Seminar:
Aktuelle Arbeiten des Data Mining

Prof. Dr. Katharina Morik
Informatik LS8

Das Internet, Fabriken, (private) Mediensammlungen, Geschäftsvorgänge hinterlassen eine Fülle von Daten. Aus diesen Daten etwas zu extrahieren, was nützlich ist für den Benutzer, für den Fertigungsprozess, für Dienstprogramme, für die Entscheidungsfindung, ist die Aufgabe von Maschinellem Lernen und Data Mining. Inzwischen sind diese Teilgebiete der *Künstlichen Intelligenz* so weit fortgeschritten, dass man sich nicht mehr so ohne weiteres einarbeiten kann. Damit ergibt sich eine Situation, die Ihnen später im Beruf oft begegnen wird:

- Ein Thema ist plötzlich in aller Munde, aber die Entwicklung dieses Gebiets haben Sie nicht verfolgt.
- Wie kann man sich einarbeiten? Man muss Fachliteratur lesen!
- Am Anfang versteht man nur Bruchstücke der Artikel – am Ende war es dann doch nicht so schwer.

Diese Situation wird mit dem Seminar geübt. Kurze Einleitungen zu den Themenblöcken werden von mir gegeben, die Literatur dazu ist aber auch als Grundlage jeweils angegeben. Das Seminar verhilft dazu, aktuelle Forschungsergebnisse zu verstehen.

Literatur zur Auswahl

Insbesondere werden in dem Seminar die folgenden Verfahren und Ansätze mit ihren aktuellen Erweiterungen vorgestellt:

- Frequent Set Mining
  Grundlage in [10], Kapitel 5.2

  1. Komprimierte Muster [19]
  2. Erweiterung auf Zeichenketten in verteilten Datenbanken [12]
  3. Erweiterung auf Sequenzen [18]

- Top Down Induction of Decision Trees
  Grundlage im Handbuch der Künstlichen Intelligenz, Kapitel 14.3

  1. Erweiterung auf verteilte Datenbanken [2]
  2. andere verteilte Data Mining Techniken: [6]

- Support Vector Machines
  Grundlage Tutorial [4], Optimierungsalgorithmus SMO [17]

  1. Erweiterung auf strukturelle Ausgaben [21]
  2. Strukturelle SVM zum Graph-Labelling [11]
  3. Verschiedene Optimierungsalgorithmen zur SVM [5]
  4. Erweiterung der SVM zum online Lernen [7]
  5. Erweiterung der SVM zum inkrementellen Lernen [13]
  6. Erweiterung der strukturellen SVM für Sequenzen [3]
  7. Erweiterung der SVM für peer-to-peer computing [1]

- Community Mining

  1. Grundlage: Soziale Netzwerke [15]
  2. Collaborative Filtering [14]
  3. Finden überlappender Communities in sozialen Netzwerken [8]
  4. Community Discovery [16]

- Subgruppenentdeckung

  1. Schnelle Subgruppenentdeckung [9]
  2. Subgruppenentdeckung durch rekursives Partitionieren [20]

Vorgehen

Schauen Sie sich die angegebene Literatur an: welches Referat möchten Sie gern halten? Es sind weit mehr Artikel angegeben, als in ein Seminar passen. Ich habe also eine Auswahlliste gegeben. Allerdings ist die Liste geordnet, so dass die Artikel aufeinander aufbauen. Beachten Sie bei Ihrer Auswahl diese Ordnung – wenn niemand die vorangegangenen Artikel referieren möchte, können Sie den

Artikel nicht auswählen! Die Artikel sind zu sehen auf der LEHRE-Seite von www-ai.cs.tu-dortmund.de.

Ein Referat zu halten, bedeutet:

- Den Originaltext lesen, einige der zitierten Aufsätze und ggf. Hintergrundmaterial in Form von Tutorials lesen.
- Eine Präsentation des Textes im Seminar abhalten.
- Eine schriftliche Ausarbeitung des Originaltextes mit Bezügen zu den anderen Referaten abgeben. Dabei gilt die Regel: so viele Tage nach Abschluss des Semesters, wie Sie sich Zeit für die Ausarbeitung nehmen, nehme ich mir nach Ihrer Abgabe Zeit für die Begutachtung, die zu einem Schein oder Nachbesserungen führt.

Beginn und Anmeldung

Das Seminar beginnt am **Dienstag, 14.4.2009**. Die Anmeldung erfolgt durch Erscheinen im R 113 des GB IV an diesem ersten Termin. Eine zusätzliche Anmeldung ist nicht erforderlich.

Voraussetzungen

Voraussetzung für das Seminar ist ein abgeschlossenes Grundstudium und die Wahlpflichtveranstaltung "Darstellung, Erwerb und Verarbeitung von Wissen". Das Seminar passt gut zu den Vorlesungen "Wissensentdeckung" und "Maschinelles Lernen", die aber nicht vorausgesetzt werden.

References

[1] Hock Hee Ang, Vivekanand Gopalkrishnan, Steven C.H. Hoi, and Wee Keong Ng. Cascade RSVM in peer-to-peer networks. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Proces. ECML PKDD 2008*. Springer, 2008.

[2] Kanishka Bhaduri, Ran Wolff, C. Gianella, and Hillol Kargupta. Distributed decision tree induction in peer-to-peer systems. *Statistical Analysis and Data Mining Journal*, 1(2):85 – 103, 2008.

[3] Antoine Bordes, Nicolas Usunier, and Leon Bouttou. Sequence labelling SVMs trained in one pass. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Procs. ECML PKDD*, pages 146–161. Springer, 2008.

[4] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[5] Olivier Chapelle, Vikas Sindhwani, and Sathiya Keerthi. Optimization techniques for semi-supervised support vector machines. *J. Machine Learning Research*, 9, 2008.

[6] Kamalika Das, Kanishka Bhaduri, Kun Liu, and Hillol Kargupta. Distributed identification of top-l inner product elements and its application in a peer-to-peer network. *IEEE Transactions on Knowledge and Data Engineering*, 2008.

[7] Hal Daume and Daniel Marcu. Learning as search optimization: Approximate large margin methods for structured prediction. In *Procs. ICML*, 2005.

[8] Steve Gregory. A fast algorithm to find overlapping communities in networks. In Walter Daelemands, Bart Goethals, and Katharina Morik, editors, *Procs. ECML PKDD*, pages 408 – 423. Springer, 2008.

[9] Henrik Grosskreutz, Stefan Rüping, and Stefan Wrobel. Tight optimistic estimates for fast subgroup discovery. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Procs. ECML PKDD*, pages 440 – 456. Springer, 2008.

[10] Jiawei Han and Micheline Kamber. *Data Mining – Concepts and Techniques*. Morgan Kaufmann, 2 edition, 2006.

[11] Thoralf Klein, Ulf Brefeld, and Tobias Scheffer. Exact and approximate inference for annotating graphs with structural SVMs. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Procs. ECML PKDD*, pages 611 – 623. Springer, 2008.

[12] Adrian Kügel and Enno Ohlebusch. A space efficient solution to the frequent string mining problem for many databases. *Data Mining Knowledge Discovery*, 17:24 – 38, 2008.

[13] Pavel Laskov, Christian Gehl, Stefan Krüger, and Klaus-Robert Müller. Incremental support vector learning: Analysis, implementation and applications. *J. Machine Learning Research*, 7, 2006.

[14] Heng Luo, Changyong Niu, Ruimin Shen, and Carsten Ullrich. A collaborative filtering framework based on both local user similarity and global user similarity. *Machine Learning*, 72(3):231–245, 2008.

[15] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Procs. ISWC*. Springer, 2005.

[16] Spiros Papadimitriou, Jimeng Sun, Christos Faloutsos, and Philip S. Yu. Hierarchical, parameter-free community discovery. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *Procs. ECML PKDD 2008*. Springer, 2008.

[17] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Scholkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 12. MIT-Press, 1999.

[18] Chedy Raissi, Toon Calders, and Pascal Poncelet. Mining conjunctive sequential patterns. *Data Mining and Knowledge Discovery*, 17(1):77–93, 2008.

[19] Arno Siebes, Jilles Vreeken, and Matthijs van Leeuwen. Item sets that compress. In *Procs. SIAM Int. Conference on Data Mining*, 2006.

[20] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *J. Machine Learning Research*, 10, 2009.

[21] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large Margin Methods for Structured and Interdependent Output Variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.