Laboratório Clusterização com WEKA Explorer

Faça o download dos datasets **car-browsers.arff*** e **iris.arff*** , e execute clusterização conforme as páginas abaixo, onde encontram-se tutoriais.:

*disponível em: www.inf.ufpr.br/menotti/am-231/data.zip

1. Preparando os dados para classificação

- 1. Inicie uma sessão do Weka ou execute em linha de comando: *java –jar weka.jar*.
- 2. Quando a GUI Chooser surgir, selecione o Explorer a partir das quatro opções do lado direito.

		Weka Explorer		
		Preprocess Carofy Chatter Associats Select attributes [th		
		Open file Open URL Open DB	Generate	
		FRM		
		Choose Name		
		Current relation Relation: None Instances: None Attributes: None	Name: None Masing: None Distinct: None	Type: None Unique: None
		Attributes		
🕝 Weka GUI Chooser		Al Nove Doort 1		
Program Visualization Iools Help				
	Applications			
WEKA	Explorer			Visualize All
The University of Waikato	Experimenter			
10 M				
Waikato Environment for Knowledge Analysis	KnowledgeFlow			
(c) 1999 - 2014		E Bartove		
The University of Waikato	Simple CLI			
Hamilton, New Zealand		Welcome to the Welca Explorer		

- 3. Estamos no **Preprocess** agora. Clique no botão **Open** para abrir a caixa de diálogo padrão através da qual você pode selecionar um arquivo. Escolha o arquivo **telco_lab3.csv**.
- 4. Você pode ignorar atributos irrelevantes durante o processo de Clustering, como custIds. Para identificar atributos redundantes, poderíamos verificar a correlação a partir da Visualização do dataset na aba Visualize. Pode-se ver que os atributos age and agecat estão correlacionados. Um deles deve ser ignorado. Nós mantemos o atributo age para fins de agrupamento; também mantemos ed (removendo edcat), então temos 8 atributos restantes para Clustering (vamos ignorar custIds, agecat e edcat quando realizamos o Clustering).



5. Antes de fazer o Clustering com Weka, precisamos normalizar seus valores de dados numéricos (use o filtro Normalize). Como temos o rótulo de classe, gostaríamos de defini-lo como nominal antes da normalização. Esta informação será usada para avaliar o desempenho do clustering.

SimpleKMeans

 Para executar o Clustering no dataset dados, clique na guia Cluster e escolha o algoritmo SimpleKMeans. Definimos k = 2 para este dataset. Clique em Classes to clusters evaluation e selecione o último atributo (churn) como classe. Clique em Store Clusters for visualization. Clique em Ignore attributes e seleciona custIds, agecat, edcat e o último atributo churn (classe, que não será clusterizado). Em seguida, clique em Start.



```
🕌 14:34:03 - SimpleKMeans
                                               - 10
kMeans
-----
Number of iterations: 3
Within cluster sum of squared errors: 2000.5871625331147
Missing values globally replaced with mean/mode
Cluster centroids:
                    Cluster#
         Full Data
                                   1
Attribute
                         0
             (5000)
                     (2482)
                              (2518)
...........
                     0.5049
                               0.4958
region
             0.5004
townsize
            0.4218 0.4184 0.4252
                      0
gender
            0.5036
                               1
                   0.4788
                             0.4729
            0.4758
age
ed
            0.5025 0.5027
                              0.5024
             0.043 0.0435
income
                               0.0425
debtinc
            0.231 0.2302 0.2317
Time taken to build model (full training data) : 0.09 seconds
=== Model and evaluation on training set ===
Clustered Instances
    2482 ( 50%)
0
1
    2518 ( 50%)
Class attribute: churn
Classes to Clusters:
  0 1 <-- assigned to cluster
1839 1895 | 0
 643 623 | 1
Cluster 0 <-- 1
                                1839+623 = 2462
Cluster 1 <-- 0
Incorrectly clustered instances : 2462.0 49.24 %
                                                   >
```

2. Você poderia visualizar os resultados da clusterização clicando com o botão **Direito** na lista de resultados e escolher **Visualize clusters assignments**. Você pode selecionar uma combinação diferente de dois atributos como *X* e *Y*.

the state of the						
reprocess Classify Cluster Associate 3	elect attributes Visualize					
Oustere						
Choose SimpleKMeans -N2 - "web	a.coreucidearListance -R fiist-last	e' -1 500 -5 10				
Cluster mede	Clusterer ourput					
Ollac training act	logr	C.4758	0.4788	0.4729		3.
O suppled test set	ed	C.5025	0.5027	0.5024		
	income	0.040	0.0435	0.0425		
O Percentage split	S. Inn Gestine	0.231	0.2302	0.251		
() Classes to clusters evaluation						
(Num) churn	*					
Store clusters for visualization						
	Time taken	to build noce	L ituli trai	ning date) :	0.23 second	15
Ignore attributes	and Medal a	nd ownlungton	an employing			
			in training			
Start	CLustered In	nstances				
Result list (right-dick for options)						
(4:47:39 - Simple Clearns		(=0.5)				
	View in main window					- 1
	View in separate window					
	Save result buller					
	Delete result huffer					
	the second second second					
	lead medd	d	d to cluster			
	Save model					
	Re evaluate model on ru	irrent test set				
	Visualize cluster assigned	cats				
	Vieweller ber					
	Visualize use					
	in a second s	clustered ans	stances :	2462.0	49.24 %	
	INCOLLECTIV					
	INCOTTACT					
	anior receiv					>

3. Você pode salvar os resultados da clusterização clicando no botão **Save** quando estiver visualizando o resultado da clusterização.



4. Os resultados são salvos em um arquivo **.arff**. Você pode usar o próprio Weka para abrí-lo e ver os resultados no novo dataset.

Cluster Nominal	churn Nominal	debtinc Numeric	income Numeric	ed Numeric	age Numeric	gender Numeric	townsize Numeric	region Numeric	Instance_number Numeric	No.
duster 1	1	0.257	0.020	0.529	0.032	1.0	0.25	0.0	0.0	1
duster0	0	0.431	0.005	0.647	0.065	0.0	1.0	1.0	1.0	2
duster 1	0	0.229	0.024	0.470	0.803	1.0	0.75	0.5	2.0	3
duster0	0	0.132	0.010	0.588	0.081	0.0	0.5	0.75	3.0	1
duster0	0	0.039	0.013	0.588	0.131	0.0	0.25	0.25	4.0	5
dusterU	U	0.12993	0.092	U.64/	0./54	0.0	0.75	U./5	5.0	>
duster 1	0	0.044	0.06391	0.470	0.557	1.0	1.0	0.25	6.0	7
duster 1	0	0.334	0.082	0.588	0.42623	1.0	0.75	0.5	7.0	
duster 1	0	0.060	0.006	0.352	0.786	1.0	0.5	0.25	8.0	-
duster0	1	0.095	0.070	0.294	0.47541	0.0	0.25	0.25	9.0	10
duster 1	1	0.199	0.035	0.764	0.672	1.0	0.0	0.75	10.0	11
duster 1	1	0.020	0.009	0.117	0.245	1.0	0.75	0.25	11.0	2
duster0	0	0.064	0.06015	0.235	0.42623	0.0	0.25	1.0	12.0	3
duster0	0	0.243	0.050	0.705	0.655	0.0	0.25	0.5	13.0	4
duster 1	0	0.227	0.007	0.823	0.885	1.0	0.0	0.25	14.0	15
cluster 1	0	0.215	0.013	0.411	0.786	1.0	0.0	0.5	15.0	.6
duster 1	0	0.220	0.152	0.647	0.639	1.0	0.0	0.0	16.0	7
duster0	0	0.24826	0.390	0.470	0.737	0.0	0.25	1.0	17.0	8
duster 1	0	0.111	0.013	0.294	0.163	1.0	1.0	1.0	18.0	9
duster 1	0	0.352	0.012	0.588	0.983	1.0	0.0	0.0	19.0	20

5. Se o dataset não tiver classe definida, quando você realizar a clusterização no dataset, escolha **Use Training Dataset** como Cluster mode.

0 v	/eka Explorer 🛛 🗕 🗖
Preprocess Classify Cluster Associate Select attributes Clusterer	i Visualize
Choose SimpleKMeans -N 2 -A "weka.core.Euclide	anDistance -R first-last" -I 500 -S 10
Cluster mode	Clusterer output
Use training set	=== Run information ===
Supplied test set Set	Scheme:weka.clusterers.SimpleKMeans -N 2 -A "w
Classes to dusters evaluation	Instances: 5000
(Num) debtinc	Attributes: 7
Store dusters for visualization	townsize
	gender
Ignore attributes	ed
Start Stop	income
Result list (right-click for options)	debtinc
09:41:39 - SimpleKMeans 09:50:52 - SimpleKMeans 09:54:32 - SimpleKMeans	=== Model and evaluation on training set ===
	kMeans
	Number of iterations: 3 Within cluster sum of squared errors: 2000.587
Status OK	Log

<u></u>	09:54:32	- SimpleKN	1eans –		×	
kMeans						^
Number of i	terations. 3					
Within clus	ter sum of sou	ared error	a: 2000.5871	62533	113	
Missing val	ues globally	replaced wit	th mean/mode	-		
Cluster cer	troids:					
		Cluster#				
Attribute	Full Data	0	1			
	(5000)	(2482)	(2518)			
region	0.5004	0.5049	0.4958			
townsize	0.4218	0.4184	0.4252			
gender	0.5036	0	1			
age	0.4758	0.4788	0.4729			
ed	0.5025	0.5027	0.5024			
income	0.043	0.0435	0.0425			
debtinc	9.9542	9.9199	9.9879			
Time taken === Model a	to build mode: and evaluation	l (full trai on training	ining data) g set ===	: 0.0)9 s	
Clustered I	Instances					
1 2462	(508)					
1 2310	1 20.61					Y
<					>	

DBScan

 Agora vamos abordar o algoritmo DBScan. Da mesma forma que fizemos para SimpleKMeans, clique em Classes to clusters evaluation e selecione o último atributo (churn) como classe. Clique em Store Clusters for visualization. Clique em Ignore attributes e seleciona custIds, agecat, edcat e o último atributo churn (classe, que não será clusterizado). Em seguida, clique em Start. Inicialmente vamos usar os valores default para epsilon e minPoints.

🕲 🖨 🕕 Weka Explorer					
Preprocess	Classify Cluster As	sociate Select attributes Visualize			
Clusterer					
Choose	DBSCAN -E 0.9 -M 6 -I weka.	clusterers.forOPTICSAndDBScan.Databases.SequentialDatabase -D weka.c	lustere		
Cluster mod	😣 🗉 weka.gui.Generico	DbjectEditor			
🔾 Use trair	weka.clusterers.DBSCAN		^		
Supplied	About				
O Percenta	Basic implementation of DE	SCAN clustering algorithm that should *not* be used More			
Classes 1	as a reference for runtime exist! Clustering of new ins	tances is not supported.	ds		
(Nom) ch					
Store clu	database_Type	weka.clusterers.forOPTICSAndDBScan.Databases.SequentialDatabase			
	database_distanceType	weka.clusterers.forOPTICSAndDBScan.DataObjects.EuclideanDataObject			
	ensilon	0.9			
Star		-			
Result list (r	minPoints	6			
23:14:32 - DE	Open	Save OK Cancel			
		0 1 < assigned to cluster			
		1895 1839 0 623 643 1			
		Cluster 0 < 0 Cluster 1 < 1			
		Incorrectly clustered instances : 2462.0 49.24 %			
		Incorrectly clustered instances : 2402.0 45.24 %	=		
			-		
Chabur					
OK		Log	x 0		

2. Experimente alterar o valor de **epsilon** para 0.3 e de **minPoints** para 100 e verifique o novo resultado da **Clusterização**.

😣 🖱 🗉 🛛 Weka Explorer					
Preprocess Classify Cluster Associat	e Select attributes Visualize				
Clusterer					
Choose DBSCAN -E 0.3 -M 100 -I weka.clust	terers.forOPTICSAndDBScan.Databases.SequentialDatabase -D weka.clust				
Cluster mode	Clusterer output				
O Use training set					
O Supplied test set Set	Time taken to build model (full training data) : 4.99 seconds				
O Percentage split % 66					
Classes to clusters evaluation	=== Model and evaluation on training set ===				
(Nom) churn 🔫	Clustered Instances				
Store clusters for visualization	0 448 (42%)				
Ignore attributes	1 200 (25%) 2 349 (33%)				
Start Stop	Unclustered instances : 3943				
Result list (right-click for options)	Class attribute: churn				
23:14:32 - DBSCAN	Classes to Clusters:				
23:17:30 - DBSCAN	0 1 2 < assigned to cluster 343 181 261 0 105 79 88 1				
	Cluster 0 < 0 Cluster 1 < No class Cluster 2 < 1				
	Incorrectly clustered instances : 626.0 12.52 %				
Status OK	Log 💉 x 0				