АНАЛИЗ КАЧЕСТВА ОЦЕНКИ РЕЗУЛЬТАТОВ СОЦИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ ПО НЕПОЛНЫМ ДАННЫМ

Ючков А. К., Хаджинова К. А., Навроцкий А. А.
Кафедра информационных технологий автоматизированных систем,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: 0-kael007-0@proton.me, xju2005@gmail.com, navrotsky@bsuir.by

Рассматривается проблема неполных данных в социологических исследованиях, предлагаются нейросетевые подходы для их анализа. Исследуются архитектуры нейросетей, такие как многослойные перцептроны и рекуррентные сети, которые могут улучшить качество социологических выводов.

Постановка задачи

Социологические исследования помогают понять общественные процессы, но часто собранные данные бывают неполными, что приводит к искажению результатов [1].

Неполные данные возникают по различным причинам: респонденты могут не ответить на вопросы, данные могут быть потеряны или некорректно записаны. Это создает сложности в анализе, так как традиционные статистические методы не всегда эффективны. Нейросети, благодаря своей гибкости и способности обучаться на больших объемах информации, могут предложить более гибкие решения.

Среди различных типов нейросетей наиболее популярными являются многослойные перцептроны *(MLP)* и рекуррентные нейронные сети *(RNN)*. *MLP* хорошо подходят для обработки структурированных данных, тогда как *RNN* эффективны для последовательных данных, что может быть полезно при анализе временных рядов в социологии.

I. Методика исследования

Для анализа неполных данных в социологических исследованиях была разработана методология, включающая несколько ключевых этапов [2], позволяющих оценить эффективность нейросетевых моделей в обработке как полных, так и неполных данных.

Исходной базой данных для исследования послужил набор данных, собранный в рамках социологического опроса [3], изначально не содержащий пропусков в ответах респондентов и включающий разнообразные переменные, такие как демографические характеристики, мнения по социальным вопросам и поведенческие данные. Для моделирования условий реальных социологических исследований были искусственно созданы пропуски в данных, выбраны следующие методы удаления: *случайное* (пропуски были введены случайным образом в определенный процент ответов; что позволяло имитировать различные уровни неполноты данных, с которыми могут столкнуться исследователи на практике) и *систематическое* (в некоторых случаях пропуски создавались целенаправленно для определенных переменных для оценки влияния).

Для сравнения эффективности обработки полных и неполных данных выбраны две модели нейросетей:

1. Многослойный перцептрон: модель выбрана за ее способность обрабатывать структурированные данные и выявлять сложные зависимости между переменными [4].
2. Рекуррентная нейронная сеть: модель выбрана для анализа последовательных данных, что особенно важно для временных рядов и динамических изменений во мнениях респондентов [5].

Перед обучением моделей данные прошли несколько этапов подготовки. Все числовые переменные были нормализованы для обеспечения однородности масштабов, что важно для нейросетей, так как они чувствительны к масштабу входных данных. Категориальные переменные преобразованы в числовой формат с использованием методов кодирования, таких как *one-hot encoding* или *label encoding*.

Для создания нескольких полных наборов данных использовались методы иммутации. Пропущенные значения заменялись на случайные из распределения существующих данных или на средние по соответствующим переменным. В обучении моделей применялись маски, указывающие на отсутствие значений в определенных ячейках для учета пропусков и адаптации к условиям неполноты.

Модели *MLP* и *RNN* обучены на различных наборах данных: сначала обучались на *полных наборах данных* без пропусков для установления базового уровня производительности, затем они обучались на наборах с *различными уровнями пропусков.* Для каждой модели проводилось несколько экспериментов с различными конфигурациями гиперпараметров (количество слоев, количество нейронов в каждом слое, скорость обучения и т.д.).

Эффективность моделей оценивалась с использованием следующих метрик: *точность* (до-

ля правильно предсказанных значений от общего числа предсказаний), *полнота* (способность модели находить все положительные примеры), *F1-мера* (гармоническое среднее между точностью и полнотой, что позволяет учитывать баланс между этими двумя метриками).

Проводился анализ ошибок для выявления закономерностей в неправильных предсказаниях, что помогло понять, какие типы неполноты данных наиболее критичны для каждой модели.

В ходе исследования проведена оценка эффективности нейросетевых моделей – многослойных перцептронов и рекуррентных нейронных сетей – в обработке как полных, так и неполных данных. Модели обучались на наборах с различными уровнями пропусков. Для каждой модели проводилось несколько экспериментов с различными конфигурациями гиперпараметров (количество слоев, количество нейронов в каждом слое, скорость обучения и т.д.).

II. Основные результаты

При обучении моделей на полных данных обе модели продемонстрировали высокую точность и согласованность результатов. *RNN* модель также показала высокую точность, приблизительно 93%, что подтверждает ее способность работать с временными рядами и последовательными данными. *MLP* модель достигла точности около 95% на текстовом наборе, что свидетельствует о ее способности эффективно выявлять сложные зависимости между переменными.

При переходе к анализу неполных данных результаты значительно варьировались в зависимости от уровня пропусков и выбранной модели:

- **10% пропусков.** Точность *MLP* снизилась на 90%. Модель все еще показывала хорошие результаты, используя методы иммутации для заполнения пропусков. Точность *RNN* также снизилась до 88%. Однако модель продемонстрировала способность учитывать временные зависимости, что помогло сохранить качество предсказаний;
- **20% пропусков.** Точность *MLP* упала до 85%. Это снижение связано с тем, что модель менее устойчива к случайным пропускам, особенно когда они касаются ключевых переменных. Точность *RNN* составила 82%. Несмотря на снижение производительности, модель все еще сохраняла способность анализировать временные ряды, что помогало в некоторых случаях предсказывать отсутствующие значения;
- **30% пропусков.** Точность *MLP* снизилась до 78%, что подчеркивает уязвимость модели к высокой степени неполных данных и делает ее менее предпочтительной для таких условий. Точность *RNN* составила 75%. Несмотря на значительное снижение производительности, *RNN* показала лучшие результаты по сравнению с *MLP* в условиях высокой неполноты, что свидетельствует о ее устойчивости к отсутствующим данным.

Выводы по результатам анализа:

RNN показала большую устойчивость к неполноте данных по сравнению с *MLP*, что связано с ее способностью учитывать предшествующие состояния и временные зависимости.

Использование методов иммутации оказало положительное влияние на производительность *MLP* при низком уровне неполноты. Однако при увеличении пропусков эффективность этих методов значительно снизилась.

В ходе анализа ошибок было установлено, что *MLP* чаще всего ошибалась при предсказании значений для переменных с высокой дисперсией. *RNN* же показывала более стабильные результаты даже при наличии значительных пропусков, но иногда не могла корректно предсказать изменения во временных рядах при резких скачках данных.

Визуализация результатов представлена на рисунке 1.

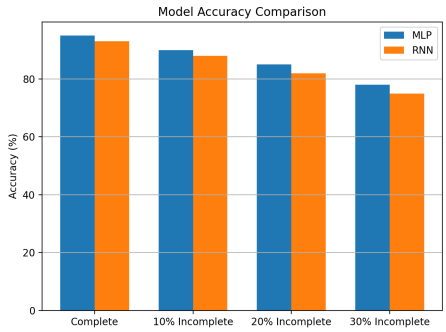

Рис. 1 – Сравнение точности моделей

Результаты исследования подтверждают, что неполнота данных существенно влияет на результаты социологических исследований. Нейросети могут эффективно обрабатывать такие данные, но выбор модели и метода обработки критически важен. *MLP* лучше подходит для статических данных, а *RNN* – для динамических изменений.

1. National Library of Medicine [Electronic resource] : The prevention and handling of the missing data. – Mode of access: `https://https://pmc.ncbi.nlm.nih.gov/articles/PMC3668100/`. – Date of access: 20.10.2024
2. Gordon D, Petousis P, Zheng H, Zamanzadeh D and Bui AA (2021) TSI-GNN: Extending Graph Neural Networks to Handle Missing Data in Temporal Settings. Front. Big Data 4:693869. doi: 10.3389/fdata.2021.693869
3. United Nations [Electronic resource] : International Migrant Stock 2020. – Mode of access: `https://www.un.org/development/desa/pd/content/international-migrant-stock`. – Date of access: 18.10.2024
4. TensorFlow [Electronic resource] : An end-to-end platform for machine learning. – Mode of access: `https://www.tensorflow.org/`. – Date of access: 20.10.2024
5. Keras [Electronic resource] : Keras – simple, flexible, powerfull. – Mode of access: `https://keras.io/`. – Date of access: 20.10.2024