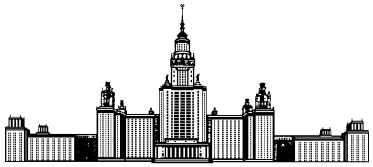
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Методы восстановления пропусков в данных»

Выполнил:

студент 3 курса 317 группы

Каюмов Эмиль Марселевич

Научный руководитель:

д.ф-м.н., профессор

Дьяконов Александр Геннадьевич

Москва, 2016

Содержание

1 Введение

В практических задачах анализа данных выборки часто содержат в себе пропущенные значения. Их причины могут быть различными, например, неответ респондента на конкретный вопрос, отказ работы датчика для измерений показателя при низких температурах, ошибки в программном обеспечении при записи данных.

За редким исключением алгоритмы машинного обучения не работают с выборками, имеющими пропущенные значения. Поэтому возникает необходимость перейти к данным, не имеющим пропусков, для дальнейшней работы с ними. Существуют различные подходы к решению данной задачи, которые различаются по своей природе, области применимости и вычислительной сложности.

В данной работе описаны основные методы восстановления пропусков в данных, произведено их сравнение на нескольких наборах данных, включая как данные с искусственно созданными пропусками, так и с данными, имеющими натуральные пропущенные значения.

2 Методы восстановления пропусков

2.1 Постановка задачи

Имеется матрица объектов-признаков, часть значений которой пропущены. Необходимо получить матрицу объектов-признаков без пропущенных значений с целью дальнейшего применения алгоритма машинного обучения.

2.2 Игнорирование объектов с пропущенными значениями

Простейшим методом решения проблемы пропущенных значений в выборке является игнорироваение объектов, имеющих пропуски. Такой метод применим только в том случае, когда малая часть объектов выборки имеет пропущенные значения. Преимуществом данного подхода является простота и невозможность испортить данные путем замены пропусков. В случае достаточно большого размера выборки метод может показывать хорошие результаты. Альтернативным вариантом в случае наличия пропусков в небольшом количестве признаков является удаление таких признаков из выборки.

2.3 Замена специальным значением

Другим простейшим методом является замена пропусков на специальное заранее определенное значение такое, как, например, 0 или -1. Данный подход позволяет не уменьшать размер выборки, однако может вносить значения, сильно отличающие от настоящих.

Для дальнейшего применения методами, основанными на деревьях, разумно заполнять пропуск с помощью значения, не встречающегося в выборке, например, -1 для неотрицательных значений признаков. Для методов, чувствительным к масштабу признаков, пропуск заменяется с помощью 0.

2.4 Замена самым частым или средним значением

Еще одним простым методом восстановления пропусков является замена на моду или среднее значение по конкретному признаку. В случае категориального

признака все пропуски заменяются на наиболее часто встречающее значение, в случае количественного признака – на среднее значение по признаку. Данный метод, в отличие от двух предыдущих, учитывает имеющиеся данные и усредняет их. Преимуществом подхода является простота, однако на практике возникает проблема в определении, является ли конкретный признак категориальным или количественным, особенно при их большом количестве.

2.5 Замена с помощью сингулярного разложения

В машинном обучении сингулярное разложение используется для приближения матрицы матрицей меньше ранга. В данном методе сингулярное разложение применяется сначала для сокращения размерности матрицы, после чего пропущенные значения заменяются по восстановленной из сингулярного разложения матрице меньшего ранга.

Для начальной инициализации используется замена пропущенных значений средним значением признака. Далее итеративно к матрице объектов и признаков применяется сингулярное разложение с отбрасыванием наименьших сингулярных чисел, после чего заменяются пропущенные значения по восстановленной матрице до сходимости или достижения максимального заданного числа итераций.

2.6 Замена с помощью метода ближайших соседей

Из предположения о том, что близкие объекты по значениям среди заполненных признаков близки в признаках, значение которых может быть пропущено (гипотеза компактности), возникает применение метода k ближайших соседей для восстановления пропусков в данных. Реализация аналогична использованию классичесокого метода k ближайших соседей за исключением того, предсказывается сразу несколько пропущенных признаков для каждого объекта.

2.7 Замена с помощью случайного леса

Заменить пропущенные значения в конкретном признаке можно, предсказав его по другим признакам с помощью одного из алгоритмов машинного обучения. Так

как среди признаков, по которым производится обучение, имеются пропущенные значения, то необходимо их изначально заменить с помощью одного из простейших методов восстановления пропусков.

В данном случае инициализацию пропущенных значений производиться с помощью замены средним значением по признаку, в качестве алгоритма для уточнения пропусков используется случайный лес. Для каждого признака, имеющего пропущенные значения, производится обучение по объектам, не имеющим пропусков в данном признаке, для оставшихся объектов производится замена значений данного признака с помощью обученного алгоритма. Эта процедура повторяется в течение нескольких итераций до сходимости или до максимального установленного числа итераций.

2.8 Замена с помощью линейной регрессии

Используется стратегия, аналогичная предыдущему методу, за исключением того, что вместо случайного леса используется линейная регрессия.

2.9 Замена с помощью EM-алгорима

Имея информацию о распределении данных в выборке, можно восстановить пропущенные значения. Используется EM-алгоритм для восстановления параметров нормального распределения. По средним значениям и матрице ковариации признаков вычисляются коэффициенты регрессии, с помощью которых вычисляются значения на местах пропусков по другим признакам этого объекта.

В качестве начальной инициализации пропущенных значений используется заполнение средним значением каждого признака. Далее итеративно восстанавливаются параметры распределения и уточняются пропущенные значения до сходимости или достижения максимального заданного числа итераций.

2.10 Замена с помощью метода k средних

Аналогично методу k ближайших соседей предполагается, что близкие по одним признакам объекты должны быть близки и по другим признакам. Однако в отличие

от метода k ближайших соседей ищутся не ближайшие соседи для каждого объекта с пропущенными значениями, а используется информация о центре кластера, в который попал конкретный объект с пропусками. Заметим, что для разбиения на кластеры необходима начальная инициализация пропущенных значений.

В данном случае выполняется инициализация пропущенных значений с помощью замены средним значением по признаку, кластеризация производится методом k средних. Пропущенные значения заменяются на соответствующие им значения центра кластера, в который попал каждый объект с пропусками. Данная процедура производится в течение нескольких итераций до сходимости или по достижению максимального заданного числа итераций.

2.11 Алгоритм ZET

Алгоритм ZET предложен Загоруйко Н.Г. [8]. В его основе лежат предложения об избыточности данных в таблице, локальной компактности и линейной зависимости.

Восстановление пропущенных значений происходит следующим образом: для каждого пропущенного значения в нормализованных данных выбираются компетентные строки и столбцы в заданном количестве. Под компетентностью для строк понимается величина, равная частному количества непропущенных значений в паре строк к их расстоянию по непропущенным значениям, для столбцов — произведение количества непропущенных значений в паре столбцов и модуля их коэффициента корреляции. Далее по известным значениям в строке и столбце с заменяемым пропущенным значением подбирается показатель степени учета компетентности во взвешенной сумме. Пропущенное значение заменяется взвешенной суммой предсказаний линейной регрессии по компетентным строкам и столбцам с коэффициентами степени компетентности этих строк и столбцов. Предсказание по строкам и столбцам усредняется.

2.12 Особенности реализации методов

1. При применении большинства методов можно получить нецелое число, в таком случае замена для категориальных признаков является некорректной. Можно вручную задавать маски для признаков, указывая, какие из них

являются категориальными, для дальнейшего округления. Однако данный вариант является трудоемким в случае большого числа признаков и не универсален. Более простым решением данной проблемы является округление до ближайшего значения конкретного признака среди имеющихся. В таком случае категориальные признаки не будут заполнены дробными значениями, а количественные признаки не получат большого искажения. Во всех дальнейших экспериментах применяется такой подход.

2. При замене пропущенных значений на практике нередко используется подход, связанный с добавлением нового признака, характеризующего наличие пропусков в конкретном признаке. Применение такого подхода во всех методах восстановления пропущенных значений будет рассмотрено отдельно.

3 Эксперименты с данными

Реализация всех описанных методов и эксперименты расположены в репозитории на Github [9].

В качестве алгоритмов машинного обучения, на основе которых сравниваются методы восстановления пропусков, использовались случайный лес, логистическая регрессия и метод k ближайших соседей. Такой выбор мотивирован тем, что эти алгоритмы имеют разную природу.

3.1 Исходные данные и условия эксперимента

Для сравнения качества восстановления пропусков в данных использовались шесть наборов данных, взятых из UCI Machine Learning Repository [3], три из которых имели натуральные пропущенные значения. Использование данных без пропусков мотивировано тем, что в таком случае можно управлять количеством пропущенных значений. Это позволяет сравнить методы для различных долей количества пропусков среди имеющихся значений. Кроме того, информация об истинных значениях пропусков позволяет сравнить восстановленные значения напрямую, а не только через качество работы алгоритма машинного обучения на полученных данных.

При использовании данных без пропусков необходимо создать пропущенные значения искусственно. Для этого существуют различные стратегии. В статье [6] предлагается создавать пропуски для случайных выбранных объектов в нескольких признаках, имеющих максимальную связь с целевой переменной по хи-квадрат критерию. В других статьях [1, 5, 7] удаляют значения случайно в некотором подмножестве признаков. Существуют и более сложные стратегии создания пропущенных значения, например, пропущенное значение в одном признаке, если в другом значение выше некоторого порога. Однако невозможно подобрать метод, имитирующий все возможные сценарии пропущенных значений. По этой причине и по причине простоты в данной работе использовалось создание пропущенных значений случайным образом по следующей схеме:

1. По заданным долям выбираются подмножества объектов и признаков, в которых будут создаваться пропуски.

2. Для каждого значения, входящего в подмножества объектов и признаков, с некоторой вероятностью создаётся пропуск так, чтобы доля пропусков во всем наборе данных соотвествовала заданной.

В качестве наборов данных, не имеющих натуральные пропуски, использовались:

- KRKP (KingRook vs KingPawn chess game) – необходимо предсказывать победит ли команда белых в шахматном матче по описанию фигур на доске. 3196 объектов и 36 признаков, все из которых категориальные.

- Creditg (German Credit Data) – необходимо предсказать низкий или высокий кредитный риск у человека. 1000 объектов и 20 признаков, среди которых есть и категориальные, и количественные.

- Segment (Image Segmentation) – необходимо классифицировать изображения по признакам выского уровня. 2310 объектов и 19 признаков, все из которых количественные.

Наборы данных с натуральными пропусками:

- Horse (Horse Colic) – необходимо предсказать, необходимо ли лошади хирургическое вмешательство. 300 объектов и 22 признаки, среди которых есть и категориальные, и количественные. 30% пропущенных значений.

- Votes (Congressional Voting Records) – необходимо предсказать, к какой партии относился голосующий в конгрессе. 435 объектов и 16 признаков, среди которых все категориальные. 6% пропущенных значений.

- Cancer (Breast Cancer Wisconsin) – необходимо предсказать, доброкачественная или злокачественная опухоль у пациента. 699 объектов и 9 признаков, все из которых количественные. Менее 1% пропущенных значений.

Использовались следующие параметры каждого из методов восстановления (при возможности их задания):

1. Замена специальным значением: для применения случайного леса пропуск заменялся -1, для логистической регрессии и метода ближайшего соседа — 0.

2. Сингулярной разложение: ранг аппроксимирующей матрице в два раза меньше количества признаков, максимальное число итераций равно 10.

3. Метод k ближайших соседей: k = 5, метрика пространства L2.

4. Случайный лес: 10 деревьев, максимальное число итераций — 3.

5. Линейная регрессия: максимальное число итераций — 3.

6. EM-алгоритм: 1 смесь нормального распределения с полной матрицей ковариации.

7. Метод k средних: 8 кластеров, максимальное число итераций — 3.

8. ZET: число компетентных строк — 6, число компетентных столбцов — 4.

Для всех методов, за исключением заменой средним значением и модой, производилось округление до ближайшего значения в массиве.

Для оценки качества классификации использовалась точность, для оценки близости восстановленного набора данных к истинному (в случае с созданием искусственных пропусков) использовалось среднеквадратичное отклонение. Все данные были предварительно нормализованы.

3.2 Результаты эксперимента

3.2.1 Восстановление искусственных пропусков

Для измерения качества восстановления пропусков создавались наборы данных с долями пропусков от 2.5% до 15% с шагов в 2.5% в четверти наиболее важных признаков для обученного случайного леса. По 10-фолдовой стратифицированной кросс-валидации измерялись точность классификации и среднеквадратичное отклонение от истинных данных. Результаты усреднялись по десяти измерениям. Из сравнения был исключен метод, игнорирующий объекты с пропущенными значениями, по причине возможной работы только с малой долей пропусков в объектах.

Зависимость точности классификации от доли пропущенных значений изображена на графике [1], зависимость среднеквадратичного отклонения между восстановленными и истинными данными от доли пропущенных значений на графике [2].

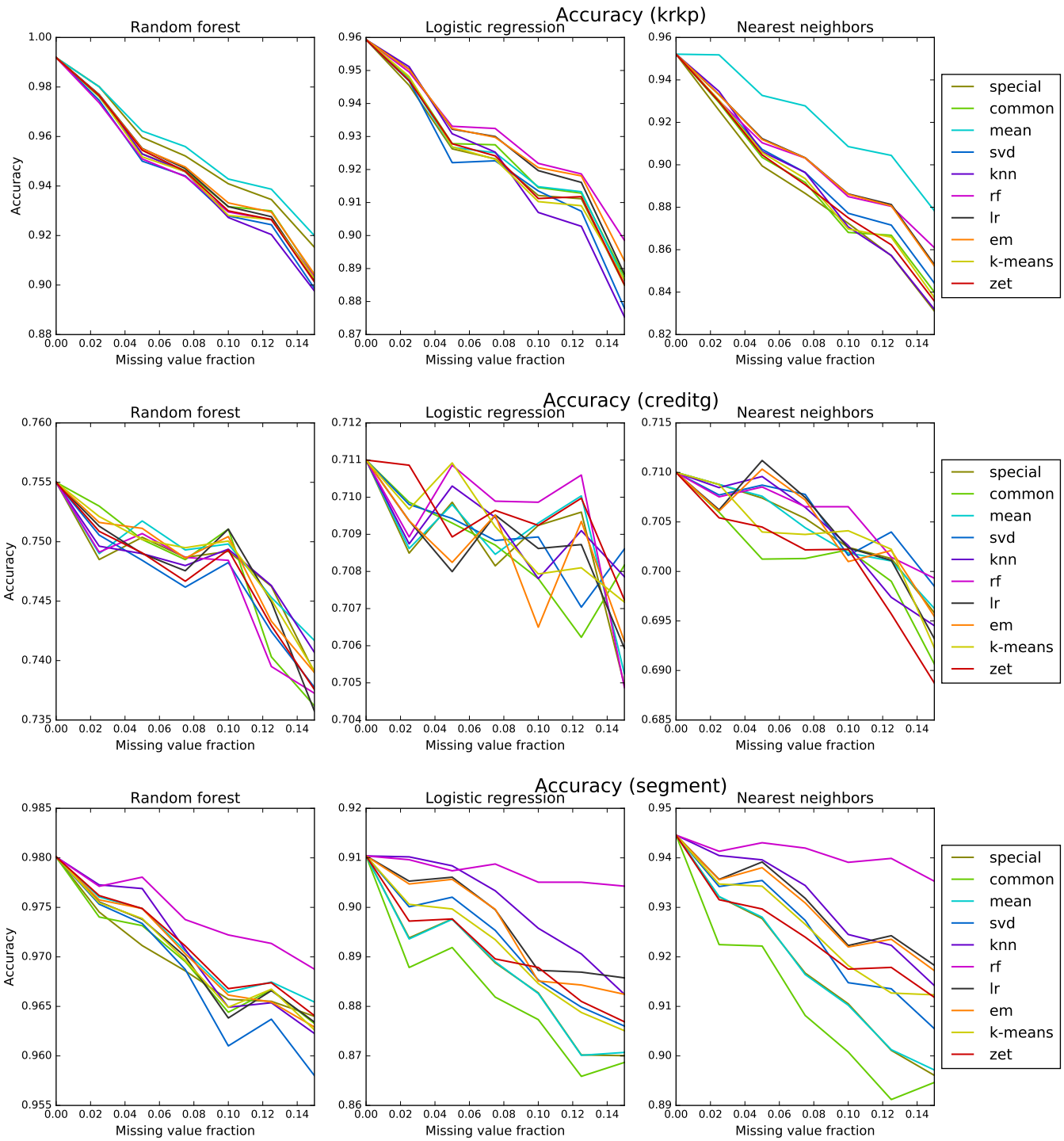


Рис. 1: Зависимость точность классификация восстановленных данных от доли пропущенных значений.

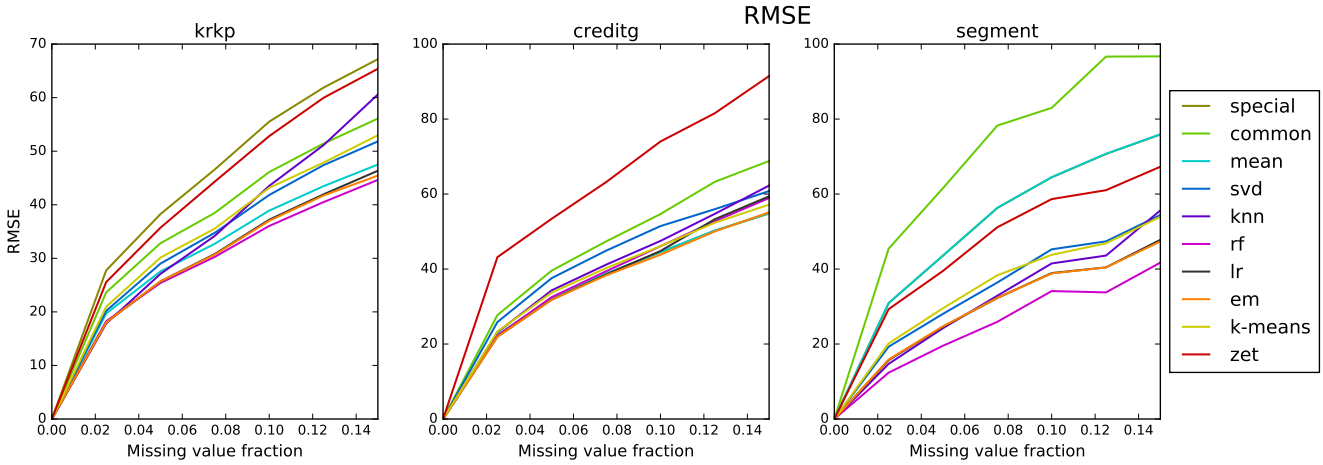


Рис. 2: Зависимость среднеквадратичного отклонения между восстановленными и истинными данными от доли пропущенных значений.

Для KRKP в двух из трех случаев лучше всего справился метод заполнения средним значением признака. Кроме того, хорошие результаты показали методы восстановление с помощью случайного леса и EM-алгоритма. В наборе данных Credit-g ни один из методов не опережает по качеству другие, поэтому можно сделать вывод, что в некоторых задачах не имеет разницы, каким методов восстанавливать пропущенные значения. Кроме того видна нестабильность в результате, что можно объяснить тем, что при восстановлении пропусков может вноситься лишняя информация в выборку. Для Segment лучший результат показало восстановление пропусков с помощью случайного леса. Хороший результат показывают методы, основанные на k ближайших соседей, EM-алгоритме и линейной регрессии. Низкий результат при замене модой признака объясняется тем, что все признаки были количественными.

В восстановлении пропусков для всех наборов данных хуже всего себя показывал метод восстановления, основанный на сингулярном разложении. Замена специальным значением показала не очень хороший результат при применении логистической регрессии и метода k ближайших соседей.

3.2.2 Восстановление натуральных пропусков

Для данных с натуральными пропусками возможно только сравнение конечных результатов работы алгоритмов классификации, поэтому результаты приведены в таблице [3]. Лучшие и близкие к ним результаты каждого столбца выделены жирным шрифтом.

Заметим, что нет метода, который бы всегда справлялся лучше остальных. Хорошие результаты показывает метод, основанный на замене самым частым значением и методе k средних. Для датасета Cancer близкие друг к другу результаты объясняются тем, что в нем малое количество пропусков.

Datasets	Horse			Votes			Cancer		
Methods	RF	LR	kNN	RF	LR	kNN	RF	LR	kNN
Ignore	-	-	-	0.9517	0.9527	0.9225	0.9591	**0.9694**	0.9694
Special	**0.8599**	0.8004	0.8400	0.9586	0.9584	0.9217	0.9557	0.9686	**0.9719**
Common	**0.8532**	0.8135	0.8270	**0.9633**	**0.9608**	0.9264	0.9542	0.9686	0.9685
Mean	0.8433	0.8004	0.8400	**0.9632**	0.9562	0.9401	0.9585	0.9686	**0.9714**
SVD	0.8201	0.8097	**0.8601**	0.9495	0.9540	0.9309	**0.9628**	0.9686	0.9700
kNN	0.8434	0.8166	0.8101	0.9517	0.9587	0.9240	**0.9628**	0.9686	0.9700
RF	0.8203	0.8065	0.8133	0.9490	0.9539	0.9240	0.9600	0.9686	0.9700
LR	0.8339	**0.8196**	0.8266	0.9565	0.9517	0.9332	**0.9628**	0.9686	0.9700
EM	0.8366	**0.8197**	0.8266	0.9518	0.9563	0.9357	**0.9628**	0.9686	0.9700
k-means	0.8464	0.8167	0.8432	0.9424	**0.9608**	**0.9423**	**0.9628**	0.9686	0.9700
ZET	0.8466	0.8097	0.8134	0.9516	**0.9630**	0.9218	0.9571	0.9686	0.9700

Рис. 3: Точность классификации восстановленного датасета с натуральными пропусками.

3.2.3 Добавление индикатор пропущенного значения

Дополнительно на данных с натуральными пропусками было проведено сравнение с подходом, в котором кроме замены пропущенного значения добавляется индикатор того, что в этом признаке было пропущенное значение. Для каждого из примененных методов машинного обучения результаты усреднились по наборам данных и

применяемым методам восстановления пропусков. Результаты можно увидеть на графике [4].

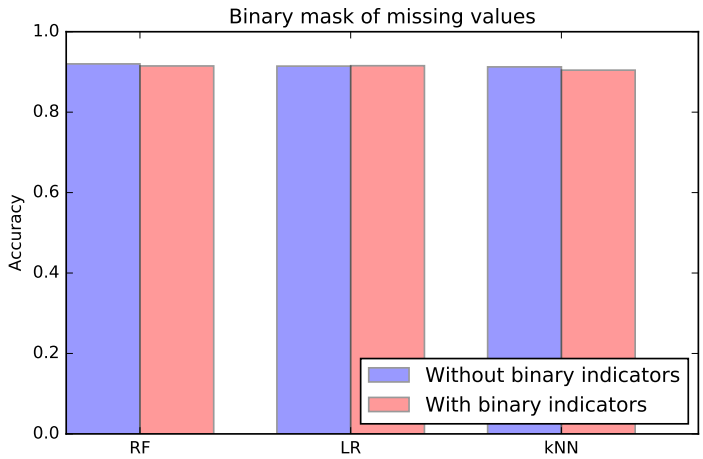


Рис. 4: Сравнение среднего качества классификации каждого из методов без и с добавлением индикатора пропущенного значения.

Качество классификации меняется незначительно. Только для логистической регрессии есть небольшой прирост, для двух других методов точность понижается.

4 Заключение

В работе были рассмотрены наиболее распространенные методы восстановления пропусков в данных, произведено их сравнение между собой. Эксперименты показали, что нет метода, который бы превосходил по качеству все остальные и мог бы быть универсальным подходом в любой задаче.

Выбор метода заполнения пропусков может зависеть от типов признаков, в которых существуют пропуски, от количества объектов, имеющих пропущенные значения, и от причины их возникновения. В каждой задаче необходим индивидуальный подбор метода обработки пропущенных значений.

Список литературы

[1] *Gupta A., Lam M.* The weight decay backpropagation for generalizations with missing values // *Annals of Operations Research.* — 1998. — Vol. 78. — Pp. 165–187.

[2] Imputing missing data for gene expression arrays / T. Hastie, R. Tibshirani†, G. Sherlock et al.

[3] *Lichman M.* UCI machine learning repository. — 2013. `http://archive.ics.uci.edu/ml`.

[4] *Luengo J., Garcia S., Herrera F.* On the choice of the best imputation methods for missing values considering three groups of classification methods // *Knowledge and Information Systems.* — 2012. — Vol. 32. — Pp. 77–108.

[5] Towards missing data imputation: A study of fuzzy k-means clustering method / D. Li, J. Deogun, W. Spaulding, B. Shuart.

[6] *Wohlrab L., Furnkranz J.* A comparison of strategies for handling missing values in rule learning. — 2009.

[7] *Zhou X.-Y., Lim J. S.* Em algorithm with gmm and naive bayesian to implement missing values // *Advanced Science and Technology Letters.* — 2014. — Vol. 46. — Pp. 1–5.

[8] *Загоруйко.* Прикладные методы анализа данных и знаний. — Институт математики, 1999.

[9] Реализация и эксперименты. `https://github.com/emilkayumov/missing-value`.