

Вестник Череповецкого государственного университета. 2023. № 5 (116). С. 39–51.
Cherepovets State University Bulletin, 2023, no. 5 (116), pp. 39–51.

Научная статья

УДК 004.85

<https://doi.org/10.23859/1994-0637-2023-5-116-3>

EDN: CVVBLH

Гибридный алгоритм активного обучения для определения намерений в диалоговых системах с неразмеченными и шумными данными

Даниил Игоревич Дранга

Национальный исследовательский технологический университет МИСиС,
Москва, Россия

dranga.daniil@gmail.com, <https://orcid.org/0000-0002-3929-8161>

Аннотация. В статье исследуется оптимизация отбора данных для решения задачи классификации намерений в целеориентированных диалоговых системах. Для обучения модели требуются данные, размеченные ассессорами, извлеченные из неразмеченных диалогов. Процесс отбора объектов для аннотации представляет особую сложность, поскольку целевой набор намерений представляет собой лишь часть из возможных намерений, присутствующих в клиентских сообщениях. Дополнительную трудность представляет тот факт, что множество сообщений не несут в себе конкретных намерений. В статье предлагается гибридный алгоритм, учитывающий особенности задачи и объединяющий в себе методы обучения без учителя и с учителем, чтобы решить проблему отбора данных для разметки. Эксперименты на реальных диалогах показали эффективность алгоритма по сравнению с альтернативными подходами. Алгоритм применим к различным моделям классификации.

Ключевые слова: активное машинное обучение, модульные диалоговые агенты, кластеризация, классификация намерений, неразмеченные данные, шумные данные

Благодарность. Работа выполнена под руководством кандидата технических наук А. С. Кожаринова

Для цитирования: Дранга Д. И. Гибридный алгоритм активного обучения для определения намерений в диалоговых системах с неразмеченными и шумными данными // Вестник Череповецкого государственного университета. 2023. № 5 (116). С. 39–51.
<https://doi.org/10.23859/1994-0637-2023-5-116-3>

Hybrid active learning algorithm for intent classification in dialogue systems with unlabeled noisy data

Daniil I. Dranga

National University of Science and Technology MISiS,

Moscow, Russia

dranga.daniil@gmail.com, <https://orcid.org/0000-0002-3929-8161>

Abstract. This article investigates the optimization of data selection for the task of intent classification in goal-oriented dialogue systems. Training the model requires data labeled by assessors extracted from unlabeled dialogues. Selecting objects for labeling is challenging because the target intent set represents only a subset of the potential intent present in customer messages. An additional difficulty is the fact that many messages lack intent. The article proposes a hybrid algorithm that takes into account the task specificities and combines unsupervised and supervised learning methods to address the data selection problem for labeling. Experiments on real dialogues demonstrate the algorithm effectiveness compared to alternative approaches and its applicability to various classification models.

Keywords: active machine learning, modular dialogue agents, clustering, intent classification, unlabeled data, noisy data

Acknowledgments. The research was conducted under the supervision of A. S. Kozharinov, Candidate of Technical Sciences.

For citation: Dranga D. I. Hybrid active learning algorithm for intent classification in dialogue systems with unlabeled noisy data. *Cherepovets State University Bulletin*, 2023, no. 5 (116), pp. 39–51 (In Russ.). <https://doi.org/10.23859/1994-0637-2023-5-116-3>

Введение

Диалоговые агенты являются особым классом программного обеспечения, предназначенным для решения определенных задач путем взаимодействия с пользователями в диалоговой форме. Их применение в различных отраслях обуславливает потребность в универсальных методах и технологиях.

Целеориентированный диалоговый агент (например, чат-бот в банковской сфере) помогает клиентам управлять услугами, открывать новые продукты и предоставлять информацию о комиссиях. Главная задача такого агента – обеспечить максимальное число корректно решенных запросов.

Современные диалоговые агенты основаны на моделях обработки естественного языка. Их обучение требует представительных и аннотированных выборок, что влечет затраты на разметку данных. Чтобы сократить эти затраты, разрабатываются методы активного обучения¹, направленные на максимизацию качества модели при ограниченном количестве запросов на разметку. Активное обучение является областью машинного обучения, которая нацелена на максимизацию качества модели, используются размеченные данные с ограничениями количества обращений к асес-

¹ Demir E., et al. Unsupervised Active Learning For Video Annotation. ICML Active Learning Workshop 2015. 2015; Perez F., Lebrete R., Aberer K. Weakly supervised active learning with cluster annotation. arXiv preprint arXiv:1812.11780. 2018; Wang T., Chen S., Jia R. One-round active learning. arXiv preprint arXiv:2104.11843. 2021.

сору для разметки неразмеченных объектов. Из неразмеченных данных выбираются наиболее информативные объекты для аннотации. В контексте создания диалогового агента классификатор пользовательских намерений является основным компонентом, поскольку он определяет основное предполагаемое желание клиента на основе сообщения. Автоматизация обработки клиентских обращений без участия человека требует точного определения всех намерений, выраженных в ходе диалога. На данный момент наибольшую точность в этой задаче обеспечивают методы на основе нейронных сетей¹, в частности, с применением метрического обучения.

Предлагаемый автором алгоритм применим к широкому спектру моделей классификации, включая нейронные сети с разнообразными архитектурами, деревья решений, градиентный бустинг над решающими деревьями и др. В экспериментах рассматривается подход, основанный на сиамских нейронных сетях², с целью получения эмбедингов пользовательских сообщений.

Сбор и разметка данных для обучения классификатора намерений имеют специфические особенности, поскольку анализируемый набор намерений является подмножеством всех возможных намерений в клиентских сообщениях. Многие обращения не содержат явных намерений и рассматриваются как шум. Современные методы активного обучения недостаточно развиты для решения подобных задач и не адаптированы для классификации пользовательских намерений в диалоговых сообщениях.

Можно выделить три ключевые особенности в этой задаче:

- при выделении информативных сообщений отбираются сообщения клиентов с высокой неуверенностью модели, не содержащие намерений. После разметки таких сообщений они не попадают в обучающую выборку и не повышают качество работы модели;
- при использовании взвешивания объектов по плотности репрезентативность других групп сообщений не сохраняется. Формируется однородная выборка с низкой вариативностью. Это приводит к формированию похожих сообщений, которые не проходят множество процедурных проверок и не дают желаемого результата;
- подходы активного обучения в этой задаче часто отбирают кластера сообщений, которые содержат намерения, где чат-бот не умеет обрабатывать (и по этой причине их нет в обучающей выборке). Разметка таких сообщений также не улучшает работу модели, так как они не добавляются в обучающие данные.

Статьи³ представляют комбинацию методов обучения без учителя и активного обучения для аннотирования видеоматериалов, сенсорных данных и спутниковых

¹ Shen Y., et al. Enhancing the generalization for Intent Classification and Out-of-Domain Detection in SLU. arXiv preprint arXiv:2106.14464. 2021; Liu B., Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling. arXiv preprint arXiv:1609.01454. 2016.

² Ren F., Xue S. Intention detection based on siamese neural network with triplet loss. IEEE Access. 2020. Vol. 8. P. 82242–82254.

³ Demir E., et al. Unsupervised Active Learning For Video Annotation. ICML Active Learning Workshop 2015. 2015; Souza V. et al. Unsupervised active learning techniques for labeling training sets: an experimental evaluation on sequential data. Intelligent Data Analysis. 2017. Vol. 21. No. 5.

изображений. В 2019 году исследователь из Amazon Research Федор Жданов предложил применять кластеризацию с использованием метода k -средних и отбор k объектов, наиболее приближенных к центру кластера, с целью обеспечения разнообразия внутри пакета в задаче активного обучения для классификации частично размеченных данных¹.

Автор текущей работы развивает направление методов активного обучения с учетом и плотности распределения, и разнообразия данных. В статье предлагается применение алгоритма кластеризации DBSCAN для группировки данных, выбора опорных точек и фильтрации шума, что редко используется в современных исследованиях. Хотя активное обучение широко применяется в работах по обработке естественного языка², его использование в задаче классификации пользовательских намерений недостаточно исследовано, и модификации алгоритмов, учитывающие специфику этой задачи, не разработаны.

В работе представлен новый гибридный алгоритм, специально разработанный для эффективной классификации пользовательских намерений. Он включает в себя методы обучения без учителя, фильтрацию текстов диалогов на основе близости и алгоритмы активного обучения. Разработанный алгоритм был протестирован на реальных данных и показал высокое качество работы, что делает его применимым для решения практических задач. Алгоритм имеет модульную структуру, которая позволяет адаптировать его под смежные задачи.

Основная часть

Постановка задачи классификации намерений

Рассмотрим постановку задачи классификации намерений. На вход классификатору подаются векторные представления сообщений клиентов $x \in R^{EMB_DIM}$ (x – векторное представление, R – множество вещественных чисел, EMB_DIM – размерность вектора), которые находятся в семантическом признаковом пространстве (эмбединги). Эти представления обладают свойством близости для семантически близких сообщений. На основе выборки обращений с известными намерениями разрабатывается алгоритм a , который может классифицировать намерения в новых сообщениях.

Для автоматизации обработки обращений клиентов без необходимости в привлечении человека необходимо точно определять все намерения в диалоге. Это приво-

P. 1061–1095; Gal Y., Islam R., Ghahramani Z. Deep bayesian active learning with image data. International conference on machine learning. PMLR, 2017. P. 1183–1192.

¹ Zhdanov F. Diverse mini-batch active learning. arXiv preprint arXiv:1901.05954. 2019.

² Hoi S. C. H., Jin R., Lyu M. R. Large-scale text categorization by batch mode active learning. Proceedings of the 15th international conference on World Wide Web. 2006. Pp. 633–642; Lewis D. D. A sequential algorithm for training text classifiers: Corrigendum and additional data. Acm Sigir Forum. New York, NY, USA: ACM, 1995. Vol. 29. No. 2. P. 13–19; Tong S., Koller D. Support vector machine active learning with applications to text classification. Journal of machine learning research. 2001. Vol. 2. November. P. 45–66; Zhou Meng, Zechen Li, and Pengtao Xie. Self-supervised regularization for text classification // Transactions of the Association for Computational Linguistics. 2021. No. 9. P. 641–656.

дит к задаче минимизации функционала качества, которым является доля неправильно определенных намерений $\mathcal{J}(a, x) = [a(x) \neq o(x)]$, где $o(x)$ – истинное намерение пользователя и $a(x)$ – определенное классификатором.

Автор описывает задачу классификации намерений при условии, что каждое обращение содержит одно единственное намерение, которое должно быть определено классификатором. На практике соблюдение этого условия обеспечивается системами диалога высокого уровня, которые занимаются обработкой более сложных случаев.

Необходимо собрать обучающую выборку для классификатора намерений, для этого ассессоры размечают сообщения из большого массива неразмеченных данных. Эти сообщения могут содержать различные намерения, включая новые или шум. Ресурсы на разметку ограничены, поэтому необходимо отобрать наиболее полезные объекты для пополнения обучающей выборки с целью улучшения качества модели.

Рассмотрим оптимизацию процесса накопления размеченных данных для классификатора намерений.

Постановка задачи активного обучения

Активное обучение оптимизирует качество модели машинного обучения с ограниченным количеством обращений к оракулу для разметки неразмеченных данных. Это достигается через итеративный процесс: модель обучается на небольшом начальном датасете, затем из большого набора данных выбираются наиболее значимые примеры для разметки оракулом. После разметки эти примеры добавляются в размеченный датасет и процесс повторяется. Активное обучение, включающее отбор объектов из выборки, обычно завершается по одному из двух критериев: исчерпание бюджета на разметку оракулом или достижение требуемого качества модели. Рис. 1 иллюстрирует процесс активного обучения с отбором объектов из выборки.

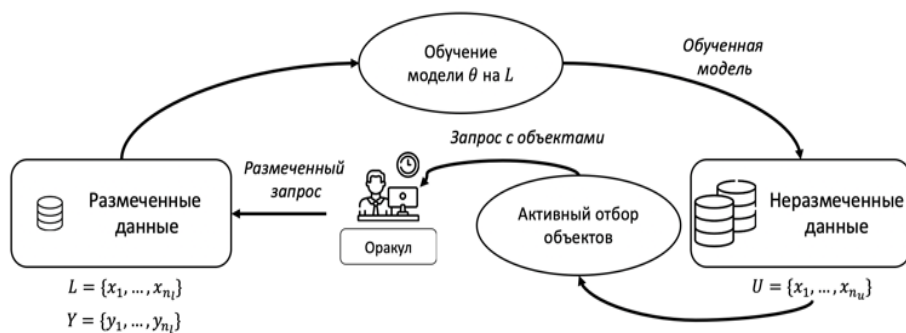


Рис. 1. Схема работы активного обучения

В контексте задачи активного обучения для классификации интенций определяется множество объектов $X = \{x_1, x_2, \dots, x_n\}$, $x_i \in R^{EMB_DIM}$, где каждый элемент x_i – является векторным представлением текстового сообщения клиента в пространстве признаков. Множество допустимых ответов обозначается как $Y = \{class_1, class_2, \dots, class_l, class_{l+1}, \dots, class_{l+kn}, noise\}$, а его подмножество, состоя-

щее из интенций, добавленных в размеченную выборку агентом, определяется как $Y_c = \{class_1, class_2, \dots, class_1\}$.

Оракул $O: X \rightarrow Y$ – это функция, которая принимает объект из множества X и возвращает метку из множества Y . Использование этой функции стоит определенное количество бюджета. Доступный бюджет ограничен $Budget$.

Рассмотрим множество неразмеченных объектов $U \subset X = \{x_1, \dots, x_m\}$. Также имеется размеченное множество объектов $L \subset X = \{x_1, \dots, x_m\}$, для которых известны метки $y_{ci} = O(x_i) \in Y_c$. Пары объектов $XL = (x_i, y_{ci})$, где $x_i \in L$ и $y_{ci} \in Y_c$, составляют обучающую выборку с классами интенций. Объекты, для которых оракул восстанавливает метку из Y_c , добавляются в размеченное множество объектов L . Также имеется размеченное множество объектов $K \subset X = \{x_1, \dots, x_{nk}\}$, для которых значения $y_{ci} = O(x_i) \in Y_c$ известны. Пары объектов $KK = (x_i, y_i)$, где $x_i \in K$ и $y_{ci} \in Y_c$, составляют контрольную выборку с классами интенций.

Предположим, что задан алгоритм классификации $A_w(x)$ с набором параметров $w \in W$, приближающий целевую функцию оракула $O(x)$ для объектов обучающей выборки X_b, Y_b . В наличии также имеется множество неразмеченных данных U . Цель состоит в разработке алгоритма $\theta_N(U): U \rightarrow B$, где $B = \{1, 0\}$, а N – размер бюджета, $N \leq Budget$, позволяющего выбрать подмножество неразмеченных данных $U_L = \theta(U)$ мощности N из множества U . Выбранное подмножество U_L передается на разметку, после чего алгоритм $A_w(x)$ обучается с использованием новых размеченных данных.

Алгоритм $\theta_N(U)$ должен выбрать из множества неразмеченных данных U подмножество U_L размером N и разметить его оракулом, чтобы дообучить алгоритм $A_w(x)$ и достичь максимального роста качества классификации. Цель алгоритма – минимизировать итоговый функционал качества алгоритма $\mathcal{Q}(A_w) = [A_w(x) \neq o(x)]$ на контрольной выборке K , т. е. максимизировать долю правильно определенных объектов.

Предлагаемый подход

В статье предлагается гибридный итеративный алгоритм активного обучения для решения задачи классификации набора классов с неполным покрытием на основе неразмеченных зашумленных данных. Алгоритм описан в контексте классификации намерений и включает 4 основных этапа. Он строит последовательность $\theta_N^1(U), \theta_N^2(U), \dots, \theta_N^I(U)$, где I – количество итераций с обучением модели, которая по завершении работы сходится к итоговой модели $\theta_N(U)$.

Первый этап включает в себя основанную на плотности пространственную кластеризацию (DBSCAN) для данных с шумом с отметкой опорных объектов и шумовых точек, инициализируя алгоритм с небольшими значениями e и $minPts$, подобранными исходя из данных. Далее, на втором этапе, происходит оценка информативности каждого объекта, основываясь на энтропии классификации в качестве показателя неопределенности модели и близости объектов к уже определенным. На третьем этапе происходит отбор объектов с наиболее высокой информационной ценностью для дальнейшего обучения модели. На четвертом этапе модель дообучается

на выбранных объектах, и качество оценивается на контрольной выборке. Процесс повторяется до стабилизации качества модели или исчерпания бюджета обучения.

Предлагаемый алгоритм итеративно обучается на неразмеченных и зашумленных данных в задаче классификации набора классов. Инициализация с небольшой величиной параметра ϵ для алгоритма DBSCAN с итеративным увеличением позволяет начать с наиболее плотных кластеров и постепенно охватывать разреженные области, обучая модель на все большем количестве объектов до стабилизации качества или исчерпания бюджета разметки.

В первой части алгоритма используется метод обучения без учителя для группировки объектов в кластеры. С помощью алгоритма DBSCAN выделяются кластеры похожих по смыслу запросов, в то время как шумовые объекты исключаются из рассмотрения. Если при заданных параметрах ϵ и $minPts$ не удастся выделить кластеры, то значение $minPts$ уменьшается. Поиск продолжается до тех пор, пока не будет найден хотя бы один содержательный кластер, или до тех пор, пока не будут рассмотрены все неразмеченные данные. В этом случае все оставшиеся данные считаются шумом для заданного уровня ϵ .

Во второй части алгоритма (в строках 9–11) производится оценка ценности каждого объекта. Алгоритм для активного обучения включает два этапа: оценку степени неуверенности классификации модели и оценку семантической близости с уже существующими намерениями. Информативность объекта связана с неуверенностью модели в классификации. Для учета этого модель обучается на имеющихся размеченных данных и используется для всех неразмеченных данных. Затем оценивается степень неуверенности модели для каждого объекта с помощью расчета энтропии классификации по формуле (1):

$$ENT_x = \sum P_{\Theta}(y_i|x) \log P_{\Theta}(y_i|x).$$

В алгоритме происходит оценка расстояния между каждым сообщением из множества U и элементами множества L с использованием косинусного сходства векторов. Сообщения, семантические представления которых имеют близкое расположение к объектам размеченной выборки, получают более высокий приоритет для аннотирования. Вычисление косинусного сходства основано на следующей формуле (2):

$$d(x, y) = (x \times y) / (|x| |y|). \quad (2)$$

Algorithm 1: Гибридный метод активного обучения в задаче классификации заданного набора классов с неполным покрытием по неразмеченным зашумленным данным

Input:
 N – количество классов в размеченной выборке
 Sim – порог для включения объекта в «просмотренную» область
 ϵ – инициальная область поиска для *DBSCAN*
 $minPts$ – инициальное количество объектов для образования кластера
 $Budget$ – количество объектов, на разметку которых рассчитан бюджет

Output:
 Модель θ

Data:
 L – множество размеченных объектов
 U – множество неразмеченных объектов
 V – множество просмотренных объектов (изначально пустое)

```

1 while Budget > 0 do
2     clusters =  $\emptyset$ 
3     while ( $minPts > 0$ ) and ( $Length(set(clusters)) \leq 1$ ) do
4          $Clusters_U, CorePoints_U \leftarrow DBSCAN(\epsilon, minPts)$  on  $U$ 
5          $clusters \leftarrow set(Clusters_U)$ 
6         if  $Length(set(clusters)) \leq 1$  then
7              $minPts = minPts - 1$ 
8          $U_f \leftarrow U[Clusters_U \neq Noise]$ 
9          $\theta \leftarrow$  Обучаем модель на данных  $L$ 
10         $ENT_{U_f} \leftarrow$  Считается по формуле (1)
11         $DIST_{U_f \rightarrow L} \leftarrow$  Для каждого объекта  $U_f$  считается расстояние до  $L$  по (2)
12        for  $cl$  in clusters do
13             $Cl_{ENT}[cl] \leftarrow \mathbb{E}[ENT_{U_f} | Clusters_{U_f} = cl]$ 
14             $Cl_{DIST}[cl] \leftarrow \mathbb{E}[DIST_{U_f} | Clusters_{U_f} = cl]$ 
15             $Cl_{ENT}[cl] \leftarrow Count([Clusters_{U_f} = cl])$ 
16            Рассчитываются ранги оценок:  $Cl_{ENT}^{Rank}[cl]$ ,  $Cl_{DIST}^{Rank}[cl]$  и  $Cl_{SIZE}^{Rank}[cl]$ 
17             $Cl_{Total}^{Rank}[cl] \leftarrow Cl_{ENT}^{Rank}[cl] + Cl_{DIST}^{Rank}[cl] + Cl_{SIZE}^{Rank}[cl]$ 
18        CorePointsFromSortedClusters  $\leftarrow$  Сортируются  $CorePoints_U$  по RankTotal
19        соответствующему объекту кластеру
20        for CorePoint in CorePointsFromSortedClusters do
21            if Budget > 0 then
22                Удаляем CorePoint из  $U$ .
23                Budget = Budget - 1
24                if CorePoint относится к существующему в  $L$  классу then
25                    CorePoint добавляется в  $L$ 
26                else
27                    CorePoint добавляется в  $V$ 
28            else
29                 $\theta \leftarrow$  Обучаем модель на данных  $L$ 
30                Возвращаем модель  $\theta$ 
31         $DIST_{U_f \rightarrow V} \leftarrow$  Для каждого объекта  $U_f$  считается расстояние до  $V$  по
32        (2)
33         $U = U [DIST_{U_f \rightarrow V} > Sim]$ 
34        Удаляем из  $U$  объекты, расстояние от которых до  $V$  меньше  $Sim$ 
35         $\epsilon \leftarrow \epsilon * (1 + \epsilon_{delta})$ 
36     $\theta \leftarrow$  Обучаем модель на данных  $L$ 
37    Возвращаем модель  $\theta$ 
    
```

Рис. 2. Псевдокод алгоритма

На третьем этапе (строки 12–17) происходит оценка важности кластеров, включающая три составляющих: среднее значение энтропии классификации, среднее значение расстояния до объектов обучающей выборки и количество объектов в кластере. Для сопоставимости и ранжирования кластеров по значимости составляющие заменяются на соответствующие ранги. Итоговая оценка важности кластера определяется как среднее арифметическое трех рангов.

В заключительном этапе алгоритма (строки 18–29) производится обход кластеров согласно их оценкам важности, после чего выбираются ключевые точки из каждого кластера до исчерпания доступного бюджета или обработки всех обнаруженных кластеров. Выбранные ключевые точки передаются для аннотации. Размеченные объекты разделяются на два множества: множество L для сообщений, относящихся к существующим намерениям, и множество V для просмотренных объектов без намерений. Затем объект удаляется из неразмеченного множества.

По завершении каждого прохода цикла значение e увеличивается, и алгоритм запускает новую итерацию с обновленной моделью и уменьшенным множеством неразмеченных объектов. Цикл продолжается до тех пор, пока средства на аннотацию объектов не будут исчерпаны. В результате получается окончательная обученная модель.

Результаты экспериментов

Для оценки эффективности предложенного алгоритма были использованы реальные данные классификатора намерений диалогового агента, который был обучен на тематике банковских услуг. В качестве исходных данных применялось неразмеченное множество, состоящее из реальных сообщений клиентов, отправленных через различные каналы коммуникации с банком.

Сообщения клиентов преобразовывались в векторное представление (эмбединги) с использованием сиамской нейронной сети, предварительно обученной на задаче сопоставления вопросов и ответов. Использовались представления из последнего полносвязного слоя нейронной сети в качестве эмбедингов сообщений. Полученные представления сохраняют семантическую структуру сообщений, и сообщения с семантически близким содержанием имеют меньшее расстояние в признаковом пространстве. В качестве классификатора применялась мультиномиальная логистическая регрессия с l2-регуляризацией, которая показала более высокое качество классификации на перекрестной проверке по сравнению с рассмотренными альтернативными подходами.

Обучающая выборка содержит 150 намерений и 9000 сообщений. Ниже в таблице приведены примеры репрезентативных и легко разделяемых данных из обучающей выборки:

Таблица

Примеры сообщений с пользовательскими намерениями

Текст пользовательского сообщения	Намерение
Привет. Хочу открыть у вас дебетовый вклад	Открытие вклада
Как я могу получить карту банка на дом?	Доставка карты
Украла карту, как можно ее деактивировать	Заблокировать карту
Какое время работы ближайшего банкомата?	Запрос информации о банкоматах
Хочу узнать, как изменить лимит по карте	Изменить лимит по карте
Нужно сменить лимит на карточке, как это сделать?	Изменить лимит по карте
При попытке совершить покупку с кредитной картой снимается комиссия. Как ее убрать?	Убрать комиссию
Последние несколько дней я столкнулся с проблемой с вашим мобильным приложением. Все переводы проходят медленно, и время от времени мне приходится ждать смс-подтверждения. Что мне делать?	Проблема с мобильным приложением

В исходной обучающей выборке было в среднем 100 примеров на намерение, варьирующихся от 10 до 600. Малый размер выборки также усложнял задачу классификации. В экспериментах сравнивались различные подходы: предложенный гибридный алгоритм, случайный выбор, выбор объектов с максимальной неуверенностью модели и энтропийной классификацией, выбор с взвешенной плотностью и энтропией, а также отбор на основе кластеризации с использованием DBSCAN. Оценка точности проводилась на отложенной выборке из 6000 сообщений при обучающей выборке в 5000 объектов (запросов к оракулу).

Отложенная выборка состояла из реальных клиентских сообщений, аналогичных контрольной выборке. Множества намерений для контрольной и обучающей выборок были идентичны. Из 1 000 000 неразмеченных сообщений отобраны размеченные примеры. В некоторых образцах отсутствовали намерения, например, «1234» или «Александр Александрович». Определенные намерения, такие как «Хочу заблокировать кредит» или «Могу ли передать долг сыну?», не были включены в подмножество из 150 размеченных образцов. В роли оракула выступали эксперты с глубокими знаниями в предметной области. Выбранные примеры были размечены и проверены для обеспечения качества аннотации.

Методы отбора по неуверенности модели и взвешенный отбор по неуверенности были использованы как алгоритмы, обеспечивающие высокое качество в большинстве задач активного обучения. Однако в эксперименте из 5000 выбранных объектов ни одно сообщение, содержащее намерение из обучающей выборки, не было обнаружено с использованием данных подходов. В результате качество алгоритма оставалось неизменным, и в обучающую выборку не добавлялось ни одного сообщения. Алгоритмы отбирали сообщения с равномерным распределением уверенности моде-

ли в классификации. Такие сообщения представляли собой либо редкие шумы, либо большие скопления сообщений без целевых намерений (например, сообщения на иностранном языке или кластер сообщений «Петров Петр Петрович»). В данной задаче классические алгоритмы активного обучения демонстрируют эффективность ниже, чем случайное сэмплирование. По этой причине они не представлены в итоговом сравнении.

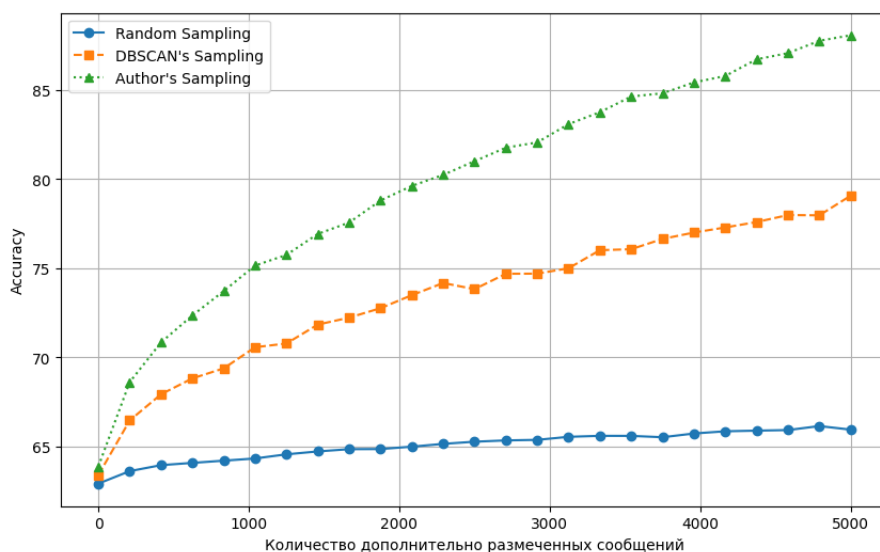


Рис. 3. Результаты сопоставления с альтернативными методами в ходе эксперимента

В качестве одного из алгоритмов для сравнения был выбран диапазон объектов с использованием сэмплирования после кластеризации, поскольку он учитывает разнообразие и плотность данных объектов, являясь хорошим индустриальным базовым решением. Были также рассмотрены альтернативные подходы для учета семантического распределения сообщений, включая тематическое моделирование. Однако экспертная оценка разделения объектов и точность работы по индексу Adjusted Rand оказались выше при использовании кластеризации алгоритмом DBSCAN. Данный алгоритм был включен в анализ.

Изучение задач активного обучения в контексте аналогичных приложений требует значительных инвестиций времени и финансов. В будущих работах автор планирует исследовать применимость адаптированных наборов данных для сопоставления разнообразных подходов с предложенным алгоритмом в разных областях применения и сценариях активного обучения.

Выводы

Автор предлагает гибридный алгоритм итеративного активного обучения, который объединяет в себе использование кластеризации для учета структуры данных и

оценку энтропии неразмеченных объектов, близких к существующей выборке. Этот алгоритм обеспечивает эффективное применение активного обучения для классификации подмножества классов с шумными данными.

Представленный алгоритм обладает следующими преимуществами: независимость от модели классификации, способность работать с шумными данными, эффективность при обработке несбалансированных выборок и возможность использования при неизвестном количестве классов.

Результаты тестирования демонстрируют его эффективность в задаче классификации намерений для диалогового банковского агента на основе эмпирических данных, обеспечивая высокое качество решения практической задачи.

Список источников

- Demir E., et al. Unsupervised Active Learning For Video Annotation // ICML Active Learning Workshop. 2015. 2015.
- Gal Y., Islam R., Ghahramani Z. Deep Bayesian active learning with image data. International conference on machine learning // PMLR. 2017. P. 1183–1192.
- Hoi S. C. H., Jin R., Lyu M. R. Large-scale text categorization by batch mode active learning. Proceedings of the 15th international conference on World Wide Web. 2006. P. 633–642.
- Lewis D. D. A sequential algorithm for training text classifiers: Corrigendum and additional data // Sigir Forum. 1995. Vol. 29, No. 2. P. 13–19.
- Liu B., Lane I. Attention-based recurrent neural network models for joint intent detection and slot filling. 2016. P. 685–689. DOI:10.21437/Interspeech.2016-1352.
- Perez F., Lebre R., Aberer K. Weakly supervised active learning with cluster annotation, 2018.
- Ren F., Xue S. Intention detection based on Siamese neural network with triplet loss. IEEE Access. 2020. Vol. 8. P. 82242–82254.
- Shen Y., et al. Enhancing the generalization for Intent Classification and Out-of-Domain Detection in SLU, 2021.
- Souza V., et al. Unsupervised active learning techniques for labeling training sets: an experimental evaluation on sequential data. Intelligent Data Analysis. 2017. Vol. 21. No. 5. P. 1061–1095.
- Tong S., Koller D. Support vector machine active learning with applications to text classification. Journal of machine learning research. 2001. Vol. 2. November. P. 45–66.
- Wang T., Chen S., Jia R. One-round active learning. arXiv preprint arXiv:2104.11843. 2021.
- Zhdanov F. Diverse mini-batch active learning. arXiv preprint arXiv:1901.05954. 2019.
- Zhou, Meng, Zechen Li, and Pengtao Xie. Self-supervised regularization for text classification // Transactions of the Association for Computational Linguistics. 2021. No. 9. P. 641–656.

References

- Demir E., et al. Unsupervised active learning for video annotation. *ICML Active Learning Workshop 2015*, 2015.
- Gal Y., Islam R., Ghahramani Z. Deep Bayesian active learning with image data. International conference on machine learning. *PMLR*, 2017, pp. 1183–1192.
- Hoi S. C. H., Jin R., Lyu M. R. Large-scale text categorization by batch mode active learning. *Proceedings of the 15th international conference on World Wide Web*, 2006, pp. 633–642.

- Lewis D. D. A sequential algorithm for training text classifiers: Corrigendum and additional data. *Sigir Forum*, Vol. 29, no. 2, Fall 1995, pp. 13–19.
- Liu B., Lane I. *Attention-based recurrent neural network models for joint intent detection and slot filling*, 2016, pp. 685–689. DOI:10.21437/Interspeech.2016-1352.
- Perez F., Lebre R., Aberer K. *Weakly supervised active learning with cluster annotation*, 2018.
- Ren F., Xue S. Intention detection based on Siamese neural network with triplet loss. *IEEE Access*, 2020, Vol. 8, pp. 82242–82254.
- Shen Y., et al. *Enhancing the generalization for Intent Classification and Out-of-Domain Detection in SLU*, 2021.
- Souza V., et al. Unsupervised active learning techniques for labeling training sets: an experimental evaluation on sequential data. *Intelligent Data Analysis*, 2017, Vol. 21, no. 5, pp. 1061–1095.
- Tong S., Koller D. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2001, Vol. 2, November, pp. 45–66.
- Wang T., Chen S., Jia R. *One-round active learning*. 2021.
- Zhdanov F. *Diverse mini-batch active learning*. 2019.
- Zhou Meng, Zechen Li and Pengtao Xie. Self-supervised regularization for text classification. *Transactions of the Association for Computational Linguistics*, 2021, no. 9, pp. 641–656.

Сведения об авторах

Даниил Игоревич Дранга – аспирант; <https://orcid.org/0000-0002-3929-8161>, dranga.daniil@gmail.com, Национальный исследовательский технологический университет МИСиС (д. 4, Ленинский проспект, 119049 Москва, Россия); **Daniil I. Dranga** – Postgraduate student; <https://orcid.org/0000-0002-3929-8161>, dranga.daniil@gmail.com, National University of Science and Technology MISiS (4, Leninskii pr., 119049 Moscow, Russia).

Статья поступила в редакцию 11.05.2023; одобрена после рецензирования 26.06.2023; принята к публикации 31.07.2023.

The article was submitted 11.05.2023; Approved after reviewing 26.06.2023; Accepted for publication 23.07.2023.