УДК 004.93'12

ОБЗОР МЕТОДОВ ПРЕДОБРАБОТКИ, ИСПОЛЬЗУЕМЫХ ДЛЯ РЕШЕНИЯ ЗАДАЧ КЛАССИФИКАЦИИ В УСЛОВИЯХ НЕПОЛНОТЫ ДАННЫХ

К. А. Майков, д.т.н., профессор МГТУ им. Н.Э. Баумана; maikov@bmstu.ru
П. А. Гаврилов, магистрант МГТУ им. Н.Э. Баумана; gavrilov.pavel.a@yandex.ru

Рассматривается задача классификации в условиях неполноты данных. ***Целью работы*** *является исследование функциональных возможностей и ограничений ряда известных методов предобработки, используемых для заполнения пропусков. Описаны типы пропусков данных. Приведено описание групп методов, используемых для решения рассматриваемой задачи. Рассмотрены статистические методы заполнения пропусков: средним арифметическим, медианой, модой; метод горячей колоды. Представлены результаты сравнительного анализа ряда методов заполнения отсутствующих данных с использованием алгоритма k-ближайших соседей в качестве классификатора. Качество классификации оценивается с помощью метода 10-кратного скользящего контроля. Обоснован выбор программного обеспечения для проведения численных экспериментов. Результаты проведённых экспериментов показывают, что при отсутствии 5 – 20 % значений признака анализируемые методы обеспечивают схожие результаты, а при отсутствии 30 – 40 % значений метод заполнения горячей колоды показывает более низкие оценки скользящего контроля, чем методы заполнения средним арифметическим, медианой и модой. В то же время при отсутствии 40 % значений метод заполнения медианой превосходит другие рассматриваемые методы.*

Ключевые слова: *машинное обучение, классификация, отсутствующие данные, пропущенные данные, предобработка данных.*

Введение

Актуальность работы обусловлена тем, что большой класс практических и учебных задач классификации, как правило, характеризуется неполнотой исходных данных, например [1, 2]:

– в результате социального анализа часть данных отсутствует или является недостоверной [1, 3-4];

– при решении задач медицинской диагностики результаты некоторых анализов могут отсутствовать [1, 5];

– формируя систему оценки кредитоспособности заёмщика, необходимо учитывать, что часть информации о нём может отсутствовать [1, 6].

Под неполным набором данных будем понимать такой набор, в котором хотя бы одно значение признака отсутствует, а под полным – такой набор, в котором все значения присутствуют.

Задача классификации в условиях неполноты данных включает в себя две подзадачи: обработку исходного набора данных и собственно классификацию. По способу решения данных подзадач выделяют четыре группы методов [1].

1. Из исходного неполного набора данных формируется модифицированный набор путём удаления объектов, у которых отсутствует значение хотя бы одного признака. Затем на основе данных полученного набора производится классификация.
2. В исходной выборке отсутствующие значения признаков заполняются значениями, полученными на основе имеющихся данных. Полученная полная выборка используется для осуществления классификации.
3. На основе исходного неполного набора данных моделируется функция плотности вероятностей, которая затем подаётся на вход соответствующего алгоритма классификации.
4. Исходный неполный набор данных подаётся на вход классификатора, учитывающего возможное отсутствие значений признаков, при этом отсутствие рассматривается как одно из возможных значений.

В данной работе рассматривается вторая группа, что позволяет использовать методы классификации, предназначенные для обработки полных наборов данных и практически апробированные на большом количестве задач [1]. При использовании данного подхода подзадачи (обработка исходных данных и классификация) решаются раздельно. Во время решения первой подзадачи выполняется предобработка данных (заполняются пропуски), а затем вторая подзадача (построение классификатора) решается с использованием полученного полного набора данных. Если у классифицируемого объекта отсутствует значение признака, объект также проходит этап предобработки, после чего классификатор определяет его принадлежность к некоторому классу. На рисунке 1 представлены описанные выше группы и некоторые их методы.

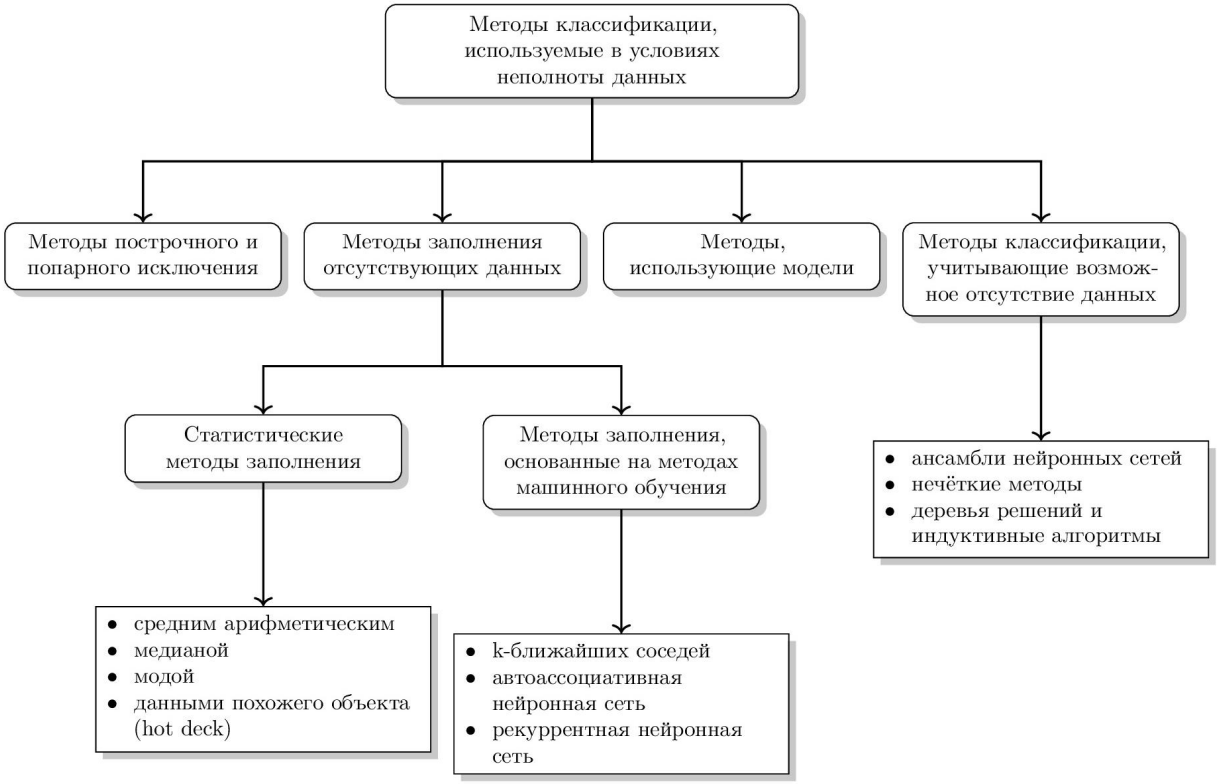

Рисунок 1 — Методы классификации, используемые в условиях неполноты данных

Типы пропусков данных. Способ обработки отсутствующих данных следует выбирать в соответствии с причинами отсутствия значений признаков [1]. Выделяют следующие типы пропусков [1].

1. Полностью случайные (missing completely at random; MCAR). Вероятность появления таких пропусков не зависит ни от значений самого измеряемого признака, ни от значений других признаков. Примером возникновения пропусков данного типа является потеря образца крови пациента, следствием которой является невозможность измерения ряда значений признаков.
2. Частично случайные (missing at random; MAR). Вероятность появления таких пропусков не зависит от значений непосредственно измеряемого признака, но обусловлена значениями других признаков. В качестве примера возникновения можно привести противопоказание некоторых анализов пациентам с поставленными диагнозами, при которых проведение данных анализов может нанести вред организму.
3. Неслучайные (not missing at random; NMAR). Вероятность появления неслучайных пропусков в данных зависит от значений самого признака. Если значение измеряемого признака не попадает в диапазон чувствительности измерительного прибора, то такие пропуски принято называть неслучайными.

При первых двух типах (MCAR, MAR) во время анализа данных причину пропусков можно проигнорировать [1], используя более простые методы обработки отсутствующих данных. В данной работе рассматриваются задачи классификации, в исходных данных которых все

пропуски относятся к первым двум типам (MCAR, MAR).

Методы заполнения, основанные на статистическом анализе. Одним из наиболее простых для реализации методов является метод заполнения средним арифметическим [1, 7]; отсутствующее значение признака j у объекта x заполняется средним арифметическим известных значений данного признака у других объектов

$$\widetilde{x}^{\,j} = \frac{1}{N_{obs}} \sum_{i=1}^{N_{obs}} x_i^j,$$

где N_{obs} – количество известных значений искомого признака, j – порядковый номер признака, а суммирование ведётся по объектам, содержащим значение искомого признака.

Следующим рассматриваемым методом является метод заполнения медианой [8], в соответствии с которым отсутствующее значение заменяется на следующее значение признака:

$$M^j = \begin{cases} x_i^j,\ i = \dfrac{n+1}{2},\ n = 2k, \\ \dfrac{x_i^j + x_{i+1}^j}{2}, i = \dfrac{n}{2}, n = 2k+1, \end{cases}$$

где j определено выше, n – количество объектов в наборе данных, а значения x^j отсортированы по возрастанию или убыванию.

Также рассматривается метод заполнения модой [8], при использовании которого отсутствующее значение признака заменяется на наиболее часто встречающееся значение данного признака.

Более сложным методом заполнения отсутствующих данных является метод горячей колоды (hot deck) [1, 8]. В соответствии с данным методом отсутствующее значение признака объекта x_1 заполняется соответствующим значением признака полного объекта x_2 при условии, что расстояние между указанными объектами является минимальным. Мера расстояния выбирается исходя из условий решаемой задачи [8].

Основной недостаток описанных выше методов заполнения отсутствующих значений заключается в игнорировании корреляций между значениями различных признаков, что в ряде случаев приводит к потерям в качестве классификации [1].

Метод классификации k-ближайших соседей. Благодаря хорошей интерпретируемости результатов и простоте реализации, при проведении экспериментов для решения подзадачи классификации был выбран метод классификации k-ближайших соседей [7]. Поскольку в решаемых тестовых задачах все признаки – количественные, целесообразно использовать метрику Минковского [7]

$$p(x, x_i) = \sqrt{\sum \left(x^j - x_2^j\right)^2},$$

где $x = \left(x^1, \ldots, x^n\right)$ – вектор объекта x, $x_i = \left(x_i^1, \ldots, x_i^n\right)$ – вектор объекта x_i. Экспериментальным путём количество ближайших соседей (параметр k), используемых для определения класса объекта, было выбрано равным трём.

Оценка качества методов. Как показано в [7], для оценки качества классификации алгоритма a на выборке X^l целесообразно использовать следующий функционал:

$$Q\left(a, X^l\right) = \frac{1}{l} \sum_{i=1}^{l} L(a, x_i),$$

где l – количество объектов тестовой выборки, $L(a,x) = \left[a(x) \neq \dot{y}(x)\right]$ – функция потерь, $\dot{y}$ – неизвестная целевая зависимость, a – обученный алгоритм.

Для более точной оценки (усреднённой на k различных парах тренировочных и тестовых наборов) используется оценка k-кратного скользящего контроля [7]. Для её нахождения исходная выборка разбивается на k непересекающихся частей, затем поочерёдно обучение происходит на $k-1$ частях, а на оставшейся оценивается качество с помощью описанного выше функционала. После этого вычисляется среднее арифметическое k полученных оценок. Согласно [9] параметр k выбран равным 10.

Описание среды проведения экспериментов. Численные эксперименты производились на компьютере Dell Latitude E7440 с процессором 2.1 ГГц Intel Core i7, 8 Гб оперативной памяти, операционной системой Arch Linux 64bit.

Благодаря выразительности и наличию протестированных библиотек, реализующих методы машинного обучения, для программирования вычислительных экспериментов был выбран язык Python 2.7.9 [10]. Библиотека scikit-learn 0.16.1 [11], включающая в себя набор инструментов для решения задач машинного обучения, была выбрана из-за наличия необходимой функциональности и эргономичного в использовании интерфейса.

Результаты экспериментов

В данном разделе исследуется зависимость качества классификации от использования одного из следующих методов заполнения: средним арифметическим, медианой, модой; метод горячей колоды (hot deck).

Оценка качества обученных алгоритмов классификации производилась описанным выше методом скользящего контроля. В качестве метода классификации используется описанный выше метод k-ближайших соседей.

Для сравнения функциональных особенностей исследуемых методов замены используются следующие наборы данных:

– тестовый набор;

– содержащий отсутствующие данные набор Pima Indians Diabetes Dataset из репозитория UCI [12].

Тестовый набор данных состоит из 100 объектов, имеющих по два признака. Каждый из объектов набора принадлежит одному из двух классов. Согласно [1] из исходного полного набора были получены пять новых наборов путём удаления 5, 10, 20, 30 и 40 % значений первого признака. Исходный набор данных изображен на рисунке 2.

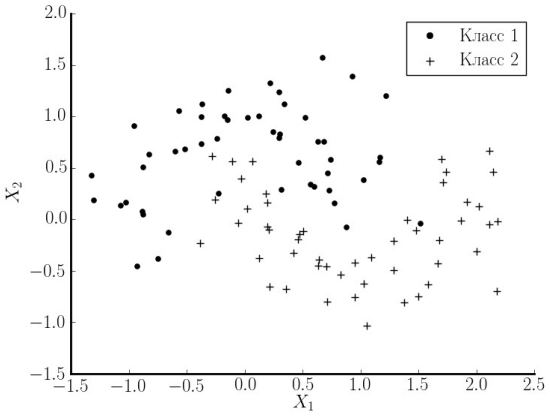

Рисунок 2 — Полный набор данных задачи бинарной классификации объектов с двумя признаками

На рисунке 3 изображен набор данных после удаления значений первого признака у 30 % объектов. Объекты, у которых отсутствует значение первого признака, изображены на рисунке горизонтальными отрезками.

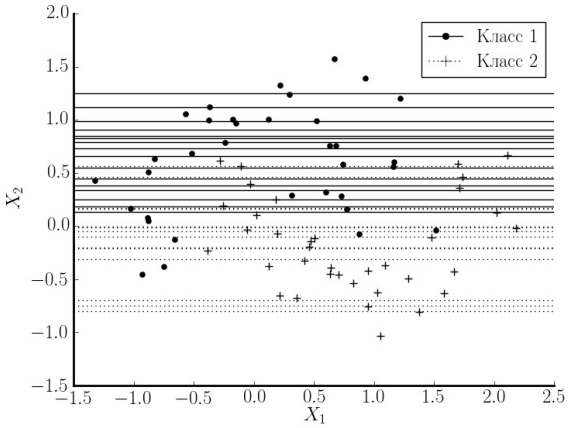

Рисунок 3 — Набор данных, в котором у 30 % объектов отсутствует значение первого признака. Горизонтальные линии обозначают объекты, у которых отсутствует значение первого признака

В таблице 1 представлены результаты классификации методом k-ближайших соседей после заполнения пропущенных значений одним из описанных выше методов. Из полученных результатов видно, что:

– при отсутствии 5 – 20 % значений первого признака анализируемые методы обеспечивают схожие результаты;

– при отсутствии 30 – 40 % значений метод заполнения горячей колоды показывает более низкие оценки скользящего контроля, чем методы заполнения средним арифметическим, медианой и модой;

– при отсутствии 40 % значений метод заполнения медианой превосходит другие рассматриваемые методы.

Таблица 1 — Оценки качества решения тестовой задачи классификации с использованием различных методов заполнения

Отсутств. значения %	Метод заполнения			
	средним	медианой	модой	горячей колоды
5	0.95	0.95	0.95	0.94
10	0.91	0.91	0.91	0.91
20	0.91	0.91	0.90	0.91
30	0.89	0.89	0.88	0.83
40	0.85	0.86	0.85	0.83

Тестирование методов заполнения выполнено на наборе данных Pima Indians Diabetes Dataset из 768 объектов, имеющих 8 признаков и метку класса, показывающую наличие или отсутствие заболевания сахарным диабетом второго типа. В таблице 2 представлена информация об отсутствии значений признаков в Pima Indians Diabetes Dataset.

Таблица 2 — Процент отсутствующих значений признаков в наборе данных Pima Indians Diabetes

Признак	1	2	3	4	5	6	7	8
% отсут. значений	0	1	5	30	49	1	0	0

Полученные в результате проведённых экспериментов оценки скользящего контроля, которые представлены в таблице 3, показывают, что метод заполнения горячей колоды превосходит в качестве классификации методы заполнения средним арифметическим, медианой и модой.

Таблица 3 — Оценки качества решения задачи классификации Pima Indians Diabetes с использованием различных методов заполнения

Метод заполнения			
средним	медианой	модой	горячей колоды
0.71	0.70	0.71	0.73

Низкие значения оценок качества классификации в таблице 3 обусловлены тем, что используемые методы предобработки игнорируют взаимосвязи между значениями признаков. Таким образом, представляется целесообразным построение комбинированного метода, объединяющего функциональные возможности исследованных методов с возможностями методов, учитывающих наличие корреляций между значениями признаков.

Заключение

Результаты выполненных численных экспериментов показывают, что при отсутствии 5 – 20 % значений признака анализируемые методы обеспечивают схожие результаты, а при отсутствии 30 – 40 % значений метод заполнения горячей колоды показывает более низкие оценки скользящего контроля, чем методы заполнения средним арифметическим, медианой и модой. В то же время при отсутствии 40 % значений метод заполнения медианой превосходит другие рассматриваемые методы.

Анализ функциональных ограничений рассматриваемых методов показал целесообразность построения метода, объединяющего в себе функциональные возможности рассмотренных методов и возможности методов, учитывающих наличие корреляций между значениями признаков.

Библиографический список

1. **Garcia-Laencina P. J., Sancho-Gomez J. L., Figueiras-Vidal A. R.** Pattern classification with missing data: a review // Neural Computing and Applications. 2010. Vol. 19. No. 2. Pp. 263-282.

2. **Демидова Л. А., Кираковский В. В., Пылькин А. Н.** Принятие решений в условиях неопределённости. М.: Горячая линия – Телеком, 2012, 288 с.

3. **Загниева И. К.** Решение проблемы неполноты данных массовых опросов // Российская социология завтрашнего дня. 2008. Вып. 3. С. 84-95.

4. **Rubin D. B.** Multiple Imputation for Nonresponse in Surveys. New York: Wiley, 1987, 258 p.

5. **Liu P., Elia E.-D., Lei L., Vasilakis C., Chountas P., Huang W.** An Analysis of Missing Data Treatment Methods and Their Application to Health Care Dataset // Proceedings of the First International Conference on Advanced Data Mining and Applications. Berlin: Springer-Verlag, 2005. Pp. 583-590.

6. **Kofman P., Sharpe I. G.** Using Multiple Imputation in the Analysis of Incomplete Observations in Finance // Journal of Financial Econometrics. 2003. Vol. 1. No. 2. Pp. 216-249.

7. **Воронцов К. В.** Математические методы обучения по прецедентам (теория обучения машин) // URL: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения: 19.02.2016).

8. **Загниева И. К., Тимонина Е. С.** Сравнение эффективности алгоритмов заполнения пропусков в данных в зависимости от используемого метода анализа // Мониторинг общественного мнения. 2014. № 1 (119). С. 41-55.

9. Introduction to Statistical Learning: With Applications in R / Ed. by **G. Casella, S. Fienberg, I. Olkin**. New York: Springer-Verlag, 2014. 426 p.

10. **Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.** Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. Pp. 2825-2830.

11. Язык программирования python // URL: http://python.org (дата обращения: 18.02.2016).

12. Pima Indians Diabetes Data Set // URL: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes (дата обращения: 18.02.2016).

UDC 004.93'12

PREPROCESSING METHODS FOR CLASSIFICATION WITH MISSING DATA: REVIEW

K. A. Maykov, PhD (technical sciences), full processor, BMSTU, Moscow; maikov@bmstu.ru
P. A. Gavrilov, master student, BMSTU, Moscow; gavrilov.pavel.a@yandex.ru

Classification task with missing data is considered. The aim of this work is to investigate features and limitations of a number of known preprocessing methods used for handling missing values. Missing data mechanisms have been described. Different approaches in pattern classification with missing data have been shown. Statistical imputation methods (mean, median, mode and hot deck imputation) have been considered. The results of the comparative analysis of a number of imputation methods have been presented using k-nearest neighbor algorithm as a classifier. The quality of the classifier is evaluated by 10-fold cross-validation. The choice of software for numerical experiments has been justified. As it has been shown in the obtained results, all the above methods provide similar results for 5 – 20% missing data percentages, and hot deck imputation provides lower cross-validation scores than mean, median and mode imputation methods for 30 – 40% missing data percentages. In the same time median imputation exceeds other reviewed methods for 40% missing data percentage.

Key words: *machine learning, classification, missing data, data preprocessing.*

References

1. **Garcia-Laencina P. J., Sancho-Gomez J. L., Figueiras-Vidal A. R.** Pattern classification with missing data: a review. *Neural Computing and Applications.* 2010, vol. 19, no. 2, pp. 263-282.

2. **Demidova L. A., Kirakovskij V. V., Pyl'kin A. N.** *Prinjatie reshenij v uslovijah neopredeljonnosti* (Decision Making under Uncertainty). Moscow, Gorjachaja linija – Telekom, 2012, 288 p. (in Russian).

3. **Zagnieva I. K.** Reshenie problemy nepolnoty dannyh massovyh oprosov. *Rossijskaja sociologija zavtrashnego dnja.* 2008, issue 3, pp. 84-95 (in Russian).

4. **Rubin D. B.** Multiple Imputation for Nonresponse in Surveys. New York, Wiley, 1987, 258 p.

5. **Liu P., Elia E.-D., Lei L., Vasilakis C., Chountas P., Huang W.** An Analysis of Missing Data Treatment Methods and Their Application to Health Care Dataset. Proceedings of the First International Conference on Advanced Data Mining and Applications. Berlin, Springer-Verlag, 2005, pp. 583-590.

6. **Kofman P., Sharpe I. G.** Using Multiple Imputation in the Analysis of Incomplete Observations in Finance. *Journal of Financial Econometrics.* 2003, vol. 1, no. 2, pp. 216-249.

7. **Voroncov K. V.** Matematicheskie metody obuchenija po precedentam (teorija obuchenija mashin), URL: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (accessed: 19.02.2016).

8. **Zagnieva I. K., Timonina E. S.** Sravnenie effektivnosti algoritmov zapolnenija propuskov v dannyh v zavisimosti ot ispol'zuemogo metoda analiza. *Monitoring obshhestvennogo mnenija.* 2014, no 1 (119), pp. 41-55 (in Russian).

9. Introduction to Statistical Learning: With Applications in R, ed. by **G. Casella, S. Fienberg, I. Olkin**. New York, Springer-Verlag, 2014, 426 p.

10. **Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.** Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011, vol. 12, pp. 2825-2830.

11. Python programming language, URL: http://python.org (accessed: 18.02.2016).

12. Pima Indians Diabetes Data Set, URL: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes (accessed: 18.02.2016).