# Восстановление данных на основе ассоциативных правил

А. Д. Богданов*, А. П. Рыжов‡

Предложен интерпретируемый метод импутации категориальных данных на основе ассоциативных правил. Эксперименты на Car Evaluation, Nursery и Mushrooms показали его конкурентоспособность с MissForest и превосходство над импутацией модой.

**Ключевые слова:** качество данных, пропущенные значения, импутация, ассоциативные правила, категориальные признаки, Apriori.

## 1. Введение

Проблема качества данных является определяющим фактором успешности прикладных систем машинного обучения. Ежегодные мировые издержки от использования некачественных данных исчисляются триллионами долларов [1], при этом ценность представляют исключительно очищенные и верифицированные выборки [2]. Критичность этой проблемы подтверждается высокой уязвимостью алгоритмов к зашумлению: искажение или «отравление» всего лишь 0,5% обучающих данных способно более чем вдвое снизить качество работы моделей [3–5].

В то же время в реальных прикладных задачах качество данных радикально отличается от нормативных требований. Так, в медицинских информационных системах доля пропусков в сырых выборках может достигать 58% [6], а пренебрежение этапом предобработки приводит к провалу подавляющего большинства индустриальных ИИ-проектов [7]. Среди дефектов данных пропуски являются наиболее массовыми. Настоящая работа посвящена исследованию проблемы импутации пропущенных категориальных значений в табличных наборах данных на основе интерпретируемого аппарата ассоциативных правил.

В докладе рассмотрены 5 основных подходов к обработке пропущенных значений [8]: **удаление данных** (попарное и списочное) с риском

* *Богданов Арсений Денисович* — студент, Московский государственный университет имени М. В. Ломоносова, e-mail: bogdanov.arseny.2005@gmail.com.

*Bogdanov Arseny Denisovich* — Student, Lomonosov Moscow State University.

‡ *Рыжов Александр Павлович* — д. т. н., профессор, Московский государственный университет имени М. В. Ломоносова, e-mail: ryjov@cs.msu.ru.

*Ryzhov Alexander Pavlovich* — D.Sc. in Engineering, Professor, Lomonosov Moscow State University.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

смещения и потери информации; **заполнение статистическими оценками** (мода, среднее), часто игнорирующее межпризнаковые корреляции; **методы машинного обучения** со вспомогательной моделью, чувствительные к гиперпараметрам и выбросам; **нейронные сети**, требующие больших объемов обучающих данных; **обработка на основе обучения** с извлечением латентных представлений, характеризующаяся высокой сложностью и низкой прозрачностью.

Большинство подходов обладают высокой вычислительной сложностью или низкой интерпретируемостью. Особый интерес представляют данные с категориальными признаками, требующими предварительной кодировки. Целью данной работы является разработка и сравнение с имеющимися методами (импутация модой и MissForest [9]) алгоритма по интерпретируемому заполнению пропущенных категориальных значений на основе ассоциативных правил вида «Если ..., то ...».

## 2. Алгоритм импутации пропущенных значений

Предлагаемый алгоритм включает следующие этапы: 1) перевод строк таблицы в транзакции приписыванием префикса-названия колонки к значению для исключения коллизий одинаковых значений в разных колонках; 2) поиск ассоциативных правил с единичным термом в правой части алгоритмом Apriori: $\text{supp} = 0.01$ при малом числе признаков, иначе $\text{supp} = 0.1$; $\text{conf} = 0.5$ (данные пороги подобраны эвристически); 3) для каждого пропуска сопоставление заполненных признаков строки с левыми частями правил и импутация правой частью наиболее подходящего правила; 4) при отсутствии подходящего правила — импутация модой как частным случаем правила длины 1 с наибольшей поддержкой; 5) обратное преобразование данных в табличный вид с удалением служебных префиксов.

## 3. Экспериментальное исследование

Эксперименты проводились на наборах данных Car Evaluation, Nursery и Mushrooms с искусственно внесенными полностью случайными (MCAR) пропусками в объеме 15%, 20% и 30% от всех значений. Сравнивались предложенный алгоритм, MissForest и импутация модой. Оценка качества выполнена по метрикам Accuracy weighted и Precision weighted (табл. 1).

## 4. Результаты и выводы

При 15–20% пропусков на данных Car Evaluation и Nursery предложенный алгоритм превосходит MissForest по Accuracy, а на Car Evaluation —

Таблица 1: Сравнение методов по метрикам Accuracy и Precision (weighted)

| | | **Accuracy** | | | **Precision** | | |
|---|---|---|---|---|---|---|---|
| Датасет | Пропуск | Acc. прав. | MissForest | Мода | Acc. прав. | MissForest | Мода |
| Car Evaluation | 15% | 0.3672 | 0.3222 | 0.3211 | 0.3575 | 0.3251 | 0.1703 |
| | 20% | 0.3608 | 0.3369 | 0.3292 | 0.4614 | 0.3299 | 0.1633 |
| | 30% | 0.3360 | 0.3565 | 0.3360 | 0.2107 | 0.3487 | 0.2107 |
| Mushrooms | 15% | 0.6623 | 0.7718 | 0.5755 | 0.7060 | 0.7707 | 0.5322 |
| | 20% | 0.6694 | 0.7854 | 0.5798 | 0.6888 | 0.7839 | 0.5356 |
| | 30% | 0.6288 | 0.7932 | 0.5749 | 0.6495 | 0.7920 | 0.5329 |
| Nursery | 15% | 0.4207 | 0.3468 | 0.3105 | 0.3266 | 0.3467 | 0.1283 |
| | 20% | 0.3999 | 0.3637 | 0.3107 | 0.3092 | 0.3678 | 0.1454 |
| | 30% | 0.3635 | 0.3803 | 0.3136 | 0.2331 | 0.3933 | 0.1150 |

также по Precision, подтверждая способность правил улавливать межпризнаковые связи. Экспериментально метод показывает нижнюю границу точности на уровне моды (например, при 30% на Car Evaluation). Восстановление логически обосновано импликациями с метриками supp и conf, формируя базу знаний о структуре данных. Минимальный порог $\text{supp}_{\min}$ отсекает шумовые транзакции, обеспечивая устойчивость к выбросам, при этом алгоритм работает напрямую с категориальными значениями без предварительного кодирования.

## Список литературы


[1] Redman T. C., *Bad Data Costs the U.S. $3 Trillion Per Year*, Harvard Business Review, 2016.

[2] Chahal H., Fedasiuk R., Flynn C., *Messier than Oil: Assessing Data Advantage in Military AI*, CSET, 2020.

[3] Rando J., Tramèr F., *Universal Jailbreak Backdoors from Poisoned Human Feedback*, 2024, arXiv: 2311.14455.

[4] Wan A., Wallace E., Shen S., Klein D., *Poisoning Language Models During Instruction Tuning*, 2023, arXiv: 2305.00944.

[5] Souly A., Rando J., Chapman E. et al., *Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples*, 2025, arXiv: 2510.07192.

[6] Горный Б. Э., Рыжов А. П., Строгалов А. С. и др., "Оценка риска неблагоприятного клинического исхода методами углубленного анализа данных", *Интеллектуальные системы. Теория и приложения*, **25**:2 (2021), 23–45.

[7] Nanda MIT, *The GenAI Divide: State of AI in Business*, 2025 `https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf`.

[8] Zhou Y., Aryal S., Bouadjenek M. R., *A Comprehensive Review of Handling Missing Data*, 2024, arXiv: 2404.04905.

[9] Stekhoven D. J., Bühlmann P., "MissForest—non-parametric missing value imputation for mixed-type data", *Bioinformatics*, **28**:1 (2012), 112–118, arXiv: 1105.0828.

Статья поступила 26 сентября 2026 г.

**Data Imputation Based on Association Rules**

A. D. Bogdanov, A. P. Ryzhov

An interpretable method for categorical data imputation based on association rules is proposed. Experiments on Car Evaluation, Nursery, and Mushrooms benchmarks demonstrate its competitiveness with MissForest and superiority over mode imputation.

**Keywords:** data quality, missing values, data imputation, association rules, categorical features, Apriori.

# References

[1] Redman T. C., *Bad Data Costs the U.S. $3 Trillion Per Year*, Harvard Business Review, 2016.

[2] Chahal H., Fedasiuk R., Flynn C., *Messier than Oil: Assessing Data Advantage in Military AI*, CSET, 2020.

[3] Rando J., Tramèr F., *Universal Jailbreak Backdoors from Poisoned Human Feedback*, 2024, arXiv: 2311.14455.

[4] Wan A., Wallace E., Shen S., Klein D., *Poisoning Language Models During Instruction Tuning*, 2023, arXiv: 2305.00944.

[5] Souly A., Rando J., Chapman E. et al., *Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples*, 2025, arXiv: 2510.07192.

[6] Gornyi B. E., Ryjov A. P., Strogalov A. S. et al., "Assessment of Adverse Clinical Outcome Risk via In-Depth Data Analysis Methods", *Intelligent Systems. Theory and Applications*, **25**:2 (2021), 23–45 (In Russian).

[7] Nanda MIT, *The GenAI Divide: State of AI in Business*, 2025 `https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf`.

[8] Zhou Y., Aryal S., Bouadjenek M. R., *A Comprehensive Review of Handling Missing Data*, 2024, arXiv: 2404.04905.

[9] Stekhoven D. J., Bühlmann P., "MissForest—non-parametric missing value imputation for mixed-type data", *Bioinformatics*, **28**:1 (2012), 112–118, arXiv: 1105.0828.

Received on September 26, 2026