Обеспечение применения контролируемого обучения к немаркированным данным о материнском здоровье с использованием кластеризации на основе псевдометок

И. А. Бессмертный
Университет ИТМО
bessmertny@itmo.ru

Б. Ч. Енкомариам
Университет ИТМО
checkt.birsh@gmail.com

***Аннотация*. Материнское здоровье остается одной из ключевых проблем общественного здравоохранения, особенно в условиях ограниченных ресурсов, где доступ к качественным медицинским услугам и надежным данным ограничен. Одной из основных проблем является отсутствие размеченных данных, что затрудняет применение методов контролируемого машинного обучения для прогнозирования. В данной работе предлагается фреймворк кластеризации с использованием псевдометок для анализа немаркированных данных о материнском здоровье. Подход объединяет неконтролируемое обучение, генерацию псевдометок и контролируемую классификацию с использованием итеративного механизма обратной связи. Алгоритм K-средних применяется для выявления скрытой структуры данных и формирования псевдометок. Предложенный метод улучшает качество классификации за счет повышения разделимости классов и представления признаков. Экспериментальные результаты демонстрируют эффективность подхода при работе с шумными и неполными медицинскими данными, что делает его практичным решением для систем здравоохранения с ограниченными ресурсами.**

Ключевые слова: материнское здоровье, машинное обучение, контролируемое обучение, немаркированные данные, псевдомаркировка, кластеризация, неконтролируемое обучение, классификация, полу-контролируемое обучение

I. Введение

Материнское здоровье остается важнейшей проблемой общественного здравоохранения, особенно в районах с ограниченными ресурсами, где ограничен доступ к качественным медицинским услугам и надежной информационной инфраструктуре. Во многих регионах африканских стран местные медицинские учреждения регулярно собирают данные о состоянии материнского здоровья; однако эти наборы данных часто не имеют маркировки, неполны и неоднородны, что ограничивает их полезность для прогнозного анализа и поддержки принятия решений. Отсутствие аннотированных результатов в сочетании с проблемами качества данных, такими как пропущенные значения и несоответствия, создает значительные барьеры для применения традиционных методов контролируемого машинного обучения.

Недавние достижения в области машинного обучения продемонстрировали эффективность подходов, основанных на данных, в задачах прогнозирования в области здравоохранения; однако большинство существующих подходов в значительной степени основаны на маркированных наборах данных, получение которых является дорогостоящим и трудоемким процессом, особенно в условиях ограниченных ресурсов, когда не хватает экспертных комментариев. Методы обучения под наблюдением и без присмотра, включая кластеризацию и псевдомаркировку, стали многообещающими альтернативами для использования немаркированных данных путем извлечения скрытых структур и создания суррогатных меток [1], [2]. В частности, алгоритмы кластеризации, такие как K-means, позволяют группировать похожие записи о пациентах, обеспечивая основу для создания псевдомаркировок и последующей классификации [3]. Концепция псевдомаркировки, представленная как простая, но эффективная стратегия обучения под наблюдением, продемонстрировала большой потенциал в преобразовании немаркированных данных в структурированные учебные задачи [4], [5]. Несмотря на эти достижения, существующие исследования часто рассматривают кластеризацию и классификацию как независимые процессы или применяют псевдомаркировку в хорошо организованных наборах данных [6]. Ограниченное количество исследований было сосредоточено на интегрированных платформах, которые сочетают псевдомаркировку на основе кластеризации с итеративными механизмами уточнения, особенно для зашумленных данных о состоянии материнского здоровья в реальном мире. Кроме того, надежность таких подходов при обработке неполных и некачественных медицинских записей остается недостаточно изученной [7].

Исходя из этих проблем, в данном исследовании предлагается гибридная система обучения без присмотра и под наблюдением врача, которая использует псевдомаркировки для обеспечения прогностического моделирования на основе полностью немаркированных местных данных о состоянии материнского здоровья за восемь лет.

Соответственно, в этом исследовании будут даны ответы на следующие исследовательские вопросы: RQ1: Как можно получить значимые прогностические метки из полностью немаркированных наборов данных о состоянии материнского здоровья, используя псевдомаркировку на основе кластеризации? RQ2: В

какой степени псевдомаркировка на основе кластеризации улучшает эффективность контролируемых классификационных моделей в прогнозировании состояния материнского здоровья? RQ3: Насколько надежна предлагаемая система при применении к зашифрованным и неполным записям о состоянии материнского здоровья в медицинских учреждениях с ограниченными ресурсами?

II. Связанные работы

Проблема обучения на основе немаркированных медицинских данных широко изучалась, особенно с помощью методов неконтролируемого обучения, полу-контролируемого обучения и псевдомаркировки. Неконтролируемое обучение широко используется для извлечения скрытых закономерностей из наборов медицинских данных, не требуя маркировки результатов.Например, [8] исследовали использование методов кластеризации для выявления сопутствующих заболеваний, связанных с беременностью, по данным о выписке из больницы. Автор применил K-Modes and Self-Organizing Maps (SOM) для группировки записей пациентов на основе сходства. Чтобы преодолеть ограничения, связанные с немаркированными данными, в медицинском машинном обучении все чаще используются псевдомаркировки. Псевдомаркировка (PL) – это широко используемый метод полууправляемого обучения (SSL), при котором используется сама модель для получения искусственных меток для немаркированных данных [9], [10], [11], [12]. Из рассмотренной литературы вытекает несколько ключевых замечаний, в том числе: (1). Неконтролируемые методы эффективны для выявления закономерностей в данных о матери и медицинском обслуживании, но не обладают прогностическими возможностями (2). Методы псевдомаркировки позволяют проводить обучение на основе немаркированных данных, но в значительной степени зависят от качества маркировки, и (3). Гибридные методы, сочетающие кластеризацию и классификацию, обладают большим потенциалом, но все еще недостаточно изучены в наборах данных о материнском здоровье, особенно в условиях нехватки ресурсов, включая страны Африки к югу от Сахары.

Таким образом, в отличие от существующих исследований, которые либо полагаются на помеченные данные, либо используют кластеризацию исключительно для выявления закономерностей, в этой работе, во-первых, используется неконтролируемое машинное обучение в качестве механизма псевдомаркировки, позволяющего проводить прогностическое моделирование на основе некачественных, немаркированных данных об истории здоровья матери в условиях ограниченных ресурсов, где аннотации соответствуют действительности. недоступны. Затем демонстрируется, как использование псевдометок в качестве входных данных для классов улучшает производительность прогнозирования за счет представления объектов и возможности их разделения.

III. Методы исследования

A. *Общий обзор предлагаемых рамок*

Предлагаемый подход представляет собой систему псевдомаркировки на основе итеративной кластеризации для прогнозирующего моделирования с использованием данных немаркированных обследований состояния материнского здоровья, собранных в медицинских учреждениях с ограниченными ресурсами.

Платформа сочетает в себе обучение без контроля, генерацию псевдотегов, контролируемое обучение и механизмы обратной связи на основе оценки для постепенного повышения эффективности прогнозирования. Поскольку теги проверки недоступны, платформа использует кластеризацию для создания суррогатных исходных данных для мониторинга.

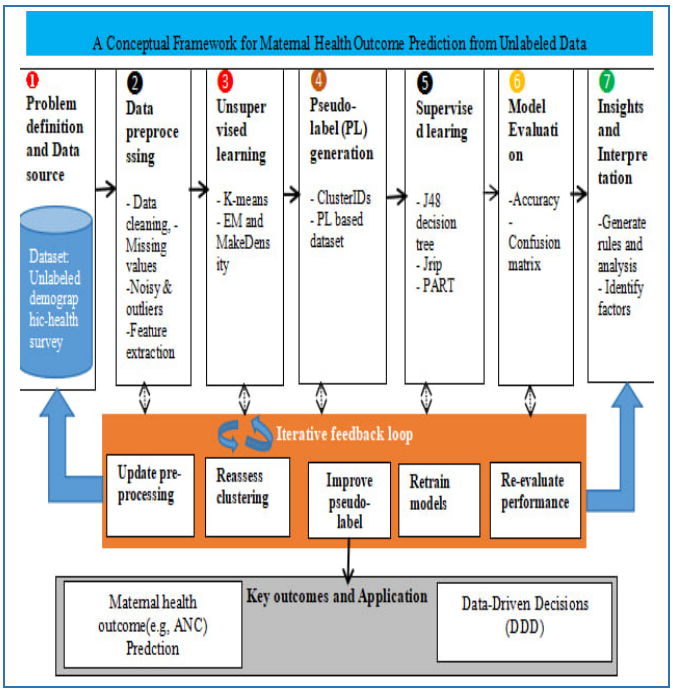

Рис. 1. Концептуальная основа для прогнозирования состояния материнского здоровья с итеративным циклом обратной связи с использованием немаркированных демографических данных

B. *Описание набора данных*

Выборочный набор данных, использованный в этом исследовании, состоит из реальных данных о состоянии материнского здоровья, полученных в ходе демографического обследования, проведенного в Университете Гондара, Эфиопия. Этот набор данных представляет собой записи за восемь лет, по своей сути не имеет маркировки и отражает реальное демографическое и клиническое состояние, характеризующееся отсутствием значений для нескольких атрибутов, несогласованными форматами записи данных, сочетанием категориальных и числовых характеристик и отсутствием допустимых обозначений классов.

C. *Предварительная обработка данных*

Для обеспечения стабильности модели были применены следующие этапы предварительной обработки: избыточные и сильно противоречивые записи были удалены вручную, вычисление пропущенных значений с использованием стратегии mean/mode, обработка зашумленных данных и выбросов с использованием методов mode и ожидаемой максимизации (EM), кодирование категориальных переменных выполнено, и пропущенные значения были уменьшены более чем на 50%. В консультации с

экспертами в предметной области и в соответствии с [13] из общего числа 40 атрибутов; было сокращено количество 15 нерелевантных атрибутов, 8 менее важных атрибутов, основанных на более низком коэффициенте усиления [14], и 4 пропущенных значения, превышающих 50%. Наконец, для этого исследования было использовано в общей сложности 5990 предварительно обработанных записей с 13 выбранными атрибутами.

IV. Эксперименты и результаты

Предложенный фреймворк был реализован с использованием стандартных библиотек машинного обучения, основанных на средстве машинного обучения Waikato Environment for Knowledge Analysis (WEKA). Соответственно, для целей машинного обучения без контроля использовались оба алгоритма кластеризации разделов, в частности алгоритмы кластеризации на основе K-средних (на основе расстояния) и EM (на основе модели), а также алгоритмы кластеризации на основе плотности. Затем, в качестве прогностической модели, для этого исследования были выбраны три хорошо известных алгоритма машинного обучения с контролем, такие как дерево решений J48 и PART & JRip с указанием правил, основываясь на их хорошей производительности во время тестирования.

Для измерения качества псевдомаркировки были использованы кластерная сумма квадратичных ошибок (SSE), количество итераций (i), временная сложность и мнение эксперта в предметной области, основанное на [15]. В качестве оценочных показателей для прогностической модели использовались точность, прецизионность, запоминание, показатель F1, коэффициент достоверности (TP), матрица путаницы и ROC. В качестве стратегии валидации было принято разделение тренировочного теста на 80/20 из-за его относительно низкой погрешности и вариаций [16] для целей сравнения.

A. Генерация и анализ псевдомаркировки

Модель кластеризации с использованием K-средних значений с K= 3, евклидовым расстоянием и начальным значением 10 была выбрана в качестве оптимальной модели кластеризации на основе более низкой суммы квадратов ошибок (SSE) внутри кластера и более быстрой сходимости по сравнению с методами, основанными на EM и плотности. Это указывает на то, что результирующие кластеры являются компактными и внутренне согласованными, что приводит к созданию более надежных суррогатных меток. Стабильность решения для кластеризации обеспечивает воспроизводимость назначений псевдометок. Кроме того, выявленные кластеры потенциально соответствуют различным скрытым характеристикам состояния материнского здоровья, что улучшает возможность разделения на классы и повышает эффективность классификации.

B. Контролируемое обучение

Используя псевдомаркировки в качестве входных данных, экспериментальные результаты показывают, что PART и превосходит J48 и JRip при обучении на данных с псевдомаркировкой. Это можно объяснить природой псевдометок, генерируемых с помощью кластеризации, которые по своей сути содержат структурные закономерности наряду со степенью шума и неопределенностью границ. В отличие от J48, который строит глобальное дерево решений и чувствителен к несоответствиям меток, и JRip, который опирается на жесткую индукцию правил, PART сочетает обучение на основе правил с частичным построением дерева решений. Этот гибридный механизм позволяет ИТ-отделу лучше улавливать локализованные области принятия решений и обрабатывать перекрывающиеся границы кластеров. В результате PART более устойчив к шумам псевдометок и более эффективно использует скрытую структуру, закодированную в выходных данных кластеризации, что приводит к повышению эффективности прогнозирования.

C. Сравнение моделей прогнозирования

ТАБЛИЦА I. СРАВНЕНИЕ МОДЕЛЕЙ ПРОГНОЗА

Модель классификации	Общая точность				Время
	Правильная классификация		Неправильная классификация		
	No	%	No	%	
Unpruned J48 decision tree со всеми атрибутами с использованием 80-ого разбиения	1495	99.87%	2	0.13%	0.01
Pruned PART rule induction со всеми атрибутами с использованием 80-ого разбиения	1496	99.93 %	1	0.07%	0.07
Pruned JRip rule induction со всеми атрибутами с использованием 80-ого разбиения	1494	99.81 %	3	0.19%	0.27

Как показано в табл. 1, классификатор ДЕТАЛЕЙ достиг точности 99,93% при обучении на псевдометодах, полученных в результате кластеризации. Это указывает на то, что псевдомаркировки внутренне непротиворечивы и структурно надежны, что позволяет классификатору практически идеально их изучать и может быть эффективно аппроксимировано моделями, основанными на правилах. Однако такая высокая точность также отражает способность модели воспроизводить задания кластеризации, которые могут не точно предсказывать клинически подтвержденные результаты. Таким образом, несмотря на то, что результаты подтверждают структурную согласованность псевдометок, для оценки эффективности прогнозирования в реальных условиях необходима дальнейшая проверка с помощью меток, подтверждающих достоверность данных.

V. Обсуждение

В начале этого исследования необходимо ответить на три исследовательских вопроса, и давайте обсудим, как на эти вопросы были даны ответы в этом исследовании.

A.RQ1: Получение прогностических меток из немаркированных данных – “Как можно получить значимые прогностические метки из полностью немаркированных наборов данных о состоянии материнского здоровья, используя псевдомаркировку на основе кластеризации?” – Это исследование

демонстрирует, что значимые прогностические метки могут быть получены с помощью кластеризации по K-среднему значению, которая группирует записи о состоянии материнского здоровья в структурно согласованные кластеры на основе сходства признаков.

B.RQ2: Влияние на эффективность классификации-- В какой степени псевдомаркировка на основе кластеризации улучшает производительность контролируемых моделей классификации?- Результаты показывают, что псевдомаркировка значительно повышает эффективность классификации. PART достиг точности 99,93%, что указывает на высокую степень обучаемости структуры псевдометок, которая преобразует проблему без маркировки в структурированную контролируемую задачу.

C.RQ3: Надежность в условиях зашумленных и низкокачественных данных. Насколько надежна предлагаемая система в применении к зашифрованным и неполным записям о состоянии здоровья матерей? Предлагаемая система устойчива к недостаткам данных и подходит для медицинских учреждений с ограниченными ресурсами, что демонстрирует надежность несколькими способами: (1) Эффективная предварительная обработка позволяет избежать пропусков и несогласованности данных; (2) Кластеризация фиксирует базовую структуру, несмотря на шум, и (3) Классификатор деталей демонстрирует высокую производительность, что указывает на устойчивость к шуму от псевдометок.

В целом, исследование подтверждает, что псевдомаркировка на основе кластеризации обеспечивает жизнеспособный путь для прогностического моделирования в наборах данных о материнском здоровье без маркировки, повышает эффективность классификации за счет структурированного контроля и поддерживает надежность в зашумленных реальных средах обработки данных. Полученные результаты подчеркивают потенциал псевдомаркировки как масштабируемого решения для медицинской аналитики в условиях, когда данные с маркировкой недоступны, обеспечивая основу для будущих исследований в области прогнозирования состояния материнского здоровья на основе данных.

VI. Будущие исследования и заключение

В этом исследовании была предложена основанная на кластеризации система псевдомаркировки, позволяющая проводить прогностическое моделирование на основе полностью немаркированных локальных наборов данных о состоянии материнского здоровья. Результаты демонстрируют, что псевдомаркировки, полученные на основе K-средних, могут эффективно преобразовывать немаркированные данные в структурированную задачу обучения, позволяя контролируемым моделям, особенно PART, достигать высокой согласованности прогнозирования. Полученные результаты подтверждают, что кластеризация улучшает разделение классов и поддерживает надежное обучение даже в условиях шума и неполноты данных, типичных для условий с ограниченными ресурсами.

Наблюдаемое превосходство PART позволяет предположить, что псевдомаркированные данные привносят структурные закономерности наряду с присущим им шумом и неопределенностью границ. В будущей работе будут изучены стратегии псевдомаркировки, учитывающие шум, включая взвешенные по достоверности и вероятностные подходы к маркировке, для повышения надежности маркировки.

Список литературы

[1] Xiaojin Zhu , Andrew B. Goldberg, Introduction to Semi-Supervised Learning, doi.org/10.1007/978-3-031-01548-9, 2009, Springer Cham, O. Chapelle, B. Schölkopf, and A. Zien, Eds., Semi-Supervised Learning. Cambridge, MA, USA: MIT Press, 2006.

[2] Jin, X., Han, J. K-Means Clustering. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning, 2011, Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_425.

[3] Y. Grandvalet and Y. Bengio, “Semi-supervised learning by entropy minimization,” in Advances in Neural Information Processing Systems (NeurIPS), 2005.

[4] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co–training," in Proc. of the 11th COLT, 1998, pp. 92–100.

[5] Wei Y, Deng Y, Sun C, Lin M, Jiang H, Peng Y. Deep learning with noisy labels in medical prediction problems: a scoping review. J Am Med Inform Assoc. 2024 Jun 20;31(7):1596-1607. doi: 10.1093/jamia/ocae108.

[6] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S. Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertaintyaware pseudo-label selection framework for semi-supervised learning. CoRR, abs/2101.06329, 2021.

[7] Chang J, Sarkar IN. Using Unsupervised Clustering to Identify Pregnancy Co-Morbidities. AMIA Jt Summits Transl Sci Proc. 2019 May 6;2019:305-314. PMID: 31258983; PMCID: PMC6568081.

[8] D.-H. Lee, “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” in ICML Workshop on Challenges in Representation Learning, 2013.

[9] Weiwei Shi, Yihong Gong, Chris Ding, Zhiheng Ma, Xiaoyu Tao, and Nanning Zheng. Transductive semi-supervised deep learning using min-max features. In ECCV, pp. 311–327, 2018.

[10] Eric Arazo, Diego Ortego, Paul Albert, Noel E O’Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In IJCNN, pp. 1–8. IEEE, 2020.

[11] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In CVPR, pp. 5070–5079, 2019.

[12] Jiawei Su, Zhiming Luo, Sheng Lian, Dazhen Lin, Shaozi Li,Mutual learning with reliable pseudo label for semi-supervised medical image segmentation,Medical Image Analysis,Volume 94,2024,103111,ISSN 1361 8415; doi.org/10.1016/j.media.2024.103111

[13] Aniket K. Shahade, Priyanka V. Deshmukh, Pritam H. Gohatre, Kanchan S. Tidke, Rohan Ingle, Method for fetal ultrasound image classification using pseudo-labelling with PCA-KMeans and an attention-augmented MobileNet-LSTM model, Methods X, Volume 15, 2025, 103563, ISSN 2215-0161, https://doi.org/10.1016/j.mex.2025.103563.

[14] Cios Krzysztof J, Pedrycz Witold, Swiniarski Roman W, Kurgan Lukasz A. Data Mining: A Knowledge Discovery Approach. New York, USA: Springer Science Business Media LLC; 2007.

[15] S. Theodoridis and K. Koutroubas. Pattern Recognition. Academic Press, 1999.