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

Interner Bericht

Eine Studie zur kostensensitiven Klassifikation unausgeglichener Datensätze mit Support-Vektor-Maschinen

Sebastian Schnitzler, Tatjana Eitrich

FZJ-ZAM-IB-2006-05

März 2006

(letzte Änderung: 28.3.2006)

Eine Studie zur kostensensitiven Klassifikation unausgeglichener Datensätze mit Support-Vektor-Maschinen

Sebastian Schnitzler, Tatjana Eitrich

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

Zusammenfassung Support-Vektor-Maschinen sind ein mächtiges Verfahren zur Klassifikation von Daten. Dementsprechend viele verschiedene Parameter sind bei der Anwendung zu setzen und zu beachten. Wir beschäftigen uns mit den Auswirkungen, die diese Parameter auf die Ergebnisse haben und suchen charakteristische Punkte im Parameterraum. Dazu werden unterschiedliche Kerne und Algorithmen vorgestellt und die zugehörigen Parameter in den Modellen modifiziert. Zusätzlich werden mit der Schwellwertverschiebung und dem Oversampling Methoden zur weiteren Verbesserung der Ergebnisse vorgestellt und bewertet.

1 Einleitung

Support-Vektor-Maschinen (SVM's) sind ein modernes Verfahren zur Klassifikation von Daten, die auf dem einfachen Gedanken der Trennung von Klassen durch Hyperebenen basieren [CST00]. In Abbildung 1 ist das in dieser Arbeit betrachtete Problem der binären Klassifikation dargestellt. Dabei gibt es genau zwei Klassen von Punkten, die wir im Folgenden stets als positiv und negativ bezeichnen werden. Bei allen Vorteilen, die das Verfahren mit sich bringt, stellt man dennoch fest, dass die Wahl der Modellparameter eine große Herausforderung ist. In dieser Arbeit führen wir beispielhaft an einem Datensatz eine SVM-Parameterstudie durch. Dabei werden verschiedene Modelle und Algorithmen untersucht, für die dann der Einfluß der Parameterwahl analysiert wird. Ein wichtiger Aspekt der Tests ist die Frage nach markanten Punkten im Parameterraum. Dazu gehören Unstetigkeiten, d.h. bei kleiner Änderung eines Wertes ändert sich das Ergebnis stark, sowie Sättigungsgebiete, wo eine weitere Änderung eines Wertes keinen Einfluß auf die Klassifikation hat.

Die Methode der Support-Vektor-Maschinen gehört zu den Verfahren des überwachten Lernens. Basierend auf einem sogenannten Trainingsdatensatz, der aus einer endlichen Menge von Eingabe-Ausgabe-Paaren (EA-Paaren) besteht, wird die in den Daten enthaltene Funktionalität erkannt und als Modell festgehalten. Dieses Modell besteht im Allgemeinen aus einer linearen oder nichtlinearen Klassifikationsfunktion. Jedes EA-Paar besitzt ein Ausgabe-Label, welches dem Algorithmus zeigt, zu welcher Klasse der jeweilige Punkt gehört.

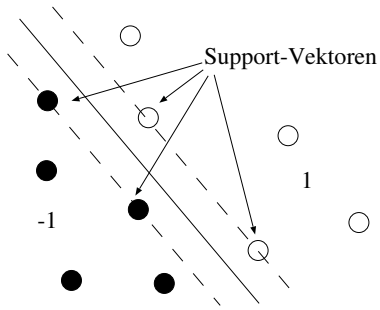

Abbildung 1. Datentrennung durch Hyperebene.

In dieser Arbeit untersuchen wir den Einfluß von Modellen und Parametern auf die Güte der erlernbaren Funktionen. Zu diesem Zweck ist es notwendig, eine Methode zur Bewertung von Ergebnissen zur Verfügung zu stellen. Dazu teilen wir den zur Verfügung stehenden Datensatz auf in Trainings- und Testdaten, siehe dazu Abbildung 2. Die Trainingsdaten werden ausschließlich zur Modellierung verwendet, wohingegen die Testdaten diesem Prozess fernbleiben müssen und für den Qualitätstest zur Verfügung stehen. Die den Tests zugrunde liegenden Maße werden in Kapitel 2 vorgestellt.

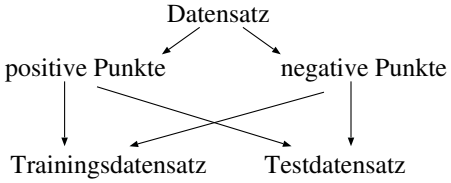

Abbildung 2. Aufteilung der Daten.

Ein wichtiger Aspekt dieser Arbeit ist die Frage nach kostensensitiver Klassifikation. Anwendungen sind häufig dadurch gekennzeichnet, dass eine Klasse stark unterbesetzt ist, die Anzahl der Punkte im Datensatz also signifikant kleiner ist als die Anzahl der Punkte in der anderen Klasse. Zusätzlich läßt sich feststellen, dass oft genau diese Klasse von verstärktem Interesse ist, was sich dadurch äußert, dass Fehlklassifikationen dieser Punkte um ein Vielfaches unerwünschter sind als im anderen Fall. Ein Beispiel dafür sind HIV-Tests [RKC+02]. Der untersuchte Datensatz fällt in dieses Raster und stellt den Anwender vor große Probleme, denn die klassischen Lernmethoden tendieren dazu, die seltenen positiven Punkte zugunsten der großen negativen Punktewolke zu ignorieren. Im schlimmsten Fall erhält man Ergebnisse ohne einen einzigen kor-

rekt klassifizierten positiven Punkt. Unsere Modellierung ist daher stark auf die Erkennung der wenigen positiven Punkte ausgerichtet und wir werden zeigen, zu welchen Kosten sich das mittels SVM's realisieren lässt.

2 Grundlagen der Parameterstudie

Wir beschreiben die für das Verständnis der in den folgenden Kapiteln dargestellten Ergebnisse notwendigen Grundlagen. Bei der Analyse der Ergebnisse unterscheiden wir zwischen den beiden Klassen, sodass es für Erfolg und Misserfolg jeweils zwei Ausprägungen gibt. Auf dieser Grundlage lassen sich Gütekriterien heranziehen. Die folgenden Definitionen gelten für diese Arbeit.

- Die Anzahl der positiven Punkte des Testdatensatzes bezeichen wir als pos.
- Die Anzahl der negativen Punkte des Testdatensatzes bezeichen wir als neg.
- Positive Punkte, die korrekt in die positive Klasse eingeordnet werden, heißen richtig positiv und werden im Folgenden mit rp bezeichnet.
- Negative Punkte, die korrekt in die negative Klasse eingeordnet werden, heißen richtig negativ und werden im Folgenden mit rn bezeichnet.
- Positive Punkte, die fälschlicherweise in die negative Klasse eingeordnet werden, heißen falsch negativ und werden im Folgenden mit fn bezeichnet.
- Negative Punkte, die fälschlicherweise in die positive Klasse eingeordnet werden, heißen falsch positiv und werden im Folgenden mit fp bezeichnet.
- Sensitivität (se) ist der Anteil richtig positiver Punkte in der Menge der positiven Testpunkte. Wir definieren also
$$\mathrm{se} := \frac{\mathrm{rp}}{\mathrm{pos}} \ .$$
- Spezifität (sp) ist der Anteil richtig negativer Punkte gemessen an der Gesamtzahl negativer Punkte und wird definiert als
$$\mathrm{sp} := \frac{\mathrm{rn}}{\mathrm{neg}} \ .$$
- Die Genauigkeit (ge) gibt den Gesamtanteil richtiger Testpunkte an, d.h.
$$\mathrm{ge} := \frac{\mathrm{rp} + \mathrm{rn}}{\mathrm{pos} + \mathrm{neg}} \ .$$
- Der Wert der Präzision (pr) kombiniert die Klassen und gibt an, welcher Anteil der als positiv klassifizierten Punkte tatsächlich positiv ist. Die Definition lautet
$$\mathrm{pr} := \frac{\mathrm{rp}}{\mathrm{rp} + \mathrm{fp}} \ .$$

– Das F-Maß (fm) ist ein Gütemaß, welches sich als Indikator für gute Klassifikationsergebnisse bewährt hat [EL05]. Es wird gebildet aus Sensitivität und Präzision nach der Vorschrift

$$\mathrm{fm} := 2\frac{\mathrm{pr} \cdot \mathrm{se}}{\mathrm{pr} + \mathrm{se}} \ .$$

Im Rahmen der Arbeit mit SVM's gibt es unterschiedliche Algorithmen, die man für das Training benutzen kann. Der Nearest-Point-Algorithmus (NPA) [KSBM00] und der Sequential-Minimal-Optimierer (SMO) [Pla99], welche zwei unterschiedliche Modelle repräsentieren, bilden die Grundlage dieser Arbeit. Eine ausführliche Beschreibung der Methoden ist in [Eit03] zu finden. Die Algorithmen setzen den sogenannten *softmargin*-Ansatz des SVM-Lernens um, bei dem Trainingsfehler bis zu einem bestimmten Maß toleriert werden, um eine möglichst glatte Funktion zu lernen. Die Idee dahinter ist, dass eine an reale Anwendungsdaten zu stark angepasste Funktion ohne Fehler auf den Trainingsdaten kaum geeignet ist, um neue Daten zu klassifizieren. Die beiden Algorithmen unterscheiden sich in der Art und Weise, wie Fehler bestraft werden. Das von SMO umgesetzte Verfahren summiert die Fehler einfach auf, wobei NPA eine quadratische Summierung beinhaltet. Das impliziert, dass Fehler in der Nähe der Trennfunktion, die kleiner als 1 sind, abgeschwächt werden, wohingegen weiter entfernte Punkte mit Abständen größer als 1 mit einem höheren Wert einfließen. Die beiden Algorithmen führen im Allgemeinen zu ähnlichen Ergebnissen. Eine vergleichende Studie ist jedoch nicht bekannt, sodass wir diesen Aspekt in dieser Arbeit kurz beleuchten werden.

Der untersuchte Datensatz gehört zur Familie der CYP-Datensätze [KE04]. Er enthält 263 Punkte, von denen 48 (19%) positiv sind. Aus dem Datensatz wurden 185 Trainingspunkte ausgewählt, deren Klassenverhältnis ebenfalls 19% zu 81% beträgt (35 positive, 150 negative Punkte). In den Testdatensatz wurden 13 positive und 65 negative Punkte, also insgesamt 78 Punkte aufgenommen. Der Originaldatensatz hat insgesamt 557 Variablen. Für unsere Tests wird eine reduzierte Variante verwendet. Mittels eines auf der Hauptkomponentenanalyse [Jol86] basierenden Verfahrens zur Variablenselektion, welches in [McC84] beschrieben ist, wurden drei Teildatensätze zu 5, 10 und 20 Variablen erstellt und analysiert. In dieser Arbeit präsentieren wir Ergebnisse für den Datensatz mit 10 Variablen. Verwendet wurde eine eigene SVM-Implementierung, die beispielsweise in [Eit03] beschrieben ist.

3 Kerne und ihre Parameter

Für die Anwendung von Support-Vektor-Maschinen ist es notwendig, eine sogenannte Kernfunktion zu wählen. Die Kernfunktion dient als Ähnlichkeitsmaß innerhalb des SVM-Optimierungsproblems. Während der Lernphase werden Ähnlichkeiten zwischen je zwei Trainingspunkten gemessen. In diesem Kapitel untersuchen wir drei Kerne. Dabei soll gezeigt werden, welche Ergebnisse sich mit diesen Kernen erzielen lassen und welchen Einfluß ihre internen Parameter auf die Güte der Klassifikation haben.

3.1 Gaußkern

Der Gaußkern ist definiert als [CST00]

$$K^G(\mathbf{x}, \mathbf{y}) := \exp\left(\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right). \tag{1}$$

Dabei ist der Parameter $\sigma \in \mathbb{R}_+$ variabel wählbar. In den folgenden Abschnitten werden Ergebnisse dokumentiert, die unter Verwendung dieses Kerns entstanden sind.

Variation des Parameters σ im Kern Die ersten Tests beschäftigen sich mit der Wahl des Parameters σ im Gaußkern. Für σ-Werte zwischen 0.5 und 20.0 wurde jeweils ein Training und ein Test durchgeführt. Tabelle 1 zeigt die Ergebnisse der Testläufe. Die Tabelle verdeutlicht den Bereich, in dem σ angesetzt werden sollte. Sinnvolle Resultate sind nur für σ-Werte zwischen 0.5 und 3.0 beim NPA und 0.5 und 1.5 beim SMO-Algorithmus zu erkennen. Große Werte führen wegen der falschen Einordnung von beinahe allen positiven Punkten zu einer Sensitivität von annähernd Null und somit zu einer für uns unbrauchbaren Klassifikation. Es herrscht zwar eine beinahe konstante Fehleranzahl, aber aus der zunehmend höheren Fehleranzahl bei der Einordnung der positiven Punkte folgt eine Verschlechterung der Ergebnisse. Damit wird deutlich, warum die Genauigkeit allein nicht als Gütemaß eingesetzt werden sollte, wenn man den Klassen unterschiedliche Bedeutungen zuspricht. Vergleicht man die beiden Algorithmen untereinander, so erkennt man bei beiden eine ähnliche Struktur in der Fehlerentwicklung, wobei man für den SMO-Algorithmus den Wert des Parameters σ niedriger ansetzen muß als für den NPA. Die fett gedruckten Werte

Algorithmus	NPA							SMO						
σ-Wert	0.5	1.0	2.0	3.0	5.0	10.0	15.0	0.5	1.0	1.5	1.75	2.0	3.0	5.0
Fehler ges.	12	15	13	13	13	13	13	14	13	13	13	13	13	13
Fehler fn	9	7	6	8	12	13	13	11	7	7	9	12	13	13
Fehler fp	3	8	7	5	1	0	0	3	7	6	4	1	0	0
Sensitivität	0.31	0.46	0.54	0.39	0.01	0.00	0.00	0.15	0.46	0.46	0.31	0.01	0.00	0.00
Spezifität	0.95	0.88	0.89	0.92	0.99	1.00	1.00	0.95	0.89	0.91	0.94	0.99	1.00	1.00
Genauigkeit	0.85	0.81	0.83	0.83	0.83	0.83	0.83	0.82	0.82	0.83	0.83	0.83	0.83	0.83
F-Maß	0.40	0.44	**0.52**	0.44	0.13	0.00	0.00	0.22	0.46	**0.48**	0.38	0.13	0.00	0.00

Tabelle 1. Ergebnisse für verschiedene σ-Werte des Gaußkerns.

zeigen die im Sinne unseres Gütemaßes besten Ergebnisse für beide Algorithmen. Dabei fällt, dass es sich jeweils um Ergebnisse mit der besten Sensitivität handelt.

Variation der Fehlergewichtung durch den Parameter C Als nächstes betrachten wir den Parameter $C \in \mathbb{R}_+$, der bestimmt, wie stark Fehler in den Trainingsdaten gewichtet werden sollen. Ein großer Wert für C bestraft Fehler mehr als ein niedrigerer Wert. Das Risiko dabei besteht jedoch in einer zu großen Anpassung an den Trainingsdatensatz. Dies nennt man Übertrainieren des Systems. Als Algorithmus wählen wir NPA. σ wird konstant mit 1.0 gewählt. Somit ist der Test für $C = 1.0$ bereits bekannt. Tabelle 2 führt die erzielten Resultate auf. Man erkennt, dass eine stärkere Gewichtung der Fehler zu besseren Resultaten führt. Dieser Effekt ist jedoch nicht durch beliebig hohe C-Werte weiter fortführbar. Stattdessen verbessert eine stärkere Fehlergewichtung die Resultate nur bis zu einer oberen Schranke. Die Tabelle zeigt, dass ab einem bestimmten Wert für C die Güte der Tests wieder sinkt. An diesem Punkt beginnt dann die Überanpassung der Klassifikationsfunktion an die Trainingspunkte. Bei $C = 10.0$ wird mit einem F-Maß von 0.60 bei 4 falsch negativen und 8 falsch positiven Punkten das beste Ergebnis erreicht. Bei weiter steigendem Wert für C bleibt die Sensitivität zwar gleich, aber eine Zunahme von falsch positiven Punkten führt zu einer Qualitätsverschlechterung in der Klassifikation. Die gleichen Tests wurden auch mit dem SMO-Algorithmus durchgeführt. Hierbei entsprechen die erzielten Resultate im Allgemeinen denen mit dem NP-Algorithmus. Erneut wird die höchste Sensitivität bei 4 falsch negativen Punkten erreicht und das beste Ergebnis bei zugleich 8 falsch negativen Punkten erzielt.

C-Wert	0.5	1.0	2.0	5.0	10.0	20.0	50.0	200.0
Fehler ges.	14	15	15	14	12	13	14	15
Fehler fn	8	7	7	5	4	4	4	4
Fehler fp	6	8	8	9	8	9	10	11
Sensitivität	0.39	0.46	0.46	0.62	0.69	0.69	0.69	0.69
Spezifität	0.91	0.88	0.88	0.86	0.88	0.86	0.85	0.83
Genauigkeit	0.82	0.81	0.81	0.82	0.85	0.83	0.82	0.81
F-Maß	0.42	0.44	0.44	0.53	0.60	0.58	0.56	0.55

Tabelle 2. Ergebnisse für verschiedene C-Werte der NPA-basierten SVM mit Gaußkern ($\sigma = 1$).

3.2 Slaterkern

Als zweiten Kern betrachten wir den Slaterkern. Der Slaterkern ist dem Gaußkern sehr ähnlich, und wird definiert durch

$$K^S(\mathbf{x}, \mathbf{y}) := \exp\left(\frac{\|\mathbf{x} - \mathbf{y}\|}{2\sigma^2}\right). \tag{2}$$

Wieder ist $\sigma \in \mathbb{R}_+$ ein variierbarer Parameter. Um eine Vergleichbarkeit mit dem Gaußkern zu schaffen, untersuchen wir ähnliche Werte wie zuvor.

Algorithmus	NPA					SMO			
σ-Wert	0.5	1.0	2.0	3.0	5.0	0.5	0.75	1.0	2.0
Fehler ges.	14	14	13	14	13	14	14	16	13
Fehler fn	11	7	6	11	13	12	8	12	13
Fehler fp	3	7	7	3	0	2	6	4	0
Sensitivität	0.16	0.46	0.54	0.16	0.00	0.16	0.39	0.16	0.00
Spezifität	0.95	0.89	0.89	0.95	1.00	0.97	0.91	0.94	1.00
Genauigkeit	0.82	0.82	0.83	0.82	0.83	0.82	0.82	0.80	0.83
F-Maß	0.22	0.46	0.52	0.22	0.00	0.13	0.42	0.11	0.00

Tabelle 3. Ergebnisse für verschiedene σ-Werte des Slaterkerns.

Variation des Parameters σ im Kern Die Ergebnisse der Variation des Parameters σ in Tabelle 3 sind denen bei der Untersuchung des Gaußkerns sehr ähnlich. Erneut wird für den NPA das beste Ergebnis bei der Wahl von $\sigma = 2.0$ mit 6 falsch negativen und 7 falsch positiven Punkten erzielt. Auch die restliche Fehlerentwicklung ist nahezu analog zum Gaußkern, sodass keine neuen Erkentnisse bei dieser Testreihe gewonnen worden konnten. Allerdings ist zu beachten, dass die Ergebnisse bei Verwendung des Slaterkerns im Allgemeinen etwas schlechter sind als mit dem Gaußkern.

Variation der Fehlergewichtung durch den Parameter C Genau wie für den Gaußkern betrachten wir als nächstes die Fehlergewichtung durch den Parameter C. Anders als beim Gaußkern sind die Auswirkungen einer Variation des

C-Wert	1.0	2.0	3.0	5.0	10.0	20.0	50.0	200.0
Fehler ges.	14	14	14	14	14	14	14	14
Fehler fn	7	6	6	6	6	6	6	6
Fehler fp	7	8	8	8	8	8	8	8
Sensitivität	0.46	0.54	0.54	0.54	0.54	0.54	0.54	0.54
Spezifität	0.89	0.88	0.88	0.88	0.88	0.88	0.88	0.88
Genauigkeit	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
F-Maß	0.46	0.50	0.50	0.50	0.50	0.50	0.50	0.50

Tabelle 4. Ergebnisse für verschiedene C-Werte der NPA-basierten SVM mit Slaterkern ($\sigma = 1$).

Parameters C in diesem Falle sehr moderat. Die Fehleranzahl bleibt konstant 14 und auch die Verteilungen der Fehler in den Klassen bleibt weitestgehend konstant. Zwar gibt es für $C > 1.0$ eine Verbesserung der Testresultate im Vergleich zur Standardwahl $C = 1.0$, doch sie ist sehr gering und verändert sich nicht durch weitere Variation von C. Die Wahl des Algorithmusses spielt dabei

keine Rolle. Sowohl NPA als auch der SMO-Algorithmus zeigen hier das gleiche Verhalten.

3.3 Polynomialkern

Als dritten und letzten Kern betrachten wir den Polynomialkern. Er ist definiert durch

$$K^P(\mathbf{x}, \mathbf{y}) := (1 + \langle \mathbf{x}, \mathbf{y} \rangle)^d \ . \tag{3}$$

$d \in \mathbb{N}_+$ ist frei wählbar. Der Polynomialkern arbeitet also mit einem anderen Parameter als der Gauß- und der Slaterkern. Statt der Weite des Kerns σ verändern wir daher den ganzzahligen Parameter d, der den Grad des Kernpolynoms bestimmt.

Polynomialgrad Der Polynomialgrad d des Kernes ist frei wählbar und in der Regel eine natürliche Zahl. Eine interessante Studie zum Polynomialkern, in der der Parameter d als rationale Zahl definiert ist, findet man in [RZI98]. Diese Idee wird in unserer Arbeit jedoch nicht weiter betrachtet, da die zugehörigen Tests hinter den Erwartungen zurückgeblieben sind. In Tabelle 5 werden

Algorithmus	NPA					SMO					
d-Wert	1	2	3	4	5	1	2	3	4	5	6
Fehler ges.	14	18	67	21	61	13	13	15	20	21	26
Fehler fn	13	8	7	6	1	13	11	6	8	3	6
Fehler fp	1	10	60	15	60	0	2	9	12	18	20
Sensitivität	0.00	0.39	0.46	0.54	0.92	0.00	0.15	0.54	0.38	0.77	0.54
Spezifität	0.99	0.85	0.08	0.77	0.08	1.00	0.97	0.86	0.82	0.72	0.69
Genauigkeit	0.82	0.77	0.14	0.73	0.22	0.83	0.83	0.81	0.74	0.73	0.67
F-Maß	0.00	0.36	0.15	0.40	0.28	0.00	0.24	0.48	0.33	0.49	0.35

Tabelle 5. Ergebnisse für verschiedene Grade des Polynomialkerns.

die Ergebnisse bei Verwendung des Polynomialkernes ohne Veränderung weiterer Parameter dargestellt. Betrachtet man die Ergebnisse nun untereinander, so fällt beim Nearest-Point-Algorithmus besonders der Unterschied zwischen geradem und ungeradem Grad auf. Ungerade Werte für den Parameter d führen zu erheblich schlechteren Ergebnissen als gerade Werte. Für geraden Polynomialgrad ist beispielsweise die Fehleranzahl im Test maximal 21, wohingegen die Fehlerzahl für ungeraden Polynomialgrad bis zu 67 groß wird. Diese Erkenntnis zieht sich durch alle Untersuchungen und wird auch in der Literatur bestätigt. Bei Verwendung des SMO-Algorithmus kann man dieses unterschiedliche Verhalten nicht beobachten. Hier sind die ungeraden Polynomgrade sogar diejenigen, die die besseren Ergebnisse erzielen. Wegen einer deutlich besseren Sensitivität

erreicht man für $d = 5$ das beste Resultat, das der SMO-Algorithmus erzielt. Insgesamt sind die Ergebnisse mit dem Polynomialkern wesentlich unstetiger als bei Verwendung der anderen Kerne. Einige Tests führten zu schlechten Ergebnissen.

C-Wert	1.0	2.0	3.0	5.0	10.0	20.0	50.0	75.0	125.0	200.0
Fehler ges.	18	20	18	23	24	21	23	36	20	21
Fehler fn	8	8	6	6	6	6	4	8	12	12
Fehler fp	10	12	12	17	18	15	19	28	8	9
Sensitivität	0.39	0.39	0.54	0.54	0.54	0.54	0.71	0.39	0.08	0.08
Spezifität	0.85	0.82	0.82	0.74	0.72	0.77	0.69	0.57	0.88	0.86
Genauigkeit	0.77	0.74	0.77	0.71	0.69	0.73	0.71	0.54	0.74	0.73
F-Maß	0.36	0.33	0.44	0.38	0.37	0.40	0.44	0.22	0.09	0.09

Tabelle 6. Ergebnisse für verschiedene C-Werte der NPA-basierten SVM mit Polynomialkern.

Variation der Fehlergewichtung durch den Parameter C Die schlechten Resultate - auch für den Polynomgrad 2, der im Allgemeinen als Standardwert verwendet wird, motivieren eine Untersuchung, ob sie sich durch eine Variation des Parameters C doch noch verbessern lassen können. Dazu haben wir für den Grad 2 mit NPA noch weitere Tests durchgeführt, die in Tabelle 6 aufgeführt sind. Da der Rechenzeitaufwand für weitere Tests mit dem SMO-Algorithmus bei großen Werten für C und Verwendung des Polynomialkerns zu hoch war, wurde auf eine eingehendere Untersuchung mit dem SMO-Algorithmus verzichtet. Analog zu den anderen Kernen verändern wir nun also wieder den Parameter C als Indikator für die Fehlergewichtung. Dabei wird eine deutliche Verbesserung der Testergebnisse durch Optimierung des Parameters C in Bezug auf den Polynomialgrad 2 sichtbar. Zwar steigt die Fehleranzahl leicht an, aber als positiver Effekt wird die Fehlerzahl bei der Klassifikation von positiven Punkten durch optimales Auswählen von C halbiert. Auch hier ist die Verbesserung nicht beliebig weit fortführbar. Bei besonders großen Werten für C nimmt die Sensitivität wieder stark ab.

4 Unterscheidung von Fehlern erster und zweiter Art

Die Untersuchungen zu den Veränderungen bei der Fehlertolerierung mit den verschiedenen Kernen haben gezeigt, dass eine variable Fehlerbestrafung durch den Parameter C ein sinnvolles Mittel zur Optimierung der Ergebnisse ist. Wir vertiefen die Untersuchungen zur Fehlertolerierung noch weiter. Im Allgemeinen unterscheiden Support-Vektor-Maschinen keine Fehlerarten, da alle Fehler einfach oder quadratisch aufaddiert werden. Es kann jedoch Sinn machen, Fehler erster Art anders zu bestrafen als Fehler zweiter Art. Als Fehler erster Art gelten

C^+-Wert	1.0	2.0	5.0	10.0	20.0	50.0
C^--Wert	1.0	1.0	1.0	1.0	1.0	1.0
Fehler ges.	15	16	15	15	16	17
Fehler fn	7	5	3	3	3	3
Fehler fp	8	11	12	12	13	14
Sensitivität	0.46	0.62	0.77	0.77	0.77	0.77
Spezifität	0.88	0.83	0.82	0.82	0.80	0.79
Genauigkeit	0.81	0.80	0.81	0.81	0.80	0.78
F-Maß	0.44	0.50	0.57	0.57	0.56	0.54

Tabelle 7. Ergebnisse für stärkere Gewichtung der Fehler erster Art.

C^+-Wert	10.0	10.0	10.0	10.0	10.0
C^--Wert	1.0	5.0	10.0	50.0	100.0
Fehler ges.	15	15	12	12	12
Fehler fn	3	4	4	4	4
Fehler fp	12	11	8	8	8
Sensitivität	0.77	0.69	0.69	0.69	0.69
Spezifität	0.82	0.83	0.88	0.88	0.88
Genauigkeit	0.81	0.81	0.85	0.85	0.85
F-Maß	0.57	0.55	0.60	0.60	0.60

Tabelle 8. Ergebnisse für stärkere Gewichtung der Fehler zweiter Art.

dabei falsch negative Punkte und als Fehler zweiter Art gelten falsch positive Punkte. Eine Differenzierung zwischen den Fehlern kann einfach realisiert werden und macht insbesondere bei unausgeglichenen Datensätzen Sinn, da Fehler in der (typischerweise kleineren) positiven Punkteklasse schwerwiegender sind als Fehler in der negativen Punkteklasse. Dazu splittet man den Parameter C auf in C^+ und C^-, wobei C^+ die Fehler erster Art und C^- die Fehler zweiter Art gewichtet [EL05].

Die Ergebnisse der unterschiedlichen Fehlergewichtung sind in den Tabellen 7 und 8 dargestellt. In diesem Abschnitt stellen wir die Ergebnisse des NP-Algorithmusses vor. Dabei dokumentiert Tabelle 7 eine Variation der Bewertung von Fehlern erster Art und Tabelle 8 eine Variation der Bewertung von Fehlern zweiter Art. Erkennbar ist eine sehr unterschiedliche Ergebnisqualität je nach Wahl der zusammenspielenden Parameter. In beiden Fällen erkennt man durch höhere Gewichtung eine Senkung der Fehlerzahl in der zugehörigen Klasse. Dabei werden allerdings teilweise zusätzliche Fehler in der anderen Klasse in Kauf genommen. Da dieser negative Effekt jedoch häufig nicht überwiegt, lässt sich die Klassifikation somit insgesamt besser optimieren als bei einer gleichen Bewertung aller Fehler. Die Unterscheidung der Fehler in Fehler erster und zweiter Art ist also ein sinnvolles Mittel um gute Ergebnisse zu erzielen. Besonders wenn nur eine Klasse von erheblichem Interesse ist oder die Klassengrößen unterschiedlich

C^- \ C^+	1.0	2.0	4.0	6.0	8.0	10.0	20.0	50.0	100.0
	NPA								
1.0	0.44	0.50	0.57	0.57	0.57	0.57	0.56	0.54	0.54
2.0	0.42	0.44	0.58	0.53	0.56	0.56	0.56	0.56	0.56
4.0	0.30	0.48	0.53	0.56	0.55	0.53	0.53	0.53	0.53
6.0	0.33	0.44	0.55	0.58	0.58	0.56	0.53	0.53	0.53
8.0	0.33	0.44	0.55	0.60	0.60	0.58	0.55	0.53	0.53
10.0	0.33	0.44	0.55	0.60	0.60	0.60	0.56	0.55	0.53
	SMO								
1.0	0.46	0.49	0.53	0.53	0.53	0.53	0.53	0.53	0.53
2.0	0.00	0.47	0.55	0.57	0.57	0.57	0.53	0.53	0.53
4.0	0.00	0.38	0.53	0.56	0.63	0.55	0.53	0.53	0.53
6.0	0.00	0.13	0.43	0.52	0.55	0.56	0.53	0.53	0.53
8.0	0.00	0.12	0.35	0.48	0.48	0.55	0.48	0.48	0.48
10.0	0.00	0.12	0.35	0.46	0.55	0.52	0.50	0.55	0.55

Tabelle 9. Verhalten des F-Maßes bei Variation von C^+ und C^-.

sind, kann eine unterschiedliche Gewichtung der Fehler bei bestimmten Daten zu einer erheblichen Verbesserung der Ergebnisse führen. Als besonders gute Wahl ist dabei $C^+ = 5.0$ und $C^- = 1.0$ hervorzuheben. Das Verhältnis von C^+ und C^- spiegelt dabei in etwa das Größenverhältnis der beiden zugrundeliegenden Klassen wieder. Dieses Ergebnis bestätigt die Aussage, dass die Gewichtung in etwa dem Umfang der Daten entsprechen sollte [Mar01].

Die gerade gewonnenen positiven Erkenntnisse motivieren dazu, die Auswahl der Fehlergewichtung noch tiefgreifender zu untersuchen. Dazu haben wir weitere Tests mit verschiedenen Fehlergewichtungen gemacht. Verwendet wurden beide Algorithmen, wobei stets der Gaußkern mit Parameter $\sigma = 1.0$ gewählt wurde und verändert wurde die Gewichtung der Fehler erster und zweiter Art durch gleichzeitige Variation von C^+ und C^-. Betrachtet man Tabelle 9, in der wir die Werte des F-Maßes in den Tests angeben, so bestätigt sich noch einmal die positive Auswirkung einer getrennten Fehlergewichtung. Deutlich ist aber auch, dass sich die Klassifikationsergebnisse für besonders hohe Werte der Parameter nicht mehr stark ändern, sondern sich auf ein bestimmtes Fehlerverhalten einpendeln. Insgesamt zeigt diese zweite Untersuchung noch einmal, dass eine unterschiedliche Gewichtung der Fehler ein sinnvolles Mittel ist, um gute Ergebnisse zu erzielen.

5 Methoden zur sensitiven Beeinflussung der Hypothesen

Zusätzlich zu der im letzten Kapitel untersuchten Möglichkeit, die Sensitivität durch unterschiedliche Fehlertoleranzen zu erhöhen, existieren weitere Möglichkeiten zur kostensensitiven Klassifikation. Wir stellen im Folgenden zwei weitere

Methoden dazu vor. Bei der ersten Methode wird die SVM-Hyperebene geändert. Bei der zweiten Methode sollen die Daten a priori verbessert werden, um der SVM die Möglichkeit zu geben, die gewünschte Klasse automatisch besser zu erkennen, als das mit den Originaldaten der Fall wäre.

5.1 Schwellwertverschiebung

Die Zielfunktion der Support-Vektor-Maschinen berechnet reelle Werte, welche dann mittels ihres Vorzeichens eine Klassifikation ergeben. Beim Betrachten der konkreten Funktionswerte der Ergebnisse fällt auf, dass die berechneten Werte in der negativen Klasse bei richtiger Einordnung betragsmäßig groß sind, während fehlerhaft eingeordnete positive Punkte Funktionswerte nahe bei Null besitzen. Sie sind vom Sprung zur positiven Klassifikation nicht weit entfernt. Durch eine affine Verschiebung der Entscheidungsebene ließe sich also das Ergebnis verbessern, da dann durch Änderung der Vorzeichen aus einigen falsch negativen Punkten korrekt klassifizierte positive Punkte werden würden. Da wir die positive Klasse bevorzugen, hoffen wir, Ergebnisse in die gewünschte Richtung optimieren zu können. Dabei sollten gleichzeitig nur wenige neue falsch positive Punkte hinzukommen. Wir werden im Folgenden untersuchen, ob eine Verschiebung der Hyperebene positive Effekte für unseren Datensatz hat oder nicht. Diese Art von Verschiebung der Entscheidungsebene nennt man Schwellwertverschiebung [ZHZ06]. Sie ist in Abbildung 3 grob dargestellt.

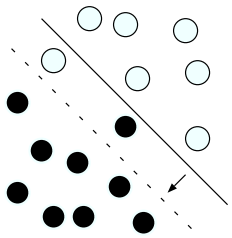

Abbildung 3. Schwellwertverschiebung.

Um den positiven Effekt der Verschiebung zu erkennen, ist es sinnvoll, einzelne Tests genauer zu betrachten. Wir verwenden den NP-Algorithmus mit Parametern $C = 1.0$ und $\sigma = 1.0$. In Tabelle 10 sind jeweils die verwendeten Schwellwertveränderungen angegeben, wobei auch negative Werte getestet wurden. Es ist deutlich erkennbar, welche Auswirkungen die Verschiebung der Entscheidungsebene bei üblichen Parameterkombinationen hat. Wenn man die Ebene um einen geeigneten Faktor in eine Richtung verschiebt, so sinkt die Anzahl der Fehler in der Klasse, zu deren Gunsten man die Verschiebung durchgeführt hat. So liegt die Anzahl der falsch negativen Punkte in unserem Beispiel

Schwellwertabweichung	-0.5	-0.2	0.0	0.2	0.5
Fehler ges.	12	14	15	15	21
Fehler fn	12	8	7	5	2
Fehler fp	0	6	8	10	16
Sensitivität	0.08	0.39	0.46	0.62	0.85
Spezifität	1.00	0.91	0.88	0.85	0.75
Genauigkeit	0.85	0.82	0.81	0.81	0.77
F-Maß	0.14	0.42	0.44	0.52	0.55

Tabelle 10. Klassifikationsänderungen durch variable Schwellwerte.

bei einer Schwellwertabweichung von 0.5 nur bei 2. Allerdings ist auch zu beachten, dass gleichzeitig die Anzahl der falsch positiven Punkte auf 16 steigt. Insgesamt gibt es nun also eine größere Anzahl an Fehlern, aber da die Klassen unterschiedliche Wichtigkeit besitzen, verbessert man dennoch das erzielte Ergebnis.

Durch Verschiebung der Entscheidungsebene wird also das Resultat erheblich verändert, wobei man in eine Richtung Verbesserungen und in der anderen Richtung Verschlechterungen erzielt. Deutlich ist die Verbesserung der Qualität der Resultate durch eine positive Schwellwertverschiebung ausgehend von Null. Die Anzahl der falsch positiven Punkte steigt nicht viel stärker als die Anzahl der falsch negativen Punkte sinkt. Unter Berücksichtigung der unterschiedlichen Bewertung der Fehler ist also das erzielte Resultat besser. In allen unseren Tests hat sich gezeigt, dass die Schwellwertverschiebung zu sehr guten Optimierungen der Ergebnisse führen kann. Gerade weil Fehler in den beiden Klassen häufig unterschiedlich gewertet werden, ist diese Methode ein sehr sinnvolles Hilfsmittel zur Ergebnisoptimierung und es macht insbesondere Sinn eine positiven Schwellwertänderung zu wählen, um die Sensitivität zu erhöhen.

Schwellwerte und Kreuzvalidierung Bisher sind die genutzten Parameter immer einzeln vom Benutzer der Support-Vektor-Maschine angegeben worden. In der Regel gibt der Benutzer jedoch keine bestimmten Parameter vor, sondern lässt den Computer aus einer vorgegebenen Auswahl von möglichen Parameterkombinationen die beste auswählen. Die Parameterauswahl geschieht im Rahmen einer sogenannten zehnfachen Kreuzvalidierung [Eit03]. Diese funktioniert folgendermaßen: Aus den Trainingspunkten werden für eine beliebige Parameterkombination zehnmal jeweils 10% für eine Kontrolle zurückgehalten und mit den anderen 90% wird eine Klassifikationsfunktion trainiert. Dann werden die zurückgehaltenen Punkte klassifiziert und die dabei entstandenen Fehler aufaddiert. Anhand der Gesamtfehlerzahl aller 10 Tests in beiden Klassen wird dann das F-Maß für die bestimmte Parameterkombination berechnet. Die Parameterkombination mit dem besten F-Maß wird dann für das endgültige Training mit allen Punkten verwendet und ermittelt somit die Klassifikationsfunktion, mit der die Punkte im Test klassifiziert werden.

Schwellwert	nach der Optimierung			bei der Optimierung		
Testnummer	1	2	3	1	2	3
Fehler ges.	19	28	38	19	19	22
Fehler fn	2	1	1	2	2	2
Fehler fp	17	27	37	17	17	20
Sensitivität	0.85	0.92	0.92	0.85	0.85	0.85
Spezifität	0.74	0.58	0.43	0.74	0.74	0.69
Genauigkeit	0.76	0.64	0.51	0.76	0.76	0.72
F-Maß	0.54	0.46	0.39	0.54	0.54	0.50

Tabelle 11. Vergleich von Ergebnissen bei unterschiedlichem Einsatz der Schwellwertverschiebung.

Nachdem die Schwellwertverschiebung nach dem Training zu guten Ergebnissen geführt hat, stellt sich nun die Frage, ob es Sinn macht, diese Technik auch schon vor der Parameterwahl in der Validierung durchzuführen. Das wird im Folgenden getestet. Im Rahmen der Tests gab es dazu die Möglichkeiten, die Schwellwertverschiebung entweder erst im finalen Test oder bereits im Rahmen der Validierung einfließen zu lassen. Es ist zu untersuchen, ob diese zweite Möglichkeit die Testergebnisse verbessern kann. Mittels der folgenden 3 typischen Beispieldurchläufe haben wir die beiden Methoden verglichen:

1. NP-Algorithmus mit Gaußkern und einer Schwellwertänderung von 0.2,
2. NP-Algorithmus mit Gaußkern und einer Schwellwertänderung von 0.4,
3. SMO-Algorithmus mit Slaterkern und einer Schwellwertänderung von 0.4.

Die möglichen Werte für σ, C^+und C^- wurden aus einer Datei ausgelesen und die SVM-Software hat die optimale Parameterkombination durch zehnfache Kreuzvalidierung bestimmt. Bei den finalen Tests wurden die in Tabelle 11 dargestellten Ergebnisse erzielt. Zusammenfassend läßt sich sagen:

- Ohne Schwellwertverschiebung während der Validierung erhielt man im Training zu jedem Datensatz eine bestimmte Parameterkombination. Die zugehörigen Tests ergaben dann Ergebnisse mit einem F-Maß zwischen 0.39 und 0.54.
- Wurde die Schwellwertverschiebung hingegen schon bei der Kreuzvalidierung berücksichtigt, sind andere Parameterkombinationen favorisiert worden. Dann ergaben die finalen Tests F-Maße zwischen 0.50 und 0.54.

Mittels der Methode der Schwellwertverschiebung bei der Parameteroptimierung wird also kein signifikant besseres Ergebnis erreicht als zuvor. Allerdings ist die Streuung der Ergebnisse geringer, also sind die Ergebnisse stabiler als zuvor. Die Frage ob eine Schwellwertverschiebung also nur nach dem letzten Training oder schon vorher gemacht werden sollte, lässt sich mit den durchgeführten Tests nicht endgültig klären. Wir tendieren jedoch dahin, dass es nicht unbedingt notwendig ist, eine Verschiebung schon vor dem letzten Training durchzuführen, da die erzielten Ergebnisse den erbrachten Aufwand nicht rechtfertigen.

5.2 Oversampling

Arbeitet man mit unterschiedlichen Klassengrößen, so haben die verschiedenen Klassen auch einen unterschiedlich großen Einfluss auf das Ergebnis. Prinzipiell tendieren Lernverfahren dazu, die größere Klasse besser zu erkennen. Auf stark unausgeglichenen Daten versagen einfache Lernmethoden mitunter komplett. Darum nutzt man Resampling-Methoden, um die Klassengrößen aneinander anzupassen [Liu04]. Generell unterscheidet man zwischen Oversampling und Undersampling. Vergrößert man die kleinere Klasse, so spricht man von Oversampling, verkleinert man die größere Klasse, so spricht man von Undersampling. Die Techniken werden für die Trainingsdaten unabhängig vom verwendeten Algorithmus durchgeführt. Im Rahmen der Tests haben wir uns für die Untersuchung der Oversampling-Methode entschieden. Der Grund dafür ist, dass der Datensatz ohnehin sehr klein ist, und dass Undersampling die Gefahr birgt, zu viele Informatione zu verlieren. Der Datensatz hatte ursprünglich 185 Trainingspunkte, von denen 35 positiv und 150 negativ waren. Im Rahmen des Oversamplings wurde die Klasse mit den positiven Punkten auf die 3-fache Größe ausgeweitet. In einem neuen Datensatz standen dann also 255 Punkte für das Training zur Verfügung. 105 Punkte davon waren positiv. Die Tests zielten darauf ab, zu zeigen, ob man mittels Oversampling die Ergebnisse tatsächlich verbessern kann oder nicht. Für beide Algorithmen sind 4 sinnvolle Parameterkombinationen gewählt worden, die mit dem unveränderten und dem modifizierten Datensatz trainiert wurden. Die Fehlerverteilungen in den Tests werden in Tabelle 12 gegenübergestellt.

Erkennbar ist ein positiver Effekt des Oversamplings, besonders bei Verwendung des SMO-Algorithmus. Bei allen 4 Parametersettings wurde die Sensitivität und daraus folgend auch das F-Maß nach einem Oversampling besser. Vor allem fällt auch auf, dass zwischen den Ergebnissen bei den unterschiedlichen Parameterkombinationen nach Verwendung von Oversampling weitaus weniger Varianz vorliegt als beim ursprünglichen Datensatz. Allerdings wird global gesehen kein besseres Ergebnis erreicht. Betrachtet man nämlich auch die Ergebnisse mit dem NP-Algorithmus, so fällt auf, dass durch das Oversampling ab einem bestimmten Punkt keine Ergebnisverbesserung mehr stattfindet. In einem Fall hat sich das Ergebnis sogar verschlechtert. Man kann also sagen, dass durch Oversampling keine besseren Ergebnisse als durch andere Methoden mit Hilfe des unveränderten Datensatzes erreicht werden. Allerdings wird dieses Ergebnis nach einem Oversampling für weitaus mehr Parameterkombinationen und somit schneller erreicht. Die Aspekte der Robustheit und der Zeit, die man bei der Parameteroptimierung sparen kann, möchten wir als positiven Effekt festhalten.

6 Fazit

Eine sinnvolle Parameterwahl ist ein wichtiger Bestandteil bei der Anwendung von Support-Vektor-Maschinen. Nur bei der Arbeit mit einer stimmigen Kombination der Parameter kann eine gute Klassifikation erfolgen. Die Modifikation und Auswahl der Parameter ist dabei nicht willkürlich, sondern durch charakteristische Merkmale des untersuchten Datensatzes und der vorgegebenen Ziel-

	NPA							
C^+	1.0		5.0		10.0		10.0	
C^-	1.0		1.0		1.0		10.0	
Oversampling	-	+	-	+	-	+	-	+
Fehler fn	7	3	3	3	3	3	4	4
Fehler fp	8	12	12	12	12	12	8	11
Sensitivität	0.46	0.77	0.77	0.77	0.77	0.77	0.71	0.71
Spezifität	0.88	0.82	0.82	0.82	0.82	0.82	0.88	0.83
Genauigkeit	0.81	0.81	0.81	0.81	0.81	0.81	0.85	0.81
F-Maß	0.44	0.57	0.57	0.57	0.57	0.57	0.60	0.55
	SMO							
C^+	1.0		5.0		10.0		10.0	
C^-	1.0		1.0		1.0		10.0	
Oversampling	-	+	-	+	-	+	-	+
Fehler fn	7	3	4	3	4	3	5	4
Fehler fp	7	12	12	12	12	12	10	11
Sensitivität	0.46	0.77	0.71	0.77	0.71	0.77	0.62	0.71
Spezifität	0.89	0.82	0.82	0.82	0.82	0.82	0.85	0.83
Genauigkeit	0.82	0.81	0.80	0.81	0.80	0.81	0.81	0.81
F-Maß	0.46	0.57	0.53	0.57	0.53	0.57	0.52	0.55

Tabelle 12. Vergleich von Ergebnissen mit und ohne Oversampling.

setzung der Klassifikation vorgegeben. Wichtige Methoden zur Ergebnisoptimierung sind insbesondere die Unterscheidung der Fehler in zwei Klassen und die Schwellwertverschiebung. Oversampling hingegen ist nur für bestimmte Datensatztypen sinnvoll.

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.

[Jol86] I. T. Jolliffe. *Principal component analysis.* Springer, New York, 1986.

[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.

[Liu04] A. Y. Liu. The effect of oversampling and undersampling on classifying imbalanced text datasets. Master's thesis, University of Texas at Austin, 2004.

[Mar01] F. Markowetz. Support vector machines in bioinformatics. Master's thesis, University of Heidelberg, 2001.

[McC84] G. P. McCabe. Principal variables. *Technometrics*, 26:137–144, 1984.

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

[RKC+02] M. L. Rekart, M. Krajden, D. Cook, G. McNabb, T. Rees, and J. Isaac-Renton. Problems with the fast-check HIV rapid test kits. *Canadian Medical Association Journal*, 167(2), 2002.

[RZI98] R. Rossius, G. Zenker, and A. Ittner. A short note about the application of polynomial kernels with fractional degree in support vector learning. *Lecture Notes In Computer Science*, 1398:143–148, 1998.

[ZHZ06] X.-Y. Liu Z.-H. Zhou. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, January 2006.