Transformationswissen in Support-Vektor-Maschinen

Vortrag zu Grundlagen der Bild-erzeugung und Bildanalyse WS 02/03

Ablauf

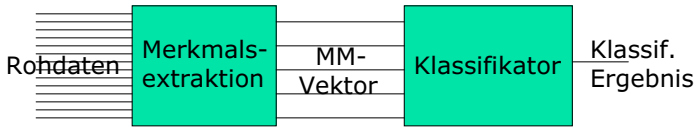
- Einleitung
- Transformationswissen
- Support-Vektor-Maschinen
- Tangenten in SVM (1)
  - Virtuelle Supportvektoren
- Tangenten in SVM (2)
  - Tangentendistanz
  - Tangentendistanz-Kerne
- Anwendung USPS-Ziffern
- Zusammenfassung

Ablauf

- **Einleitung**
- Transformationswissen
- Support-Vektor-Maschinen
- Tangenten in SVM (1)
  - Virtuelle Supportvektoren
- Tangenten in SVM (2)
  - Tangentendistanz
  - Tangentendistanz-Kerne
- Anwendung USPS-Ziffern
- Zusammenfassung

Motivation

- Grundsatz:
  Verbesserung der Generalisierungsfähigkeit eines Klassifikationssystems durch Einbringen von a-priori Wissen in die Mustererkennungskette



- Möglichkeiten des Einbringens von Vorwissen:
  - In Merkmalsextraktion
  - In Wahl des Klassifikatortyps
  - In Klassifikatorerzeugung

Motivation

- Beispiel Buchstabenerkennung
  - Wissen um Bedeutungserhaltung bei
    Änderung der Schreibgeschwindigkeit, -richtung,
    Verschiebung, Skalierung, Scherung
  - Einbringen in die Merkmalsextraktion:
    Invariante Repräsentation durch Normalisierung
  - MATLAB-Demo
  - beliebige Klassifikatoren anwendbar
  - es existieren viele Arten von Vorwissen, die nicht direkt in die Merkmalsextraktion eingebracht werden können
    z.B: leichte Rotationen, nichtlineare Verzerrungen, Rauschen

Motivation

- Dies führt zur Frage:
  Wie kann a-priori Wissen in Klassifikatorerzeugung eingebracht werden?
- ... Frage natürlich viel zu allgemein:
- Vielfalt an Klassifikatoren:
  - Bayes-, Nächste Nachbar-, Polynomklassifikator,
  - Perceptron, Neuronale Netze
  - Support-Vektor-Maschine, ...
- Daher Beschränkung auf SVM

Motivation

- Vielfalt an Arten von a-priori Wissen
  - Herkömmlichste Art: Lernbeispiele
  - statistisches Wissen: klassenweise exakte Verteilungen, oder Modellannahmen
  - Wissen über Struktur der Daten: z.B. Nachbarschaften von Pixeln
  - Existenz eines „Ähnlichkeitsmaß" oder „Distanzmaß" seiner Objekte, z.B. DTW-Distanz bei Online HWR
  - Transformationswissen: Klassenerhalt unter bekannten Transformationen
- Daher Beschränkung auf Transformationswissen
- Frage: Wie kann Transformationswissen in SVM eingebracht werden?

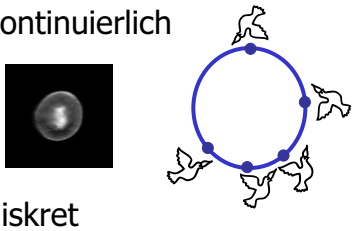
Ablauf

- Einleitung
- **Transformationswissen**
- Support-Vektor-Maschinen
- Tangenten in SVM (1)
  - Virtuelle Supportvektoren
- Tangenten in SVM (2)
  - Tangentendistanz
  - Tangentendistanz-Kerne
- Anwendung USPS-Ziffern
- Zusammenfassung

Transformationswissen

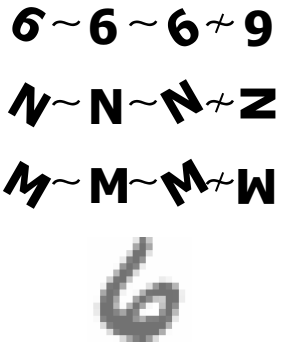
- Global parametrisierte Transformationsgruppe

kontinuierlich



diskret

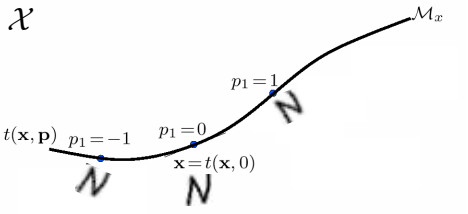
- Lokale Transformationen



....

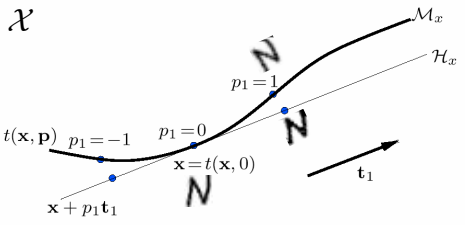
Transformationswissen

- Formalisierung:
  - Differenzierbare Abbildung $\mathbf{t}(\mathbf{x},\mathbf{p})$ mit $\mathbf{t}:\mathcal{X}\times\mathcal{P}\to\mathcal{X}, \mathcal{P}\subseteq\mathbb{R}^l$ Parametermenge, $\mathcal{X}$ Musterraum. Üblicherweise $\mathbf{t}(\mathbf{x},\mathbf{0})=\mathbf{x}$ Mannigfaltigkeit $\mathcal{M}_\mathbf{x}$ der transformierten Muster.
- Anschauung:



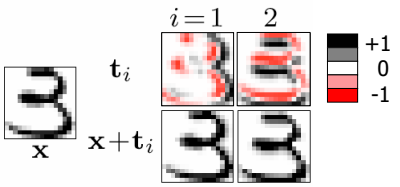
Transformationswissen

- Exakte Berechnung von $\mathcal{M}_x$ oft aufwendig, daher
- Lokale lineare Approximation in $\mathbf{x}$:
  - Tangenten $\mathbf{t}_i := \frac{\partial}{\partial p_i}\mathbf{t}(\mathbf{x},\mathbf{p})\Big|_{\mathbf{p}=\mathbf{0}}$
  - Tangentialebene $\mathcal{H}_x$ aufgespannt durch die $\mathbf{t}_i$
- Anschauung:



Transformationswissen

- Beispiel Tangentialverschiebung von Bildern
  - x-Translation und y-Translation



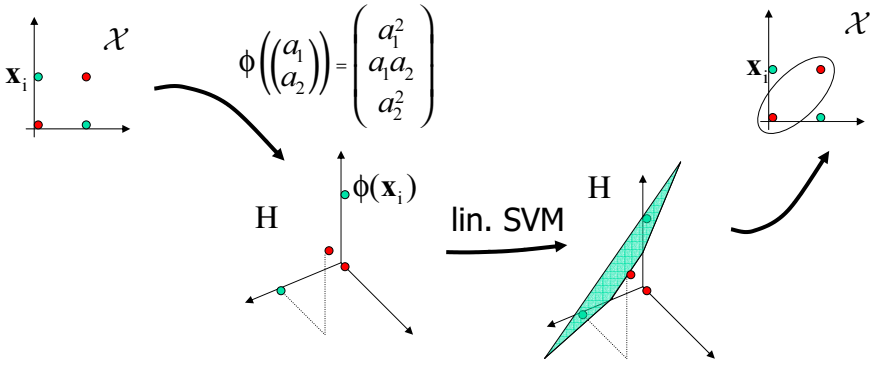
  - MATLAB-Demo

Ablauf

- Einleitung
- Transformationswissen
- **Support-Vektor-Maschinen**
- Tangenten in SVM (1)
  - Virtuelle Supportvektoren
- Tangenten in SVM (2)
  - Tangentendistanz
  - Tangentendistanz-Kerne
- Anwendung USPS-Ziffern
- Zusammenfassung

Wiederholung Nichtlineare SVM

- Nichtlineare Abbildung $\phi(\mathbf{x}) : \mathcal{X} \to \mathrm{H}$ anschließend lineare SVM
  - Beispiel XOR-Problem:



Kernel Trick

- Problem:
  Berechnungskomplexität bei Operieren auf $\phi(\mathbf{x}_i)$
- Beobachtung:
  in Training und Klassifikation treten nur Skalarprodukte $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_H$ auf.
- Trick:
  Effizient berechenbare Kernfunktion
  $$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle_H$$
  macht Kenntnis von ϕ und H überflüssig!
- Anwendbar auf Vielzahl linearer Algorithmen „kernelization"

Kernfunktionen

- Vorstellung:
  - Ähnlichkeitsmaß, verallgemeinertes Skalarprodukt
- Einfache Beispiele
  - Linear $k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$
  - Polynomial $k(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + c)^p$
  - Sigmoid $k(\mathbf{x}, \mathbf{y}) = \tanh(\kappa \langle \mathbf{x}, \mathbf{y} \rangle + \theta)$
  - Gauss-RBF $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\|\mathbf{x} - \mathbf{y}\|^2 / (2\sigma^2)\right)$
- Allgemeine Bedingung
  - Mercer Theorem oder
  - Positive Definitheit

Anwendung von SVM

- Trainingsdaten $(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathcal{X}, i = 1, \ldots, N$
  $y_i \in \{\pm 1\} \triangleq$ 2-Klassen-Klassifikation
- Wähle Kern $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, z.B. linear $\langle \vec{x}, \vec{x}' \rangle$, Gauss-RBF $e^{-\gamma \|\mathbf{x} - \mathbf{x}'\|^2}$
- Setze Train-param. C und Kern-param. (γ / p)
- Lösung eines Quadratischen OP ergibt
  $$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b\right)$$
- Angenehme Eigenschaften
  - Spärlichkeit, wenige $\alpha_i \neq 0 \Rightarrow \mathbf{x}_i$ „Support-Vektor"
  - Optimalität, gute Generalisierung

A-priori Wissen in SVM

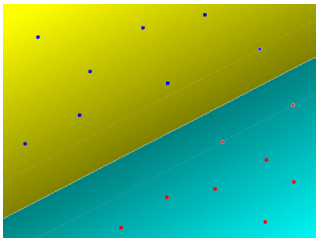
- Modifikation der Datenrepräsentation
  - invariante Merkmale
- Modifikation der Trainingsdaten
  - VSV Methode [Schö96]
- Modifikation der Kerne
  - invariante Kerne durch Integration [Burg99]
  - jittering kernels [De01]
  - Tangentendistanz-Kerne
- Modifikation der Optimierungsfunktion
  - invariant hyperplane [Schö98, Chap00]

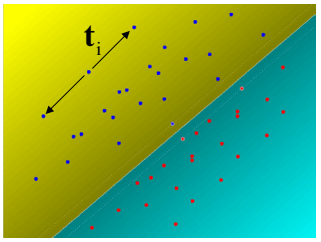
Ablauf

- Einleitung
- Transformationswissen
- Support-Vektor-Maschinen
- **Tangenten in SVM (1)**
  - **Virtuelle Supportvektoren**
- Tangenten in SVM (2)
  - Tangentendistanz
  - Tangentendistanz-Kerne
- Anwendung USPS-Ziffern
- Zusammenfassung

Virtuelle SV Methode [Schö96]

- Motivation:
  - Vervielfältigen der Trainingsdaten durch kleine Verschiebungen in Tangentialebene (=„virtuelle" Daten)
  - Originaldaten: nach Vervielfältigung:





Virtuelle SV Methode

- Problem
  - Speicherintensiv: Anzahl Lernbeispiele N'
    - +2 Beispiele für jede der l Tangenten:

      $\bullet \Rightarrow$  $\quad N' = (1+2l)\cdot N$

    - Alle Kombinationen der Tangenten:

      $\bullet \Rightarrow$  $\quad N' = 3^l \cdot N$

  - Rechenintensiv
    - SVM-Trainingsalgorithmen skalieren $\mathcal{O}\left(N^2\right)$

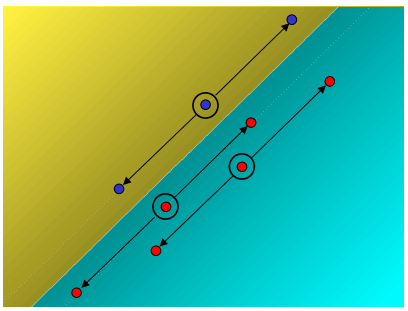
Virtuelle SV Methode

- Idee

  virtuelle Beispiele aller Daten ist ungünstig

  $\Rightarrow$ nur virtuelle Beispiele einer ausdrucksstarken Untermenge: den Supportvektoren

- Schritte
  - Trainieren
  - Extrahieren der SV
  - Erzeugen von VSV $\mathbf{t}(\mathbf{x}_i, \mathbf{p})$
  - Erneutes Training



Virtuelle SV Methode

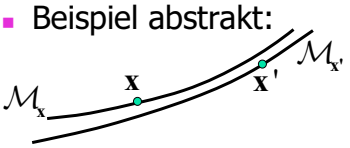
- Vorteile
  - Einfachheit,
  - alle Kerne möglich
  - diskrete Invarianzen behandelbar
- Nachteile
  - Wahl der „Größe" der Verschiebungen unklar
  - noch immer Speicher- und Laufzeitintensiv
  - resultierende SVM besitzt mehr SV als originale
  - mehrere Trainingsdurchläufe notwendig

Ablauf

- Einleitung
- Transformationswissen
- Support-Vektor-Maschinen
- Tangenten in SVM (1)
  - Virtuelle Supportvektoren
- **Tangenten in SVM (2)**
  - **Tangentendistanz**
  - **Tangentendistanz-Kerne**
- Anwendung USPS-Ziffern
- Zusammenfassung

Tangentendistanz [Sim93]

- Motivation:
  - Abstand von Mannigfaltigkeiten oft geeigneter als Punktabstände.
  - Beispiel abstrakt:



Abstand $\mathbf{x}'$ zu $\mathbf{x}$ sehr viel größer als Abstand der Mannigfaltigkeiten.

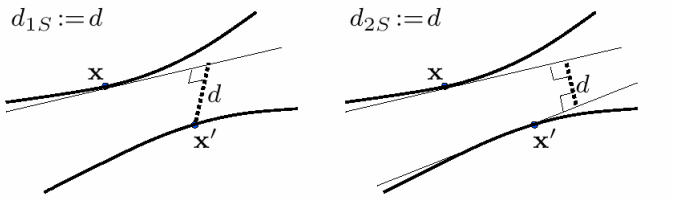
  - Beispiel konkret: $\mathbf{t}(\mathbf{x},\mathbf{p})$ sei x-Translation

$\mathbf{x} =$  $\mathbf{x}' =$ 

$\mathrm{d}(\mathbf{x},\mathbf{x}')$ maximal, jedoch $\mathrm{d}(\mathcal{M}_x,\mathcal{M}_{x'}) = 0$

Tangentendistanz

- Lokale lineare Approximation:
  - Abstand der Tangentialebenen statt Mgfk.
  - Symmetrische 2-seitige Distanz $\mathrm{d}_{2S}(\mathbf{x},\mathbf{x}')$
  - Nichtsymmetrische 1-seitige Distanz $\mathrm{d}_{1S}(\mathbf{x},\mathbf{x}')$
  - Beispiel in d=3 Dimensionen:
    Tangentialräume = Geraden (so i.a. windschief)
    allgemeiner: $\dim(\mathcal{H}_{\bar{x}}) \le \mathrm{d}-2$

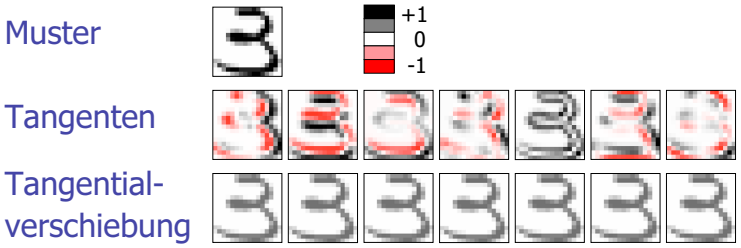


Tangentendistanz

- Berechnung
  - Beispiel $$d_{1S}(\mathbf{x},\mathbf{x}') := \min_{\mathbf{p}} \left\| \mathbf{x} + \sum_{i=1}^{l} p_i \mathbf{t}_i - \mathbf{x}' \right\|$$
  - Berechnung der Tangenten durch
    exakte Differentiation oder
    Vorwärts-/Rückwärts-/zentrale Differenzen
    $$\mathbf{t}_i \approx \mathbf{t}(\mathbf{x},\mathbf{e}_i) - \mathbf{t}(\mathbf{x},\mathbf{0})$$
  - Orthogonalisierung der Tangenten
  - Orthogonale Projektion
  - Komplexität $\mathcal{O}(l^2)$ wegen Orthogonalisierung
    insbesondere d_{2S} 4-fachen Aufwand wie d_{1S}

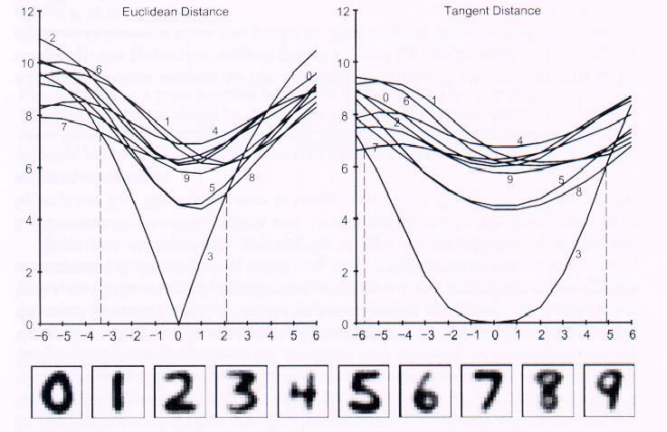
Tangentendistanz

- Anwendungsspezifische Details:
  7 Tangenten for OCR
  - x-, y-Translation, Skalierung, Rotation, Linienstärke, axiale und diagonale hyperbolische Transformationen
  - e.g:



Tangentendistanz

- Experimentelle lokale Invarianz von TD



Tangentendistanz

- Resultat
  - Neue Abstandsmaße basierend auf Tangenteninformation
  - Anwendbarkeit in allen distanzbasierten Methoden, z.B. nächste Nachbar-Klass., RBF-Netzwerk,...

Tangentendistanz-Kerne [Ha02]

- Motivation:
  - TD mehrmals „Rekordverfahren“:
    - K-Nächste Nachbar Klassifikator [Sim93]
    - statistischer Ansatz [Ke00]
  - TD noch nicht in SVM-Ansätze verwendet
  - Idee: Ersetzen von Euklidischer Distanz durch TD

z.B.

Gauss-RBF $k^{RBF}(\mathbf{x},\mathbf{x}') := e^{-\gamma\|\mathbf{x}-\mathbf{x}'\|^2}$

$\Rightarrow k^{RBF}_{2S}(\mathbf{x},\mathbf{x}') := e^{-\gamma d_{2S}(\mathbf{x}-\mathbf{x}')^2}$

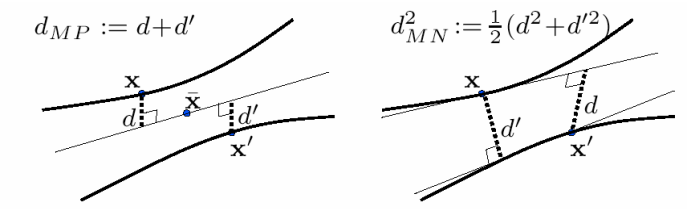
Neg. quadr. Eukl. Distanz $k^{ND}(\mathbf{x},\mathbf{x}') := -\|\mathbf{x}-\mathbf{x}'\|^{\gamma}, \quad \gamma \in (0,2]$

$\Rightarrow k^{ND}_{lS}(\mathbf{x},\mathbf{x}') := -d_{lS}(\mathbf{x},\mathbf{x}')^{\gamma}$

Tangentendistanz-Kerne

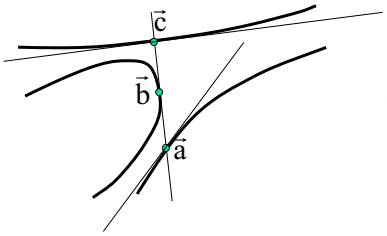
- Problem: d_{lS} nicht symmetrisch.

  $\Rightarrow$ symmetrische TD-Varianten definieren:
- d_{MP} : Verwendung von Tangenten im Mittelpunkt
- d_{MN} : Mittel von 2 einseitigen Distanzen



Tangentendistanz-Kerne

- Erkenntnis: TD-Kerne i.a. nicht pos. definit
- Wesentlicher Grund:
  TD verletzt Dreiecksungleichung
- Beispiel:

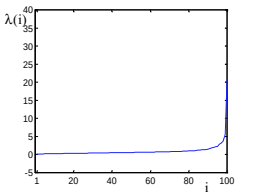


$$d_{2S}(\vec{a},\vec{c}) > d_{2S}(\vec{a},\vec{b}) + d_{2S}(\vec{b},\vec{c})$$

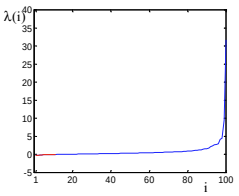
- Folge: Verlust der Optimalität, jedoch nicht notwendigerweise Generalisierungsfähigkeit!

Tangentendistanz-Kerne

- Empirische Beobachtung:
  - Lediglich schwache pd-Verletzung auf echten Daten, typische Spektren von Kernmatrizen:



idealer pd Fall des standard Kerns k^{RBF}

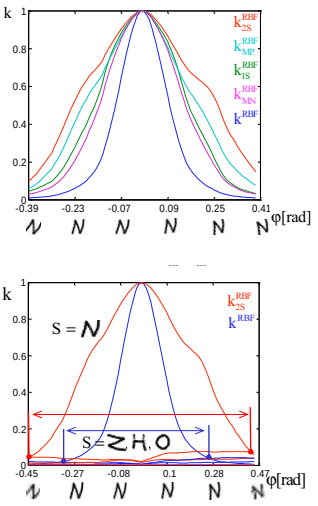


nicht pd Fall des TD-Kerns k_{2S}^{RBF}

  - Keine Konvergenzprobleme, gute Erkennungsergebnisse

Tangentendistanz-Kerne

- Verhalten unter lokalen Transformationen:
  - z.B. kleine Rotationen R_φ eines Beispiels N
  - $\Rightarrow$verbesserte Robustheit. Darstellung von $k_*^{RBF}\left(N, R_\varphi(N)\right)$
  - $\Rightarrow$verbesserte Unterscheidbarkeit. Darstellung von $k_*^{RBF}\left(S, R_\varphi(N)\right)$



Ablauf

- Einleitung
- Transformationswissen
- Support-Vektor-Maschinen
- Tangenten in SVM (1)
  - Virtuelle Supportvektoren
- Tangenten in SVM (2)
  - Tangentendistanz
  - Tangentendistanz-Kerne
- **Anwendung USPS-Ziffern**
- Zusammenfassung

Anwendung USPS-Ziffern

- Details zu den Daten:
  - Standard Benchmark-Datensatz
  - 7191 Training-, 2007 Testbeispiele von handgeschriebenen Ziffern in Form von 16x16 Graubildern.
  - Beispiele:

  - Normierung der Daten $[0,1]$, $\|\cdot\|_2 \leq 1$

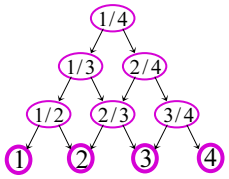
Anwendung USPS-Ziffern

- Erkennungsraten aus der Literatur [Ha02]:

Method	Error rate [%]
Human Performance [13]	2.5
Neural Net (LeNet1) [14]	4.2
SVM, no invariance [11]	4.0
SVM, VSV-method [12]	3.2
k-Nearest Neighbour [14]	*5.7
k-NN + TD [14]	*2.5
TD + kernel densities [7]	2.4
* := extended training set.	

Anwendung USPS-Ziffern

- Details zu SVM-training:
  - Basiskern k^{RBF}
  - 7 Tangenten von Simard
  - 2 symmetrische TD-Kerne $k^{RBF}_{MN}, k^{RBF}_{MP}$
  - VSV Methode für Basiskern, 14 virtuelle Beispiele
  - DAGSVM [Pla00] als Multiklassenlösung



  - SVM-Light [Joa99] für 2-Klassen Teilprobleme
  - 9-23 Parameter-Variationen (C, γ) pro Kern

Anwendung USPS-Ziffern

- Erkennungsraten:

Kernel	Error rate [%]	γ	C	# param. sets
k^{RBF}	4.6	8	10	14
k^{RBF} VSV	3.6	8	1	23
k^{RBF}_{MP}	3.5	20	10	9
k^{RBF}_{MN}	3.4	10	10	9

  - Verwendung von Tangenten ist dem Basiskern überlegen
  - TD-Kerne liefern vergleichbare Erkennung wie die VSV-Methode

Anwendung USPS-Ziffern

- Komplexitäten der Modelle:

Kernel	Training-time [s]	Test-time [s]	Average # SVs
k^{RBF}	199	228	175
k^{RBF} VSV	4521	1232	1094
k^{RBF}_{MP}	2437	3394	232
k^{RBF}_{MN}	3159	3950	133

  - Verwendung von Tangenten führt zu Erhöhung der Laufzeiten
  - TD-Kerne schneller als VSV während Training, langsamer beim Testen
  - TD-Kerne erzeugen kleine Modelle

Ablauf

- Einleitung
- Transformationswissen
- Support-Vektor-Maschinen
- Tangenten in SVM (1)
  - Virtuelle Supportvektoren
- Tangenten in SVM (2)
  - Tangentendistanz
  - Tangentendistanz-Kerne
- Anwendung USPS-Ziffern
- **Zusammenfassung**

Zusammenfassung

- Grundsatz „Einbringen von Vorwissen zur Erkennungsverbesserung"
- Existierende Vielfalt von Vorwissen und allgemeine Methoden, diese in Klassifikationssysteme einzubringen
- Detaillierte Darstellung von Formen von Transformationswissen
- Detaillierte Schilderung von 2 Methoden, wie dies in SVM eingebracht werden kann:
  VSV-Methode und TD-Kerne
- Experimenteller Vergleich der Verfahren anhand Ziffernerkennung

Referenzen

- [Burg99] C. Burges. Geometry and invariance in kernel based methods. In *Advances in Kernel Methods – Support Vector Learning*, pages 89-116. MIT Press, 1999.
- [Chap00] O. Chapelle and B. Schölkopf. Incorporating invariances in nonlinear SVMs. In *Neural Information Processing Systems*, 2000.
- [De02] D. DeCoste and B. Schölkopf. Training invariant support vector machines. *Machine Learning*, 16(1):161-190,2002.
- [Ha02] B.Haasdonk, D. Keysers, „Tangent Distance Kernels for Support Vector Machines", ICPR 2002, Proc. 16th Int. Conf. Pat. Rec., 2002.
- [Joa99] T. Joachims. Making large-scale SVM learning practical. In *Advances in Kernel Methods – Support Vector Learning*, pages 169-184. MIT Press, 1999.
- [Ke00] D. Keysers, J. Dahmen, T. Theiner, and H. Ney. Experiments with an extended tangent distance. In *Proceedings 15th International Conference on Pattern Recognition, vol. 2*, pages 38-42. IEEE Computer Society, 2000.
- [Pla00] J. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin dags for multiclass classification. In *Advances in Neural Information Processing Systems, 12*, pages 547-553. MIT Press, 2000.
- [Schö95] B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In Proceedings First International Conference on Knowledge Discovery & Data Mining, pages 252-257. AAAI Press, 1995.
- [Schö96] B. Schölkopf, C. Burges, and V. Vapnik. Incorporating invariances in support vector learning machines. In *Artificial Neural Networks – ICANN'96, LNCS, 1112*, pages 47-52. Springer, 1996.
- [Schö02] B. Schölkopf, A. Smola, „Learning with kernels", MIT Press, 2002.
- [Sim93] P. Simard, Y. LeCun, J. Denker, „Efficient pattern recognition using a new transformation distance", In Advances in Neural Information Processing Systems, 5, pp. 50-58, 1993.
- [Sim98] P. Simard, Y. LeCun, J.S. Denker, and B. Victorri. Transformation invariance in pattern recognition – tangent distance and tangent propagation. In *LNCS, 1524*, pages 239-274. Springer, 1998.