KNIME TUTORIAL

Anna Monreale KDD-Lab, University of Pisa

Email: annam@di.unipi.it

What is KNIME?

- KNIME = Konstanz Information Miner
- Developed at University of Konstanz in Germany
- Desktop version available free of charge (Open Source)
- Modular platform for building and executing workflows using predefined components, called nodes
- Functionality available for tasks such as standard data mining, data analysis and data manipulation
- Extra features and functionalities available in KNIME by extensions
- Written in Java based on the Eclipse SDK platform

KNIME resources

- Web pages containing documentation
 - <a>www.knime.org tech.knime.org tech.knime.org
 - installation-0
- Downloads
 - knime.org/download-desktop
- Community forum
 - tech.knime.org/forum
- Books and white papers
 - knime.org/node/33079

Installation and updates

- Download and unzip KNIME
 - No further setup required
 - Additional nodes after first launch
- Workflows and data are stored in a *workspace*
- New software (nodes) from update sites
 - <u>http://tech.knime.org/update/community-contributions/</u> <u>realease</u>



PRODUCTS / APPLICATIONS / PARTNERS / SERVICES / RESOURCES / COMPANY

>

You are here: / Home / Download KNIME Desktop & SDK





Download KNIME Desktop & SDK

Download the latest KNIME Deskop and KNIME SDK version 2.8.2 for Windows, Linux, and Mac OS X.

KNIME Desktop

The KNIME Desktop version is intended for end users and provides everything needed to immediately begin using KNIME as well as extend KNIME with extension packages developed by others. The downloads also contain the KNIME quickstart guide.

Windows

Usually unzipping the archive somewhere on your hard drive is sufficient for the installation of KNIME. However, under Windows problems with the built-in unzip utility sometimes truncate file names. Therefore we offer self extracting archives:

- KNIME for Windows 32bit (self-extracting archive)
- KNIME for Windows 64bit (self-extracting archive)

If you are using a proper unzipper and want to use zip archives instead, you can find them here.

Linux

For Linux a 32 and 64bit build are available:

- KNIME for Linux 32bit
- KNIME for Linux 64bit

Mac OS X

Since KNIME 2.3.0 we are proud to announce a fully supported KNIME build for Mac OS X. It requires a 64bit Intelbased architecture with Java 1.6:

Workspace

- The workspace is the directory where all your workflows and preferences are saved in the next KNIME session.
- The workspace directory can be located anywhere on your hard-disk.
- By default, the workspace directory is "[KNIME]
 \workspace". But, you can change it, by changing the path requested at the beginning, before starting the KNIME working session.



Download Extensions

- From the Top Menu, select
 Help -> Software Updates
- In the "Software Updates" window, select Tab Available Software
- Open the sites and select the extensions
- Click the **Install** button on the top right
- Restart KNIME
- In the Node Repository you can see the new nodes

software opuates and add-ons		
Installed Software Available Software		
hune filter text		
abbe used and		10.0K
Name	Version	^
Image: Imag		Properties
KNIME Base Chemistry Types & Nodes	2.2.0.0026420	
KNIME CDK Chemistry Feature	2.2.0.0026420	
KNIME Chemistry Add-Ons	2.2.0.0026420	Add Ste
KNIME Core	2.2.0.0026430	
KNIME Distance Matrix Feature	2.2.0.0026420	Manage Sites
KNIME External Tool Node	2.2.0.0026420	
KNIME HTML/PDF Writer Feature	2.2.0.0026420	
KNIME Itemset Mining	2.2.0.0026420	Refresh
KNIME JFreeChart	2.2.0.0026420	
KNIME LIBSVM integration	2.2.0.0026420	-
KNIME Math Expression Extension (JEP)	2.2.0.0026420	
KNIME Public Server View	2.2.0.0026420	
CDL KNIME Python Scripting	2.2.0.0026420	
CIV C KNIME R Statistics Integration	2.2.0.0026420	
KNIME R Statistics Integration (Binaries)	2.2.0.0026420	-
KNIME Weka Data Mining Integration	2,2,0,0026420	×
Show only the latest versions of available software		
Include items that have already been installed		
Open the 'Automatic Updates' preference page to set up an automatic upd	ate schedule.	
		Close

What can you do with KNIME?

Data manipulation and analysis

• File & database I/O, filtering, grouping, joining,

Data mining / machine learning

WEKA, R, Interactive plotting

Scripting Integration

• R, Perl, Python, Matlab ...

Much more

• Bioinformatics, text mining and network analysis

KNIME Workflow

- KNIME does not work with scripts, it works with workflows.
- A workflow is an analysis flow, which is the sequence of the analysis steps necessary to reach a given result:
 - 1. Read data
 - 2. Clean data
 - 3. Filter data
 - 4. Train a model



- KNIME implements its workflows graphically.
- Each step of the data analysis is executed by a little box, called a node.
- A sequence of nodes makes a workflow.

Import/export of workflow

- Workflows can be imported and exported as .zip files
 - With or without the underlying data
 - File → Import KNIME workflow...
 - File → Export KNIME workflow...





Create a new workflow group

🚵 Workflow Projects 🛛 🗘 🖗 🔲	⊈ ~ - □	🔺 0: KNIME_project 🛛	
🗉 📥 KNIME_project			🔺 Create new workflow group
New KNIME workflow			New Knime workflow group Creates a new workflow group, which allows for the grouping of several KNIME workflows.
Open		File Reader	
Copy			Name of the workflow group to create: Chapter2
Timport KNIME workflow		adult data set	
Favorite Not			
🗄 🕖 Last us 🔊 Configure			
Node Repos	- 8		
Cancel Execution	$\overline{\nabla}$		
 IO ID 	,		
🗄 🗟 Data Ma 📷 Edit Report			Finish Cancel
E ∑ Statistic Refresh			

Create a new workflow

Re Bar We Seek Ro He Image: See See See See See See See See See S	A KNIME	
Image: Section of the workflow Image: Section of the workflow to create: Image: Section of the workflow Image: Section of the workflow to create: Image: Section of the workflow Image: Section of the workflow to create: Image: Section of the workflow Image: Section of the workflow to create: Section of the workflow to create: Section of the work	File Edit View Search Run Help	
Image: Section	i 📸 • 🔜 🐘 i 🛷 • i ฎ - 🖓 • i 🎭 i 🗛 •	
Intervention Intervention Interventintervente Interventinte	🚵 Worlflow Projects 🗇 🗘 👰 😑 🤹 🎽 🗖	
Einish Cancel	Image: Number State workform Open Copy Passe Image: Number workform Image: Number workform.	New KNIME Workflow Wizard Create a new KNIME workflow. Name of the workflow to create: my_first_workflow Workspace destination: /Chapter2 Browse
		Finish Cancel

KNIME nodes: Overview

Node = basic processing unit of KNIME workflow which performs a particular task



Ports

- Data Port: a white triangle which transfers flat data tables from node to node
- Database Port: Nodes executing commands inside a database are recognized by their database ports (brown square)
- PMML Ports: Data Mining nodes learn a model which is passed to the referring predictor node via a blue squared PMML port



Database Connection Reader



Decision Tree Learner



Other Ports

- Whenever a node provides data that does not fit a flat data table structure, a general purpose port for structured data is used (dark cyan square).
- All ports not listed above are known as "unknown" types (gray square).



Node Creation



Node Operations



I/O Operations



ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes.

CSV (Comma-Separated Values) file stores tabular data (numbers and text) in plain-text form.

Read data from file



tings Flow	ariables Memor	ry Policy				
ter ASCII dat	a file location: (p	ress 'Enter' to upda	ate preview)-			
valid LIRL:	Hata/Liser%-20M	leeting/Templates/	adult%20data	%20set/adult.da	ta 🖌 🛛 Bro	W/SP
Tana orter	2010/0201102011	eccedent i compracest.		TOE DOOL OF OUR		
		Preserve	user settings (or new location		
		- rieserie	asor secongs i	or norriocación		
sic Settings-					_	
read row ID	s	Column del	imiter: ,	*	Adva	nced
read column	headers	🔽 ignore :	spaces and tal	s		
		lava-st	vie comments	Sin	ale line commen	
			,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,		gio into common	··
eview						
	Click column h	header to change o	olumn propert	ies (* = name/ty	pe user settings	;)
ow ID	l age	S workclass	🛔 fnlwgt	S education	educati	S marital
Row0	39	State-gov	77516	Bachelors	13	Never-married A
Row1	50	Self-emp-no	83311	Bachelors	13	Married-civ E
Row2	38	Private	215646	HS-grad	9	Divorced H
Row3	53	Private	234721	11th	7	Married-civ H
Row4	28	Private	338409	Bachelors	13	Married-civ P
Row5	37	Private	284582	Masters	14	Married-civ E
Row6	49	Private	160187	9th	5	Married-spo C
Row7	52	Self-emp-no	209642	HS-grad	9	Married-civ E
Row8	31	Private	45781	Masters	14	Never-married P
Row9	42	Private	159449	Bachelors	13	Married-civ E
Row10	37	Private	280464	Some-college	10	Married-civ E
Row11	30	State-gov	141297	Bachelors	13	Married-civ P
Row12	23	Private	122272	Bachelors	13	Never-married A
Row13	32	Private	205019	Assoc-acdm	12	Never-married 5
	40	Private	121772	Assoc-voc	11	Married-civ C
Row14	34	Private	245487	7th-8th	4	Married-civ T
Row14 Row15	1.00.00	Self-emp-no	176756	HS-grad	9	Never-married F
Row14 Row15 Row16	25	Drivento	186824	HS-grad	9	Never-married M
Row14 Row15 Row16 Row17	32	Privace		11 Likes	7	Married-civ S
Row14 Row15 Row16 Row17 Row18	25 32 38	Private	28887	1101		
Row14 Row15 Row16 Row17 Row18 Row19	25 32 38 43	Private Self-emp-no	28887 292175	Masters	14	Divorced E
Row14 Row15 Row16 Row17 Row18 Row19 Row20	25 32 38 43 40	Private Self-emp-no Private	28887 292175 193524	Masters Doctorate	14 16	Divorced E Married-civ P
Row14 Row15 Row16 Row17 Row18 Row19 Row20 Row21	25 32 38 43 40 54	Private Self-emp-no Private Private	28887 292175 193524 302146	Masters Doctorate HS-grad	14 16 9	Divorced E Married-civ P Separated C
Row14 Row15 Row16 Row17 Row18 Row19 Row20 Row21 Row22	25 32 38 43 40 54 35	Private Self-emp-no Private Private Federal-gov	28887 292175 193524 302146 76845	Masters Doctorate HS-grad 9th	14 16 9 5	Divorced E Married-civ P Separated C Married-civ F
Row14 Row15 Row16 Row17 Row18 Row19 Row20 Row21 Row22 Row23	25 32 38 43 40 54 35 43	Private Private Self-emp-no Private Private Federal-gov Private	28887 292175 193524 302146 76845 117037	Masters Doctorate H5-grad 9th 11th	14 16 9 5 7	Divorced E Married-civ P Separated C Married-civ F Married-civ T

Read data from file

- Click in the column name
 - Change column name
 - Change type

Dialog - 0:1	1 - File Reader				
e					
Settings class	Variables Menon	🛃 New setti	ngs for col	umn 🔀	
Piow	variables memory	Column Proper	ties		
Enter ASCII da	ata file location: (pr		de colore le c		
		DONTINEU	ide column in o	output table	
valid URL:	file:/C:/data/Use	•	Name: fir	al weight	adi 🗸 🛛 Browse
		1			
			Type: Strin	9 📉 h	
Davis California		wine cost			
basic Secongs		miss, valu	Je pattern:		
read row I	Ds				Advance
read colum	n headers		_		
				omain	ala basa sana ang
		L		pn	gie line comment:
			OK	Cancel	
Preview			- +		
	Click column h	eader to change (column propert	ties (* = name/typ	pe user settings)
Row ID	age	S workclast	rniwgt	education	education-num
Row0	39	State-gov	77516	Bachelors	13
Row1	50	Self-emp-no	83311	Bachelors	13
Row2	38	Private	215646	HS-grad	9
Row3	53	Private	234721	11th	7
Row4	28	Private	338409	Bachelors	13
Row5	37	Private	284582	Masters	14
Row6	49	Private	160187	9th	5
Row7	52	Self-emp-no	209642	HS-grad	9
Row8	31	Private	45781	Masters	14
Row9	42	Private	159449	Bachelors	13
Row10	37	Private	280464	Some-college	10
Row11	30	State-gov	141297	Bachelors	13
Row12	23	Private	122272	Bachelors	13
Row13	32	Private	205019	Assoc-acdm	12
	1.0				1 × ×

Table Data

Row	v ID	Colur	nn Head	ler				, <u> </u>		_				
L						Integ	zer		String					
							5							
			/		C	lata t	type		data type	2				
			/											
🔺 File	Table - (0:1 - File Rea	ader											×
File														
Table "ad	dult.data" ·	Rows: 32561	Spec - Columns	: 15 Propertie:	Flow Variat	bles			1					
Row ID		(age	S workclass	S final we	S education		lucati	S marital	S ocupa	S relation	S race	S sex	capital	
Row		39	State-gov	77516	Bachelors	13	N	lever-married	d Adm-clerical	Not-in-family	White	Male	2174	^
Row1	\supset	50	Self-emp-no	83311	Bachelors	13	M	larried-civ	. Exec-manag	Husband	White	Male	0	
Row2		38	Private	215646	HS-grad	9	D	ivorced	Handlers-cle	Not-in-family	White	Male	0	
Row3		53	Private	234721	11th	7	M	1arried-civ	. Handlers-cle	Husband	Black	Male	0	
Row4		28	Private	338409	Bachelors	13	M	larried-civ	. Prof-specialty	Wife	Black	Female	0	
Row5		37	Private	284582	Masters	14	M	larried-civ	. Exec-manag	Wife	White	Female	0	
Row6		49	Private	160187	9th	5	M	1arried-spo	. Other-service	Not-in-family	Black	Female	0	
Row7		52	Self-emp-no	209642	HS-grad	9	Μ	larried-civ	. Exec-manag	Husband	White	Male	0	
Row8		31	Private	45781	Masters	14	N	lever-married	d Prof-specialty	Not-in-family	White	Female	14084	
Row9	1	42	Private	159449	Bachelors	13	M	1arried-civ	. Exec-manag	Husband	White	Male	5178	
Row1	0	37	Private	280464	Some-college	10	M	1arried-civ	. Exec-manag	Husband	Black	Male	0	
Row1	1	30	State-gov	141297	Bachelors	13	M	larried-civ	 Prof-specialty 	Husband	Asian-Pac-Is	Male	0	
Row12	2	23	Private	122272	Bachelors	13	N	lever-marrie	d Adm-clerical	Own-child	White	Female	0	
Row1	3	32	Private	205019	Assoc-acdm	12	N	lever-married	d Sales	Not-in-family	Black	Male	0	
Row1	4	40	Private	121772	Assoc-voc	11	Μ	larried-civ	. Craft-repair	Husband	Asian-Pac-Is	Male	0	
Row1	5	34	Private	245487	7th-8th	4	M	larried-civ	. Transport-m	Husband	Amer-Indian	Male	0	
Row1	6	25	Self-emp-no	176756	HS-grad	9	N	lever-married	d Farming-fish	Own-child	White	Male	0	
Row1	7	32	Private	186824	HS-grad	9	N	lever-married	d Machine-op	Unmarried	White	Male	0	
Row1	8	38	Private	28887	11th	7	Μ	larried-civ	. Sales	Husband	White	Male	0	
Row1	9	43	Self-emp-no	292175	Masters	14	D	ivorced	Exec-manag	Unmarried	White	Female	0	~
		<				111							>	

Other input nodes: CSV Reader

Dialog - 0:11 - CSV Reader(FP-AR)	
CSV Reader Flow Variables Memory Policy /Users/annamonreale/Desktop/MAINS/ite Browse	

CSV Writer

		•
Vriter options:		
Write column he	ader	
Don't write colu	mn headers if file exists	
Write row ID		
Compress output	ut file (gzip)	
If file exists		
	<u> </u>	Abr

Data Manipulation

- Three main sections
 - **Columns**: binning, replace, filters, normalizer, missing values, ...
 - Rows: filtering, sampling, partitioning, ...
 - Matrix: Transpose

Statistics node

Statistics

- For all numeric columns computes statistics such as
- minimum, maximum, mean, standard deviation, variance, median, overall sum, number of missing values and row counts
- For all nominal values counts them together with their occurrences.

Options Flow Variables Memory Policy Calculate median values (computationally expensive) Nominal values Exclude Column(s): Select all search hits add all >> + Col0 + Col1 Add all >>		Dialog – 0:14 – Statistics	5
Calculate median values (computationally expensive) Nominal values Exclude Column(s): Select all search hits add all >> Column(s): Select all search hits i Col0 i Col1 Col2 Col3 Max no. of most frequent and infrequent values (in view): 20 \$ Max no. of possible values per column (in output table): 1,000 \$	Optio	ns Flow Variables Mem	nory Policy
Nominal values Exclude Column(s): Select all search hits add all >> Col0 Col1 Col2 Col3 Col3 Max no. of most frequent and infrequent values (in view): 20 Max no. of possible values per column (in output table): 1,000	Calculate	e median values (computatio	onally expensive)
Column(s): Search Select all search hits add >> Col0 Col1 Col2 Col2	Nominal values	- Select	Include
i Col0 add all >> i Col1 i Col2 i Col2 i Col2 i Col2 i Col2 i Col3 i Col3 Max no. of most frequent and infrequent values (in view): 20 (‡) Max no. of possible values per column (in output table): 1,000 (‡)	Column(s): Search	add >>	Column(s): Search
Max no. of most frequent and infrequent values (in view): 20 (‡) Max no. of possible values per column (in output table): 1,000 (‡)	i Col0 i Col1	add all >> </td <td>i Col2 i Col3</td>	i Col2 i Col3
Max no. of most frequent and infrequent values (in view): 20 (‡) Max no. of possible values per column (in output table): 1,000 (‡)		<	
Max no. of possible values per column (in output table): 1,000	Max no. of most frequen	t and infrequent values (in vi	iew): 20 🔹
	Max no. of possible valu	es per column (in output tab	ble): 1,000 (*)

Correlation Analysis

- Linear Correlation node computes for each pair of selected columns a correlation coefficient, i.e. a measure of the correlation of the two variables
 - Pearson Correlation Coefficient
- Correlation Filtering node uses the model as generated by a Correlation node to determine which columns are redundant (i.e. correlated) and filters them out.
 - The output table will contain the reduced set of columns.



Linear Correlation Correlation Filter

Data Views

- Box Plots
- Histograms, Pie Charts, Scatter plots, ...
- Scatter Matrix

Mining Algorithms

- Clustering
 - Hierarchical
 - K-means
 - Fuzzy –c-Means
- Decision Tree
- Item sets / Association Rules
 - Borgelt's Algorithms (Extension)
- Weka (Extension)

EXERCISES

Anna Monreale

KDD-Lab, University of Pisa

Email: annam@di.unipi.it

DATA MANIPULATION

Anna Monreale KDD-Lab, University of Pisa Email: annam@di.unipi.it

Data Manipulation

See Workflow on the course website

http://didawiki.cli.di.unipi.it/doku.php/dm/mains.santanna.dm4crm.2012

MARKET BASKET ANALYSIS

Anna Monreale KDD-Lab, University of Pisa Email: annam@di.unipi.it

Market Basket Analysis

 Problem: given a database of transactions of customers of a supermarket, find the set of frequent items copurchased and analyze the association rules that is possible to derive from the frequent patterns.

Frequent Patterns and AR in KNIME

• Use the nodes implementing the Borgelt's Algorithms:



• Item Set Finder node provides different algorithms:

- Apriori (Agrawal et al. 1993)
- FPgrowth (frequent pattern growth, Han et al 2000)
- RElim (recursive elimination)
- SaM (Split and Merge)
- JIM (Jaccard Item Set Mining)
- AR Learner uses Apriori Algorithm

- Given the output of the Item set Finder node sometimes you cannot see all the components of the itemset
 - we need to transform it in a string and
 - then, we can also write the result in a file



• First we need to split the collection

				🔲 💛 💟 Dialog – 0:38 – Split Collection Column
CSV Reader	Create Collection Column	ltem Set Finder (Borgelt)	Split Collection Column	Settings Flow Variables Memory Policy
FP-AR_Product	Node 31	Node 29	Node 38	() ItemSet Replace input column Determine most specific type Element Count Policy Best effort Use input table information Count in advance

 Second we combine the columns that have to compose the itemset (string)

CSV Reader	Create Collection Column	ltem Set Finder (Borgelt)	Split Collectio Column	n Column Combiner			
	—-> <mark>}∰</mark> >—	—⊳ <mark>@</mark> ⊳—	> <mark> </mark> >	> <mark>}}</mark> >			
FP-AR_Product	Node 31	Node 29	Node 38	Create Itemset as String			
					Dia	log – 0:39 – Column	Combiner
					Settings	Flow Variables	Memory Policy
				Delimiter	3		
				💿 Quote Character	" 🗌 Quote alv	vays	
				O Replace Delimiter by			
				Name of appended column	Itemset		
				Exclude		Select	Include
				Column(s):	Search	add >>	Column(s): Search
				Select all search hits			Select all search hits
				ltemSetSize		add all >>	For Split Value 1
				I ItemSetSupport			I Split Value 2
						< < remove	Split Value 3
							I Split Value 5
						< remove all	Split Value 6

?

Cancel

Apply

OK

- The combiner does not eliminate the missing values "?"
- The combined itemsets contain a lot of "?"



OK

Apply

Cancel

?

Before writing in a file eliminate the split columns



.. The output table

File

Filtered table - 0:41 - Column Filter(Elimine some columns)

F	il	e	

		Table "default"	- Rows: 139122	Spec - Columns: 4 Properties Flow Variables
Row ID	$ +\rangle$	▼ ItemSetSize ItemS	etSupport D Relative	ltemSetSup S Itemset
Row94237	7	3	0.03	16864,30459,233740,15786,265109,311540,85800
Row102226	7	3	0.03	253300,7697,45168,15506,36369,72989,85800
Row35465	6	3	0.03	39071,68523,14635,31560,75153,85800
Row63365	6	3	0.03	228263,38950,37860,76174,65616,224434
Row63811	6	3	0.03	2334354,76174,265109,31560,75153,85800
Row65867	6	3	0.03	52006,265111,221614,265109,75153,85800
Row68210	6	3	0.03	31555,14845,45168,31560,85800,75153
Row72720	6	3	0.03	287124,236490,243821,75153,31560,85800
Row78817	6	3	0.03	30958,7697,257536,25227,228164,56674
Row81349	6	3	0.03	27008,30459,65125,16722,48067,265109
Row84546	6	3	0.03	269468,30459,233740,52769,265109,311540
Row84610	6	3	0.03	269468,233740,16281,48067,265109,85800
Row86734	6	3	0.03	28467,16281,72989,221614,31560,75153
Row89111	6	3	0.03	26308,15506,243821,31560,75153,85800
Row89246	6	3	0.03	76288,40287,56674,48067,75153,265109
Row90026	6	3	0.03	2335012,67463,68523,221614,265109,85800
Row94238	6	3	0.03	16864,30459,233740,15786,265109,311540
Row94239	6	3	0.03	16864,30459,233740,15786,265109,85800
Row94241	6	3	0.03	16864,30459,233740,15786,311540,85800
Row94245	6	3	0.03	16864,30459,233740,311540,265109,85800
Row94253	6	3	0.03	16864,30459,15786,265109,311540,85800
Row94342	6	3	0.03	16864,233740,15786,48067,265109,311540

Now you can see all the items in a set!!!

Now we can complete the workflow with the CSV Writer



CUSTOMER SEGMENTATION

Anna Monreale KDD-Lab, University of Pisa Email: annam@di.unipi.it

Customer Segmentation

- **Problem:** given the dataset of RFM (Recency, Frequency and Monetary value) measurements of a set of customers of a supermarket, find a high-quality clustering using K-means and discuss the profile of each found cluster (in terms of the purchasing behavior of the customers of each cluster).
- Applying also the Hierarchical clustering and compare the results
- Provide a short document (max three pages in pdf, excluding figures/plots) which illustrates the input dataset, the adopted clustering methodology and the cluster interpretation.

DATA

- Dataset filename: rfm_data.csv.
- Dataset legend: for each customer, the dataset contains
 - date_first_purchase: integer that indicates the date of the first purchase of the customer
 - date_last_purchase: integer that indicates the date of the last purchase of the customer
 - Number of purchases: number of different purchases in terms of receipts
 - Amount: total money spent by the customer
- Need to vompute the columns for
 - *Recency*: no. of days since last purchase
 - Frequency: no. of visits (shopping in the supermarket) in the observation period
 - *Monetary value*: total amount spent in purchases during the observation period.

Clustering in KNIME

- Data normalization
 - Min-max normalization
 - Z-score normalization
- Compare the clustering results before and after this operation and discuss the comparison

\setminus	Noi	rmali	zer
/	-Þ	- ÌÍ-	Þ
	N	ode 1	J 13

00	Dialog – 2:2 – No	rmalizer
	Methods Flow Variables	Memory Policy
O Min-Max Normalization	Min: 0.0	
	Max: 1.0	
Z-Score Normalization		
O Normalization by Decimal Scalin	Ig	
Do not normalize	Select	Normalize
Column(s):	Search add >>	Column(s): Search
Select all search hits		Select all search hits
DATA_PRIMA_SPESA	add all >>	DATA_ULTIMA_SPESA
	< remove	
	<pre><< remove a</pre>	

K-Means

- Two options
 - K-means in Mining section of Knime

K-means in Weka section of Knime





 The second one allows the SSE computation useful for finding the best k value

CHURN ANALYSIS

Anna Monreale KDD-Lab, University of Pisa Email: annam@di.unipi.it

Churn Analysis

- **Problem**: Problem: given a dataset of measurements over a set of customers of an e-commenrce site, find a high-quality classifier, using decision trees, which predicts whether each customers will place only one or more orders to the shop.
- Provide a short document (max three pages in pdf, excluding figures/plots) which illustrates the input dataset, the adopted clustering methodology and the cluster interpretation.

Data

- Filename: OneShotCustomersEX.csv
 - Contains transactions from 15,000 online customers
- In the web page of the course you can download the attribute description
- The class of the data is **Customer Typology** that can be
 - one shot = only 1 purchase
 - loyal = more than one purchase

Decision Trees in Knime

- For Classification by decision trees
 - Partitioning of the data in training and test set



On the training set applying the learner

• On the test set applying the predictor



Decision

Evaluation of our classification model



\varTheta 😑	Cont	fusion Matrix	x - 0:13 - Scorer
File Hili	te		
File Hili Tipologia One Shot Loyal	te One Shot 2713 0	Loyal 0 2387	
Correc	t classified:	5,100	Wrong classified: 0
	100	0/	F 0.0/

00	Accuracy statistics – 0:13 – Scorer								
File									
			Table "default" - Rows: 3	Spec – Columns: 10	Properties	Flow Variables			
Row ID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	D Recall	D Precision	D Sensitivity	D Specifity	D F-measure
One Shot	2713	0	2387	0	1	1	1	1	1
Loyal	2387	0	2713	0	1	1	1	1	1
Overall	?	?	?	?	?	?	?	?	?
	(*******)) 4 1