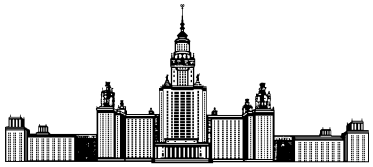
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

«Название курсовой работы»

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 и другие.

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

Если написать классификацию MCAR, MAR, NMAR, то это, наверное, должно быть здесь. Но тогда надо объяснять, какой метод на что рассчитан (что сомнительно в какой-то мере).

Возможно, стоит делить на группы: простейшие методы (первые три), основанные на предсказаниях (логистическая регрессия, случайный лес, ...), основанные на разложениях и подобных вещах (SVD, EM, K-Means, ...).

Здесь же про постановку задачи, обозначения.

2.1 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.

2.2 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.

стоит ли писать о том, что для деревьев логично работать с -1 в категориальном случае, а для линейного такая замена может исказить результат ?

2.3 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.

2.4 Замена с помощью SVD

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.

2.5 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.

2.6 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.

2.7 Замена с помощью линейной регрессии

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

2.8 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.

2.9 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.

2.10 Особенности реализации методов

1. При применении большинства методов можно получить дробное число, в таком случае такая 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 методах используется 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 рассмотрено отдельно.

3. ...

3 Эксперименты с данными

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

В качестве алгоритмов машинного обучения, на основе которых сравниваются методы восстановления пропусков, использовались случайные леса, логистическая регрессия и метод 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 репозитория UCI (ссылка), 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 (ссылка) предлагается создавать пропуски для случайных выбранных объектов в нескольких признаках, имеющих максимальную связь с целевой переменной по хи-квадрат критерию. В других статьях (ссылки) 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.

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

3.2 Результаты эксперимента

Графики и таблицы с результатами.

3.3 Обсуждение и выводы

В каких случаях что использовать. Что лучше работает, что не стоит использовать.

4 Заключение

Список литературы