Методы восстановления пропусков в данных

Каюмов Эмиль

ММП ВМК МГУ

Спецсеминар
«Алгебра над алгоритмами и эвристический поиск закономерностей»

25 ноября 2015

План

Содержание

Зачем это нужно?

Большинство реальных данных имеют пропущенные значения.

- Ошибки при записи.
- Ошибки при измерении.
- Невозможность сбора.

Далеко не все алгоритмы умеют работать с неполными данными.

MCAR

- Missing completely at random

$$P(M_i|X_{osb}, X_{mis}, \theta) = const$$

Вероятность пропуска не зависит ни от значений наблюдаемых, ни от значений пропущенных данных.

Пример: среди пациентов только у случайной части измерили массу.

MAR

- Missing at random

$$P(M_i|X_{osb}, X_{mis}, \theta) = f(X_{obs}, \theta)$$

Вероятность пропуска зависит от значений наблюдаемых, но не от значений пропущенных данных.

Пример: среди пациентов масса измеряется только у тех, у кого высокое давление.

NMAR

- Missing not at random

$$P(M_i|X_{osb}, X_{mis}, \theta) = f(X_{obs}, X_{mis}, \theta)$$

Вероятность пропуска зависит от значений и наблюдаемых, и от значений пропущенных данных.

Пример: среди пациентов взмешивают только тех, кто имеет избыточную массу.

Вывод

Важно понимать:

- Из какого источника и как были получены данные.
- Какой тип пропущенных значений соответсвует каждому признаку.

Содержание

Простейшие методы

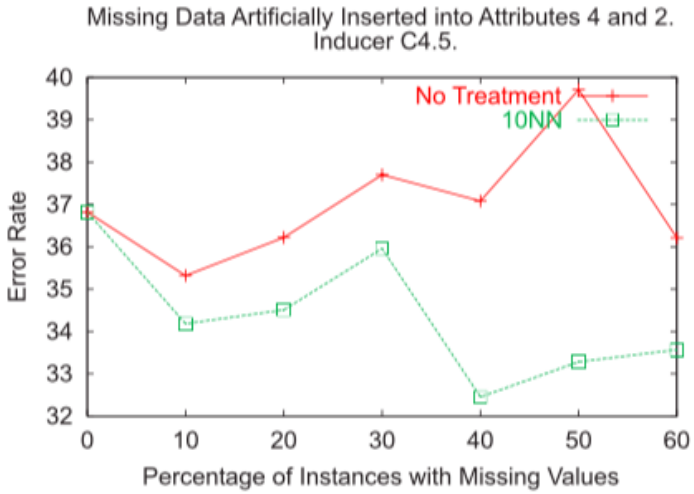
- Удаление объектов с пропущенными значениями (можно удалять не объекты, а признаки).
  - Ничего не испортим, но что если данных и так мало?
- Замена случайным значением.
- Замена специальным значением (индикатор пропущенного значения).
  - Как понимать специальное значение в случае вещественного признака?

Простые методы

- Замена средним значением признака.
- Замена медианой признака.
- Замена модой признака.
- Можно вычислять по каждому классу в отдельности.
  - Можно сделать только на обучающей выборке.
- Размножить выборку всеми возможными значениями пропущенного признака.
  - Необходимы эвристики для объединения результатов размноженного объекта.

Метод ближайших соседей

- Необходимо определить метрику и число соседей.
- Подходит и для категориальных признаков, и для непрерывных.
- Просто в случае одного пропущенного атрибута, но необходимы эвристики для множественных замен (например, искать только среди полностью заполненных объектов).



- Per Jonsson, Claes Wohlin «An Evaluation of k-Nearest Neighbour Imputation Using Likert Data».
- Gustavo Batista, Maria Carolina Monard «A Study of K-Nearest Neighbour as an Imputation Method».

Closest Fit

- Аналогично методу ближайших соседей, но с заданной метрикой, учитывающей тип признака.
- Простое использование в случае множественной замены.

$$dist(x', x'') = \sum_{i=1}^{d} dist(x_i', x_i'')$$

$$dist(x_i', x_i'') = \begin{cases} 0, & x_i' = x_i'' \\ 1, & x_i' \neq x_i'' \ if \ categorical \\ \frac{|x_i' - x_i''|}{max(x_i) - min(x_i)}, & x_i' \neq x_i'' \ if \ numerical \end{cases}$$

- Jerzy Grzymala-Busse, Witold Grzymala-Busse, and Linda Goodwin «A Closest Fit Approach to Missing Attribute Values in Preterm Birth Data».

Нейронная сеть

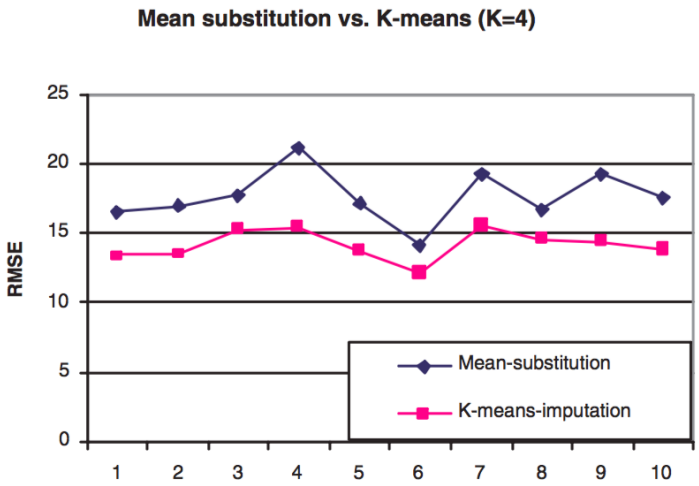
- В отличие от других алгоритмов для предсказывания, одна сеть – один паттерн пропущенных данных.
- Вход – признаки, известные в данном паттерне, выход – неизвестные признаки паттерна.

	Average CCRs	
Data set	Training (%)	Test (%)
EPS data set		
Complete	99.09	58.00
Reconstruction methods:		
bp	95.65	74.35[a]
regression	96.09	67.39
average	95.22	64.35
zero	96.09	71.30

- Amit Guptaa, Monica Lam «The weight decay backpropagation for generalizations with missing values».

Метод k средних

1. Выбираем центры кластеров как случайные k объектов без пропусков в данных.
2. Действуем по стандартному алгоритму k средних.
3. Заполняем пропущенные значения как соответвующие значения центров кластеров или самих объектов кластера.



- Необходимо определять метрику и число кластеров.
- Метрика должна учитывать пропущенные значения.

- Dan Li, Jitender Deogun, William Spaulding, Bill Shuart «Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method».

Метод нечётких k средних (Fuzzy K-means)

- Теперь объект принадлежит не конкретному кластеру, а каждому в определённой степени.

Принадлежность x_i кластеру k:

$$U(v_k, x_i) = \frac{d(v_k, x_i)^{2/(m-1)}}{\sum_{j=1}^{K} d(v_j, x_i)^{2/(m-1)}}$$

Центроиды кластеров:

$$v_k = \frac{\sum_{i=1}^{N} U(v_k, x_i) x_i}{\sum_{i=1}^{N} U(v_k, x_i)}$$

Пропущенные значения:

$$x_{i,j} = \sum_{k=1}^{K} U(v_k, x_i) v_{k,j}$$

RMSE:

	Manhattan Distance	Euclidean Distance	Cosine-based Distance
K-means	13.37	14.08	17.65
Fuzzy K-means	11.12	11.77	14.99

- Dan Li, Jitender Deogun, William Spaulding, Bill Shuart «Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method».

Кроме того

- EventCovering - аппроксимация смеси распределений одним дискретным с минимальными потерями информационного критерия. Опирается на rough set и кластеризует объекты.
  - Andrew Wong, David Chiu «Synthesizing statistical knowledge from incomplete mixed-mode data».
- Максимум правдоподобия и EM-алгоритм.
  - A. P. Dempster, N. M. Laird, D. B. Rubin «Maximum Likelihood from Incomplete Data via the EM Algorithm».
  - James Honaker, Gary King «What to Do about Missing Values in Time-Series Cross-Section Data».
- SVM
  - Feng Honghai, Chen Guoshun, Yin Cheng, Yang Bingru, Chen Yumei «A SVM Regression Based Approach to Filling in Missing Values».

Содержание

Эксперимент 1.

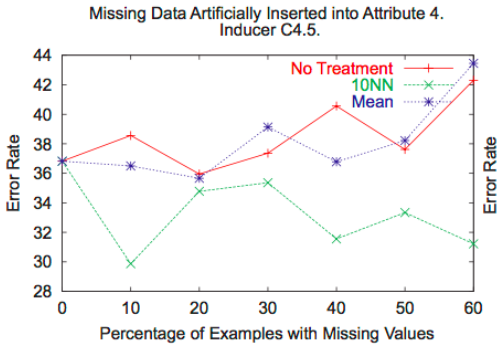
- Gustavo Batista, Maria Carolina Monard «An Analysis of Four Missing Data Treatment Methods for Supervised Learning».

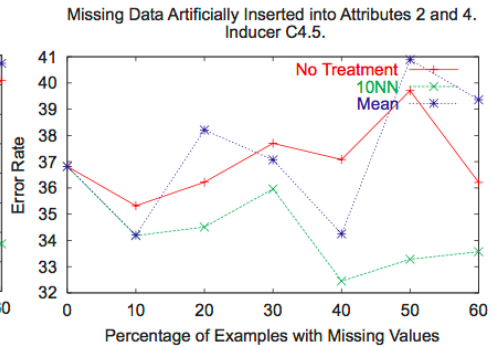
Data set	# Instances	#Duplicate or conflicting (%)	#Attributes (quanti., quali.)	Class	Class %	Majority Error
bupa	345	4 (1.16%)	6 (6,0)	1 2	42.03% 57.97%	42.03% on value 2
cmc	1473	115 (7.81%)	9 (2,7)	1 2 3	42.70% 22.61% 34.69%	57.30% on value 1
pima	769	1 (0.13%)	8 (8,0)	0 1	65.02% 34.98%	34.98% on value 0
breast	699	8 (1.15%)	9 (9,0)	2 4	65.52% 34.48%	34.48% on value 2

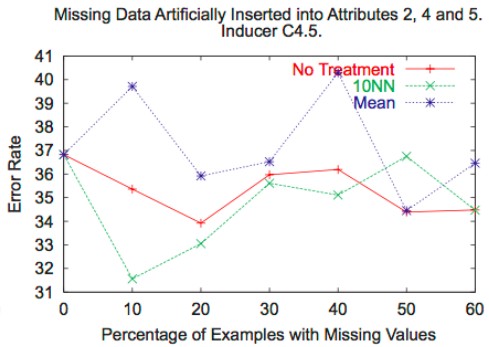
- C4.5
- 10-fold CV.
- Искуственная порча данных.

Эксперимент 1.

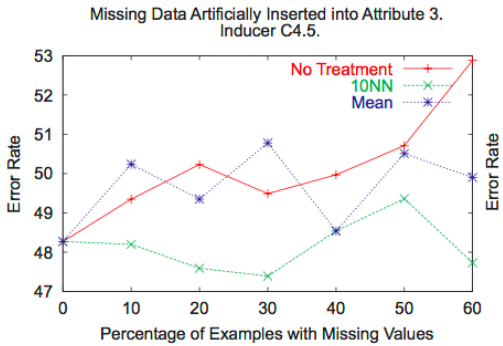
- bupa:

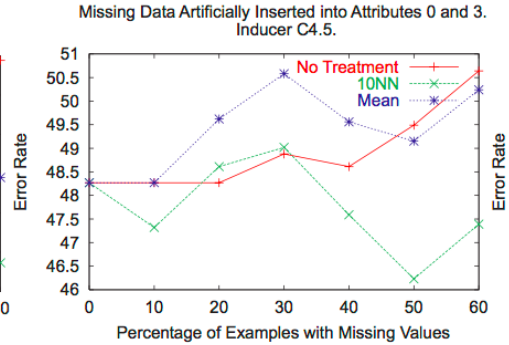


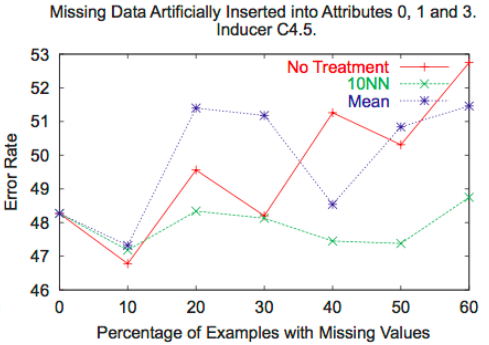




- cmc:

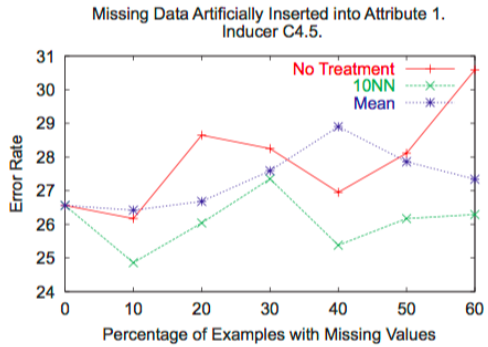


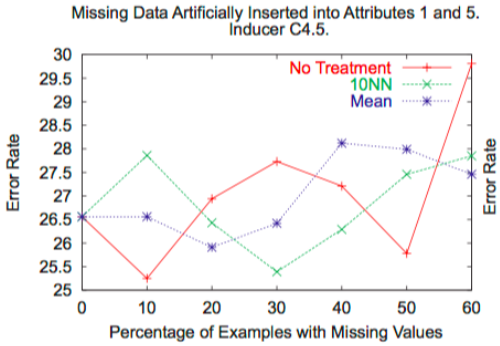


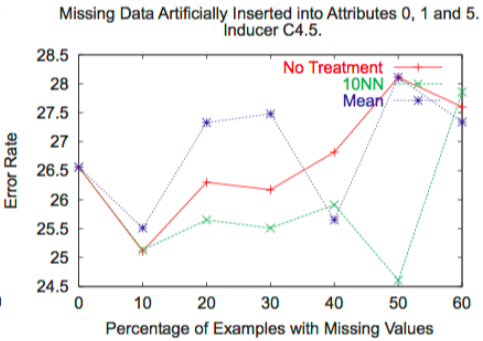


Эксперимент 1.

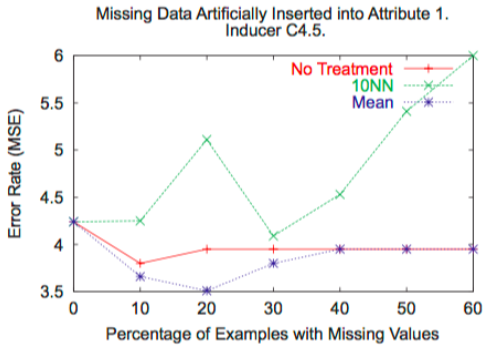
- pima:

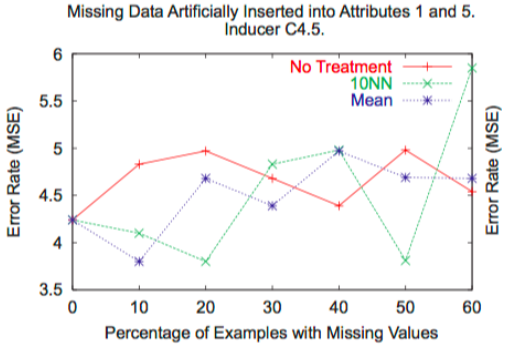


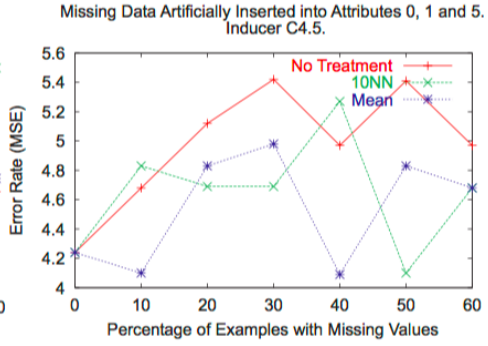




- breast:







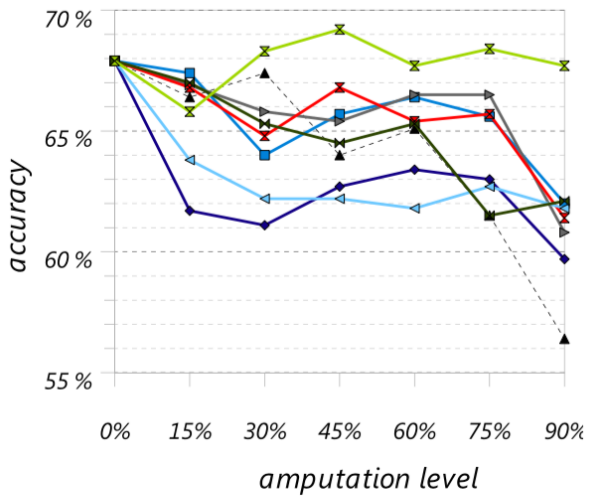
Эксперимент 2.

- Lars Wohlrab, Johannes Furnkranz «A Comparison of Strategies for Handling Missing Values in Rule Learning».

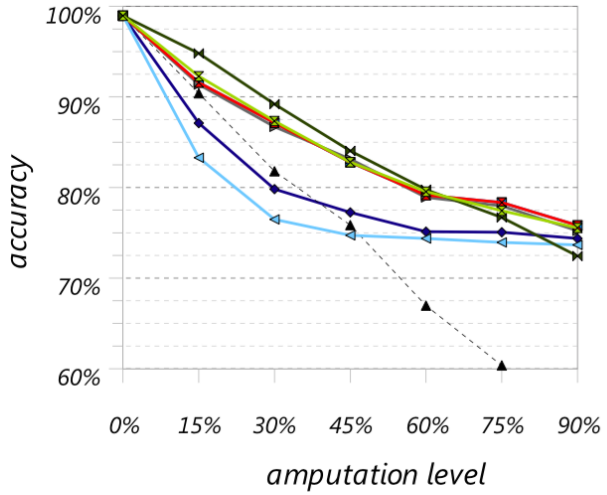
- Simple separate-and-conquer rule-learner (похож на CN2) - алгоритм на идее «правило индукции» (rule induction) как и деревья решений.
- 10-fold CV.
- Искусственная порча данных.

1 Delete Strategy
2 Ignored Value Strategy
3 Any Value Strategy
4 Special Value Strategy
5 Common Value Strategy
6 Pessimistic Value Strategy
7 Predicted Value Strategy
8 Distributed Value Strategy

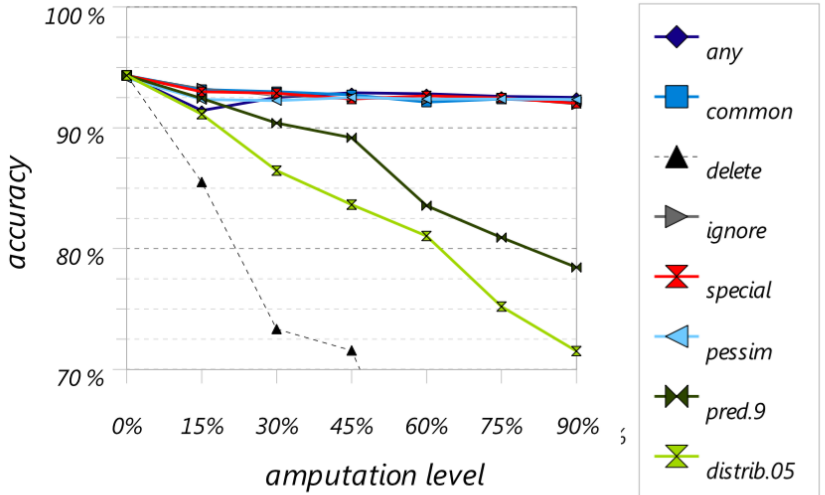
Эксперимент 2.



5.1.1: achieved accuracies (credit-g)



5.1.3: achieved accuracies (KRKP)



5.1.5: achieved accuracies (segment)

Содержание

Что реализовано?

- R: Amelia II, Mice – хорошо.
- Python: Pandas, Numpy, scikit-learn – слабо.
- Matlab – средне.

Неупомянутые статьи и книги

- Julian Luengo, Salvador Garcia, Francisco Herrera «A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: The good synergy between RBFNs and EventCovering method».
- Xi-Yu Zhou, Joon Lim «Replace Missing Values with EM algorithm based on GMM and Naive Bayesian».
- Benjamin Marlin «Missing Data Problems in Machine Learning».