



Mining High-Speed Data Streams

Pedro Domingos & Geoff Hulten
Department of Computer Science & Engineering
University of Washington

Datum : 22.11.2006

Seminar: Maschinelles Lernen und symbolische Ansätze

Vortragender: Kei Ogata

Mining High-Speed Data Streams

■ Übersicht

1. **Einführung**
2. **Hoeffding trees**
3. **Das VFDT System**
4. **Ausblick**
5. **Fazit**

Mining High-Speed Data Streams

■ Einführung

Was ist ein High-Speed Data Stream?

Ein Data Stream, der eine so hohe Datendichte hat, dass er mit konventionellen Methoden nicht verlustfrei verarbeitet werden kann.

Wo entstehen High-Speed Data Streams?

- Transaktionen im Einzelhandel, in Banken und in der Telekommunikationsbranche
- Web + Ubiquitous Computing

- 1.
- 2.
- 3.
- 4.
- 5.

Mining High-Speed Data Streams

■ Einführung

Wie sollte ein High-Speed Data Stream verarbeitet werden?

Die Daten treffen sequenziell ein



Ein inkrementeller Lerner wäre sinnvoll

ABER:

- Keine Garantie für die Übereinstimmung mit dem Batch-Modell
(falls garantiert, dann langsamer als der Batch-Algorithmus)
- Abhängigkeit von der Datenreihenfolge

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Einführung

Welche Eigenschaften sollte ein Lerner haben, der ein High-Speed Data Stream verarbeitet?

Ein Lerner für High-Speed Data Streams sollte:

- ... für eine unendlich große Datenmengen konzipiert sein
- ... jedes Datum nur maximal ein Mal aufrufen müssen
- ... bei der Verarbeitung eines Datum möglichst schnell sein
- ... Daten online verarbeiten können

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Einführung

Merkmale eines Entscheidungsbaum-Lerners:

- In den Knoten befinden sich Tests
- In den Blättern steht die Klassifizierung
- Gute Attribute sollten möglichst weit oben stehen

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Einführung

Beispiele für Entscheidungsbaum-Lerner

- ID3, C4.5 und CART (Classification and Regression Trees)

- **Nachteil**: Trainingsdaten müssen alle gleichzeitig im Arbeitsspeicher vorhanden sein



Die maximale Größe des Baumes wird durch den Arbeitsspeicher begrenzt

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Einführung

Alternative Entscheidungsbaum-Lerner:

- SLIQ, SPRINT
 - Vorteil: Trainingsdaten müssen **nicht** alle gleichzeitig im Arbeitsspeicher vorhanden sein
 - Nachteil: Trainingsdaten müssen alle auf der Festplatte vorhanden sein und mehrmals von dort gelesen werden
-  Die maximale Größe des Baumes wird durch den Festplattenspeicher begrenzt

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Hoeffding trees

Die Hoeffding Schranke

- Gegeben: eine reelle Zufallsvariable r mit der oberen Schranke R
- r wird n -Mal beobachtet und der Durchschnitt \bar{r} gebildet
- Mit der Wahrscheinlichkeit $1-\delta$ ist der wahre Durchschnitt

mindestens $\bar{r} - \varepsilon$ mit $\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}}$

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

1.
2.
3.
4.
5.

■ Hoeffding trees

Die Hoeffding Schranke

- $G(X_i)$ sei ein heuristisches Maß für das Attribut i
- Gesucht: das Attribut X_a , wofür $\bar{G}(X_i)$ den Maximalwert hat, nachdem n Beispiele gelernt wurden
- Sei $\Delta\bar{G} = \bar{G}(X_a) - \bar{G}(X_b) \geq 0$ der Unterschied zwischen dem besten und zweitbesten Attribut
- Wenn n Beispiele gelernt wurden und $\Delta\bar{G} > \varepsilon$, dann garantiert die Hoeffding Schranke mit der Wahrscheinlichkeit $1 - \delta$, dass das Attribut X_a das beste Attribut ist

Mining High-Speed Data Streams

■ Hoeffding trees

Wie werden Hoeffding trees aufgebaut?

- Traversiere für jedes Datum aus dem Stream den bereits aufgebauten Baum und finde die Stelle an die es gehört
- Falls die Klasse des neuen Datums nicht mit der der anderen übereinstimmt:
- Berechne $\Delta\bar{G} = \bar{G}(X_a) - \bar{G}(X_b) \geq 0$ und splitte den Knoten mit dem Attribut X_a auf

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Hoeffding trees

Welche Eigenschaften haben Hoeffding trees?

- Entscheidungsbaum
- Komplexität beim Lernen:
im worst-case proportional zur Anzahl der Attribute
- Annähernd hohe Ähnlichkeit zum Ergebnis eines Batch-Lerners (bei genügend großer Anzahl an Trainingsdaten)
- Die Wahrscheinlichkeit, dass Hoeffding trees und konventionelle Entscheidungsbäume unterschiedliche Tests in den Knoten haben, sinkt mit der Anzahl der Trainingsdaten exponentiell.

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Hoeffding trees

Welche Komplexitäten haben Hoeffding trees?

- d : Anzahl der Attribute
- v : Anzahl der Werte, die ein Attribut annehmen kann
- c : Anzahl der Klassen
- l : Anzahl der Blätter

- Erstellung: $O(dvc)$
- Speicherplatz: $O(ldvc)$

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Das VFDT System

Was ist das VFDT System?

- Very Fast Decision Tree Lerner
- Baut einen Hoeffding tree auf
- Man kann Information Gain oder Gini-Index als Maß für die Heuristik verwenden

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Das VFDT System

Vorteile des VFDT Systems:

- Behandlung von Attributen mit ähnlichem G
- Wiederberechnung von G
- Organisation des Arbeitsspeicher
- Behandlung schlechter Attribute
- Initialisierung
- Rescan der Beispiele

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Das VFDT System

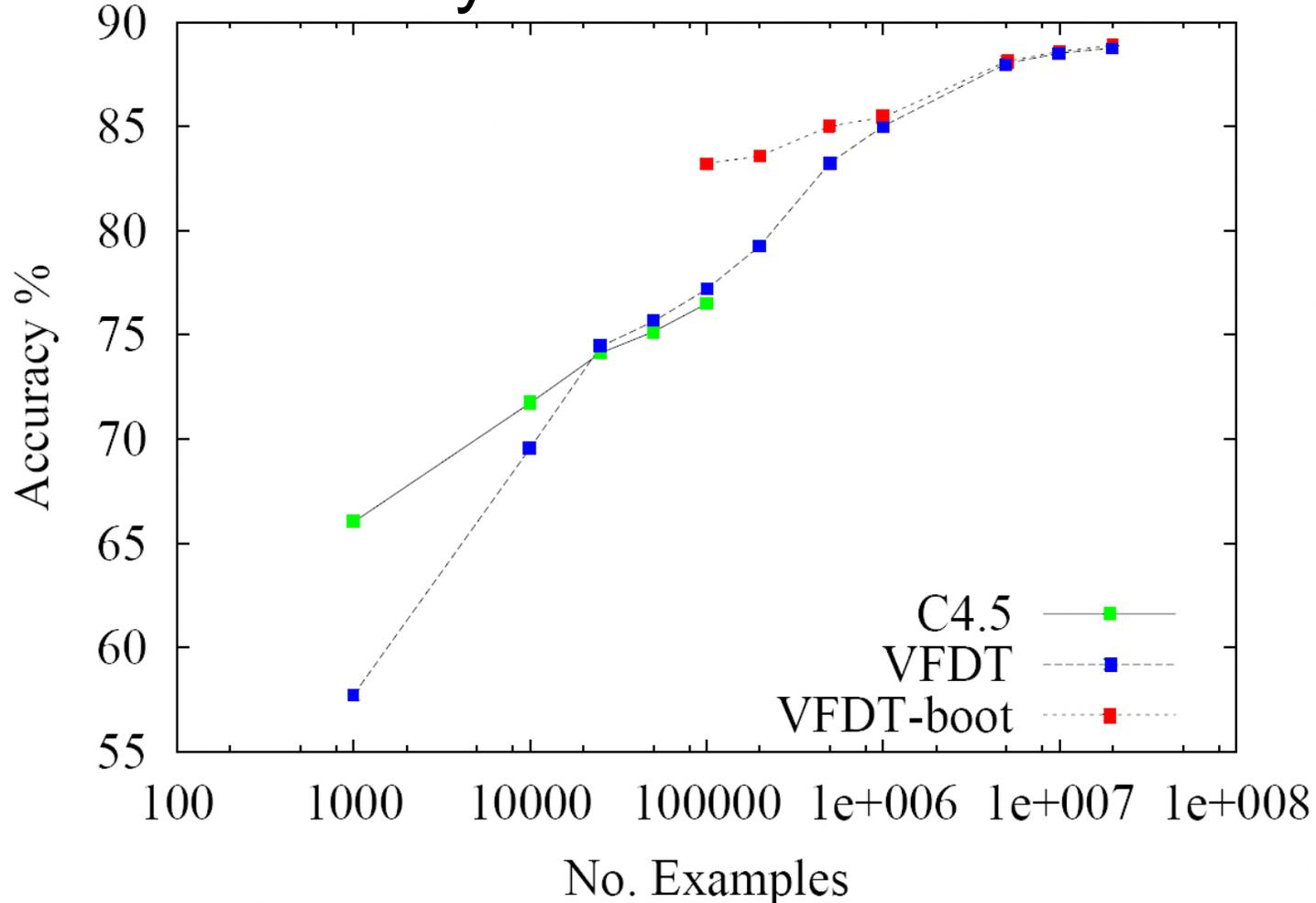
Experimente mit dem VFDT System

- Verwendung von synthetischen Daten
- Heuristisches Maß von VFDT war information gain
- Verglichen mit C4.5
- 40MB Hauptspeicher
- Noise von 0 % ~ 30 %
- 50k verschiedene Testdaten pro Durchlauf

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

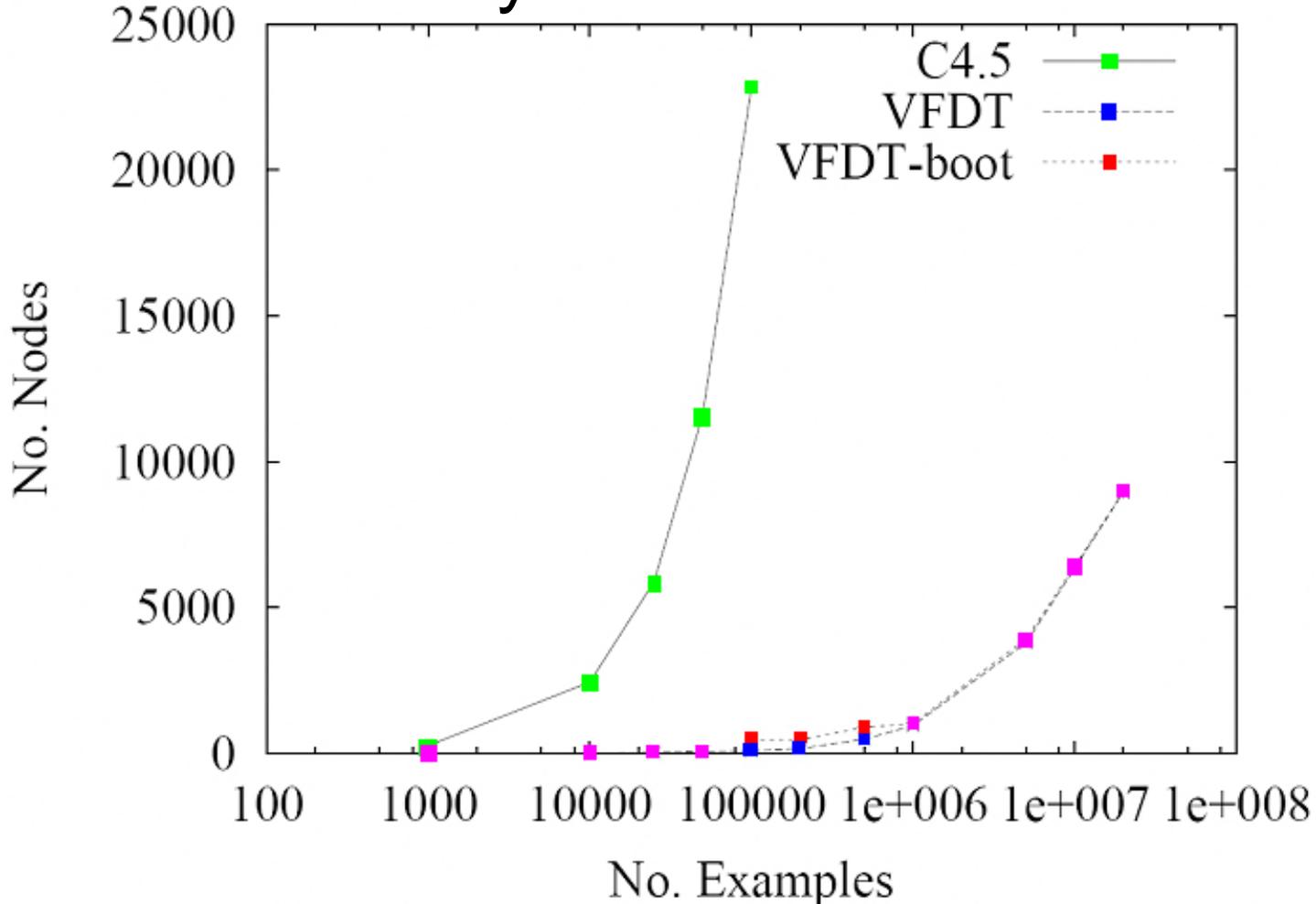
■ Das VFDT System



- 1.
- 2.
- 3.
- 4.
- 5.

Mining High-Speed Data Streams

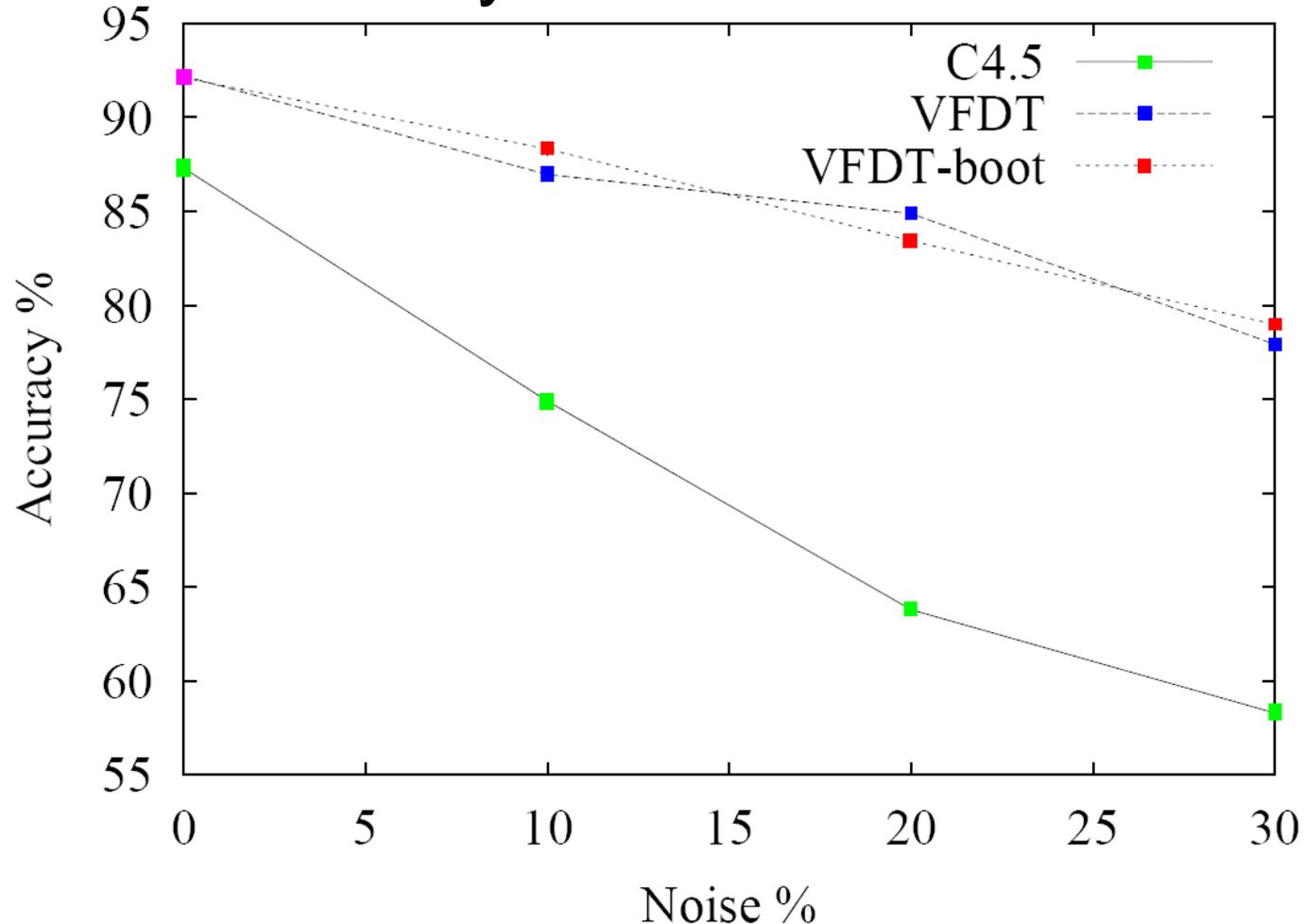
■ Das VFDT System



- 1.
- 2.
- 3.
- 4.
- 5.

Mining High-Speed Data Streams

■ Das VFDT System



- 1.
- 2.
- 3.
- 4.
- 5.

Mining High-Speed Data Streams

■ Ausblick

- Vergleich mit SPRINT/SLIQ und ID5R
- Verwendung für Weblog Daten oder IDS
- Verwendung von numerischen Attributen
- Verwendung von Kosten
- Verwendung in Umgebungen, in denen sich das Konzept mit der Zeit ändert
- Verwendung in Umgebungen, in denen sich noch nicht einmal das Ergebnis in den Speicher passt

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Fazit

- Hoeffding trees sind geeignet, um sehr schnell inkrementell zu lernen
- Hoeffding trees haben eine asymptotische Ähnlichkeit zu Bäumen, die im batch-Verfahren gelernt wurden
- VFDT ist ein auch in Experimenten erfolgreich getestetes System zu Verarbeiten von High-Speed Data Streams

1.
2.
3.
4.
5.

Mining High-Speed Data Streams

■ Quellenangaben

- Mining HighSpeed Data Streams
Pedro Domingos & Geoff Hulten
- Data Mining auf Datenströmen
Andreas M. Weiner
Integriertes Seminar SS 2005 TU Kaiserslautern



Mining High-Speed Data Streams

Vielen Dank
für Eure
Aufmerksamkeit