Ranking im Vergleich mit Hyperrectangle und Normalisierung als Verfahren zur Klassifizierung von Daten

Bachelorarbeit im Fach Informatik
am Institut für Informatik
Universität Zürich

Prof. Dr. Abraham Bernstein
Betreuer: Dr. Jörg-Uwe Kietz

Vorgelegt von
Patrick Leibundgut
Hochdorf, Luzern, Schweiz
Matrikelnummer: 05-913-041

1. September 2010

Abstract

For the classification of instances you can use different methods. The use of geometric distance or semantic distance for the kNN method provides a different result depending on the distribution of the attributes. The semantic distance is because of its incorrect interpretation of the distance significantly less correct with the classifications and thus proves to be unsuitable for a classification. The comparison of the results of the ranking and the normalization as pre-processing methods shows, that the ranking got better results in the classification as the normalization with skew distributed attributes. The normalisation performs better for attributes, that are not skew distributed,.

Zusammenfassung

Für die Klassifizierung von Instanzen kann man verschiedene Verfahren anwenden. Die Verwendung von geometrischem Abstand oder semantischem Abstand für das kNN Verfahren liefert in Abhängigkeit von der Verteilung der Attribute unterschiedliche Ergebnisse. Der semantische Abstand ergibt wegen ungeeigneter Betrachtung des Abstandes signifikant schlechtere Klassifizierungen und erweist sich daher als ungeeignet für eine Klassifizierung. Der Vergleich von Ranking und Normalisierung als Preprocessing Methoden ergibt, dass das Ranking bei schief verteilten Attributen bessere Ergebnisse bei der Klassifizierung liefert als bei Normalisierung. Bei nicht schief verteilten Attributen liefert die Normalisierung bessere Resultate.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

Das Ziel dieser Arbeit ist ein Vergleich zwischen dem Ranking und der Normalisierung als Preprocessing Varianten auf die Attribute der Verteilung vor der Anwendung der kNN Klassifikation (geometrischer Abstand). Ebenso wird ein Vergleich zwischen dem erwähnten Ranking und der kNN Klassifikation mit dem Hyperrectangle Verfahren (semantischer Abstand) zur Klassifikation von Datensätzen gemacht.

In dieser Arbeit wird die Bearbeitung von Attributen bei der Klassifizierung mit dem k nächsten Nachbarn (kNN) Verfahren untersucht. Gibt es einen Zusammenhang zwischen den Resultaten mit der ursprünglichen Verteilung der einzelnen Attribute und deren Veränderung mittels Normalisierung, Ranking oder semantischer Ähnlichkeit vor dem Beginn der Klassifizierung? Um dies zu überprüfen, wurde ein Programm in Java geschrieben. Es verfügt über Funktionen vom Einlesen der Datensätze, über die Bearbeitung der Attribute mit Normalisierung und Ranking, bis hin zur Klassifizierung und der anschliessenden Auswertung der Richtigkeit der Klassifizierung.

Bei dieser Untersuchung interessiert vor allem der Zusammenhang zwischen Ranking und Normalisierung, welche vor der Klassifizierung der Instanzen auf die einzelnen, schief verteilten Attribute angewendet werden.

Die beiden zu überprüfenden Hypothesen untersuchen die folgenden Zusammenhänge:

Als erstes den Zusammenhang der kNN Verfahren mit Ranking als Preprocessing und dem Hyperrectangle Verfahren bei schief verteilten Attributen. Die erste Hypothese sagt aus, dass sich Ranking als Preprocessing Methode mit anschliessender kNN Klassifizierung ähnlich verhält wie dies das Hyperrectangle Verfahren macht.

Und Zweitens wird der Zusammenhang zwischen Ranking und Normalisierung als Preprocessing für die Attribute untersucht. Die zweite Hypothese besagt, dass bei schief verteilten Attributen das Ranking als Preprocessing Methode besser abschneidet als die Normalisierung - und bei nicht schief verteilten Attributen gilt der umgekehrte Fall.

1.1 Vorgehensweise

Diese Arbeit geht dem Vergleich von Ranking als Preprocessing Methode mit anschliessender kNN Klassifikation und dem Hyperrectangle Verfahren mit integrierter kNN Klassifikation auf die Spur. Untersucht werden der Effekt und die Effizienz bei schief verteilten Attributen in Datensätzen.

Zu diesem Zweck wurden 16 Datensätze gesammelt, die alle schief verteilte sowie nicht schief verteilte Attribute beinhalten. Um die Datensätze für das Programm gebrauchen zu können, mussten vor der Verwendung gewisse Transformationen und Anpassungen vorgenommen werden. Wie im Abschnitt 3 mit den Ergebnissen ersichtlich ist, sind pro Datensatz fünf Durchläufe gemacht worden: Ein Datensatz besteht aus allen Attributen und je zwei Datensätze bestehen nur aus den schief verteilten Attributen, beziehungsweise den nicht schief verteilten Attributen. Die Bestimmung der Schiefe erfolgt einmal durch die normalisierte Schiefe und das zweite Mal durch die Schiefe, welche mit dem 3. zentralen Moment bestimmt wird. Für das Programm wurden die Datensätze getrennt und einzeln abgespeichert, damit es diese ohne Probleme verarbeiten konnte. Für die verschiedenen Durchläufe wurden die Datensätze anhand der Schiefe der Attribute aufgeteilt und in das Programm gespeist. So wurde aufgrund der einzelnen Verteilungen der Attribute die Entscheidung getroffen, ob sie im Datensatz der schiefen Attribute oder dann in jenem der nicht schiefen Attribute stehen werden.

Das Programm wurde in Java geschrieben und mit den nötigen Fähigkeiten versehen. Das Programm kann die Datensätze einlesen, sie verarbeiten und die Resultate auswerten und ausgeben. Java als Programmiersprache wurde für dieses Programm gewählt, weil die Möglichkeit für ein Ausführen des Programms auf Rechnern mit verschiedenen Plattformen gegeben war. Als Nachteil stellte sich die Geschwindigkeit der Kommunikation mit dem SQL Server heraus, wie auch Einbussen in der Verarbeitungsgeschwindigkeit. Der Versuch der Optimierung der Geschwindigkeit mittels SQL Abfragen beim Hyperrectangle Verfahren schlug fehl. Das alternative Vorgehen war noch weniger performant als die erste iterative Variante.

Um die Resultate interpretieren zu können, war eine genaue Analyse der Datensätze von Nöten. Dies machte in einem ersten Schritt mit der Identifizierung von schief verteilten Attributen das Java Programm. In einem nächsten detaillierteren Schritt wurde mit Hilfe

vom Data Audit in SPSS Clementine[1] ein Output mit hilfreichen und nützlichen Daten zu den einzelnen Datensätzen produziert.
Die Betrachtung und Beurteilung der Resultate folgte danach unter Einbezug der oben aufgeführten Methoden und Ergebnisse.
Nach dem Beschrieb der Vorgehensweise der Berechnung folgen nun Definitionen von Schiefe und eine Beschreibung der Datensätze. In einem nächsten Kapitel werden die Klassierungsfunktionen des Java Programms theoretisch erklärt. Darauf folgen die Ergebnisse mit einer abschliessenden Interpretation.

1.2 Schiefe Verteilung

In diesem Kapitel wird dem Begriff Schiefe nachgegangen und die essentiellen Bereiche erklärt. Für dieses Kapitel wurde das Buch von Schira (Schira, 2009) verwendet.
Unter der Schiefe wird die Neigungsstärke der statistischen Verteilung der einzelnen Attribute verstanden. Es gibt eine positive (rechtsschiefe/linkssteil) und eine negative (linksschiefe/rechtssteil) Neigung. Zur Bestimmung der Schiefe der einzelnen Attribute wurden zwei verschiedene Masse verwendet. Beide Masse reagieren stark auf Ausreisser, weil die Abweichungen vom Mittelwert mit der 3. Potenz auf das Mass der Schiefe einwirken. Als Konsequenz findet man keine allgemein akzeptierten Grenzwerte für die Unterscheidung zwischen schief und nicht schief.

1.2.1 Schiefe mit dem 3. zentralen Moment

Um die Schiefe herauszufinden, wurde in dieser Arbeit als ersten Ansatz das zentrale Moment 3. Ordnung genommen. Dies weist bei einer symmetrischen Verteilung den Wert 0 auf. Wird ein Wert grösser als 0 erreicht, bedeutet dies eine Rechtsschiefheit der Verteilung. Ist der Wert des Moments kleiner als 0, dann liegt eine linksschiefe Verteilung vor.

[1] http://www.spss.com/software/modeling/modeler-pro/

Das 3. zentrale Moment ist wie folgt definiert:

$$M_3^Z = E(X - \mu)^3$$

$M_3^Z: 3. zentrales\ Moment$

$X: Zufallsvariable$

$E(\dots): Erwartungwert$

$\mu = E(X): Erwartungwert\ von\ X$

Oder anders mit einer Summenformel geschrieben[2]:

$$M_3^Z = \frac{\sum_{i=1}^{n}(x_i - m)^3}{(n-1)}$$

$M_3^Z: 3. zentrales\ Moment$

$n: Anzahl\ der\ Werte$

$x_i: aktueller\ Wert$

$m: Mittelwert\ der\ Zufallsvariabeln$

In Java gibt es eine Math Library von Apache[3], mit verschiedenen Packages für die Berechnung von mathematischen Grössen und, hier wesentlich, Methoden für statistische Grössen. Für das Programm wurde die Klasse „ThirdMoment"[4] verwendet. Diese berechnet eine Statistik vom 3. zentralen Moment von Listen mit Gleitkommazahlen. Diese Implementation der Klasse wurde mit Datenreihen, welche sowohl normalverteilte, wie auch schief verteilte Attribute hatten, getestet. Sie liefert stets die erwarteten, wie oben beschriebenen Werte. Für linksschiefe Verteilungen ist der Wert negativ, bei einer symmetrischen Verteilung 0 und im Falle einer Rechtsschiefheit positiv. Eine genaue Berechnung des Wertes zeigt die nachfolgende iterative Formel, welche genau der obigen Summenformel entspricht:

[2] http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm
[3] http://commons.apache.org/math/api-1.1/overview-summary.html
[4] org.apache.commons.math.stat.descriptive.moment.ThirdMoment

$$M_{3,new} = M_{3,old} - 3 * \frac{dev}{n} * M_2 + (n-1)(n-2) * \frac{dev^3}{n^2}$$

$$M_{2,new} = M_{2,old} + dev^2 * \frac{(n-1)}{n}$$

$M_{3,new}; M_{3,old}: 3. zentrales\ Moment\ nach\ und\ vor\ Durchlauf$
$M_{2,new}; M_{2,old}: 2. zentrales\ Moment\ nach\ und\ vor\ Durchlauf$
$dev: aktueller\ Wert - vorheriger\ Mittelwert$
$n: Anzahl\ Werte\ inklusive\ aktuellem\ Wert$

1.2.2 Normalisierte Schiefe

Das 3. zentrale Moment alleine ist noch kein geeignetes Mass für die Asymmetrie einer Verteilung, da die Masseinheit der Zufallsvariablen die Streuung beeinflusst. Um dies zu unterdrücken, wird dieses 3. zentrale Moment auf dem Wertebereich normalisiert. Danach bekommt man die normierte Schiefe (Skewness), dessen Formel wie folgt aussieht:

$$Skewness = \frac{n}{(n-1)(n-2)} * \sum_{i=1}^{n} \frac{(x_i - m)^3}{s^3}$$

$n: Anzahl\ der\ Werte$
$xi: aktueller\ Wert$
$m: Mittelwert$
$s: Standardabweichung$

Die obige Variante der Schiefe wird auch vom Programm SPSS Clementine verwendet und ist auf den Data Audits der Datensätze in der zweitletzten Spalte sichtbar (norm. Schiefe). Im Abschnitt 1.2.4 folgt ein solcher Data Audit als anschauliches Beispiel. Für das Programm wurde dafür die Klasse „Skewness"[5] von der zuvor verwendeten Apache Library verwendet und implementiert.

[5] org.apache.commons.math.stat.descriptive.moment.Skewness

1.2.3 Vergleich der beiden Masse für Schiefe

Für die Arbeit wurden beide Masse für die Schiefe betrachtet. In beiden Fällen wurde ein fixer Grenzwert, der eine Unterteilung in schief oder nicht schief vornimmt, für das Mass der Schiefe bestimmt und alle Datensätze an diesem Wert gemessen. Eine genauere Erklärung zum Ablauf im Programm findet man im Kapitel 2.1, Datensatzanalyse.
Was sagen diese beiden Masse für Schiefe nun aus? Sind beide geeignet für das Problem dieser Arbeit? Für diese Arbeit ist es nicht sinnvoll, ein nicht normiertes Mass für die Bestimmung der Verteilung der Attribute zu nehmen. Denn dann fliesst ein zusätzlicher Faktor, der Wertebereich der Attribute, mit in die Ergebnisse. Das kann die Ergebnisse stark verändern, was man beim späteren Vergleich der beiden Ergebnisse mit unterschiedlichen Verwendungen der Schiefe klar erkennt.
Die richtige Entscheidung ist folglich die über den Wertebereich der Attribute normierte Schiefe, wobei zu beachten ist, dass Ranking und Normalisierung ebenfalls eine Form der Normierung über den Werteberich darstellt. Die normierte Schiefe gibt eine vergleichbare Schiefe über alle Attribute und Datensätze.
Die nächste Frage ist die Entscheidung über den Grenzwert der Schiefe. Für die Arbeit wird eine Schiefe von ±1 als Grenze angenommen. Der Frage, ob dieser bestimmte Wert aussagekräftig ist oder nicht wird in dieser Arbeit nicht nachgegangen. Eine weiterführende Frage wäre, wie sich die Resultate bei einem unterschiedlichen Grenzwert der Schiefe verhalten würden. Der vermutete Zusammenhang wäre eine geringere Richtigkeit der Klassifizierung bei sehr kleinen und sehr grossen Grenzwerten für die Schiefe und eine höhere Richtigkeit bei den Grenzwerten für die Schiefe dazwischen.

1.2.4 Veranschaulichung der Schiefe

Ein typisches Beispiel für eine schiefe, linkssteile Verteilung ist der Lohn. Es gibt viele Leute mit einem Lohn in tieferen Regionen um CHF 5‘000.- pro Monat und nur wenige mit einem Lohn über CHF 50‘000.- pro Monat. In der untenstehenden Abbildung sieht man links eine solche linkssteile Verteilung mit einer stetigen Kurve. Rechts ist das Umgekehrte, eine rechtssteile Verteilung zu sehen. Auf der horizontalen Achse ist der Wert abgetragen und auf der vertikalen Achse deren Häufigkeit.

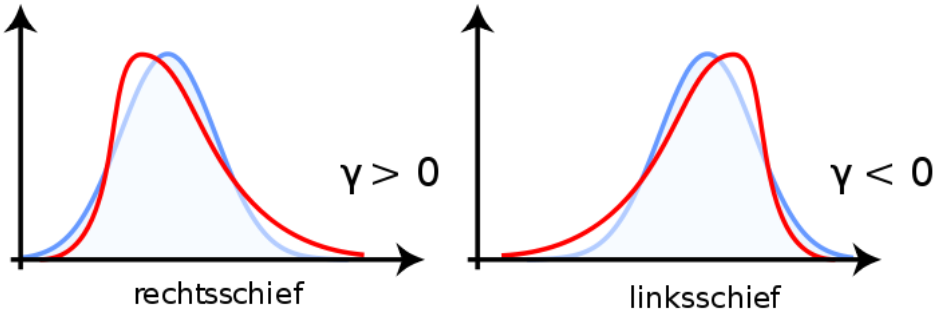


Abbildung 1: Linkssteile und rechtssteile stetige Verteilung[6]

In der nachfolgenden Abbildung sieht man einen Output vom Programm SPSS Clementine, in welchem man Informationen zu den Attributen eines Datensatzes herauslesen kann. Als Beispiel wird der Datensatz „diabetes" gewählt. Man sieht, dass es acht verschiedene Attribute gibt. Die erste Spalte ist der Name des Attributes. In der nächsten Spalte ist die Verteilung der einzelnen Instanzen in einem Histogramm abgebildet. Die dritte Spalte stellt den Bereich der Attributwerte dar. Diesen jeweiligen Wert erhält man indem man den maximalen Wert minus den minimalen Wert rechnet. Nach dem Mittelwert und der Standardabweichung folgt die normierte Schiefe, die durch das Programm SPSS Clementine selbst berechnet wird. Dieser Wert entspricht genau dem vom Java Programm errechneten Wert für die normierte Schiefe. Die letzte Spalte zeigt den vom Java Programm errechneten Wert für das 3. zentrale Moment der Verteilung.

Für das Java Programm gilt ein Attribut als schief, wenn der Wert der Schiefe mit dem 3. zentralen Moment („Schiefe z.Mom") über ±500'000 liegt. Für die normierte Schiefe wird der Grenzwert von ±1 gewählt. Diese Bestimmung der Grenzwerte beruht auf einer empirischen Untersuchung der jeweiligen Werte der Schiefe der vorliegenden Datensätze.

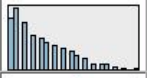
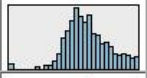
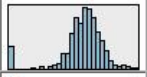
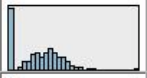
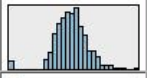
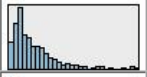
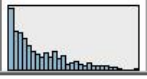
Wenn man jetzt im Beispiel von „diabetes" schaut, sind bei Verwendung des 3. zentralen Momentes die Attribute 2, 3, 5 und 8 schief verteilt, denn der Wert der Schiefe liegt über dem Grenzwert. Dieser Grenzwert wird für alle Datensätze verwendet.

Für den zweiten Durchlauf mit der normierten Schiefe sind die Attribute 3, 5, 7 und 8 schief verteilt. Alle besitzen einen absoluten Wert der Schiefe, welcher über 1 liegt.

[6] Quelle : http://de.wikipedia.org/wiki/Schiefe_(Statistik)

Man erkennt bereits hier, dass die Ergebnisse für die beiden unterschiedlichen Masse für die Schiefe nicht identisch sein werden. Es werden in beiden Fällen nicht die gleichen Attribute als schief angeschaut.

Data Audit von diabetes

	Feld	Diagramm	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	number of times pregnant		17	3.845	3.370	0.902	26390.0
2	plasma glucose concentration		199	120.895	31.973	0.174	4344412.8
3	diastolic blood pressure		122	69.105	19.356	-1.844	-10227402.64
4	triceps skin fold thickness		99	20.536	15.952	0.109	339652.0
5	2-Hour serum insulin		846	79.799	115.244	2.272	2660566924.0
6	BMI		67.100	31.993	7.884	-0.429	-160830.3
7	diabetes pedigree function		2.342	0.472	0.331	1.920	53.4
8	age		60	33.241	11.760	1.130	1405510.0

Vom Java Programm berechnet

Abbildung 2: Datenaudit von SPSS für Datensatz diabetes

1.3 Datensätze

Alle Datensätze, welche in dieser Evaluation verwendet wurden, sind auf Onlinedatenbanken zugänglich. Für diese Arbeit wurden Datensätze von der UCI-Repository[7], dem KDDCUP[8] und LIBSVM[9] einer weiteren Sammlung von Datensätzen diverser Onlineseiten verwendet.

Die Datensätze sind Datenreihen mit einer Klassifizierung. Sie bestehen alle aus verschiedenen Attributen und einer Klassifizierungsvariablen. Dabei sind alle Attribute

[7] http://archive.ics.uci.edu/ml/
[8] http://www.sigkdd.org/kddcup/index.php
[9] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#fourclass

Zahlenwerte und es gibt keine fehlenden Werte. Die Klassifizierungsvariable ist ebenfalls eine Zahl und befindet sich immer an letzter Stelle in einer Zeile. Es gibt Datensätze mit zwei oder mehreren Klassen.

Alle Datensätze liegen in einer CSV-Datei vor, welche die einzelnen Attribute und Klassenvariablen mit einem Komma trennt. Die Trennung der unterschiedlichen Instanzen erfolgt durch einen Zeilenumbruch. Diese genaue Formatierung muss gegeben sein, dass das Programm die einzelnen Datensätze korrekt einliest.

Ein wichtiges Kriterium an den Datensatz ist, dass ein Teil der Attribute eine schiefe Verteilung aufweisen muss. Es dürfen nicht alle Attribute schief sein, denn dann kann der Einfluss des Loslösens der schiefen Attribute bei den verschiedenen Testverfahren nicht untersucht werden.

1.4 Probleme

Bei der Datensatzbeschaffung stellte sich die Suche nach geeigneten Daten als schwierig heraus. Es liessen sich viele verschiedene Datensätze finden. Datensätze, welche den Anforderungen für diese Arbeit genügten, waren jedoch Mangelware. Es brauchte eine intensive Suche im Internet in den Datenbanken, um 16 verschiedene Datensätze zu bekommen. Jeder in Frage kommende Datensatz wurde heruntergeladen, formatiert und danach analysiert. Falls der Datensatz keine schief verteilten Attribute aufwies, war die Arbeit umsonst und die Suche musste fortgesetzt werden.

Das Hyperrectangle Verfahren (beschrieben in Kapitel 2.3.2, Semantischer Abstand) stellte sich als das interessanteste der drei Verfahren der gesamten Arbeit heraus. Es brauchte drei unterschiedliche Programmieransätze, um am Schluss zu einem brauch- und anwendbaren Programm zu kommen. Die erste rein iterative Variante war zu langsam, funktionierte aber sehr genau. Beim zweiten Versuch wurde das Programm mit der Datenbankanbindung noch langsamer und musste nochmals überdacht werden. Danach zog man einen Kompromiss in Betracht. Man beschränkte sich - auf Kosten der Genauigkeit - auf die sieben ähnlichsten Instanzen und wertete nur diese aus. Somit konnten die Schleifen früher durch hinzugefügte Abbruchbedingungen unterbrochen werden, was zu einer Zeitersparnis im Programm führte und den Rechenaufwand reduzierte.

2. Implementierung

Für diese Arbeit wurde ein Standalone-Programm in Java erstellt. Das Programm wurde in der Entwicklungsumgebung Eclipse[10] geschrieben und kann als Executable-Jar[11] auch ausserhalb der Umgebung und plattformenunabhängig auf anderen Rechnern ausgeführt werden.

Das Programm liest die Datensätze, welche als CSV-Dateien[12] vorliegen, ein und verarbeitet sie je nach Aufruf und gegebenen Parametern. So ist es möglich, das Programm ohne Änderung auf viele verschiedene Datensätze anzuwenden. Als nächstes führt das Programm wahlweise die Normalisierung oder das Ranking auf den einzelnen Attributen aus, um sie anschliessend mit dem kNN Verfahren (siehe Kapitel 2.3) zu klassifizieren oder benutzt die semantische Ähnlichkeit mit dem Hyperrectangle Verfahren. Als Letztes evaluiert das Programm die Klassifizierung anhand der ursprünglichen und „richtigen" Klasse. Zu beachten ist, dass die ursprüngliche Klassierung der Datensätze eine Gewichtung der Attribute enthalten kann, während die vorgestellten, respektive. die analysierten Verfahren jedoch eine Gleichgewichtung unterstellen. Die Auswertung der Resultate wird dann in eine Datei geschrieben und sowohl auf die Festplatte, wie auch auf einen FTP-Server gespeichert. Dieser Vorgang ermöglicht die zentrale Betrachtung der Ergebnisse nach der Ausführung des Programms an verschiedenen Rechnern.

2.1 Datensatzanalyse

Bevor die Datensätze in die Klassifizierungsverfahren des Programms gesteckt werden, muss sichergestellt sein, dass die Daten den Anforderungen genügen. Dies wird mit einer Analyse des Datensatzes durch ein schnelles, selbstgeschriebenes Java Programm getan. Hierzu werden alle Attribute auf ihre Verteilungsschiefe hin untersucht. Danach wird das Ergebnis in eine Datei geschrieben. In dieser Datei sieht man eine Zusammenfassung ausschlaggebender Kriterien. Es sind dies die gesamte Anzahl der Attribute und die Anzahl der schief verteilten Attribute. Als weitere Information für das Erstellen der drei Datensätze

[10] http://www.eclipse.org/

[11] Eine JAR-Datei ist ein Archiv mit zusätzlichen Metainformationen. Damit können Javaprogramme verpackt und gestartet werden.

[12] Textdatei mit Werten, die durch Kommas getrennt sind. (Comma Seperated Values)

(alle Attribute, nur schiefe Attribute, ohne schiefe Attribute) steht in dieser Infodatei explizit, welche Attribute schief verteilt sind.
Nachfolgend sieht man den Inhalt einer solchen Infodatei. Im Dateinamen ist der Datensatzname vermerkt.

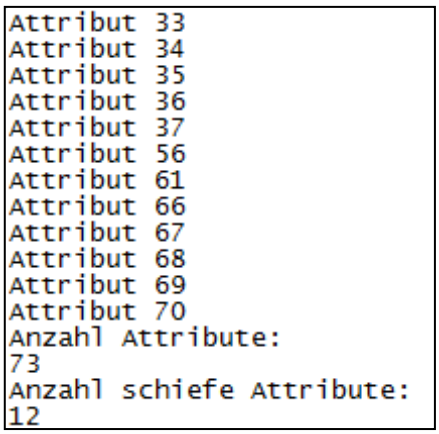


Abbildung 3: Analysedatei von onehr

2.2 Preprocessing Verfahren

2.2.1 Normalisierung

Die Normalisierung ist die erste und eine häufig verwendete Methode, um Datensätze vor dem Klassifizieren zu bearbeiten. Bei der Normalisierung werden die einzelnen Attribute auf Werte zwischen 0 und 1 transformiert. Es gibt durchaus andere Skalenwerte für das Intervall. Am häufigsten jedoch, und dies ist in dieser Arbeit ebenfalls der Fall, wird ein Intervall zwischen 0 und 1 gewählt. Die grösste Zahl wird dann die 1 und die Kleinste die 0. Alle Werte dazwischen werden proportional zu ihrem Ausgangswert Zahlen zwischen 0 und 1 annehmen. (Witten & Frank, 2005)
Durch die Normalisierung werden die Datensätze vergleichbarer und sehr grosse Zahlen haben nicht mehr eine so grosse Gewichtung bei der Berechnung des Abstandes.
Die Normalisierung eignet sich vor allem bei gleichverteilten oder sogar normalverteilten Daten. Bei Extremwerten ist das Verfahren nicht so geeignet, da diese auch im kleineren Intervall extrem sind. Bei der Normalisierung wird die Relation zwischen den Werten nicht verändert, nur die absolute Distanz wird kleiner.

Beispiel zur Normalisierung:

Datenreihe vor der Normalisierung:

$$[\mathbf{3}, \mathbf{5}, \mathbf{1}, \mathbf{6}, \mathbf{9}, \mathbf{9}, \mathbf{10}, \mathbf{15}, \mathbf{5}, \mathbf{9}, \mathbf{38}]$$

Datenreihe nach der Normalisierung:

$$[\ \mathbf{0.05}, \mathbf{0.10}\ , \mathbf{0}, \mathbf{0.14}, \mathbf{0.21}, \mathbf{0.21}, \mathbf{0.24}, \mathbf{0.38}, \mathbf{0.10}, \mathbf{0.21}, \mathbf{1}]$$

2.2.2 Ranking

Das Ranking, oder auch Rangordnung zu Deutsch, ist eine weitere Methode für die Bearbeitung der Attribute vor der Klassifizierung. Ranking funktioniert, indem alle Werte in eine Reihenfolge gebracht und diese dann so durchnummeriert werden. Falls zwei oder mehrere Werte gleich sind, bekommen sie als Wert den Durchschnitt der Ränge, welche sie in der Reihenfolge gehabt hätten. Somit verkürzen sich die einzelnen Abstände zwischen weit entfernten Werten. Das Kriterium ist nicht der absolute Wert einer Zahl, sondern die Rangierung. Deshalb ist der Abstand zwischen zwei Werten, welche zwei Einheiten auseinander liegen, aber direkt nacheinander folgen genau gleich gross, wie der Abstand zwischen zwei Werten welche 100 Einheiten auseinander liegen und ebenfalls in der Rangordnung hintereinander stehen.
Aus den vorher genannten Gründen macht das Rangieren der Werte vor allem bei schief verteilten Daten Sinn, denn es ist robust gegenüber Extremwerten und Nichtlinearitäten. (Kladroba, 2005)

Beispiel zum Ranking:

Datenreihe vor dem Ranking:

$$[\mathbf{3}, \mathbf{5}, \mathbf{1}, \mathbf{6}, \mathbf{9}, \mathbf{9}, \mathbf{10}, \mathbf{15}, \mathbf{5}, \mathbf{9}, \mathbf{38}]$$

Datenreihe in aufsteigender Reihenfolge:

$$[\ \mathbf{1}, \mathbf{3}, \mathbf{5}, \mathbf{5}, \mathbf{6}, \mathbf{9}, \mathbf{9}, \mathbf{9}, \mathbf{10}, \mathbf{15}, \mathbf{38}]$$

Datenreihe nach dem Ranking:

$$[\ \mathbf{2}, \mathbf{3.5}, \mathbf{1}, \mathbf{5}, \mathbf{7}, \mathbf{7}, \mathbf{9}, \mathbf{10}, \mathbf{3.5}, \mathbf{7}, \mathbf{11}]$$

2.3 kNN Verfahren

Das kNN Verfahren ist ein Klassifikationsverfahren. Um die Klasse der jeweiligen Instanzen zu bestimmen, berücksichtig der kNN Algorithmus die **k n**ächsten **N**achbarn (**k n**earest **n**eighbours). Dieses Verfahren gehört zur Kategorie „Lazy Learning", da das Lernen nur aus dem Abspeichern von Trainingsinstanzen besteht.
So werden zu einer Testinstanz die k nächsten Nachbarn gesucht. Als Anzahl der nächsten Nachbarn k wird üblicherweise die nächst höhere ungerade Zahl der Anzahl der verschiedenen Klassen gewählt. Dies wäre also bei vier verschiedenen Klassen k der Wert 5.
Sobald man die nächsten Nachbarn gefunden hat, wird die häufigste oder wahrscheinlichste Klasse der Nachbarn der untersuchten Testinstanz zugewiesen. Sind zum Beispiel drei der fünf Trainingsinstanzen in der Klasse 3, so wird auch die Testinstanz diese Klasse erhalten. (Witten & Frank, 2005)

2.3.1 Geometrischer Abstand

Der geometrische Abstand wird zwischen der Testinstanz und den jeweiligen Trainingsinstanzen berechnet. Um den geometrischen Abstand zu berechnen, nimmt man die Werte aller Dimensionen (Attribute) der Test- und der Trainingsinstanz. Mit der Formel für den Euklidischen Abstand kann der jeweilige Abstand zwischen den beiden Instanzen berechnet werden. Die Formel für den Euklidischen Abstand ist die folgende:

$$s = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

$s : Euklidischer\ Abstand$
$n : Anzahl\ Attribute$
$a, b : Werte\ der\ Attribute\ für\ die\ Test - und\ Trainingsinstanz$

Mit dem errechneten Abstand der beiden Instanzen kann nun bestimmt werden, wie ähnlich sich die beiden Instanzen sind. Für das kNN Verfahren werden nun die k nächsten

Nachbarn genommen. Je kleiner der geometrische Abstand zwischen den beiden Instanzen ist, desto näher sind sie sich.

Beispiel zum geometrischen Abstand:

In der untenstehenden Abbildung ist eine mögliche Situation für ein kNN Verfahren, das auf geometrischem Abstand basiert, dargestellt. Auf den beiden Achsen sind zwei Attribute abgebildet und es gibt zwei verschiedene Klassen: „+" und"-". Um nun eine Klassifikation für die neue Instanz (roter Punkt) vorzunehmen, betrachtet man die nächsten Nachbarn (umkreiste Instanzen). In diesem Beispiel sind die fünf nächsten Nachbarn entscheidend. Diese haben zweimal die Klasse „+" und dreimal die Klasse „-". Die neue Instanz wird also die Klasse „-" erhalten und ist so klassifiziert.

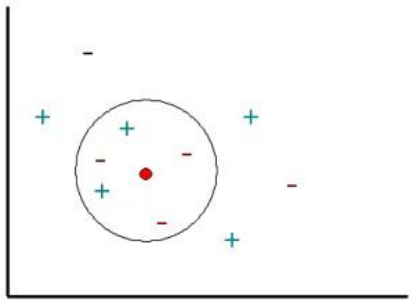


Abbildung 4: kNN Verfahren

2.3.2 Semantischer Abstand

Das Hyperrectangle Verfahren nutzt die semantische Ähnlichkeit. Bei diesem Verfahren werden die Attribute nicht verändert (kein Preprocessing), sondern sie werden bei der Klassifikation direkt angewendet. Für die Berechnung des semantischen Abstands werden in den vorliegenden und gespeicherten Trainingsinstanzen die ähnlichsten Instanzen gesucht und danach die Klassifizierung aufgrund dieser vorgenommen. Um die ähnlichsten Instanzen zu finden, sucht man nach den Trainingsinstanzen, welche die kleinste Anzahl anderer Trainingsinstanzen im aufgespannten Hyperrectangle haben. Das Hyperrectangle ist ein Vieleck, das sich zwischen der jeweiligen Test- und Trainingsinstanz aufspannt. Die Dimension entspricht der Anzahl Attribute. Hat man dieses Hyperrectangle gebildet, zählt man alle sich darin befindenden anderen Trainingsinstanzen. Falls man keine findet, sind sich die getestete Trainingsinstanz und die Testinstanz sehr ähnlich, im

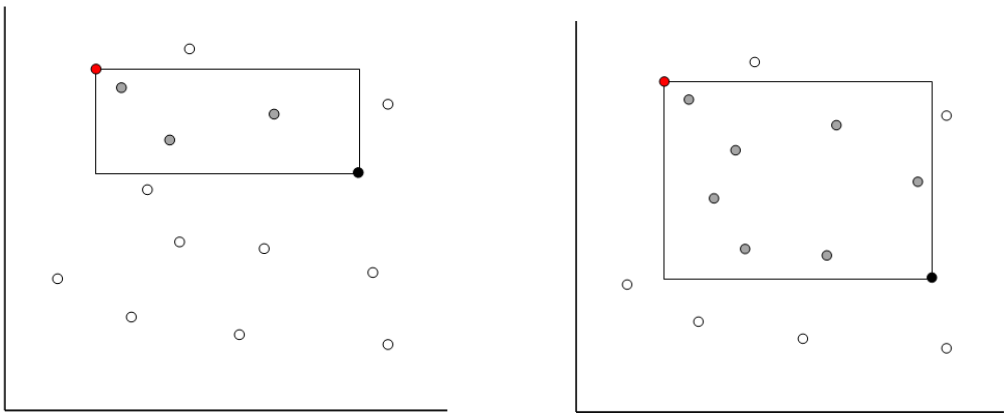
umgekehrten Fall mit vielen Trainingsinstanzen im Vieleck nicht. Im Programm zur Berechnung dieser Hyperrectangle werden jeweils die fünf ähnlichsten Trainingsinstanzen zur Klassifizierung verwendet. Findet das Programm fünf Trainingsinstanzen mit keiner anderen Trainingsinstanz im Hyperrectangle, dann wird abgebrochen und aus diesen die Klasse bestimmt. Dies wurde aus Gründen der Effizienz des Algorithmus so gewählt. (Gao, 2007)

Beispiel zum semantischen Abstand:

In den beiden untenstehenden Abbildungen sieht man zwei Beispiele für die Bestimmung der Ähnlichkeit mittels semantischem Abstand. Hier ist der Vergleich zwischen einer Testinstanz (roter Punkt) und zwei verschiedenen Trainingsinstanzen aufgezeigt. Die ganze Prozedur muss danach auf alle Trainingsinstanzen angewendet werden, da die k ähnlichsten (nächsten) Nachbarn für das kNN Verfahren gefunden werden sollen.

Für die Bestimmung des semantischen Abstandes wird nun ein Hyperrectangle, hier im zwei dimensionalen Fall ist es ein Rechteck, aufgespannt mit den beiden Instanzen als gegenüberliegende Eckpunkte. Danach werden alle eingeschlossenen Punkte in diesem Rechteck gezählt. Je weniger andere Instanzen sich in diesem Rechteck befinden, desto kleiner ist der semantische Abstand und desto ähnlicher sind sich die beiden getesteten Instanzen.

Für das kNN Verfahren werden die k Trainingsinstanzen mit dem kleinsten semantischen Abstand zur Testinstanz zur Bestimmung der Klasse genommen.



Der semantische Abstand ist 3

Der semantische Abstand ist 7

Abbildung 5: Beispiele für die Berechnung des semantischen Abstandes

2.4 10-Fold-Cross-Validation

Für die oben genannten Klassifizierungen durch die k nächsten Nachbarn braucht es eine Aufteilung der Datensätze in Trainings- und Testinstanzen. Üblicherweise wird dies durch eine Teilung der Datensätze in zehn gleichgrosse Teile erreicht, wobei immer ein Teil als Testinstanzen dient und die anderen neun Teile als Trainingsinstanzen. Danach wird das Ganze zehn Mal ausgeführt, bis alle Teile einmal als Testinstanz gewertet und klassifiziert wurden. Als grosser Vorteil wird die einmalige Klassifizierung aller Instanzen gesehen, was bei einer zufälligen Aufteilung des Datensatzes nicht zwingend gegeben ist. (Witten & Frank, 2005)

Im Programm wird die Aufteilung in die zehn Teile durch eine Hashfunktion erreicht. So kann garantiert werden, dass für alle Tests genau die gleichen Teile genommen wurden und die gleichen Voraussetzungen vorliegen.

Beispiel für eine 10-Fold-Cross-Validation:

Abbildung 6: 10-Fold-Cross-Validation Illustration

2.5 Evaluation

Die Evaluation erfolgt durch das Vergleichen der tatsächlichen, bereits im Datensatz vorhandenen Klasse mit dem errechneten Klassenwert aus der Ausführung der Algorithmen. Besteht eine Übereinstimmung der Klassen, wird es als korrekt taxiert, andernfalls nicht. Diese Überprüfung erfolgt intern in einer HashMap Datenstruktur und die Zusammenfassung der Resultate wird in eine Datei geschrieben und dem Benutzer

zugänglich gemacht. Diese Datei wird auf der lokalen Festplatte und auf einem entfernten FTP-Server gespeichert. In dieser Datei stehen die Anzahl richtig klassifizierter Instanzen, die Anzahl falsch klassifizierter Instanzen und der Prozentsatz der richtig klassifizierten Instanzen. Zur Veranschaulichung ist der Inhalt einer solchen Output Datei in der folgenden Abbildung dargestellt. Dies ist der Output vom ersten der zehn verschiedenen Folds für den Datensatz „magic04“.

```
Normalisation:
Right Calssified: 15708
Wrong Calssified: 5052
Share of Right Classified: 0.7566473988439306
Ranking:
Right Calssified: 16392
Wrong Calssified: 4368
Share of Right Classified: 0.7895953757225433
Hyper:
Right Calssified: 6576
Wrong Calssified: 14184
Share of Right Classified: 0.31676300578034683
```

Abbildung 7: Outputdatei von magic04 Fold 1

3. Ergebnisse

3.1 Interpretation

Nachfolgend werden nur die Zahlen von der Normalisierung und dem Ranking miteinander verglichen, denn das Hyperrectangle Verfahren wird im nachfolgenden Unterkapitel separat besprochen.

Die Ergebnistabellen sind in drei Sektionen unterteilt. In jeder Sektion gibt es zwei verschiedene Tabellen mit den gleichen Datensätzen. Der Unterschied liegt in der Definition der Verteilungsschiefe der Attribute. In der jeweils ersten Tabelle ist die Schiefe mit dem 3. zentralen Moment ausschlaggebend und in der zweiten Tabelle dann die Schiefe, welche auf den Wertebereich normalisiert ist.

In den ersten beiden Tabellen (Sektion 1) befinden sich die Datensätze, welche die Hypothese unterstützen, dass bei schief verteilten Attributen das Ranking als Preprocessing Methode besser abschneidet als die Normalisierung. In der dritten und vierten Tabelle (Sektion 2) sieht man die Datensätze, welche diese Hypothese vermeintlich nicht unterstützen. In diesen ersten vier Tabellen gibt es fett gedruckte Zahlen, was bedeutet, dass sie signifikant grösser ist als die nicht fette(n) Zahl(en) in derselben Zeile. Um zu entscheiden, ob das Ergebnis signifikant besser ist, wurde ein einseitiger T-Test durchgeführt. Die Grenze wurde bei 0.05, also 5%, angesetzt. Dies wird in der Wissenschaft häufig so verwendet. Ist der P-Wert also unter der 0.05 Grenze so wird dieser als signifikant besser angenommen und fett gedruckt. In den letzten beiden Tabellen (Sektion 3) sind nur noch Datensätze mit nicht signifikanten Grössenunterschieden in den beiden Durchläufen mit nur schief verteilten und ohne schief verteilte Attribute.

3.2 Ergebnistabellen mit Anteil an korrekter Klassifizierungen

Tabelle 1: Ergebnisse mit Schiefe mit 3. zentralem Moment, welche die Hypothese unterstützen

Datensatz		Anz. Attribute	Anz. Attr sn > T *	Normalisierung	Ranking	Hyperrectangle
Page-blocks	Mit allen Attributen	10	8	0.93356441 (0.00935314)	**0.95446386 (0.00797704)**	0.40428813 (0.06019444)
	Ohne schiefe Attribute	2	0	0.88605993 (0.01445809)	0.89573123 (0.00729693)	0.56797502 (0.02066543)
	Nur schiefe Attribute	8	8	0.92674783 (0.01122237)	**0.95094447 (0.00776751)**	0.20323377 (0.03733812)
diabetes	Mit allen Attributen	8	4	0.67383023 (0.04073624)	0.68225673 (0.07338201)	0.40488613 (0.04386316)
	Ohne schiefe Attribute	4	0	**0.643706 (0.04751516)**	0.59080745 (0.0361991)	0.59517598 (0.07229545)
	Nur schiefe Attribute	4	4	0.66383023 (0.06795273)	0.68811594 (0.03804134)	0.30763975 (0.04625376)
code-rna	Mit allen Attributen	8	2	0.82979115 (0.0781161)	**0.88905897 (0.00825515)**	0.59483955 (0.01304472)
	Ohne schiefe Attribute	6	0	0.54420824 (0.03599642)	0.50730784 (0.07583646)	0.58129995 (0.01736931)
	Nur schiefe Attribute	2	2	0.75200810 (0.05124471)	**0.84938597 (0.00873918)**	0.38550868 (0.01267964)
optdigits	Mit allen Attributen	64	5	**0.95252493 (0.0078678)**	0.93152887 (0.00716684)	0.0992431 (0.00979439)
	Ohne schiefe Attribute	59	0	**0.95367435 (0.00789899)**	0.93065852 (0.01436128)	0.0992431 (0.00979439)
	Nur schiefe Attribute	5	5	0.39674302 (0.03666587)	0.3912435 (0.02307643)	0.04402349 (0.01317152)

* Anzahl der Attribute mit einer Schiefe sn grösser als der Grenzwert T

Tabelle 2: Ergebnisse mit normalisierter Schiefe, welche die Hypothese unterstützen

Datensatz		Anz. Attribute	Anz. Attr sn > T	Normalisierung	Ranking
Page-blocks	Mit allen Attributen	10	9	**0.93356441 (0.00935314)**	0.95446386 (0.00797704)
	Ohne schiefe Attribute	1	0	**0.80153846 (0.01727602)**	0.70967033 (0.02062892)
	Nur schiefe Attribute	9	9	0.85318681 (0.08805727)	0.77934066 (0.15263152)
diabetes	Mit allen Attributen	8	4	0.84605634 (0.02328165)	**0.88884105 (0.01361717)**
	Ohne schiefe Attribute	4	0	0.66285714 (0.04957872)	0.64428571 (0.07452039)
	Nur schiefe Attribute	4	4	0.57714286 (0.05479295)	**0.64428571 (0.07452039)**
winequality-red	Mit allen Attributen	11	5	0.47597544 (0.03469041)	0.48144544 (0.02696235)
	Ohne schiefe Attribute	6	0	**0.4630137 (0.04579888)**	0.41712329 (0.03853649)
	Nur schiefe Attribute	5	5	0.41027397 (0.04089882)	0.39315068 (0.03877247)
eighthr	Mit allen Attributen	72	10	0.90833333 (0.02590027)	0.9077381 (0.01689423)
	Ohne schiefe Attribute	62	0	0.91964286 (0.02266599)	0.93035714 (0.01753456)
	Nur schiefe Attribute	10	10	0.89345238 (0.0233841)	**0.92857143 (0.01706811)**
transfusion	Mit allen Attributen	4	3	0.64852941 (0.09389526)	0.67647059 (0.03733222)
	Ohne schiefe Attribute	1	0	**0.70588235 (0.06756887)**	0.63235294 (0.04650408)
	Nur schiefe Attribute	3	3	0.71911765 (0.05115449)	0.76176471 (0.07028803)

Tabelle 3: Ergebnisse mit Schiefe mit 3. zentralem Moment, welche die Hypothese vermeintlich verwerfen

Datensatz		Anz. Attribute	Anz. Attr sn > T	Normalisierung	Ranking	Hyperrectangle
Spambase	Mit allen Attributen	57	4	0.84605634 (0.02328165)	**0.88884105 (0.01361717)**	0.39421498 (0.0007932)
	Ohne schiefe Attribute	53	0	0.84892145 (0.01982534)	**0.87878978 (0.01412725)**	0.39421498 (0.0007932)
	Nur schiefe Attribute	4	4	0.69759281 (0.03111444)	**0.74875872 (0.0241412)**	0.51350618 (0.01446524)
magic04	Mit allen Attributen	10	7	0.75449643 (0.01161589)	**0.78717238 (0.00658236)**	0.30981395 (0.00854473)
	Ohne schiefe Attribute	3	0	0.61263084 (0.02007937)	**0.62928707 (0.00789353)**	0.48487719 (0.01184829)
	Nur schiefe Attribute	7	7	0.7333872 (0.01487064)	**0.7697058 (0.01131085)**	0.306170763 (0.00750862)
winequality-red	Mit allen Attributen	11	2	0.47597544 (0.03469041)	0.48144544 (0.02696235)	0.41748701 (0.05991665)
	Ohne schiefe Attribute	9	0	0.46489844 (0.04128417)	**0.50198866 (0.03363917)**	0.35968824 (0.02923749)
	Nur schiefe Attribute	2	2	0.39205479 (0.03840664)	0.39820501 (0.01510689)	0.3741521 (0.03821974)
winequality-white	Mit allen Attributen	11	3	0.42645438 (0.02300937)	**0.47877513 (0.01957396)**	0.35056281 (0.02222619)
	Ohne schiefe Attribute	8	0	0.41185116 (0.02474974)	**0.47204162 (0.02401064)**	0.31394669 (0.01386178)
	Nur schiefe Attribute	3	3	0.32314405 (0.02418722)	**0.3813327 (0.03844246)**	0.35795536 (0.04552151)

Tabelle 4: Ergebnisse mit normalisierter Schiefe, welche Hypothese vermeintlich verwerfen

Datensatz		Anz. Attribute	Anz. Attr sn > T	Normalisierung	Ranking
optdigits	Mit allen Attributen	64	33	**0.95252493 (0.0078678)**	0.93152887 (0.00716684)
	Ohne schiefe Attribute	31	0	**0.98275862 (0.00620761)**	0.10689655 (0.0192431)
	Nur schiefe Attribute	33	33	**0.75287356 (0.01953647)**	0.10431034 (0.0101415)
magic04	Mit allen Attributen	10	4	0.75449643 (0.01161589)	**0.78717238 (0.00658236)**
	Ohne schiefe Attribute	6	0	**0.72815029 (0.03722636)**	0.62595376 (0.08634928)
	Nur schiefe Attribute	4	4	**0.55184971 (0.04621471)**	0.35196532 (0.00018279)
winequality-white	Mit allen Attributen	11	5	0.42645438 (0.02300937)	**0.47877513 (0.01957396)**
	Ohne schiefe Attribute	6	0	**0.38542601 (0.02773289)**	0.13654709 (0.07667484)
	Nur schiefe Attribute	5	5	**0.35493274 (0.0305798)**	0.05964126 (0.04985898)
onehr	Mit allen Attributen	72	10	0.9422619 (0.01945052)	0.95119048 (0.0183572)
	Ohne schiefe Attribute	62	0	0.95 (0.01734268)	**0.96904762 (0.01183844)**
	Nur schiefe Attribute	10	10	0.94821429 (0.01013654)	**0.96845238 (0.01192128)**
heart	Mit allen Attributen	13	4	0.728 (0.1041894)	0.73183333 (0.07191151)
	Ohne schiefe Attribute	9	0	**0.74 (0.09092121)**	0.54 (0.11352924)
	Nur schiefe Attribute	4	4	**0.62 (0.08485281)**	0.412 (0.08854377)
wine	Mit allen Attributen	13	2	0.94375 (0.05472469)	0.93161765 (0.06220374)
	Ohne schiefe Attribute	11	0	**0.97058824 (0.0571662)**	0.42352941 (0.07744478)
	Nur schiefe Attribute	2	2	**0.56470588 (0.0968556)**	0.42941176 (0.09625833)
german nummer	Mit allen Attributen	24	15	0.63296703 (0.04892544)	0.63736264 (0.04775972)
	Ohne schiefe Attribute	9	0	**0.68241758 (0.04222793)**	0.33516484 (0.12870129)
	Nur schiefe Attribute	15	15	**0.63076923 (0.03917873)**	0.57362637 (0.05856225)

Tabelle 5: Ergebnisse mit Schiefe mit 3. zentralen Moment, welche keine Signifikanz aufweisen

Datensatz		Anz. Attribute	Anz. Attr sn > T	Normalisierung	Ranking	Hyperrectangle
image-segmentation	Mit allen Attributen	19	12	0.14345550 (0.00505807)	0.13874346 (0.00508809)	0.14345550 (0.00505807)
	Ohne schiefe Attribute	7	0	0.14345550 (0.00505807)	0.14293194 (0.00856750)	0.14293194 (0.00496693)
	Nur schiefe Attribute	12	12	0.14043222 (0.02241071)	0.14188482 (0.00717445)	0.01361257 (0.00505807)
eighthr	Mit allen Attributen	72	12	0.90833333 (0.02590027)	0.90773810 (0.01689423)	0.93035714 (0.01753456)
	Ohne schiefe Attribute	60	0	0.90416667 (0.02727002)	0.90773810 (0.01802173)	0.93035714 (0.01753456)
	Nur schiefe Attribute	12	12	0.89345238 (0.02388381)	0.87797619 (0.02864986)	0.92976190 (0.01702191)
onehr	Mit allen Attributen	72	12	0.94226190 (0.01945052)	0.95119048 (0.0183572)	0.96904762 (0.01183844)
	Ohne schiefe Attribute	60	0	0.95300789 (0.0135714)	0.95420189 (0.01083685)	0.96966399 (0.0113698)
	Nur schiefe Attribute	12	12	0.93273810 (0.0231472)	0.93750000 (0.01780194)	0.96904762 (0.01183844)
transfusion	Mit allen Attributen	4	3	0.64852941 (0.09389526)	0.67647059 (0.03733222)	0.36029412 (0.10722864)
	Ohne schiefe Attribute	1	0	0.62058824 (0.09321462)	0.66029412 (0.03697651)	0.42941176 (0.05270463)
	Nur schiefe Attribute	3	3	0.65588235 (0.08524339)	0.67058824 (0.04112976)	0.36029412 (0.10722864)
heart	Mit allen Attributen	13	3	0.728 (0.1041894)	0.73183333 (0.07191151)	0.54116667 (0.1215437)
	Ohne schiefe Attribute	10	0	0.732 (0.09938639)	0.74066667 (0.0798579)	0.533 (0.11556036)
	Nur schiefe Attribute	3	3	0.57633333 (0.09580974)	0.58933333 (0.09630904)	0.4305 (0.09717634)
wine	Mit allen Attributen	13	2	0.94375000 (0.05472469)	0.93161765 (0.06220374)	0.09191176 (0.06551503)
	Ohne schiefe Attribute	11	0	0.92610294 (0.04863359)	0.89522059 (0.07725936)	0.09227941 (0.08168515)
	Nur schiefe Attribute	2	2	0.75294118 (0.08115669)	0.72242647 (0.10633202)	0.07389706 (0.04863359)
bupa	Mit allen Attributen	6	4	0.56028226 (0.10256178)	0.58649194 (0.10307243)	0.46159274 (0.06696606)
	Ohne schiefe Attribute	2	0	0.51612903 (0.10224445)	0.58266129 (0.09694322)	0.42308468 (0.10570112)
	Nur schiefe Attribute	4	4	0.54445565 (0.11192033)	0.56078629 (0.09089248)	0.38064516 (0.12364697)
german nummer	Mit allen Attributen	24	3	0.63296703 (0.04892544)	0.63736264 (0.04775972)	0.7010989 (0.04265473)
	Ohne schiefe Attribute	21	0	0.60769231 (0.04749208)	0.58791209 (0.04152302)	0.7010989 (0.04265473)
	Nur schiefe Attribute	3	3	0.60549451 (0.0415876)	0.5978022 (0.03243362)	0.59450549 (0.04588269)

Tabelle 6: Ergebnisse mit normalisierter Schiefe, welche keine Signifikanz aufweisen

Datensatz		Anz. Attribute	Anz. Attr sn > T	Normalisierung	Ranking
code-rna	Mit allen Attributen	8	1	0.82979115 (0.0781161)	**0.88905897 (0.00825515)**
	Ohne schiefe Attribute	7	0	0.61298263 (0.15483716)	0.56891417 (0.06222131)
	Nur schiefe Attribute	1	1	0.33336891 (0.00051552)	0.33361793 (0.00055112)
image-segmentation	Mit allen Attributen	19	12	0.1434555 (0.00505807)	0.13874346 (0.00508809)
	Ohne schiefe Attribute	7	0	0.1434555 (0.00505807)	0.14031414 (0.01455962)
	Nur schiefe Attribute	12	12	0.14132497 (0.00125058)	0.14188482 (0.00717445)
bupa	Mit allen Attributen	6	4	0.56028226 (0.10256178)	0.58649194 (0.10307243)
	Ohne schiefe Attribute	2	0	0.540625 (0.09438443)	0.48125 (0.09793883)
	Nur schiefe Attribute	4	4	0.6 (0.07186745)	0.5 (0.13176157)

3.2.1 Vergleich der Ergebnisse mit den beiden Massen für Schiefen

Die Datensätze sind sowohl mit der normalisierten Schiefe als auch mit der Schiefe, die als Mass das 3. zentrale Moment nimmt, ausgewertet. Erwartungsgemäss liefern die beiden verschiedenen Masse für die Schiefe unterschiedliche Ergebnisse.

Die beiden ersten Tabellen (Sektion 1) zeigen die Datensätze, welche die Hypothese, dass das Ranking als Preprocessing Methode gegenüber der Normalisierung als Preprocessing Methode vor dem Klassifizieren mit dem kNN Verfahren bei schief verteilten Attributen besser geeignet ist, unterstützen.

Der Datensatz „code-rna" ist im ersten Fall mit der nicht normalisierten Schiefe Hypothesen unterstützend. Im zweiten Fall, mit der normalisierten Schiefe, ist nur noch ein Attribut schief verteilt. Der Durchlauf mit den nur schiefen Attributen, welcher im ersten Fall noch signifikant bessere Resultate für das Ranking aufwies, ist nun nicht mehr signifikant und deutlich schlechter. Hier liegt das Problem bei der Anzahl der Attribute, die den Unterschied ausmachen. Der gleiche Umstand tritt im nächstgenannten Datensatz „optdigits" auf. Dieser ist ebenfalls ohne normalisierte Schiefe signifikant hypothesenunterstützend und im anderen Falle nicht. Die Anzahl der schief verteilten Attribute variiert auch hier sehr stark. So lässt sich auch dieser Unterschied erklären.

In den Tabellen 3 und 4 (Sektion 2) sind die Datensätze, welche die Hypothese vermeintlich verwerfen. Mit der nicht normalisierten Schiefe sind dies vier Datensätze und mit der normalisierten Schiefe sind es sieben. Dieser doch recht grosse Unterschied ist wiederum auf die unterschiedliche Bestimmung der Verteilungsschiefe der einzelnen Attribute zurückzuführen. So sind die Datensätze nicht mehr dieselben und liefern auch nicht die gleichen Ergebnisse. Der Datensatz „spambase" ist nur in der Schiefe mit dem 3. zentralen Moment vorhanden, da er in der Betrachtung mit der normalisierten Schiefe nur schief verteilte Attribute aufweist.

Die Hypothese, dass das Ranking besser auf schief verteilten Attributen funktioniert als dies die Normalisierung macht, wird hier nicht bestätigt. In fast allen Fällen ist entweder das Ranking bei beiden Datensatzsplits (nur schiefe und ohne schiefe Attribute) besser oder dann ist die Normalisierung bei beiden besser.

Das Abschneiden der beiden Preprocessing Varianten ist nicht nur von der Verteilung der Attribute beeinflusst. Ein weiterer Faktor, der die Ergebnisse der beiden Verfahren tangiert,

ist der Wertebereich der einzelnen Attribute. Es macht einen grossen Unterschied, wenn die Werte der Attribute eng beieinander auf einem kleinen Intervall liegen oder sie auf einem grossen Intervall sind und auch grössere Ausreisser besitzen. Denn die Stärke vom Ranking liegt in der Behandlung von Ausreissern. Es ist resistent gegen Ausreisser im Gegenteil zur Normalisierung, welche weniger gut geeignet ist, wenn Ausreisser vorhanden sind.

3.2.2 Normalisierung im Vergleich mit Ranking

Die zweite Hypothese besagt, dass das Ranking besser funktioniert als die Normalisierung/Standardisierung, falls die Attribute schief verteilt sind. Um diese Hypothese zu bestätigen oder sie zu verwerfen, muss man die Ergebnisse genau betrachten und alle beeinflussenden Umstände in Betracht ziehen.

Es gibt Datensätze, welche die Hypothese unterstützen, solche die sie weder bestätigen noch verwerfen und solche die sie vermeintlicher Weise verwerfen. Einfluss auf die Verfahren haben unter anderem der Anteil der schiefen Attribute der Datensätze, die Anzahl verschiedener Klassen, die Anzahl der Instanzen in den jeweiligen Datensätzen und der Einfluss der einzelnen Attribute. Unter dem Einfluss der Attribute ist die Aussagekraft von Attributen gemeint. Es gibt Attribute welche eine klare Unterscheidung zwischen den einzelnen Klassen haben und andere, bei welchen man die Klassenzugehörigkeit nur schwer bestimmen kann. Falls nun die entscheidenden Attribute eines Datensatzes normalverteilt sind und die andern die Klassifikation nicht wesentlich beeinflussen, ist die Normalisierung tendenziell im Vorteil gegenüber dem Ranking. Die vorhergehende Normalisierung ist für solche Verteilungen besser geeignet als das Ranking, was auch die Mehrheit der Datensätze in den Ergebnistabellen bestätigt. Das Ranking hat den umgekehrten Vorteil und ist das günstigere Verfahren, wenn die schief verteilten Attribute ein stärkeres Gewicht auf die Klassifizierung haben.

Ein weiteres Kriterium für die Abwägung der Ergebnisse ist die Aufteilung des Datensatzes in einen Teil mit nur nicht schief verteilten Attributen und einen anderen Teil mit nur schief verteilten Attributen. Möglicherweise sind nur nicht massgebende Attribute im einen oder anderen Teil vorhanden und verschlechtern so die Ergebnisse erheblich.

Bei allen Datensätzen in der ersten Sektion (Tabelle 1 und 2) ist ersichtlich, dass das Ranking für den dritten Durchlauf mit den nur schief verteilten Attributen besser war oder die Normalisierung beim zweiten Durchlauf ohne die schief verteilten Attribute besser war. Mindestens einer dieser beiden Fakten ist signifikant. Beim Durchlauf mit allen Attributen ist es vom Datensatz abhängig, welche Methode das bessere Ergebnis liefert.
In der zweiten Sektion (Tabelle 3 und 4) sind die verwerfenden Datensätze zu sehen. Beim ersten Datensatz ist die Normalisierung entscheidend besser als das Ranking beim Versuch mit nur den schief verteilten Attributen. Dieses Ergebnis sieht ein wenig extrem aus. Dies lässt sich an der Struktur des Datensatzes erklären. Alle nicht schief verteilten Attribute sind in einem Intervall mit einem Bereich von 0-29.22. Diese Attribute sind nicht entscheidend für die Klassifizierung und werden nicht stark gewichtet. Die anderen Werte der Attribute sind über einen sehr grossen Bereich verteilt. Die Normalisierung funktioniert viel besser. Die Attribute sind nicht so schief verteilt, wie man auf den Histogrammen erkennt (siehe Anhang). Deshalb ist dieser Datensatz mit Vorsicht zu geniessen.
Beim zweiten und dritten Datensatz tritt ein ähnliches Problem auf. Die nicht schief verteilten Attribute sind ebenfalls ein wenig schief. Darum auch das bessere Abschneiden des Rankings. Da der Bereich für die Werte aber sehr klein ist, wird das Attribut nicht als schief gewertet. Wenn man also dies im Hinterkopf hat, kann man sagen, dass auch diese beiden Datensätze die Hypothese unterstützen.
In der letzten Sektion (Tabelle 5 und 6) sind alle Ergebnisse, die keinen signifikanten Unterschied aufweisen, aufgeführt. Die meisten von ihnen weisen aber eine klare Tendenz in Richtung der vorher besprochenen Hypothese auf.

3.2.3 Hyperrectangle im Vergleich mit Ranking

Des Weiteren überprüfen wir die Hypothese, dass sich Ranking mit anschliessendem kNN Verfahren und das Hyperrectangle Verfahren vergleichbar zueinander verhalten. Diese Verfahren sollten bei schief verteilten Attributen besser abschneiden als bei der Verwendung von eher symmetrisch verteilten Attributen.
Um dies zu überprüfen, betrachtet man die beiden letzten Spalten mit den Ergebnissen zum Ranking und dem Hyperrectangle Verfahren aus den jeweils ersten Tabellen pro

Sektion (Tabelle 1, 3 und 5) mit der Schiefe, die mit dem 3. zentralen Moment gerechnet wurde.

Es fällt auf, dass das Hyperrectangle Verfahren viel schlechter abschneidet als das Ranking. In fast allen überprüften Datensätzen ist die Richtigkeit der Klassifikation mit dem geometrischen Abstand viel höher als beim Hyperrectangle Verfahren. Daraus lässt sich schliessen, dass das Hyperrectangle Verfahren sich nicht gut für eine Klassifikation eignet. Der Grund ist folgender: Beim Ranking wird der durchschnittliche Abstand zu allen Attributen für die Klassifikation hergenommen. Dem gegenüber verwendet das Hyperrectangle Verfahren den minimalen Abstand zu allen Attributen. So hat jede Instanz mindestens zwei benachbarte Instanzen mit dem Abstand 0. Falls nun eine ebenfalls sehr ähnliche Instanz in nicht allen Attributen ähnlicher ist, wird diese nicht gezählt und so eine falsche Unähnlichkeit vorgetäuscht. Die Betrachtung von allen Attributen als Einheit, weist sich bei diesem Verfahren als Problem aus. Deshalb ist das Hyperrectangle Verfahren in dieser Art nicht für die Klassifikation zu gebrauchen und es hat kein ähnliches Verhalten wie dies das Ranking aufweist. Um diesen Sachverhalt zu veranschaulichen, betrachtet man die folgenden zwei Abbildungen.

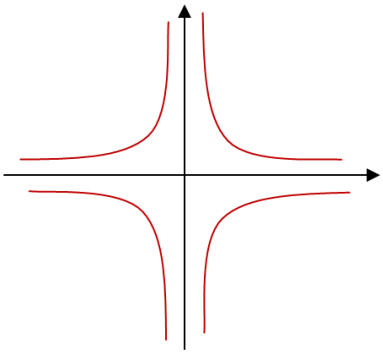


Abbildung 9: Semantischer Abstand

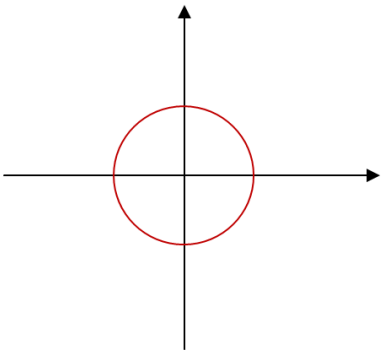


Abbildung 8: Geometrischer Abstand

In beiden Abbildungen 9 und 10 stehen die Achsen für je ein Attribut. Auf der linken Seite ist der semantische Abstand abgebildet, welcher im Hyperrectangle Verfahren Verwendung findet. Auf der rechten Seite ist der geometrische Abstand, der im anderen kNN Verfahren benutzt wird. Die roten Linien sind im linken Bild alle Instanzen mit dem Abstand 0 zur gegebenen Instanz (im Nullpunkt). Im Bild auf der rechten Seite sind alle Instanzen mit dem gleichen Abstand zur gegebenen Instanz auf dem Kreis.

4. Fazit

In dieser Arbeit wurde die Hypothese getestet, dass bei Datensätzen mit schief verteilten Attributen der semantische Abstand mit dem Hyperrectangle Verfahren sich ähnlich verhält wie der geometrische Abstand mit dem kNN Verfahren nach Ranking als Preprocessing Methode.

Diese Hypothese musste verworfen werden. Das Verfahren, welches den semantischen Abstand mittels Hyperrectangles verwendet, stellte sich als nicht tauglich für diese Art von Klassifizierung heraus. Das Hauptproblem ist die Verwendung des Abstandes, welcher nicht der durchschnittliche Abstand aller Attribute ist, sondern sich aus der Anzahl sich im Hyperrectangle befindenden anderen Instanzen berechnet.

Des Weiteren wurde die Hypothese getestet, dass das Ranking als Preprocessing Methode gegenüber der Normalisierung als Preprocessing Methode vor dem Klassifizieren mit dem kNN Verfahren bei schief verteilten Attributen besser geeignet ist.

Im Gegensatz zur ersten Hypothese kann diese Hypothese nicht verworfen werden. Keiner der getesteten 16 Datensätze verwarf die Hypothese, wenn man die Attribute und Einflüsse genau betrachtet. Die meisten unterstützen die Hypothese. So kann man sagen, dass sich das Ranking bei schief verteilten Attributen besser als Preprocessing Variante eignet als die Normalisierung. Das anschliessende kNN Verfahren liefert mit dem vorhergegangenen Ranking die besseren Ergebnisse.

Bei nicht schief verteilten Attributen ist die Normalisierung dem Ranking vorzuziehen.

5. Literaturverzeichnis

Gao, B. J. (2007). *Hyper-Rectangle-Based Discrimiative Data Generalization And Applications In Data Mining.* Burnaby (CAN): Simon Frazer University.

Kladroba, A. (2005). *Statistische Methoden zur Erstellung und Interpretraion von Rankings und Ratings.* Berlin: Verlag für Wissenschaft und Forschung.

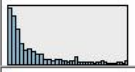
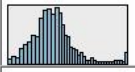
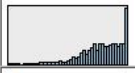
Schira, J. (2009). Statistische Methoden der VWL und BWL. München: Pearson Studium.

Witten, I. H., & Frank, E. (2005). *Data Mining, Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufmann Publishers.

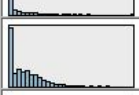
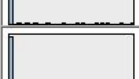
A Anhang

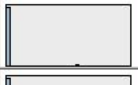
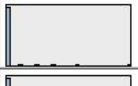
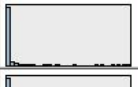
A.1 Data Audits

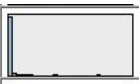
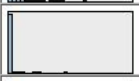
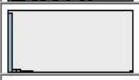
Data Audit von page-blocks

	Feld	Diagramm Stichpr.	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	height		803	10.570	19.474	20.404	752981005.5
2	lenght		546	88.416	114.703	2.145	16171807680.9
3	area		143986	1211.670	5031.259	19.086	12146345342838954
4	eccen		536.993	12.884	29.111	7.387	910543600.4
5	p_black		0.948	0.368	0.171	1.598	39.8
6	p_and		0.938	0.786	0.170	-0.858	-21.0
7	mean_tr		536.000	4.928	15.926	16.553	334102965.8
8	blackpix		33010	372.450	1313.976	13.309	150879960308707.8
9	blackand		46126	746.331	1940.331	11.071	4041354128439052
10	wb_trans		3211	106.785	168.977	5.693	137248191320.1

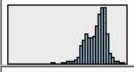
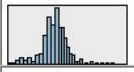
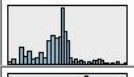
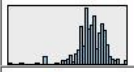
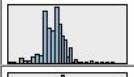
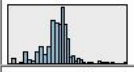
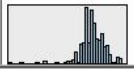
Data Audit von spambase

	Feld	Diagramm Stichpr.	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	word_freq_make		4.540	0.105	0.305	5.676	743.0
2	word_freq_address		14.280	0.213	1.291	10.087	99694.8
3	word_freq_all		5.100	0.281	0.504	3.009	1772.9
4	word_freq_3d		42	0.062	1.373	26.344	327486.9
5	word_freq_our		10.000	0.312	0.673	4.747	6639.0
6	word_freq_over		5.880	0.096	0.274	5.957	562.4
7	word_freq_remove		7.270	0.114	0.391	6.766	1865.8
8	word_freq_internet		11.110	0.105	0.401	9.725	2884.8
9	word_freq_order		5.260	0.090	0.279	5.226	519.7
10	word_freq_mail		18.180	0.239	0.645	8.488	10460.4
11	word_freq_receive		2.610	0.060	0.202	5.510	207.4
12	word_freq_will		9.670	0.542	0.862	2.867	8435.6
13	word_freq_people		5.550	0.094	0.301	6.956	872.5
14	word_freq_report		10.000	0.059	0.335	11.755	2035.2
15	word_freq_addresses		4.410	0.049	0.259	6.971	555.9
16	word_freq_free		20.000	0.249	0.826	10.764	27870.1
17	word_freq_business		7.140	0.143	0.444	5.689	2290.3
18	word_freq_email		9.090	0.185	0.531	5.414	3729.5
19	word_freq_you		18.750	1.662	1.775	1.592	40961.0
20	word_freq_credit		18.180	0.086	0.510	14.603	8894.3
21	word_freq_your		11.110	0.810	1.201	2.436	19390.3
22	word_freq_font		17	0.109	0.974	10.308	49503.1
23	word_freq_000		5.450	0.102	0.350	5.714	1129.2
24	word_freq_money		12.500	0.094	0.443	14.687	5856.5

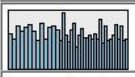
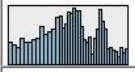
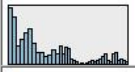
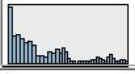
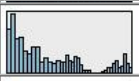
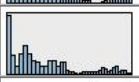
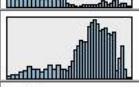
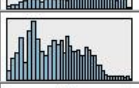
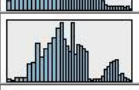
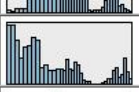
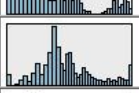
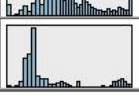
25	word_freq_hp		20.830	0.550	1.671	5.717	122723.3
26	word_freq_hpl		16	0.181	0.786	7.641	20372.7
27	word_freq_george		33	0.705	3.343	5.852	1008470.0
28	word_freq_650		9.090	0.125	0.539	6.607	4745.5
29	word_freq_lab		14	0.064	0.538	13.446	10919.9
30	word_freq_labs		5	0.058	0.377	7.987	2906.1
31	word_freq_telnet		12.500	0.065	0.403	12.669	3823.8
32	word_freq_857		4	0.026	0.276	12.230	1720.4
33	word_freq_data		18.180	0.097	0.556	13.190	10418.9
34	word_freq_415		4	0.027	0.276	12.188	1722.2
35	word_freq_85		20.000	0.105	0.532	15.231	10560.0
36	word_freq_technology		7.690	0.097	0.403	7.673	2302.8
37	word_freq_1999		6.890	0.137	0.423	5.323	1858.6
38	word_freq_parts		8	0.008	0.201	31.466	1396.1
39	word_freq_pm		11	0.043	0.372	15.706	4552.9
40	word_freq_direct		4.760	0.065	0.350	9.147	1801.9
41	word_freq_cs		7	0.030	0.317	14.932	2727.6
42	word_freq_meeting		14	0.100	0.706	10.760	19604.0
43	word_freq_original		3.570	0.046	0.224	7.629	393.3
44	word_freq_project		20.000	0.079	0.622	18.772	20767.7
45	word_freq_re		21.420	0.301	1.012	9.146	43545.5
46	word_freq_edu		22.050	0.180	0.911	10.123	35203.8
47	word_freq_table		2	0.002	0.051	32.600	40.5
48	word_freq_conference		10	0.017	0.251	25.510	2115.3
49	word_freq_;		4.385	0.039	0.243	13.709	909.7
50	word_freq_(		9.752	0.139	0.270	13.584	1234.2

51	word_freq_[		4.081	0.017	0.109	21.084	126.9
52	word_freq_!		32.478	0.269	0.816	18.658	46556.4
53	word_freq_$		6.003	0.076	0.246	11.163	763.0
54	word_freq_#		19.829	0.044	0.429	31.062	11303.4
55	capital_run_length_average		1101.500	5.192	31.729	23.762	3490102426.2
56	capital_run_length_longest		9988	52.173	194.891	30.765	1047136369383.4
57	capital_run_length_total		15840	283.289	606.348	8.710	8927816011338.8

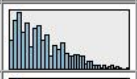
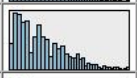
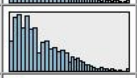
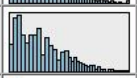
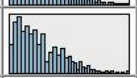
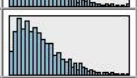
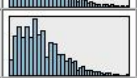
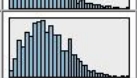
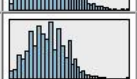
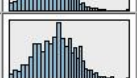
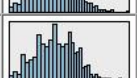
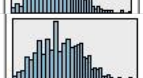
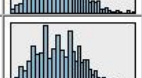
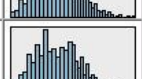
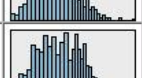
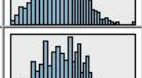
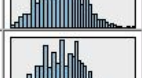
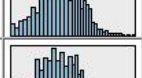
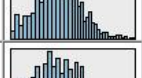
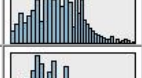
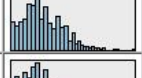
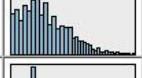
Data Audit von code-rna

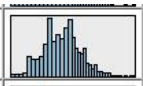
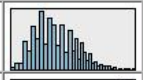
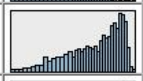
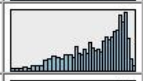
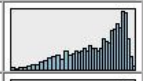
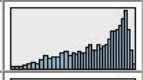
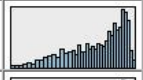
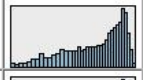
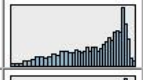
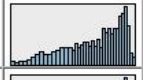
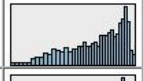
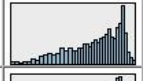
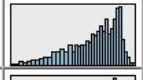
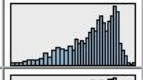
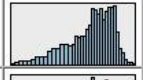
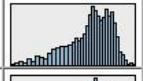
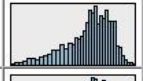
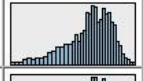
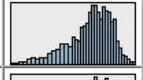
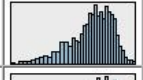
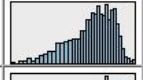
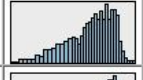
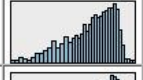
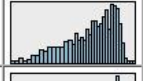
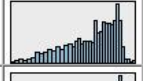
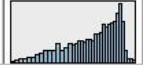
	Feld	Diagramm Stichpr.	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	deltaG_total value		1844	-577.936	164.134	-0.644	-87966643859.5
2	lenght of schroter sequence		66	116.385	8.904	-3.987	-87010209.3
3	'A' frequencies of seq 1		0.374	0.226	0.034	0.486	0.6
4	'U' frequencies of seq 1		0.313	0.214	0.045	0.703	2.0
5	'C' frequencies of seq 1		0.278	0.265	0.034	-0.680	-0.8
6	'A' frequencies of seq 2		0.374	0.222	0.030	0.840	0.7
7	'U' frequencies of seq 2		0.313	0.212	0.029	-0.123	-0.1
8	'C' frequencies of seq 2		0.278	0.270	0.028	-0.112	-0.1

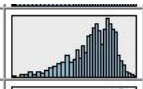
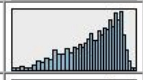
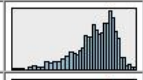
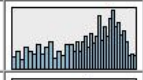
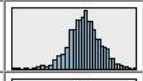
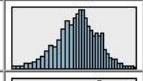
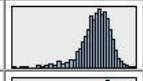
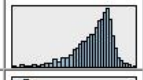
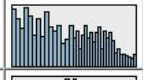
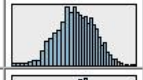
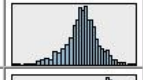
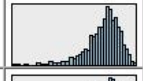
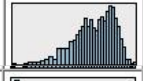
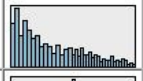
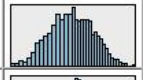
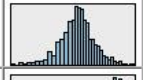
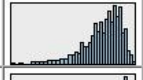
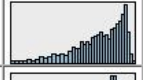
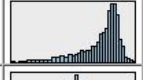
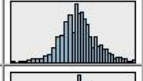
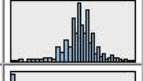
Data Audit von image-segmentation

	Feld	Diagramm Stichpr.	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	region-centroid-col		253	124.940	72.859	0.058	47351370.4
2	region-centroid-row		240	123.483	57.431	0.111	44083407.1
3	region-pixel-count		0	9.000	0.000	--	0.0
4	short-line-density-5		0.333	0.015	0.041	2.782	0.4
5	short-line-density-2		0.222	0.005	0.024	5.505	0.2
6	vedge-mean		29.222	1.891	2.649	5.532	215764.0
7	vedge-sd		991.718	5.708	44.989	14.443	2757947571.2
8	hedge-mean		44.722	2.407	3.470	5.140	450303.4
9	hedge-sd		1386.329	7.904	53.471	16.560	5309127119.2
10	intensity-mean		143.444	37.048	38.135	1.275	148250010.0
11	rawred-mean		137.111	32.807	34.995	1.332	119693029.5
12	rawblue-mean		150.889	44.206	43.510	1.125	194333169.8
13	rawgreen-mean		142.556	34.131	36.304	1.381	138529011.5
14	exred-mean		59.556	-12.723	11.588	-0.884	-2885048.5
15	exblue-mean		94.444	21.474	19.654	0.419	6672874.6
16	exgreen-mean		58.556	-8.751	11.607	0.783	2566296.7
17	value-mean		150.889	45.162	42.901	1.135	187867610.8
18	saturation-mean		1.000	0.427	0.228	0.946	23.7
19	hue-mean		5.957	-1.365	1.544	1.786	13789.8

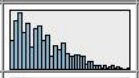
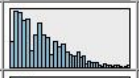
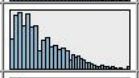
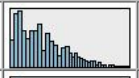
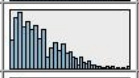
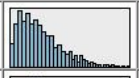
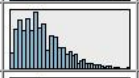
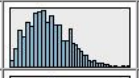
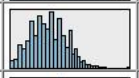
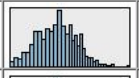
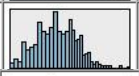
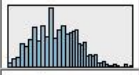
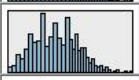
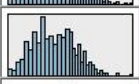
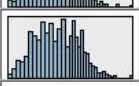
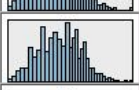
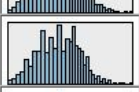
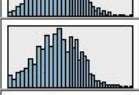
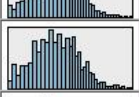
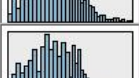
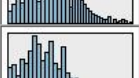
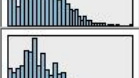
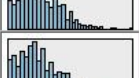
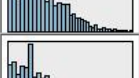
Data Audit von eighthr

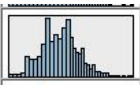
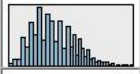
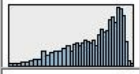
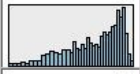
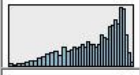
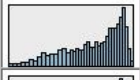
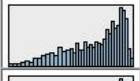
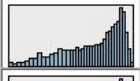
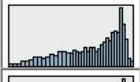
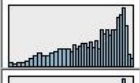
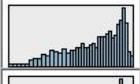
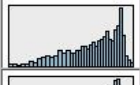
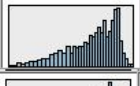
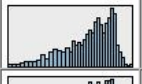
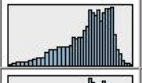
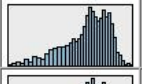
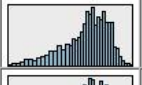
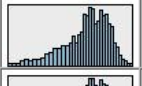
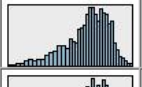
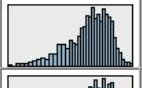
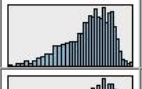
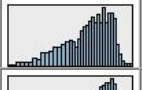
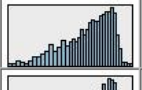
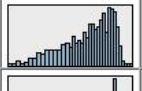
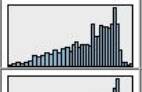
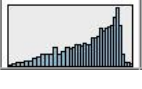
	Feld	Diagramm	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	WSR0		6.900	1.629	1.253	1.006	3649.9
2	WSR1		6.900	1.570	1.244	1.119	3970.4
3	WSR2		7.100	1.537	1.219	1.141	3809.6
4	WSR3		6.700	1.520	1.195	1.114	3504.4
5	WSR4		7.200	1.515	1.188	1.169	3610.2
6	WSR5		7.400	1.534	1.159	1.203	3450.9
7	WSR6		7.200	1.632	1.138	1.098	2981.6
8	WSR7		7.500	2.038	1.154	0.738	2094.0
9	WSR8		9.100	2.526	1.172	0.485	1441.8
10	WSR9		8.200	2.843	1.208	0.303	986.7
11	WSR10		8.600	2.982	1.291	0.308	1222.6
12	WSR11		8.700	3.030	1.379	0.369	1781.4
13	WSR12		8.900	3.057	1.410	0.387	2001.4
14	WSR13		9.500	3.114	1.431	0.381	2059.6
15	WSR14		9.000	3.185	1.420	0.307	1617.5
16	WSR15		8.900	3.235	1.372	0.257	1226.2
17	WSR16		8.500	3.202	1.274	0.217	826.2
18	WSR17		7.900	2.944	1.236	0.231	804.7
19	WSR18		7.400	2.582	1.243	0.386	1366.0
20	WSR19		7.100	2.301	1.230	0.517	1776.3
21	WSR20		8.700	2.105	1.223	0.677	2281.9
22	WSR21		9.300	1.953	1.227	0.892	3038.1
23	WSR22		7.700	1.814	1.247	0.984	3515.0
24	WSR23		8.300	1.718	1.281	1.050	4069.4

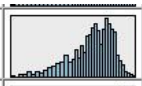
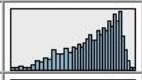
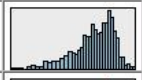
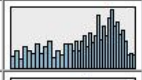
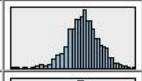
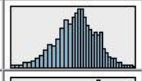
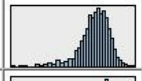
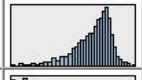
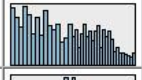
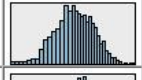
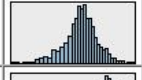
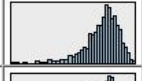
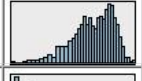
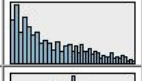
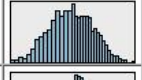
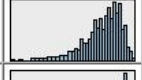
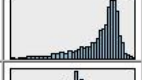
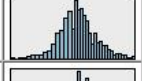
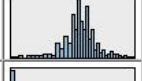
25	WSR_PK		8.400	4.176	1.174	0.539	1607.8
26	WSR_AV		5.900	2.316	0.919	0.746	1069.2
27	T0		31.700	18.982	6.791	-0.676	-390285.6
28	T1		31.100	18.661	6.857	-0.679	-403339.1
29	T2		31.400	18.382	6.919	-0.687	-419400.0
30	T3		30.400	18.135	6.977	-0.689	-431187.7
31	T4		30.800	17.920	7.023	-0.689	-440186.0
32	T5		30.500	17.787	7.084	-0.695	-455681.7
33	T6		30.900	17.924	7.287	-0.677	-483112.6
34	T7		32.400	18.772	7.617	-0.648	-528071.0
35	T8		33.300	20.148	7.599	-0.636	-514738.6
36	T9		35.000	21.586	7.462	-0.661	-506477.3
37	T10		37.600	22.821	7.373	-0.696	-514280.4
38	T11		38.300	23.739	7.306	-0.703	-505786.6
39	T12		40.100	24.369	7.251	-0.687	-483147.0
40	T13		40.400	24.784	7.172	-0.662	-450504.1
41	T14		40.100	25.038	7.098	-0.630	-415276.7
42	T15		39.600	25.037	7.034	-0.596	-382253.6
43	T16		40.500	24.714	6.987	-0.588	-369581.2
44	T17		40.500	23.938	6.946	-0.560	-346092.9
45	T18		38.000	22.805	6.852	-0.558	-331330.5
46	T19		36.000	21.735	6.686	-0.596	-328796.9
47	T20		34.400	20.941	6.618	-0.626	-334664.1
48	T21		33.000	20.355	6.612	-0.641	-341846.6
49	T22		32.700	19.847	6.644	-0.655	-354406.6
50	T23		31.700	19.398	6.700	-0.664	-368307.3

51	T_PK		39.900	25.888	6.859	-0.643	-382582.9
52	T_AV		33.300	21.159	6.749	-0.619	-350717.3
53	T85		29.000	13.714	4.760	-0.668	-132904.5
54	RH85		0.990	0.576	0.255	-0.496	-15.1
55	U85		34.090	1.976	4.516	0.158	26771.2
56	V85		38.290	1.941	6.091	0.060	25036.3
57	HT85		285.000	1533.700	35.322	-0.679	-55161879.0
58	T70		26.100	6.074	3.800	-0.754	-76233.6
59	RH70		0.990	0.399	0.262	0.322	10.7
60	U70		42.580	5.165	6.327	0.157	73317.2
61	V70		49.220	1.010	6.292	0.053	24354.7
62	HT70		330.000	3148.364	46.654	-0.852	-159616329.2
63	T50		23.100	-10.501	3.805	-0.423	-42974.9
64	RH50		0.990	0.300	0.245	0.843	22.8
65	U50		56.280	9.821	9.343	0.165	247802.7
66	V50		56.410	0.647	7.352	0.314	229951.1
67	HT50		485	5822.426	75.711	-0.831	-665108744.2
68	KI		98.750	10.680	20.171	-0.780	-11799717.8
69	TT		69.250	37.689	11.007	-1.300	-3197760.3
70	SLP		355	10165.476	52.056	0.506	131514351.5
71	SLP_		275	-0.836	34.135	0.296	21724417.3
72	Precp		20.650	0.359	1.263	7.232	26838.7

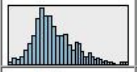
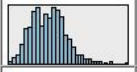
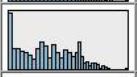
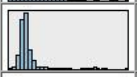
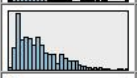
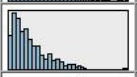
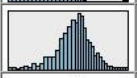
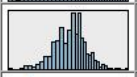
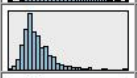
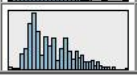
Data Audit von onehr

	Feld	Diagramm	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	WSR0		6.900	1.629	1.253	1.006	3649.6
2	WSR1		6.900	1.570	1.243	1.120	3970.7
3	WSR2		7.100	1.537	1.219	1.141	3808.5
4	WSR3		6.700	1.520	1.195	1.114	3503.7
5	WSR4		7.200	1.515	1.187	1.170	3610.8
6	WSR5		7.400	1.533	1.158	1.204	3452.9
7	WSR6		7.200	1.632	1.137	1.099	2984.3
8	WSR7		7.500	2.039	1.154	0.737	2091.0
9	WSR8		9.100	2.526	1.172	0.484	1438.6
10	WSR9		8.200	2.844	1.208	0.303	984.1
11	WSR10		8.600	2.982	1.291	0.307	1218.8
12	WSR11		8.700	3.030	1.379	0.368	1777.4
13	WSR12		8.900	3.057	1.410	0.387	1998.3
14	WSR13		9.500	3.115	1.431	0.380	2055.2
15	WSR14		9.000	3.185	1.419	0.306	1616.2
16	WSR15		8.900	3.235	1.372	0.256	1221.1
17	WSR16		8.500	3.203	1.274	0.216	822.8
18	WSR17		7.900	2.945	1.236	0.230	801.1
19	WSR18		7.400	2.582	1.243	0.385	1362.2
20	WSR19		7.100	2.302	1.230	0.516	1772.6
21	WSR20		8.700	2.106	1.223	0.675	2280.9
22	WSR21		9.300	1.954	1.227	0.890	3034.8
23	WSR22		7.700	1.815	1.246	0.983	3511.1
24	WSR23		8.300	1.718	1.281	1.049	4066.1

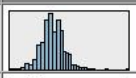
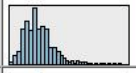
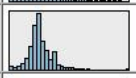
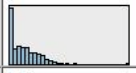
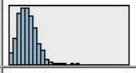
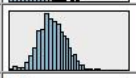
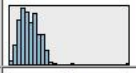
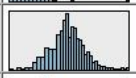
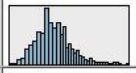
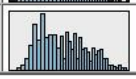
25	WSR_PK		8.400	4.177	1.174	0.538	1605.2
26	WSR_AV		5.900	2.317	0.919	0.745	1067.7
27	T0		31.700	18.986	6.791	-0.677	-390911.3
28	T1		31.100	18.665	6.857	-0.679	-403983.2
29	T2		31.400	18.386	6.919	-0.687	-420058.5
30	T3		30.400	18.139	6.977	-0.689	-431861.2
31	T4		30.800	17.924	7.023	-0.690	-440877.8
32	T5		30.500	17.791	7.084	-0.696	-456392.1
33	T6		30.900	17.928	7.287	-0.678	-483882.4
34	T7		32.400	18.777	7.618	-0.648	-528898.9
35	T8		33.300	20.152	7.599	-0.637	-515583.6
36	T9		35.000	21.590	7.463	-0.662	-507290.9
37	T10		37.600	22.826	7.373	-0.696	-515073.1
38	T11		38.300	23.744	7.306	-0.704	-506545.2
39	T12		40.100	24.374	7.251	-0.688	-483891.4
40	T13		40.400	24.789	7.173	-0.663	-451221.5
41	T14		40.100	25.042	7.099	-0.630	-415986.8
42	T15		39.600	25.040	7.034	-0.596	-382948.6
43	T16		40.500	24.718	6.987	-0.588	-370262.6
44	T17		40.500	23.942	6.946	-0.561	-346761.8
45	T18		38.000	22.809	6.853	-0.559	-331971.1
46	T19		36.000	21.739	6.687	-0.597	-329391.4
47	T20		34.400	20.945	6.618	-0.627	-335237.4
48	T21		33.000	20.359	6.612	-0.642	-342420.4
49	T22		32.700	19.851	6.644	-0.656	-354985.1
50	T23		31.700	19.402	6.700	-0.665	-368900.9

51	T_PK		39.900	25.892	6.859	-0.644	-383227.2
52	T_AV		33.300	21.164	6.749	-0.619	-351318.4
53	T85		29.000	13.718	4.762	-0.668	-133013.1
54	RH85		0.990	0.576	0.255	-0.495	-15.1
55	U85		34.090	1.978	4.516	0.156	26590.8
56	V85		38.290	1.944	6.090	0.059	24587.6
57	HT85		285.000	1533.693	35.314	-0.678	-55119733.6
58	T70		26.100	6.077	3.801	-0.753	-76313.2
59	RH70		0.990	0.399	0.262	0.322	10.7
60	U70		42.580	5.164	6.326	0.157	73469.3
61	V70		49.220	1.011	6.290	0.053	24230.2
62	HT70		330.000	3148.377	46.645	-0.853	-159754874.9
63	T50		23.100	-10.499	3.805	-0.424	-43082.3
64	RH50		0.990	0.300	0.245	0.842	22.7
65	U50		56.280	9.816	9.343	0.166	249432.2
66	V50		56.410	0.647	7.350	0.314	230013.7
67	HT50		485	5822.459	75.705	-0.832	-665939025.0
68	KI		98.750	10.678	20.165	-0.780	-11794344.9
69	TT		69.250	37.685	11.006	-1.299	-3195403.8
70	SLP		355	10165.446	52.058	0.506	131794407.0
71	SLP_		275	-0.850	34.130	0.297	21794705.3
72	Precp		20.650	0.359	1.262	7.233	26840.3

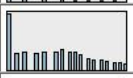
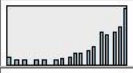
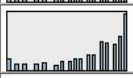
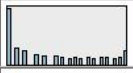
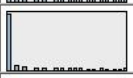
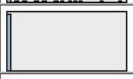
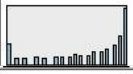
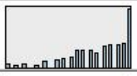
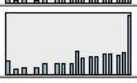
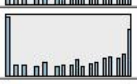
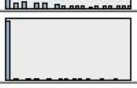
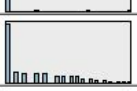
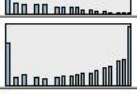
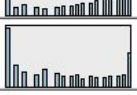
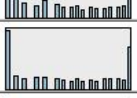
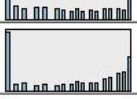
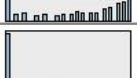
Data Audit von winequality-red

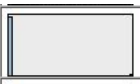
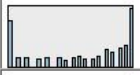
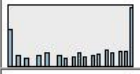
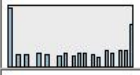
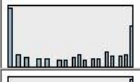
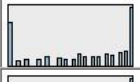
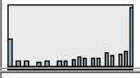
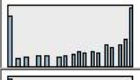
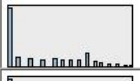
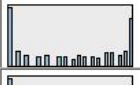
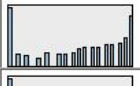
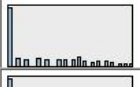
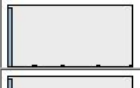
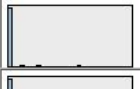
	Feld	Diagramm	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	fixed acidity		11.300	8.320	1.741	0.983	8278.4
2	volatile acidity		1.460	0.528	0.179	0.672	6.2
3	citric acid		1.000	0.271	0.195	0.318	3.8
4	residual sugar		14.600	2.539	1.410	4.541	20311.5
5	chlorides		0.599	0.087	0.047	5.680	0.9
6	free sulfur dioxide		71	15.874	10.458	1.250	2284309.3
7	total sulfur dioxide		283	46.467	32.895	1.516	86099503.9
8	density		0.014	0.997	0.002	0.071	0.0
9	pH		1.270	3.311	0.154	0.194	1.1
10	sulphates		1.670	0.658	0.170	2.429	18.9
11	alcohol		6.500	10.423	1.066	0.861	1662.7

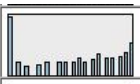
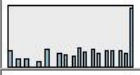
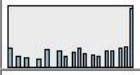
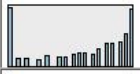
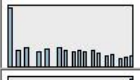
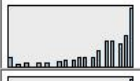
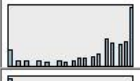
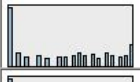
Data Audit von winequality-white

	Feld	Diagramm Stichpr.	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	fixed acidity		10.400	6.855	0.844	0.648	1905.4
2	volatile acidity		1.020	0.278	0.101	1.577	7.9
3	citric acid		1.660	0.334	0.121	1.282	11.1
4	residual sugar		65.200	6.391	5.072	1.077	687953.1
5	chlorides		0.337	0.046	0.022	5.023	0.3
6	free sulfur dioxide		287	35.303	17.000	1.405	33873630.3
7	total sulfur dioxide		431	138.358	42.494	0.390	146796107.2
8	density		0.052	0.994	0.003	0.978	0.0
9	pH		1.100	3.188	0.151	0.458	7.7
10	sulphates		0.860	0.490	0.114	0.977	7.1
11	alcohol		6.200	10.514	1.231	0.487	4445.9

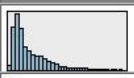
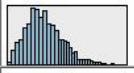
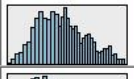
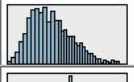
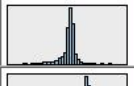
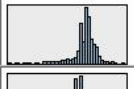
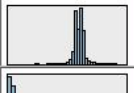
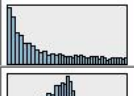
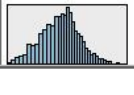
Data Audit von optdigits

	Feld	Diagramm Stichpr.	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	block1		0	0.000	0.000	--	0.0
2	block2		8	0.301	0.867	3.880	9658.3
3	block3		16	5.482	4.632	0.480	182267.2
4	block4		16	11.806	4.260	-1.241	-366483.6
5	block5		16	11.451	4.538	-1.051	-374981.9
6	block6		16	5.505	5.613	0.631	426327.0
7	block7		16	1.387	3.371	2.799	409764.0
8	block8		16	0.142	1.052	9.448	41970.7
9	block9		5	0.002	0.089	50.414	133.8
10	block10		15	1.961	3.052	1.690	183588.7
11	block11		16	10.577	5.435	-0.733	-449637.2
12	block12		16	11.715	4.012	-0.792	-195293.6
13	block13		16	10.625	4.788	-0.615	-257703.3
14	block14		16	8.296	5.936	-0.147	-117781.8
15	block15		16	2.200	4.062	2.027	519031.8
16	block16		15	0.152	0.989	8.385	30963.9
17	block17		5	0.005	0.120	30.131	198.2
18	block18		16	2.596	3.454	1.310	206199.0
19	block19		16	9.581	5.886	-0.488	-380191.8
20	block20		16	6.735	5.918	0.327	259279.3
21	block21		16	7.187	6.143	0.197	174555.3
22	block22		16	8.048	6.291	-0.119	-113188.2
23	block23		16	2.046	3.582	2.033	356811.8
24	block24		8	0.049	0.435	10.850	3422.6

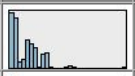
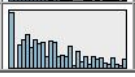
25	block25		1	0.001	0.032	30.879	4.0
26	block26		16	2.336	3.086	1.208	135626.5
27	block27		16	9.239	6.128	-0.372	-326933.0
28	block28		16	9.134	5.903	-0.296	-232774.7
29	block29		16	9.673	6.283	-0.491	-465212.0
30	block30		16	7.868	6.002	-0.016	-13560.1
31	block31		16	2.340	3.625	1.501	273145.2
32	block32		2	0.003	0.065	23.132	23.8
33	block33		1	0.001	0.036	27.608	5.0
34	block34		15	2.043	3.212	1.546	195616.2
35	block35		16	7.659	6.260	0.034	32316.9
36	block36		16	9.238	6.190	-0.342	-309817.8
37	block37		16	10.348	5.920	-0.675	-534675.0
38	block38		16	9.200	5.879	-0.375	-291437.5
39	block39		14	2.913	3.486	0.850	137657.7
40	block40		0	0.000	0.000	--	0.0
41	block41		7	0.027	0.316	14.073	1699.4
42	block42		16	1.406	2.934	2.729	263393.9
43	block43		16	6.457	6.505	0.327	343726.3
44	block44		16	7.187	6.469	0.140	145225.3
45	block45		16	7.922	6.316	-0.018	-16948.7
46	block46		16	8.675	5.806	-0.279	-208510.5
47	block47		16	3.510	4.369	1.004	319840.4
48	block48		6	0.020	0.214	15.607	581.6
49	block49		10	0.018	0.269	23.845	1775.2
50	block50		16	0.820	2.009	3.437	106452.7

51	block51		16	7.869	5.667	-0.081	-56228.5
52	block52		16	9.886	5.142	-0.380	-197161.2
53	block53		16	9.765	5.315	-0.361	-207119.7
54	block54		16	9.283	5.941	-0.443	-354760.6
55	block55		16	3.744	4.902	1.134	510301.8
56	block56		12	0.148	0.768	6.918	11960.1
57	block57		1	0.000	0.016	61.830	1.0
58	block58		10	0.283	0.928	4.459	13614.4
59	block59		16	5.856	4.980	0.421	198659.0
60	block60		16	11.943	4.335	-1.287	-400493.2
61	block61		16	11.461	4.992	-1.139	-541136.2
62	block62		16	6.700	5.776	0.232	170969.5
63	block63		16	2.106	4.028	2.056	513369.6
64	block64		16	0.202	1.151	8.006	46595.6

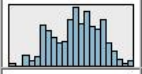
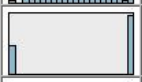
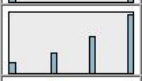
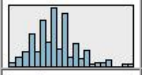
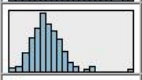
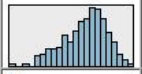
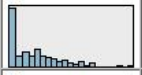
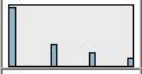
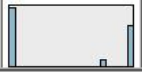
Data Audit von magic04

	Feld	Diagramm Stichpr.	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	fLength		329.894	53.250	42.365	2.014	2911680163.3
2	fWidth		256.382	22.181	18.346	3.372	395922172.4
3	fSize		3.382	2.825	0.473	0.876	1757.4
4	fConc		0.880	0.380	0.183	0.486	56.5
5	fConc1		0.675	0.215	0.111	0.686	17.6
6	fAsym		1033.157	-4.332	59.206	-1.046	-4130051609.0
7	fM3Long		570.101	10.546	51.000	-1.123	-2833123414.5
8	fM3Trans		385.746	0.250	20.827	0.120	20638092.1
9	fAlpha		90.000	27.646	26.104	0.851	287817772.2
10	fDist		494.278	193.818	74.732	0.230	1822242264.7

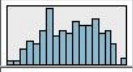
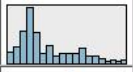
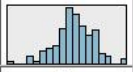
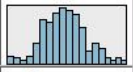
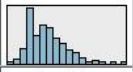
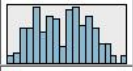
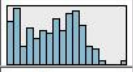
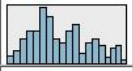
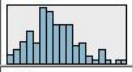
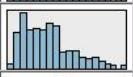
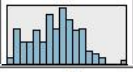
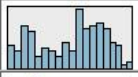
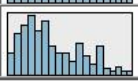
Data Audit von transfusion

	Feld	Diagramm	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	Recency		74	9.507	8.095	1.880	743253.5
2	Frequency		49	5.515	5.839	3.211	476341.0
3	Monetary		12250	1378.676	1459.827	3.211	7442828658899.2
4	Time		96	34.282	24.377	0.749	8087721.2

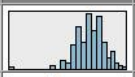
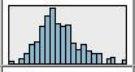
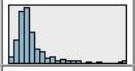
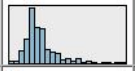
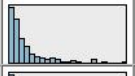
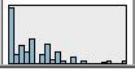
Data Audit von heart

	Feld	Diagramm	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	f1		48	54.433	9.109	-0.164	-33019.4
2	f2		1	0.678	0.468	-0.765	-21.0
3	f3		3	3.174	0.950	-0.879	-201.2
4	f4		106	131.344	17.862	0.723	1099496.3
5	f5		438	249.659	51.686	1.184	43641268.8
6	f6		1	0.148	0.356	1.992	24.0
7	f7		2	1.022	0.998	-0.045	-11.9
8	f8		131	149.678	23.166	-0.528	-1751773.0
9	f9		1	0.330	0.471	0.729	20.3
10	f10		6.200	1.050	1.145	1.263	506.5
11	f11		2	1.585	0.614	0.543	33.6
12	f12		3	0.670	0.944	1.210	271.7
13	f13		4	4.696	1.941	0.287	560.6

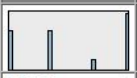
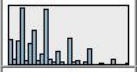
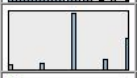
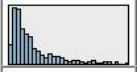
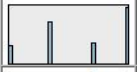
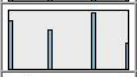
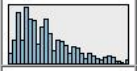
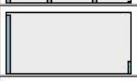
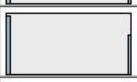
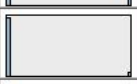
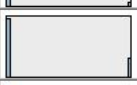
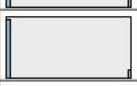
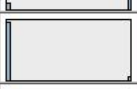
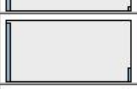
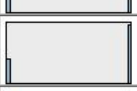
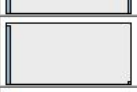
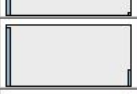
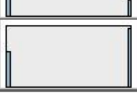
Data Audit von wine

	Feld	Diagramm	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	alcohol		3.800	13.001	0.812	-0.051	-4.8
2	malic acid		5.060	2.336	1.117	1.040	253.7
3	ash		1.870	2.367	0.274	-0.177	-0.6
4	alcalinity of ash		19.400	19.495	3.340	0.213	1388.7
5	magnesium		92	99.742	14.282	1.098	559957.5
6	total phenols		2.900	2.295	0.626	0.087	3.7
7	flavanoids		4.740	2.029	0.999	0.025	4.4
8	nonflavanoid phenols		0.530	0.362	0.124	0.450	0.2
9	proanthocyanins		3.170	1.591	0.572	0.517	17.0
10	color intensity		11.720	5.058	2.318	0.869	1894.0
11	hue		1.230	0.957	0.229	0.021	0.0
12	od280/od315		2.730	2.612	0.710	-0.307	-19.2
13	proline		1402	746.893	314.907	0.768	4196384418.5

Data Audit von bupa

	Feld	Diagramm	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	mean corpuscular volume		38	90.159	4.448	-0.388	-11691.6
2	alkphos		115	69.870	18.348	0.754	1592046.0
3	sgpt alamine		151	30.406	19.512	3.063	7783539.8
4	sgot		77	24.643	10.064	2.293	799515.8
5	gammagt		292	38.284	39.255	2.866	59291991.3
6	drinks		20.000	3.455	3.338	1.544	19634.7

Data Audit von germannummer

	Feld	Diagramm	Bereich	Mittelwert	Std.Abw.	norm. Schiefe	Schiefe z.Mom
1	f1		3	2.577	1.258	0.007	13.8
2	f2		68	20.903	12.059	1.094	1912935.4
3	f3		4	2.545	1.083	-0.012	-15.1
4	f4		182	32.711	28.253	1.949	43824796.9
5	f5		4	2.105	1.580	1.017	3998.2
6	f6		4	3.384	1.208	-0.118	-206.9
7	f7		3	2.682	0.708	-0.305	-108.0
8	f8		3	2.845	1.104	-0.273	-365.4
9	f9		3	2.358	1.050	0.046	52.7
10	f10		56	35.546	11.375	1.021	1498023.9
11	f11		2	2.675	0.706	-1.827	-639.7
12	f12		3	1.407	0.578	1.273	244.6
13	f13		1	1.155	0.362	1.909	90.4
14	f14		1	1.404	0.491	0.392	46.2
15	f15		1	1.037	0.189	4.913	33.0
16	f16		1	0.234	0.424	1.258	95.4
17	f17		1	0.103	0.304	2.616	73.4
18	f18		1	0.907	0.291	-2.807	-68.7
19	f19		1	0.041	0.198	4.637	36.1
20	f20		1	0.179	0.384	1.677	94.3
21	f21		1	0.713	0.453	-0.943	-87.2
22	f22		1	0.022	0.147	6.527	20.6
23	f23		1	0.200	0.400	1.502	96.0
24	f24		1	0.630	0.483	-0.539	-60.6