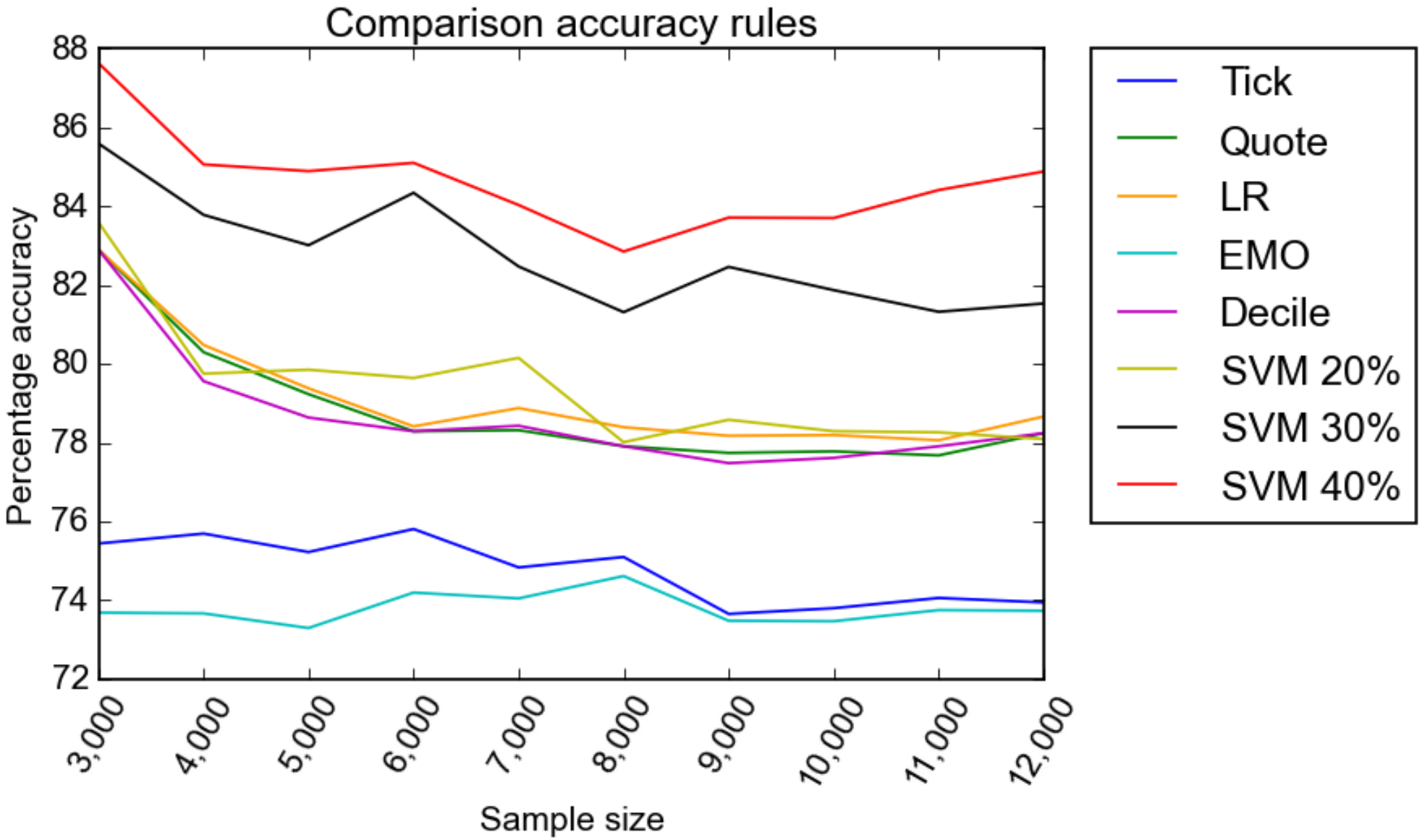
Parameter Optimization

- Need to choose two parameters: (C, γ).
- No a priori knowledge about what values of C and γ will work
- Solution: simple grid search on the parameter space (C, γ) for different powers of 2 for both parameters: $(2^{-3}, 2^{-13}), (2^{-1}, 2^{-11}), (2^{1}, 2^{-9}), \ldots . (2^{15}, 2^{5})$.

Parameter Optimization (cont.)



Results


Comparison accuracy rules
Percentage accuracy
88
86
84
82
80
78
76
74
72
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
11,000
12,000
Sample size
Tick
Quote
LR
EMO
Decile
SVM 20%
SVM 30%
SVM 40%

Conclusions

SVMs Advantages

1. easily trained and can handle vast amounts of data
2. reliable and highly accurate for trade direction classification, as shown by our experiments.
3. fast predictions imply viable alternative for real time order (trade) classification problems
4. independent of any hypothesis about the structure or functioning of a market
5. can be used in a wide variety of distinct markets.

Conclusions (cont.)

SVMs Disadvantages

1. same as with any data-driven approach: does not provide the user with an explanation of the underlying mechanism at work

2. you get no simple rules like Tick rule or Quote rule either

Conclusions (cont.)

Two key points for SVM training: feature and parameter selection.

Both of these tasks can be automated to result in a highly accurate model as compared to previous classification rules available in the literature

We showed that for a particular data set SVM outperforms all other proposed rules.

Future Research

1. Test our method on more stocks and other exchanges than NASDAQ and compare the results
2. Increase the efficiency and speed by parallelization
3. Test other machine learning formalisms: trees, logistic regression, and neural networks