Auto Model と Turbo Prep のオートクリーニング機能を再現

- "Quality Measures"によるデータクリーニングー

Auto Model と TurboPrep の両方に、データセットから不要な特徴量を削除できる「オートクリーニング」機能があります。 RapidMiner は特徴量を削除する必要があるかどうかを判断するために「品質測定」(Quality Measures)を実行します。この「品質測定」をご存じでない場合は、以下を参照してください:

- Correlation: 説明変数と目的変数の線形相関測定
- ID-ness: 特徴量の値が独自的(ユニーク) であるかの測定
- **Stability:**特徴量の値がどの程度同一であるかの測定
- Missing: 特徴量の欠落値の%測定
- Text-ness: 特徴量に含まれるフリーテキストの量の測定

Auto Model では、「品質測定」は「入力の選択」ページで使用されます:

		•	Rest	ART (BACK	> NEXT	-	•		Select Inputs Here the focus is on the quality of your data, specifically the quality of each column of data. You may want to consider discarding data columns (Attributes) that provide less value.
			Deselect Red	Selected: 8 / To Deselect Yellow	otal: 11	X Deselect All			How do you know which Attributes are valuable, and which are worthless? A key point is that you're looking for patterns. Without some variation in the data and some discernible
Selected	Status 🕆	Quality	Name	Correlation	ID-ness	Stabillity	Missing	Text-ness	patterns, the data is not likely to be useful. A quick summary of things to look out for (more details below) includes:
	•		Name	8.63%	99.85%	0.15%	0.00%	78.67%	CC) Columns that too closely mirror the target column.
	•		Ticket Number	8.46%	70.97%	0.84%	0.00%	35.54%	 (I) Columns where nearly all values an different, (5) Columns where nearly all values a identical
	•	-	Cabin	5.10%	14.21%	2.03%	77.46%	27.31%	 (M) Columns with missing values, (T) Columns which look like they contain free text.
X	٠		Life Boat	59.37%	2.06%	8.02%	62.87%	3.20%	To help you make a decision, we indicate the Attribute value with a color-coded status bubble (red /yellow / green). Details are provided by th multiplane (C) US (M T) As a concert of the
Z	•		Passenger Class	9.76%	0.23%	54.16%	0.00%	2.39%	is a good idea to deselect at least hose Attributes that have a red status bubble. The ing for the machine learning model will only include the selected Attributes.
2		-	Sex	27.95%	0.15%	64.40%	0.00%	2.15%	





Turbo Prep では最初に「Clense」ボタンをクリックしてから、左側の「Auto Cleansing」 オプションをクリックすることで利用できます:

Turbo Prep							
Data Sets	Titanic Add new data sets	on the left. Details fo	or the selected d	ata are shown below.	You can change the d	ata with the following	g actions.
+ LOAD DATA	🗙 TRANSFORM	VI CLEANS	E 🖩 GENI pare your data fo	ERATE SPIVO	T 🐎 MERGE at RapidMiner perform	an automatic data c	leansing.
Titanic							
//Samples/data/Titanic Rows: 1,309 Columne: 12	Passenger Cl Category	Name _{Category}	Sex Category	Age Number	No of Sibling Number	No of Parents	Ticket N Category
Last Change: None	First	Allen, Miss. Eli	Female	29	0	0	24160
	First	Allison, Master	Male	0.91670000	1	2	113781
Turbo Prep							
Cleanse	Titanic						
	Select a column to c	lean (hold Shift for se	electing a range of	of columns; Ctrl for (de-)selecting multiple col	umns; Alt to select al	Il columns
0 columns selected	Je commit cl	CANC	EL				
AUTO CLEANSING		Ilmu					lu
	ing of your data which	ch solves the most co	mmon data qual	lity problems. Especia	Ily useful before mach	ne learning models	are used.
	First	Allen, Miss. Eli	Female	29	0	0	24160
REMOVE CORRELATED	First	Allison, Master	Male	0.91670000	1	2	113781





	Defir	ne Target Impro	ve Quality Char			Summary	
		•	•	•		-0	
s table is just for y	your information	. RapidMiner will autor	matically remove the c	olumns highlighted b	elow since they have	e a very low quality f	or machine learning.
also replace all r	missing values f	for the remaining colur	nns.				
D-like					(ID-like		(Many missing
i me tegory	Sex Category	Age Number	No of Sibling	No of Parents	Ticket Number Category	Passenger F Number	Cabin Category
llen, Miss. Eli	Female	29	0	0	24160	211.33750000	B5
llison, Master	Male	0.91670000	1	2	113781	151.55000000	C22 C26
llison, Miss. H	Female	2	1	2	113781	151.55000000	C22 C26
llison, Mr. Hud	Male	30	1	2	113781	151.55000000	C22 C26
llison, Mrs. Hu	Female	25	1	2	113781	151.55000000	C22 C26
nderson, Mr	Male	48	0	0	19952	26.55000000	E12
ndrews, Miss	Female	63	1	0	13502	77.95830000	D7
ndrews, Mr. T	Male	39	0	0	112050	0	A36
ppleton, Mrs	Female	53	2	0	11769	51.47920000	C101
rtagaveytia, Mr	Male	71	0	0	PC 17609	49.50420000	?

そこから、「品質測定」は「Improve Quality」ステップで使用されます:

これらの2つのオプションは便利ですが、クリーニングのために特徴量を選択する方 法をあまり制御できません。

この便利な機能を自分のプロセスで使用するには、「Quality Measures」オペレーター を使用できます。 このオペレーターは、データの特徴量について、Auto Model および TurboPrep と同じ方法で計算します。 このオペレーターをフィルタリングと組み合わ せると、お好みに合わせてクリーニングプロセスを制御できます。 見てみましょう。



実装の内容



まず、属性の1つに「ラベル」の役割を与えるようにしてください。 そうしないと、 「相関」(Correlation)の品質測定値を作成できません。

Row No.	Survived	Passenger	Name	Sex	Age	No of Sibling	No of Parent	Ticket Numb	Passenger F	Cabin	Port of Emb	Life Boat
1	Yes	First	Allen, Miss. E	Female	29	0	0	24160	211.33750000	B5	Southampton	2
2	Yes	First	Allison, Mast	Male	0.91670000	1	2	113781	151.55000000	C22 C26	Southampton	11
3	No	First	Allison, Miss	Female	2	1	2	113781	151.55000000	C22 C26	Southampton	?
4	No	First	Allison, Mr. H	Male	30	1	2	113781	151.55000000	C22 C26	Southampton	?
5	No	First	Allison, Mrs	Female	25	1	2	113781	151.55000000	C22 C26	Southampton	?
6	Yes	First	Anderson, Mr	Male	48	0	0	19952	26.55000000	E12	Southampton	3
7	Yes	First	Andrews, Mis	Female	63	1	0	13502	77.95830000	D7	Southampton	10
8	No	First	Andrews, Mr	Male	39	0	0	112050	0	A36	Southampton	?
9	Yes	First	Appleton, Mrs	Female	53	2	0	11769	51.47920000	C101	Southampton	D
10	No	First	Artagaveytia,	Male	71	0	0	PC 17609	49.50420000	?	Cherbourg	?
11	No	First	Astor, Col. Jo	Male	47	1	0	PC 17757	227.52500000	C62 C64	Cherbourg	?
12	Yes	First	Astor, Mrs. Jo	Female	18	1	0	PC 17757	227.52500000	C62 C64	Cherbourg	4
13	Yes	First	Aubart, Mme	Female	24	0	0	PC 17477	69.30000000	B35	Cherbourg	9
14	Yes	First	Barber, Miss	Female	26	0	0	19877	78.85000000	?	Southampton	6
15	Yes	First	Barkworth, Mr	Male	80	0	0	27042	30	A23	Southampton	в
16	No	First	Baumann, Mr	Male	?	0	0	PC 17318	25.92500000	?	Southampton	?
17	No	First	Baxter, Mr. Qu	Male	24	0	1	PC 17558	247.52080000	B58 B60	Cherbourg	?
18	Yes	First	Baxter, Mrs. J	Female	50	0	1	PC 17558	247.52080000	B58 B60	Cherbourg	6
19	Yes	First	Bazzani, Miss	Female	32	0	0	11813	76.29170000	D15	Cherbourg	8
	1917	1000			12223	22	12			1221	12201020000	

ExampleSet (1,309 examples, 1 special attribute, 11 regular attributes)



その後、データは特徴量フィルタリングのプロセスに入ります。 フィルタリングプロ セスをこの場合このように作成しましたが、お好みで変えても構いません。



「Select Subprocess」オペレーター(ここでは Filter Attributes と名前を変えています) 内にこのように 3 つの全体的なステップに分けてオペレーターを配置します:

- 1. データを複製します (Multiply オペレーター)
- 2. 「Quality Measures」オペレーターで、フィルタリングする特徴量を選択します
- 3. 「Select by Weights」オペレーターを使用して特徴量をフィルタリングします

「Weight Attributes by Quality Measures」と呼ばれた 2 番目のステップで、ほとんどの 作業を行っています。







最初に、品質測定値が計算されて、結果(以下に表示)には、元のデータの各特徴量の行と、その特徴量の品質測定値に関する情報が含まれます。

Row No.	Attribute	Correlation	ID-ness	Stabillity	Missing	Text-ness
1	Passenger Class	0.09763710	0.00229183	0.54163484	0	0.02392666
2	Name	0.08632738	0.99847212	0.00152788	0	0.78673797
3	Sex	0.27951638	0.00152788	0.64400306	0	0.02145149
4	Age	0.00308164	0.05500382	0.04493308	0.20091673	0
5	No of Siblings or Spouses on Board	0.00077424	0.00534759	0.68067227	0	0
6	No of Parents or Children on Board	0.00683260	0.00611154	0.76546982	0	0
7	Ticket Number	0.08463433	0.70970206	0.00840336	0	0.35536542
8	Passenger Fare	0.05966562	0.02444614	0.04587156	0.00076394	0
9	Cabin	0.05095917	0.14209320	0.02033898	0.77463713	0.27311488
10	Port of Embarkation	0.00905755	0.00229183	0.69931140	0.00152788	0.04739947
11	Life Boat	0.59374771	0.02062643	0.08024691	0.62872422	0.03197988

その後、「Filter Examples」オペレーターを使用してフィルター処理する必要がありま す。 そうすれば、保持したい特徴量だけが残ります。 次のように式を書いてありま す:

Edit Expression: parameter expression

Expression

 ((Stabillity > 0.9) ||
 (Missing > 0.7) ||
 (Correlation < 0.0001) ||
 4 (Correlation > 0.95) ||
 5 ([ID-ness] > 0.9 && [Text-ness] < 0.85)
 6
</pre>

Info: Expression is syntactically correct.





使用される値は、AutoModel の[Select Inputs]ページの[Help View]ボックスに表示される 情報に基づいています。 ただし、必要に応じて変更できます。

		Load C)ata Select Task Pi	repare Target Se	elect Inputs Mo	del Types Re	sults		(i) Select Inputs
			≪ RESTA	IRT (BACK) NEXT				Select inputs Here the focus is on the quality of your data, specifically the quality of each column of data. You may want to consider discarding data columns (Attributes) that provide less value.
			Deselect Red	Deselect Yellow	Select All	X Deselect All			How do you know which Attributes are valuable, and which are worthless? A key point is that you're looking for patterns. Without some variation in the data and some discernible
Selected	Status †	Quality	Name	Correlation	ID-ness	Stabillity	Missing	Text-ness	patterns, the data is not likely to be useful. A quick summary of things to look out for (more details below) includes:
	•		Name	8.63%	99.85%	0.15%	0.00%	78.67%	 (C) Columns that too closely mirror the target column,
		-	Ticket Number	8.46%	70.97%	0.84%	0.00%	35.54%	(i) Columns where nearly all values are different, (S) Columns where nearly all values are
		-		(F-100)			-		 Identical, (M) Columns with missing values,
		-	Cabin	5.10%	14.21%	2.03%	77.40%	27.31%	 (T) Columns which look like they contain free text.
	۲		Life Boat	59.37%	2.06%	8.02%	62.87%	3.20%	To help you make a decision, we indicate the Attribute value with a color-coded status bubble (red /yellow / green.) E platilis are provided by the maintiv bars (C1)/S/M/T). As a general rule, it
Z	•		Passenger Class	9.76%	0.23%	54.16%	0.00%	2.39%	Is a good idea to deselect at least those Attributes that have a red status bubble. The input for the machine learning model will only include the selected Attributes.
2	•		Sex	27.95%	0.15%	64.40%	0.00%	2.15%	∞ •

その後、「Generate Attributes」オペレーターを使い、新しく weight 列を作成します。

Row No.	Attribute	Correlation	ID-ness	Stabillity	Missing	Text-ness	weight
1	Passenger Class	0.09763710	0.00229183	0.54163484	0	0.02392666	1
2	Sex	0.27951638	0.00152788	0.64400306	0	0.02145149	1
3	Age	0.00308164	0.05500382	0.04493308	0.20091673	0	1
4	No of Siblings or Spouses on Board	0.00077424	0.00534759	0.68067227	0	0	1
5	No of Parents or Children on Board	0.00683260	0.00611154	0.76546982	0	0	1
6	Passenger Fare	0.05966562	0.02444614	0.04587156	0.00076394	0	1
7	Port of Embarkation	0.00905755	0.00229183	0.69931140	0.00152788	0.04739947	1
8	Life Boat	0.59374771	0.02062643	0.08024691	0.62872422	0.03197988	1





次に、「ExampleSet to Weights」オペレーターを使用して、サンプルセットを「weights vector」に変換しました。

attribute	weight
Passenger Class	1
Sex	1
Age	1
No of Siblings or Spouses on Board	1
No of Parents or Children on Board	1
Passenger Fare	1
Port of Embarkation	1
Life Boat	1

最後に、「weights vector」は、「Select by Weights」オペレーターが関連する特徴量を 選択するために使用します。

Row No.	Survived	Passenger Class	Sex	Age	No of Siblings or Spouses on Board	No of Parents or Children on Board	Passenger Fare	Port of Embarkation	Life Boat
1	Yes	First	Female	29	0	0	211.33750000	Southampton	2
2	Yes	First	Male	0.91670000	1	2	151.55000000	Southampton	11
3	No	First	Female	2	1	2	151.55000000	Southampton	?
4	No	First	Male	30	1	2	151.55000000	Southampton	?
5	No	First	Female	25	1	2	151.55000000	Southampton	?
6	Yes	First	Male	48	0	0	26.55000000	Southampton	3
7	Yes	First	Female	63	1	0	77.95830000	Southampton	10
8	No	First	Male	39	0	0	0	Southampton	?
9	Yes	First	Female	53	2	0	51.47920000	Southampton	D
10	No	First	Male	71	0	0	49.50420000	Cherbourg	?
11	No	First	Male	47	1	0	227.52500000	Cherbourg	?
12	Yes	First	Female	18	1	0	227.52500000	Cherbourg	4
13	Yes	First	Female	24	0	0	69.30000000	Cherbourg	9
14	Yes	First	Female	26	0	0	78.85000000	Southampton	6
15	Yes	First	Male	80	0	0	30	Southampton	в
16	No	First	Male	?	0	0	25.92500000	Southampton	?
17	No	First	Male	24	0	4	247.52080000	Cherbourg	?
18	Yes	First	Female	50	0	1	247.52080000	Cherbourg	6
19	Yes	First	Female	32	0	0	76.29170000	Cherbourg	8
~~					•	•	75 04470000	AL	

この出力は、必要に応じてさらに処理したり、モデリングで使用したりできます。

以上のやり方で通常のプロセスでも「品質測定」を実行出来るようになりますので、 是非、試してみてください。



