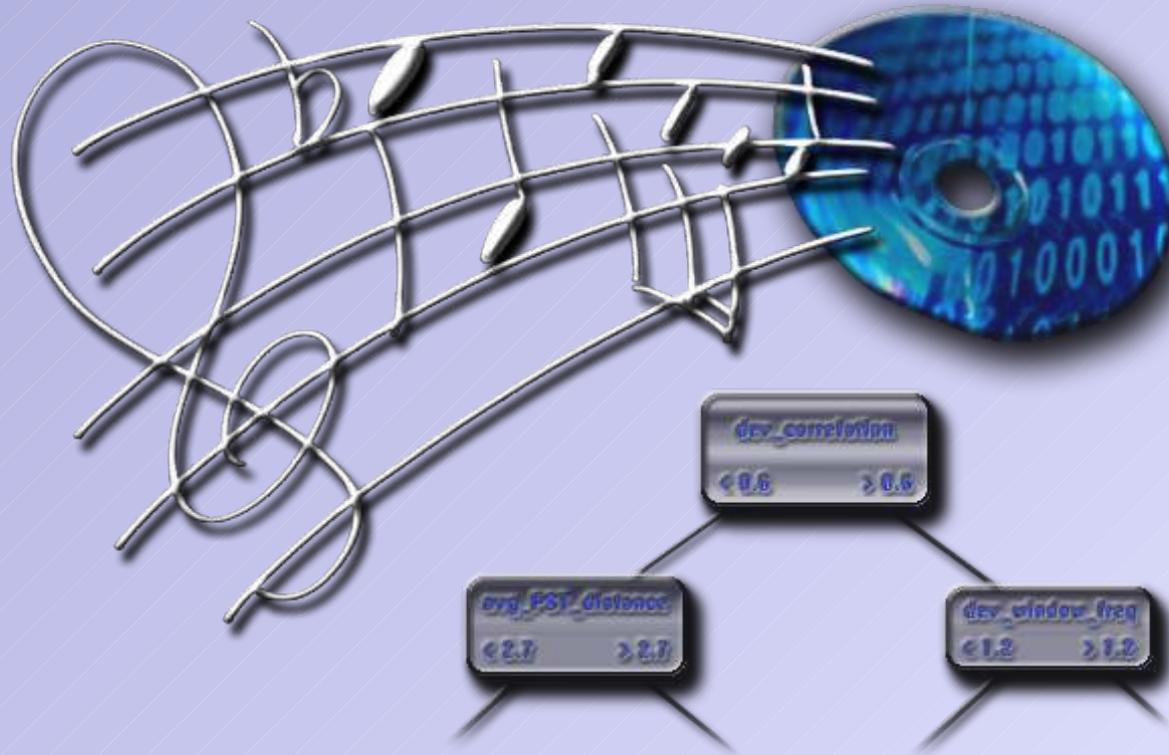


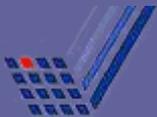


Merkmalsgenerierung und Merkmalsraumtransformationen





-
- Warum Merkmalsgenerierung?
 - Merkmalsraumtransformationen
 - Manuelle Merkmalsgenerierung
 - PCA und Kernfunktionen als Generierung
 - Anwendung: Merkmalsextraktion aus Audiodaten
 - Automatisierte Merkmalsraumtransformationen als Suche





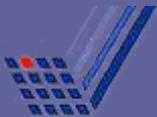
Gegeben:

Beispiele X in LE

- die anhand einer Wahrscheinlichkeitsverteilung P auf X erzeugt wurden und
- Wobei jedes x mit einem Funktionswert $y = f(x)$ versehen ist.

H die Menge von Funktionen in LH .

Ziel: Eine Hypothese $h(X) \in H$, die das erwartete Fehlerrisiko $R(h)$ minimiert.





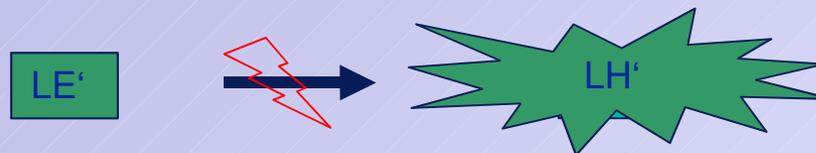
Beispiele für Hypothesen



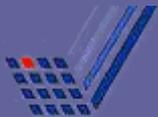
- Regeln
- Entscheidungsbäume
- Funktionen

Problem:

Ungenügende Eingaberepräsentation verhindert das Finden optimaler Hypothesen.



Hinzufügen weiterer
Merkmale vergrößert
den Suchraum für
mögliche Hypothesen!





Lehrstuhl für
Künstliche
Intelligenz

Bei-Spiel

A	B	Klasse
10	10	+
20	10	-
5	10	-
25	5	+
25	3	-
25	7	-
6	20	+
...



Ingo Mierswa



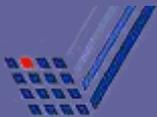
Merkmalsgenerierung: Fläche

Neues Merkmal:

$$\text{Fläche} = A * B$$

Laenge und Breite
Flaeche

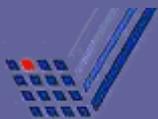
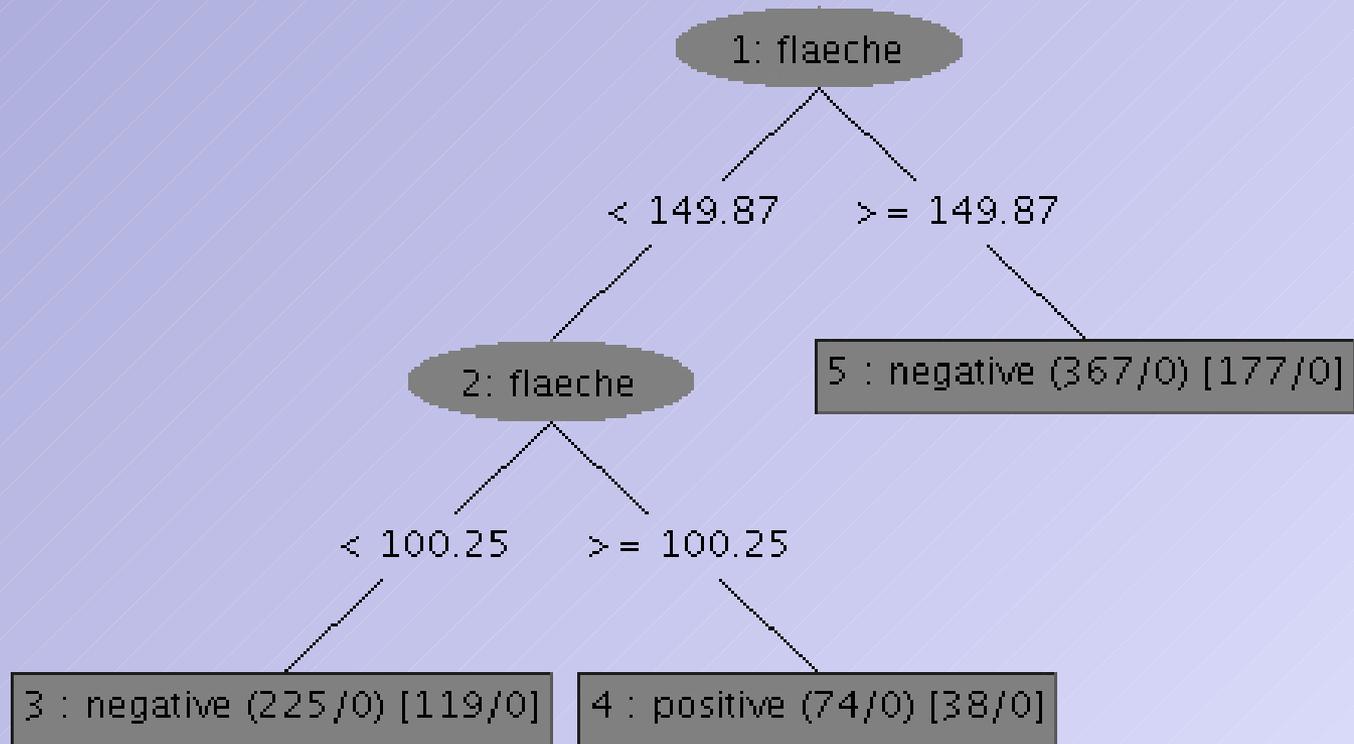
A	B	Fläche	Klasse
10	10	100	+
20	10	200	-
5	10	50	-
25	5	125	+
25	3	75	-
25	7	175	-
6	20	120	+
...



DT mit Merkmal Fläche



Lehrstuhl für
Künstliche
Intelligenz



Ingo Mierswa

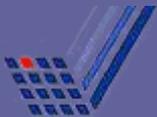


Verbesserung der Performanz

Generierte Merkmale erlauben u.U. Hypothesen, die die Daten viel besser generalisieren:

	<i>Precision</i>	<i>Recall</i>
<i>Ohne Fläche</i>	66,33 %	58,04%
<i>Mit Fläche</i>	99,12 %	100 %

(1000 Beispiele, A und B in $[0,30]$, 10-fach kreuzvalidiert)

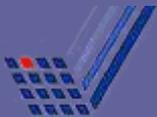




Warum Merkmalsgenerierung?

Merkmalsgenerierung ist nötig, da...

- Zusammenhänge zwischen Merkmalen explizit dargestellt werden können
- Hypothesen leichter verständlich werden
- Die Vorhersageperformanz erhöht werden kann (Flann und Dietterich, 1986)
- Lernkomplexität verringert werden kann (Fawcett und Utgoff, 1992)
- Geforderte Hypothesensprachen benutzt werden können (z.B. lineare Modelle oder einfache Entscheidungsbäume)





Merkmalsraumtransformationen

Zusammenfassung von Selektion und Generierung:

Gegeben sind n Instanzen $(\vec{x}_i, y_i) \in X \times Y$

Die Dimensionen von X heißen Merkmale, X heißt Merkmalsraum.

Gesucht ist eine Transformation des Merkmalsraums X in X' ,
so dass Lernverfahren einfacher geeignete Hypothesen
finden können.

Merkmalsraumtransformation



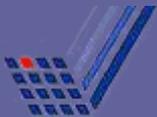


Merkmalsraumtransformationen fassen Selektion und Generierung zusammen.

Wettkampf:

Selektion verkleinert den Merkmalsraum,
Generierung fügt weitere Merkmale hinzu.

Nicht jedes neu generierte Merkmal ist auch sinnvoll!





Manuelle Merkmalsgenerierung



- Explizite Darstellung von „Ahnungen“ durch Hinzufügen von Merkmalen
- Entscheidender Vorverarbeitungsschritt für das maschinelle Lernen
- Äußerst mühsam: ca. 80 % der Arbeitszeit





Principal Component Analysis (Wdh.)

Verringerung der Dimension durch Konstruktion neuer Merkmale

- Möglicherweise korrelierte Merkmale werden in eine (kleinere) Anzahl unkorrelierter Merkmale (*principal components*) überführt
- Jedes zusätzliche Merkmal bildet eine zunehmende Menge der Datenvarianz ab





Sind Transformationen noch nötig?

- Standard-Kernfunktionen nutzen den Kernel-Trick zur effizienten Abbildung in einer höherdimensionalen Raum.
- Häufig genügt jedoch keine Standard-Funktion den vorliegenden Daten.
- Erster Schritt: Der Versuch, günstige Merkmale zu generieren.
- Diese Erkenntnisse helfen beim Design einer neuen Kernfunktion (Mercer-Bedingungen!)





Merkmalsextraktion

Generiert man neue Merkmale als Funktion einer Reihe, so spricht man von *Merkmalsextraktion*.

Sinnvoll, wenn:

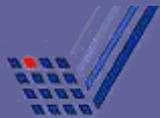
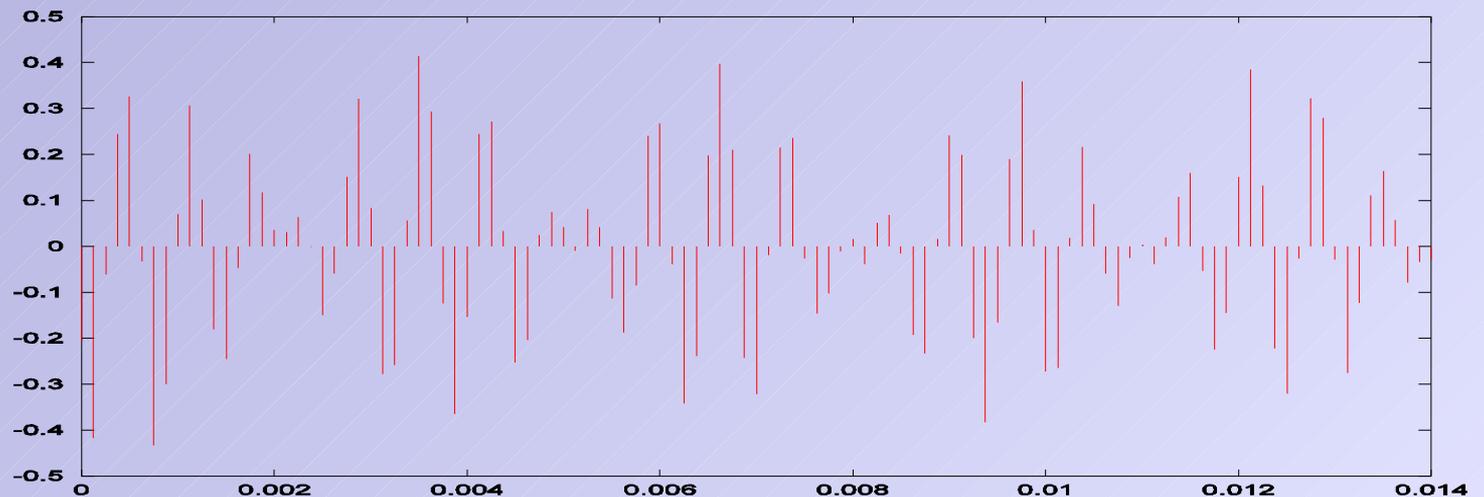
- Die i -ten Werte jeder Instanz nicht korreliert sind
- Translationen der Reihe Vorhersagen ändern (Translationsinvarianz gewünscht)
- Die (implizite) Ordnung der Merkmale explizit dargestellt werden soll
- Die konkrete Form der Reihe nicht die entscheidenden Informationen liefert





Anwendung: Audiodaten

- Musikstücke sind akustische Wellen und damit reellwertige Funktionen der Zeit.
- Digitalisierung approximiert kontinuierliche Funktion.
- Zu jedem Zeitpunkt ist die Elongation der Welle gegeben:





Musikdaten als Zeitreihen

.....

Diese Diskretisierung der akustischen Welle erlaubt die Darstellung eines Musikstückes als

- univariate,
- endliche,
- äquidistante

Zeitreihe mit $7,938 \cdot 10^6$ sample points für ein Lied von drei Minuten. Daher:

Eine Extraktion aussagekräftiger Merkmale ist für eine Weiterverarbeitung durch maschinelle Lernverfahren unabdingbar!





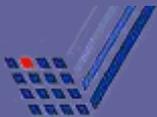
Merkmale in der Zeitdimension



- Merkmale, die sich ohne Basistransformation extrahieren lassen.
- Sei im Folgenden $\tilde{x} = (x_i)_{i \in \{1, \dots, n\}}$ die gegebene Zeitreihe.

- *Mittlere Lautstärke:*

$$LS(\tilde{x}) = \frac{1}{n} \sum_{i=1}^n |x_i|$$

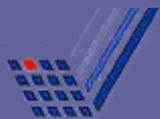




- Transformation *CharacteristicaExtractor* liefert die Extrema einer Reihe. Die übrigen Werte der Reihe sind nach der Anwendung 0.
- Sei $E_x = \{x_i \mid x_i \neq 0\}$ die Menge der Extrema. O.B.d.A. seien die Werte von E_x mit e_1, \dots, e_m bezeichnet.
- Die durchschnittliche Extremadifferenz ist dann:

$$AvgDiff(\tilde{x}) = \frac{1}{|E_x| - 1} \sum_{i=2}^{|E_x|} (\text{index}(e_i) - \text{index}(e_{i-1}))$$

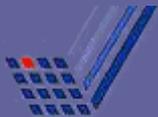
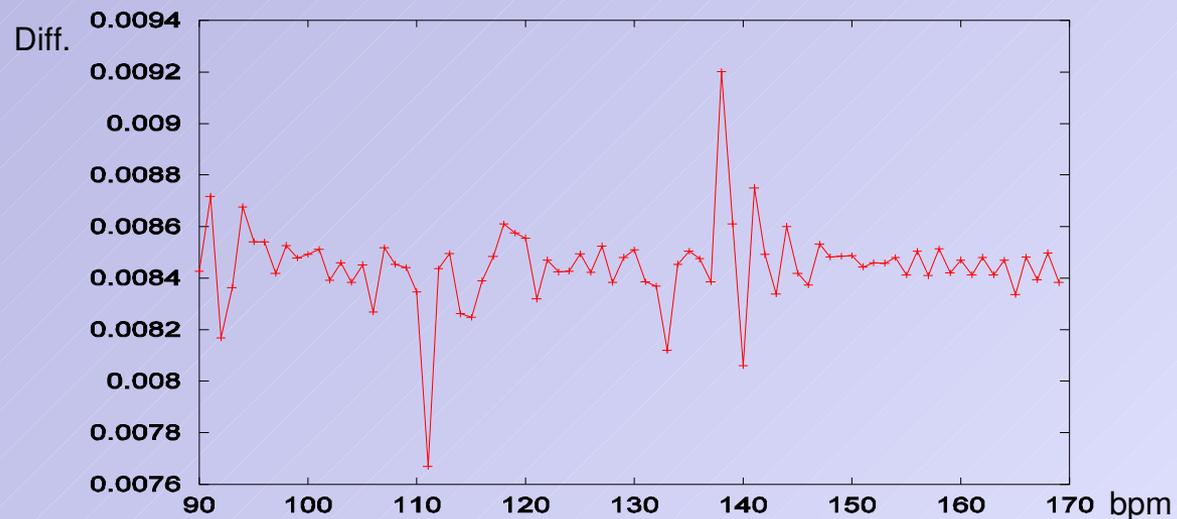
- Entspricht der Kette





Tempo durch Autokorrelation

- *Phasenverschiebung* des Musikstückes um einen Takt.
- Bestimmung der *Autokorrelation* durch Berechnung der gemittelten Differenz zum Original.
- Iteriere Berechnung für alle Geschwindigkeiten im Intervall zwischen 90 und 170 bpm:

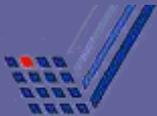




Merkmale in der Frequenzdimension

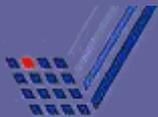


- Merkmale, die nach einer Basistransformation in den Raum der harmonischen Schwingungen ermittelt werden.
- Anwendung einer *FourierTransformation* (FFT).
- Die Funktion *PeakFinder* liefert Peaks im Frequenzspektrum, d.h. Höhe, Frequenz und Breite der k höchsten Peaks werden als Merkmal bereitgestellt.



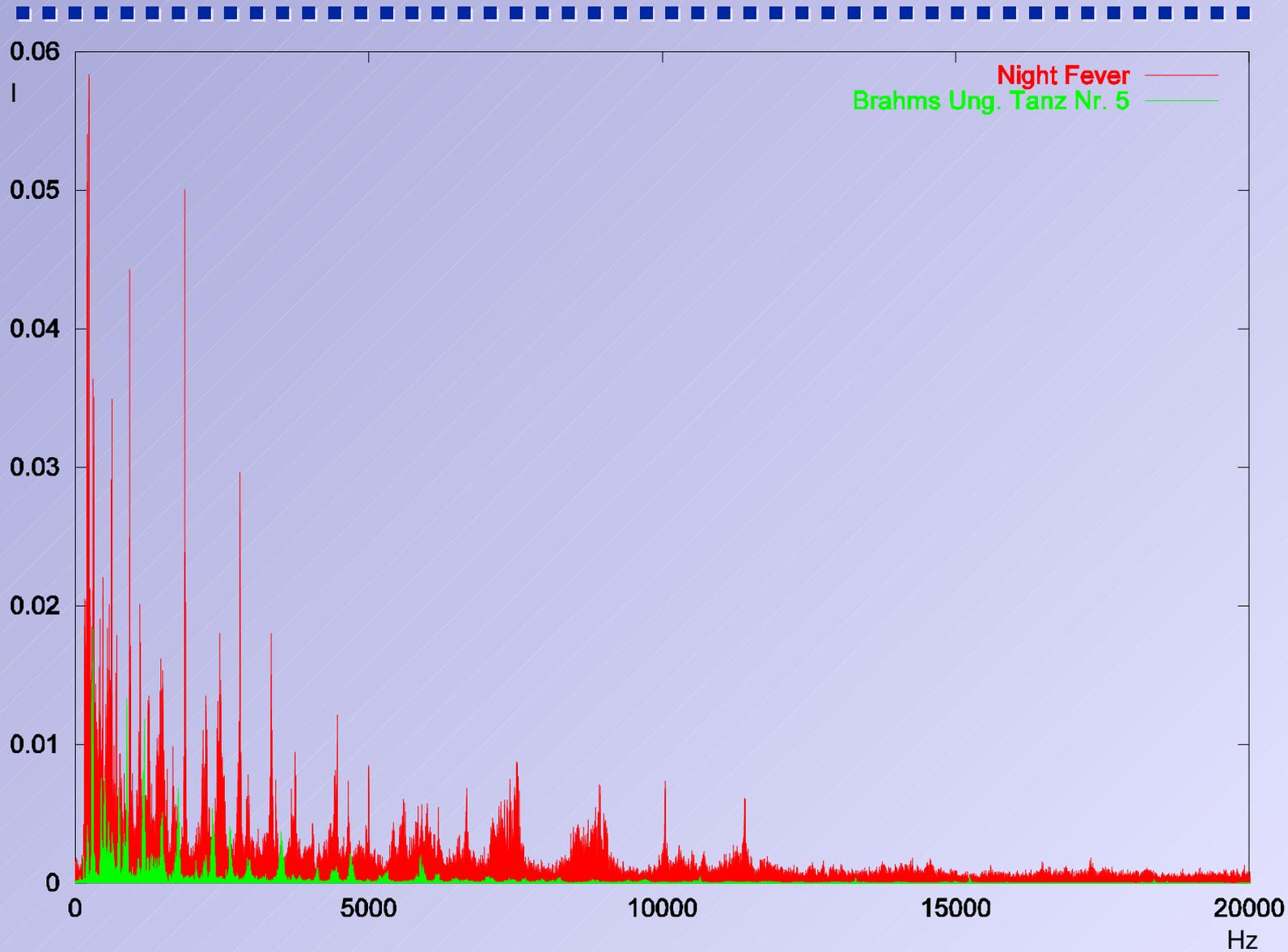


Lehrstuhl für
Künstliche
Intelligenz



Ingo Mierswa

Peaks im Frequenzspektrum



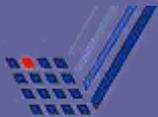
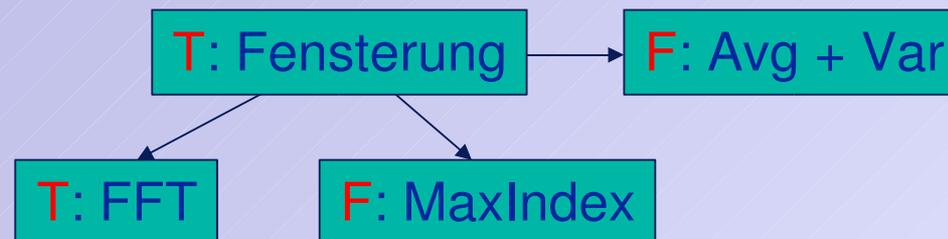


Stärkste Frequenz in Zeitfenstern

- Bei einer vollständigen FT geht die Information verloren, wann genau die Frequenzen auftreten.
- Verschieben eines *Zeitfensters* der Breite w mit Schrittweite s und Berechnung der Frequenz des stärksten Peaks nach einer FT liefert die Zeitreihe:

$$y_j = \text{index} \left(\max \left(FT \left((x_i)_{i \in \{j \cdot s, \dots, j \cdot s + w\}} \right) \right) \right)$$

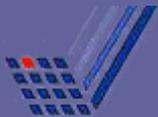
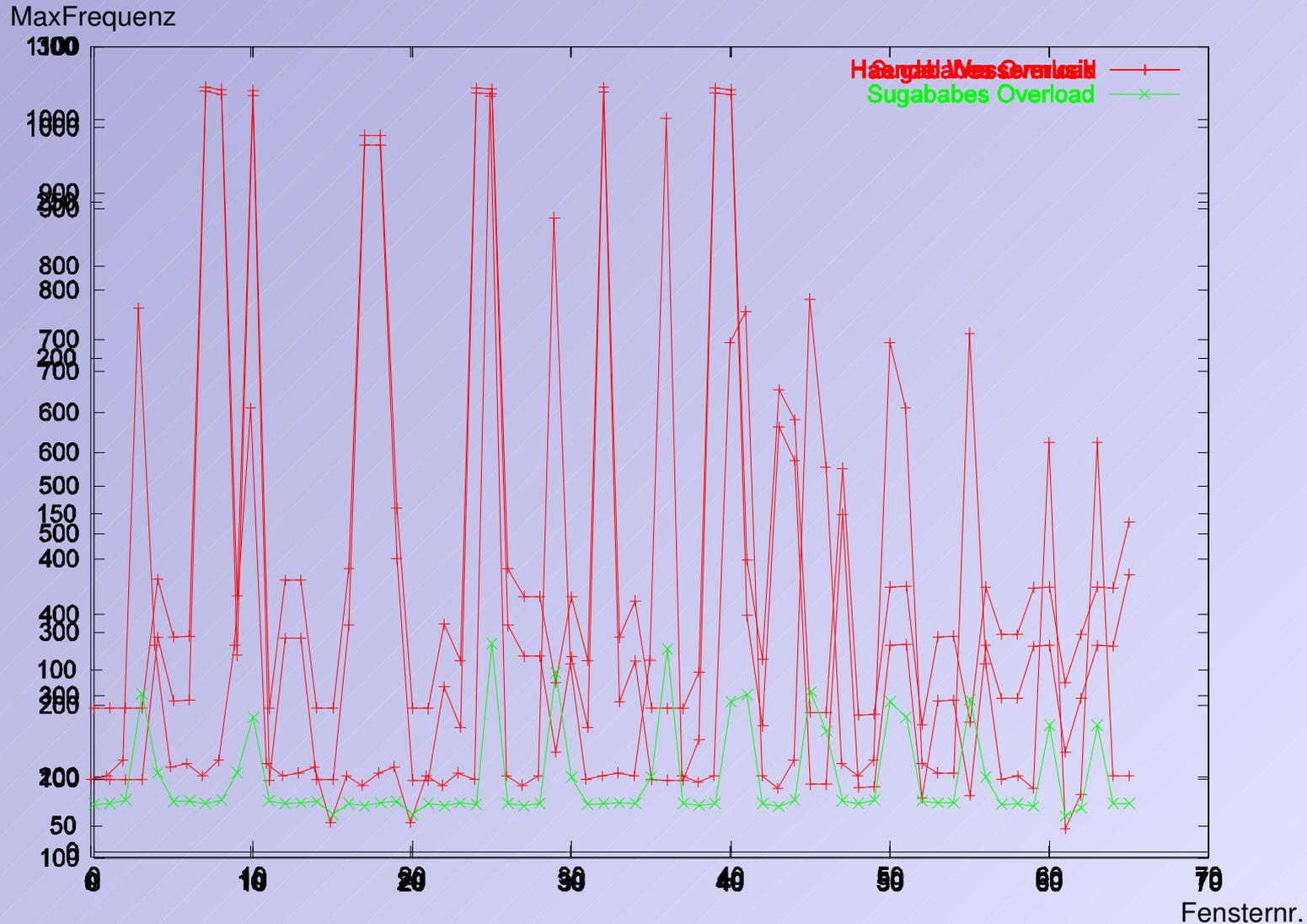
- Durchschnitt und Varianz der neuen Reihe liefern neue Merkmale. Entspricht dem Methodenbaum:





Lehrstuhl für
Künstliche
Intelligenz

Stärkste Frequenz in Zeitfenstern



Ingo Mierswa

Auszeichnungen



Lehrstuhl für
Künstliche
Intelligenz

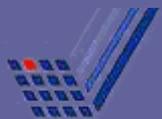
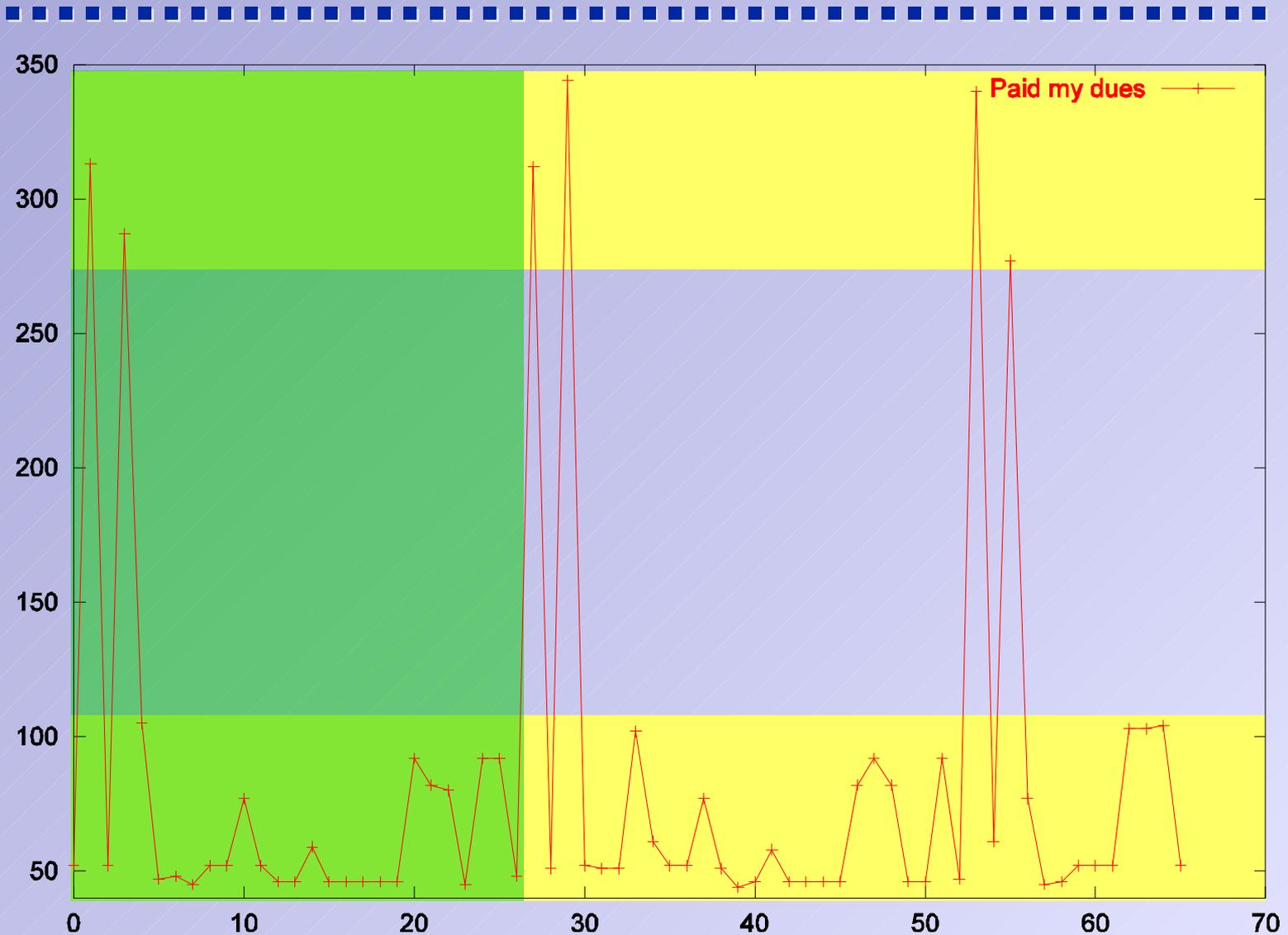


- Nach der gefensterten FT können mit einem *k-Means Clustering* Intervalle im Frequenzbereich gesucht werden.
- Ein Intervall besteht aus Start- und Endwert, Intervalltyp und Stärke (Anzahl der enthaltenen Punkte).
- Diese Werte bilden weitere Merkmale.
- Basierend auf diesen Intervallen ist das Finden weiterer Intervalle in der Zeitdimension möglich.





Intervalle im Frequenzbereich





Rekonstruktion des Zustandsraums



- Musik als nichtlineares dynamisches System.
- Die Kenntnis aller *Zustandsvariablen* erlaubt eine vollständige Beschreibung des Systems.
- *Dimension* eines solchen Systems entspricht der Anzahl der Zustandsvariablen.
- Der *Zustandsraum* besitzt als Basisvektoren die Zustandsvariablen eines Systems.
- Messung der Zustandsvariablen erlaubt direkte Transformation in den Zustandsraum.
- Jedoch: Die Zustandsvariablen sind oft nicht explizit bekannt.





- Ausweg: Konstruktion eines topologisch identischen Raumes – des *Phasenraums*.
- Vektoren sind zeitlich verzögerte Ausschnitte der Reihe: $\vec{p}_i = (x_i, x_{i+d}, x_{i+2d}, \dots, x_{i+(m-1)d})$
- d ist die Verzögerung (*delay*) und m die Dimension des entstehenden Phasenraums. Die Menge

$$P_{d,m} = \{ \vec{p}_i \mid i = 1, \dots, n - (m-1)d \}$$

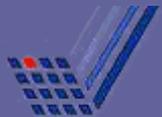
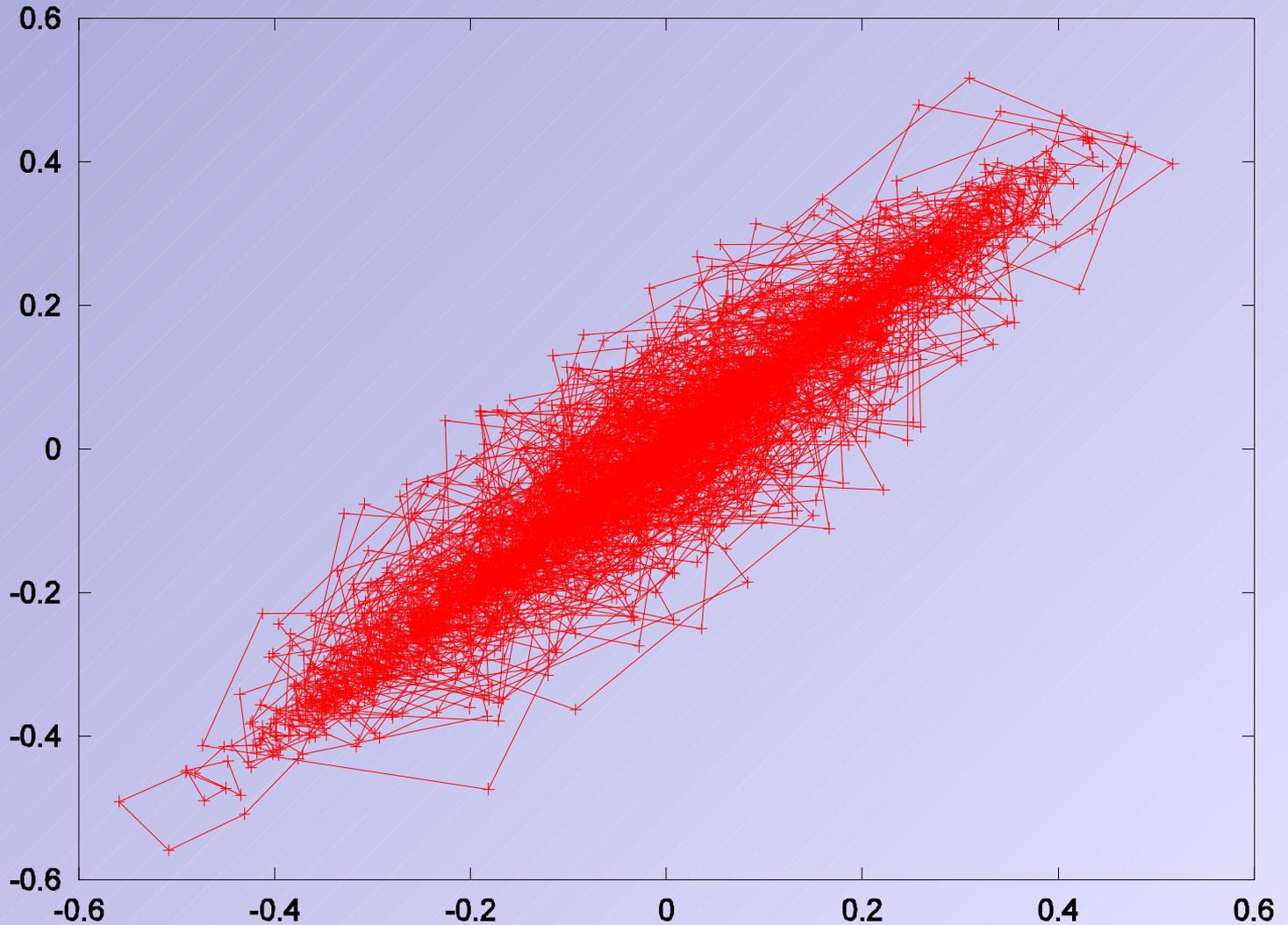
bildet die Transformierte der Reihe im Phasenraum.





Lehrstuhl für
Künstliche
Intelligenz

Beispiel: PST bei Popmusik

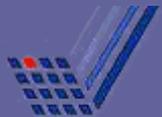
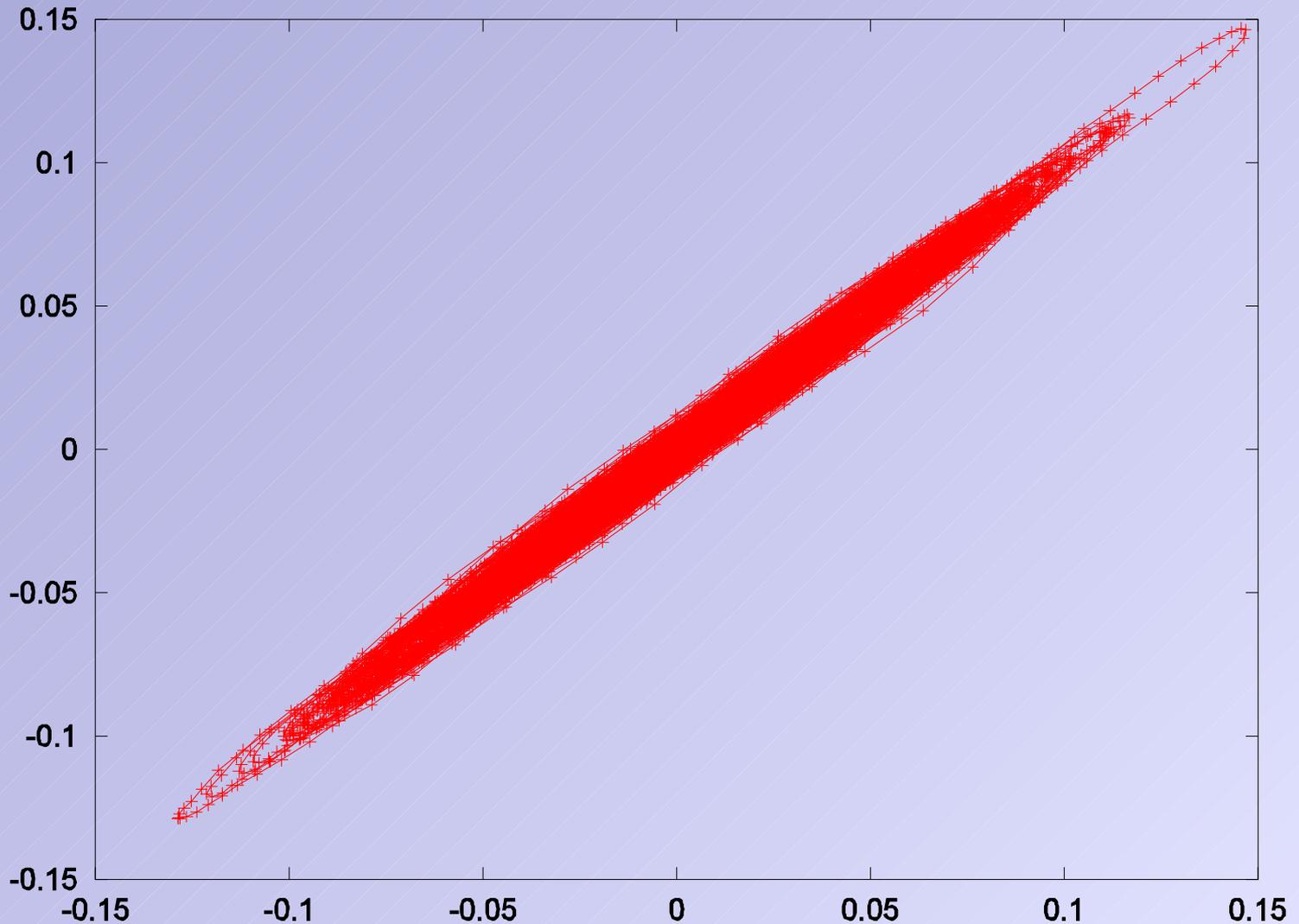


Ingo Mierswa



Lehrstuhl für
Künstliche
Intelligenz

Beispiel: PST bei klassischer Musik

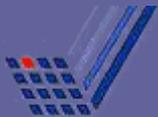
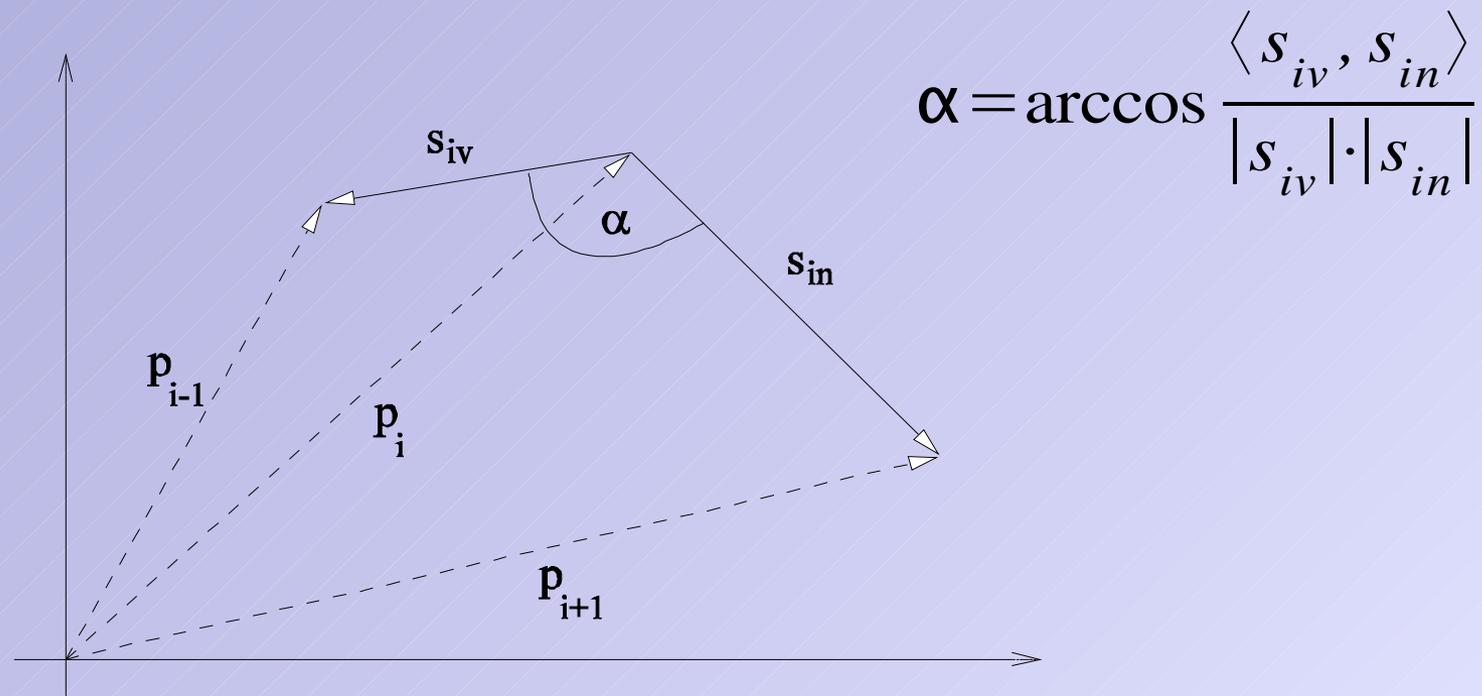


Ingo Mierswa



Winkel Transformation

Motiviert durch die Unterschiede bei Klassik und Pop bildet *AngleTransformation* eine neue Reihe bestehend aus den Winkeln zwischen Teilstücken:





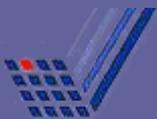
Winkel im Phasenraum



- Der Durchschnitt über die Winkel zwischen allen Teilstücken nach einer Transformation in den Phasenraum bildet somit ein weiteres Merkmal:

$$\text{AvgAngle}(\tilde{x}) = \frac{1}{|P_{d,m}|^{-2}} \sum_{i=2}^{|P_{d,m}|-1} \arccos \frac{\langle s_{iv}, s_{in} \rangle}{|s_{iv}| \cdot |s_{in}|}$$

- Wie zuvor wird auch die Varianz als Merkmal ermittelt.
- Kleine Winkel im Phasenraum entstehen bei gleichmäßiger Änderung *aller* Zustandsvariablen.



Länge der Teilstücke



Lehrstuhl für
Künstliche
Intelligenz



- Im Phasenraum fallen deutlich die unterschiedlichen Längen der Teilstücke für verschiedene Genres auf. Erneut bilden wir Durchschnitt und Varianz:

$$\text{AvgLength}(\tilde{x}) = \frac{1}{|P_{d,m}| - 1} \sum_{i=2}^{|P_{d,m}|} |s_{iv}|$$

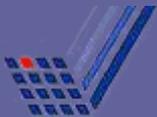
- Jedes Element im Phasenraum entspricht einer Zustandskonfiguration des Systems. Je enger Elemente liegen, desto ähnlicher sind die Werte der Zustandsvariablen zu den Zeitpunkten.





■ Es ergeben sich die Merkmale:

- Mittlere Lautstärke
- Durchschnittliche Extremadifferenz und Varianz
- Tempo sowie Varianz der Autokorrelation
- k höchste Peaks nach Fouriertransformation
- Durchschnitt und Varianz der stärksten Frequenzen im Zeitverlauf
- i Frequenzintervalle
- Durchschnitt und Varianz der Winkel nach Phasenraumtransformation
- Durchschnitt und Varianz der Abstände nach Phasenraumtransformation





Automatisierte Transformation

Problem:

Manuelle Merkmalsraumtransformationen äußerst aufwendig!

- Suche nach optimaler Transformation entspricht Suche im Raum aller möglichen Merkmalsräume (Suchraum)
- Jeder Suchpunkt entspricht einem Merkmalsraum
- Benötigt wird ein Startzustand, eine Abbruchbedingung und eine Suchstrategie





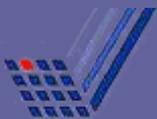
Wrapper-Ansätze (Wdh.)

Die Bewertungsfunktion ist die Performanz desjenigen Lernalgorithmus, der zum Lernen optimiert werden soll.

„Innere“ Kreuzvalidierung:

- Trainingsdaten Lernen
- Testdaten Lernen

Eine zusätzliche äußere Kreuzvalidierung kann die Performanz des Suchverfahrens messen.



Bergsteigerverfahren

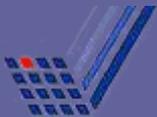


Lehrstuhl für
Künstliche
Intelligenz

Erweiterung des klassischen Bergsteigerverfahrens
Forward Selection (Aha und Bankert, 1996):

- Starte mit der leeren Menge
- Füge dieser Menge in jeder Iteration:
 - ein ursprüngliches Merkmal hinzu
 - oder ein neu generiertes Merkmal hinzu
- ...bis die Performanz nicht weiter steigt.

Genau wie im reinen Selektionsfall leidet das Verfahren unter lokalen Optima durch Interaktivität zwischen den Merkmalen.



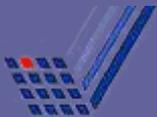
Ingo Mierswa



Evolutionäre Algorithmen

Evolutionäre Algorithmen haben sich bereits für die reine Merkmalsselektion als gut skalierbares Suchverfahren erwiesen (Bala etal, 1995)

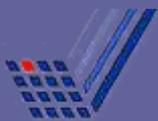
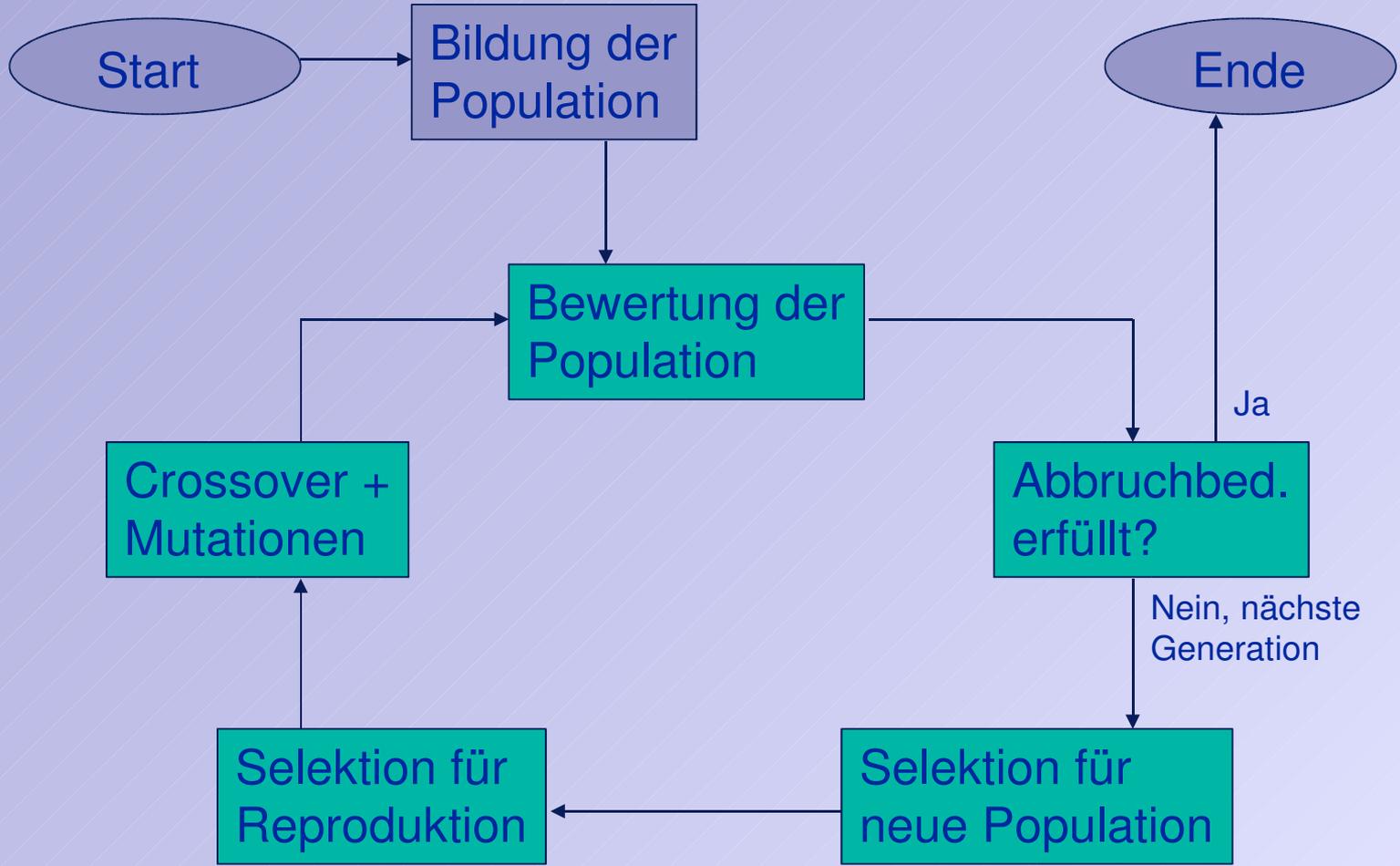
- Merkmalsinteraktivitäten sind durch Kreuzungen weniger problematisch
- Individuen sind Merkmalsräume, d.h. sowohl Originalmerkmale als auch neu generierte
- Berechnung der Fitness durch innere Kreuzvalidierung (Wrapper)



Aufbau des GGA



Lehrstuhl für
Künstliche
Intelligenz

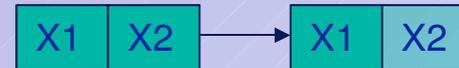




- Hinzufügen eines neu konstruierten Merkmals



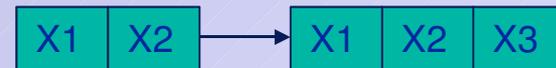
- Deselektieren eines Merkmals



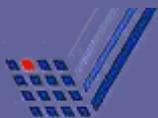
- (Re-)Selektieren eines Merkmals



- Hinzufügen eines Original-Merkmals



- Löschen eines Merkmals





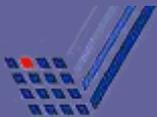
Generierender genetischer Algorithmus (Ritthoff/etal/2002):

- Generierung, Deselektion und Selektion
- Unabhängige Wahrscheinlichkeiten

YAGGA (Fischer/Mierswa):

- Generierung, Hinzufügen von Originalmerkmalen, Löschen
- Wahrscheinlichkeit für Löschung doppelt so groß wie für Hinzufügen und Generieren

Im Mittel bleibt Länge erhalten, bis sich längere oder kürzere Individuen als besser erweisen.





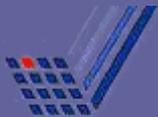
Vermeidung von Introns

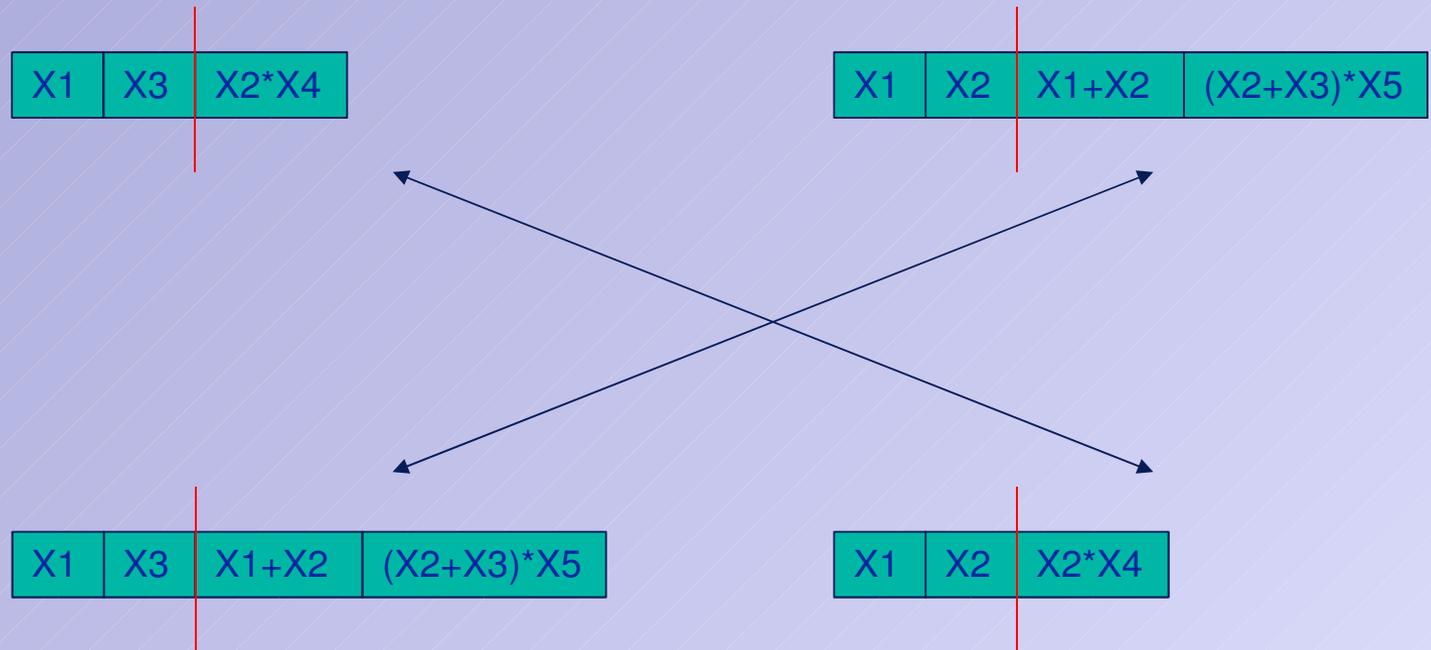


- Introns verursachen Repräsentationsaufblähung (Langdon et al, 1999 sowie Banzhaf et al, 1998)
- *Hier*: Merkmale ohne Informationsgewinn

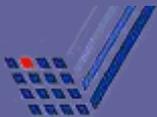
Auswege:

- Beachtung von Reflexivität, Kommutativität und Äquivalenz (Theorem von Richardson, 1968)
- Bestrafung von aufgeblähten Merkmalsräumen in Fitnessfunktion (Soule et al, 1996)
- Mutationswahrscheinlichkeiten wie bei YAGGA





- Kreuzungen erlauben schnelle Suche durch Merkmalsräume
- GAs zur Selektion und Generierung auch für große Datensätze gut geeignet

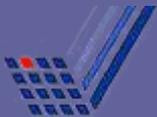




Abgrenzung zur klassischen GP



- Bei der klassischen genetischen Programmierung werden mathematische Beschreibungen explizit erzeugt (Koza, 1992)
- Häufig ist der Zusammenhang zwischen Merkmalen jedoch von größerem Interesse
- Durch die Generierung neuer Merkmale und Verwendung eines Lernverfahrens robuster gegenüber Rauschen



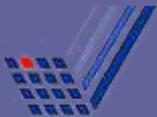
Zusammenfassung



Lehrstuhl für
Künstliche
Intelligenz



- Merkmalsauswahl und –generierung transformieren den Merkmalsraum
- Wesentlicher – aber mühsamer! – Schritt zur Steigerung des Verständnis und der Performanz
- Sowohl PCA als auch Kernfunktionen sind Transformationen des Merkmalsraums
- Automatisierte Suche im Raum der Merkmalsräume, z.B. mit Hilfe von EAs
- Äußerst komplex, daher Vermeidung von Introns essentiell



Ingo Mierswa