

Clusteranalyse – Evaluierung

Teil 2: Extrinsische Evaluierung

Intrinsische Evaluierung

Bewertung von Ergebnissen beim Clustern

Verschiedene Initialisierungen können unterschiedliche Ergebnisse liefern.
Um ein „bestes“ Ergebnis auszuwählen, braucht man ein Evaluationsmaß.



„Intrinsische Evaluierung“

d.h. es werden keine äußeren Faktoren berücksichtigt

- Clusteranalyse gehört zu den nicht überwachten Lernverfahren
- D.h. es gibt kein „korrektes“ Ergebnis, mit dem man das Ergebnis vergleichen kann

Bewerte, wie sauber die Cluster voneinander getrennt sind

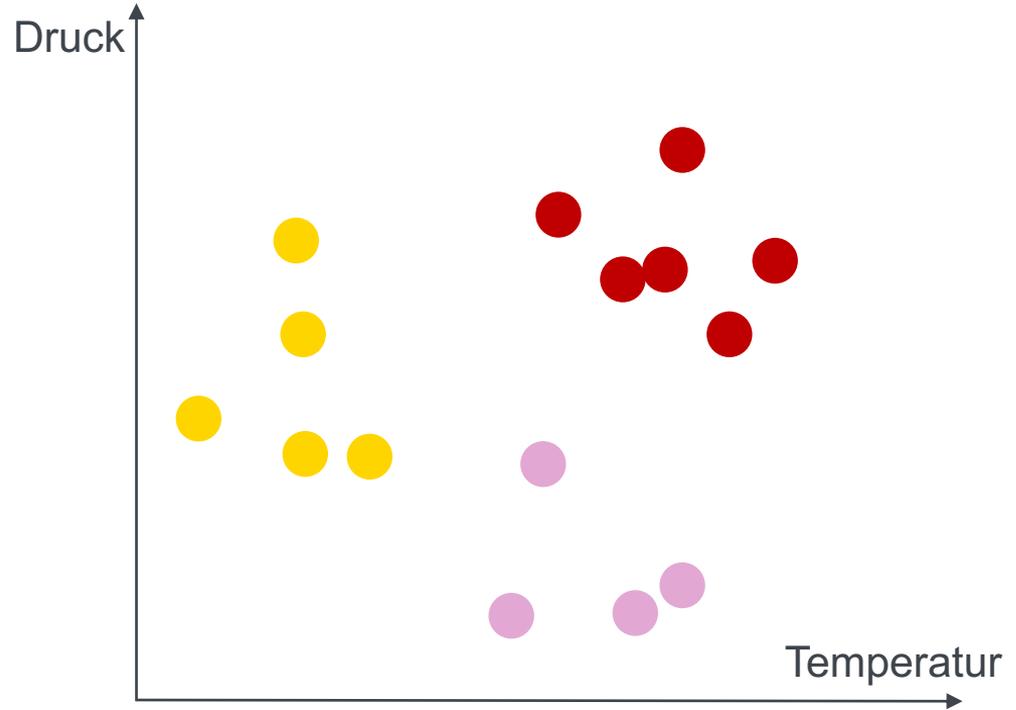
Sind alle Punkte im passendsten Cluster?

Extrinsische Evaluierung

Wenn es doch ein „richtig“ und „falsch“ gibt

- Manchmal hofft man, bei der Clusteranalyse auf bekannte Klassifizierungen zu kommen
- Z.B. aufgrund von Messwerten auf Fehlertypen

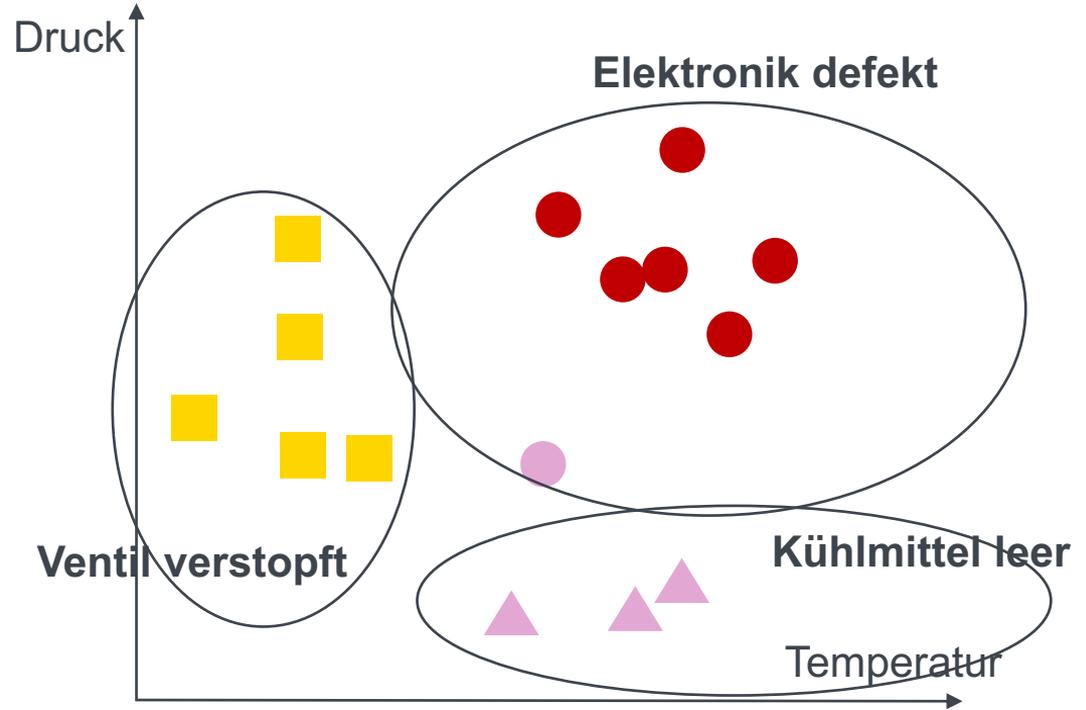
- Ventil verstopft
- Elektronik defekt
- △ Kühlmittel leer



Wenn es doch ein „richtig“ und „falsch“ gibt

- Manchmal hofft man, bei der Clusteranalyse auf bekannte Klassifizierungen zu kommen
- Z.B. aufgrund von Messwerten auf Fehlertypen

- Ventil verstopft
- Elektronik defekt
- △ Kühlmittel leer



Bei der extrinsischen Evaluierung von Clustering sind für alle Punkte korrekte (extern vorgegebene) Klassen bekannt.

Extrinsische Evaluierung

- Ziel: vergleiche Zuordnung zu Clustern mit Zuordnung zu vorgegebenen Klassen
- Fast dasselbe Problem wie bei Klassifikation!
- Nur, dass die Cluster keine Namen haben

Idee:
Konfusionsmatrix!

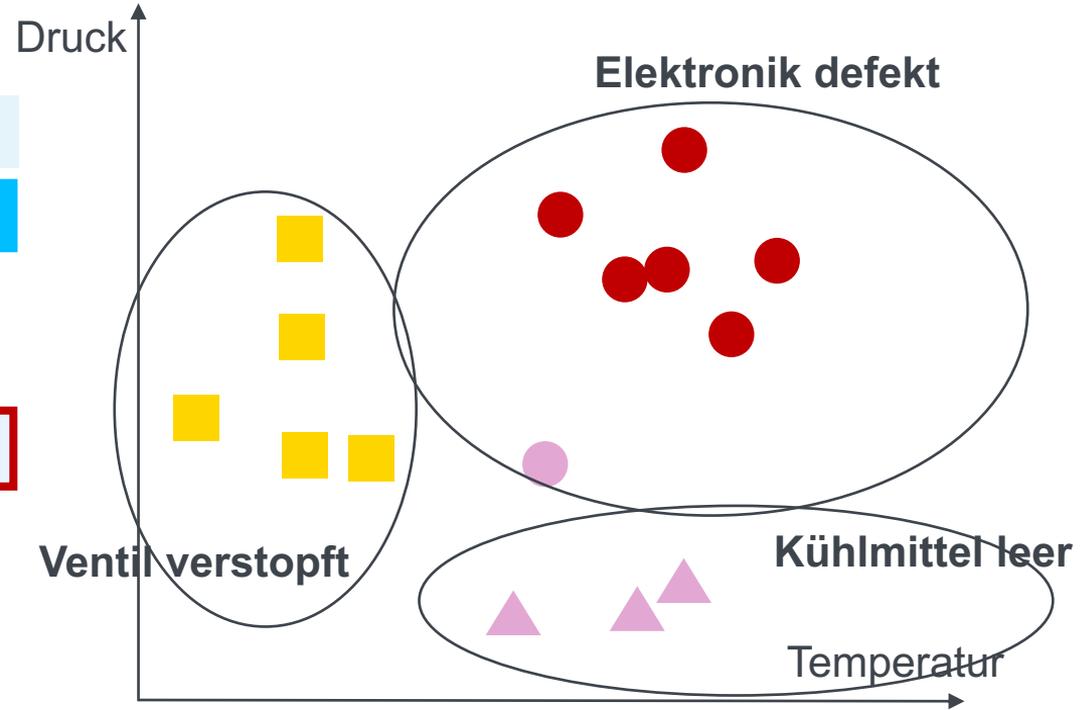
		Vorhersage		
		Katze	Hund	Goldfisch
Tatsächliche Klasse	Katze	93	7	0
	Hund	14	186	0
	Goldfisch	9	11	20

Wenn es doch ein „richtig“ und „falsch“ gibt

		Cluster		
		gelb	rot	rosa
Tatsächliche Klasse	Ventil	5	0	0
	Elektronik	0	6	1
	Kühlmittel	0	0	3



Nur möglich, wenn gleich viele Klassen wie Cluster

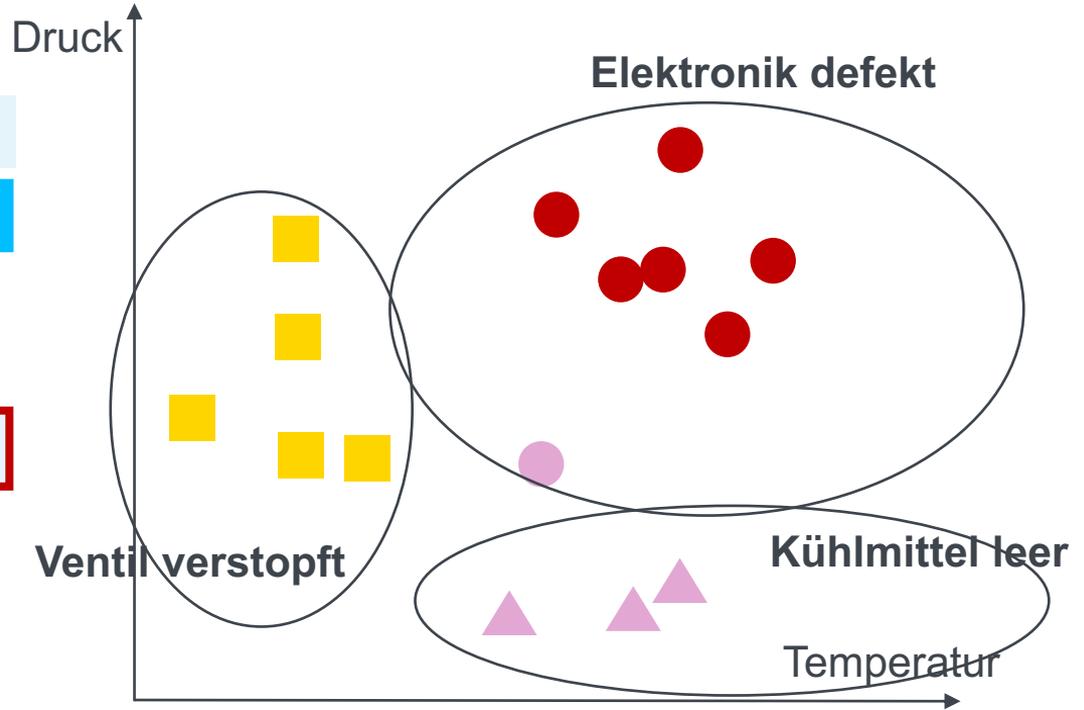


Wenn es doch ein „richtig“ und „falsch“ gibt

		Cluster		
		rot	rosa	gelb
Tatsächliche Klasse	Ventil	0	0	5
	Elektronik	6	1	0
	Kühlmittel	0	3	0



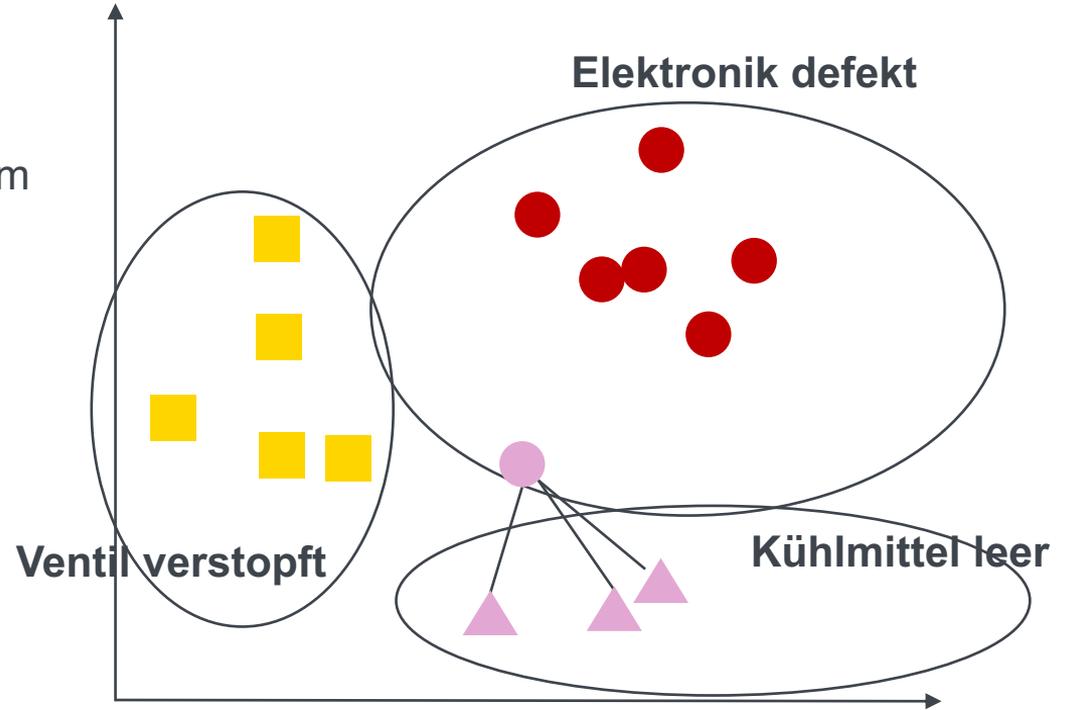
Nur möglich, wenn man weiß, welche Klasse welchem Cluster entspricht



Die Lösung: Paarweise Auswertung

- 15 Punkte: 105 Paare von Punkten
- Nicht erwünscht:
 - Paare mit verschiedenen Fehlern im selben Cluster
 - Paare mit gleichem Fehler in verschiedenen Clustern

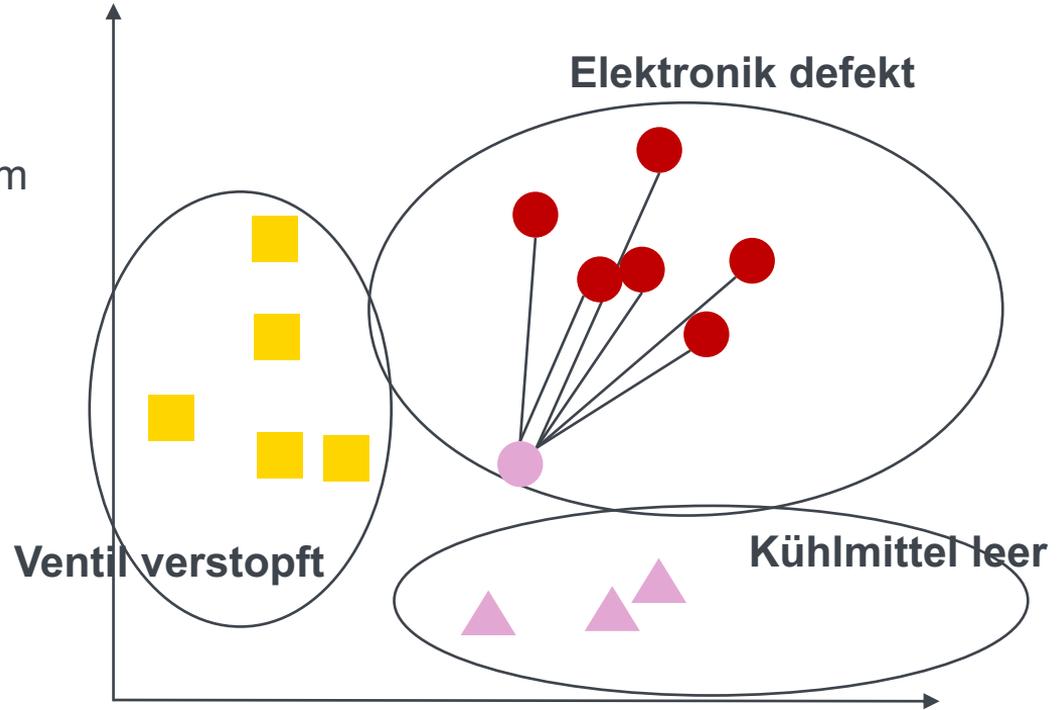
		Cluster	
		versch.	gleich
Fehler	verschieden		3
	gleich		



Die Lösung: Paarweise Auswertung

- 15 Punkte: 105 Paare von Punkten
- Nicht erwünscht:
 - Paare mit verschiedenen Fehlern im selben Cluster
 - Paare mit gleichem Fehler in verschiedenen Clustern

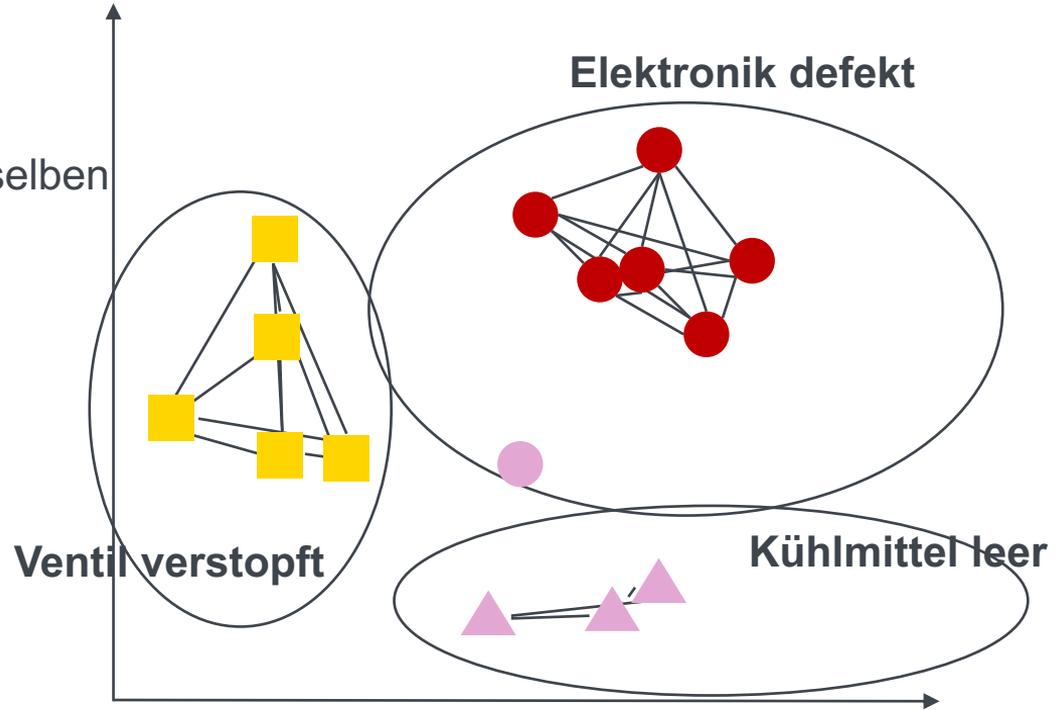
		Cluster	
		versch.	gleich
Fehler	verschieden		3
	gleich	6	



Die Lösung: Paarweise Auswertung

- 15 Punkte: 105 Paare von Punkten
- Erwünscht:
 - Paare im selben Cluster, mit dem selben Fehler
 - Paare in verschiedenen Clustern, mit verschiedenen Fehlern

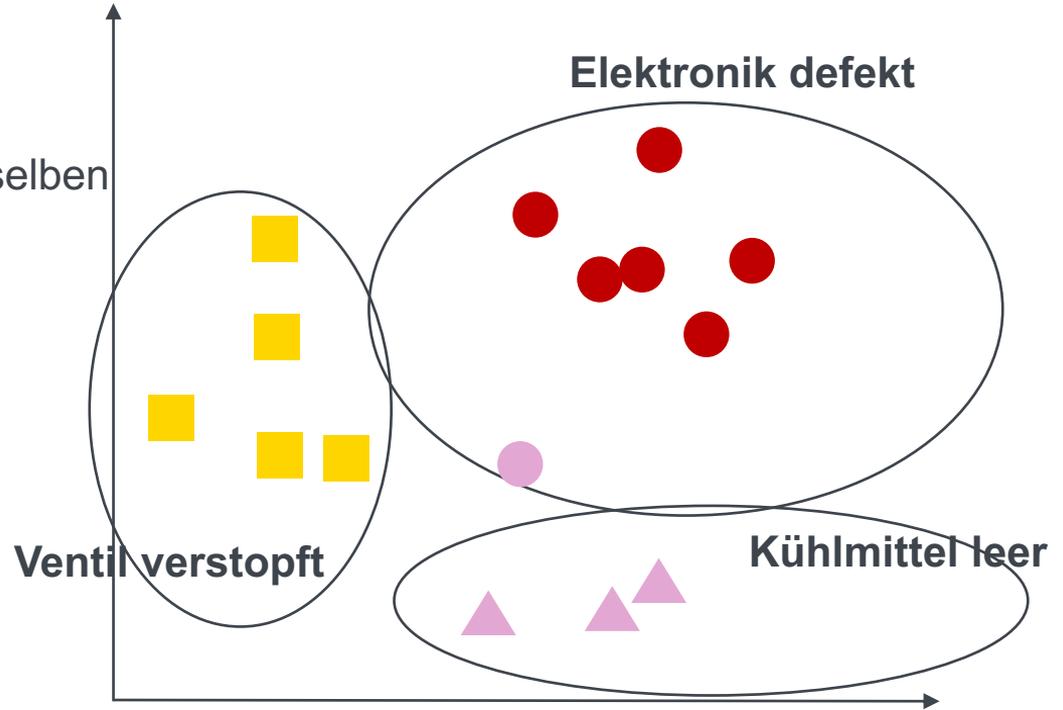
		Cluster	
		versch.	gleich
Fehler	verschieden		3
	gleich	6	28



Die Lösung: Paarweise Auswertung

- 15 Punkte: 105 Paare von Punkten
- Erwünscht:
 - Paare im selben Cluster, mit dem selben Fehler
 - Paare in verschiedenen Clustern, mit verschiedenen Fehlern

		Cluster	
		versch.	gleich
Fehler	verschieden	68	3
	gleich	6	28



Es gibt eine ganze Reihe von extrinsischen Evaluationsmaßen, die die korrekte Zuordnung mithilfe von Paaren von Punkten bewerten.

Dabei sollten Paare von Punkten derselben Klasse möglichst auch im selben Cluster liegen und umgekehrt.

Rand Index

- Wie Accuracy bei der Klassifikation
- Anzahl der korrekten Fälle geteilt durch Anzahl aller Fälle

$$\frac{68 + 28}{68 + 28 + 3 + 6} = \frac{96}{105} = 0,91$$

Rand Index 0,91

		Cluster	
		verschieden	gleich
Fehler	verschieden	68	6
	gleich	3	28

Rand Index

- Wie Accuracy bei der Klassifikation
- Anzahl der korrekten Fälle geteilt durch Anzahl aller Fälle

$$\frac{68 + 28}{68 + 28 + 3 + 6} = \frac{96}{105} = 0,91$$

		Paar im selben Cluster	
		negativ	positiv
Gleicher Fehler	negativ	richtig negativ	falsch positiv
	positiv	falsch negativ	richtig positiv

Selbe Formel wie Accuracy bei binärer Klassifikation

Rand Index

$$RI = \frac{RN + RP}{RN + RP + FN + FP}$$

Paare von Punkten, die im selben Cluster liegen und derselben Klasse angehören, nennt man „richtig positiv“ (RP).

Paare von Punkten, die in verschiedenen Clustern liegen und verschiedenen Klassen angehören, gelten als „richtig negativ“ (RN).

Paare von Punkten, die im selben Cluster liegen, aber verschiedenen Klassen angehören, nennt man „falsch positiv“ (FP).

Paare von Punkten, die in verschiedenen Clustern liegen, aber derselben Klasse angehören, heißen „falsch negativ“ (FN).

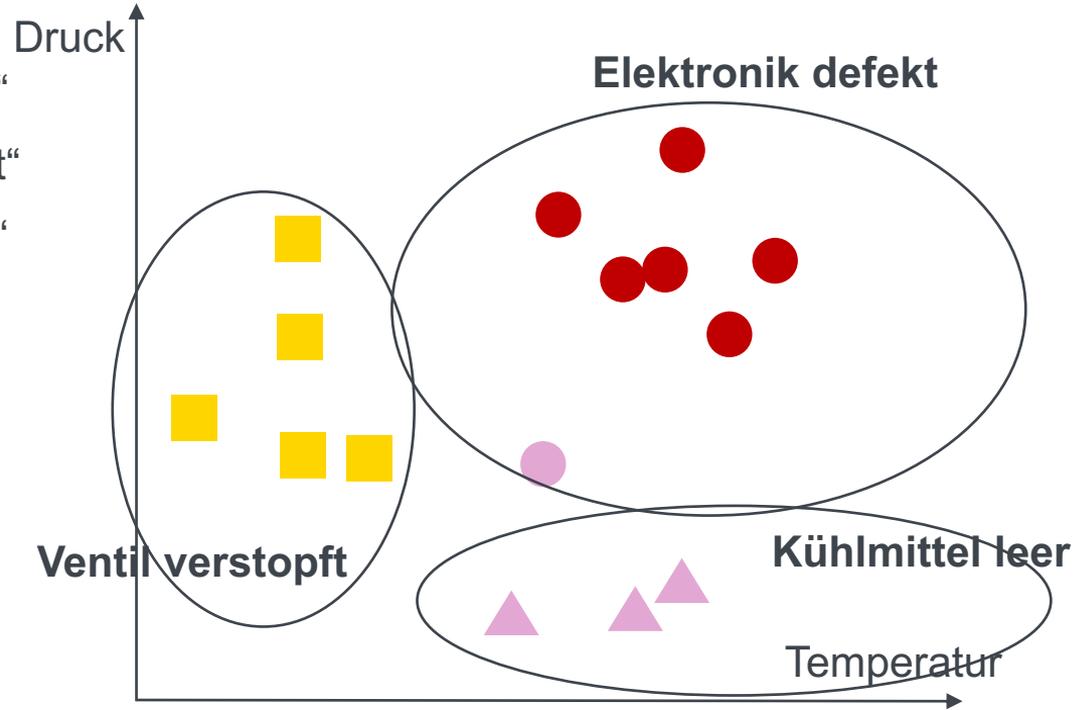
Der Rand Index wird mit derselben Formel berechnet wie die Accuracy bei der binären Klassifikation:

$$\text{RI} = \frac{\text{RP} + \text{RN}}{\text{RP} + \text{RN} + \text{FP} + \text{FN}}$$

Er erfasst also die Anzahl der „richtigen“ Fälle, geteilt durch die Anzahl aller Fälle.

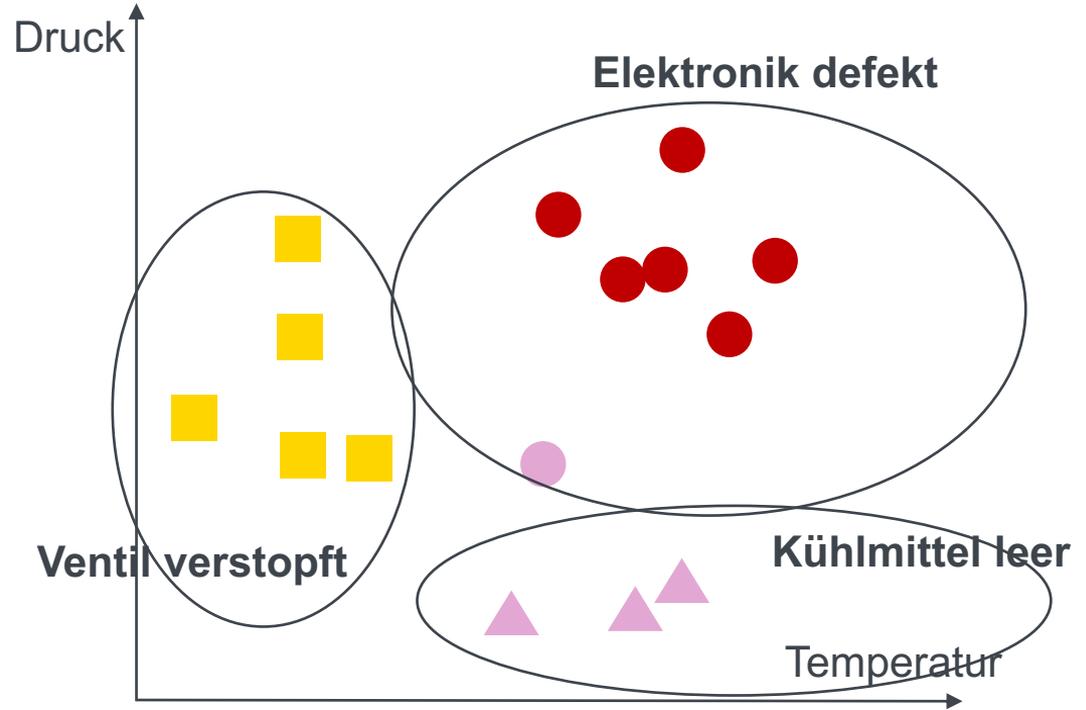
Purity

- Suche häufigste Klasse im Cluster
 - Gelb: 5 Fälle von „Ventil verstopft“
 - Rot: 6 Fälle von „Elektronik defekt“
 - Rosa: 3 Fälle von „Kühlmittel leer“
- Summe
 - 14 Fälle
- Punkte insgesamt
 - 15 Fälle
- Purity (Reinheit) = $\frac{14}{15} = 0.93$



Kritische Bewertung

- Clusteranalyse sollte nur dann zum Einsatz kommen, wenn das Ziel **NICHT** darin liegt, die Daten zu klassifizieren
- Denn: Überwachte Lernverfahren sind viel besser, wenn Klassen gelernt werden sollen!
- Clusteranalyse eher zum Entdecken neuer Klassen
- ⇒ Abgleich mit bekannten Klassen dann weniger sinnvoll



Evaluierung bei der Clusteranalyse

Intrinsische Evaluierung

- Bewertung konkreter Ergebnisse bei der Clusteranalyse (z.B. ob die Cluster gut voneinander abgegrenzt sind)
- Vergleich verschiedener Durchläufe einer Clusteranalyse (z.B. bei unterschiedlichen Initialisierungen oder unterschiedlichen Clusteranzahlen)
- Vergleich verschiedener Clusterverfahren für konkrete Daten

Extrinsische Evaluierung

- Vergleich verschiedener Clusterverfahren (z.B. hinsichtlich ihrer Eignung für bestimmte Arten von Daten)
- Weniger für Bewertung konkreter Ergebnisse (wenn Klassen vorliegen, lieber gleich überwachte Verfahren verwenden)

Dr. Antje Schweitzer

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung



Universität Stuttgart

Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering



IHK Industrie- und Handelskammer
Reutlingen

Reutlingen | Tübingen | Zollernalb



IHK Region Stuttgart



IHK Industrie- und Handelskammer
Karlsruhe



Lizenzbestimmungen

“Clusteranalyse – Evaluierung, Teil 2: Extrinsische Evaluierung” von Antje Schweitzer, KI B³ / Uni Stuttgart

Das Werk - mit Ausnahme der folgenden Elemente:

- Logos der Verbundpartner und des Förderprogramms
- im Quellenverzeichnis aufgeführte Medien

ist lizenziert unter:

 [CC BY 4.0 \(https://creativecommons.org/licenses/by/4.0/deed.de\)](https://creativecommons.org/licenses/by/4.0/deed.de)

(Namensnennung 4.0 International)

Quellenverzeichnis

Titelfoto: [Omar Flores \(https://unsplash.com/de/@designedbyflores\)](https://unsplash.com/de/@designedbyflores), ohne Titel, auf [Unsplash \(https://unsplash.com/de/fotos/blaue-rote-und-weisse-grafiken-IQT_bOWtysE\)](https://unsplash.com/de/fotos/blaue-rote-und-weisse-grafiken-IQT_bOWtysE), lizenziert unter [Unsplash-Lizenz \(https://unsplash.com/license\)](https://unsplash.com/license). Bildausschnitt verändert.