第3章

データの前処理

概要

ジェリーはインターネットのデザインと広告の小さな会社のマーケティングのマネー ジャーです。ある日、上司からインターネットを使用しているユーザーに関する情報 を新たに取得するように依頼がありました。この情報を使用して、企業はどのような 人がインターネットを使用しているかを確認し、どのようにすれば企業のサービスを 顧客へ販売できるかどうかを判断したいと考えていました。

ジェリーはオンライン調査を作成し、いくつかの人気のある Web サイト上へオンライ ン調査へのリンクを配置しました。分析を始めるために、2週間以内に十分なデータ を収集しましたが、それらのデータを非正規化する必要があることや、データセット の中に欠損値や無効な値が含まれていることを発見し、ジェリーは分析を開始する前 に、データを整備する必要があることに気づきました。

学習目標

- データの前処理の概念と目的を説明する
- 欠損値の対処方法を列挙する
- 一貫性のないデータを明確にし、対処する
- データ行の削除の目的を説明し、データ行の削除を実施する
- 変数の削減の目的を説明し、変数の削減を実施する

ハンズオン

これから実際に PC 上で操作していきますので、RapidMiner がインストールされている ことが前提です。また、この本の関連サイトへアクセスするためには、インターネッ ト接続が必要です。下記の関連サイトからは、各章で使用されるすべてのデータセッ トのコピーをダウンロード出来ます。

https://sites.google.com/site/dataminingforthemasses3e/

Data Mining for the Mass 🗙			/	mjn —			
C 🛆 🔒 Secure	https://sites.google.com	/site/dataminingforthem	nasses3e/		* 1		
Data Mi	ining for	the		Search this	site		
Masses.	.3e						
1140000)	00						
Resources							
Second Edition Site	•••••				•••••		
Chapter Slides	Walasma to the same ani	n wah aita fas tha haak D	ate Mining for the Manage	Third Edition If you are			
First Edition Site	looking for the companion	s of the book, use the navi	gation to the left.				
RapidMiner	tooking for the companion websites for pror editions of the book, use the navigation to						
The R Project	The third edition of the bo	ok was prepared using Ra	apidMiner 9.0 and R 3.5 wi	th R Studio 1.1.4. The da	ta		
R Studio	sets below are compatible with prior editions of the b	e with these software vers look all data sets are stor	ions, and match the examp ed in either Comma Separ	ples given in the book. As ated Values (csv) or Tex) t		
	(.txt) format for simplicity's	s sake.					
	Downloads	/					
	Chapter 3 Data	<u>Chapter 4 Data</u> <u>Chapter 4 Exercise</u>	<u>Chapter 5 Data</u> <u>Chapter 5 Exercise</u>	<u>Chapter 6 Data</u> <u>Chapter 6 Exercise</u>			
	<u>Chapter 7 Training</u> <u>Chapter 7 Scoring</u> <u>Chapter 7 Exercise</u>	<u>Chapter 8 Scoring</u>	<u>Chapter 9 Training</u> <u>Chapter 9 Scoring</u>	<u>Chapter 10 Trainin</u> Chapter 10 Scoring	g		
	<u>Chapter 10 Training</u> <u>Chapter 10 Scoring</u> <u>Chapter 10 Exercise</u>	<u>Chapter 12 Data</u>					

図 3-4. Data Mining for the Masses, Third Edition 関連サイト

上記サイトにて、「Chapter 3 data」のダウンロードリンクをクリックすることで第3 章のデータセット(Chapter03DataSet.csv)をダウンロードすることが可能です。

ータの読み込みと欠損値の対処について

データの前処理の最初のタスクは**欠損値(Missing data)**の除去です。図 3-9 をご覧く ださい。欠損値はゼロや他の値と違い空白になっています。値は不明、未定義、また は未定です。欠損値はデータベースの世界では、null と呼ばれることもあります。デ ータマイニングの目的に応じて、欠損値をそのままにするか、他の値に置き換えるこ とが可能です。

9	vwInternetU	ser - DataMiningFo	orTheMasses - (OpenOffice.org Ba	se: Table Data View										- 0 X
Eil	e <u>E</u> dit <u>V</u> iev	v Insert <u>T</u> ools	<u>W</u> indow <u>H</u> elp)											&
1															
	Gender	Race	Birth_Year	Marital_Status	Years_on_Internet	Hours_Per_Day	Preferred_Browser	Preferred_Search_Engine	Preferred_Email	Read_News	Online_Shopping	Online_Gaming	Facebook	Twitter	Other_Social_Network
	M	White	1972	М	8	1	Firefox	Google	Yahoo	Y	N	N	Y	N	
	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y	N	
	F	African American	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	γ		Y	N	-
	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N	Y	Χ
	M	White	1954	M	2	3	Internet Explorer	Bing	Hotmail	Y	Y	N	Y	N	
	M	African American	1982	D	15	4	Internet Explorer	Google	Yahoo	Y	N	Y	N	N	
	M	African American	1981	D	11	2	Firefox	Google	Yahoo		Y	_	Y	Y	LinkedIn
	M	White	1977	S	3	3	Internet Explorer	Yahoo	Yahoo	Y			Y	99	LinkedIn
	lF	African American	1969	М	6	2	Firefox	Google	Gmail	N	Y	N	N	N	

図 3-9: 調査データ内の欠損値

例えば、Other Social Network 変数は、一般的な SNS 以外でユーザーが使用している SNS についての自由記述欄ですが、欠損値となっているユーザーは単に回答していな いだけと考えられるため、欠損値はそのままにしておきます。一方で、Online Gaming 変数には、オンラインゲームを利用"Y"もしくは未利用"N"のどちらかの回答がありま すが、欠損値となっているものもありますので対処する必要があります。

下記チュートリアルでは、RapidMiner での欠損値の対処方法について学びます。以下の手順に従ってデータにアクセスし、欠損値を修正します。

 RapidMiner Studio を起動していない場合は、デスクトップにあるアイコンをダブ ルクリックして起動して下さい。起動すると次の画面が表示されますので、新 しくプロセスを作成するために、Blank ボタンをクリックして下さい。

^{*} portions of this book are adapted from Data Mining for the Masses, by Matthew North, copyright 2020.



 RapidMiner は、様々な種類のデータソースへアクセスすることが可能ですが、 Studio の Free 版を使用している場合は、レコード数が 10,000 件以下に制限され ます。本チュートリアルの目的としては 10,000 件以下で十分ですが、ビジネス においては、RapidMiner ライセンスをアップグレードすることをお勧めします。 RapidMiner の画面上の左側にリポジトリがあります。これから、リポジトリを 新たに作成する方法をご紹介します。まず、下記のようにリポジトリタブの右 側にあるアイコン をクリックし、Create repository メニューをクリックしま す。



図 3-11. リポジトリを作成する

 New local repository が選択された状態で Next ボタンをクリックすると、下記(図 3-12)の画面が表示されますので、「RapidMinerBook」と入力し、ディレクトリ のパスはそのままで Finish ボタンをクリックします。

New Reposito	ry X
<u>A</u> lias:	RapidMinerBook
Root directory:	✓ Use standard location
	C:\Users\10589759\RapidMiner\repositories\RapidMinerBook
	$\leftarrow \underline{P}revious \longrightarrow \underline{N}ext \qquad \boxed{\mathbb{P}revious} \qquad \underline{\mathbb{P}revious} \qquad \underline{\mathbb{P}reviou$

図 3-12. リポジトリの名前とディレクトリのパスを設定

クリックすると、下記のように RapidMinerBook リポジトリが作成されていることを確認出来ます。

Repository	×	
	🕂 Import Data	≡▼
🕨 👅 Trainin	g Resources (connected)	
🕨 🔁 Sample	es	
🕨 🚨 Comm	unity Samples (connected))
🕨 📒 DB		
🕨 🌉 Local F	Repository (10589759)	
🕨 🌉 RapidM	MinerBook (10589759)	
Cloud F	Repository (disconnected)	

図 3-13. 新しいリポジトリが作成された

5) 次の画像内の矢印で示されているオペレータとリポジトリとパラメータ はよく 使用されます。リポジトリは取り込んだデータを保存する場所です。オペレー タはデータマイニングを実施するためのツールが配置されており、パラメータ

を変更することでオペレータの動作を変更することが可能です。 中央の領域は デザイン画面と呼ばれており、ここでモデルを作成します。



図 3-14. RapidMiner でよく使われるツール

6) 関連サイトからダウンロードした Chapter03DataSet.csv の変数は15個で、11 レコードあります。これからこのデータを取り込みますが、もし関連サイトから ら Chapter03DataSet.csv をダウンロードされていない場合は、 関連サイトからデ ータをダウンロードします。次に、上記の図 3-14 のリポジトリタブの下にある Import Data ボタンをクリックします。そして、下記の図 3-15の画面が表示され たら My Computer ボタンをクリックします。



図 3-15. CSV ファイルを読み込む

7) 図 3-15 に表示されている'My Computer'ボタンを押して、第3章のデータセット をダウンロードして保存したディレクトリへ移動します。正しいディレクトリ へ移動するために小さな矢印のついたフォルダのアイコンを使う必要があるか もしれません。

	Sele	ct the data location.	\backslash
Chapter Data Sets			
Bookmarks	File Name	Size	Type Last Modified
★ Last Directory	Chapter03DataSet.csv	1 KB M	licrosoft Excel Comma Se May 31, 2018

8) 図 3-16のように、Chapter03DataSet.csv をクリックし、Next を選択します。

< 1	leader Row			1 🗘	File Encod	ding	windows-	1252 •	· 🗸 U:	se Quotes	•		
tart	Row			1 ‡	Escape C	haracter	١	۱		Trim Lines			
olu	mn Separato		omma 🚏	¥	Decimal C	Character			si	kip Comment:	s #		
1	Gender	Race	Birth Ve	Marital	Vears o	Hours D	Dreferre	Dreferre	Dreferre	Read N	Online	Online	Facel
2	M	White	1972	Maritai	8	1	Firefox	Google	Yahoo	Y	N	N	Y
3	M	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y
4	F	African A	1977	s	6	2	Firefox	Yahoo	Yahoo	Y	Y		Y
5	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N
6	М	White	1954	М	2	3	Internet	Bing	Hotmail	Y	Y	N	Y
7	М	African A	1982	D	15	4	Internet	Google	Yahoo	Y	N	Y	Ν
8	м	African A	1981	D	11	2	Firefox	Google	Yahoo		Y		Y
9	М	White	1977	S	3	3	Internet	Yahoo	Yahoo	Y			Y
0	F	African A	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	Ν
1	м	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y		Y	Y
2	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y

図 3-17. データ形式の指定

9) デフォルトでは、カンマ区切りとして認識します。もし、データセットがカンマ以外を区切り文字として使用している場合には、これを変更することができ

ます。変更する場所は、図 3-17 の黒い矢印で示されています。もし、データに カンマが含まれている場合は、データにない区切り文字、例えばタブ、セミコ ロン、パイプ(|)、または他の記号などを使えば、意図しない場所で列が区切ら れることを回避することができるでしょう。それでもデータを正しく区切れな い場合は、正規表現(Regular Expression)を使いましょう。データの両端に"(ク ォーテーションマーク)を使用することもできます。これにより、"(クォー テーションマーク)内にあるカンマはデータの一部として認識され、"(クォ ーテーションマーク)の外にあるカンマは列の区切り文字として扱われます。 もし、CSV ファイルがこのようなデータであれば、'Use Quotes'のボックスをチ ェックすることを忘れないようにしましょう。Chapter03DataSet.csv ファイル内 のデータは列内のデータに"(クォーテーションマーク)も含まれておらず、 また、列内のデータにカンマも含まれていないので、'Use Quotes'ボックスのチ ェックの有無による影響はありません。

デフォルトの取込設定では、CSV データセットの一行目をヘッダー行として指 定しており、一行目が列名として認識されます。もし、CSV データセットの一 行目の列が列名を含んでいない場合には、データセットをインポートする時に、 図 3-18 の左端の矢印の Header Row のチェックが外れていることを確認しましょ う。そうすれば、一行目のレコードがヘッダー行として扱われることはありま せん。

RapidMiner チュートリアル(Tutorial for RapidMiner) 第3章

		\checkmark			- Op	sony you	i aatu io						
/ F	leader Row	*		1 🖕	File Encod	ding	windows-	1252 •	· 🗸 U	se Quotes	¥ -		
tart	Row			1 ‡	Escape C	haracter	۱		rim Lines				
olur	nn Separato	ır C	comma ","	•	Decimal Character		Skip Comments		#				
1	Gender	Race	Birth_Ye	Marital	Years_o	Hours_P	Preferre	Preferre	Preferre	Read_N	Online	Online	Facel
2	м	White	1972	м	8	1	Firefox	Google	Yahoo	Y	N	N	Y
3	м	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	N	N	Y
4	F	African A	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y		Y
5	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	N	N
6	м	White	1954	м	2	3	Internet	Bing	Hotmail	Y	Y	N	Y
7	м	African A	1982	D	15	4	Internet	Google	Yahoo	Y	Ν	Y	N
8	м	African A	1981	D	11	2	Firefox	Google	Yahoo		Y		Y
9	м	White	1977	S	3	3	Internet	Yahoo	Yahoo	Y			Y
10	F	African A	1969	M	6	2	Firefox	Google	Gmail	N	Y	N	Ν
11	м	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y		Y	Y
12	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	N	N	Y

図 3-18. CSV データの列のプレビュー

10) データのプレビューで、それぞれの列が正しく区切られていることを確認した ら、Nextをクリックします。

				Format your o	olumns.			
	Date format MMM d,	yyyy h:mm:ss a z	•	Replace errors wit	h missing values ①			
	Gender 🔅 🔻	Race 🔅 🔻	Birth_Year ♦ ▼ integer	Marital_Sta	Years_on_I •	Hours_Per • •	Preferred •	Preferred
1	м	White	1972	М	8	1	Firefox	Google
2	м	Hispanic	1981	S	14	2	Chrome	Google
3	F	African American	1977	S	6	2	Firefox	Yahoo
4	F	White	1961	D	8	6	Firefox	Google
5	М	White	1954	М	2	3	Internet Explorer	Bing
6	м	African American	1982	D	15	4	Internet Explorer	Google
7	М	African American	1981	D	11	2	Firefox	Google
8	м	White	1977	S	3	3	Internet Explorer	Yahoo
9	F	African American	1969	м	6	2	Firefox	Google
0	м	White	1987	S	12	1	Safari	Yahoo
1	F	Hispanic	1959	D	12	5	Chrome	Google

図 3-19. 列名を設定

11) 図 3-19 の画面で、RapidMiner はそれぞれの列のデータ型を推測します。データ 型とは数値、文字列、または日付のように列が持つデータの種類のことです。

これらはこの画面の各列の名前の右側にある歯車のアイコンの横の下三角▼を クリックすると変更できます。歯車アイコンによって列の名前やロールを変え ることも、また、除外する列を選択することもできますが、今回はデフォルト の設定のままにします。後の章で、さまざまな種類のデータやデータマイニン グのタスクに適したデータ型やロールについての設定をさらに扱います。もし、 データが日付を含んでいる場合、必要であれば、Date format(日付フォーマッ ト)を変更して、RapidMiner にデータの日付のフォーマットをどのように読ま せるのか必ず教えましょう。これらの各機能は図 3-20 の黒い矢印で示してあり ます。設定はデフォルトのままで、Nextをクリックし、先に進みます。

	Date format MMM d,	yyyy h:mm:ss a z	•	Replace	e errors wit	h missing values ①			
	Gender 🎄 🔻 polynominal	Race polynominal	 Birth_Year Change Type 	 Marital_9 polynominal 	Sta	Years_on_I ¢ ▼ integer	Hours_Per	Preferred 💠 🔻	Preferred
1	м	White 🗶	Change Role	binominal		8	1	Firefox	Google
2	М	Hispanic	Rename column	real		14	2	Chrome	Google
3	F	African Ameri	Exclude column	integer		6	2	Firefox	Yahoo
4	F	White	1961	date_time	\mathbf{i}	8	6	Firefox	Google
5	М	White	1954	date		2	3	Internet Explorer	Bing
6	М	African America	an 1982	time		15	4	Internet Explorer	Google
7	М	African America	an 1981	D		11	2	Firefox	Google
3	М	White	1977	S		3	3	Internet Explorer	Yahoo
9	F	African America	an 1969	м		6	2	Firefox	Google
0	М	White	1987	S		12	1	Safari	Yahoo
1	F	Hispanic	1959	D		12	5	Chrome	Google
	1								

図 3-20. データ型、ロール、日付フォーマットの設定と列のインポート

12) 図 3-21 のように、最後のステップでは RapidMinerBook リポジトリをデータの保 存場所として選択し、データセットに Chapter03DataSet と名前をつけます。

Finish をクリックすると、私たちが作りたい様々なタイプのデータマイニング プロセスでこのデータセットを利用できるようになります。

Import Data - Where to store the data?			×
w	here to store the data?		
Local Repository (10589759)			
RapidMinerBook (10589759) Cloud Repository (disconnected)			
,			
×			
Name Chapter03DataSet			
Location //RapidMinerBook/Chapter03DataSet			
		Einish Cance	

図 3-21. リポジトリの選択とデータセット名の設定

13) Finish をクリックすると、結果画面(Result Perspective)、または結果ビュー(Result view)と呼ばれるビューにデータが表示されます。

Interprocess	– RapidMine	Studio Edu	cational 9.0.001 @ L22	2419AA									- 🗆 ×
Eile Edit Proce	ss <u>V</u> iew <u>C</u>	onnections	Cloud Settings	Extensions	<u>H</u> elp								
	-		•		Views:	Design	Results T	urbo Prep A	uto Model		Find data,	operato	rsetc 🔎 All Studio 🔻
Result History		Example	Set (//RapidMiner	Book/Chapte	r03DataSet)	\times \land	、 、						Repository ×
	ExampleSe	t (11 examp	les, 0 special attribut	les, 15 regular	attributes)		\mathbf{X}		Filter (11 / 1	1 examples): al	I	•	🕒 Import Data 🛛 = 👻
Data	Row No.	Gender	Race	Birth_Year	Marital_Stat	Years_on_In	Hours_Per	Preferred_B	Preferred_S	Preferred_E	Read_Ne	Onl	Training Resources (connection)
0010	1	м	White	1972	М	8	1	Firefox	Google	Yahoo	Y	N	Samples
	2	м	Hispanic	1981	S	14	2	Chrome	Google	Hotmail	Y	Ν	Community Samples (cont B DB
Σ	3	F	African American	1977	S	6	2	Firefox	Yahoo	Yahoo	Y	Y	 Local Repository (1058975)
Statistics	4	F	White	1961	D	8	6	Firefox	Google	Hotmail	N	Y	🕶 🌅 RapidMinerBook (1058975)
	5	М	White	1954	М	2	3	Internet Explo	Bing	Hotmail	Y	Y	Chapter03DataSet (108
	6	м	African American	1982	D	15	4	Internet Explo	Google	Yahoo	Y	N	Cloud Repository (disconne)
Charts	7	М	African American	1981	D	11	2	Firefox	Google	Yahoo	?	Y	\backslash
	8	М	White	1977	S	3	3	Internet Explo	Yahoo	Yahoo	Y	?	
	9	F	African American	1969	М	6	2	Firefox	Google	Gmail	N	Y	
Advanced	10	М	White	1987	S	12	1	Safari	Yahoo	Yahoo	Y	?	
Charts	11	F	Hispanic	1959	D	12	5	Chrome	Google	Gmail	Y	Ν	

図 3-22. 結果画面(Result Perspective)または結果ビュー(Result view)

14) 図 3-23 の画面の左側では、データの記述統計量(Σ アイコン)を見たり、データ のチャートやアノテーションを作れることがわかるでしょう。例えば、Statistics

をクリックしたら、すぐにインターネット歴の平均が 8.818 年であるとか (Years_on_Internet 変数をクリックすると、標準偏差(Deviation)が 4.332 であるな ども見られます)、Preferred_Search_Engine 変数の統計から Google(7)が最も人気 のサーチエンジンであることなどがすぐにわかるでしょう。

Result History		ExampleSet (//RapidMinerB	ook/Ch	apter03DataSet)	×				
		Name	$\left \cdot\right $	Туре	Missing	Statistics	Filter (15 / 15 attributes):	Search for Attributes	
Data	~	Gender		Polynominal	0	F (4)	Most M (7)	Values M (7), F (4)	
Statistics	~	Race		Polynominal	0	Least Hispanic (2)	Most White (5)	Values White (5), African Americar	
<u> </u>	~	Birth_Year		Integer	0	^{Min} 1954	Max 1987	Average 1972.727	
Charts	~	Marital_Status		Polynominal	0	Least M (3)	Most D (4)	Values D (4), S (4),[1 more]	
Advanced	~	Years_on_Internet		Integer	0	Min 2	Max 15	Average 8.818	
Charts	~	Hours_Per_Day		Integer	0	Min 1	Max 6	Average 2.818	
Annotations	~	Preferred_Browser		Polynominal	0	Least Safari (1)	Most Firefox (5)	Values Firefox (5), Internet Explore	
	~	Preferred_Search_Engine		Polynominal	0	Least Bing (1)	Google (7)	Values Google (7), Yahoo (3),[1	
	~	Preferred_Email		Polynominal	0	Least Gmail (2)	Most Yahoo (6)	Values Yahoo (6), Hotmail (3),[1	
	~	Read_News		Polynominal	1	Least N (2)	Most Y (8)	Values Y (8), N (2)	

図 3-23. 記述統計量

15) 画面上部にある Design ビューボタンをクリックし、デザイン画面(またはビュー)へ切り替えます。そして、RapidMinerBook レポジトリ下の Chapter03DataSet
 を選択して、プロセスウィンドウヘドラッグ&ドロップしましょう。

Repository ×	Process
🕂 Import Data 🛛 🗉 🔻	Process
Training Resources (connected)	Process
🕨 🔁 Samples	
Community Samples (connected)	D inp
🕨 📕 DB	Retrieve Chapter03DataSet
Local Repository (10589759)	out
RapidMinerBook (10589759)	
Chapter03DataSet (10589759 - v1, 8/4/18 3:1:	
Cloud Repository (disconnected)	
< >	

図 3-24. リポジトリ下のデータをプロセスへ追加

16) プロセスウィンドウにある長方形のものをオペレーターと言います。 プロセス ウィンドウにデータセットをドラッグ&ドロップすることによって、 Chapter03DataSet にアクセスする Retrieve オペレータを追加しました。Retrieve オ ペレータをクリックして選択すると、このオペレータが Chapter03DataSet にアク セスされていることが右側にあるパラメータパネルでわかります。オペレータ が選択されている時は、オペレータの名前と外枠がオレンジに変わります。 Retrieve オペレータは単純にデータセットをリポジトリから取得し、プロセスウ ィンドウで利用できるようにしてくれます。オペレータの端とプロセスウィン ドウの端にある小さな半円は、ボートと呼ばれています。Retrieve オペレータの 右側にアウトプット (out)ボートが、プロセスウィンドウの右側に結果(res)ボー トが見えるでしょう。この2つのポートを線でつなぐことができます。out ポー トをクリックしてから、res ポートをクリックするとポート同士が線でつながり ます。

Process ×		
Process	100% 🔎 🔑 🔎 🛃 🧃	$\overline{\mathbf{O}}$
Process		
Dinp		res
Retrieve Chapter03DataSet	\int	res (
out out		

図 3-25. out ポートを res ポートへ接続

17) オペレータを組み合わせて線でつなぎ、データマイニングの流れを構築します。 データマイニングプロセスを実行してその結果を見るには、図 3-26 の矢印にあ

る青い三角の実行ボタンをクリックします。これにより、先ほどのデータセッ トのインポートが完了した時に見られたような結果画面に切替わります。



図 3-26. プロセスを実行し結果を確認

18) 画面の最上部にある Design と Results ボタンを使うことにより、デザイン画面と 結果画面を切り替えることができます。結果画面は、どこに欠損値があり、そ れをどう処理するかを決めるのに非常に役立ちます。Online Game 変数を例にあ げましょう。結果画面の Statistics タブでは、'N'が6件、'Y'が2件、そして欠損 値 (Missing) が3件あることがわかります。

	1	• •	Views:	Design	Results	Turbo Prep Auto Model	
Result History							
		Name	• - Туре		Statistics	Filter (15 / 15 attributes)): Search for Attn
Data	~	Preferred_Browser	Polynominal	0	Least Safari (1)	Most Firefox (5)	Values Firefox (5), Inte
Statistics	~	Preferred_Search_Engine	Polynominal	0	Least Bing (1)	Most Google (7)	Values Google (7), Ya
	~	Preferred_Email	Polynominal	0	Least Gmail (2)	Most Yahoo (6)	Values Yahoo (6), Hot
Charts	~	Read_News	Polynominal	1	Least N (2)	^{Most} Y (8)	Values Y (8), N (2)
Advanced Charts	~	Online_Shopping	Polynominal	2	Least N (4)	Most Y (5)	Values Y (5), N (4)
	~	Online_Gaming	Polynominal	▼ 3	Least Y (2)	(N (6)	Values N (6), Y (2)
Annotations	~	Facebook	Polynominal	0	Least N (3)	^{Most} Y (8)	Values Y (8), N (3)

図 3-27. 結果画面の記述統計量

中心的傾向の測定値 — 平均、中央値、最頻値(モード) — はデータ内に見 つかる欠損値を置き換えるのに使用されることがあります。最頻値(モード)、 または最も一般的な値で欠損値を置き換えることもできます。もちろん、これ は最頻値(モード)や最も一般的な値がすべての観測の正確な値であることを 前提としていて、これは間違っている場合があります。ですから、これから RapidMiner でどのように欠損値を扱うかを、現在の例のデータ中の欠損値を変 更して見せますが、それが欠損値を扱う上で常に適切な方法ではないというこ とに注意して下さい。RapidMiner で Online_Gaming 変数内の3件の欠損値を'N' に変更するためには、画面上部の Design ボタンを選択し、デザイン画面へ戻り ます。



図 3-28. 欠損値(missing values)を処理するオペレータを検索

19) 図 3-28 のオペレータパネルで該当のオペレータを見つけるために、フォルダの 階層を移動したり、検索することができます。RapidMiner はたくさんのオペレ ータを提供しているため、時々、自分が欲しい物を探すのが難しい場合があり ます。図 3-28 の中で黒い矢印で示されている便利な検索ボックスへキーワード を入力することで、該当のオペレータを見つけやすくするができます。この検 索ボックスに'missing'とタイプしてみましょう(図 3-28 参照)。RapidMiner が自動

でオペレータ群の中からオペレータの名前または概要(description)にこのワード が含まれるものを探します。検索するといくつか見つかりましたが、私たちは、 欠損値を置き換えたいと思っていますので、Cleansing フォルダーの中の Missing サブフォルダにある Replace Missing Values オペレータを選択し、図 3-29 のよう にプロセスウィンドウの中の線の上にドラッグしましょう。マウスのカーソル を線上に合わせると、線が少し太く変わります。そして、マウスボタンを離す と、オペレータが接続されます。Replace Missing Values オペレータを離しても接 続されない場合には、手動で線をつなぐことが可能です。Retrieve オペレータの out ポートをクリックしてから、Replace Missing Values オペレータの exa ポートを クリックするだけでオペレータ同士が接続されます。exa は Example Set のこと を表しています。'examples'というワードは RapidMiner の中では、データセット における行として使われています。プロセスを実行したときに、結果が表示さ れるように、Replace Missing Values オペレータの右側にある exa ポートから結果 (res)ポートに接続されていることを確認しましょう(図 3-29)。



図 3-29. Replace Missing Values オペレータを追加

20) RapidMiner の中で、オペレーターが 選択されているときには、オペレータの外 枠の長方形がオレンジ色になります(図 3-29 で Replace Missing Values オペレータ が選択されているのが見られます)。この時に、選択されたオペレータのパラメ

ータやプロパティを変更することができます。今回のチュートリアルでは、 Online_Gaming 変数の欠損値3つ全てを'N'で置き換えることに決めました。な ぜならこれがこの変数における最も多い回答(モード)だからです。これを行う ために、図 3-30 のように attribute filter type を 'single' に変更します。すると attribute と ラベルのついたドロップダウンリストが現れますので、' Online_Gaming' 変数を選択します。次に default とついたドロップダウンリスト の候補の中から'value'を選択します。これにより、replenishment value ボックス が表示されます。図 3-30 のように'N'をこのボックスにタイプします。選択した オプションに応じて使用可能なオプションが変わるため、全てのオプションを 表示するには、画面を広げるか、パラメータパネルの右側にある縦方向のスク ロールバーを使用する必要がある場合があるので、注意しましょう。



図 3-30. 欠損値のパラメータ設定

21) パラメータパネルの中には、他にもたくさんのオプションが用意されています。 ここでは、全てを見てみることはしませんが、自由に試して見てください。例 えば、'attribute filter type'の 'subset'を使用することで、データセット内の複数

(subset)の変数を変更することができますが、今回は'single'のままにしておいて 下さい。パラメータをセットしたら、実行ボタンをクリックしましょう。プロ セスが実行され、結果画面に切り替わりますので、次に Statistics アイコンをク リックします。結果画面が図 3-31 のように見えるはずです。

Result History		📕 ExampleSet (Replace Missi	ng Val	ues) ×				
		Name	ŀł	Туре	Missing	Statistics	Filter (15 / 15 attributes):	Search for Attributes
Data	~	Online_Gaming		Polynominal	0 ◀	Least Y (2)	N (9)	Values N (9), Y (2)
Statistics	~	Gender		Polynominal	0	F (4)	Most M (7)	Values M (7), F (4)
	~	Pace		Polynominal	0	Least Hispanic (2)	Most White (5)	Values White (5). Afric

図 3-31. 欠損値を変更した結果

22) Online_Gaming 変数がリストの一番上に移動し、欠損値(Missing)が0になったことが見えるでしょう。次に Data アイコン(statistics アイコンのすぐ上にある) をクリックしましょう。Online_Gaming 変数には'Y'と 'N'の値のみが入力されていることが分かります。この変数の欠損値を全て置換することに成功しました。 一方で、Online_Shopping 変数には2つの欠損値があります。疑問符(?)は、欠損値を表しています。この変数では、nulll値を最頻値(モード)で置き換えず、レコードごとデータを削除します。

レコードの削除

デザイン画面に切り替えます。 次は、データセット内のレコードを削除する方法につ いて説明します。

 オペレータパネル内の検索ボックスに、'filter'という単語を入力します。これは、この例で使用する Filter Example オペレータを見つけるのを助けてくれます。 Filter Examples オペレータをドラッグして、Replace Missing Values オペレータのすぐ後に接続します。プロセスウィンドウは図 3-32 のようになります。



図 3-32. フィルターを追加

Filter Examples オペレーターを選択した状態で、パラメータパネルを確認します。
 Add Filters…ボタンをクリックします。すると、図 3-33 のように Create Filters 画
 面が表示されますので、最初のドロップダウンリストで Online_Shopping 変数を
 選択します。次に2番目のドロップダウンリストで 'is missing' を選択します。

^{*} portions of this book are adapted from Data Mining for the Masses, by Matthew North, copyright 2020.

🧼 Create Filters: filters	×
Create Filters: filters Defines the list of filters to apply.	
Online_Shopping	*/ *
Match all Match any Preselect comparators	Add Entry

図 3-33. フィルターを作成

3) OK をクリックして実行し、結果(Result)画面に切り替わると、Online_Shopping に欠損値があった2件しか表示されていないことがわかります。他のすべての レコードは削除されているため、これは意図したものではありません。私たち は Online_Shopping の値が欠損している 2つのレコードを削除し、他のレコード を残したいと考えています。これを直す方法は 2つあります。 デザインビュー に切り替えます。 パラメータパネルに、'invert filter'というチェックボックスが 表示されています。



図 3-34. invert filter オプション

4) これをクリックしてプロセスを再実行すると、Online_Shopping 変数の欠損値以 外の値を持つ9件のレコードが保持され、2件がフィルターにより削除されてい

ることがわかります。または、Add Filters… ボタンをもう一度クリックして Create Filter ウィンドウを再度開き、今度は 'is missing' ではなく 'is not missing' を 選択して、問題を修正することもできます。これでも同じ結果が得られます。 1 つの Filter Examples オペレータで複数のフィルターを追加できることも覚えて おいてください。Create Filters ウィンドウを開いた状態で、Add Entry ボタンを クリックすれば、他の変数のフィルターを設定するための 2 行目が作成されま す。さまざまな条件でフィルタリングを試してみてください。

フィルタリング以外の方法でもデータを削減することができます。 以下の手順に従っ て、RapidMiner でデータセットのサンプリングを実施しましょう。

 すぐ前に紹介した検索のテクニックを使用して、オペレータ検索機能を使用して Sample というオペレータを見つけ、これをプロセスへ追加します。パラメー タパネルで、サンプルを'relative'(相対)サンプルになるように設定し、サンプル 比率(sample ratio)フィールドに.5と入力して、結果のデータセットが全体の 50% を保持するように指定します。 画面は図 3-35 のようになります。



図 3-35. データセットのランダムサンプリング(50%)

 2) ここでプロセスを実行すると、結果画面には Filter Examples オペレーターが Online_Shopping の値を持たないレコードを削除した後に残った 9 件のデータか ら、ランダムに選択された 4 件または 5 件のデータのみがあるのがわかるでし ょう。

このように、データセット内のレコード数を減少させてデータを削減させる方法は多 く存在し、さまざまな理由があることがわかります。次に、不整合(inconsistent)データ の処理に移りますが、フィルタリングを実施した際に不整合データを削除してしまっ たので元の状態に戻す方法について説明します(図 3-37 のような状態にします)。デ ザイン画面に戻り、Sampleオペレータをクリックします。次に、右クリックして[削除 (Delete)]を選択するか、キーボードの Delete キーを押します。次に Filter Examples オペ レータをクリックします。 削除するのではなく、これを無効にします。無効にするに は、右クリックして Enable Operator オプションを選択するか(トグルスイッチでオン /オフを切り替える仕様)、またはキーボードで Ctrl + E を押します(こちらもトグ ル)。 無効にした Filter Examples オペレータを、メインプロセスの隅にドラッグして、 邪魔にならないようにします。結果(res)ポートに繋がっていた線が消えている場合は、 Replace Missing Values オペレータの exa ポートから res ポートに再接続します。プロセス が図 3-37 のような状態になっていることを確認して下さい

不整合データの対処

不整合データは欠損値とは違います。不整合データとは、値は存在しても、それが妥 当な値または意味のある値ではないことを指します。Twitter 変数の中には不整合デー タが存在します(図 3-36)。

ie_Sho	Facebook	Twitter	Other_Socia
	Y	Ν	?
	Y	Ν	?
	Y	Ν	?
	Ν	Y	?
	γ	Ν	?
	N	N ?!?!	?
	γ	Y /	LinkedIn
	γ	99 🗡	LinkedIn
	N	Ν	?
	Y	Ν	MySpace

図 3-36. Twitter 変数の不整合データ

99は何をあらわしているのでしょう?Twitter 変数の有効な値は「Y」と「N」の2つだ けのようです(つまり、回答者は Twitter を使用しているか、使用していないかのどち らかです)。これらの値は統一されていないので、意味がありません。 データマイニ ングを行う者として、Online_Shopping に欠損値があったレコードの場合と同様に、レ コードをフィルターで除外するか、特定の値を他の値で置き換えられるオペレータを 使用するかを選択することができます。

 もし他の画面を開いていたら、デザイン画面へ戻りましょう。図 3-37 のような 画面になるように、Sample オペレータを削除(Delete)し、Filter Examples オペ レータを削除もしくは無効(Ctrl + E)にしてあることを確認しましょう。



図 3-37. Filter Examples オペレータを無効にし、Sample オペレータを削除

- Replace Missing Value オペレータは、Online_Gaming 変数の値のみを変更するだけですので、データセットのレコードを削除することはないため、削除する必要はありません。図 3-38 のようにオペレータパネルの検索機能を使用して、 Replace(置換)と呼ばれるオペレータを検索し、プロセスへ追加します。
- パラメータパネルで、attribute filter type を 'single' に変更し、attribute ~ 'Twitter' を指定します。 replace what のフィールドへ、これから私たちが置換したいと思 っている値、99を入力しましょう。そして、回答者の約 80%が Twitter を使用し ていないと回答しているので、今回は、'replace by'フィールドへ最頻値(モード) の「N」を入力します。



図 3-38. Replace オペレータのパラメータ設定

4) プロセスを実行し、Statistics(統計)タブをクリックします。 図 3-39 を見ると、 Twitter 変数には「N」の値が 9つ、「Y」の値が 2つあることがわかります。

Result History		📒 ExampleSet (Replace)	×					
		Name	١H	Туре	Missing	Statistics	Filter (15 / 15 attributes):	Search for Attribut
Data	~	Twitter		Polynominal	0	Leest Y (2)	Most N (9)	Values N (9), Y (2)
Statistics	~	Online_Gaming		Polynominal	0	Least Y (2)	Nost N (9)	Values N (9), Y (2)
				Determined	0	Least	Most	Values M (7) E (4)

図 3-39. 不整合な値を整合性のある値で置換

変数の削減

多くのデータセットにおいて、与えられた質問への回答を出すのに無関係な変数があ ることがわかるでしょう。第4章では、相関関係または特定の変数間の関係の強さを 評価する方法について説明します。データセット内の特定の変数が、特定の質問に答 えるのに関係ないという理由だけで、それらの変数が絶対に関心のある変数にならな いということではないということを覚えておいてください。 このような理由から、こ の第3章の前半でデータセットをインポートするときにすべての変数を取り込みまし た。興味のない変数や無関係な変数は、次の手順で簡単に除外できます。

 デザイン画面に戻ります。オペレータの検索フィールドで、'Select Attributes' と 入力します。Select Attributes オペレータがオペレータのエリアの中のフォルダ の階層の中に現れますので、ドラッグして Replace オペレータと結果ポート (res) の間に収まるようにします。画面は図 3-40 のようになります。

^{*} portions of this book are adapted from Data Mining for the Masses, by Matthew North, copyright 2020.



図 3-40. データセットの変数を選択

- 2) パラメータパネルで、attribute filter type を 'subset' に設定し、Select Attributes ボタ
 - ンをクリックします。 図 3-41 のようなウィンドウが表示されます。

Ø Select Attributes: attributes	×
Select Attributes: attributes The attribute which should be chos	en.
Attributes	Selected Attributes
Search	Search
Birth_Year A Facebook Gender Hours_Per_Day Marital_Status Online_Gaming Online_Shopping Other_Social_Network	•
Preferred_Browser Preferred_Email Preferred_Search_Engine Race Read_News	
	Apply K Cancel

図 3-41. 変数選択画面

 青い右矢印と左矢印を使用して、変数を選択できます。この例では、Birth_Year、 Gender、Marital_Status、Race、そして Years_on_Internet を選択し、右矢印を使用 して右側の Selected Attributes (選択した変数)エリアへ移動させます。一度に複数 の変数を選択するには、コントロール(Ctrl)またはシフト(Shift)キーを押しなが

ら、選択したい、または除外したい変数をクリックします。Apply ボタンを選択しプロセスを実行すると、選択した変数のみが結果画面に表示されます。

その他のデータ操作

この章ではこれまで、RapidMiner リポジトリにインポートした第3章のデータセット を使用して作業してきました。しかし、時間の経過とともに変化する可能性のあるデ ータセットがあり、それが増大したり変化したりする場合ではどうでしょうか。この 例では、データセットを RapidMiner リポジトリへインポートせずに、直接 CSV ファイ ルを読み取る方法についてご紹介します。

1) オペレータの検索ボックスで、「Read」と入力します。



図 3-42. RapidMiner で Read 系のオペレータを検索

 Read CSV オペレータをプロセスウィンドウにドラッグ&ドロップします。この 例では、前に使用した全てのオペレータをプロセスから削除しています (Retrieve、Replace Missing Values、Replace、Select Attributes オペレータなど)。



図 3-43. Read CSV オペレータのパラメータ

パラメーターを表示できるように、Read CSV オペレータが選択されてい 3) ることを確認してください。データファイルの区切り記号が第3章のデータセ ットのようにカンマである場合は、column separators パラメーターをカンマ (,) に変更します。次に csv file ボックスを無視して、代わりに Import Configuration Wizard ボタンをクリックしてください。これにより、データインポートウィザ ードが開きます。 既存のリポジトリに追加するでも、新しいリポジトリを作成 するのでもなく、コンピューター上またはネットワーク上のどこかにある.csv ファイルへアクセスすることが可能です。まず、最初のステップで第3章のデ ータファイルを選択し Next ボタンをクリックします。そして、Column Separator がカンマとなっていることを確認し、データファイルの一行目が列名として正 しく認識されていることを確認後に Next ボタンをクリックします。データ型と ロールの設定は何も変更せずに Finish ボタンをクリックします。図 3-43 のよう に、Read CSV オペレータの出力ポート(out)が res ポートへ接続されていることを 確認してから、プロセスを実行します。結果画面に表示される内容が、以前に

リポジトリヘインポートしたデータとまったく同じであることが確認できるでしょう。

この先の章のチュートリアルでは、データファイルのインポート(import)をするのでは なく読み取り(read)をします。

新機能について

2018 年の夏に、RapidMiner は RapidMiner Studio アプリケーションの一部として新機 能 TurboPrep をリリースしました。TurboPrep を使用すると、データを見ながら前処理 ができます。また、TurboPrep は、データの異常またはエラーを自動的に修正する自動 クレンジング機能も備えています。さらに、ピボットテーブルを作成したり、データ を可視化することができます。

TurboPrep 機能の詳細は RapidMiner の Web サイトで確認することができます: <u>https://www.rapidminer.jp/rapidminer-studio/turboprep-automodel/</u> RapidMiner Studio をダウンロードしてインストールすると、TurboPrep を試すこともで きます。RapidMiner Studio のインストール方法は下記ブログで紹介されています。

RapidMiner の始め方~10step でできる簡単インストール方法~

https://www.ksk-

anl.com/blog/rapidminer%E3%81%AE%E5%A7%8B%E3%82%81%E6%96%B9%EF%BD%9E10step %E3%81%A7%E3%81%A7%E3%81%8D%E3%82%8B%E7%B0%A1%E5%8D%98%E3%82%A4%E 3%83%B3%E3%82%B9%E3%83%88%E3%83%BC%E3%83%AB%E6%96%B9%E6%B3%95%EF%B D%9E/

※本チュートリアルは 2020年04月28日 時点のものです。