Bauhaus-Universität Weimar
Fakultät Medien
Web Technology & Information Systems

Abschlussarbeit
zur Erlangung des akademischen Grades
BACHELOR OF SCIENCE

INTRINSISCHE PLAGIATERKENNUNG MIT AUSREIßERN

von

TSVETOMIRA BOYCHEVA PALAKARSKA

Betreuer
Prof. Dr. Benno Stein
Dr. Sven Meyer zu Eißen

März 2008

За
мама, тати,
Йоана и Денис

Erklärung

Hiermit versichere ich, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat.

Ort, Datum Unterschrift

Kurzfassung

Gegenstand dieser Arbeit ist die Erkennung von plagiierten Textabschnitten in einem Dokument. Schwerpunkt der Plagiatanalyse bildet die intrinsische Plagiaterkennung, bei der keine Dokumentensammlung zum Vergleich vorliegt. Grundlage der intrinsischen Plagiaterkennung ist das Erkennen von Stilwechseln innerhalb eines Dokumentes. Der Schreibstil eines Autors, der mittels Stilmerkmalen beschrieben wird, wird abschnittsweise analysiert. Als Stilmerkmale dienen unter anderem die durchschnittliche Satzlänge oder die Anzahl der Adjektive im betrachteten Abschnitt. Der Abschnitt wird somit durch m verschiedene Stilmerkmale als m-dimensionaler Vektor repräsentiert. Die Stilmerkmale werden anschließend mit denen des Dokumentes verglichen. Unterscheidet sich ein Stilmerkmal von dem des Dokumentes, so liegt ein Stilwechsel vor.

Ausreißer werden verwendet, um Stilwechsel festzustellen. Ziel ist es, solche Ausreißer zu entdecken. Dazu werden Funktionen zur Ausreißererkennung verwendet. Es wird erwartet, das dadurch eine höhere Genauigkeit bei der Klassifikation von vermeintlich plagiierten Textabschnitten erzielt werden kann.

In der vorliegenden Arbeit werden Funktionen zur Ausreißererkennung vorgestellt und analysiert. Zur Auswertung liegt eine Dokumentensammlung von insgesamt 760 Dokumenten vor. Es zeigt sich, dass die Anforderung einer hohen Genauigkeit erfüllt ist. Ergebnisse belegen, dass mit steigender Anzahl erkannter Ausreißer innerhalb eines Abschnittes sich die Genauigkeit erhöht.

Inhaltsverzeichnis

Seite

Erklärung . . . iv
Kurzfassung . . . vi
Inhaltsverzeichnis . . . ix
Abbildungsverzeichnis . . . xi
Tabellenverzeichnis . . . xiii

Kapitel

1 Einleitung . . . 1
2 Statistische Grundlagen . . . 3
2.1 Der Wahrscheinlichkeitsbegriff . . . 3
2.2 Zufallsvariablen und ihre Verteilungen . . . 4
2.3 Kenngrößen von Verteilungen . . . 6
2.4 Diskrete Verteilungen . . . 7
2.5 Stetige Verteilungen . . . 8
2.6 Grundgesamtheit, Stichprobe, Schätzer und Tests . . . 12
2.6.1 Schätzer . . . 13
2.6.2 Statistische Tests . . . 14
2.6.2.1 χ^2- Test . . . 15
2.6.2.2 Kolmogorov-Smirnov-Test . . . 16
3 Plagiaterkennung . . . 21
4 Modell zur Erfassung von Stiländerungen als Ausreißer . . . 25
5 Funktionen zur Eine-Klassen-Klassifikation von Plagiatstellen als Ausreißer . . . 29
5.1 Quantifizierung der Normalverteilungsannahme . . . 32
5.2 Lineare Funktionen . . . 34
5.3 Regressionsfunktionen . . . 35
5.4 Relative Abweichungen durch die Dichtefunktion der Normalverteilung . . . 39
5.5 Ausreißer außerhalb des verstärkten Interquartilbereiches der Dichtefunktion der Normalverteilung . . . 40
6 Evaluierung der Funktionen zur Eine-Klassen-Klassifikation . . . 43

6.1 Gütemaße zur Bewertung der Funktionen 43
6.2 Support-Vektor-Maschine (SVM) 44
6.3 Experimente 45
6.4 Analyse der Funktionen 46
7 Zusammenfassung und Ausblick 51
Anhang
A Übersicht von Verfahren zur Eine-Klasse-Klassifikation 53
B Übersicht von Verfahren zur Ausreißererkennung 57
Literaturverzeichnis 64

Abbildungsverzeichnis

2.1 Dichte- und Verteilungsfunktion einer stetig verteilten Zufallsvariablen . . 8
2.2 Dichte- und Verteilungsfunktionen von Normalverteilungen 10
2.3 Dichte- und Verteilungsfunktionen der Standardnormalverteilung 10

3.1 Taxonomie von Plagiatvergehen nach Meyer zu Eissen u. a. (2007) 21

5.1 Plagiaterkennung durch Eine-Klassen-Klassifikation (Aktivitätsdiagramm) 30
5.2 Plagiaterkennung durch Eine-Klassen-Klassifikation als Zwei-Schritt-Prozess 31
5.3 Varianzen für zwei Stichprobenvariablen 33
5.4 Normal-Quantil-Plots für zwei Stichprobenvariablen 34
5.5 Lineare Funktionen und absolute Häufigkeiten für zwei Stichprobenvariablen 35
5.6 Anpassung der Regressionsgeraden für zwei Stichprobenvariablen 37
5.7 Ursprüngliche und modifizierte Regressionsgeraden für zwei Stichprobenvariablen 38
5.8 Vergleich zwischen den Geraden der linearen Funktion und den Regressionsgeraden 39
5.9 Dichtefunktionen von zwei Stichprobenvariablen. 40

6.1 Berechnungsschema von Precision und Recall 43
6.2 Precision und Recall für 500 Wörter langen Textabschnitten 47
6.3 Precision und Recall für 1000 Wörter langen Textabschnitten 48
6.4 Precision und Recall für 1500 Wörter langen Textabschnitten 48
6.5 Precision und Recall für 2000 Wörter langen Textabschnitten 48
6.6 Precision und Recall der Eine-Klassen-SVM 49

Tabellenverzeichnis

2.1 Wahrscheinlichkeitsaussagen bzgl. einer Zufallsvariablen nach Rinne (1997) 5
2.2 Wahrscheinlichkeiten als Flächenanteile in bestimmten Intervallen 11
2.3 Kritische Werte der D-Statistik D_n nach Mason u. Bell (1986) 18

6.1 Ergebnisse der Eine-Klassen-SVM 49

A.1 Übersicht von Eine-Klassen-Klassifikations-Methoden 55

B.1 Übersicht verschiedenen Verfahren zur Ausreißererkennung 63

1 Einleitung

Die Menge an Informationen, die durch das World Wide Web zugänglich ist, bietet Benutzern die Möglichkeit, aus einem ständig wachsenden Informationsbestand Wissen zu gewinnen. Insbesondere gestattet der uneingeschränkte Informationsfluss einem Benutzer gezielt Ideen ausfindig zu machen und als eigene auszugeben. Diese stammen beispielsweise aus dem akademischen Bereich, der Wissenschaft, dem Produktdesign oder der Musikbranche. Der Vorgang der Aneignung fremden geistigen Eigentums ohne diesen als solchen anzugeben, wird als Plagiieren bezeichnet. Plagiate resultieren aus diesem Vorgehen und repräsentieren die übernommenen Ideen.

In dieser Arbeit werden Plagiate innerhalb von Textdokumenten betrachtet, die mit Hilfe von Ausreißern offengelegt werden sollen. Bei der intrinsischen Plagiatanalyse eines Dokumentes liegen keine externen Quellen zum Vergleich vor, d.h. ein Plagiatvergehen wird nur anhand des untersuchten Dokumentes festgestellt. Dadurch lassen sich beispielsweise Plagiate erkennen, die aus einem Dokument stammen, das elektronisch nicht zur Verfügung steht. Bei der intrinsischen Plagiatanalyse sind nur Informationen über die Dokumentbestandteile des angeblichen Autors gegeben, wohingegen Informationen über Dokumentbestandteile anderer Autoren fehlen. Diese Situation wird als Eine-Klasse-Klassifikation bezeichnet. Ziel ist es, durch die gegebenen Informationen Ausreißer zu bestimmen.

Die intrinsische Plagiaterkennung mit Ausreißern basiert auf dem Verfahren der Stilanalyse (Meyer zu Eißen u. Stein (2006)). Dabei wird das untersuchte Dokument durch eine Stilrepräsentation dargestellt. Die Stilrepräsentation quantifiziert den Schreibstil des Autors durch Erfassung verschiedener Stilmerkmale, z.B. durchschnittliche Satzlänge oder Anzahl der Abkürzungen im Dokument. Innerhalb des Dokumentes werden mit Hilfe der Stilmerkmale Schreibstilunterschiede analysiert. Weicht der Schreibstils eines Abschnittes bezüglich eines Stilmerkmals von dem des Dokumentes ab, so spricht diese Abweichung für einen Ausreißer. Ein Ausreißer ist also ein Zeichen dafür, dass der Abschnitt plagiiert ist. Für die Erkennung solcher Ausreißer und der Feststellung eines Plagiatvergehens werden schließlich in dieser Arbeit verschiedene Funktionen vorgestellt.

Das zweite Kapitel führt statistische Grundlagen ein, auf die in den späteren Kapiteln zurückgegriffen wird. Kapitel 3 stellt im Anschluss verschiedene Methoden zur Plagiaterkennung vor. Das nächste Kapitel führt eine Repräsentation zur Erfassung von Stiländerungen innerhalb eines Dokumentes ein, um in Kapitel 5 Funktionen zur Eine-Klasse-Klassifikation von plagiierten Stellen als Ausreißer vorzustellen. Anschließend werden die Funktionen in Kapitel 6 evaluiert und die Ergebnisse diskutiert. Eine Zusammenfassung und ein Ausblick schließen die Arbeit ab.

Anhang A bietet eine Übersicht über verschiedene Eine-Klassen-Klassifizierungsverfahren. Anhang B stellt zudem einige Ausreißererkennugstrategien zusammen.

2 Statistische Grundlagen

In diesem Kapitel werden grundlegende statistische Begriffe eingeführt, die für die weitere Betrachtung vorausgesetzt werden. Zunächst wird der Wahrscheinlichkeitsbegriff eingeführt. Im zweiten Teil dieses Kapitels werden Zufallsvariablen und ihre Verteilungen vorgestellt. Im dritten Abschnitt werden Kenngrößen von Verteilungen eingeführt. Diskrete und stetige Verteilungen werden in den nächsten beiden Abschnitten behandelt. Für beide Verteilungstypen werden jeweils exemplarische Verteilungen vorgestellt. Das Kapitel schließt mit der Einführung der Begriffe Grundgesamtheit, Stichprobe, Schätzer und statistische Tests.

2.1 Der Wahrscheinlichkeitsbegriff

Ausgangspunkt ist ein Zufallsexperiment. Ein Zufallsexperiment ist ein beliebig oft wiederholbarer Vorgang, der unter gleichen Bedingungen erfolgt und dessen Ausgang nicht mit Sicherheit vorhersagbar ist. Sei ω_i mit $i = 1, 2, \ldots, n$ ein mögliches Ergebnis eines Zufallsexperimentes, ein sogenanntes Elementarereignis, so definiert die Menge $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ den Ergebnisraum des Zufallsexperimentes. Jedem Versuchsausgang ist höchstens ein Elementarereignis $\omega \in \Omega$ zugeordnet. Jede Teilmenge A eines Ergebnisraums Ω heißt Ereignis. Ein Ereignis A tritt ein, wenn der Versuchsausgang ω ein Element der Menge A ist, es gilt also $\omega \in A$. Die Potenzmenge aller Ereignisse $\mathcal{P}(\Omega)$ wird als Ereignisraum bezeichnet.

Der klassische Laplacesche Wahrscheinlichkeitsbegriff ist wie folgt formuliert. Die Wahrscheinlichkeit $P(A)$ eines Ereignisses A wird definiert durch:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der für A günstigen Elemenratereignisse}}{\text{Anzahl aller möglichen Elemenratereignisse}}$$

Der axiomatische Wahrscheinlichkeitsbegriff legt die formal-mathematischen Eigenschaften der Ereignisse fest. Die axiomatische Definition der Wahrscheinlichkeit nach Kolmogorov ist wie folgt aufgebaut. Gegeben seien der Ergebnisraum Ω und der Ereignisraum $\mathcal{P}(\Omega)$. Jedem Ereignis $A \in \mathcal{P}(\Omega)$ wird eine reelle Zahl $P(A)$ durch die Funktion

$P : \mathcal{P}(\Omega) \rightarrow \mathbb{R}$ zugeordnet. $P(A)$ wird als die Wahrscheinlichkeit des Ereignisses A bezeichnet. Sie wird durch folgende Axiome charakterisiert:

- Axiom I: $\forall\ A \in \mathcal{P}(\Omega) :\ 0 \leq P(A) \leq 1$.

  Die Wahrscheinlichkeit für das Eintreten des Ereignisses A ist eine reelle Zahl zwischen 0 und 1. Der Definitionsbereich der Funktion P ist Ω, während der Wertebereich das Intervall $[0; 1]$ darstellt.

- Axiom II: $P(\Omega) = 1$.

  Die Wahrscheinlichkeit für das Eintreten des (sicheren) Ereignisses Ω ist 1.

- Axiom III: $A, B \in \mathcal{P}(\Omega) \wedge A \cap B = \emptyset \rightarrow P(A \cup B) = P(A) + P(B)$.

  Die Wahrscheinlichkeit für das Eintreten des Ereignisses A oder B ist gleich der Summe der Wahrscheinlichkeiten der einzelnen Ereignisse, wenn diese sich gegenseitig ausschließen (Storm (1995)).

2.2 Zufallsvariablen und ihre Verteilungen

Eine Zufallsvariable wird als Ergebnis eines Zufallsexperimentes betrachtet. Eine Zufallsvariable X ist eine Funktion, die jedem Elementarereignis $\omega \in \Omega$ eindeutig eine reelle Zahl x so zuordnet, dass für jedes x die Wahrscheinlichkeit $P(X(\omega) \leq x)$ definiert ist (Rinne (1997)). x bzw. x_i heißt Realisierung oder Wert der Zufallsvariablen X an der Stelle ω. Der Bezug auf ω wird im Allgemeinen weggelassen: $X = x$ bzw. $X = x_i$. Eine Zufallsvariable X ist diskret, wenn sie endlich viele oder abzählbar unendlich viele Realisierungen x_i annimmt. Eine Zufallsvariable X ist stetig, wenn sie jeden beliebigen reellen Zahlenwert x_i eines Intervalls annimmt. Über eine Zufallsvariable werden ausschließlich Wahrscheinlichkeitsaussagen getroffen, siehe Tabelle 2.1.

Das Verteilungsgesetz einer Zufallsvariable ist durch eine Verteilungsfunktion gegeben. Die Verteilungsfunktion $F(x)$ einer Zufallsvariablen X gibt die Wahrscheinlichkeit für die Realisierung einer Zufallsvariablen im Bereich von $-\infty$ bis zu einer oberen Grenze x an:

$$F(x) = P(X \leq x),\ x \in \mathbb{R}$$

Die Verteilungsfunktion besitzt folgende Eigenschaften (Rinne (1997)):

1. $0 \leq F(x) \leq 1$.
2. $F(x)$ ist monoton steigend in x, d.h. $F(x_2) \geq F(x_1)\ \forall\ x_2 > x_1$.

Wahrscheinlichkeit	**Erklärung**
$P(X = x)$	Einzelwahrscheinlichkeit, wobei X genau den Wert einer Realisierung x annimmt.
$P(X \leq x)$	Erreichungswahrscheinlichkeit von x, X ist höchstens x.
$P(X > x)$	Überschreitungswahrscheinlichkeit von x, X ist größer x.
$P(x_a \leq X \leq x_b)$	Wahrscheinlichkeit für Realisierungen im Intervall $[x_a; x_b]$.
$P(X < x_a \vee X > x_b)$	Wahrscheinlichkeit für Realisierungen außerhalb des Intervalls $[x_a; x_b]$.

Tabelle 2.1 *: Wahrscheinlichkeitsaussagen bezüglich einer eindimensionalen Zufallsvariablen nach Rinne (1997). Im Fall der Einzelwahrscheinlichkeit handelt es sich um eine diskrete Zufallsvariable, in den anderen Fällen um stetige Zufallsvariablen.*

3. $F(x)$ ist rechtsseitig stetig.

4. $F(x) \to 0$ für $x \to -\infty$;

   $F(x) \to 1$ für $x \to +\infty$.

Für eine diskrete Zufallsvariable sei $p(x_i)$ die Wahrscheinlichkeitsfunktion. Folgender Zusammenhang zwischen der Wahrscheinlichkeits- und Verteilungsfunktion ist gegeben:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p(x_i)$$

Für eine stetige Zufallsvariable sei $f(x)$ die Wahrscheinlichkeitdichte. Die Beziehung zwischen Dichte- und Verteilungsfunktion lautet:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \mathrm{f}(t)\ \mathrm{d}t$$

Für die Wahrscheinlichkeitdichte gilt:

$$\int_{-\infty}^{+\infty} \mathrm{f}(x)\ \mathrm{d}x = 1$$

Da die Wahrscheinlichkeit, dass eine Zufallsvariable einen bestimmten Wert x_i annimmt, gleich 0 ist, wird die Wahrscheinlichkeit, dass diese in einem endlichen Intervall $[a; b]$ liegt, betrachtet:

$$P(a \leq X \leq b) = F(b) - F(a) = \int_{a}^{b} \mathrm{f}(x)\ \mathrm{d}x$$

Die Wahrscheinlichkeit dieses Ereignisses ist gleich der Fläche unter der Kurve der Wahrscheinlichkeitdichte f zwischen $x = a$ und $x = b$.

Falls eine Wahrscheinlichkeit α vorgegeben ist, so dass

$$P(X > x) = \alpha$$

gilt, wird die zugehörige Abszisse $x = x_\alpha$ Quantil der Verteilung genannt. Das bedeutet, dass der Flächeninhalt unter der Dichtefunktion rechts von x_α gleich α ist.

2.3 Kenngrößen von Verteilungen

Kenngrößen sind Maßzahlen, die das Zufallsexperiment qualitativ beschreiben. Sie charakterisieren eine Zufallsvariable und deren Wahrscheinlichkeitsverteilung. Die Lage der Wahrscheinlichkeitsverteilung wird durch Lageparameter beschrieben. Ihre Variabilität wird dagegen durch Streuungsparameter spezifiziert.

Gegeben seien eine Zufallsvariable X und ihre Realisierungen, die Werte x_i. Weiterhin definieren $p(x_i)$ die Wahrscheinlichkeitsfunktion und $f(x)$ die Dichtefunktion für eine diskrete bzw. stetige Zufallsvariable.

Der Erwartungswert ist ein Lageparameter. Dieser bezeichnet das Zentrum, um das sich die Werte der Zufallsvariable gruppieren. Der Erwartungswert μ wird wie folgt definiert:

$$\mu = E(X) = \begin{cases} \sum_i x_i p(x_i), & \text{wenn } X \text{ eine diskrete ZV ist} \\ \int_{-\infty}^{+\infty} \mathrm{x} f(x)\, \mathrm{d}x, & \text{wenn } X \text{ eine stetige ZV ist} \end{cases} \tag{2.1}$$

ZV steht für den Begriff der Zufallsvariable.

Ein Quantil ist ebenfalls ein Lageparameter. Ein Quantil ist ein bestimmter Punkt einer aufsteigend geordneten Liste der Werte der Zufallsvariable. Bei einer Unterteilung dieser Liste in n gleich große Teile ergeben sich $n - 1$ Quantile. Abhängig von der Wahl von n ist die Rede von Quartilen ($n = 4$), Quintilen ($n = 5$), Dezilen ($n = 10$) oder Perzentilen ($n = 100$). Die Differenz zweier Quantile stellt einen Steuungparameter dar. Die so entstandenen Intervalle bestimmen letztendlich Streubereiche.

Die Varianz ist ein Streuungsparameter. Dieser beschreibt wie stark die Werte einer Zufallsvariable um den Erwartungswert streuen. Sie ist ein Maß für die Abweichung der Zufallsvariable vom Erwartungswert. Sie wird auch als mittlere quadratische Abweichung

vom Erwartungswert bezeichnet. Die Varianz σ^2 wird wie folgt definiert:

$$\sigma^2 = V(X) = E\left[(X-\mu)^2\right] = \begin{cases} \sum_i x_i^2 p(x_i) - \mu^2, & \text{wenn } X \text{ eine diskrete ZV ist} \\ \int_{-\infty}^{+\infty} \mathrm{x}^2 f(x)\, \mathrm{d}x - \mu^2, & \text{wenn } X \text{ eine stetige ZV ist} \end{cases} \tag{2.2}$$

Die positive Quadratwurzel $\sigma = \sqrt{V(X)}$ heißt Standardabweichung.

2.4 Diskrete Verteilungen

Die empirische Verteilung $F_n(x)$ einer Zufallsvariable X bezüglich ihrer Realisierungen x_i mit $i = 1, 2, \ldots, n$ ist gegeben durch:

$$F_n(x) = \frac{H_n(x)}{n}, \tag{2.3}$$

dabei bezeichnet $H_n(x)$ die Anzahl der Werte x_i, die höchstens gleich x sind. Die empirische Verteilungsfunktion an der Stelle x gibt somit die relative Häufigkeit dafür an, dass X Werte annimmt, die höchstens gleich x sind. Die relative Häufigkeit ist dabei der Quotient der Anzahl des Auftretens eines Wertes (= absolute Häufigkeit) und n.

Eine äquivalente Darstellung der empirischen Verteilungsfunktion lautet:

$$F_n(x) = \begin{cases} 0, & \text{wenn } x < x_{(1)} \\ \frac{i}{n}, & \text{wenn } x_{(i)} \le x < x_{(i+1)}, \quad i = 1, 2, \cdots, n-1 \\ 1, & \text{wenn } x_{(n)} \le x \end{cases} \tag{2.4}$$

Voraussetzung für diese Definition ist, dass die Werte der Größe nach aufsteigend geordnet werden. Schließlich entsteht eine Rangliste:

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n-1)} \le x_{(n)} \tag{2.5}$$

$x_{(i)}$ bezeichnet einen Wert an der i-ten Position (Storm (1995)). Die empirische Verteilungsfunktion ist eine monoton wachsende Treppenfunktion, wobei die Treppenhöhe an den beobachteten Werten um den Betrag $\frac{1}{n}$ wächst, wenn alle Werte voneinander verschieden sind. Sind dagegen r Werte gleich, so wächst sie an dem entsprechenden Wert um $\frac{r}{n}$.

Die Häufigkeitsverteilung von X ist durch Auszählung ihrer Werte gegeben. In diesem Fall wird zwischen absoluten und relativen Häufigkeiten unterschieden. Die graphische Darstellung der Häufigkeitsverteilung ist als Histogramm darstellbar. Dabei werden die

Werte in Klassen zusammengefasst und die relativen Häufigkeiten bezüglich dieser Klassen berechnet.

Hinzuzufügen ist, dass auch eine stetige Zufallsvariable durch die empirische Verteilungsfunktion und die Häufigkeitsverteilung dargestellt werden kann. Dazu ist aber eine Diskretisierung ihres Wertebereiches notwendig.

2.5 Stetige Verteilungen

Gegeben seien eine stetige Zufallsvariable X und ihre Realisierungen x_i mit $i = 1, 2, \ldots, n$.

Die gleichmäßige stetige Verteilung, die sogenannte Rechteckverteilung, besitzt die Wahrscheinlichkeitsdichte $f(x)$ und Verteilungsfunktion $F(x)$, wobei die Zufallsvariable Werte aus dem Intervall $[a; b]$ annimmt:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{wenn } \; a \leq x \leq b \\ 0, & \text{wenn } \; x < a, \; x > b, \; a < b \end{cases}$$

$$F(x) = \begin{cases} 0, & \text{wenn } \; x \leq a \\ \frac{x-a}{b-a}, & \text{wenn } \; a < x \leq b \\ 1, & \text{wenn } \; b < x \end{cases}$$

Ein beliebiger Wert aus diesem Intervall tritt mit gleicher Wahrscheinlichkeit auf. Abbildung 2.1 zeigt die Dichte- und Verteilungsfunktion einer auf $[a; b]$ gleichmäßig stetig verteilten Zufallsvariablen.

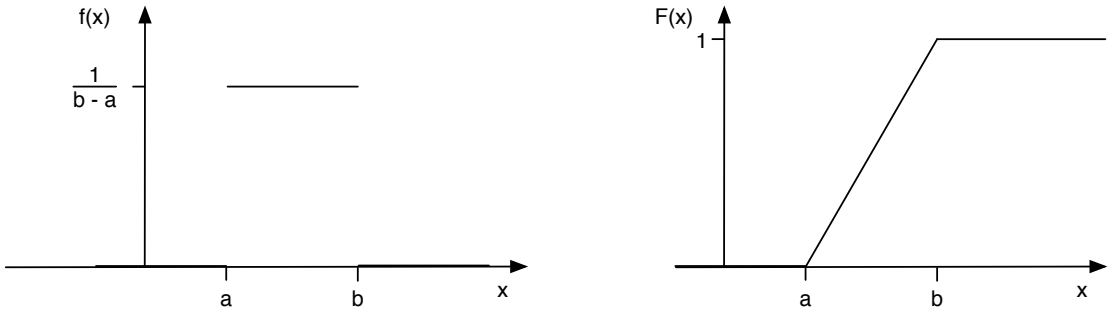


Abbildung 2.1 *: Dichte- und Verteilungsfunktion einer gleichmäßig stetig verteilten Zufallsvariablen.*

Der Normalverteilung wird eine besondere empirische Bedeutung zugesprochen, da viele Zufallsvariablen in der Praxis einem Verteilungsmodell angehören, deren Form nä-

herungsweise einer Normalverteilung entspricht. Zufallsvariablen treten mir einer relativen Häufigkeit auf, so dass das zugehörige Histogramm durch die Dichtefunktion der Normalverteilung approximiert wird.

Der zentrale Grenzwertsatz (Beyer u. a. (1991)) rechtfertigt die Modellierung von Zufallsvariablen durch eine Normalverteilung. Seien $X_1, X_2, \ldots, X_n$ unabhängige und identisch verteilte Zufallsvariablen mit dem Erwartungswert μ und der Varianz $\sigma^2 \neq 0$. Die Summe dieser Zufallsvariablen $S_n = \sum_{i=1}^{n} X_i$ ist selbst eine Zufallsvariable. Für S_n gilt der zentrale Grenzwertsatz: S_n ist für $n \to \infty$ annähernd normalverteilt. Daraus folgt auch, dass die Zufallsvariable $\bar{X} = \frac{S_n}{n}$ ebenfalls gegen eine Normalverteilung konvergiert. $\bar{X}$ ist nicht nur dann normalverteilt, wenn die Zufallsvariablen $X_1, X_2, \ldots, X_n$ normalverteilt sind, sondern bei „großen“ n auch dann, wenn diese keiner Normalverteilung folgen.

Eine stetige Zufallsvariable X unterliegt einer Normalverteilung mit den Kenngrößen Erwartungwert μ, und Varianz σ^2, wenn ihre Dichte durch

$$f(x) = \varphi\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty \tag{2.6}$$

gegeben ist.

Die zugehörige Verteilungsfunktion lautet:

$$F(x) = \Phi\left(x; \mu, \sigma^2\right) = P(X \leq x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \tag{2.7}$$

Die Dichtefunktion ist durch den Verlauf einer Glockenkurve charakterisiert. Diese nähert sich asymptotisch der Abszissenachse bei abfallenden und steigenden Werten x mit $x \to -\infty$ bzw. $x \to +\infty$ an. Die Dichtefunktion besitzt bei $x = \mu$ ihr Maximum und hat den Wert $\frac{1}{\sigma\sqrt{2\pi}}$. Sie ist symmetrisch bezüglich $x = \mu$. Die Abszissen der Wendepunkte sind $x_{1,2} = \mu \pm \sigma$. Eine Veränderung des Erwartungswertes μ bewirkt eine Parallelverschiebung der Dichtefunktion entlang der x-Achse. Eine Veränderung von σ bewirkt dagegen eine Stauchung der Dichtefunktion. Je größer σ ist, desto flacher verläuft die Dichtefunktion, siehe Abbildung 2.2.

Die Wahrscheinlichkeit einer Zufallsvariable X wird durch die Verteilungsfunktion bestimmt. Anschaulich gibt der Flächenanteil unter der Dichtefunktion im betrachteten Bereich die Wahrscheinlichkeit in diesem an. Die Gesamtfläche zwischen der Dichtefunktion und der x-Achse beträgt 1. Je größer σ ist, desto geringer ist die Wahrscheinlichkeit für Werte, die weit entfernt von μ sind. Umgekehrt gilt, je kleiner σ ist, desto stärker ist die Konzentration der „Wahrscheinlichkeitsmasse“ in der Umgebung von μ. Abbildung 2.3 stellt den Zusammenhang zwischen der Dichte- und Verteilungsfunktion am Beispiel

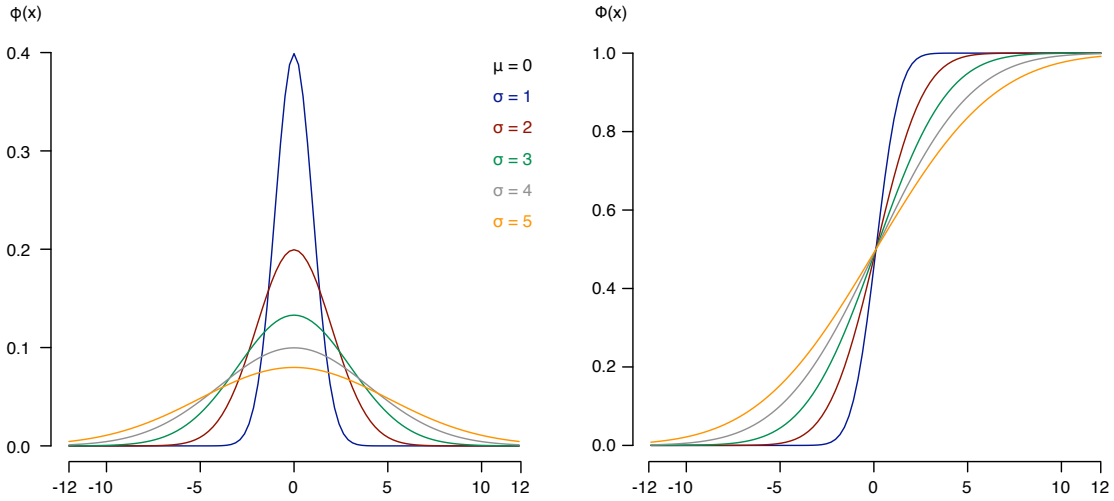


Abbildung 2.2 *: Dichtefunktion (φ) und Verteilungsfunktionen (Φ) von Normalverteilungen für unterschiedliche Standardabweichungen σ mit gleichem Erwartungswert μ.*

einer Normalverteilung dar. Tabelle 2.2 fasst zudem einige Wahrscheinlichkeitswerte für bestimmte Intervalle zusammen. Diese Wahrscheinlichkeiten gelten für alle Normalverteilungen bezüglich den entsprechenden Kenngrößen. Folgende Berechnungsformeln sind

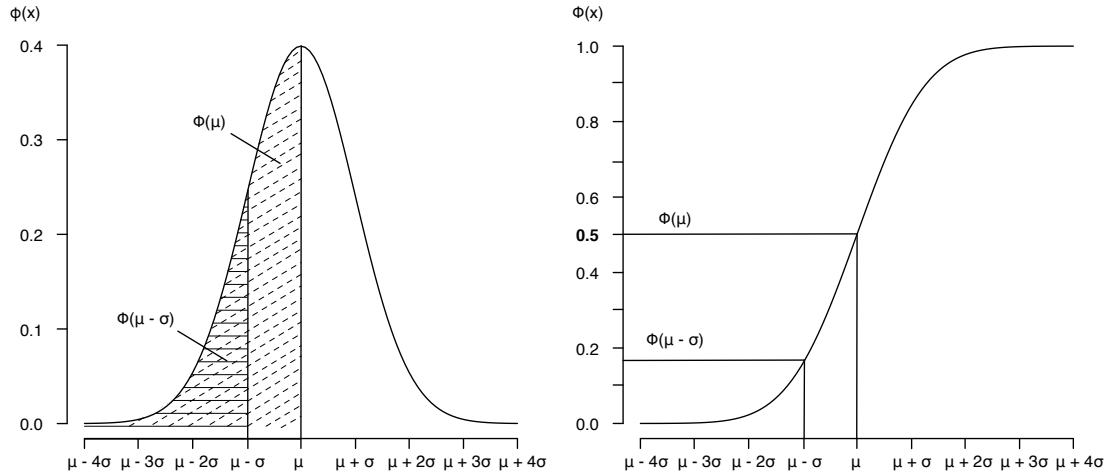


Abbildung 2.3 *: Dichtefunktion (φ) und Verteilungsfunktionen (Φ) der Standardnormalverteilung. Die Wahrscheinlichkeiten entsprechen der Fläche unter der Dichtefunktion.*

für die Wahrscheinlichkeit einer normalverteilten Zufallsvariablen X gegeben, wobei a

Intervall	**Flächenanteil (Wahrscheinlichkeit)**
$[\mu - \sigma; \mu]$	0.34
$[\mu; \mu + \sigma]$	0.34
$[\mu - \sigma; \mu + \sigma]$	0.68
$[\mu + \sigma; \mu + 2\sigma]$	0.14
$[\mu - \sigma; \mu - 2\sigma]$	0.14
$[\mu - 2\sigma; \mu + 2\sigma]$	0.95
$[\mu - 3\sigma; \mu + 3\sigma]$	0.99

***Tabelle 2.2** : Wahrscheinlichkeiten als Flächenanteile in bestimmten Intervallen. Die Angaben sind ungefähr auf* 1% *genau (Sachs u. Hedderich (2006)).*

und b reelle Zahlen sind:

$$\begin{aligned} P(a \leq X \leq b) &= \Phi\left(b; \mu, \sigma^2\right) - \Phi\left(a; \mu, \sigma^2\right) \\ P(X \leq b) &= \Phi\left(b; \mu, \sigma^2\right) \\ P(X \geq a) &= 1 - P(X < a) = 1 - \Phi\left(a; \mu, \sigma^2\right) \end{aligned} \tag{2.8}$$

Eine Normalverteilung mit $\mu = 0$ und $\sigma = 1$ heißt standardisierte Normalverteilung oder Standardnormalverteilung. Die Dichtefunktion ist gegeben durch:

$$f(x) = \varphi(x; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \tag{2.9}$$

Die Verteilungsfunktion lautet:

$$F(x) = \Phi(x; 0, 1) = P(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt \tag{2.10}$$

Jede Normalverteilung mit beliebigem Erwartungswert und Varianz lässt sich in die Standardnormalverteilung überführen, indem die Zufallsvariable X folgendermaßen standardisiert wird:

$$U = \frac{X - \mu}{\sigma} \tag{2.11}$$

Für diese Zufallsvariable gilt:

$$f(x) = \frac{1}{\sigma} \varphi\left(\frac{x - \mu}{\sigma}; 0, 1\right) \tag{2.12}$$

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}; 0, 1\right) \tag{2.13}$$

Für beide Funktionen einer Normalverteilung, der Dichte- und der Verteilungsfunktion, gibt es Tabellen, aus deren für vorgegebene x-Werte der dazugehörige Funktionswert abzulesen ist. Die Werte der Verteilungsfunktion lassen sich auch approximativ berechnen, z.B. durch folgende polynomiale Approximation nach Abramowitz u. Stegun (1964):

$$\Phi(x; 0, 1) = 1 - \varphi(x; 0, 1)\left(a_1 t + a_2 t^2 + a_3 t^3\right) + \epsilon(x), \quad t = \frac{1}{1+px}, \quad |\epsilon(x)| < 10^{-5}$$
$$p = 0.33267, \; a_1 = 0.4361836, \; a_2 = -0.1201676, \; a_3 = 0.9372980 \tag{2.14}$$

2.6 Grundgesamtheit, Stichprobe, Schätzer und Tests

Eine Zufallsvariable X mit der zugehörigen Wahrscheinlichkeitsverteilung wird als Grundgesamtheit bezeichnet. Die Grundgesamtheit beschreibt folglich die Menge aller möglichen Realisierungen der Zufallsvariable. Eine n-elementige Teilmenge einer Grundgesamtheit von X heißt konkrete Stichprobe. Jede Realisierung wird als Element der Stichprobe bezeichnet. Wird das Zufallsexperiment wiederholt, so sind die Realisierungen von Stichprobe zu Stichprobe unterschiedlich. Das bedeutet, dass der erste Wert der ersten Stichprobe unterschiedlich von dem ersten Wert der zweiten Stichprobe ist. Somit ist der erste Wert ebenfalls eine Zufallsvariable. Analog werden auch die restlichen Werte als Zufallsvariablen betrachtet. Diese heißen Stichprobenvariablen. Aus der in einer Stichprobe enthaltenen Informationen werden Aussagen über die Grundgesamtheit gemacht. Folgende Fälle treten hierbei auf:

1. Schätzen der Kenngrößen der Grundgesamtheit.

   Falls nur der Verteilungstyp der Grundgesamtheit bekannt ist, werden die unbekannten Kennwerte dieser Verteilung geschätzt.

2. Prüfen von Hypothesen.

   Hypothesen über die Verteilung oder deren Kenngrößen sind mit den aus der Stichprobe entnommenen Informationen so in Verbindung zu bringen, dass darüber entschieden werden kann, ob sie mit dieser Information vereinbar sind (Beyer u. a. (1991)).

Im folgenden Abschnitt wird das Schätzen der Kenngrößen der Grundgesamtheit betrachtet, um im Anschluss das Prüfen von Hypothesen zu behandeln.

2.6.1 Schätzer

Durch Näherungswerte, sogenannte Schätzwerte, werden unbekannten Kenngrößen bestimmt. Die Vorschrift zur Berechnung eines Schätzwertes wird als Schätzfunktion oder Punktschätzung bezeichnet. Diese ist eine Funktion der Stichprobe. Eine Schätzfunktion sollte folgenden Kriterien genügen (Beichelt u. Montgomery (2003)):

1. Erwartungstreue (Unverzerrtheit).

   Eine Schätzfunktion $\hat{\theta}$ eines Parameters θ ist erwartungstreu oder unverzerrt, wenn der Erwartungswert von $\hat{\theta}$ gleich dem Parameter θ ist. $\hat{\theta}$ ist asymptotisch erwartungtreu, falls bei wachsendem Stichprobenumfang der Grenzwert des Erwartungswertes von $\hat{\theta}$ gleich dem Parameter θ ist.

2. Konsistenz.

   Eine Schätzfunktion $\hat{\theta}$ eines Parameters θ ist konsistent, wenn $\hat{\theta}$ mit wachsendem Stichprobenumfang in Wahrscheinlichkeit gegen θ konvergiert.

3. Wirksamkeit.

   Seien $\hat{\theta}_1$ und $\hat{\theta}_2$ erwartungstreue Schätzfunktionen für θ. $\hat{\theta}_1$ heißt wirksamer oder effizienter als $\hat{\theta}_2$, wenn für ihre Varianzen $V(\hat{\theta}_1)$ und $V(\hat{\theta}_2)$ $V(\hat{\theta}_1) < V(\hat{\theta}_2)$ gilt. Das Verhältnis $\eta = \frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)}$ wird als relativer Wirkungsgrad von $\hat{\theta}_2$ in Bezug auf $\hat{\theta}_1$ bezeichnet.

4. Asymptotische Normalverteilung.

   Eine Schätzfunktion $\hat{\theta}$ ist bei wachsendem Stichprobenumfang asymptotisch normalverteilt, wenn die standardisierte Schätzfunktion gegen die standardisierte Normalverteilung strebt.

Das arithmetische Mittel $\bar{X}$ einer Stichprobe vom Umfang n, siehe Formel 2.15, ist eine erwartungstreue, konsistente, effiziente und asymptotische normalverteilte Schätzfunktion für den unbekannten Erwartungswert μ der Zufallsvariable in der Grundgesamtheit.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.15}$$

Die Stichprobenvarianz oder empirische Varianz, siehe Formel 2.16, ist eine erwartungstreue und asymptotisch effiziente Schätzfunktion für die unbekannte Varianz σ^2 der Zufallsvariable in der Grundgesamtheit.

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \tag{2.16}$$

Die positive Quadratwurzel von s^2 wird als empirische Standardabweichung bezeichnet.

Eine geeignete Schätzfunktion für die unbekannte, theoretische Verteilungsfunktion der Grundgesamtheit stellt die empirische Verteilungsfunktion dar, siehe Formel 2.3. Der Satz von Gliwenko besagt, dass die empirische Verteilungsfunktion der Stichprobe vom Umfang n für $n \to \infty$ mit Wahrscheinlichkeit gleichmäßig für alle x gegen die theoretische Verteilungsfunktion der Grundgesamtheit konvergiert (Beyer u. a. (1991)). Daraus folgt, dass für „große“ n die empirische Verteilungsfunktion eine geeignete Schätzfunktion für die theoretische Verteilungsfunktion ist.

2.6.2 Statistische Tests

Das Prüfen von Hypothesen erfolgt durch statistische Tests. Diese prüfen die Verteilung einer Grundgesamtheit einer Stichprobenvariable oder ihren Kenngrößen anhand einer gezogenen Stichprobe. Die aus der Stichprobe entnommenen Informationen werden herangezogen, um Entscheidungen über Hypothesen zu treffen. Eine Hypothese ist eine Annahme über unbekannte Eigenschaften der Grundgesamtheit. Es wird zunächst eine Nullhypothese H_0 aufgestellt, deren Annahme oder Ablehnung durch die Überprüfung der Stichprobe festgelegt wird. Neben H_0 können auch eine oder mehrere Alternativhypothesen behandelt werden.

Falls die Verteilungsfunktion der Grundgesamtheit bekannt ist und die Hypothese über unbekannte Kenngrößen dieser Verteilung aufgestellt wird, so handelt es sich um eine Parameterhypothese. Statistische Tests, die diese Hypothese prüfen, heißen Parametertests. Bezieht sich dagegen die Hypothese auf die unbekannte Verteilung der Grundgesamtheit, so wird sie als Verteilungshypothese oder nichtparametrische Hypothese bezeichnet. Entsprechend wird von nichtparametrischen, parameterfreien oder verteilungsfreien Tests gesprochen.

Aufgrund der Tatsache, dass der Ausgang des Tests auf Informationen der Stichprobe basiert, können zwei Fehler begangen werden:

- Fehler erster Art: die Nullhypothese wird abgelehnt, obwohl sie richtig ist.
- Fehler zweiter Art: die Nullhypothese wird angenommen, obwohl sie falsch ist.

Die Wahrscheinlichkeit den Fehler erster Art zu begehen, heißt Irrtumswahrscheinlichkeit. Diese nimmt den vorzugebenden Wert α an. Typische Werte für α sind 0.05 oder 0.01. Die statistische Sicherheit ist somit durch $(1 - \alpha)$ definiert. Durch die Irrtumswahrscheinlichkeit wird der Ablehnungsbereich, der sogenannte kritische Bereich, für die

Hypothese festgelegt. Wird α kleiner als 0.01 gewählt, wird der Nachweis einer falschen Hypothese schwieriger. Dadurch wächst in der Folge die Wahrscheinlichkeit eines Fehlers zweiter Art. Dieser Fehler heißt Güte oder Trennschärfe des Prüfverfahrens und nimmt den Wert β an.

Für die Annahme oder Ablehnung der Nullhypothese wird eine Prüfgröße verwendet. Die Prüfgröße ist eine Stichprobenfunktion einer zur betrachteten Grundgesamtheit gehörenden Stichprobe. Steht die Prüfgröße im Widerspruch zu dem Ablehnungsbereich, besteht kein Grund zur Ablehnung der Nullhypothese.

Im Folgenden werden zwei parameterfreie Tests, der χ^2- und der Kolmogorov-Smirnov-Test, zur Überprüfung der Hypothese, dass die Grundgesamtheit der betrachteten Stichprobe einer bestimmten Verteilung folgt, behandelt. Der χ^2-Test ist für beliebige Verteilungen durchführbar, wogegen der Kolmogorov-Smirnov-Test nur bei stetigen Verteilungsfunktionen anwendbar ist. In den folgenden Abschnitten wird beschrieben wie beide Tests die Verteilungsfunktion einer Grundgesamtheit auf eine Normalverteilung überprüfen.

2.6.2.1 χ^2- Test

Durch den χ^2-Test wird die unbekannte Verteilungsfunktion der Grundgesamtheit der Stichprobe auf die Normalverteilungsfunktion der gleichen Grundgesamtheit geprüft. Die Nullhypothese H_0 sagt folglich aus, ob die Stichprobe in der Grundgesamtheit normalverteilt ist:

$$H_0 : F(x) = \Phi\left(x; \mu, \sigma^2\right)$$

Dabei ist F die unbekannte Verteilungsfunktion der Stichprobe und Φ die Verteilungsfunktion der Normalverteilung mit den Parametern μ und σ^2, siehe Formel 2.7. Beide Kenngrößen werden aus der Stichprobe geschätzt, d.h. für den Erwartungswert wird der Mittelwert durch Formel 2.15 berechnet und für die Varianz wird die empirische Varianz durch Formel 2.16 verwendet.

Der Wertebereich der Stichprobe wird in k äquidistanten Klassen unterteilt. Die unbekannte Verteilungsfunktion wird durch die absolute Häufigkeit h_m mit $(m = 1, 2, \ldots k)$ der Werte in den Klassen repräsentiert. Diese gibt die Anzahl der Werte in jeder Klasse an, wie Abschnitt 2.4 beschrieben.

Mit Hilfe von $\Phi\left(x; \mu, \sigma^2\right)$ lässt sich die Wahrscheinlichkeit p_m dafür berechnen, dass ein Wert x in die m-te Klasse fällt. p_m gibt also den Flächenanteil unter der Normalverteilungskurve zwischen der oberen und unteren Klassengrenze einer Klasse m an. Um

diese Fläche zu ermitteln, werden die Stichprobenwerte standardisiert, um auf eine Standardnormalverteilung prüfen zu können.

Für p_m gilt nach den Formeln 2.8 folgende Gleichung:

$$p_m = P\left(x \in k_m\right) = P\left(\frac{o_{m-1} - \mu}{\sigma} \leq x \leq \frac{o_m - \mu}{\sigma}\right) = \Phi\left(\frac{o_m - \mu}{\sigma}\right) - \Phi\left(\frac{o_{m-1} - \mu}{\sigma}\right)$$

Dabei ist o_m die obere und o_{m-1} die untere Klassengrenze der m-ten Klasse. Das Produkt von n und p_m gibt die theoretische Häufigkeit an. Die Bedingung
$np_m \geq 5$ muss erfüllt sein. Ist sie nicht erfüllt, werden Klassen zusammengefasst, indem die entsprechenden empirischen und theoretischen Häufigkeiten jeweils aufsummiert und die Klassengrenzen angepasst werden.

Die Differenz zwischen der empirischen und der theoretischen Häufigkeit gibt die Abweichung zwischen der empirischen und theoretischen Verteilung der Stichprobe an. Die Prüfgröße χ^2_S stellt die Summe der normierten Abweichungen dar und wird wie folgt berechnet:

$$\chi^2_S = \sum_{m=1}^{k} \frac{\left(h_m - np_m\right)^2}{np_m} \tag{2.17}$$

Die Prüfung auf Normalverteilung schließt ein, durch die vorgegebene Irrtumswahrscheinlichkeit α das Quantil $\chi^2_{\alpha;k-r-1}$ der χ^2-Verteilung aus entsprechenden Tabellen zu entnehmen. Die Freiheitsgrade sind durch $(k - r - 1)$ gegeben, r ist dabei die Anzahl der geschätzten Parameter der angenommenen Verteilungsfunktion. Es gilt $r = 2$ wegen der Schätzung von μ und σ^2.

Der Quantilwert ist ebenfalls nach folgender Formel zu berechnen. f bezeichnet die Freiheitsgrade und z_q ist dabei das Quantil der Ordnung p der standardisierten Normalverteilung.

$$\chi^2_{f,p} \approx \frac{\left(\sqrt{2 \cdot f - 1} + z_p\right)^2}{2}$$

Die Nullhypothese wird abgelehnt, falls:

$$\chi^2_S > \chi^2_{\alpha;k-r-1}$$

2.6.2.2 Kolmogorov-Smirnov-Test

Die Nullhypothese überprüft die unbekannte Verteilungsfunktion auf eine Normalverteilung. Der Test basiert auf dem maximalen Abstand zwischen der empirischen Verteilungsfunktion und der theoretischen Verteilung, der Normalverteilung.

Die empirische Verteilung ist durch Formel 2.3 gegeben. Die theoretische Verteilung wird durch die Fläche unter der Standardnormalverteilung bestimmt. Anschließend werden die Funktionswerte der entsprechenden Tabellen entnommen oder approximiert.

Im nächsten Schritt werden Differenzen D'_i und D''_i zwischen der empirischen und der theoretischen Verteilungsfunktion an den Sprungstellen $x_{(i)}$ berechnet. $x_{(i)}$ gibt den Wert an der i-ten Position der Rangliste an.

$$D'_i = |F\left(x_{(i-1)}\right) - \Phi\left(x_{(i)}; \mu, \sigma^2\right)| = |\frac{i-1}{n} - P\left(x_{(i)}\right)|$$
$$D''_i = |F\left(x_{(i)}\right) - \Phi\left(x_{(i)}; \mu, \sigma^2\right)| = |\frac{i}{n} - P\left(x_{(i)}\right)|$$

Der größte Wert dieser Differenzen stellt die sogenannte D-Statistik D_n dar:

$$D_n = \max_i \left(D'_i, D''_i\right)$$

Die D-Statistik wird mit einem kritischen Wert verglichen, um eine Entscheidung über die Nullhypothese zu treffen. Der kritische Wert ist aus Tabellen zu entnehmen oder für beliebige Werte der Irrtumswahrscheinlichkeit α nach folgender Formel

$$D_{\alpha,n} = \sqrt{\frac{-0.5 \cdot \ln \frac{\alpha}{2}}{n}}$$

zu berechnen. Die Nullhypothese wird abgelehnt, wenn die D-Statistik größer als der kritische Wert ist.

Voraussetzung für den Test ist die Stetigkeit der hypothetischen Verteilung, da in diesem Fall die Verteilung der Testgröße davon unabhängig ist. Die Größen D'_i, D''_i und D_n sind verteilungsfrei, da sie nur vom Stichprobenumfang abhängen. Außerdem muss die theoretische Verteilungsfunktion vollständig spezifiziert werden, da der Fehler zweiter Art sonst groß wird. Falls die Parameter der hypothetischen Funktion, hier Mittelwert und Varianz, aus der Stichprobe geschätzt werden, werden die kritischen Werte „konservativ“ oder ungenau.

Eine Variante des Kolmogorov-Smirnov Tests ist der Lillifors Test. Dieser verläuft nach dem gleichen Schema, wobei die Bestimmung der kritischen Werte exakter ist. Lillifors bestimmte diese durch Simulationsstudien. Die Resultate, also die simulierten Quantile, wurden in Tabellen festgehalten. Die kritischen Werte wurden dabei auf Basis von 1000 zufälligen Beispielen und für bestimmte Werte von α $(0.2, 0.15, 0.10, 0.05, 0.01)$ erstellt.

Abdi u. Molin (2007) präsentieren ein Verfahren zur Berechnung der kritischen Werte als eine Funktion des Stichprobenumfangs und der Irrtumswahrscheinlichkeit. Dieses Papier stellt eine alternative Methode vor, um den Ausgang des Tests zu entscheiden. Bei der

Testdürchführung werden nicht die kritischen Werte bestimmt, sondern die Wahrscheinlichkeit einer bestimmten Prüfgröße. Ist diese kleiner als die Irrtumswahrscheinlichkeit, wird die Nullhypothese abgelehnt. Diese Methode ist auf zwei Nachkommastellen genau. Hierfür wird zunächst Größe A berechnet:

$$
\begin{aligned}
A =& \frac{-\left(b_1 + n\right) + \sqrt{\left(b_1 + n\right)^2 - 4b_2\left(b_0 - D_n^{-2}\right)}}{2b_2} \\
b_0 =& 0.08861783849346, \\
b_1 =& 1.30748185078790, \\
b_2 =& 0.08861783849346
\end{aligned}
$$

Schließlich wird die Wahrscheinlichkeit der Prüfgröße $P\left(D_n\right)$ bestimmt.

$$
\begin{aligned}
P\left(D_n\right) =& - 0.37782822932809A^0 + 1.67819837908004A^1 \\
& - 3.02959249450445A^2 + 2.80015798142101A^3 \\
& - 1.39874347510845A^4 + 0.40466213484419A^5 \\
& - 0.06353440854207A^6 + 0.00287462087623A^7 \\
& + 0.00069650013110A^8 - 0.00011872227037A^9 \\
& + 0.00000575586834A^{10} + \epsilon
\end{aligned}
$$

Falls $P\left(D_n\right)$ kleiner als die Irrtumswahrscheinlichkeit ist, wird die Nullhypothese abgelehnt, anderenfalls angenommen.

Mason u. Bell (1986) haben ebenfalls Formeln für die Bestimmung der kritischen Werte entwickelt. Diese sind nur für bestimmte α angegeben, siehe Tabelle 2.3.

α	$\frac{Wert}{D_n}$
0.01	1.035
0.05	0.895
0.10	0.819
0.15	0.775
0.20	0.741

Tabelle 2.3 *: Kritische Werte für D_n nach Mason u. Bell (1986).*

Gegenüberstellung des χ^2- und des Kolmogorov-Smirnov-Tests Bei dem χ^2-Test werden unbekannte Parameter geschätzt. Werden diese bei dem Kolmogorov-Smirnov-Test geschätzt, ist die Variante von Lilliefors zu benutzen. Bei dem χ^2-Test gehen Informationen über die Stichprobe durch die Klasseneinteilung verloren, da nach Bosch (1993) nicht die Verteilung, sondern daraus abgeleitete Klassenwahrscheinlichkeiten getestet werden, während der Kolmogorov-Smirnov-Test die Stichprobenwerte ohne Informationsverlust über die empirische Verteilungsfunktion benutzt.

Der Kolmogorov-Smirnov-Test ist bei kleinerem Stichprobenumfang dem χ^2-Test vorzuziehen. Der χ^2-Test sollte dagegen bei großem Stichprobenumfang benutzt werden, da die Testgröße asymptotisch χ^2-verteilt ist. Der χ^2-Test ist konsistent und unverfälscht, der Kolmogorov-Smirnov-Test ist konsistent, aber nicht unverfälscht. Die Konsistenz sichert in diesem Zusammenhang, dass der Fehler zweiter Art mit wachsendem Stichprobenumfang gegen Null konvergiert. Die Unverfälschtheit besagt, dass die Wahrscheinlichkeit, eine wahre Nullhypothese abzulehnen, nie größer ist, als wenn die Nullhypothese falsch ist (Bosch (1993)).

3 Plagiaterkennung

Ausgangspunkt bei der Plagiaterkennung ist ein verdächtiges Dokument, welches auf mögliche Plagiatstellen hin untersucht wird. Bei der Analyse des Dokumentes wird dieses entweder lokal oder global betrachtet. Bei der erstgenannten Herangehensweise werden die Textabschnitte des Dokumentes untersucht, bei letztgenannten das ganze Dokument als eine Einheit. Dementsprechend werden Erkennungsmethoden Kategorien der lokalen und der globalen Dokumentenanalyse zugeordnet. Diese Erkennungsmethoden ziehen bei der Untersuchung eines Dokumentes eine Dokumentensammlung zum Vergleich heran, um ein Plagiatvergehen festzustellen. Im Fall, dass keine Dokumentensammlung vorliegt, wird versucht Plagiate intrinsisch, also aus dem untersuchten Dokument heraus, zu erkennen. Meyer zu Eissen u. a. (2007) fassen unterschiedliche Plagiatvergehen und korrespondierende Erkennungsmethoden in einer Taxonomie in Abbildung 3.1 zusammen.

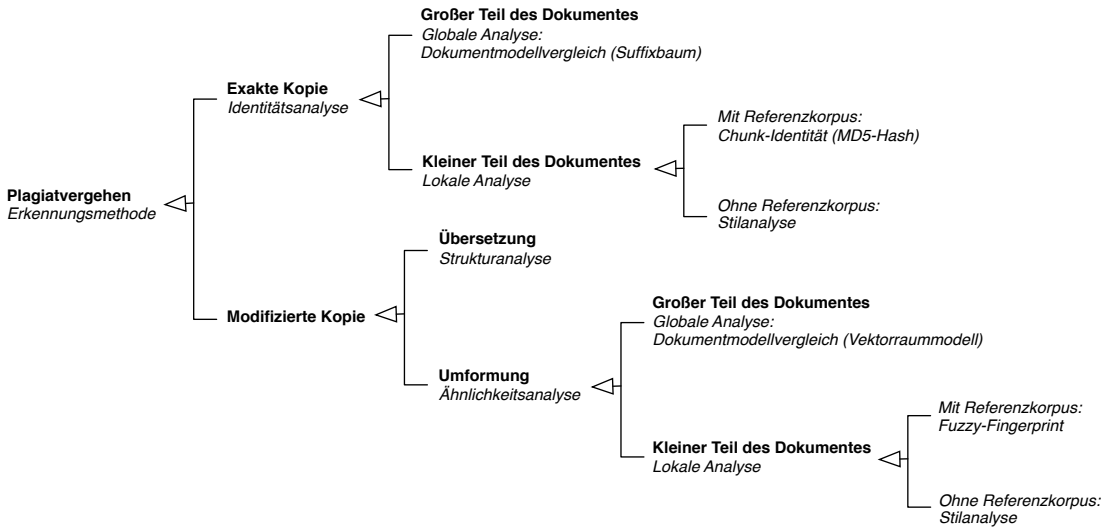


Abbildung 3.1 *: Taxonomie von Plagiatvergehen in Verbindung mit entsprechenden Erkennungsmethoden nach Meyer zu Eissen u. a. (2007).*

Die Plagiaterkennung mit einer Dokumentensammlung ist zudem als drei-stufiger Prozess darstellbar (Stein u. a. (2007)):

1. Heuristisches Retrieval.

   Abschnitte eines untersuchten Dokumentes werden in einem Referenzkorpus gesucht. Die zurückgegebenen Kandidatendokumente, falls vorhanden, enthalten die entsprechenden Originalpassagen.

2. Detaillierte Analyse.

   Die möglicherweise plagiierten Abschnitte jedes Kandidatendokumentes werden mit dem Anfragedokument verglichen. Wird eine hohe Ähnlichkeit festgestellt, bestätigt dies den Plagiatverdacht.

3. Wissensbasierte Nachbearbeitung.

   Die ähnlichen Abschnitte werden bezüglich ordnungsgemäßer Angaben wie Quellennachweise oder Zitate untersucht. Entsprechend wird ein Plagiatvergehen festgestellt oder verneint.

Erkennungsmethoden werden im zweiten Schritt des Erkennungsprozesses, der detaillierten Analyse, angewandt. Die Plagiatenerkennung ohne Referenzkorpus basiert auf der Untersuchung des Schreibstils eines Autors, der Stilanalyse. Diese bildet den Schwerpunkt dieser Arbeit und wird im Kapitel 4 vorgestellt. Im Folgenden wird entsprechend der Taxonomie 3.1 eine Übersicht der Erkennungsmethoden gegeben. Deren Ziel ist es, identische oder ähnliche Dokumente zu finden. Diese deuten, unter Verwendung einer Dokumentensammlung, auf mögliche Plagiate hin.

Chunk Identity Dieses Verfahren berechnet mittels einer Hashfunktion Hashwerte für jeden Abschnitt eines Dokumentes. Zwei Abschnitte sind identisch, wenn die berechneten Hashwerte übereinstimmen.

Fuzzy-Fingerprinting Beim Fuzzy-Fingerprint-Verfahren handelt es sich um einen hashbasierten Ansatz, welcher eine Variante des Similarity-Hashings darstellt. Es wird hierbei davon ausgegangen, dass bei einer Kollision der Hashwerte die korrespondierenden Abschnitte eine hohe Ähnlichkeit zueinander aufweisen. Der Ansatz der Chunk Identity wird auf diese Weise für ähnliche Zeichenfolgen erweitert. Eine detaillierte Beschreibung der Vorgehensweise zur Plagiaterkennung mittels Fuzzy-Fingerprinting ist in Stein u. Meyer zu Eißen (2006) zu finden.

Vektorraummodell Basierend auf der Vereinigungsmenge aller in einer Menge von Dokumenten enthaltenen Wörter wird ein Wörterbuch gebildet. Dessen Einträge spannen einen Vektorraum auf. Jedes Dokument wird durch einen Wortvektor im Vektorraum repräsentiert. Ein Wortvektor enthält somit Informationen über die im Dokument enthaltenen und nicht enthaltenen Wörter. Sollen zwei Dokumente miteinander verglichen werden, so wird der Kosinus des Winkels zwischen den normierten Wortvektoren berechnet. Je kleiner der Winkel ausfällt, desto ähnlicher sind beide Dokumente. Das normierte Skalarprodukt beider Wortvektoren heißt Kosinus-Ähnlichkeitsmaß. Andere Maße, die im Vektorraummodell zur Anwendung kommen, sind zum Beispiel das Pseudo-Kosinusmaß, das Dice-Maß, das Überdeckungsmaß oder das Jaccard-Maß (Ferber (2003)).

Suffixbaummodell Das Suffixbaummodell berücksichtigt die Beziehung der Wörter eines Dokumentes untereinander, das bedeutet, dass die natürliche Reihenfolge der Wörter im Gegensatz zum Vektorraummodell beibehalten wird. Ein Suffixbaum ist eine graphbasierte Datenstruktur. Ein Dokument d wird als Folge von Wörtern $d = w_1 \dots w_m$ betrachtet. Der i-te Suffix des Dokumentes ist diejenige Teilfolge von d, die mit dem Wort w_i anfängt. Ein Suffixbaum von d ist ein beschrifteter Baum, der jeden Suffix von d entlang des Pfades enthält, der durch Kanten mit den Wörtern des Suffixes beschriftet ist. Die Ähnlichkeit zweier Dokumente wird durch Überlappungen von Zeichenfolgen im gemeinsamen Suffixbaum der beiden Dokumente festgelegt. Als Ähnlichkeitsmaß dient ein angepasstes Jaccard-Maß. Dieses Maß erfasst Übereinstimmungen der Wortreihenfolge. Das Maß setzt die Schnittmenge und die Vereinigung der Kantenmengen mit verschiedener Beschriftung ins Verhältnis. Die Beschriftung ist davon abhängig, ob die Kanten der jeweiligen Menge während des Einfügens von Suffixen des jeweiligen Dokuments im Suffixbaum traversiert werden (Meyer zu Eißen u. a. (2005)).

Plagiatvergehen, die durch Übersetzung eines Dokumentes oder Teile davon in einer anderen Sprache entstehen, werden durch eine sprachübergreifende Ähnlichkeitsanalyse erkannt. Potthast u. a. (2008) stellen hierzu ein Verfahren zur sprachübergreifenden Plagiaterkennung, die Cross-Language-Explicit-Semantic-Analyse (CL-ESA), vor.

Cross-Language-Explicit-Semantic-Analyse Bei diesem Verfahren ermöglicht ein sprachübergreifendes Konzeptraummodell den Inhalt von verschiedensprachigen Dokumenten einheitlich darzustellen und zu vergleichen. Ein Konzept ist ein sprachunabhängiges Merkmal, für dieses demnach in jeder Sprache eine Repräsentation existiert. Grundlage des Konzeptraummodells ist eine multilinguale Wissensmenge, die Konzepte in verschie-

denen Sprachen enthält. Es werden sogenannte Stützvektoren erstellt, die das Konzept in jeder Sprache repräsentieren. Zur Berechnung des Konzeptvektors für ein Dokument werden diejenigen Stützvektoren verwendet, die die Konzepte der Sprache des Dokumentes darstellen. Um eine Ähnlichkeit zwischen zwei verschiedensprachigen Dokumenten zu bestimmen, wird wiederum der Kosinus des Winkels zwischen den Konzeptvektoren berechnet.

4 Modell zur Erfassung von Stiländerungen als Ausreißer

Die intrinsische Plagiaterkennung basiert auf dem Verfahren der Stilanalyse, das Stiländerungen innerhalb eines betrachteten Dokumentes erfasst. Dieses wird in Bezug auf Änderungen im Schreibstil analysiert. Zu diesem Zweck werden explizit die Abschnitte des Dokumentes untersucht. Ein Ausreißer wird dabei als Stilwechsel interpretiert, der potentiell plagiierte Stellen darstellt.

Der Schreibstil jedes Abschnittes wird mit Hilfe quantifizierbarer Stilmerkmale berechnet. Diese basieren nach Meyer zu Eissen u. a. (2007) auf folgenden semiotischen Charakteristika:

1. Auf Zeichenebene definierte Textstatistiken.

   Beispiele hierfür sind die Anzahl der Kommata, die Anzahl der Fragezeichen und die Wortlänge.

2. Auf Satzebene definierte syntaktische Merkmale.

   Untersucht wird die Satzlänge und es werden Statistiken über die Benutzung von Funktionswörter erhoben.

3. Wortart-Merkmale zur Quantifizierung der Benutzung von Wortklassen.

   Auszugsweise ist die Anzahl der Adjektive und die der Pronomen zu nennen.

4. Geschlossene Mengen, die bestimmte Wortklassen zählen.

   Beispiele hierfür sind die Anzahl der Stoppwörter und die der Fremdwörter.

5. Strukturelle Merkmale, die die Textorganisation beschreiben.

   Interessant sind hier beispielsweise die Paragraph- und die Kapitellänge.

Zur Feststellung von Schreibstiländerungen wird eine formal-mathematische Sicht für das Dokument benötigt. Stein u. Meyer zu Eissen (2007) stellen hierzu eine Stilmodellrepräsentation für das Dokument vor. Sei d das untersuchte Dokument und sei weiterhin

$s_1, \ldots, s_n$ eine Zerlegung von d in n nicht unterbrochene sich nicht überlappende Abschnitte. Sei $\sigma_1, \ldots, \sigma_m$ eine Menge von Stilmerkmalen, von denen jedes einen Stilaspekt des Schreibstils quantifiziert, wobei jedes Merkmal einen reellen Wert zu einem Abschnitt $s \subseteq d$ zuordnet:

$$\sigma_i : s \mapsto \sigma_i(s) \in \mathbf{R}$$

Die Stilmodellrepräsentation $\mathbf{s}$ eines Abschnittes s wird als m-dimensionaler Vektor dargestellt:

$$\mathbf{s} = \begin{pmatrix} \sigma_1(s) \\ \vdots \\ \sigma_m(s) \end{pmatrix}, s \subseteq d$$

Die Hypothese ist, dass wenn ein Abschnitt $s^- \subset d$ plagiiert ist, unterscheidet sich seine Stilmodellrepräsentation $\mathbf{s}^-$ deutlich von der Stilmodellrepräsentation $\mathbf{s}^+$, die die nicht-plagiierten Abschnitte $s^+ \subset d$ beschreibt. Um diesen Unterschied zu erfassen, wird für jeden Abschnitt s der Vektor $\mathbf{s_\Delta}$ der relativen Abweichungen seiner Stilmerkmale vom Mittelwert des Dokuments wie folgt berechnet:

$$\mathbf{s_\Delta} = \begin{pmatrix} \frac{\sigma_1(s)-\sigma_1(d)}{\sigma_1(d)} \\ \vdots \\ \frac{\sigma_m(s)-\sigma_m(d)}{\sigma_m(d)} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \; x_i = \frac{\sigma_i(s) - \sigma_i(d)}{\sigma_i(d)}, \; s \subseteq d \tag{4.1}$$

Ein Beispiel zur Verdeutlichung dieses Sachverhalts. Ein Diplomand schreibt überwiegend kurze Sätze. Ein Abschnitt in seiner Arbeit besteht jedoch aus langen Sätzen. Das bedeutet, dass bezüglich des Merkmals „Satzlänge“ die Stilmodellrepräsentation des betrachteten Abschnittes stark von der Stilmodellrepräsentation des ganzen Dokumentes abweicht. Diese Abweichung ist als Ausreißer interpretierbar. Werden bezüglich verschiedene Stilmerkmale Ausreißer in diesem Abschnitt entdeckt, so bekräftigen diese die Aussage, dass dieser plagiiert ist. Ziel dieser Arbeit ist es, solche Ausreißer innerhalb eines Abschnittes zu erkennen, um über diesen ein korrektes Plagiaturteil zu fällen.

Für die weitere Betrachtung wird das untersuchte Dokument durch eine $m \times n$-Matrix, siehe Matrix 4.3, dargestellt. Dabei gibt m die Anzahl der Stilmerkmale und n die Anzahl der Abschnitte des Dokumentes an. Die Vektoren der Form 4.1 sind die Zeilenvektoren der Matrix. Ein Element $x_{j,i}$ stellt dabei den Wert des j-ten Abschnittes für die Abweichung bezüglich des i-ten Stilmerkmals dar:

$$x_{j,i} = \frac{\sigma_i(s_j) - \sigma_i(d)}{\sigma_i(d)}, \; s_j \subseteq d \tag{4.2}$$

$$d = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-1,1} & x_{n-1,2} & \cdots & x_{n-1,m} \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix} \tag{4.3}$$

5 Funktionen zur Eine-Klassen-Klassifikation von Plagiatstellen als Ausreißer

Die Plagiatanalyse verfolgt das Ziel Plagiate innerhalb eines Dokumentes automatisch zu erkennen. Eine Möglichkeit zur Erkennung ist die Anwendung von Lernverfahren aus dem Gebiet des maschinellen Lernens. Diese Verfahren lernen aus einer Menge von Beispielen mit bestimmten Klassenzugehörigkeiten eine Funktion f, um zukünftig unbekannten Beispielen eine Klasse zuzuordnen. Der Lernvorgang ist als Trainieren und die Zuordnung als Klassifikation bekannt. Im Fall der Plagiatanalyse sind zwei Klassen vorhanden – die der nicht-plagiierten Abschnitte und die der plagiierten Abschnitte. Ein Zwei-Klassen-Lernverfahren wäre eine Lösung für die Zuordnung von Abschnitten zu einer Klasse. Jedoch existiert keine repräsentative Menge von plagiierten Abschnitten. Plagiatvergehen werden zudem kaum veröffentlicht und es ist schwierig „typische“ Plagiate zu sammeln. Die zweite Klasse, die der plagiierten Abschnitte, fehlt demnach. Aus diesem Grund wird ein Lernverfahren zur Eine-Klassen-Klassifikation, insbesondere Funktionen dieser, betrachtet.

Die intrinsische Plagiaterkennung ist ein solches Eine-Klasse-Klassifizierungsproblem. Charakteristisch ist, dass ausschließlich Informationen über eine Klasse gegeben sind – Informationen über die sogenannte Zielklasse. Dagegen gehören alle anderen Informationen einer Ausreißerklasse an. Im Fall der intrinsischen Plagiaterkennung formen die Dokumentbestandteile des vermutlichen Autors, die nicht-plagiierten Abschnitte, die Zielklasse. Informationen über Dokumentbestandteile beliebiger Autoren, die plagiierten Abschnitte, sind demgegenüber der Ausreißerklasse zugeordnet. Die Elemente der Zielklasse sind zahlreich vorhanden, während die der Ausreißerklasse unbekannt sind und fehlen können.

Die Aufgabe eines Lernverfahrens zur Eine-Klassen-Klassifikation besteht im Kontext der intrinsischen Plagiaterkennung darin, alle Abschnitte zu identifizieren, die zu der Ausreißerklasse, der „Plagiatklasse“, gehören. Verfahren, die diese Aufgabe leisten, basieren

ausschließlich auf Elementen aus der Zielklasse. Anhand dieses Wissens muss ein solches Verfahren in der Lage sein, Ausreißer- von Zielklassenelementen zu trennen. Die Funktion, die durch das Verfahren bestimmt wird, ist bei der Ausreißererkennung[1] einzusetzen. Die Ausreißererkennung beantwortet dabei die Frage, ob es sich bei einem untersuchten Wert des Abschnittes um eine Plagiatstelle, also um einen Ausreißer, handelt.

Der Prozess der Ausreißererkennung verläuft dabei in zwei Schritten: einer Vorbereitung und einer Auswertung. Diese sind an den Vorgang des Trainierens und Klassifizierens angelehnt. Bei der Vorbereitung werden ausschließlich Elemente der Zielklasse benötigt, während bei der Auswertung sowohl Ziel- als auch Ausreißerklassenelemente vorliegen müssen. Beide Schritte nehmen die Klassenelemente in deren Stilmodellrepräsentation, insbesondere dem Stiländerungsvektor (siehe Formel 4.1) entgegen. Ausreißerklassenelemente werden im untersuchten Dokument künstlich konstruiert, indem ursprüngliche Abschnitte modifiziert oder fremde Abschnitte eingefügt werden. Jeder Abschnitt erhält zusätzlich eine Information über seine Klassenzugehörigkeit.

Das Aktivitätsdiagramm in Abbildung 5.1 stellt das im Folgenden eingesetzte Verfahren dar. Ziel ist es, Funktionen zur Ausreißererkennung für dieses Verfahren zu bestimmen.

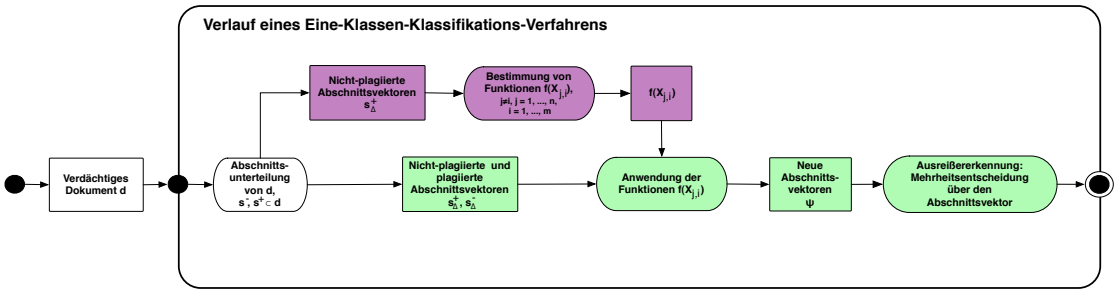


Abbildung 5.1 *: Plagiaterkennung durch Eine-Klassen-Klassifikation als Zwei-Schritt-Prozess. Die lila eingefärbte Komponenten stellen die Vorbereitung, die grünen die Auswertung dar.*

Im Vorbereitungsschritt werden Funktionen aus den Zielklassenelementen bestimmt. Dies erfordert eine Untersuchung der Komponenten der Abschnittsvektoren. Die Vektorkomponente für bestimmte Stiländerungsmerkmale $x_{j,i}$ sind von Abschnitt zu Abschnitt verschieden. Das bedeutet, dass die Werte $x_{j+1,i}, x_{j+2,i}, \ldots, x_{n,i}$ Realisierungen einer Stichprobenvariablen $X_{j,i}$ beschreiben. In Anlehnung an die Dokumentenmatrix stellen die Spaltenvektoren die Realisierungen der konkreten Stichprobenvariablen dar. Im Vorbereitungsschritt wird für jede Stichprobenvariable $X_{j,i}$ eine Funktion $f\left(X_{j,i}\right)$ berechnet. Im Auswertungsschritt wird anschließend für jeden Abschnittswert $x_{j,i}$ durch die

[1] engl. Outlier Detection

entsprechende Funktion einen Funktionswert $\psi_{j,i}$ bestimmt. Somit wird für jeden j-ten Abschnitt $\mathbf{s_\Delta}$ ein neuer Vektor Ψ_j erzeugt. Anhand des Abschnittsvektors Ψ_j erfolgt die Ausreißererkennung. Eine Mehrheitsentscheidung bezüglich der Anzahl der gefundenen Ausreißer im Abschnitt bestimmt schließlich, ob der betrachtete Abschnitt plagiiert ist oder nicht. Die Bestimmung der Ausreißer erfolgt durch einen Vergleich jeder Vektorkomponente mit einem Schwellwert. Dieser liegt grundsätzlich bei 0.5. Ist eine Vektorkomponente größer als dieser Schwellwert, so wird angenommen, dass ein Ausreißer in dem betrachteten Abschnitt vorliegt. Abhängig von der Anzahl der festgestellten Ausreißer wird der Abschnitt als ein Ausreißerklassenelement angesehen, siehe hierzu Abbildung 5.2.

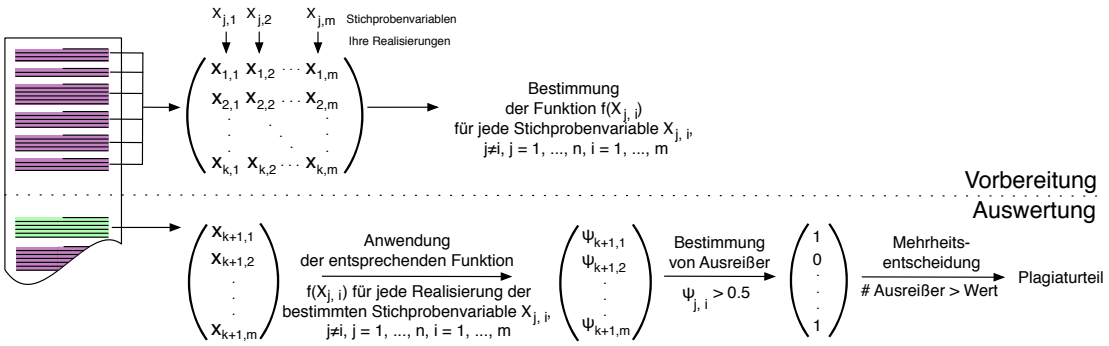


Abbildung 5.2 *: Zuerst wird für jede Stichprobenvariable der Zielklasse eine Funktion berechnet. Anschließend wird diese für die Auswertung der Abschnitte des Dokumentes benutzt. Lila markierte Abschnitte gehören der Zielklasse an, grüne dagegen sind Ausreißerklassenelemente.*

Die in diesem Kapitel vorzustellenden Funktionen stützen sich auf folgende Überlegungen. Das Dokument besitzt m Stichprobenvariablen mit einer bestimmten Wahrscheinlichkeitsverteilung. Es wird angenommen, dass die Stichprobenvariablen der Zielklasse, also die Grundgesamtheiten jeder Stichprobenvariablen, normalverteilt sind. Ausreißerklassenelemente werden als gleichverteilt betrachtet. Bei einer normalverteilten Stichprobenvariablen treten Werte im Zentrum der Verteilung relativ häufig auf, d.h. die Wahrscheinlichkeit solcher Werte ist sehr hoch, während die Wahrscheinlichkeit für Werte mit großer Abweichung vom Erwartungswert gering ist. Umgekehrt verhält sich die Betrachtung der Wahrscheinlichkeit für Ausreißer bei der entsprechenden Stichprobenvariablen: ein Wert nahe dem Erwartungswert ist unmöglich ein Ausreißer, während ein Wert mit großer Abweichung vom Erwartungswert mit hoher Wahrscheinlichkeit einen Ausreißer darstellt.

Zunächst wird in diesem Kapitel wird die Normalverteilungsannahme durch einen Gewichtsfaktor beurteilt. Im Anschluss daran werden lineare Funktionen eingeführt. Weitere lineare Funktionen, die Regressionsgeraden, werden im dritten Abschnitt präsentiert. Die Dichtefunktion der Normalverteilung wird für eine weitere Funktion im vierten Abschnitt angewandt, bevor abschließend Ausreißer bezüglich des Interquartilbereiches der Dichtefunktion der Normalverteilung betrachtet werden.

5.1 Quantifizierung der Normalverteilungsannahme

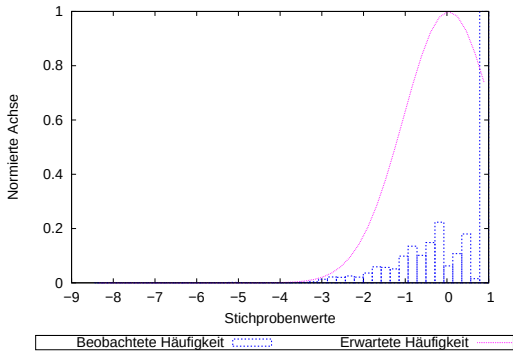
Zur Überprüfung der Normalverteilungsannahme der Stichprobenvariablen werden Tests auf Normalverteilung durchgeführt. Hierzu werden der χ^2- und der Kolmogorov-Smirnov-Test herangezogen. Im Fall, dass ein Test die Hypothese der Normalverteilung einer Stichprobenvariable ablehnt, wird die Irrtumswahrscheinlichkeit α als Quantifizierung verwendet. Wird dagegen die Hypothese angenommen, dient die statistische Sicherheit $(1 - \alpha)$ als Gewichtung. Diese ist mit der Frage nach der Genauigkeit der Untersuchung in dem Sinne verbunden, dass bei statistischen Aussagen immer ein Irrtum möglich ist. Die getroffene Quantifizierung verstärkt die Funktionswerte der Funktionen zur Ausreißererkennung. Die Quantifizierung, auch Faktor genannt, gleicht die Verletzung der Normalverteilungsannahme aus, die bei einer hohen Varianz der Stichprobenvariable auftritt. Das bedeutet, dass die Abweichung der Werte vom Erwartungswert hoch ist, wie Abbildung 5.4(a) zeigt. Dadurch „passt“ die Normalverteilung nicht zu der tatsächlichen Verteilung. Bei einer niedrigen Varianz dagegen ist die Normalverteilungsannahme erfüllt, siehe Abbildung 5.4(b).

Bei der Durchführung der Tests für alle Stichprobenvariablen wurde die Hypothese der Normalverteilung abgelehnt. Aufgrund der Tatsache, dass jede Stichprobenverteilung ein unterschiedliches „Annäherungsprofil“ an die Normalverteilung aufweist, kommt das Maß des relativen Fehlers für die Berechnung eines alternativen Faktors in Frage. Hierbei handelt es sich um eine dimensionslose Größe, die den den Quotient aus dem absoluten Fehler und dem exakten Zahlenwert bildet. Der relative Fehler wird demnach wie folgt berechnet:

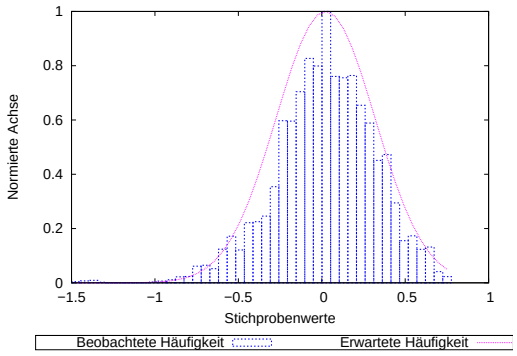
$$\text{Relativer Fehler} = \frac{\text{gemessener Wert} - \text{exakter Wert}}{\text{exakter Wert}}$$

Daraus folgt, dass der relative Fehler r für die Stichprobenvariablen und der Normalverteilung folgende Form annimmt:

$$r = \frac{\text{absolute Häufigkeit} - \text{theoretisch Häufigkeit}}{\text{theoretische Häufigkeit}} = \frac{h_m - np_m}{np_m}$$



(a) Hohe Varianz der Stichprobenvariable „Anzahl der Abkürzungen“.



(b) Niedrige Varianz der Stichprobenvariable „Durchschnittliche Satzlänge“.

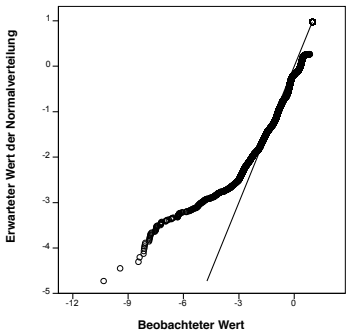
Abbildung 5.3 *: Varianzen für zwei Stichprobenvariablen. Die beobachtete Häufigkeit stellt die absoluten Häufigkeiten der Werte in den Klassen dar. Die erwartete Häufigkeit berechnet sich aus den Wahrscheinlichkeiten der Werte für jede Klasse durch die Verteilungsfunktion* Φ *und den Stichprobenumfang.*

Die Notation ist gemäß der Berechnung der Prüfgröße bei dem χ^2- Test, siehe Formel 2.17. Für den gesamten Stichprobenumfang der k-Klassen ergibt sich somit für den relativen Fehler:

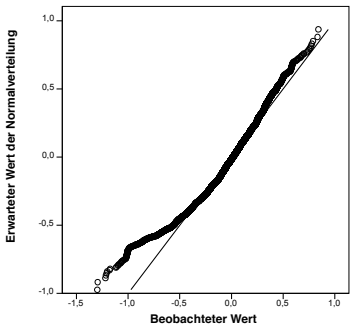
$$r = \frac{1}{k} \sum_{i=m}^{k} min\left(1, \frac{h_m - np_m}{np_m}\right), \quad r \in [0,1]$$

Damit ist der Wert der maximalen relativen Abweichung auf 1 beschränkt. Es wird berücksichtigt, dass die theoretische Häufigkeit in bestimmten Intervallen auch keinen Wert annimmt. Dem Faktor ω, der beurteilt, ob die Stichprobenvariable normalverteilt ist, wird der Wert $(1 - r)$ zugesprochen.

Eine weitere Möglichkeit zur Bestimmung eines Faktors besteht darin, die Quantile der empirischen Häufigkeit mit den entsprechenden Quantilen der Normalverteilung zu vergleichen. Graphisch entsteht auf diese Weise ein sogenannter Normal-Quantil-Plot. Hierbei werden auf der x-Achse die Quantile der beobachteten Häufigkeit aufgetragen, auf der y-Achse die der Normalverteilung. Falls beide Verteilungstypen übereinstimmen, liegen die Punkte auf der Winkelhalbierenden, siehe Abbildung 5.4.



(a) Normal-Quantil-Plot der Stichprobenvariable „Anzahl der Abkürzungen“.



(b) Normal-Quantil-Plot der Stichprobenvariable „Durchschnittliche Satzlänge“.

Abbildung 5.4 *: Normal-Quantil-Plots für zwei Stichprobenvariablen. Die Plots wurden durch die statistische Software SPSS erstellt.*

5.2 Lineare Funktionen

Eine lineare Funktion ist eine Gerade mit der Funktionsgleichung:

$$f(x) = ax + b,$$

wobei a die Steigung und b den y-Achsenabschnitt bezeichnen. Die Steigung a einer Geraden durch zwei gegebenen Punkte x_1 und x_2 lautet:

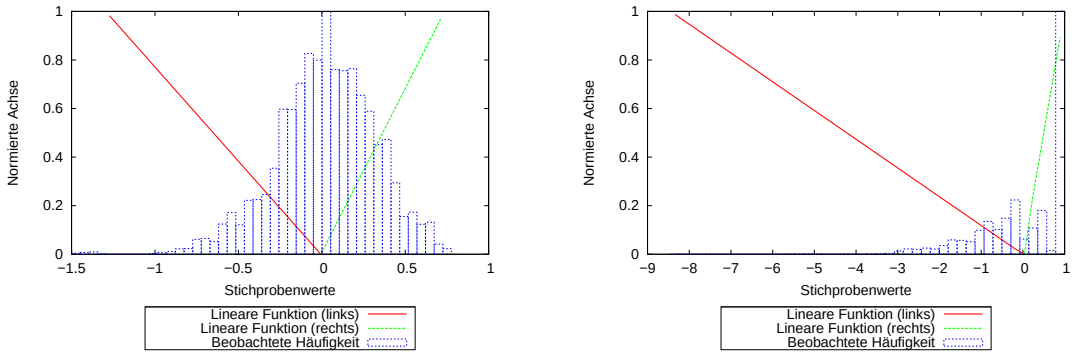
$$a = \frac{f(x_2) - f(x_1)}{x_2 - x_1} = \frac{\Delta y}{\Delta x},\ x_2 \neq x_1 \tag{5.1}$$

Der Achsenabschnitt b einer gegebenen Gerade, die durch die Punkte $(x_1, f(x_1))$ und $(x_2, f(x_2))$ geht, ist definiert durch:

$$b = \frac{x_2 f(x_1) - x_1 f(x_1)}{x_2 - x_1},\ x_2 \neq x_1 \tag{5.2}$$

In unserem Fall werden zwei Geraden auf Basis jeder Stichprobenvariablen $X_{j,i}$ der Zielklasse konstruiert. Der Wertebereich der Stichprobenvariablen wird diskretisiert und die Werte in Klassen eingeteilt. Als Konsequenz der Normalverteilungsannahme folgt, dass je weiter ein Wert vom Erwartungswert $E(X_{j,i}) = \mu$ der Stichprobenvariablen $X_{j,i}$ entfernt liegt, desto höher ist die Wahrscheinlichkeit, dass dieser ein Indiz für einen Ausreißer ist. Daher verläuft eine Gerade durch die Punkte $(x_{min}, 1)$ und $(E(X_{j,i}), 0)$, die andere durch $(E(X_{j,i}), 0)$ und $(x_{max}, 1)$. Dabei stellen x_{min} und x_{max} den kleinsten bzw. größten Wert der Stichprobenvariable dar. Abbildung 5.5 zeigt den Verlauf der

linearen Funktionen für zwei Stichprobenvariablen aus der Zielklasse in Verbindung mit den dazugehörigen absoluten Häufigkeiten.



(a) Stichprobenvariable „Durchschnittliche Satzlänge“.

(b) Stichprobenvariable „Anzahl der Abkürzungen“.

Abbildung 5.5 *: Lineare Funktionen und absolute Häufigkeiten für zwei Stichprobenvariablen. Die Geraden wurden für die Klassenmittelpunkte berechnet.*

Mittels der Gleichungen 5.1 und 5.2 lassen sich die Steigungen a_1 und a_2 und Achsenabschnitte b_1 und b_2 dieser beiden Geraden berechnen. Daraus ergeben sich die Werte 5.3 für die erste Gerade und 5.4 für die zweite.

$$a_1 = \frac{-1}{E\left(X_{j,i}\right) - x_{min}}, \quad b_1 = \frac{E\left(X_{j,i}\right)}{E\left(X_{j,i}\right) - x_{min}} \tag{5.3}$$

$$a_2 = \frac{1}{x_{max} - E\left(X_{j,i}\right)}, \quad b_2 = \frac{-E\left(X_{j,i}\right)}{x_{max} - E\left(X_{j,i}\right)} \tag{5.4}$$

Im Vorbereitungsschritt werden die Steigungen und y-Achsenabschnitte beider Geraden für jede Stichprobenvariable der Zielklasse berechnet und gespeichert. Diese Parameter werden bei dem Auswertungsschritt in der linearen Funktionsgleichung eingesetzt, um Ausreißer zu erkennen. Wenn ein untersuchter Abschnittswert größer ist als der Erwartungswert, der aus der Stichprobenvariable dieses Wertes berechnet wurde, werden die Parameter der ersten Gerade für die Bestimmung des Funktionswertes verwendet, sonst die der zweiten.

5.3 Regressionsfunktionen

Eine Funktion, die eine gegebene Menge von Punkten $(x_1, y_1), \ldots, (x_n, y_n)$ bezüglich des quadratischen Fehlers so gut wie möglich annähert, heißt Regressionsfunktion. Im Fol-

genden werden lineare Regressionsfunktionen, d.h. Geraden mit der Funktionsgleichung

$$f(x) = ax + b,$$

mit a als Steigung und b als y-Achsenabschnitt, betrachtet.

Die Ermittlung der Steigung a und des y-Achsenabschnittes b wird aus den konkreten Wertepaaren (x_i, y_i) mit $i = 1, 2, \ldots, n$ so vorgenommen, dass sich die Regressionsgerade dieser Punktwolke möglichst gut anpasst. Dazu wird die Methode der kleinsten Quadrate verwendet, bei der a und b aus folgender Forderung bestimmt werden:

$$\sum_{i=1}^{n} [y_i - f(x_i)]^2 = \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \longrightarrow Minimum \tag{5.5}$$

a und b sind so zu bestimmen, dass die Quadratsumme der Ordinatendifferenzen zwischen den Wertepaaren und den entsprechenden Punkten auf der gesuchten Gerade minimiert wird. Die Funktion 5.5 wird nach den beiden unabhängingen Variablen a und b partiell abgeleitet. Die jeweiligen Ergebnisse werden auf 0 gesetzt. Es entsteht das Gleichungssystem:

$$\frac{\partial \left[\sum_{i=1}^{n} (y_i - ax_i - b)^2\right]}{\partial b} = -2 \sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

$$\frac{\partial \left[\sum_{i=1}^{n} (y_i - ax_i - b)^2\right]}{\partial a} = -2 \sum_{i=1}^{n} (y_i - ax_i - b)\, x_i = 0$$

Durch Umformung ergeben sich folgende Gleichungen:

$$nb + a \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$b \sum_{i=1}^{n} x_i + a \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i$$

Mit

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

erhält man die Lösungen $\hat{a}$ und $\hat{b}$. Diese sind Schätzer für a und b.

$$\hat{a} = a = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{5.6}$$

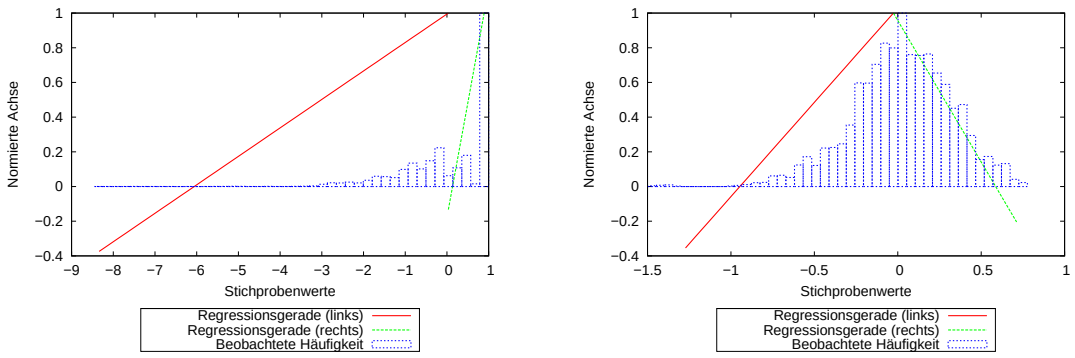
$$\hat{b} = b = \bar{y} - \hat{a}\bar{x} \tag{5.7}$$

Die Geradengleichung lautet somit:

$$\hat{y} = \hat{a}x + \hat{b} = \bar{y} + \hat{a}\,(x - \bar{x}) \tag{5.8}$$

Die Formeln sind aus Beyer u. a. (1991) entnommen.

Für die Benutzung dieser Funktion wird eine Klassenunterteilung der Stichprobenvariable $X_{j,i}$ vorausgesetzt. Seien $x_1, \ldots, x_k$ die Mittelpunkte der k Klassen und $y_1, \ldots, y_k$ die zugehörigen absoluten Häufigkeiten. Dies ermöglicht die Betrachtung von x_i und y_i als ein Wertepaar. Der Mittelpunkt der Klasse e, in der der Erwartungswert $E\,(X_{j,i})$ fällt, wird als x_e bezeichnet, $x_e \in [x_1, \ldots, x_k]$. Wie bei der linearen Funktion des vorhergehenden Abschnittes werden auch hier zwei Geraden beschrieben – die eine verläuft durch die Punkte $(x_1, y_1), \ldots, (x_e, y_e)$, die andere durch $(x_e, y_e), \ldots, (x_k, y_k)$. Die erste Gerade nähert die linke Seite des entstandenen Histogramms an, die zweite die rechte. Beide treffen sich schließlich am Punkt (x_e, y_e) und fallen jeweils bis (x_1, y_1) bei der linken Gerade bzw. bis (x_k, y_k) bei der rechten ab. Die Geraden passen sich auf diese Weise den beobachteten Häufigkeiten der Stichprobvariablen an. Abbildung 5.6 verdeutlicht dies für die zwei Stichprobenvariablen aus der Zielklasse.



(a) Stichprobenvariable „Anzahl der Abkürzungen“.

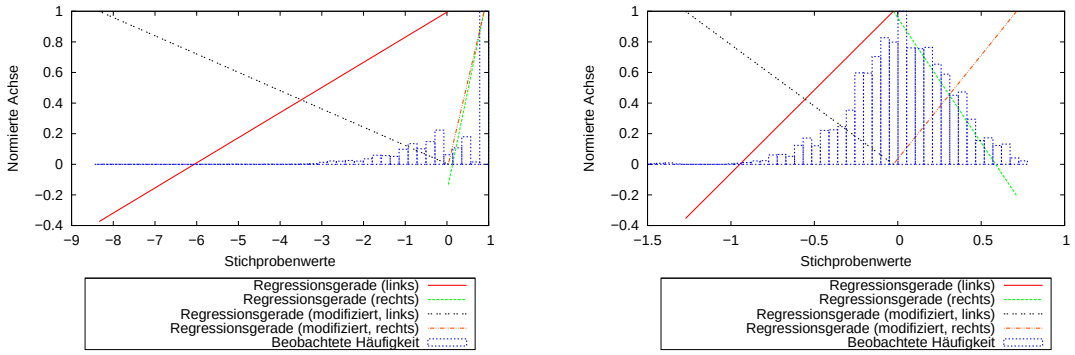
(b) Stichprobenvariable „Durchschnittliche Satzlänge“.

Abbildung 5.6 *: Anpassung der Regressionsgeraden an den beobachteten Häufigkeiten von zwei Stichprobenvariablen.*

Der Verlauf der Regressionsgeraden in Abbildung 5.6(a) lässt sich durch die Verletzung der getroffenen Annahme der Normalverteilung erklären. Die meisten Werte dieser Stichprobe liegen aufgrund der hohen Varianz nicht symmetrisch um den Erwartungswert. Hinzu kommt, dass durch die iterative Berechnung der Parameter der Regressionsgera-

den, siehe den Quotienten der Summen in Formel 5.6, jedes betrachtete Wertepaar den Verlauf der Geraden verändert. Solche Fälle sind nicht auszuschließen.

Das Ziel der Funktionen der Regressionsgeraden ist es jedoch, dass das Minimum bei (x_e, y_e) und das Maximum bei (x_1, y_1) bzw. (x_k, y_k) angenommen wird. Ein Wert nahe x_e sagt aus, dass kein Ausreißer vorliegt. Daher müssen beide Geraden in y-Richtung gespiegelt werden. Dies zieht die Veränderung des Vorzeichens der Steigungen beider Geraden nach sich. Anschließend werden die Geraden verschoben, so dass beide durch den Punkt $(x_e, 0)$ gehen. In diesem Fall wird ein neuer y-Achsenabschnitt berechnet. Diese Transformationen werden für zwei Stichprobenvariablen in Abbildung 5.7 dargestellt.



(a) Stichprobenvariable „Anzahl der Abkürzungen“.

(b) Stichprobenvariable „Durchschnittliche Satzlänge“.

Abbildung 5.7 *: Ursprüngliche und modifizierte Regressionsgeraden und absolute Häufigkeiten für zwei Stichprobenvariablen.*

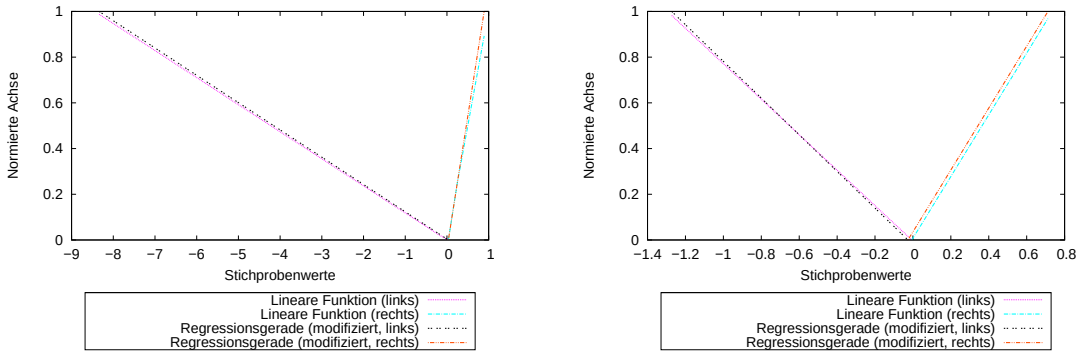
Die Steigungen $\hat{a}_1$ und $\hat{a}_2$ und Achsenabschnitte $\hat{b}_1$ und $\hat{b}_2$ der modifizierten Geraden lassen sich wie folgt für die jeweilige Gerade berechnen:

$$\hat{a}_1 = -\frac{\sum_{i=1}^{e} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{e} (x_i - \bar{x})^2}, \quad \hat{b}_1 = \hat{a}_1 x_e$$

$$\hat{a}_2 = -\frac{\sum_{j=e}^{k} (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=e}^{k} (x_j - \bar{x})^2}, \quad \hat{b}_2 = \hat{a}_2 x_e$$

Da es sich bei den Regressionsgeraden ebenfalls um lineare Funktionen handelt, fallen sie in den meisten Fällen zusammen, siehe dazu Abbildung 5.8.

Wie bei dem linearen Funktionen aus dem vorhergehenden Abschnitt werden auch hier der Anstieg und der Achsenabschnitt der beiden modifizierten Funktionen anhand der Zielklassen-Stichprobenvariablen im Vorbereitungsschritt berechnet und gespeichert, um dann bei der Auswertung diese für die entsprechenden Abschnittswerte zu benutzen. Ist



(a) Stichprobenvariable „Anzahl der Abkürzungen“.

(b) Stichprobenvariable „Durchschnittliche Satzlänge“.

Abbildung 5.8 *: Vergleich zwischen den Geraden der linearen Funktion und den Regressionsgeraden für zwei Stichprobenvariablen.*

ein untersuchter Abschnittswert größer als x_e, der aus der Stichprobenvariablen dieses Wertes berechnet wurde, werden die gespeicherten Werte der ersten Geraden verwendet, andernfalls die der zweiten.

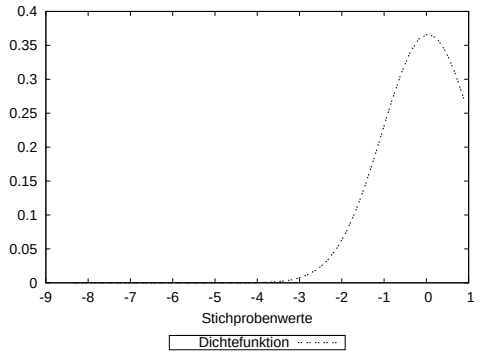
5.4 Relative Abweichungen durch die Dichtefunktion der Normalverteilung

Im Folgenden wird die Dichtefunktion der Normalverteilung $\varphi\left(x; \mu, \sigma^2\right)$ für die Bestimmung einer weiteren Funktion zur Ausreißererkennung betrachtet. Die Kenngrößen der Dichtefunktion für eine Stichprobenvariable $X_{j,i}$ sind der Erwartungswert $E\left(X_{j,i}\right)$ und die Varianz $V\left(X_{j,i}\right)$. Beide Kenngrößen werden durch die Werte der Stichprobenvariablen geschätzt. Eine Schätzfunktion für den Erwartungswert ist das arithmetische Mittel. Für die Varianz bietet die empirische Varianz einen Berechnungsweg. Abbildung 5.9 zeigt die Dichtefunktion für zwei Stichprobenvariablen.

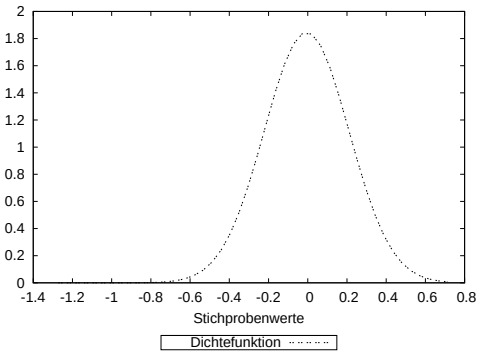
Im Vorbereitungsschritt werden die Kenngrößen der Dichtefunktion für jede Stichprobenvariable berechnet und gespeichert. Bei der Auswertung eines Abschnittes wird für jeden Abschnittswert $x_{j,i}$ folgende Funktion verwendet:

$$f\left(x_{j,i}\right) = \frac{\varphi\left(E\left(X_{j,i}\right); E\left(X_{j,i}\right), V\left(X_{j,i}\right)\right) - \varphi\left(x_{j,i}; E\left(X_{j,i}\right), V\left(X_{j,i}\right)\right)}{\varphi\left(E\left(X_{j,i}\right); E\left(X_{j,i}\right), V\left(X_{j,i}\right)\right)}$$

Diese Funktion berechnet die relative Abweichung eines Wertes vom Erwartungswert durch die Dichtefunktion. Diese Abweichung ist für Werte in der Nähe des Erwartungs-



(a) Stichprobenvariable „Anzahl der Abkürzungen“.



(b) Stichprobenvariable „Durchschnittliche Satzlänge“.

Abbildung 5.9 *: Dichtefunktionen von zwei Stichprobenvariablen.*

wertes niedrig, während sie mit steigendem Abstand wächst. Diese Tatsache wird wiederum für die Ausreißererkennung eingesetzt.

5.5 Ausreißer außerhalb des verstärkten Interquartilbereiches der Dichtefunktion der Normalverteilung

Die Dichtefunktion der Normalverteilung einer Stichprobenvariablen wird im Folgenden als Funktion für die Ausreißererkennung verwendet. Wie bei der vorhergehenden Funktion werden auch hier im Vorbereitungsschritt die Kenngrößen der Dichtefunktion für jede Stichprobenvariable berechnet und gespeichert. Bei der Auswertung werden allerdings diejenigen Funktionswerte der Dichtefunktion als Ausreißer betrachtet, die außerhalb des linksseitig und rechtsseitig verstärkten Interquartilbereiches der Dichtefunktion der entsprechenden Stichprobenvariable liegen. Der Interquartilbereich IQR wird aus der Differenz des 3. und 1. Quartils berechnet:

$$\mathrm{IQR} = Q_3 - Q_1$$

Dieser umfasst die Hälfte aller Werte. Die Berechnung des 1. und 3. Quartils erfolgt nach der Methode von Mendenhall und Sincich (Weisstein (2007)):

$$Q_1 = x_{(l)}, \quad l = \left\lfloor \frac{n+1}{4} \right\rfloor$$

$$Q_3 = x_{(u)}, \quad u = \left\lfloor \frac{3n+3}{4} \right\rfloor$$

$x_{(l)}$ und $x_{(u)}$ sind dabei die Werte an den Positionen l und u der Rangliste der Stichprobenwerte. Ein Wert, der außerhalb des 1.5-fachen des unteren und oberen Quartils fällt, heißt Ausreißer. Dieser befindet sich in folgenden Bereichen:

$$Q_3 + 1.5 \cdot \text{IQR} \leq \text{Ausreißer} \leq Q_1 - 1.5 \cdot \text{IQR}$$

6 Evaluierung der Funktionen zur Eine-Klassen-Klassifikation

Im Folgenden werden die in Kapitel 5 vorgestellten Funktionen ausgewertet. Zur Bewertung der Funktionen werden im ersten Abschnitt Gütemaße eingeführt. Im zweiten Abschnitt wird das Klassifikationsverfahren Support-Vektor-Maschine vorgestellt. Anschließend wird der Experimentaufbau beschrieben, der der Analyse der Funktionen im letzten Abschnitt zugrunde liegt.

6.1 Gütemaße zur Bewertung der Funktionen

Precision und Recall sind Gütemaße zur Bewertung und zum Vergleich von Information Retrieval Systemen (Rijsbergen (1979)). An ein System wird eine Anfrage gestellt. Dieses gibt entsprechend der Anfrage eine Menge an relevanten und irrelevanten Antworten zurück. Precision beschreibt dabei die Genauigkeit der Antwort auf die Anfrage. Sie gibt den Anteil der relevanten Antworten in der Ergebnismenge an. Recall stellt dagegen die Vollständigkeit der Antwort auf die Anfrage dar, indem die gefundenen relevanten Antworten angegeben werde. Beide Gütemaße nehmen Werte aus dem Intervall $[0;1]$ an. Deren Berechnung erfolgt nach dem Schema in Abbildung 6.1.

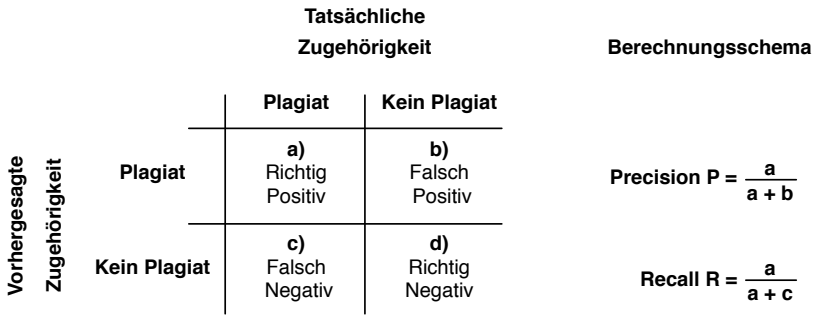


Abbildung 6.1 *: Berechnung von Precision und Recall nach Wikipedia (2008).*

Die Anfrage stellt im Fall der Plagiaterkennung die Suche nach plagiierten Stellen in einem verdächtigen Dokument dar. Die Relevanz beschreibt die Übereinstimmung der tatsächlich plagiierten Abschnitte und die vom Verfahren durch die Funktionen als plagiiert eingestuften Abschnitte. Precision bezeichnet den Anteil der plagiierten Abschnitte unter den gefundenen, Recall den Anteil der plagiierten Abschnitte, die gefunden wurden. Falls alle relevanten Abschnitte gefunden werden, beträgt der Recall 1. Die Precision ist 1, falls alle gefundenen Abschnitte relevant sind, d.h. alle Abschnitte der Ergebnismenge relevant sind. Umgekehrt gilt, falls die Ergebnismenge keine relevanten Abschnitte enthält, sind beide Gütemaße 0. Bei der Auswertung werden immer beide Werte betrachtet.

Das F-Maß kombiniert die beiden vorgestellten Gütemaße, indem dieses das harmonische Mittel beider bildet:

$$\text{F-Maß} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

6.2 Support-Vektor-Maschine (SVM)

Die Support-Vektor-Maschine ist ein Klassifikationsverfahren, das auf der statistischen Lerntheorie basiert. Anhand einer Trainingsmenge, die positiven und negativen Beispielen enthält, bzw. Ziel- und Ausreißerobjekte, wird eine trennende Hyperebene berechnet. Ziel ist es, den kürzesten Abstand von Punkten beider Seiten zur Ebene zu maximieren, um so eine optimale Hyperebene zu finden. Das Maximieren der Breite der Hyperebene ist ein Optimierungsproblem. Stützvektoren[1], die zur Trainingsmenge gehören und einen minimalen Abstand von der Hyperebene haben, bestimmen die Lage der Hyperebene. Nicht linear trennbare Daten werden in einem höherdimensionalen Raum durch eine Kernelfunktion überführt, so dass sie linear trennbar werden. Dies wird als „kernel trick“ bezeichnet. Im Testfall neuer Objekte wird die Zugehörigkeit der Elemente zu einer bestimmten Klasse durch die Hyperebene bestimmt.

Im Papier von Schölkopf u. a. (1999) wird das das Eine-Klassen-Klassifizierungsproblem durch den Support-Vektor-Algorithmus betrachtet. Die Funktionsweise einer Eine-Klassen-SVM wird dabei vorgestellt. Nach der Transformation im Merkmalsraum wird versucht, die Daten mit größtem Abstand vom Ursprung zu trennen. Anschließend wird die Standard-SVM-Technik, die Zwei-Klassen-SVM, eingesetzt. Für neue Punkte wird ihre Zuordnung durch ihre Lage bezüglich der Hyperebene bestimmt. Die Zielsetzung ist wie folgt formuliert. Gegeben sei eine Datenmenge mit einer bestimmten Wahrscheinlichkeitsverteilung P. Gesucht ist eine Teilmenge S von P, so dass die Wahrscheinlichkeit, dass ein Testpunkt

[1] engl. support vectors

aus P, der außerhalb von S liegt, durch einen zuvor spezifizierten Wert, der zwischen 0 und 1 liegt, begrenzt ist. Es wird eine binäre Funktion f konstruiert, die positive Werte innerhalb und negative Werte außerhalb von S annimmt, also im Komplement $\bar{S}$. f ist folglich positiv im Bereich, der die meisten Datenvektoren erfasst, und negativ sonst. f wird durch eine Kernel-Erweiterung im Sinne der Stützvektoren bestimmt, wobei die Trenngrenze maximiert oder äquivalent dazu die Länge der Gewichtsvektoren kontrolliert wird.

Manevitz u. Yousef (2002) stellen eine Ausreißer-SVM vor, die nicht nur den Ursprung als zugehörig zur Ausreißerklasse annimmt, wie dies bei der Eine-Klassen-SVM ist, sondern auch alle Datenpunkte die nahe genug am Ursprung liegen. Falls ein Dokumentenvektor wenige Komponenten ungleich Null besitzt, so ist dies ein Anzeichen dafür, dass das Dokument wenige Elemente mit der gewählten Teilmenge teilt. Dieser Vektor gilt als nicht repräsentativ für die Klasse und wird schließlich als Ausreißer erkannt. Die Anzahl der Merkmale mit Werten verschieden von Null und kleiner als eine vorbestimmte Grenze sind, bestimmen den Vektor als negatives Beispiel.

Vorteile von SVMs sind der Schutz vor Überanpassung an den Trainingsdaten und die Verarbeitung großer Datenmengen. Nachteilig ist die empirische Suche nach Parametern für die Kernelfunktion und der unklare Variableneinfluss (Eitrich (2003)).

6.3 Experimente

Die intrinsische Plagiaterkennung mit Ausreißern basiert auf der Untersuchung von einzelnen Abschnitten aus Dokumenten. Zur Verfügung stand die Dokumentensammlung nach Coriand (2008). Diese enthält 70 Dissertationen aus verschiedenen wissenschaftlichen Bereichen. 30 davon wurden dahingehend manipuliert, dass jeweils Plagiatstellen mit ähnlicher Thematik eingefügt wurden. Mit Hilfe der Kombinatorik entstanden aus diesen 30 Dokumenten durch Variieren der Anzahl der plagiierten Abschnitte 720 Testdokumente. Die 40 Originaldokumente wurden zu dieser Dokumentenmenge hinzugefügt, so dass sich eine Gesamtmenge von 760 Dokumenten ergab. Die Dokumente wurden in Abschnitte fester Größe unterteilt. Die Abschnittsgröße variiert von 500, 1000, 1500 und 2000 Wörtern. Die Abschnitte werden schließlich durch ihre Stilmodellrepräsentation dargestellt.

Die Abschnittsvektoren der Stilmodellrepräsentation werden für die im vorhergehenden Kapitel vorgestellten Funktionen als Eingabe verwendet. Ein Teil der Zielklassenelemente wird im Vorbereitungsschritt dazu benutzt, um die Funktionen zu spezifizieren. Im Auswertungsschritt werden anschließend auf die gleiche Anzahl von Ziel- uns Aus-

reißerklassenelemente diese Funktionen angewandt, um Ausreißer innerhalb der neuen Abschnittsvektoren, die aus den Funktionen berechnet wurden, zu finden. Die Anzahl der Ausreißer, deren Auftreten vorausgesetzt wird, um den Abschnitt als plagiiert zu bestimmen, wird bei der Analyse der Funktionen variiert. Diese liegen zwischen 1 und 6. Somit ist bestimmbar, für welche Anzahl an Ausreißern die Funktionen die beste Erkennungsleistung besitzen.

Zudem werden die durch die Funktionen transformierten Abschnittsvektoren durch Eine-Klassen-SVM klassifiziert. Diese wird mit Zielklassenelementen trainiert und mit gleicher Anzahl Ziel- und Ausreißerklassenelementen getestet. Die Implementation einer Eine-Klassen-SVM, der Libsvm (Chang u. Lin), wird dabei benutzt.

6.4 Analyse der Funktionen

Bei der intrinsischen Plagiaterkennung mit Ausreißern ist das Ziel eine hohe Precision zu erreichen. Ihre Werte sollten besser als 0.5 sein, da 0.5 schon mit einem zufälligen Ziehen erreicht werden kann. Der Recall spielt dabei eine nebensächliche Rolle. Es ist zu erwarten, dass, je mehr Ausreißer in einem Abschnitt erkannt werden, die Wahrscheinlichkeit zunimmt, den Abschnitt korrekt als plagiiert einzustufen.

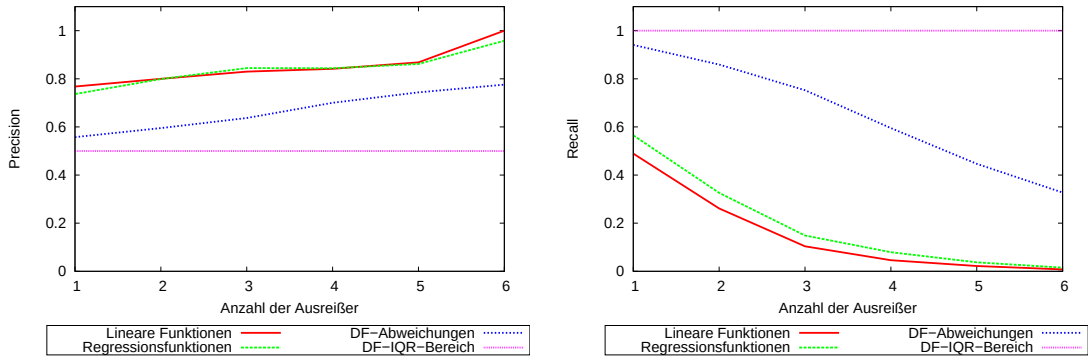
Die Dichtefunktion der Normalverteilung und die Erweiterung durch relative Abweichungen, siehe Abschnitt 5.4, wird durch die Bezeichnung DF-Abweichungen abgekürzt. Des Weiteren wird auch für die Ausreißer außerhalb des verstärkten Interquartilbereiches der Dichtefunktion der Normalverteilung, siehe Abschnitt 5.5, die Abkürzung DF-IQR-Bereich verwendet.

Aus den Abbildungen 6.2, 6.3, 6.4 und 6.5 ist zu schließen, dass für die Festlegung auf einen Ausreißer pro Abschnitt alle Funktionen bezüglich Precision und Recall die beste Retrieval-Eigenschaft liefern. Bereits ab zwei Ausreißern fällt der Recall stärker ab als die Precision steigt, so dass ab zwei Ausreißern weniger günstige Ergebnisse erzielt werden.

Die Anforderung einer hohen Precision erfüllen die Funktionen der linearen Funktionen und der Regressionsgeraden. Bei einem Ausreißer und einer Abschnittsgröße von 500, 1000 oder 1500 Wörtern sind die Funktionen DF-Abweichungen und DF-IQR-Bereich bezogen auf das F-Maß am besten: für DF-Abweichungen wird ein Wert von 0.708 bei 1500 Wörter erreicht, für DF-IQR-Bereich 0.75 bei 500 und 1000 Wörter. Bei 2000 Wörtern und einen Ausreißer sind die Funktionen DF-Abweichungen und die linearen Funktionen mit F-Maß von 0.756 bzw. 0.735 am besten.

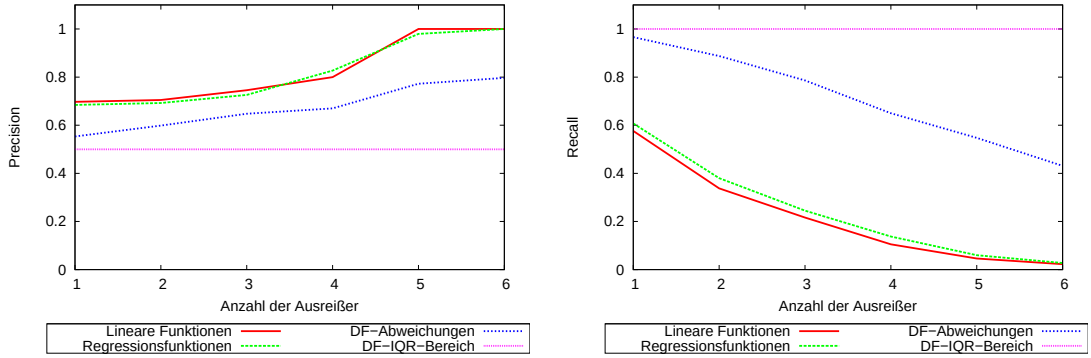
Bei der Eine-Klassen-SVM wird ein hoher Recall bei gleichzeitig geringerer Precision erreicht. Obwohl die Precision geringfügig niedriger als bei den Ausreißerfunktionen ist, ist der Recall deutlich höher. Dieser liegt für alle Abschnittsgrößen im Schnitt bei 0.911. Das F-Maß bei der Abschnittsgröße von 1500 Wörter für die DF-Abweichungen-Funktion liegt mit 0.759 deutlich höher als der Wert, der bei einem Ausreißer mittels derselben Funktion erzielt wird. Im Gegensatz dazu ist jedoch das F-Maß der DF-IQR-Bereich-Funktion bei 500 und 1000 Wörtern mit 0.69 und 0.68 unter der Ausreißererkennung mit 0.75. Bei Abschnittsgröße von 2000 Wörter ist das F-Maß für DF-Abweichungen 0.824, für die linearen Funktionen 0.85.

Zusammenfassend erzielt die Eine-Klassen-SVM bessere Ergebnisse auf den Daten für alle Funktionen außer die DF-IQR-Bereich-Funktion. Die Precision schwankt im Bereich zwischen 0.581 und 0.775. Die Funktionen haben wie erwartet eine hohe Precision bei steigender Anzahl der Ausreißer, der Recall ist dafür aber schlechter als bei der Eine-Klassen-SVM. Der Kompromiss zwischen Precision und Recall ist bei der Eine-Klassen-SVM besser.



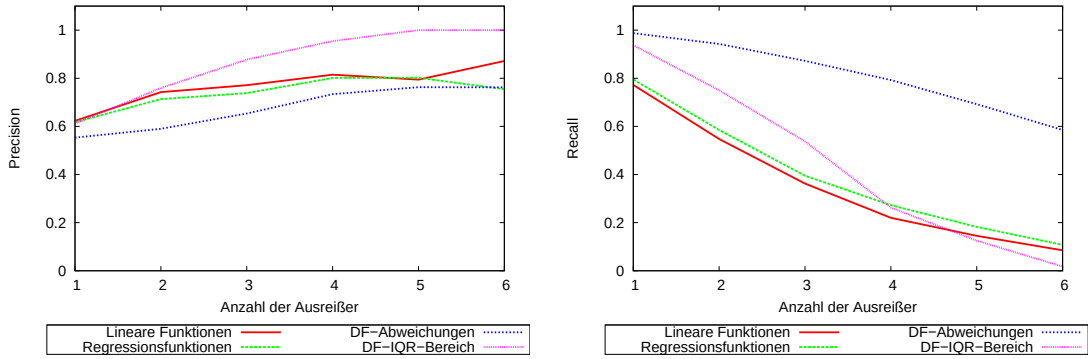
(a) Precision (b) Recall

Abbildung 6.2 *: Precision und Recall für 500 Wörter langen Textabschnitten.*



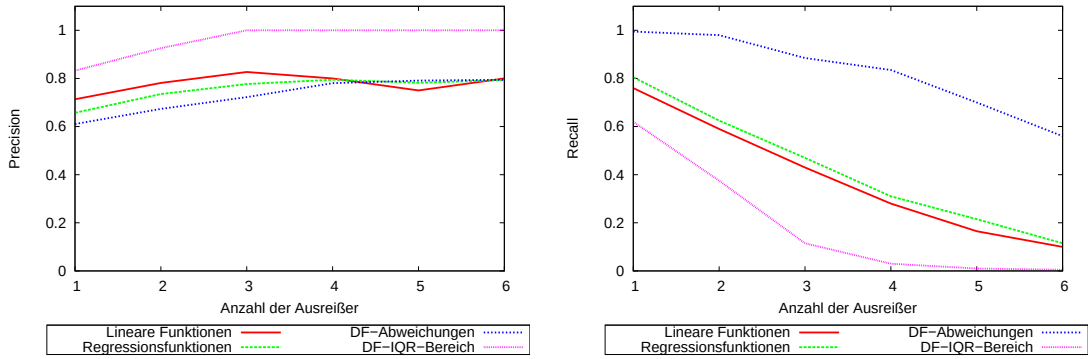
(a) Precision (b) Recall

Abbildung 6.3 *: Precision und Recall für 1000 Wörter langen Textabschnitten.*



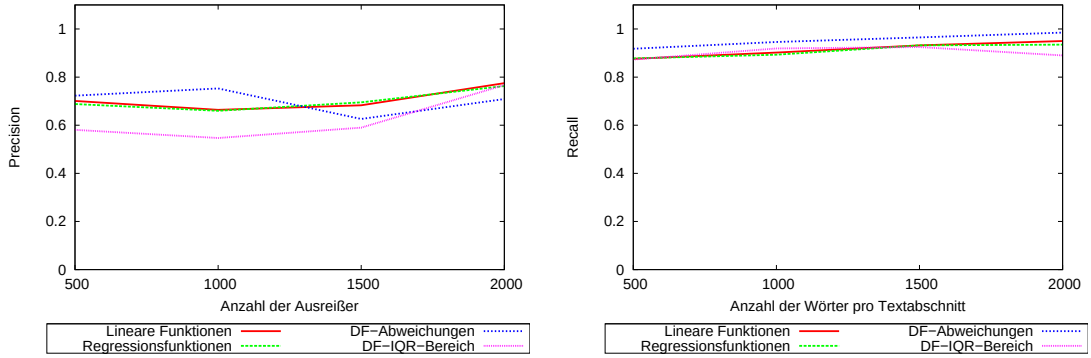
(a) Precision (b) Recall

Abbildung 6.4 *: Precision und Recall für 1500 Wörter langen Textabschnitten.*



(a) Precision (b) Recall

Abbildung 6.5 *: Precision und Recall für 2000 Wörter langen Textabschnitten.*



(a) Precision (b) Recall

Abbildung 6.6 *: Precision und Recall der Eine-Klassen-SVM.*

Abschnittsgröße	**500 W**		**1000 W**		**1500 W**		**2000 W**	
Funktion	**P**	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Lineare Funktionen	0.701	0.876	0.664	0.902	0.683	0.932	0.775	0.950
Regressionsfunktionen	0.688	0.878	0.66	0.893	0.695	0.932	0.763	0.935
DF-Abweichungen	0.723	0.918	0.753	0.946	0.626	0.965	0.709	0.985
DF-IQR-Bereich	0.581	0.873	0.547	0.919	0.590	0.925	0.767	0.89

Tabelle 6.1 *: Ergebnisse der Eine-Klassen-SVM der verschiedenen Funktionen für Textabschnitten mit 500, 1000, 1500 und 2000 Wörter.* ***W*** *steht für die Wörteranzahl eines Abschnittes,* ***P*** *für die Precision und* ***R*** *für den Recall.*

7 Zusammenfassung und Ausblick

Ziel dieser Arbeit war es, mit Hilfe von Ausreißern Plagiatstellen zu erkennen. Dazu wurde die intrinsische Plagiatanalyse durch Funktionen zur Ausreißererkennung erweitert. Die intrinsische Plagiatanalyse führt eine Stilmodellrepräsentation für ein untersuchtes Dokument ein. Diese quantifiziert den Schreibstil des Autors durch Stilmerkmale. Weicht die Stilmodellrepräsentation eines Abschnittes von der des Dokumentes ab, so ist dies ein Hinweis auf einen Stilwechsel. Ausreißer stellen diesen Wechsel dar. Um solche Ausreißer zu erkennen wurden Eine-Klassen-Klassifikations-Funktionen vorgestellt. Diese wurden durch eine gegebene Klasse, die der nicht-plagiierten Abschnitte, bestimmt und auf nicht-plagiiete und plagiierte Abschnitte angewandt. Dabei wurde die Stilmodellrepräsentation eines Abschnittes durch diese Funktionen in einem neuen Vektor überführt. Die Untersuchung der Vektorkomponenten stellt dabei Ausreißer fest. Es wurden lineare Funktionen, Regressionsfunktionen und die Dichtefunktion der Normalverteilung für zwei weitere Funktionen verwendet. Diese bestimmen relative Abweichungen der Werte der Dichtefunktion und untersuchen Ausreißer außerhalb des verstärkten Interquartilbereiches der Dichtefunktion. Bei der Auswertung ergab sich eine hohe Precision bei steigender Anzahl an erkannten Ausreißern pro Abschnitt. Für die Funktionen beträgt die Precision durchschnittlich 63.8%, wenn ausschließlich ein Ausreißer herangezogen wird. Für sechs Ausreißer erhöht sich die Precision um über 20% auf 86%. Im Gegenzug verringert sich jedoch der Recall von 80.1% auf 17.17%. Die Klassifikation durch Eine-Klassen-SVM ergibt eine durchschnittliche Precision von 68.2% bei einem Recall von 91.1%.

Zur Verbesserung der derzeitigen Ergebnisse wird vorgeschlagen, Methoden zur Kombination der Funktionen zur Ausreißererkennung anzuwenden. Regeln zur Kombination von Eine-Klassen-Klassifizierer werden in Tax u. Duin (2001) vorgestellt. Folgende Regeln sind dabei gegeben: Mittelwert-Votum, gewichtetes Mittelwert-Votum, Produkt-Votum, Mittelwert der geschätzten Wahrscheinlichkeiten und Produktkombination der Wahrscheinlichkeiten. Des Weiteren ist zu untersuchen, inwieweit Ensemble-Methoden für Eine-Klassen-Klassifizierer geeignet sind. Ensemble-Methoden betrachten Klassifizie-

rer als Gruppe, um die Nachteile der einzelnen Klassifizierer aufzuwiegen. Beispiele für solche Methoden sind Bagging, Boosting, Cascading, Random Forests und Stacking.

A Übersicht von Verfahren zur Eine-Klasse-Klassifikation

Dieses Kapitel stellt Methoden zur Lösung des Eine-Klassen-Klassifikationsproblem vor. Tabelle A.1 fasst die verschiedenen Verfahren und dazugehörige Referenzen zusammen. Detaillierte Angaben dazu sind in Tax (2001) zu finden.

Folgende Methoden zur Eine-Klasse-Klassifikation werden unterschieden:

1. Erkennungsmethoden für Ausreißer.

   Diese Methoden werden anhand ihrer Erkennungsstrategie unterteilt:

   - Methoden, die auf Klassifikation und Lernverfahren basieren.

     Diese erzeugen Ausreißer künstlich um die Zielklasse. Anschließend wird ein Klassifizierer trainiert, so dass er zwischen Ziel- und Ausreißerklasse unterscheiden kann. In Meyer zu Eißen u. Stein (2006) wird die intrinsischen Plagiaterkennung mittels Stilanalyse behandelt. Dabei wird ein Korpus mit Objekten der Ziel- und der Ausreißerklasse erzeugt. Auf dieser Grundlage werden anschließend eine Diskriminanzanalyse und die SVM angewendet.

   - Methoden für die Klassifizierungs- oder Regressionsprobleme.

     Anstatt die wahrscheinlichsten Gewichte für den Klassifizierer zu benutzen, die den Klassifizierungsfehler in einer gegebenen Trainingsmenge minimieren, wird die Wahrscheinlichkeit der Korrektheit der Gewichte für den Klassifizierer verwendet.

   - Dichtemethoden.

     Hierbei wird die Wahrscheinlichkeitsverteilung der Merkmale der Zielklasse geschätzt. Ausreißer werden als gleichverteilt angenommen. Durch Anwendung des Satzes von Bayes werden diese von der Zielklasse getrennt. Diese Vorgehensweise wird in Stein u. Meyer zu Eissen (2007) beschrieben.

2. Begrenzungsmethoden

Diese Methoden fokussieren auf die Grenze zwischen Ziel- und Ausreißerklasse. Die Berechnung der Grenze basiert auf der Entfernung der Objekte innerhalb der Zielklasse.

Ein Beispiel für eine Begrenzungsmethode ist die *k-centers*-Methode (Ypma u. Duin (1998)). Dieses Verfahren ist eine Variante des *k-means*-Clustering, wobei die Clusterzentren ausschließlich auf einem pro Trainingsobjekt beschränkt sind. Die Clusterzentren werden so auf den Trainingsobjekten platziert, dass die maximale Distanz aller minimalen Distanzen zwischen den Trainingsobjekten und den Clusterzentren minimiert wird. Die Methode ist sensitiv zu Ausreißern in der Trainingsmenge. Die Entfernung eines Testobjekts zu der Zielklasse wird berechnet, um seine Zugehörigkeit zu einer Klasse zu ermitteln.

Die Methode des *Nächsten-Nachbarn* (NN-d) wird aus einer lokalen Dichteschätzung durch den Nächsten-Nachbarn-Klassifiziers abgeleitet. Bei der NN-d Dichteschätzung wird eine Zelle, häufig eine Hypersphäre in d-Dimensionen, um das Testobjekt zentriert und seine lokale Dichte berechnet. Bei dem Eine-Klassen-Klassifizier-NN-d wird ein Testobjekt akzeptiert, wenn seine lokale Dichte größer oder gleich der lokalen Dichte seines ersten Nachbarn aus der Trainingsmenge ist.

3. Rekonstruktionsmethoden

   Diese modellieren den Erzeugungsprozess der Objekte. Wenn es möglich ist, ein Objekt im Modell zu dekodieren und seine Merkmale zu rekonstruieren, wird der Rekonstruktionsfehler dazu benutzt, die Anpassung des Objekts am Modell zu messen. Je kleiner dieser Fehler ist, desto besser passt das Objekt zum Modell und desto wahrscheinlicher handelt es sich um keinen Ausreißer.

   Als Rekonstruktionsmethoden werden das *k-means*-Clustering und die *Learning Vector Quantization* (LVQ) betrachtet. Bei beiden Methoden wird vorausgesetzt, dass die Daten geklumpt sind und durch Prototypobjekte oder Codebook-Vektoren charakterisiert werden. Meistens werden die Zielobjekte durch den nächsten Prototypvektor repräsentiert, der mittels der euklidischen Distanz gemessen wird. Bei dem k-means-Clustering wird der Fehler als die minimale Differenz zwischen dem Objekt und dem Prototypvektor minimiert. Die Entfernung eines Objektes zu der Zielklasse ist definiert als die quadratische Distanz des Objektes zum nächsten Prototypen.

   Der LVQ-Algorithmus ist eine überwachte Version des k-means-Clustering und wird hauptsächlich zur Klassifizierung verwendet. Jedes Trainingsobjekt besitzt

eine Clusterzugehörigkeit. LVQ wird auf einem neuronalen Netz trainiert, wobei der Klassifizierungsfehler minimiert wird.

Einordnung	**Verfahren**
Dichtemethoden	Gauß-Modell
	Bishop (1995)
	Mischung aus Gauß-Modellen
	Duda u. Hart (1991)
	Parzen Dichte
	Parzen (1962)
Begrenzungsmethoden	k-centers-Methode
	Ypma u. Duin (1998)
	Methode des Nächsten Nachbarn (NN-d)
	Duda u. Hart (1991)
	Support Vector Data Description (SVDD)
	Tax (2001)
Rekonstruktionsmethoden	k-means-Clustering
	Bishop (1995)
	Learning Vector Quantization (LVQ)
	Carpenter u. a. (1991)
	Self-organizing Maps (SOM)
	Kohonen (1997)
	Principal Component analysis (PCA)
	Bishop (1995)
	Mischung aus PCA
	Tipping u. Bishop (1999)
	Diabolo Netzwerke
	Hertz u. a. (1991)
	Auto-encoder Netzwerke
	Japkowicz u. a. (1995)

Tabelle A.1 Übersicht der Eine-Klassen-Klassifikations-Methoden nach Tax (2001).

B Übersicht von Verfahren zur Ausreißererkennung

Dieses Kapitel bietet eine Zusammenstellung verschiedener Verfahren auf dem Gebiet der Ausreißererkennung. Es werden Familien verwandter Verfahren und dazugehörige Beispiele kurz beschrieben. Tabelle B.1 fasst die verschiedenen Verfahren und korrespondierende Referenzen zusammen.

Verteilungsbasierte Ausreißererkennung Verteilungsbasierte Ansätze zur Erkennung von Ausreißern beziehen sich auf statistische Verteilungen wie die Gauß- oder Poisson-Verteilung. Die zugrunde liegende Verteilung wird modelliert. Die Definition der Ausreißer ist hierbei abhängig von den Eigenschaften der Verteilung.

Statistische Tests auf Ausreißern, sogenannte *Discordancy Tests*, wurden von Barnett u. Lewis (1994) für verschiedene Sachlagen entwickelt. Diese sind abhängig von der Verteilung, deren (un)bekannten Parametern, der Anzahl der Ausreißer oder dem Typ der Ausreißer. Letztere identifizieren einen Extremwert in der Datenmenge als einen Ausreißer. Folgende Tests, die in diese Kategorie fallen, setzen eine Normalverteilung voraus, so dass ein Test auf Normalverteilung diesen Tests vorausgehen sollte. Die Werte der betrachteten Stichprobe müssen zudem aufsteigend geordnet werden.

- Grubbs-Test.

  Der Grubbs-Test überprüft, ob der erste (kleinste) und der letzte (größte) Wert $x_{(1)}$ bzw. $x_{(n)}$ einer Stichprobe der Größe n Ausreißer sind. Die Hypothese lautet, dass beide Werte keine Ausreißer sind. Dazu wird die Abweichung zum arithmetischen Mittel $\bar{x}$ ins Verhältnis mit der Standardabweichung s gesetzt. Die Hypothese wird verworfen, falls die Prüfgrößen T_1 und T_2 folgende Bedingungen erfüllen:

$$T_1 = \frac{\bar{x} - x_{(1)}}{s} > T_{n;1-\alpha}$$
$$T_n = \frac{x_{(n)} - \bar{x}}{s} > T_{n;1-\alpha}$$

Die kritischen Werte $T_{n;1-\alpha}$ bei einer bestimmten Irrtumswahrscheinlichkeit α können aus entsprechenden Tabellen eingesehen werden (Insight (1997)).

- Dixon's r-Statistiken.

  Die Hypothesen entsprechen denen des Grubbs-Tests. Bei diesem Test werden Verhältnisse von Spannweiten definiert. Diese werden wieder für den ersten (1) und den letzten (n) Wert angegeben:

$$r_{kg}(1) = \frac{x_{(1+k)} - x_{(1)}}{x_{(n-g)} - x_{(1)}} \quad k = 1, 2; g = 0, 1, 2$$
$$r_{kg}(n) = \frac{x_{(n)} - x_{(n-g)}}{x_{(n)} - x_{(1+k)}} \quad g = 1, 2; k = 0, 1, 2$$

  Der Test berücksichtigt die g größten und k kleinsten Werte neben dem ersten und letzten Wert. Die Hypothesen werden bei der Irrtumswahrscheinlichkeit α abgelehnt, falls der Wert $r_{kg}(1)$ bzw. $r_{kg}(n)$ den kritischen Wert $r_{kg,n;1-\alpha}$ übersteigt. Die kritischen Werte sind Dixon (1951) zu entnehmen.

- David-Hartley-Pearson-Test.

  Die Hypothese lautet, dass der kleinste oder der größte Wert zur Stichprobe gehört. Diese wird abgelehnt, falls gilt:

$$Q = \frac{x_{(n)} - x_{(1)}}{\sqrt{\frac{1}{1-n} \cdot \sum_{i=1}^{n} \left(x_{(i)} - \bar{x}\right)^2}} > Q_{n;1-\alpha}$$

  Trifft diese Bedingung zu, wird der kleinste oder der größte Wert als Ausreißer behandelt. Die Entscheidung ist abhängig davon, welcher der beiden Werte die größte Differenz vom Mittelwert $\bar{x}$ besitzt. Die kritischen Werte bei einer bestimmten Irrtumswahrscheinlichkeit α finden sich in Hartung (1987).

Bie diesen Methoden wirkt sich nachteiig aus, dass die Verteilung a priori gegeben sein muss, also Wissen, was in den meisten Fällen nicht vorhanden ist. Ein weiterer Nachteil ist, das die Mehrzahl der Tests univariat ist, d.h. sie analysieren nur eine Zufallsvariable.

Tiefenbasierte Ausreißererkennung Jedes Datenobjekt, bzw. jeder Datenwert, wird als ein Punkt in einem m-dimensionalen Raum dargestellt. Dabei wird dem Datenobjekt ein Tiefenwert zugeordnet. Die Datenobjekte werden in Ebenen gemäß der Tiefe geordnet. Bezüglich der Ausreißererkennung sind Ausreißer Datenobjekte, die in der kleinsten

Ebene liegen (Ruts u. Rousseeuw (1996)). Dieser Ansatz hebt die Nachteile der verteilungsbasierten Methoden auf, ist aber abhängig von der Berechnung der Ebenen. Diese basiert auf der Berechnung von m-konvexen Hüllen. Dieser Prozess ist exponentiell und ist praktisch ungeeignet für Dimensionen größer zwei.

Entfernungsbasierte Ausreißererkennung Ein Datenobjekt ist ein entfernungsbasierter Ausreißer in einer Datenmenge, wenn höchstens n Datenobjekte innerhalb einer bestimmten Entfernung vom Ausreißer liegen. Diese Begriffsbildung verallgemeinert mehrere Konzepte. Eines davon basiert auf der Entfernung eines Datenpunktes zu seinem k-nächsten Nachbarn. Nach der Anordnung der Daten entsprechend ihrer Entfernung zum k-nächsten Nachbarn werden die *top-k* Punkte als Ausreißer identifiziert (Ramaswamy u. a. (2000)).

Ausreißer in einer Normalverteilung werden nach der sogenannten „Drei-Sigma-Regel“ bestimmt. Ausreißer sind in diesem Fall Werte, die im Bereich außerhalb des dreifachen der Standardabweichung um den Erwartungswert liegen. Diese Regel basiert auf der Tatsache, dass 99.73 % der Gesamtfläche in diesem Bereich liegt. Außerhalb davon befinden sich kaum gültige Beobachtungen.

Ein weiteres Beispiel ist die Beobachtung von Residuen im Regressionsmodell zur Festlegung von Ausreißern. Die Unterschiede der beobachteten und angepassten Werte werden ausgewertet. Ausreißer sind diejenigen Residuen, die große Differenzen aufweisen.

Dichtebasierte Ausreißererkennung Jedem Datenobjekt wird ein Ausreißer-Faktor zugewiesen, d.h. ein Grad, zudem ein Objekt als Ausreißer betrachtet werden kann. Dieser Grad heißt *local outlier factor* (LOF). Er ist lokal bezüglich einer begrenzten Nachbarschaft. Zur Bestimmung des Faktors wird diese Nachbarschaft und die Dichte der Objekte benutzt (Breunig u. a. (2000)).

Räumliche Ausreißererkennung Räumliche Verfahren erkennen Ausreißer daran, dass diese gegenüber ihren direkten räumlichen Nachbarn starke Abweichungen in den nichträumlichen Attributen aufweisen. Nichträumliche Attribute beziehen sich nicht auf die Position eines Objektes im Raum. Ein räumlicher Ausreißer ist eine lokale Instabilität in ihren nichträumlichen Attributen oder ein räumlich referenziertes Objekt, dessen nichträumliche Attribute extrem relativ zu seiner Nachbarschaft sind, auch wenn sie sich nicht signifikant von der gesamten Population unterscheiden (Shekhar u. a. (2003)).

Clustering Verfahren Clustering Methoden bei der Ausreißererkennung bestimmen Ausreißer als Objekte, die nach dem Clustering keinen Cluster zugeordnet wurden. Ausreißer liegen also in keinem Cluster, sie sind „Nebenprodukte“ oder Rauschen. Allgemein werden Ausreißer während des Clustering-Prozesses toleriert oder ignoriert.

Im Kontext der clusterbasierten Ausreißererkennung stellen Dantong u. a. (1999) eine Methode vor, die auf Wavelet Transformation basiert und die Ausreißer durch Entfernen von Clustern aus der ursprünglichen Datenmenge identifiziert.

Jaing u. a. (2001) präsentieren ein zwei-phasen Clustering-Algorithmus für Ausreißererkennung. In der ersten Phase wird der *k-means* (Hartigan u. Wong (1979)) Algorithmus modifiziert, indem ein neues Clusterzentrum gebildet wird, falls die neue Eingabe weit genug von allen Clusterzentren liegt. Datenpunkte im gleichen Cluster ähneln entweder allen Ausreißern oder allen Nicht-Ausreißern. In der zweiten Phase wird ein minimaler Spannbaum erzeugt. Die kleinsten Cluster, also Bäume mit der geringsten Anzahl an Knoten, werden als Ausreißer betrachtet.

Ausreißererkennung durch Betrachtung von semantischem Wissen He u. a. (a) betrachten zusätzlich Attribute, die Klasseninformationen zu den Objekten enthalten. Es wird vorausgesetzt, dass die benutzten Clustering-Algorithmen Objekte aufgrund der Ähnlichkeit jeweils demselben Cluster zuordnen und dass Objekte der gleichen Klasse ein ähnliches Verhalten aufweisen. Semantische Ausreißer werden als Datenpunkte definiert, die sich anders verhalten als die Punkte derselben Klasse. Jedem Objekt wird ein sogenannter *semantic outlier factor* (SOF) zugewiesen. Er besteht aus dem Produkt der Rate der Objekte mit der gleichen Klassenzugehörigkeit im Cluster und dem Maß, welches die durchschnittliche Ähnlichkeit des betrachteten Objektes zu den Objekten derselben Klasse aufweist. Der resultierende Wert wird ins Verhältnis gesetzt zu der Frequenz der Objektklasse in der Datenmenge.

Einordnung	**Verfahren**
Verteilungsbasiert	Ausreißer in Normalverteilungen
	Freedman u. a. (1978)
	Ausreißer in univariaten Regressionsmodellen
	Draper u. Smith (1966)
	Ausreißer in multivariaten Regressionsmodellen
	Rousseeuw u. Leroy (1987)

Einordnung	**Verfahren**
	Ausreißer in Exponentialverteilungen
	Pawlitschko (2000)
	Statistische Unterscheidbarkeitstest
	Barnett u. Lewis (1994)
	Überwachte Lernverfahren
	Yamanishi u. a. (2000) , Yamanishi u. ichi Takeuchi (2001)
Tiefenbasiert	Tiefenbasierte konvexe Hüllen/
	Tiefenkonturen/
	ISODEPTH
	Ruts u. Rousseeuw (1996)
	Verbesserte Tiefenkonturen, FDC
	Johnson u. a. (1998)
	Schältiefen
	Preparata u. Shamos (1998)
Entfernungsbasiert	Unifizierte entfernungsbasierte
	UO(p,D) bzw. DB(p,D)-Ausreißer
	Knorr u. Ng (1998) , Knorr u. Ng (1997)
	Entfernungsbasiert/Intensional Knowledge
	Knorr u. Ng (1999)
	Entfernungsbasiert zum k-nächsten Nachbarn
	Ramaswamy u. a. (2000)
	Entfernungsbasiert mit zufälliger einfacher Beschneidung
	Bay u. Schwabacher (2003)
	Extended Distance Based Outlier
	Zhixiang
Dichtebasiert	Lokale dichtebasierte Ausreißer-LOF
	Breunig u. a. (2000)
	Top-n LOF Ausreißer
	Jin u. a. (2001)
	Dichtebasierte Ausreißer in Projektionen
	Aggarwal u. Yu (2001)

Einordnung	**Verfahren**
Sequentielle Ausnahmen	Simple Deviation with Smoothing Factor
	Arning u. a. (1996)
	k-d Baum Ausreißer
	Chaudhary u. a.
Clustering Verfahren	Clustereleminierung
	Clustering mit Ausreißerbehandlung
	Jaing u. a. (2001)
	Clusterbasierte lokale Ausreißer CBLOF
	He u. a. (b)
Räumliche Ausreißer	Unified Spatial Ausreißer
	Shekhar u. a. (b) , Shekhar u. a. (a)
	Graphbasierte Spatial Ausreißer
	Shekhar u. a. (2001)
	Spatial Temporal Ausreißer
	Chen u. Li (2004)
Semantische Ausreißer	Semantische Ausreißer-SOF
	He u. a. (a)
	Klassenbasierte Ausreißer
	He u. a. (2005a)
Weitere Ansätze	Biased Sampling for Clustering and Outlier Detection
	Kollios u. a. (2003)
	Complementarity of Clustering and Outlier Detection
	Zhixiang
	Outlier Detection with Replicator Neural Networks
	Hawkins u. a. (2002)
	Hypergraph basierte Ausreißer
	Wei u. a.
	Connectivity Based Outlier
	Tang u. a. (2002)
	Local heuristic based Outlier search
	Frequent Pattern based Outlier

Einordnung	**Verfahren**
	He u. Xu (2005)
	Unified Subspace Outlier Ensemble Framework
	He u. a. (2005b)

Tabelle B.1 Übersicht verschiedenen Verfahren zur Ausreißererkennung nach Deutsch (2006).

Literaturverzeichnis

[Abdi u. Molin 2007] Abdi, Herve ; Molin, Paul: Lilliefors/Van Soest's test of normality. In: Salkind, Neil J. (Hrsg.): *Encyclopedia of Measurement and Statistics*. Sage Publications, 2007, S. 540–544

[Abramowitz u. Stegun 1964] Abramowitz, Milton ; Stegun, Irene A.: *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. ninth Dover printing, tenth GPO printing. New York : Dover, 1964. – ISBN 0–486–61272–4

[Aggarwal u. Yu 2001] Aggarwal, Charu C. ; Yu, Philip S.: Outlier Detection for High Dimensional Data. In: *SIGMOD Conference*, 2001

[Arning u. a. 1996] Arning, Andreas ; Agrawal, Rakesh ; Raghavan, Prabhakar: A Linear Method for Deviation Detection in Large Databases. In: *Knowledge Discovery and Data Mining*, 1996, 164-169

[Barnett u. Lewis 1994] Barnett, Vic ; Lewis, Toby: *Outliers in Statistical Data*. John Wiley and Sons Ltd, 1994. – ISBN 0471930946

[Bay u. Schwabacher 2003] Bay, S. ; Schwabacher, M.: *Mining distance-based outliers in near linear time with randomization and a simple pruning rule*. citeseer.ist.psu.edu/article/bay03mining.html. Version: 2003

[Beichelt u. Montgomery 2003] Beichelt, Frank ; Montgomery, Douglas: *Teubner-Taschenbuch der Stochastik. Wahrscheinlichkeitstheorie, Stochastische Prozesse, Mathematische Statistik*. 1. Teubner, 2003. – ISBN 3–319–00457–7

[Beyer u. a. 1991] Beyer ; Hackel ; Pieper ; Tiedge: *Mathematik für Ingenieure, Naturwissenschaftler, Ökonomen und Landwirte. Wahrscheinlichkeitsrechnung und mathematische Statistik*. Teubner, 1991

[Bishop 1995] Bishop, Christopher M.: *Neural Networks for Pattern Recognition*. New York, NY, USA : Oxford University Press, Inc., 1995. – ISBN 0198538642

[Bosch 1993] Bosch, Karl: *Statistik-Taschenbuch*. R. Oldenbourg Verlag, 1993

[Breunig u. a. 2000] Breunig, Markus M. ; Kriegel, Hans-Peter ; Ng, Raymond T. ; Sander, Jörg: LOF: identifying density-based local outliers, 2000, 93–104

[Carpenter u. a. 1991] Carpenter, Gail A. ; Grossberg, Stephen ; Rosen, David B.: Art 2-A: an adaptive resonance algorithm for rapid category learning and recognition. In: *Neural Netw.* 4 (1991), Nr. 4, S. 493–504. http://dx.doi.org/http://dx.doi.org/10.1016/0893-6080(91)90045-7. – DOI http://dx.doi.org/10.1016/0893–6080(91)90045–7. – ISSN 0893–6080

[Chang u. Lin] Chang, Chih-Chung ; Lin, Chih-Jen: *LIBSVM – A Library for Support Vector Machines*. http://www.csie.ntu.edu.tw/~cjlin/libsvm/. – [Letzter Zugriff am 28.01.2008]

[Chaudhary u. a.] Chaudhary, Amitabh ; Szalay, Alexander S. ; Moore, Andrew W.: *Very Fast Outlier Detection in Large Multidimensional Data Sets*. citeseer.ist.psu.edu/649786.html

[Chen u. Li 2004] Chen, Tao ; Li, Zhilin: *A Multiscale Approach to detect spatial-temporal Outliers*. Department of Land Surveying and Geo-Informatics. http://www.isprs.org/istanbul2004/comm4/papers/490.pdf. Version: 2004. – [Letzter Zugriff am 18.02.2008]

[Coriand 2008] Coriand, Franz: *Neue Verfahren zur intrinsischen Plagiatanalyse*, Bauhaus-Universität Weimar, Fakultät Medien, Mediensysteme, Diplomarbeit, Februar 2008

[Dantong u. a. 1999] Dantong, Yu ; Sheikholeslami, Gholam ; Zhang, Aidong: FindOut: Finding Outliers in Very Large Datasets. Version: 5, 1999. citeseer.ist.psu.edu/yu99findout.html. 1999 (99-03). – Forschungsbericht

[Deutsch 2006] Deutsch, Stephan: *Outlier Detection in USENET Newsgruppen*, Universität Dortmund, Diplomarbeit, 2006

[Dixon 1951] Dixon, W. J.: *Ratios Involving Extreme Values*. http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.aoms/1177729693. Version: 1951. – [Letzter Zugriff am 30.01.2008]

[Draper u. Smith 1966] Draper, N. ; Smith, H.: *Applied Regression Analysis*. John Wiley & Sons, 1966

[Duda u. Hart 1991] Duda ; Hart: *Pattern Classification and Scene Analysis*. Wiley Publication, 1991

[Eitrich 2003] Eitrich, Tatjana: *Support-Vektor-Maschinen: Künstliche Intelligenz und Statistik im Bunde*. http://www.fz-juelich.de/jsc/files/docs/vortraege/ja2003/SVM.pdf. Version: 2003. – [Letzter Zugriff am 28.01.2008]

[Ferber 2003] Ferber, Reginald: *Information Retrieval: Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web*. Heidelberg : dpunkt Verlag, 2003 http://information-retrieval.de/

[Freedman u. a. 1978] Freedman, D. ; Pisan, R. ; Purves, R.: *Statistics*. W.W. Norton, New York, 1978

[Hartigan u. Wong 1979] Hartigan, J. A. ; Wong, M. A.: Algorithm AS 136: A K-Means Clustering Algorithm. In: *Applied Statistics* 28 (1979), Nr. 1, S. 100–108

[Hartung 1987] Hartung, Joachim: *Statistik. Lehr- und Handbuch der angewandten Statistik*. R. Oldenbourg Verlag, 1987

[Hawkins u. a. 2002] Hawkins, S. ; He, H. ; Williams, G. ; Baxter, R.: *Outlier Detection Using Replicator Neural Networks*. citeseer.ist.psu.edu/hawkins02outlier.html. Version: 2002

[He u. a. a] He, Zengyou ; Deng, Shengchun ; Xu, Xiaofei: *Outlier Detection Integrating Semantic Knowledge*. citeseer.ist.psu.edu/547244.html

[He u. Xu 2005] He, Zengyou ; Xu, Xiaofei: FP-Outlier: Frequent Pattern Based Outlier Detection. In: *Comput. Sci. Inf. Syst.* 2 (2005), Nr. 1, S. 103–118

[He u. a. b] He, Zengyou ; Xu, Xiaofei ; Deng, Shengchun: *Discovering Cluster Based Local Outliers*. citeseer.ist.psu.edu/547375.html

[He u. a. 2005a] He, Zengyou ; Xu, Xiaofei ; Deng, Shengchun: An Optimization Model for Outlier Detection in Categorical Data. In: *CoRR* abs/cs/0503081 (2005). http://dblp.uni-trier.de/db/journals/corr/corr0503.html#abs-cs-0503081. – informal publication

[He u. a. 2005b] He, Zengyou ; Xu, Xiaofei ; Deng, Shengchun: A Unified Subspace Outlier Ensemble Framework for Outlier Detection in High Dimensional Spaces. In: *CoRR* abs/cs/0505060 (2005). http://dblp.uni-trier.de/db/journals/corr/corr0505.html#abs-cs-0505060. – informal publication

[Hertz u. a. 1991] Hertz, John ; Krogh, Anders ; Palmer, Richard G.: *Introduction to the theory of neural computation*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc., 1991. – ISBN 0–201–50395–6

[Insight 1997] Insight, GraphPad: *Grubbs' Test for Detecting Outliers*. http://www.graphpad.com/articles/grubbs.htm. Version: 1997. – [Letzter Zugriff am 30.01.2008]

[Jaing u. a. 2001] Jaing, M. F. ; Tseng, S. S. ; Su, C. M.: Two-phase clustering process for outliers detection. In: *Pattern Recogn. Lett.* 22 (2001), Nr. 6-7, S. 691–700. http://dx.doi.org/http://dx.doi.org/10.1016/S0167-8655(00)00131-8. – DOI http://dx.doi.org/10.1016/S0167–8655(00)00131–8. – ISSN 0167–8655

[Japkowicz u. a. 1995] Japkowicz, Nathalie ; Myers, Catherine ; Gluck, Mark A.: A Novelty Detection Approach to Classification. In: *IJCAI*, 1995, 518-523

[Jin u. a. 2001] Jin, Wen ; Tung, Anthony K. H. ; Han, Jiawei: Mining top-n local outliers in large databases. In: *Knowledge Discovery and Data Mining*, 2001, 293-298

[Johnson u. a. 1998] Johnson, Theodore ; Kwok, Ivy ; Ng, Raymond T.: Fast Computation of 2-Dimensional Depth Contours. In: *Knowledge Discovery and Data Mining*, 1998, 224-228

[Knorr u. Ng 1997] Knorr, Edwin M. ; Ng, Raymond T.: A Unified Notion of Outliers: Properties and Computation. In: *Knowledge Discovery and Data Mining*, 1997, 219-222

[Knorr u. Ng 1998] Knorr, Edwin M. ; Ng, Raymond T.: Algorithms for Mining Distance-Based Outliers in Large Datasets. In: *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, 1998, 392–403

[Knorr u. Ng 1999] Knorr, Edwin M. ; Ng, Raymond T.: Finding Intensional Knowledge of Distance-Based Outliers. In: *The VLDB Journal*, 1999, 211-222

[Kohonen 1997] Kohonen, Teuvo (Hrsg.): *Self-organizing maps*. Secaucus, NJ, USA : Springer-Verlag New York, Inc., 1997. – ISBN 3–540–62017–6

[Kollios u. a. 2003] Kollios, G. ; Gunopulos, D. ; Koudas, N. ; Berchtold, S.: Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Datasets. In: *IEEE Transactions on Knowledge and Data Engineering* 15 (2003), Nr. 5. citeseer.ist.psu.edu/kollios03efficient.html

[Manevitz u. Yousef 2002] Manevitz, Larry M. ; Yousef, Malik: One-class svms for document classification. In: *J. Mach. Learn. Res.* 2 (2002), S. 139–154. – ISSN 1533–7928

[Mason u. Bell 1986] Mason ; Bell: New Lilliefors and Srinivasan Tables with Applications. In: *Communications in Statistics: Simulation and Computation* 15 (1986), S. 451–477

[Meyer zu Eißen u. Stein 2006] Meyer zu Eissen, Sven ; Stein, Benno: Intrinsic Plagiarism Detection. In: Lalmas, Mounia (Hrsg.) ; MacFarlane, Andy (Hrsg.) ; Rüger, Stefan M. (Hrsg.) ; Tombros, Anastasios (Hrsg.) ; Tsikrika, Theodora (Hrsg.) ; Yavlinsky, Alexei (Hrsg.): *ECIR* Bd. 3936, Springer, 2006 (Lecture Notes in Computer Science). – ISBN 3–540–33347–9, 565-569

[Meyer zu Eissen u. a. 2007] Meyer zu Eissen, Sven ; Stein, Benno ; Kulig, Marion: Plagiarism Detection without Reference Collections. (2007), S. 359–366. ISBN 978–3–540–70980–0

[Meyer zu Eißen u. a. 2005] Meyer zu Eissen, Sven ; Stein, Benno ; Potthast, Martin: The Suffix Tree Document Model Revisited. In: Tochtermann, Klaus (Hrsg.) ; Maurer, Hermann (Hrsg.): *Proceedings of the 5th International Conference on Knowledge Management (I-KNOW 05), Graz, Austria*, Know-Center, Juli 2005 (Journal of Universal Computer Science). – ISSN 0948–695x, S. 596–603

[Parzen 1962] Parzen, E.: On estimation of a probability density function and mode. In: *Annals of Mathematical Statistics* 33 (1962), S. 1065–1076

[Pawlitschko 2000] Pawlitschko, J.: *the distribution of a test statistic for outlier identication in exponential samples.* citeseer.ist.psu.edu/pawlitschko00distribution.html. Version: 2000

[Potthast u. a. 2008] Potthast, Martin ; Stein, Benno ; Anderka, Maik: A Wikipedia-Based Multilingual Retrieval Model. In: *30th European Conference on IR Research, ECIR 2008, Glasgow* Bd. 4956 LNCS. Berlin Heidelberg New York : Springer, 2008 (Lecture Notes in Computer Science). – ISSN 0302–9743, S. 522–530

[Preparata u. Shamos 1998] Preparata, F. ; Shamos, M.: *Computational Geometry: An Introduction.* Springer, 1998

[Ramaswamy u. a. 2000] Ramaswamy, Sridhar ; Rastogi, Rajeev ; Shim, Kyuseok: Efficient algorithms for mining outliers from large data sets, 2000, 427–438

[Rijsbergen 1979] Rijsbergen, C. J. V.: *Information Retrieval*. Newton, MA, USA : Butterworth-Heinemann, 1979. – ISBN 0408709294

[Rinne 1997] Rinne, Horst: *Taschenbuch der Statistik*. 2. Frankfurt am Main, Thun : Verlag Harri Deutsch, 1997. – ISBN 3–8171–1559–8

[Rousseeuw u. Leroy 1987] Rousseeuw, P. J. ; Leroy, A. M.: *Robust regression and outlier detection*. New York, NY, USA : John Wiley & Sons, Inc., 1987. – ISBN 0–471–85233–3

[Ruts u. Rousseeuw 1996] Ruts, Ida ; Rousseeuw, Peter J.: Computing depth contours of bivariate point clouds. In: *Comput. Stat. Data Anal.* 23 (1996), Nr. 1, S. 153–168. http://dx.doi.org/http://dx.doi.org/10.1016/S0167-9473(96)00027-8. – DOI http://dx.doi.org/10.1016/S0167–9473(96)00027–8. – ISSN 0167–9473

[Sachs u. Hedderich 2006] Sachs, Lothar ; Hedderich, Jürgen: *Angewandte Statistik. Methodensammlung mit R*. Springer, 2006

[Schölkopf u. a. 1999] Schölkopf, Bernhard ; Platt, John C. ; Shawe-Taylor, John ; Smola, Alex J. ; Williamson, Robert C.: Estimating the support of a high-dimensional distribution. (1999). http://citeseer.ist.psu.edu/article/sch99estimating.html

[Shekhar u. a. 2003] Shekhar, S. ; Lu, C. ; Zhang, P.: *A unified approach to spatial outliers detection*. citeseer.ist.psu.edu/article/shekhar03unified.html. Version: 2003

[Shekhar u. a. a] Shekhar, Shashi ; Lu, Chang-Tien ; Zhang, Pusheng: A Unified Approach to Detecting Spatial Outliers. citeseer.ist.psu.edu/article/shekhar03unified.html

[Shekhar u. a. 2001] Shekhar, Shashi ; Lu, Chang-Tien ; Zhang, Pusheng: Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In: *the Seventh SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD 2001)*, 2001, 371-376

[Shekhar u. a. b] Shekhar, Shashi ; Zhang, Pusheng ; Huang, Yan ; Vatsavai, Ranga R.: *Trends in Spatial Data Mining*. citeseer.ist.psu.edu/694073.html

[Stein u. a. 2007] Stein, Benno ; Eissen, Sven M. ; Potthast, Martin: Strategies for Retrieving Plagiarized Documents. In: Clarke, Charles (Hrsg.) ; Fuhr, Norbert

(Hrsg.) ; KANDO, Noriko (Hrsg.) ; KRAAIJ, Wessel (Hrsg.) ; VRIES, Arjen de (Hrsg.): *30th Annual International ACM SIGIR Conference*, ACM, July 2007. – ISBN 987–1–59593–597–7, S. 825–826

[Stein u. Meyer zu Eißen 2006] STEIN, Benno ; MEYER ZU EISSEN, Sven: Near Similarity Search and Plagiarism Analysis. In: SPILIOPOULOU, M. (Hrsg.) ; KRUSE, R. (Hrsg.) ; BORGELT, C. (Hrsg.) ; NÜRNBERGER, A. (Hrsg.) ; GAUL, W. (Hrsg.): *From Data and Information Analysis to Knowledge Engineering*, Springer, 2006. – ISBN 1431–8814, S. 430–437

[Stein u. Meyer zu Eissen 2007] STEIN, Benno ; MEYER ZU EISSEN, Sven: Intrinsic Plagiarism Analysis with Meta Learning. (2007), Juli, 45-50. http://ceur-ws.org/Vol-276. – ISSN 1613–0073

[Storm 1995] STORM, Regina: *Wahrscheinlichkeitsrechnung, mathematische Statistik und statistische Qualitätskontrolle*. Fachbuchverlag Laipzig - Köln, 1995

[Tang u. a. 2002] TANG, J. ; CHEN, Z. ; FU, A. ; CHEUNG, D.: *A Robust Outlier Detection Scheme in Large Data Sets*. citeseer.ist.psu.edu/tang01robust.html. Version: 2002

[Tax u. Duin 2001] TAX, David M. J. ; DUIN, Robert P. W.: Combining One-Class Classifiers. In: *Lecture Notes in Computer Science* 2096 (2001), 299–?? citeseer.ist.psu.edu/tax01combining.html

[Tax 2001] TAX, D.M.J.: *One-class classification; Concept-learning in the absence of counter-examples*, Delft University of Technology, Diss., 2001

[Tipping u. Bishop 1999] TIPPING, Michael E. ; BISHOP, Christopher M.: Mixtures of Probabilistic Principal Component Analysers. In: *Neural Computation* 11 (1999), Nr. 2, 443-482. citeseer.ist.psu.edu/tipping98mixtures.html

[Wei u. a.] WEI, Li ; QIAN, Weining ; ZHOU, Aoying ; JIN, Wen ; YU, Jeffrey X.: *HOT: Hypergraph-based Outlier Test for Categorical Data*. citeseer.ist.psu.edu/644023.html

[Weisstein 2007] WEISSTEIN, Eric W.: *Quartile*. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Quartile.html. Version: 2007. – [Letzter Zugriff am 26.01.2008]

[Wikipedia 2008] Wikipedia: *Recall und Precision*. http://de.wikipedia.org/wiki/Recall_und_Precision. Version: 2008. – [Letzter Zugriff am 28.01.2008]

[Yamanishi u. ichi Takeuchi 2001] Yamanishi, Kenji ; Takeuchi, Jun ichi: Discovering outlier filtering rules from unlabeled data: combining a supervised learner with an unsupervised learner. In: *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA : ACM, 2001. – ISBN 1–58113–391–X, S. 389–394

[Yamanishi u. a. 2000] Yamanishi, Kenji ; Takeuchi, Junichi ; Williams, Graham J. ; Milne, Peter: On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In: *Knowledge Discovery and Data Mining*, 2000, 320-324

[Ypma u. Duin 1998] Ypma, A. ; Duin, R.: *Support objects for domain approximation*. citeseer.ist.psu.edu/ypma98support.html. Version: 1998

[Zhixiang] Zhixiang, Detection S.: *On Complementarity of Cluster and Outlier*. citeseer.ist.psu.edu/610797.html