**FORSCHUNGSZENTRUM JÜLICH GmbH**
**Zentralinstitut für Angewandte Mathematik**
**D-52425 Jülich, Tel. (02461) 61-6402**

Interner Bericht

# Gütemaße zur Optimierung von Support-Vektor-Maschinen

*Sebastian Schnitzler, Tatjana Eitrich*

FZJ-ZAM-IB-2006-06

März 2006

(letzte Änderung: 28.3.2006)

# Gütemaße zur Optimierung von Support-Vektor-Maschinen

Sebastian Schnitzler, Tatjana Eitrich

Zentralinstitut für Angewandte Mathematik
Forschungszentrum Jülich
http://www.fz-juelich.de

**Zusammenfassung** Support-Vektor-Maschinen sind ein modernes und effizientes Verfahren zur Klassifikation von Daten. Mittels des sogenannten Trainings auf einem bekannten Datensatz erstellen sie eine trennende Hyperebene zwischen den Klassen, welche in einem hochdimensionalen Merkmalsraum liegt. Dabei hängt das Optimierungsproblem von Parametern ab, die die Lage der trennenden Hyperebene beeinflussen. Diese Parameter müssen vor dem Training ausgewählt und bewertet werden. In dieser Arbeit werden Gütemaße zur Qualitätseinschätzung von Ergebnissen während der Parameteroptimierung erläutert. Wir entwickeln ein neues Gütemaß, welches auf der geometrischen Anschauung von Ergebnissen basiert und einen gewichteten Abstand zum optimalen Punkt berechnet. Es basiert auf der sogenannten Receiver-Operating-Characteristic (ROC), die zur graphischen Darstellung von Klassifikationsergebnissen geeignet ist und gute Ergebnisse sichtbar macht. In unseren Tests analysieren wir den Einfluss von Gütemaßen bei der Parameteroptimierung. Wir vergleichen die Ergebnisse unseres neuen Maßes mit denen, die durch die Nutzung traditioneller Maße erzielt werden können. Für die Analyse wird der frei verfügbare Wisconsin Breast-Cancer-Datensatz verwendet, der zu den Datensätzen gehört, bei denen Sensitivität eine große Rolle spielt.

## 1 Einleitung

Klassifikationsprobleme treten auf, sobald gegebene Datenpunkte einer von zwei oder mehr Klassen zugeordnet werden müssen. Zur Lösung eines Klassifikationsproblemes gibt es verschiedene Verfahren. Eine Methode ist die der Support-Vektor-Maschinen (SVM's) [Vap95]. Das Prinzip der Support-Vektor-Maschinen basiert auf der Trennung der verschiedenartigen Punkte durch eine optimale Hyperebene in einem hochdimensionalen Merkmalsraum. Sei dazu

$$\mathcal{T} := \left\{ \left(x^i, y_i\right) \in \mathcal{X} \times \{-1, 1\}, \quad i = 1, \ldots, t \right\} \tag{1}$$

ein Trainingsdatensatz mit $t$ Punkten in einem $n$-dimensionalen Raum $\mathcal{X}$ ($t \in \mathbb{N},\ n \in \mathbb{N}$). Jedem Vektor $x^i$ ist ein Klassenlabel $y_i$ zugeordnet. $n$ ist die Anzahl der betrachteten Variablen (Attribute). Die Anwendung von Support-Vektor-Maschinen führt auf ein quadratisches Optimierungsproblem, welches in zwei

Varianten vorliegen kann. Das sogenannte 1-Norm-Modell hat die Form

$$\min_{\alpha\in\mathbb{R}^t} \quad \frac{1}{2}\sum_{i=1}^{t}\sum_{j=1}^{t} y_i y_j \alpha_i \alpha_j \langle x^i, x^j\rangle - \sum_{i=1}^{t}\alpha_i \tag{2}$$

unter den Nebenbedingungen $y^T\alpha = 0$ und $0 \leq \alpha \leq C$. Das 2-Norm-Modell ist von der ähnlichen Gestalt

$$\min_{\alpha\in\mathbb{R}^t} \quad \frac{1}{2}\sum_{i=1}^{t}\sum_{j=1}^{t} y_i y_j \alpha_i \alpha_j \left(\langle x^i, x^j\rangle + \frac{\delta_{ij}}{2C}\right) - \sum_{i=1}^{t}\alpha_i \tag{3}$$

unter den Nebenbedingungen $y^T\alpha = 0$ und $0 \leq \alpha$. Die Modelle (2) und (3) unterscheiden sich durch die Art und Weise, wie Trainingsfehler während der Modellierung bewertet werden. Beim ersten Modell werden die Werte, um die jeder Punkt das vorgegebene Ziel verfehlt, einfach summiert, wohingegen das zweite Modell die Beträge der Abweichungen vor der Summierung noch quadriert. Der Parameter $C > 0$ bestimmt dann, wie stark die akkumulierten Fehler insgesamt gewichtet werden. In einem komplexeren Modell, welches Fehler in den Klassen unterscheidet, wird der Parameter aufgesplittet in $C = C^+$ für alle Trainingspunkte $i$ mit $y_i = 1$ sowie $C = C^-$ für die Trainingspunkte mit $y_i = -1$.

Der Lösungsvektor $\alpha^*$ führt in beiden Fällen zu einer linearen Hypothese

$$h(x) = \operatorname{sgn}\left(\sum_{i=1}^{t} y_i \alpha_i^* \langle x^i, x\rangle + b\right) . \tag{4}$$

Hierbei ist $\operatorname{sgn}(\cdot)$ eine modifizierte Signumfunktion, die definiert ist als

$$\operatorname{sgn}(a) = \begin{cases} 1 & a \geq 0 \\ -1 & a < 0 \end{cases} \qquad (a \in \mathbb{R}) \ .$$

Das zentrale Element in (2) bzw. (3) und (4) ist stets das Skalarprodukt zwischen zwei Punkten im Datenraum $\mathcal{X}$ und erlaubt uns, sogenannte Kerne zu verwenden. Die Idee dabei ist, dass die Hypothese (4) zu einer nichtlinearen Funktion wird, falls alle verwendeten Datenpunkte des Raumes $\mathcal{X}$ vor dem Training zunächst nichtlinear in einen hochdimensionalen Raum abgebildet werden. Sei dazu

$$\Phi : \mathcal{X} \to \mathcal{F} \tag{5}$$

eine Abbildung vom Datenraum in einen Hilbertraum $\mathcal{F}$ mit Dimension $d$, den sogenannten Merkmalsraum. $\Phi$ wird auch als Merkmalsabbildung bezeichnet. Dann werden (2) und (3) zu

$$\min_{\alpha\in\mathbb{R}^t} \quad \frac{1}{2}\sum_{i=1}^{t}\sum_{j=1}^{t} y_i y_j \alpha_i \alpha_j \langle \Phi(x^i), \Phi(x^j)\rangle_{\mathcal{F}} - \sum_{i=1}^{t}\alpha_i \tag{6}$$

bzw.

$$\min_{\alpha \in \mathbb{R}^t} \quad \frac{1}{2} \sum_{i=1}^{t} \sum_{j=1}^{t} y_i y_j \alpha_i \alpha_j \left( \langle \Phi(x^i), \Phi(x^j) \rangle_{\mathcal{F}} + \frac{\delta_{ij}}{2C} \right) - \sum_{i=1}^{t} \alpha_i \tag{7}$$

und die Hypothese hat die Form

$$h(x) = \operatorname{sgn} \left( \sum_{i=1}^{t} y_i \alpha_i^* \langle \Phi(x^i), \Phi(x) \rangle_{\mathcal{F}} + b \right) \ . \tag{8}$$

An dieser Stelle kann man dann Kernfunktionen zum Ersetzen der Skalarprodukte verwenden. Kernfunktionen sind Funktionen $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ für die gilt [SS02]

$$K(x^i, x^j) = \langle \Phi(x^i), \Phi(x^j) \rangle_{\mathcal{F}} \quad \forall x^i, x^j \in \mathcal{X} \ . \tag{9}$$

Für zwei beliebige Punkte des Datenraumes soll der Wert der Kernfunktion genau dem Wert des Skalarproduktes der beiden transformierten Punkte im Merkmalsraum $\mathcal{F}$ entsprechen. Übrigens ist nicht jede Abbildung $K$ mit dieser Eigenschaft auch tatsächlich ein Kern im Sinne der SVM-Theorie [Mer09]. Eine symmetrische Funktion $K(x^i, x^j)$ ist genau dann eine Kernfunktion, wenn für einen beliebigen Trainingsdatensatz $\mathcal{T}$ sowie reelle Zahlen $\gamma_1, \ldots, \gamma_t$ die Ungleichung

$$\sum_{i=1}^{t} \sum_{j=1}^{t} \gamma_i \gamma_j K(x^i, x^j) \geq 0 \tag{10}$$

erfüllt ist [Gen01]. Die Abbildung $K$ muss also positiv semidefinit[1] sein. Eine Reihe von Standard-Kernen ist bekannt [CST00]. In dieser Arbeit arbeiten wir mit

- dem Gauß-Kern (RBF-Kern)

$$K(x^i, x^j) := e^{-\frac{\|x^i - x^j\|^2}{2\sigma^2}} \quad (\sigma \in \mathbb{R}_+) \tag{11}$$

  und
- dem Slater-Kern

$$K(x^i, x^j) := e^{-\frac{\|x^i - x^j\|}{2\sigma^2}} \quad (\sigma \in \mathbb{R}_+) \ . \tag{12}$$

Es ist zu beachten, dass die Lage und Wahl der Hyperebene nicht nur von den gegebenen Trainingsdaten, sondern zusätzlich auch von verschiedenen Parametern in den Nebenbedingungen des Optimierungsproblems und in den Kernfunktionen abhängt. Es ist also ein Teil der Arbeit mit Support-Vektor-Maschinen, diese Parameter anhand der vorliegenden Trainingsdaten in einer vorgeschobenen Phase zu optimieren. Dabei stellt sich die Frage, wann ein Parametertupel hinreichend gut ist und für die anschließende Erstellung der Klassifikationsfunktion verwendet werden sollte. Dazu gibt es während der Parameteroptimierung

[1] [Gen01] nennt diese Eigenschaft *positive definite*.

interne Lern- und Testphasen. Um die dabei entstehenden Ergebnisse miteinander zu vergleichen, benötigt man charakteristische Kennzahlen, die einen internen Test beschreiben. Von dieser Problematik ausgehend definieren wir zunächst einige Bezeichnungen.

- Positive Punkte (pos), die korrekt in die Klasse 1 eingeordnet werden, heißen richtig positiv und werden im Folgenden mit rp bezeichnet.
- Analoges gilt für negative Punkte (neg), die korrekt in die Klasse $-1$ eingeordnet werden. Diese heißen richtig negativ und werden im Folgenden mit rn bezeichnet.
- Die Sensitivität (se) ist ein Maß für den Anteil richtig positiver Punkte und wird definiert durch
$$\mathrm{se} := \frac{\mathrm{rp}}{\mathrm{pos}} \ . \tag{13}$$
- Die Spezifität (sp) ist analog der Anteil richtig negativer Punkte von allen negativen Punkten und wird definiert durch
$$\mathrm{sp} := \frac{\mathrm{rn}}{\mathrm{neg}} \ . \tag{14}$$
- Die Genauigkeit (ge) gibt den Gesamtanteil aller richtig eingeordneter Punkte an und wird berechnet mittels
$$\mathrm{ge} := \frac{\mathrm{rp} + \mathrm{rn}}{\mathrm{pos} + \mathrm{neg}} \ . \tag{15}$$
- Die Präzision (pr) ist definiert durch
$$\mathrm{pr} := \frac{\mathrm{rp}}{\mathrm{rp} + \mathrm{fp}} \tag{16}$$
und besagt, wieviele der als positiv klassifizierten Testpunkte tatsächlich positiv sind.

## 2 Gütemaße

Nun sind wir in der Lage, Gütemaße zu definieren. Sie spielen eine wichtige Rolle bei der Optimierung von Parametern für Support-Vektor-Maschinen. Um eine Parameterkombination automatisch einschätzen zu können, muss einer Kombination anhand der mit ihr erzielten Trainingsergebnisse eine Bewertung zugeordnet werden. Diese Bewertung soll dann dazu dienen, die verschiedenen Parameterkombinationen untereinander zu ordnen. Gerade bei unausgeglichenen Datensätzen besteht ein Konfliktpotential zwischen den unterschiedlichen Bewertungen. Eine Bewertung ist nämlich nicht eindeutig, sondern kann sich je nach Klassifikationsziel unter anderem darin unterscheiden, ob sie Sensitivität oder Genauigkeit höher gewichtet. Als Beispiele sind daher (13) und (15) bereits einfache Gütemaße. Sie sind jedoch verbesserungswürdig, da sie nur gewisse Extreme betrachten, nämlich hohe Sensitivität oder hohe Genauigkeit im

Test. Zur Parameteroptimierung kann jedoch stets nur ein einzelnes Kriterium herangezogen werden, insbesondere wenn ein numerischer Optimierer diese Aufgabe übernehmen soll [GK04]. Gerade die Genauigkeit, die im Allgemeinen zur Qualitätsmessung herangezogen wird, birgt ein hohes Risiko bei der Modellierung von Klassifikationsfunktionen auf unausgeglichenen Datensätzen, bei denen Fehler in den verschiedenen Klassen vom Benutzer nicht gleich bewertet werden. Hier sollte das verwendete Qualitätsmaß angepasst werden. Daher betrachten wir im Folgenden kombinierte Gütemaße, die sich für kostensensitive Parameteroptimierung eignen.

– Das F-Maß

$$\mathrm{FM} := \frac{2 \cdot \mathrm{pr} \cdot \mathrm{se}}{\mathrm{pr}+\mathrm{se}} \tag{17}$$

ist ein harmonisches Maß aus Sensitivität und Präzision [Ren04].

– Das gewichtete F-Maß

$$\mathrm{FM}_\beta := \frac{\left(1+\beta^2\right) \mathrm{pr} \cdot \mathrm{se}}{\beta^2 \mathrm{pr}+\mathrm{se}} \quad (\beta > 0) \tag{18}$$

stellt über den Parameter $\beta$ eine Möglichkeit zur Verschiebung des harmonischen Gewichtes dar [EL05]. Für dieses Maß sei angemerkt, dass $\mathrm{FM} \equiv \mathrm{FM}_1$ gilt und folgendes Grenzwertverhalten beobachtet werden kann:

$$\lim_{\beta \to 0} \mathrm{FM}_\beta \equiv \mathrm{pr}$$

und

$$\lim_{\beta \to \infty} \mathrm{FM}_\beta \equiv \mathrm{se} \; .$$

– Die gewichtete Summe von Sensitivität und Präzision ist definiert als

$$\mathrm{S}_\beta := \beta \cdot \mathrm{se} + (1-\beta) \cdot \mathrm{pr} \quad (\beta \in (0,1)) \; . \tag{19}$$

Die Maße (17) bis (19) sind Beispiele dafür, wie man Testresultate sensitiv und dennoch genau bewerten kann. Sie haben gemeinsam, dass sie normiert in $[0,1]$ sind und den Wert 1 annehmen, falls eine Klassifikation komplett fehlerfrei war, also jedes Element genau in die zugehörige Klasse einordnet wurde. Diese Normierung ist im Rahmen der Vergleichbarkeit sinnvoll. Ein Gütemaß, welches nicht in diese Kategorie fällt, ist der Anreicherungsfaktor. Er besagt, um welchen Faktor die Dichte positiver Punkte in den als positiv klassifizierten Testpunkten im Gegensatz zum Gesamttestdatensatz gestiegen ist [KE04].

Neben den bisher vorgestellten Gütemaßen wird nun zusätzlich noch ein weiteres Maß eingeführt, welches auf der Receiver-Operating-Characteristic (ROC) basiert. Dies ist ein Verfahren, das ursprünglich in der Signalerkennung verwendet wurde und hilfreich zur graphischen Darstellung und zum Vergleich von

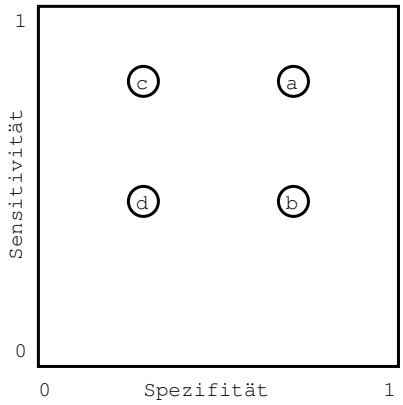


**Abbildung 1.** Beispiel für ROC-Graphik.

Klassifikationsergebnissen ist [Faw03]. Im ROC-Raum werden Sensitivität und Spezifität aufgetragen. Jedes Klassifikationsergebnis ergibt einen einzelnen ROC-Punkt, sodass man unterschiedliche Klassifikationen durch die Lage der Punkte untereinander vergleichen kann. Eine solche Graphik ist in Abbildung 1 dargestellt. Das Ergebnis a ist dabei eindeutig besser als die Ergebnisse b, c und d, jedoch kann man b und c nicht vergleichen. Ausgehend von dem Einheitsquadrat der ROC-Visualisierung entwickeln wir ein neues Gütemaß zur Bewertung von Testergebnissen. Anschaulich soll das neue Maß den Abstand zum optimalen Punkt im ROC-Raum darstellen. Der optimale Punkt hat eine fehlerfreie Klassifikation ($\mathrm{se} = 1.0,\ \mathrm{sp} = 1.0$). Für einen Abstand von 0 (optimaler Punkt) muss das Maß 1 und beim maximalen Abstand genau 0 sein. Damit wird bereits ein Vorteil gegenüber anderen Gütemaßen erkennbar. Dieses Maß soll nur dann den Wert 0 haben, wenn alle Punkte falsch eingeordnet werden. Somit gibt es auch für sehr schlechte Tests immer noch eine Vergleichbarkeit untereinander. Ausgehend vom euklidischen Abstand eines Punktes zum optimalen Eckpunkt des Quadrates in der rechten oberen Ecke definieren wir ein normiertes Abstandsmaß als

$$\mathrm{M} := \sqrt{\frac{\mathrm{se}^2 + \mathrm{sp}^2}{2}} \tag{20}$$

Selbstverständlich soll das neue Maß so flexibel sein, dass man die Klassen unterschiedlich gewichten kann, denn nicht immer sind Fehler in beiden Klassen von gleicher Bedeutung. Anschaulich gesehen ist dies zu vergleichen mit einer Streckung oder Stauchung des betrachteten Quadrates hin zu einem Rechteck, dessen Diagonale mit 1 normiert wird. Dies kann man direkt in die Gleichung für das neue Gütemaß einbauen. Wir erhalten schließlich das neue Gütemaß in der gewichteten Form als

$$\mathrm{M}_{\beta,\gamma} := \sqrt{\frac{\beta \cdot \mathrm{se}^2 + \gamma \cdot \mathrm{sp}^2}{\beta + \gamma}}\ . \tag{21}$$

Mittels $\delta := \beta/\gamma$ kann man (21) vereinfachen und es ergibt sich das neue gewichtete Abstandsmaß

$$\mathrm{M}_\delta = \sqrt{\frac{\delta \cdot \mathrm{se}^2 + \mathrm{sp}^2}{\delta + 1}} \ . \tag{22}$$

Für die Grenzwerte gelten

$$\lim_{\delta \to 0} \mathrm{M}_\delta = \mathrm{sp}$$

und

$$\lim_{\delta \to \infty} \mathrm{M}_\delta = \mathrm{se} \ .$$

## 3 Tests zur Parameteroptimierung

Die Qualität der im letzten Abschnitt vorgestellten Gütemaße wird anhand des auf [HBM98] frei verfügbaren Brustkrebs-Datensatzes der University of Wisconsin Hospitals, Madison [MW90] untersucht. Die Anzahl der Datenpunkte beträgt 699 bei 9 relevanten Attributen. Die Gewebeproben werden in gut- und bösartig unterteilt. Der Datensatz ist unausgeglichen, denn das Verhältnis der Anzahl gutartiger und bösartiger Punkte beträgt 2 zu 1. Aus diesem Grund gehen wir dazu über, die Gruppe der bösartigen Gewebe als positive Klasse zu definieren. Dadurch erhalten wir ein in zweierlei Hinsicht unausgeglichenes Optimierungsproblem:

1. Die Anzahl positiver Punkte ist signifikant kleiner, als die Anzahl negativer Punkte.
2. Ein falsch negativ klassifizierter Punkt ist mit deutlich höheren Kosten zu bewerten als ein falsch positiver Punkt.

Das Augenmerk liegt jetzt also eindeutig auf der Klasse 1 und im Folgenden werden wir untersuchen, wie die verschiedenen Gütemaße auf dieses Problem reagieren und welche Ergebnisse unseren Vorstellungen entsprechen. Von den insgesamt 699 Punkten in der Datenbank benutzen wir 324 zur Parameteroptimierung und für das spätere Training; die übrigen 375 Punkte werden für einen unabhängigen Test zur Seite gelegt. Die Parameteroptimierung basiert auf mehrmaliger Durchführung einer zehnfachen Kreuzvalidierung [Eit03]. Dabei werden aus den 324 Punkten zur Parameteroptimierung für eine beliebige Parameterkombination zehnmal jeweils 10% der Punkte für eine Kontrolle zurückgehalten und mit den anderen 90% wird eine Support-Vektor-Maschine trainiert. Dann werden die zurückgehaltenen Punkte klassifiziert und die dabei entstandenen Fehler gespeichert. Anhand der kumulierten Fehlerzahl beider Klassen aus allen 10 Tests wird dann das Gütemaß für diese Parameterkombination berechnet. Die Parameteroptimierung funktioniert dann so, dass alle gewünschten Parametertupel diesen Prozess durchlaufen. Die Parameterkombination mit dem besten Ergebnis bezüglich des gewählten Gütemaßes wird dann für ein endgültiges Training mit allen 324 Punkten verwendet. Es entsteht eine Klassifikationsfunktion, mit deren Hilfe die Testpunkte klassifiziert werden. Das vollständige Verfahren

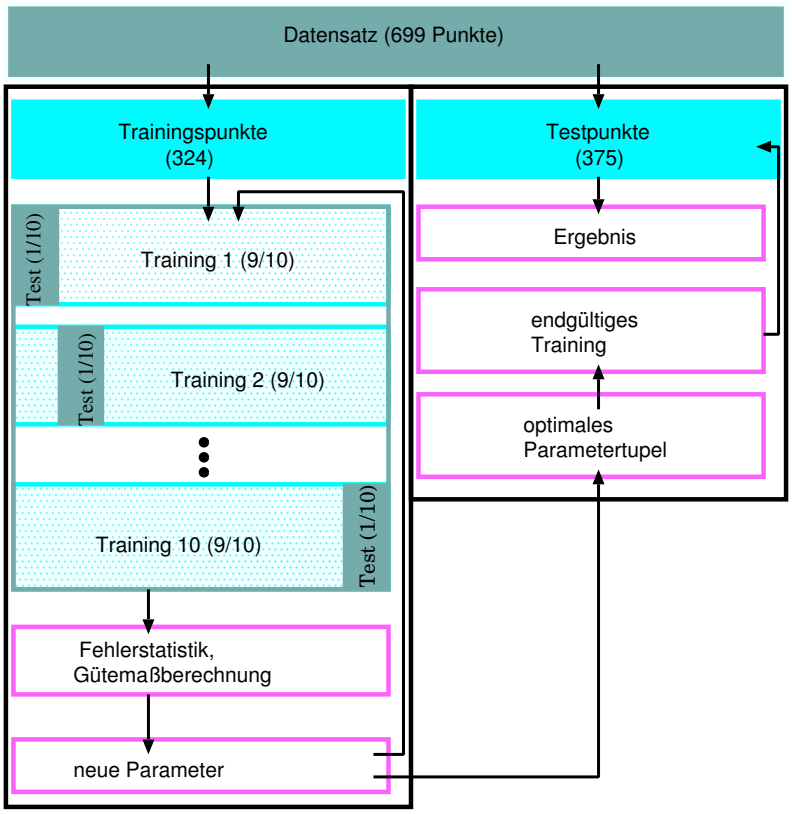


**Abbildung 2.** Parameteroptimierung mittels zehnfacher Kreuzvalidierung und finaler Test.

ist in Abbildung 2 dargestellt. Es stellt sich die Frage, für welches Gütemaß während der Kreuzvalidierung die endgültige Klassifikation am besten ist.

Unsere Tests zur Parameteroptimierung sind nach folgendem Schema durchgeführt worden: Die verschiedenen Gütemaße sind unabhängig voneinander zur Parameteroptimierung eingesetzt worden. Optimiert wurden die Parameter $\sigma$, $C^+$ und $C^-$, wobei wir für jeden Parameter 7 mögliche Werte zugelassen haben. Insgesamt wurde also auf einem $7 \times 7 \times 7$-Grid optimiert, was insgesamt 343 Möglichkeiten entspricht und zu 3773 SVM-Trainingsphasen (jeweils $10 + 1$ pro Kombination) führte. Es wurden sowohl der SMO-Algorithmus (*sequential minimal optimization*) [Pla99] für das Modell (6), als auch der NP-Algorithmus (*nearest point algorithm*) [KSBM00] für (7) untersucht. Zusätzlich sind zwei Kerne zur Verfügung gestellt worden – der Gauß-Kern (11) und der Slater-Kern (12). Mit 9 getesteten Gütemaßen ergaben sich insgesamt $3773 \cdot 4 \cdot 9 = 135828$ SVM-Trainingsphasen. Für alle Algorithmen und Kerne sind die besten Parameterkombinationen mittels der 9 Gütemaße bestimmt worden. Diese wurden

dann für das finale Training und den Test benutzt. Wir vergleichen im Folgenden ausschließlich die Testergebnisse, ohne auf die Phänomene während der Optimierung einzugehen. Es ist intuitiv klar, dass stark sensitive Maße andere Parameterwerte auswählen als die allgemeineren Maße. Wir untersuchen, wie stark sich dadurch Ergebnisse unterscheiden.

Bei unseren Tests lag der Akzent darauf, das bekannte F-Maß, welches im Zusammenhang mit Support-Vektor-Maschinen mehrfach zu guten Ergebnissen führte [MKO03,EL05], mit unserem neuen Abstandsmaß (22) zu vergleichen. Dazu haben wir bei beiden Maßen die Parameter $\beta$ und $\delta$ variiert. Zusätzlich zum Standardwert 1.0, wurden die Werte 0.5 für einen geringeren Einfluss der Sensitivität und 2.0 für einen stärkeren Einfluss getestet. Zur Vervollständigung der Tests haben wir 3 weitere der vorgestellten Gütemaße in die Studie eingebunden. Die Ergebnisse der Tests sind in den Tabellen 1 bis 4 aufgelistet. Auf den ersten Blick fällt auf, dass die Genauigkeit als Entscheidungsgrundlage bei der Parameteroptimierung dann aber nicht zu den besten Genauigkeiten im Test geführt hat.

| Gütemaß | $FM_{0.5}$ | FM | $FM_2$ | se | $S_{0.8}$ | ge | $M_{0.5}$ | $M_1$ | $M_2$ |
|---|---|---|---|---|---|---|---|---|---|
| $\sigma$ | 2.0 | 2.0 | 0.5 | 50.0 | 5.0 | 2.0 | 2.0 | 2.0 | 5.0 |
| $C^+$ | 0.5 | 20.0 | 0.5 | 50.0 | 10.0 | 20.0 | 2.0 | 20.0 | 10.0 |
| $C^-$ | 20.0 | 1.0 | 5.0 | 2.0 | 0.5 | 1.0 | 50.0 | 1.0 | 0.5 |
| Fehler 1. Art | 30 | 9 | 0 | 0 | 2 | 9 | 34 | 9 | 2 |
| Fehler 2. Art | 5 | 6 | 147 | 33 | 3 | 6 | 6 | 6 | 3 |
| Genauigkeit | 0.91 | 0.96 | 0.61 | 0.91 | 0.99 | 0.96 | 0.89 | 0.96 | 0.99 |

**Tabelle 1.** Erzielte Testergebnisse für den SMO-Algorithmus mit Gauß-Kern.

| Gütemaß | $FM_{0.5}$ | FM | $FM_2$ | se | $S_{0.8}$ | ge | $M_{0.5}$ | $M_1$ | $M_2$ |
|---|---|---|---|---|---|---|---|---|---|
| $\sigma$ | 2.0 | 2.0 | 2.0 | 50.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 |
| $C^+$ | 0.5 | 2.0 | 10.0 | 50.0 | 10.0 | 2.0 | 2.0 | 2.0 | 10.0 |
| $C^-$ | 10.0 | 0.5 | 0.5 | 1.0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Fehler 1. Art | 40 | 9 | 7 | 0 | 7 | 9 | 9 | 9 | 7 |
| Fehler 2. Art | 5 | 8 | 15 | 246 | 15 | 8 | 8 | 8 | 15 |
| Genauigkeit | 0.88 | 0.96 | 0.94 | 0.34 | 0.94 | 0.96 | 0.96 | 0.96 | 0.94 |

**Tabelle 2.** Erzielte Testergebnisse für den NP-Algorithmus mit Gauß-Kern.

Klammert man das gewichtete F-Maß mit $\beta = 2$ und die Sensitivität aus, so liegt die erzielte Genauigkeit bei allen anderen Gütemaßen zwischen 88% und 99% und ist daher insgesamt positiv zu bewerten. Die Anzahl der falsch

| Gütemaß | $FM_{0.5}$ | FM | $FM_2$ | se | $S_{0.8}$ | ge | $M_{0.5}$ | $M_1$ | $M_2$ |
|---|---|---|---|---|---|---|---|---|---|
| $\sigma$ | 5.0 | 2.0 | 10.0 | 50.0 | 10.0 | 2.0 | 2.0 | 1.0 | 10.0 |
| $C^+$ | 2.0 | 2.0 | 50.0 | 50.0 | 50.0 | 2.0 | 2.0 | 50.0 | 50.0 |
| $C^-$ | 20.0 | 20.0 | 5.0 | 5.0 | 5.0 | 20.0 | 20.0 | 5.0 | 5.0 |
| Fehler 1. Art | 38 | 7 | 4 | 0 | 4 | 23 | 23 | 16 | 4 |
| Fehler 2. Art | 1 | 6 | 5 | 246 | 5 | 1 | 1 | 5 | 5 |
| Genauigkeit | 0.90 | 0.97 | 0.98 | 0.34 | 0.98 | 0.94 | 0.94 | 0.94 | 0.98 |

**Tabelle 3.** Erzielte Testergebnisse für den SMO-Algorithmus mit Slater-Kern.

| Gütemaß | $FM_{0.5}$ | FM | $FM_2$ | se | $S_{0.8}$ | ge | $M_{0.5}$ | $M_1$ | $M_2$ |
|---|---|---|---|---|---|---|---|---|---|
| $\sigma$ | 10.0 | 1.0 | 2.0 | 50.0 | 2.0 | 1.0 | 1.0 | 1.0 | 2.0 |
| $C^+$ | 5.0 | 2.0 | 50.0 | 50.0 | 50.0 | 50.0 | 50.0 | 2.0 | 50.0 |
| $C^-$ | 50.0 | 0.5 | 1.0 | 2.0 | 1.0 | 20.0 | 20.0 | 0.5 | 1.0 |
| Fehler 1. Art | 34 | 7 | 7 | 0 | 7 | 15 | 15 | 7 | 7 |
| Fehler 2. Art | 1 | 6 | 5 | 246 | 5 | 5 | 5 | 6 | 5 |
| Genauigkeit | 0.91 | 0.97 | 0.97 | 0.34 | 0.97 | 0.95 | 0.95 | 0.97 | 0.97 |

**Tabelle 4.** Erzielte Testergebnisse für den NP-Algorithmus mit Slater-Kern.

negativen Punkte schwankt hingegen zwischen 0 und 40, was deutliche Unterschiede in der Sensitivität aufzeigt. Es ist allerdings zu beachten, dass die Tests mit mehr als 23 falsch negativen Punkten von Gütemaßen ermittelt wurden, die der Sensitivität wenig Bedeutung zugesprochen haben. Daher ist eine geringere Sensitivität verständlicherweise auch zu erwarten gewesen. Aus diesem Grund sollten diese Tests in einer Gesamtbewertung der Sensitivität ausgeklammert werden. Weiterhin fällt auf, dass bei einer Optimierung der Parameter mittels Sensitivität dann auch Ergebnisse mit 100% richtig eingeordneten positiven Punkten zu verzeichnen sind. Allerdings wird in diesem Fall die Genauigkeit viel zu gering, sie liegt nämlich unter 60%. Andererseits erlangt man durch Verwendung der Genauigkeit als Gütemaß während der Kreuzvalidierung keine Ergebnisse mit optimaler Genauigkeit. Mit 94–96% Genauigkeit kann man die dabei erzielten Ergebnisse im Vergleich zu den anderen Gütemaßen als eher mittelmäßig bezeichnen. Hervorzuheben ist die gewichtete Summe aus Sensitivität und Präzision. Die Genauigkeit lag bei den verschiedenen Settings zwischen 94% und 99% und in allen Testreihen wurden besonders gute Resultate bei der Sensitivität erzielt, die von keinem vergleichbaren Maß übertroffen wurden. Damit ist diese Summe als Gütemaß empfehlenswert und sollte noch einmal eingehend untersucht werden. Vergleicht man schließlich das bekannte gewichtete F-Maß mit dem neuen Abstandsmaß mit jeweils äquivalenter Gewichtung, so erkennt man einige Vorteile des neuen Gütemaßes. Besonders für die Testreihen unter dem Slater-Kern zeichnet sich das Abstandsmaß durch höhere Sensitivität als das zugehörige gewichtete F-Maß aus. Desweiteren erzielten die mit diesem Ab-

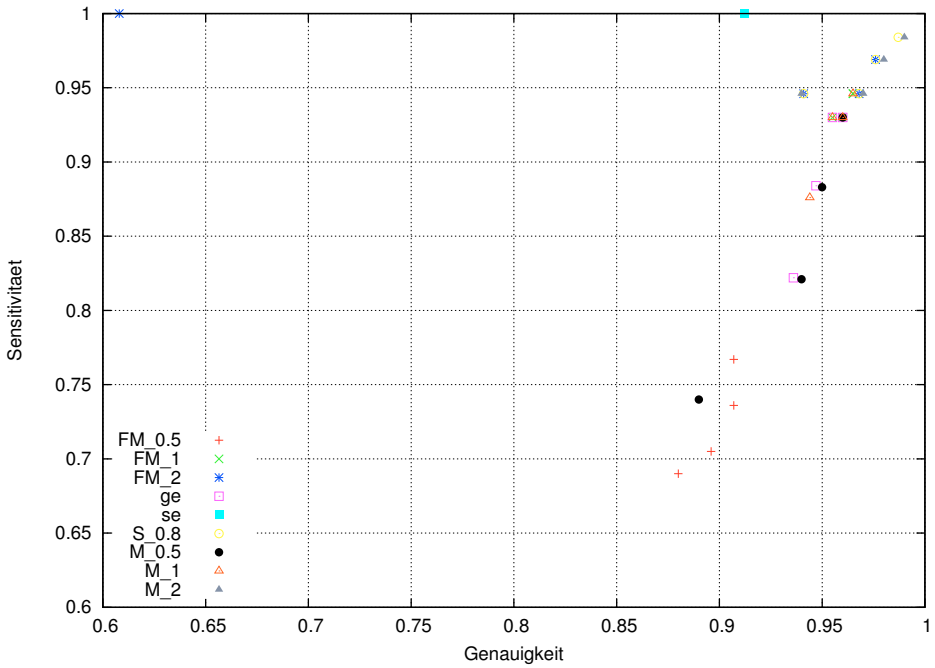


**Abbildung 3.** Sensitivität und Genauigkeit unter verschiedenen Gütemaßen.

standsmaß ausgewählten Parameterkombinationen sehr gute Gesamtergebnisse. Nur in einem von zwölf Fällen ist die Genauigkeit der durch das gewichtete F-Maß ausgewählten Parameterkombination besser.

Zur Veranschaulichung der Ergebnisse sind in Abbildung 3 die erzielten Werte für Sensitivität und Genauigkeit noch einmal graphisch dargestellt. Jeder Datensatz besteht aus 4 Punkten. Obwohl es eine starke Konzentration von Punkten in der Nähe des Randes gibt, läßt sich mit Hilfe der Abbildung eine weitere wichtige Aussage treffen. Es fällt auf, dass das F-Maß mit $\beta = 0.5$ zu den schlechtesten Ergebnissen führte. Alle vier Punkte, die die Testreihe darstellen, liegen wesentlich weiter vom (optimalen) Eckpunkt entfernt als beinahe alle anderen Punkte. Das F-Maß reagiert auf Änderung des Parameters $\beta$ sehr stark, sodass die Nutzung des flexiblen F-Maßes in der Praxis kritisch erscheint. Umso erfreulicher sind die Ergebnisse, die wir mit unserem neuen Gütemaß erzielen konnten.

## 4 Zusammenfassung

Wir haben in dieser Arbeit die Auswirkungen des Gütemaßes für die Parameteroptimierung von Support-Vektor-Maschinen untersucht. Dazu wurden verschiedene Gütemaße vorgestellt und in unabhängigen Testreihen für einen interessanten Datensatz getestet und verglichen. Erkennbar war die unterschiedliche

Auswahl der optimalen Parameter je nach ausgewähltem Gütemaß. Es war deutlich erkennbar, wie das Verfahren auf Vorlieben, die durch jedes Maß und einen optionalen Parameter gegeben waren, reagierte. Es konnte gezeigt werden, dass gewünschte Trends in Ergebnissen wiedergefunden werden, wenn ein entsprechendes Gütemaß schon während der Parametersuche eingesetzt wird. Es gibt Maße, die stabile Parameterkombinationen auswählen, mit denen im Test eine moderate Gesamtfehleranzahl zu verzeichnen ist. Der gegensätzliche Trend, also extreme Maße und Parameter können im Test zu enorm vielen Fehlern führen. Besonders positiv hervorzuheben ist das kombinierte Maß aus Sensitivität und Präzision mit dem Parameterwert 0.8, der für ein Verhältnis $4:1$ von Sensitivität zu Präzision steht. Mit diesem Maß wurde in jeder Testreihe im Test das beste Ergebnis im Hinblick auf die Genauigkeit erzielt. Aber auch das neue Abstandsmaß führte zu sehr guten und insbesondere stabilen Ergebnissen.

## Literatur


[CST00] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[Eit03] T. Eitrich. Support-Vektor-Maschinen und ihre Anwendung auf Datensätze aus der Forschung. Berichte des Forschungszentrums Jülich JUEL-4096, Forschungszentrum Jülich, 2003.

[EL05] T. Eitrich and B. Lang. Efficient optimization of support vector machine learning parameters for unbalanced datasets. *Journal of Computational and Applied Mathematics*, 2005. in press.

[Faw03] T. Fawcett. ROC graphs: notes and practical considerations for researchers. Technical Report HPL-2003-4, HP Labs, 2003.

[Gen01] M. G. Genton. Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2:299–312, 2001.

[GK04] G. A. Gray and T. G. Kolda. APPSPACK 4.0: asynchronous parallel pattern search for derivative-free optimization. Sandia Report SAND2004-6391, Sandia National Laboratories, Livermore, CA, August 2004.

[HBM98] S. Hettich, C. L. Blake, and C. J. Merz. *UCI repository of machine learning databases*, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[KE04] A. Kless and T. Eitrich. Cytochrome p450 classification of drugs with support vector machines implementing the nearest point algorithm. In J. A. López, E. Benfenati, and W. Dubitzky, editors, *Knowledge Exploration in Life Science Informatics, International Symposium, KELSI 2004, Milan, Italy*, volume 3303 of *Lecture Notes in Computer Science*, pages 191–205. Springer, 2004.

[KSBM00] S.S. Keerthi, S.K. Shevade, C. Bhattacharyya, and K.R.K. Murthy. A fast iterative nearest point algorithm for support vector machine classifier design. *IEEE Transactions on Neural Networks*, 11:124–136, 2000.

[Mer09] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society, London A*, 209:415–446, 1909.

[MKO03] D. R. Musicant, V. Kumar, and A. Ozgur. Optimizing F-measure with support vector machines. In Ingrid Russell and Susan M. Haller, editors,

*Proceedings of the Sixteenth International Florida Artificial Intelligence Research Society Conference, May 12-14, 2003, St. Augustine, Florida, USA*, pages 356–360. AAAI Press, 2003.

[MW90] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming. *SIAM News*, 23:1–18, 1990.

[Pla99] John C. Platt. *Fast training of support vector machines using sequential minimal optimization*. MIT Press, 1999.

[Ren04] J. D. M. Rennie. Derivation of the F-measure. http://people.csail.mit.edu/∼jrennie/writing, 2004.

[SS02] B. Schölkopf and A. J. Smola. *Learning with kernels*. MIT Press, 2002.

[Vap95] V. Vapnik. *The Nature of statistical learning theory*. Springer Verlag, New York, 1995.