Classifying The Real World

Anwendungen von
Support Vector Machines

Sebastian Bitzer (sbitzer@uos.de)
Sven Lauer (svlauer@uos.de)
Seminar: Neuronale Netze
University of Osnabrueck
10. 07. 2003

Überblick

- Textkategorisierung
- Bioinformatik
- Gesichtserkennung
- weitere Anwendungen

Textkategorisierung

Textkategorisierung

- Eine der „killer"-Anwendungen für SVMs
- Wachsende Informationsvielfalt (Internet, Gb-Harddrives, …) → Information sammeln ist billig, Information nutzbar machen ist teuer
- Momentan geschieht (erfolgreiche) Textkategorisierung meistens durch Menschen
- Verschiedenste Ansätze: Regelbasiert, Neuronale Netze, … und eben: SVMs
- Idee: maschinelles Lernen mit feedback (da schon grosse Mengen klassifizierter Informationen vorliegen)
- Klassifikation: klassische SVM-Aufgabe
- binär (spam/non-spam) oder mehrere Kategorien (Kombination von SVMs)

Textkategorisierung (II)

- Representation eines Textes durch Wortvektor: binär („bag of words") oder gewichtet (z. B. nach Worthäufigkeit)
- Probleme bei reinen „bag of words"-Ansätzen:
  - Reihenfolge, Beziehungen zwischen den Wörtern werden ignoriert
  - Thematisch ähnliche Texte müssen nicht unbedingt dieselben Wörter enthalten (Synonyme, Hypernyme, Hyponyme, …)
    → mögliche Lösung: „Ähnlichkeit" zwischen Wörtern definieren und in Kernel integrieren
    → anstatt diese Ähnlichkeitsbeziehung zu definieren, könnte diese auch gelernt werden (annäherende Synonymie könnte etwa definiert werden als häufiges Auftreten zweier Wörter im gleich Kontext, ohne das diese je gemeinsam auftreten)
    → andere Ansätze vorhanden, besonders interessant: Entdeckung semantischer Ähnlichkeit durch Analyse bilingualer Korpora

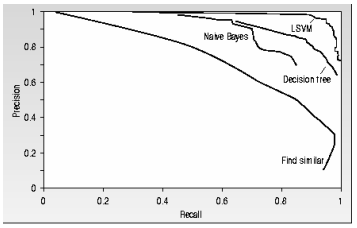
Eigenschaften von Textklassifikationsaufgaben

- Sehr viele Features (Wörter im Lexikon)
- Wenige irrelevante Features
- Aber: sparse input vectors
- Oft sind die Kategorien linear trennbar, falls nicht, kann mit entsprechendem Kernel nachgeholfen werden
- → SVM scheinen zur Textklassifikation eine gute Idee zu sein

Joachims et. al. (1998)

	Bayes	Rocchio	C4.5	k-NN	SVM (poly) degree d = 1	2	3	4	5	SVM (rbf) width γ = 0.6	0.8	1.0	1.2
earn	95.9	96.1	96.1	97.3	98.2	98.4	**98.5**	98.4	98.3	**98.5**	98.5	98.4	98.3
acq	91.5	92.1	85.3	92.0	92.6	94.6	**95.2**	95.2	95.3	95.0	95.3	95.3	**95.4**
money-fx	62.9	67.6	69.4	78.2	66.9	72.5	75.4	74.9	**76.2**	74.0	75.4	**76.3**	75.9
grain	72.5	79.5	89.1	82.2	91.3	93.1	**92.4**	91.3	89.9	**93.1**	91.9	91.9	90.6
crude	81.0	81.5	75.5	85.7	86.0	87.3	88.6	**88.9**	87.8	**88.9**	89.0	88.9	88.2
trade	50.0	77.4	59.2	77.4	69.2	75.5	76.6	77.3	**77.1**	76.9	78.0	**77.8**	76.8
interest	58.0	72.5	49.1	74.0	69.8	63.3	67.9	73.1	**76.2**	74.4	75.0	**76.2**	76.1
ship	78.7	83.1	80.9	79.2	82.0	85.4	86.0	**86.5**	86.0	**85.4**	86.5	87.6	87.1
wheat	60.6	79.4	85.5	76.6	83.1	84.5	85.2	**85.9**	83.8	**85.2**	85.9	85.9	85.9
corn	47.3	62.2	87.7	77.9	86.0	86.5	85.3	**85.7**	83.9	**85.1**	85.7	85.7	84.5
micro-avg.	**72.0**	79.9	**79.4**	**82.3**	84.2	85.1	85.9	86.2	85.9	86.4	86.5	86.3	86.2
					combined: 86.0					combined: 86.4			

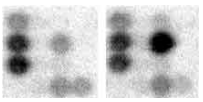
Lineare SVM gegen klassische Ansätze



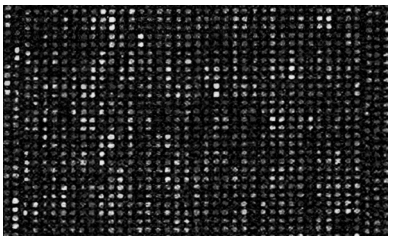
Bioinformatik

Genexpression

- Mensch und Schimpanse teilen 98,7% ihres Erbgutes
- aber Gene werden im Gehirn des Menschen bis zu viermal mehr benutzt (4-mal stärkere Genexpression)



Microarray



Methodik: http://www.bio.davidson.edu/courses/genomics/chip/chip.html

Das Klassifikationsproblem

	Gen	1.	2.	3.	4.		78.	79.
1	YAL001C	-0.38	-0.38	-0.43	-0.06	…	-0.67	-0.38
2	YAL002W	-0.3	-0.09	-0.18	-0.14	…	-0.45	-0.15
…	…	…	…	…	…	…	…	…
2467	YPR201W	-0.04	0.16	0.12	-0.38	…	-0.27	0.07

- Hefe-Gene anhand der reellwertigen Vektoren zu vorher existierenden funktionalen Klassen zuordnen (5 funktionale Klassen + 1 Kontrollklasse)

$\Rightarrow$ eine SVM pro Klasse

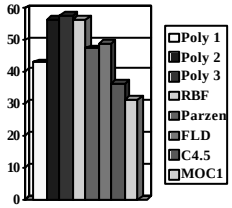
Ergebnis

- da Anzahl der negativen so hoch (z.B. 2450 gegenüber 17 pos.), ist Fehlerrate bei allen Algorithmen sehr niedrig

⇒ „cost savings“ definiert

- beste Methode für jede der 5 Klassen ist eine SVM (entweder mit höher-gradigem Polynomkernel oder RBF-Kernel)

mittlere Kosteneinsparung (cost savings)

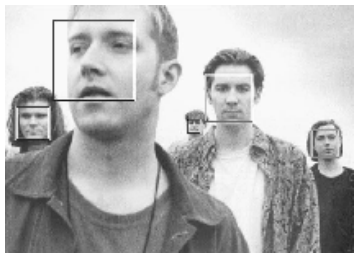


Gesichtserkennung

(face) detection

- Detection:
  - Gesichter
  - Tumore in MRI-Scans
  - Strukturfehler in produzierten Teilen
- Vorbedingung zur Identifikation einer abgebildeten Person
- Probleme:
  - starke Variabilität in zu findenden Mustern (unterschiedliche Gesichter, Gesichtsausdrücke, Brillen, Schatten)

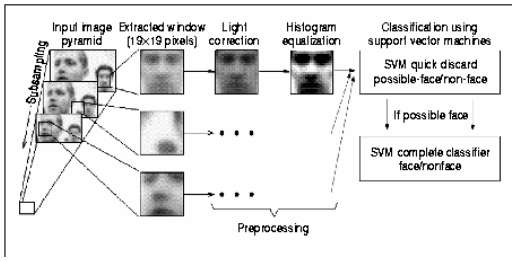
Ziel



Vorgehensweise

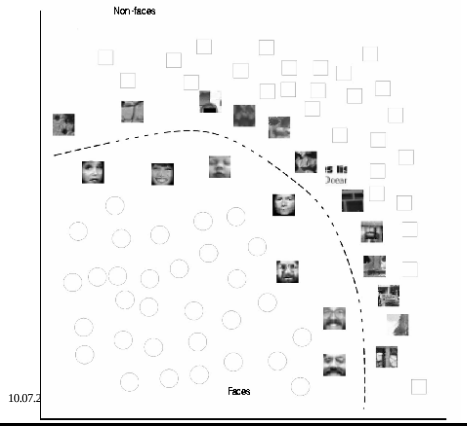
- Gleich große Fenster (hier 19x19 pixel) aus dem Bild ausschneiden
- Bild skalieren (vergrößern/verkleinern), wieder Fenster ausschneiden
- So bekommt man eine Reihe von unterschiedlich grossen Fenstern aus dem Bild
- Fenster vorverarbeiten
- SVM klassifiziert Gesicht / Nicht-Gesicht
- Falls Gesicht: markieren mit Rahmen

Vorgehensweise

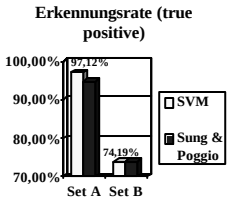


Bootstrapping

- falsch negativ klassifizierte Beispiele als negative Beispiele für weitere Trainingsdurchgänge (z.B. aus Landschaftsbildern)
- Sinnvoll, da Variation unter Nicht-Gesichtern wesentlich grösser ist als Variation unter Gesichtern



Ergebnisse



false positive	SVM	Sung & Poggio
Set A	4	2
Set B	20	11

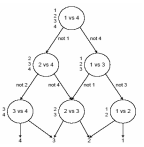
Perspektiven

- Möglicherweise auch rotierte Gesichter mit demselben classifier erkennen?
- Anwendung auf andere Klassifikationsobjekte (Tumorerkennung, …)
- Verbesserung der Performance (bessere Filterung)

Weitere Anwendungen

Handschrifterkennung

- offline vs. online
- Kernel für Sequenzen: Gaussian dynamic time warping (GDTW) kernel
- mehrere Klassen: The DAGSVM algorithm



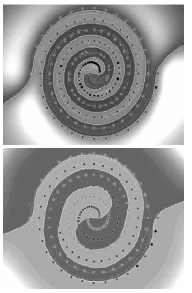
Zwei Spiralen

- KMOD (Kernel with Moderate Decreasing):

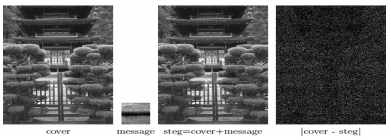
$$KMOD(x,y) = K\left[\exp\left(\frac{g}{\|x-y\|^2 + s^2}\right) - 1\right]$$

- RBF:

$$K(x,y) = \exp\left(-a\|x-y\|^2\right)$$



Steganographie-Erkennung



- Bild wird statistischer Analyse unterzogen

? Vektor

- SVM bekommt Vektor als Input und entscheidet dann, ob versteckte Nachricht enthalten, oder nicht

Zusammenfassung

- SVMs kommen mittlerweile überall vor, wo es etwas zu klassifizieren gibt
- Klassifizierung ist nicht auf zwei Klassen beschränkt
- SVMs liefern oft sehr gute Ergebnisse
- **aber**: richtige Vorverarbeitung von Daten ist essentiell (z.B. Skalierung)

References

(1) http://www.vcell.de/genomstation/genexpression.html
(2) http://filebox.vt.edu/cals/cses/marnof/mglab/Microarray.html
(3) Brown, Grundy, Lin, Cristianini, Sugnet, Furey, Ares and Haussler. *Knowledge-based analysis of microarray gene expression data by using suport vector machines.* Proc. Natl. Acad. Sci, 97:262–267, 2000.
(4) Bahlmann, Haasdonk and Burkhardt, *On-line Handwriting Recognition with Support Vector Machines--A Kernel Approach*, Proc. 8th IWFHR, 2002
(5) Platt, Cristianini and Shawe-Taylor. *Large Margin DAGS for Multiclass Classification*, Advances in Neural Information Processing Systems, 12 ed. S.A. Solla, T.K. Leen and K.-R. Muller, MIT Press, 2000
(6) Romero and Alquézar, *Maximizing the margin with feed-forward neural networks*, Proc. INNS-IEEE International Joint Conference on Neural Networks (IJCNN2002), pp.743-748., 2002
(7) Lyu and Farid, *Detecting Hidden Messages Using Higher-Order Statistics and Support Vector Machines.* Proc. 5th International Workshop on Information Hiding 340-354, 2002
(8) Osuna, Freund and Girosi, *Support Vector Machines: Training and Applications*, 1997
(9) Joachims, *Text Categorization with Support Vector Machines, Learning With Many Relevant Features*, 1998
(10) http://www.support-vector.net
(11) http://www.kernel-machines.org
(12) http://www.clopinet.com/isabelle/Projects/SVM/applist.html