Sebastian Büsch[1], Volker Nissen[2], Arndt Wünscher[3]

X-CASE Experts in Consulting and Software Engineering

Automatische Klassifizierung von Data-Warehouse-Daten für das Information Lifecycle Management

1 - X-CASE GmbH

Albert-Einstein-Straße 3
98693 Ilmenau / Thür.
Tel.: +49 3677 20 88 0
Fax: +49 3677 20 88 29
E-Mail: Sebastian.Buesch@x-case.de
Internet: www.x-case.de

2 - University of Technology Ilmenau

Faculty of Economic Sciences and Media
Institute of Business & Information Systems Engineering
P.O. Box 10 05 65
D-98684 Ilmenau

Tel.: +49 3677 69-4043
Fax: +49 3677 69-4219
E-Mail: volker.nissen@tu-ilmenau.de
Internet: http://www.tu-ilmenau.de/wid

3 - University of Technology Ilmenau, Institute of Business & Information Systems Engineering

Gliederung

X-CASE Experts in Consulting and Software Engineering

- Problemstellung
- Zielstellung und Forschungsfragen
- Methodik

- Literaturanalyse zu anwendbaren Data-Mining-Verfahren
- Durchführung der Data-Mining-Verfahren
- Auswahl und Evaluation

- Fazit

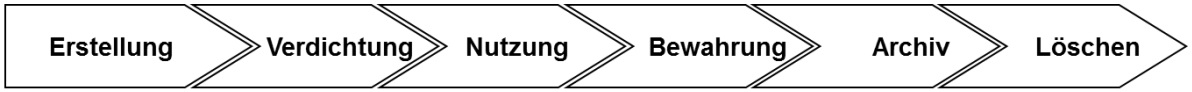
Problemstellung

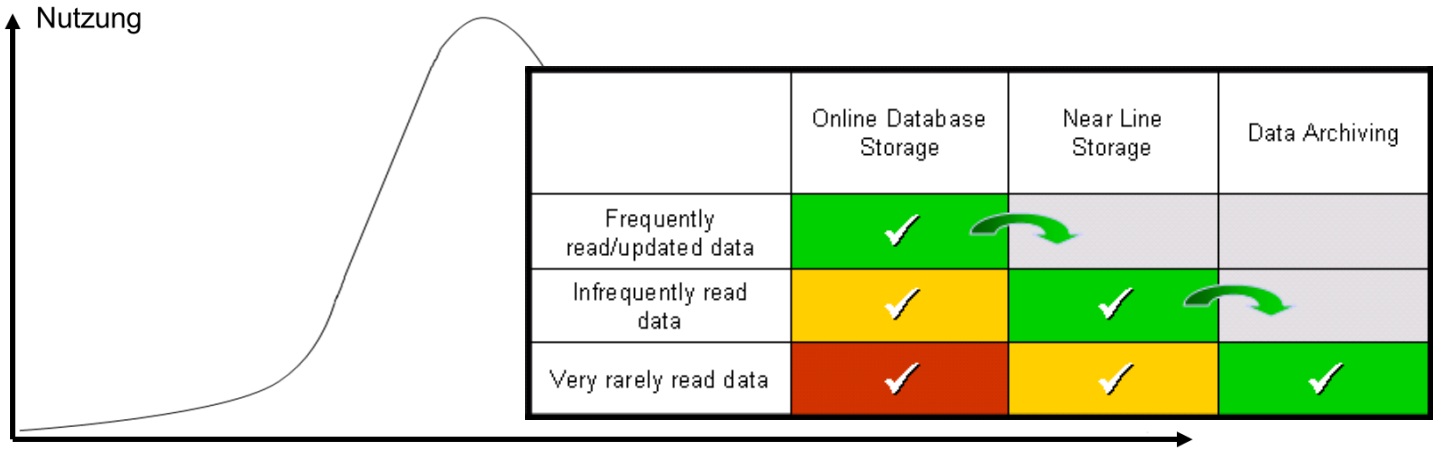
X-CASE Experts in Consulting and Software Engineering

- Data-Warehouse-Systeme (DWS) als Komponente zur Entscheidungsunterstützung im Rahmen der BI
- unterschiedliche Zugriffszeiten, SLAs oder Kosten etc. je nach verwendeten Speichermedien wie Archiv, Nearline Storage, In-Memory-Datenspeicherung usw.
- Information Lifecycle Management (ILM): aktives Datenmanagement anhand des Informationslebenszyklus primär zur Kostenreduktion
- Automatisierte Abbildung von Anwender- und Administratorenwissen zur Datenklassifikation im Sinne des ILM

Problemstellung

- Information Lifecycle Management (ILM):





Vgl. Hahne, M.: Information Lifecycle Management – Neue Architektur für das Data Warehouse. In (Gluchowski et al. Hrsg.): Schlaglichter der Wirtschaftsinformatik, GUC – Verlag, Chemnitz, 2007. S. 164.

- Beurteilung der Daten und deren Klassifizierung zentrale Herausforderung des ILM -> Identifikation geeigneter Kriterien
- Bisher vorrangig Zugriffsverläufe betrachtet
- Offen: Identifikation und Verwendung geeigneter DWS-Metadaten

Zielstellung und Forschungsfragen

X-CASE Experts in Consulting and Software Engineering

- Ziel: Entwicklung eines automatischen Klassifizierungsverfahrens für DWS-Daten nach ILM

- Forschungsfragen:
  - 1. Welche Kriterien eignen sich zur automatisierten Klassifizierung in DWS?
  - 2. Wie genau können Klassifizierungsverfahren eine Klassifizierung von Daten in DWS automatisiert vornehmen und damit Anwender- und Expertenwissen abbilden?
  - 3. Wie gut kann der entwickelte Klassifizierungsalgorithmus im Vergleich zu einfacheren Klassifikationsansätzen auf Basis von Datenzugriffen und Alter klassifizieren?

Methodik

X-CASE Experts in Consulting and Software Engineering

- Literaturanalyse zu anwendbaren Data-Mining-Verfahren nach Webster/Watson

  Jane Webster and Richard T. Watson. Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly, 26(2):xiii–xxiii, Juni 2002.

- Auswahl geeigneter Data-Mining-Verfahren durch Analogieschluss

- Anwendung Prozessmodell nach Fayyad et al.

  Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.: From data mining to knowledge discovery: An overview. In: Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (Hrsg.): Advances in knowledge discovery and data mining. Menlo Park et al. : AAAI Press, 1996, S. 1-34.

- Verfahrensauswahl und Evaluation am Fallbeispiel in der Automobilindustrie

Literaturanalyse zu anwendbaren Data-Mining-Verfahren

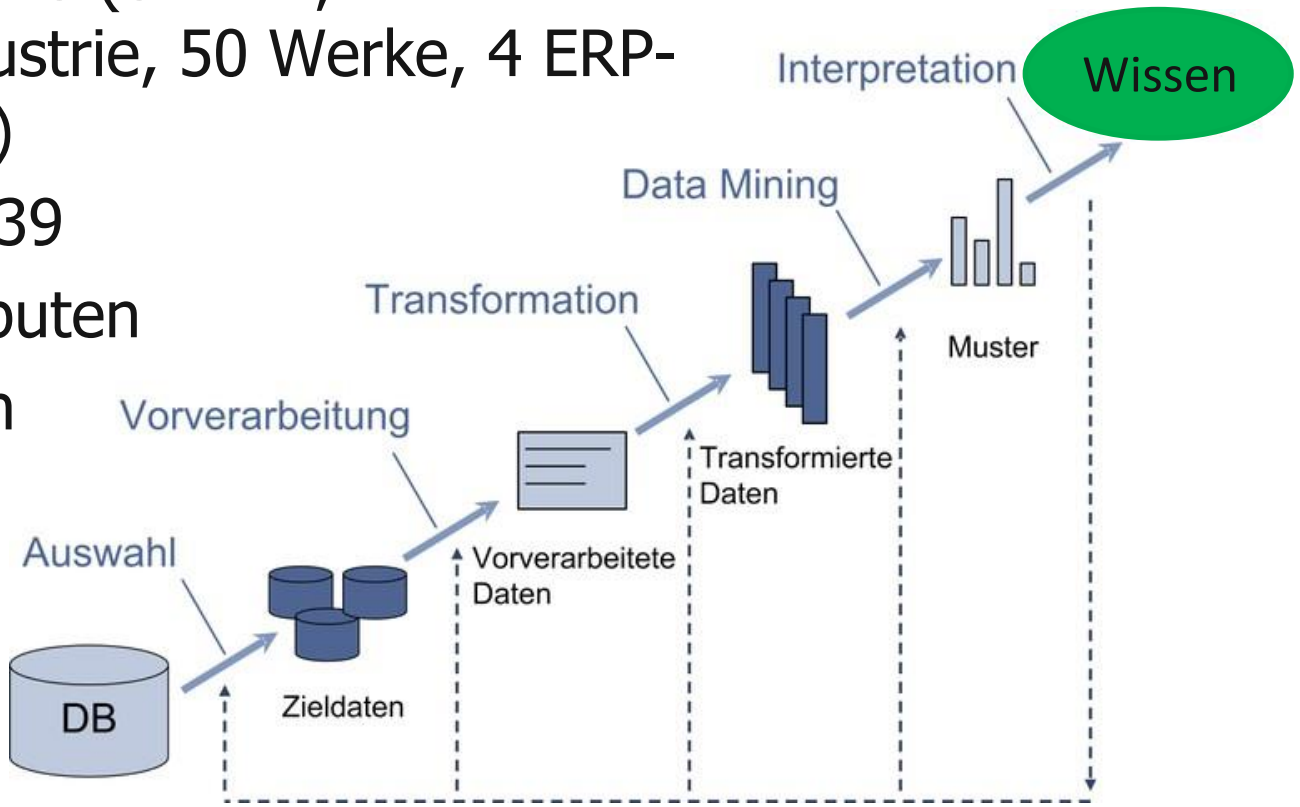
X-CASE Experts in Consulting and Software Engineering

- Suche in Journals, Datenbanken etc.
- Filterung auf Problemstellung
- 182 Fälle
- Auswahl:
  - MLP
  - SVM
  - Entscheidungsbaum

Klassen	Attribute	Instanzen	verwendetes Verfahren	Algorithmus	nominal data	numeric data	Genauigkeit	Quelle
5	41	260	SVM	SMO	x	x	93,48	(Chakchai et al., 2014, S. 92)
5	41	260	K-Nearest Neighbor	Ibk k= 5	x	x	93,076	(Chakchai et al., 2014, S. 92)
5	41	260	KNN	MLP	x	x	92,308	(Chakchai et al., 2014, S. 92)
5	41	260	Entscheidungsbaum	C4.5	x	x	90,386	(Chakchai et al., 2014, S. 92)
4	18	148	SVM	SMO	x	x	85,94	(Kotsiantis, et al., 2007)
5	41	260	Regelbasiert	RIPPER	x	x	85,314	(Chakchai et al., 2014, S. 92)
5	13	294	KNN	MLP	x	x	84,16	(Kotsiantis, et al., 2007)
5	13	294	Naive Bayes		x	x	83,95	(Kotsiantis, et al., 2007)
5	13	294	SVM	SMO	x	x	83,26	(Kotsiantis, et al., 2007)
4	18	148	Naive Bayes		x	x	83,13	(Kotsiantis, et al., 2007)
5	13	294	K-Nearest Neighbor	3NN	x	x	82,33	(Kotsiantis, et al., 2007)
4	18	148	KNN	MLP	x	x	82,26	(Kotsiantis, et al., 2007)
7	25	205	Entscheidungsbaum	C4.5	x	x	81,77	(Kotsiantis, et al., 2007)
4	18	148	K-Nearest Neighbor	3NN	x	x	81,74	(Kotsiantis, et al., 2007)
7	19	219	Naive Bayes		x	x	81,06	(Kotecha, et al., 2011)
5	13	294	Entscheidungsbaum	C4.5	x	x	80,22	(Kotsiantis, et al., 2007)
4	33	655	Entscheidungsbaum	C4.5	x	x	79,49	(Jantan, et al., 2011, S. 35)
4	33	655	K-Nearest Neighbor	K-Star	x	x	79,34	(Jantan, et al., 2011, S. 35)
5	13	294	Ripper Rule		x	x	79,26	(Kotsiantis, et al., 2007)
4	18	148	Ripper Rule		x	x	77,36	(Kotsiantis, et al., 2007)
5	41	260	Naive Bayes		x	x	76,154	(Chakchai et al., 2014, S. 92)
4	18	148	Entscheidungsbaum	C4.5	x	x	75,84	(Kotsiantis, et al., 2007)
7	25	205	Ripper Rule		x	x	74,05	(Kotsiantis, et al., 2007)
6	19	366	Genetic Programming	CGP	x	x	72,37	(Al-Madi und Ludwig, 2012, S. 82)
4	33	655	Entscheidungsbaum	Random Forest	x	x	72,21	(Jantan, et al., 2011, S. 35)
4	10	148	Genetic Programming	CF-GP	x	x	72,02	(Al-Madi und Ludwig, 2012, S. 82)
6	19	366	Genetic Programming	CF-GP	x	x	71,34	(Al-Madi und Ludwig, 2012, S. 82)
6	19	366	Genetic Programming	GP	x	x	71,33	(Al-Madi und Ludwig, 2012, S. 82)
4	10	148	Genetic Programming	CGP	x	x	71,17	(Al-Madi und Ludwig, 2012, S. 82)
4	33	655	KNN	RBFN	x	x	70,53	(Jantan, et al., 2011, S. 35)
4	10	148	Genetic Programming	GP	x	x	70,32	(Al-Madi und Ludwig, 2012, S. 82)
4	33	655	KNN	MLP	x	x	70,25	(Jantan, et al., 2011, S. 35)
4	10	148	Entscheidungsbaum	SGP	x	x	70,15	(Al-Madi und Ludwig, 2012, S. 82)
7	25	205	K-Nearest Neighbor	3NN	x	x	67,23	(Kotsiantis, et al., 2007)
6	19	366	Entscheidungsbaum	SGP	x	x	64,67	(Al-Madi und Ludwig, 2012, S. 82)
7	25	205	Naive Bayes		x	x	57,41	(Kotsiantis, et al., 2007)
7	25	205	SVM	SMO	x	x	56,55	(Kotsiantis, et al., 2007)
7	25	205	KNN	MLP	x	x	48,84	(Kotsiantis, et al., 2007)

Durchführung der Data-Mining-Verfahren

- Anhand Prozessmodell des KDD nach Fayyad et al.
- Fallbeispiel DWS (SAP BI; Automobilindustrie, 50 Werke, 4 ERP-Quellsysteme)
- Auswahl von 39 aus 115 Attributen
- 159 Instanzen
- 4 Klassen



Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.: From data mining to knowledge discovery: An overview. In: Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (Hrsg.): Advances in knowledge discovery and data mining. Menlo Park et al. : AAAI Press, 1996, S. 1-34.

Durchführung der Data-Mining-Verfahren

X-CASE Experts in Consulting and Software Engineering

- Daten-reduktion durch Attribut-selektion

Nr.	Attribute (eines Datenobjekts)	MLP-Wrapping	SMO-Wrapping	LR-Wrapping	NB Wrapping	C4.5-Wrapping	CFS	Chi-Squared Top 20	Relief Top 20
1	Tage seit der letzten Änderung	x	x	x	x	x	x	11	6
2	Alter in Tagen								14
3	Tage seit letzter Verwendung						x	13	13
4	Größe in KiloByte						x	14	
5	Anzahl Zugriffe gesamt			x		x	x	6	
6	Typ des Datenobjekts				x				19
7	Tage seit letztem Zugriff							12	12
8	Ebene nach [HS09]								4
9	empfohlene Speicherebene nach [HS09]					x		7	2
10	erhält das Objekt Daten?			x					7
11	werden Daten aus dem Objekt fortgeschrieben?				x				16
12	existieren Queries auf dem Datenobjekt?	x		x					11
13	mit Quellsystem verbunden?								18
14	Ebene nach Oßmann, 2008								10
15	Zugriffswahrscheinlichkeit nach [HS09]								3
16	ILM-Wert nach [HS09]	x	x	x	x	x		1	9
17	Anzahl Zugriffe (nur multidimensionale Abfragen)			x				4	
18	Zugriffswert nach [HS09]		x					19	5
19	Bereich		x						15
20	Anzahl Zugriffe (nur lesend für ETL)			x				5	
21	Zugriffe pro Tag						x	3	
22	Klasse des Datenobjekts							8	1
23	Verbindungsgrad		x				x	15	17
24	durschnittliche Querylaufzeit gesamt			x					
25	durschnittliche Querylaufzeit (nur Datenbank)								
26	Anzahl Querys auf dem Objekt	x	x					17	
27	Anzahl eingehende Verbindungen von anderen Objekten								
28	Anzahl ausgehende Verbindungen zu anderen Objekten		x						
29	Anzahl ausgehende Verbindungen und Querys						x		
30	Anzahl Nutzer, die auf das Objekt zugegriffen haben	x						16	
31	durschschnittliche Anzahl selektierter DB-Sätze pro Zugriff								
32	durschschnittliche Anzahl Datensätze im Bericht pro Zugriff								
33	durschnittliches Verdichtungsverhältnis (=Nr.31/Nr.32)		x					18	
34	durch. Tage seit der letzten Beladung							20	
35	durch. Tage seit der letzten Beladung (max. 1 pro Tag)								
36	Beladungen pro Tag	x					x	9	
37	Anzahl Beladungen gesamt	x						10	
38	Informationslebenszyklustyp		x	x	x	x		2	8
39	Navigationsintensität	x							20
	Summe	8	9	9	5	5	8	20	20
	Zeitdauer zur Ermittlung des Subsets in Sekunden	1820	120	90	5	4	<1	<1	<1

Durchführung der Data-Mining-Verfahren

X-CASE Experts in Consulting and Software Engineering

- Genauigkeit und AUC-Wert
- Überprüfung der relativen Attribut-wichtigkeit
- „Entfernen und Zurücklegen“
- Pruning im Entscheidungsbaum

	Entscheidungsbaum		SVM		MLP	
Datensatz	Genauigkeit	AUC	Genauigkeit	AUC	Genauigkeit	AUC
Full	74,20%	0,88	83,60%	0,91	81,10%	0,94
MLP-Wrapping	80,50%	0,87	75,50%	0,87	**83,60%**	0,94
SVM-Wrapping	80,50%	0,88	83,00%	0,91	82,40%	0,95
LR-Wrapping	83,60%	0,89	74,80%	0,85	81,10%	0,93
NB Wrapping	84,30%	0,91	76,70%	0,85	78,60%	0,90
C4.5-Wrapping	**86,80%**	0,92	73,00%	0,82	81,10%	0,92
CFS	79,20%	0,88	70,40%	0,84	78,60%	0,92
Chi-Squared Top20	82,40%	0,90	79,20%	0,87	80,50%	0,95
Relief TOP20	**86,80%**	**0,93**	**86,20%**	**0,93**	80,50%	**0,94**

		MLP	
Datensatz	fehlendes Attribut	Genauigkeit	AUC-Wert
MLP-Wrapping		83,60%	0,938

		MLP	
Datensatz	fehlendes Attribut	TP Rate	AUC-Wert
MLP-Wrappin			
MLP-Wrappin			
MLP-Wrappin			
MLP-Wrappin			
MLP-Wrappin			
MLP-Wrappin			
MLP-Wrappin			

		MLP	
Datensatz	fehlendes Attribut	TP Rate	AUC-wert
MLP5		86,20%	0,952
MLP5	Tage seit der letzten Änderung	73,60%	0,859
MLP5	existieren Queries auf den Infoprovider?	76,70%	0,89
MLP5	Anzahl Querys auf dem IP	81,80%	0,939
MLP5	Anzahl Nutzer mit realem Zugriff auf den IP	79,20%	0,919
MLP5	Beladungen pro Tag	85,50%	0,94

Durchführung der Data-Mining-Verfahren

X-CASE Experts in Consulting and Software Engineering

- Subsets nach der Attributselektion -> Forschungsfrage 1

Attribute (eines Datenobjekts)	Beschreibung / Wertebereich / Beispielausprägungen	C45	SVM18	MLP5	Anzahl
Tage seit der letzten Änderung	Zahlenwert	x	x	x	3
Existieren Queries auf dem Datenobjekt?	ja / nein; Querys sind multidimensionale Berichtsdefinitionen	x	x	x	3
Alter in Tagen	Zahlenwert	x	x		2
Tage seit letztem Zugriff	Zahlenwert	x	x		2
Zugriffswert nach Heinrich und Stelzer (2009)	numerischer, berechneter Wert anhand der Zugriffe	x	x		2
Bereich	Fachbereich / Organisationseinheit, z.B. Vetrieb, Controlling	x	x		2
Verbindungsgrad	numerischer, berechneter Wert anhand der Verbindungen zu anderen Datenobjekten, vgl. Heinrich und Stelzer (2009)	x	x		2
Navigationsintensität	beschreibt, wie oft in einem Bericht flexibel navigiert wird; Navigationsintensität von 1 bedeutet, Berichte werden nur starr ohne OLAP-Funktionen verwendet; ein Wert > 1 bedeutet einen zunehmenden Anteil flexibler Nutzung	x	x		2
Tage seit letzter Verwendung	letzte Lese- oder Schreiboperation		x		1
Typ des Datenobjekts	multidimensionaler Cube, flaches Datenobjekt		x		1
Ebene nach Heinrich und Stelzer (2009)	vgl. Heinrich und Stelzer (2009)		x		1
Empfohlene Speicherebene nach Heinrich und Stelzer (2009)	vgl. Heinrich und Stelzer (2009)		x		1
Werden Daten aus dem Objekt fortgeschrieben?	ja / nein		x		1
Mit Quellsystem verbunden?	ja / nein		x		1
Ebene nach Oßmann (2008)	vgl. Oßmann (2008)		x		1
Zugriffswahrscheinlichkeit nach Heinrich und Stelzer (2009)	vgl. Heinrich und Stelzer (2009)		x		1
Klasse des Datenobjekts	z.B. Extraktionsebene, Transformationsebene, Datenbereitstellungsebene, Berichtsebene		x		1
Anzahl Querys auf dem Objekt	Zahlenwert			x	1
Anzahl Nutzer, die auf das Objekt zugegriffen haben	Zahlenwert			x	1
Datenbeladungen pro Tag	Zahlenwert			x	1
Informationslebenszyklustyp	z.B. fallender Zugriffsverlauf, konstante Zugriffe über die Zeit		x		1

Durchführung der Data-Mining-Verfahren

X-CASE Experts in Consulting and Software Engineering

- Ergebnisse der Klassifizierung -> Forschungsfrage 2

Entscheidungsbaum

Loeschen	Offline	Nearline	Online	<-- classified as
64	3	0	1	**Loeschen**
3	35	0	3	**Offline**
0	1	3	4	**Nearline**
2	1	0	39	**Online**

SVM

Loeschen	Offline	Nearline	Online	<-- classified as
64	2	0	2	**Loeschen**
3	33	0	5	**Offline**
0	1	4	3	**Nearline**
1	0	2	39	**Online**

MLP

Loeschen	Offline	Nearline	Online	<-- classified as
63	4	0	1	**Loeschen**
0	39	0	2	**Offline**
0	2	0	6	**Nearline**
1	3	0	38	**Online**

Auswahl und Evaluation

X-CASE Experts in Consulting and Software Engineering

- Bewertung der Verfahren anhand von Fehlerpunkten
- berücksichtigt Grad der Fehleinschätzung
  - 1 Fehlerpunkt bei angrenzenden Klassen; + 1 Fehlerpunkt pro Klasse dazwischen
  - Klassen: online -> nearline -> offline -> löschen
  - Fehlklassifizierung als „löschen“ + 1 Fehlerpunkt
- Bsp. SVM (Klasse offline):

$$33 * 0 + 3 * 2 + 5 * 2 = 16$$

Loeschen	Offline	Nearline	Online	<-- classified as
64	2	0	2	**Loeschen**
3	33	0	5	**Offline**
0	1	4	3	**Nearline**
1	0	2	39	**Online**

Auswahl und Evaluation

X-CASE Experts in Consulting and Software Engineering

- Vergleich mit anderen Verfahren -> Forschungsfrage 3

- Alternative 1: nach je 90 Tagen Verlagerung auf das nächste Speichermedium (online -> nearline -> offline -> löschen)
- Alternative 2: Verteilung der Objekte nach beobachteter Zugriffshäufigkeit (top 42 online, 8 NLS, 41 offline, 68 Objekte mit den geringsten Zugriffen löschen)

	Ent.Baum	SVM	MLP	Alt. 1 (Alter)	Alt. 2 (Zugriffe)
Genauigkeit	88,7 %	88,1 %	88,1 %	51,6 %	62,3 %
Fehlerpunkte	33	34	**29**	197	98

Fazit

X-CASE Experts in Consulting and Software Engineering

- Automatisierte Datenbewertung für DWS/ILM durch Verfahren des maschinellen Lernens möglich
- Verbesserung der Ergebnisse durch Attributselektion
- Verschiedene Kriterien herangezogen, nur 2 Merkmale für alle Verfahren relevant (Tage seit letzter Änderung, existieren Querys?)
- Bestes Verfahren (MLP) erreicht Genauigkeit von 88,1%
- Verfahren deutlich besser als triviale Alternativen
- DWS-Metadaten sollten zur Bewertung verwendet werden
- Entwickelte Verfahren grundsätzlich übertragbar – eine Fallstudie außerhalb Automotive zu empfehlen

VIELEN DANK FÜR IHR INTERESSE