DOI 10.37882/2223–2966.2021.05.13

ВОССТАНОВЛЕНИЕ ПРОПУЩЕННЫХ ЗНАЧЕНИЙ В ЗАДАЧАХ КЛАССИФИКАЦИИ ПРИ НЕСБАЛАНСИРОВАННОСТИ ДАННЫХ

RECOVERING MISSING VALUES IN CLASSIFICATION TASKS WITH DATA IMBALANCES

I. Kanishchev

Summary. Missing values is considered one of the biggest challenges faced by machine learning models. It can be exacerbated by the presence of imbalanced data. Several methods have been proposed and compared, such as pattern approximation, but they do not account for the adverse conditions found in real-world databases. This paper presents a comparison of the techniques used to classify a record from a real unbalanced database with a large amount of missing data, where the main goal is to preprocess the data for recovery and select completely filled records for further application of these methods.

Algorithms such as clustering, decision tree, artificial neural networks and Bayesian classifier were compared. The results can be used to ensure that describing the problem and understanding the database are essential steps for a correct comparison of methods in a real problem.

Keywords: missing value recovery, unbalanced data, classification.

Канищев Илья Сергеевич
Аспирант, Вятский государственный университет, Киров
kanishchev.ilya@gmail.com

Аннотация. Отсутствие данных считается одной из самых больших проблем, с которыми сталкиваются модели машинного обучения. Она может усугубиться при наличии несбалансированных данных. Было предложено и сопоставлено несколько методов, например, аппроксимация шаблонов, но они не учитывают неблагоприятные условия, обнаруженные в реальных базах данных. В данной работе представлено сравнение методик, используемых для классификации записи из реальной несбалансированной базы данных с большим количеством отсутствующих данных, где основной целью является предварительная обработка данных для восстановления и выбора полностью заполненных записей для дальнейшего применения этих методов.

Было проведено сравнение таких алгоритмов, как кластеризация, дерево решений, искусственные нейронные сети и байесовский классификатор. По результатам можно убедиться, что описание проблемы и понимание базы данных являются важными шагами для правильного сравнения методов в реальной проблеме.

Ключевые слова: восстановление пропущенных значений, несбалансированные данные, классификация.

Введение

Проблема с отсутствующими данными, возможно, является наиболее распространенной проблемой, с которой сталкиваются модели машинного обучения при анализе реальных данных. Во многих приложениях — от экспрессии генов в вычислительной биологии до ответов на опросы в социальных науках — в той или иной степени присутствуют недостающие данные. Поскольку многие статистические модели и алгоритмы машинного обучения полагаются на полные наборы данных, важно правильно обрабатывать недостающие данные.

В некоторых случаях для обработки недостающих данных может быть достаточно простых подходов. Например, анализ полного случая использует только полностью известные данные и пропускает все наблюдения с пропущенными значениями для проведения статистического анализа. Это хорошо работает, если только несколько наблюдений содержат пропущенные значения, и когда данные отсутствуют полностью случайным образом, анализ полного случая не приводит к смещенным результатам [1]. С другой стороны, некоторые алгоритмы машинного обучения естественным образом учитывают недостающие данные, и нет необходимости в предварительной обработке. Во многих других ситуациях пропущенные значения необходимо вменять до проведения статистического анализа полного набора данных [2].

Методы исследования

Для анализа данных использована выборка по кредитным договорам клиентов банка, содержащая 213 201 записей о кредитных заявках, включающая положительные решения и отказы в предоставлении кредита. Общее количество признаков данных — 71, 13 из которых содержат пропущенные значения (Табл. 1.). Данные представлены за период с января 2014 года по апрель 2020.

Пусть дан набор данных $X = \{x_1, \dots, x_n\}$ с пропущенными значениями

Таблица 1. Признаки с пропущенными значениями

Признак	Количество пропущенных значений
Должность	17 342
Категория должности	2 171
Категория клиента	58 781
Образование	6 303
Основание прож-я	5 507
Отрасль работы	16 551
Подтверждение дохода	68 819
Пол	229
Располагаемый доход	773
Семейное положение	6 324
Стаж работы	2 171
Тип залога	105 552
Тип имущества заемщика	88 656

Таблица 2. Основные методы

Название	Метод	Литература
Замена на среднее значение	Mean	[1]
Замена на наиболее часто встречающееся значение	Mode	[1]
Стохастическая аппроксимация	Stochastic Regression	[3]
Интерполяция	Interpolation	[4]
Байесовский линейная регрессия	Bayesian Binary Linear Regression	[5]

$x_{i,d}, (i, d) \in M$. Цель состоит в том, чтобы найти значения отсутствующих данных, который максимально правдоподобно напоминали исходные полные данные. Таким образом, когда проводится восстановление пропущенных данных, с использованием моделей машинного обучения, результаты должны быть аналогичны тем, которые были бы получены по полному набору данных. В Табл 2. Представлены основные методы по восстановлению пропущенных значений.

Среднее значение — для каждого пропущенного значения приписывается среднее значение всех известных значений в измерении.

Замена на наиболее часто встречающееся значение — для каждого пропущенного значения приписывается наиболее часто встречающееся значение известных значений в измерении.

Стохастическая аппроксимация — это метод поиска линейной регрессии с ограничениями путем перебора набора случайных коэффициентов, соответствующих определенным ограничениям.

Интерполяция — это метод нахождения неизвестных промежуточных значений некоторой функции по имеющемуся дискретному набору ее известных значений.

Байесовский линейная регрессия — это подход в логистической регрессии, в котором статистический анализ проводится в контексте байесовского вывода.

Рассмотрим задачу линейной регрессии, где существует линейная связь между x (объясняющая переменная) и y (зависимая переменная). Также ε, который описывает случайную составляющую в линейной связи между x и y.

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

y_i определяется двумя компонентами:

Неслучайный (структурный) компонент $\alpha + \beta x_i$, где x — объясняющая переменная в i-ом наблюдении, α и β — фиксированные параметры модели. α измеряет значение, в котором линия регрессии пересекает ось

Таблица 3.

№	Метод	Процент правильно восстановленных значений
1	Замена на среднее значение	66%
2	Замена на наиболее часто встречающееся значение	70%
3	Стохастическая аппроксимация	90%
4	Интерполяция	60%
5	Байесовский линейная регрессия	93%

$$\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)x_i=0\rightarrow\sum_{i=1}^{n}y_ix_i-\alpha x_i-\beta x_i^2=0\rightarrow\sum_{i=1}^{n}y_ix_i-(\bar{y}-\beta\bar{x})x_i$$
$$\beta x_i^2=0\rightarrow\sum_{i=1}^{n}y_ix_i-\bar{y}x_i+\beta\bar{x}x_i-\beta x_i^2=0\rightarrow\sum_{i=1}^{n}(y_i-\bar{y}+\beta\bar{x}-\beta x_i)x_i=0$$
$$\sum_{i=1}^{n}y_i-\bar{y}+\beta(\bar{x}-x_i)=0\rightarrow\sum_{i=1}^{n}(y_i-\bar{y})=-\beta\sum_{i=1}^{n}(\bar{x}-x_i)\rightarrow\beta=\frac{\sum_{i=1}^{n}(y_i-$$
$$\beta=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{Coc(x,y)}{Var(x)}=X'X^{-1}X'y$$

Рис. 1.

y. β измеряет крутизну линии регрессии. Случайная составляющая ε_i.

Необходимо минимизировать сумму всех квадратов отклонений от линии регрессии.

$$\sum_{i=1}^{n}(Y_i-\widehat{Y}_i)^2=\sum_{i=1}^{n}(\widehat{\varepsilon_i})^2=min$$

Чтобы получить оптимальные параметры α и β необходимо выбрать критерии выбора. В данном случае — находим α и β для случая, когда сумма всех квадратов отклонений минимальная.

$$\min_{\alpha,\beta}\sum_{i=1}^{n}(\varepsilon_i)^2\rightarrow\min_{\alpha,\beta}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2=$$
$$=S(\alpha,\beta)$$

Для минимизации функции необходимо вычислить условия первого порядка для α и β и установить их равными нулю.

$$\frac{dS(\alpha,\beta)}{d\alpha}=-2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)=0$$

$$\frac{dS(\alpha,\beta)}{d\beta}=-2\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)x_i=0$$

Для α:

$$\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)=0\rightarrow$$
$$\rightarrow\ \alpha=\sum_{i=1}^{n}(y_i-\beta x_i)\rightarrow\alpha=\bar{y}-\beta\bar{x}$$

Для β: см. рис. 1.

Для реализации байесовского вывода используется библиотека pymc3.

Результаты исследований, их обсуждение

Для валидации описанных выше методов создана тестовая выборка, состоящая из 27 035 объектов, каждый из которых характеризуется 133 признаками. На 10% этой выборки, случайным образом будут смоделированы различные типы пропусков. Результаты применения данных методов представлены в Табл. 3.

Таким образом можно предположить алгоритм восстановления пропущенных значений:

1. Из исходной выборки отбираются подмножество данных, не имеющих пропущенных значений.
2. На этом подмножестве моделируются различные типы пропусков.
3. Смоделированные пропущенные данные восстанавливаются с использованием описанного метода
4. Метод используется для восстановления реально пропущенных данных в исходной выборке.

Выводы

Проведенное исследование показало, что 3 и 5 метод позволяют с большей точностью восстановить пропущенное значение — свыше 90%. Таким образом, применение данных методов позволит обеспечить более точный результат работы моделей классификации.

ЛИТЕРАТУРА

1. Roderick JA Little and Donald B Rubin. Statistical Analysis with Missing Data. John Wiley & Sons, 1987
2. Алексеева А.В., Донцова Ю.С., Клячкин В.Н. Восстановление пропущенных наблюдений при классификации объектов. Теоретические и практические аспекты развития отечественного авиастроения. 2014.
3. Bharali, S.; Hazarika, J. Regression Models with Stochastic Regressors: An Expository note. 2018. URL: https://doi.org/10.20944/preprints201809.0539.v1).
4. S. Thirukumaran and A. Sumathi, "Missing value imputation techniques depth survey and an imputation Algorithm to improve the efficiency of imputation," Advanced Computing (ICoAC), 2012 Fourth International Conference on, Chennai, 2012
5. Mervat Mahdy, Dina Eltelbany. On Some Results of Bayesian Regression with Missing Data. 2012. The 47th Annual Conference on Statistics, Computer Science and operations Research ISRR, Cairo-Egypt, pp. 56–69.

Журнал «Современная наука: актуальные проблемы теории и практики»

Вятский государственный университет