Способ автоматизированного формирования обучающего набора данных для алгоритмов машинного обучения классификации электронных документов

И.Д. Королев, Д.В. Акинфиев

Краснодарское высшее военное училище имени генерала армии С.М. Штеменко

Аннотация: В статье рассмотрен способ автоматизированного формирования обучающего набора данных для алгоритмов машинного обучения классификации электронных документов, отличающийся от известных формированием обучающих наборов данных на основе синтеза методов кластеризации и аугментации данных на основе расчета расстояния между объектами в многомерных пространствах.
Ключевые слова: обучение с учителем, кластеризация, распознавание образов, алгоритм машинного обучения, электронный документ, векторизация, формализованные документы.

В случае обучения с учителем данные, состоящие из входной и выходной информации, размеченной экспертами, анализируются алгоритмом машинного обучения, при этом цель обучения заключается в определении алгоритмом машинного обучения общего правила для определения соответствия между входной и выходной информацией. Важный аспект для обучения с учителем заключается в подготовке для алгоритма машинного обучения большого количества качественных обучающих наборов данных для улучшения прогнозирующей способности [1].

Для алгоритмов машинного обучения, реализующих нейронные сети и алгоритмы глубинного обучения, требуется большое количество наборов данных на этапе обучения с учителем [2-3]. Подход, заключающейся в использовании размеченных экспертами наборов данных, вследствие огромного количества необходимых данных оказывается трудоемким и дорогостоящим.

Функционирование процесса формирования обучающих наборов данных для алгоритмов машинного обучения классификации электронных документов при обучении с учителем происходит следующим образом: после

определения задач и объекта для которого разрабатывается система, реализующая алгоритмы машинного обучения, формируется корпус неразмеченных электронных документов, признаки которых на этом этапе явно неопределенны.

$$D = \{d_1, d_2, ..., d_i\}, i = \overline{1,n} \quad (1)$$

где n – количество документов в корпусе.

Происходит определение классов, на которые будет производиться разметка электронных документов, характеристики данных классов.

$$C = \{c_1, c_2, ..., c_j\}, j = \overline{1,m} \quad (2)$$

где m – количество классов.

На подбор экспертов накладываются определенные условия: наличие и занятость личного состава, компетентность экспертов в областях, на которые необходимо разметить документы. Если разрабатываемая система предполагает классификацию электронных документов для всех структурных подразделений, организации необходимо от каждого из них выделение личного состава, что связано с определенными трудностями.

Оперативность процесса формирования обучающего набора данных можно представить следующим образом:

$$Q_{фон} = \frac{1}{T_{нк} + T_{пэ} + T_{окл} + T_{рв}} \quad (3)$$

где $T_{нк}$ - время, затрачиваемое на формирование неразмеченного корпуса, $T_{пэ}$ - время, затрачиваемое выделение и подбор экспертов, $T_{окл}$ - время затрачиваемое на определение классов, $T_{рв}$ - время, затрачиваемое на разметку обучающего набора данных.

Для повышения оперативности формирования обучающего набора данных предлагается минимизировать время, затрачиваемое на выделение и

подбор экспертов и на разметку обучающего набора данных, за счет автоматизации данного процесса.

Для решения данной задачи необходимо оценить характер данных неразмеченного корпуса электронных документов. С этой целью был взят корпус электронный документов. Размер корпуса составил 3610 электронных документов. Для визуализации данных использовался алгоритм стохастического вложения соседей с t-распределением (t-SNE) [4]. Это метод нелинейного уменьшения размерности для вложения многомерных данных в пространство низкой размерности. В частности, он моделирует каждый многомерный объект с помощью двух- или трехмерной точки таким образом, что похожие объекты моделируются близлежащими точками, а непохожие объекты моделируются удаленными точками с высокой вероятностью. Алгоритм был реализован на языке программирования python 3.9 с использованием свободно распространяемой библиотеки sklearn. Перед началом работы алгоритма корпус электронных документов был векторизован методом TF-IDF [5].

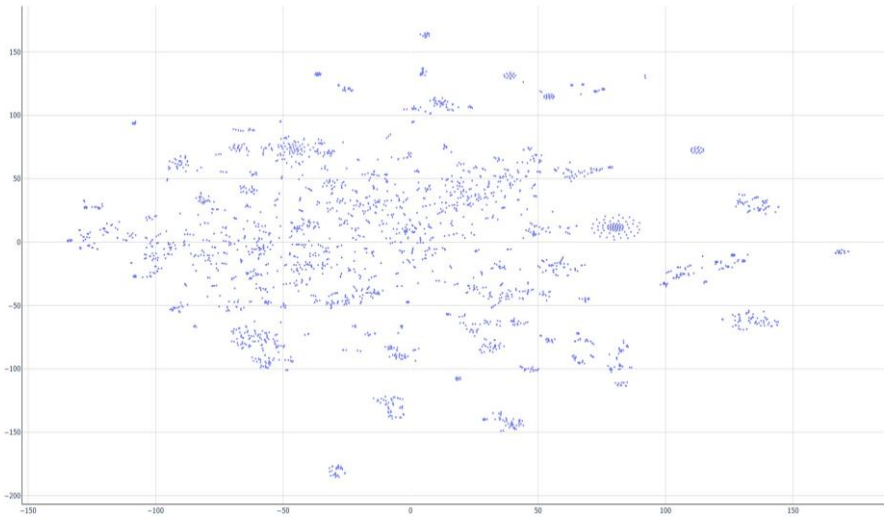

Рис. 1. – Результат работы алгоритма t-SNE для корпуса электронных документов, векторизованных методом TF-IDF

Как видно из рисунка 1, электронные документы собираются в сгущения с явно выраженным центром, что свидетельствует о возможности применения алгоритмов кластеризации, основанных на расчете метрик расстояния [6]. Выбор алгоритма должен исходить из условия, что количество кластеров известно. Вместе с тем известно и содержание кластеров, потому что существует необходимость разметки электронных документов по структурным подразделениям организации. Поэтому необходимо использование алгоритмов машинного обучения, работающих по схеме обучения с подкреплением. Для этого необходимо выделение для каждого структурного подразделения организации как минимум одного электронного документа, отражающего специфику деятельности данного структурного элемента.

На рисунке 2 представлена структурно-функциональная модель способа автоматизированного формирования обучающего набора данных для алгоритмов машинного обучения классификации электронных документов.

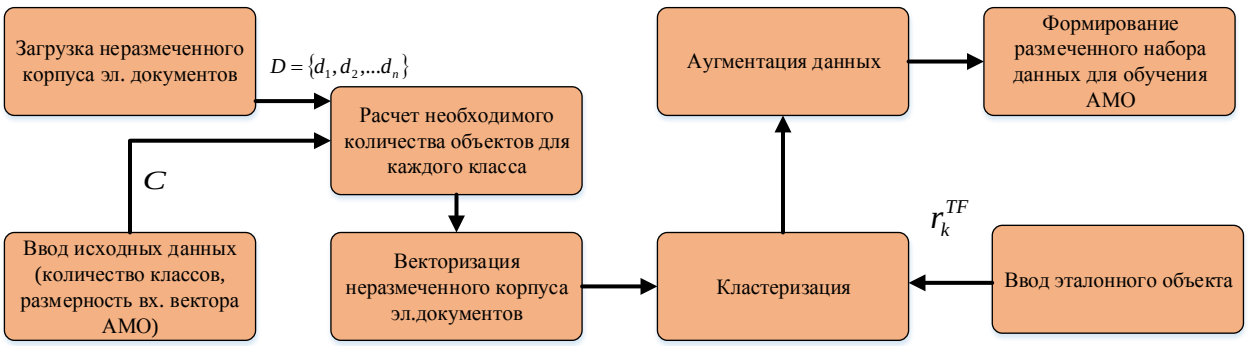

Рис. 2. – Структурно-функциональная модель способа автоматизированного формирования обучающего набора данных для алгоритмов машинного обучения классификации электронных документов.

Способ состоит из следующих этапов:

1 этап – формирование и загрузка неразмеченного корпуса электронных документов;

2 этап – ввод исходных данных;

3 этап – расчет необходимого количества объектов выборки для каждого структурного подразделения организации;

4 этап – векторизация неразмеченного корпуса электронных документов;

5 этап – ввод эталонных объектов для каждого структурного подразделения организации;

6 этап – кластеризация неразмеченного корпуса электронных документов;

7 этап – аугментация объектов обучающей выборки;

8 этап – формирование размеченного набора данных для обучения алгоритмов машинного обучения.

На этапе формирования и загрузки неразмеченного корпуса электронных документов необходимо сформировать корпус электронных документов, в который будут входить электронные документы, характерные для каждого структурного подразделения организации.

На этапе ввода исходных данных в систему загружаются количество алгоритмов машинного обучения, исходные векторы для каждого структурного подразделения организации. Исходные данные получается в процессе работы методики формирования и определения параметров алгоритмов машинного обучения классификации электронных документов по значимости информации для должностных лиц организаций [7].

На следующем этапе производится расчет необходимого количества объектов для каждого структурного подразделения организации по правилу:

$$N_{c_j} = (\left|R_{c_j}\right| * 2 + 0.1 * \left|R_{c_j}\right| * 2) \quad (4)$$

где $\left|R_{c_j}\right|$ - мощность множества ключевых слов, отражающих выполняемые задачи и функции структурного подразделения организации.

На 4 этапе производится векторизация неразмеченного корпуса электронных документов методом TF-IDF. Исходными векторами для кластеризации будут вектора структурных подразделений организации.

На 5 этапе в систему вводятся эталонные объекты для каждого структурного подразделения организации.

Кластеризация неразмеченных объектов производится с помощью алгоритма FOREL, с первичным заданием центра масс, за счет введения эталонного объекта структурного подразделения организации [8]. Введение эталонных объектов позволит ускорить процесс кластеризации, а также улучшит качество кластеризации [9].

Кластеризация будет проводиться на два кластера. Кластер «1» - объект принадлежит структурному подразделения организации, кластер «0» - объект не принадлежит структурному подразделению организации. Расстояние между объектами, для отнесения объекта к кластеру будет рассчитываться по правилу:

$$d(p,q) = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2} < R \qquad (5)$$

где p - центр кластера, q - объект набора данных, R - радиус отнесения объектов к кластеру.

Кластеризация производится только для кластера «1». Все объекты, не попавшие в кластер «1», автоматически помечаются как кластер «0». Данное изменение алгоритма FOREL связано с несбалансированностью неразмеченного корпуса объектов (к различным структурным подразделениям организации будет относится различное количество объектов). Параметр R кластеризации задается оператором. После отнесения всех объектов к кластерам, производится перерасчет центров масс. Для этого рассчитывается среднее арифметическое координат векторов для кластера

«1». Полученное значение задается, как центр масс. Затем происходит повторное отнесение объектов к кластерам, до тех пор, пока центр масс кластера «1» не перестанет изменяться. Кластеризация производится для каждого структурного подразделения организации только для класса «1». Формирование объектов класса «0» производится после подсчета достаточности размера выборки.

После кластеризации неразмеченного набора данных, производится подсчет достаточности размера выборки по правилу (4). В случае недостаточности объектов проводится аугментация данных [10]. Под аугментацией понимается искусственное создание объектов, имеющих признаки, схожие с реальными объектами.

Для выполнения аугментации выбирается случайный объект класса «1» и центр масс кластера. Координаты нового объекта тогда будут:

$$q_{но}(q_1^{цм} - \frac{q_1^{цм} - q_1^{сов}}{2}, q_2^{цм} - \frac{q_2^{цм} - q_2^{сов}}{2}, ..., q_n^{цм} - \frac{q_n^{цм} - q_n^{сов}}{2}) \quad (6)$$

где $q_{но}$ - новый объект выборки, $q^{цм}$ - центр масс кластера «1», $q^{сов}$ - случайный объект кластера «1».

После кластеризации параметр R отчерчивает вокруг центра масс сферу в многомерном пространстве для кластера «1». Каждый объект, находящийся внутри сферы, попадает в кластер «1». Соответственно, каждая точка этой сферы будет принадлежать кластеру. Новый объект размещается на середине отрезка между случайным объектом и центром масс кластера. Размерность пространства набора данных, как правило, не менее стомерного, причем значения признаков для векторизованного электронного документа не являются целыми числами. Соответственно, вероятность того, что новый объект попадет в существующий, будет стремиться к 0.

После аугментации формируется кластер «0» для каждого структурного подразделения организации следующим порядком: необходимо

выбрать объекты в количестве равном количеству объектов кластер «1», для этого из каждого класса «1» структурных подразделений организации отбираются случайным образом равное количество объектов. Данное положение объясняется тем, что для качественного обучения алгоритма машинного обучения классификации электронных документов по значимости информации для должностных лиц структурных подразделений организации необходима сбалансированная выборка, где количество объектов для каждого класса равно.

Выходными значениями способа будет размеченный набор данных для обучения алгоритмов машинного обучения классификации электронных документов по значимости информации для должностных лиц организации $(X_r^{TFIDF}, Y)^l$. В набор включаются все объекты кластера «1», и такое же количество объектов из кластеров «1» других структурных подразделений организации. Выбор объектов из других структурных подразделений организации происходит случайным образом, так формируется кластер «0». Тем самым задаётся сбалансированный набор размеченных данных для обучения алгоритма машинного обучения.

Размеченные наборы будут представлять собой массивы данных с размерностью, равной количеству объектов кластеров «0» и «1» и размерностью, равной размерности исходного вектора структурного подразделения организации.

Литература

1. Мандель И. Д. Кластерный анализ. — М.: Финансы и Статистика, 1988 – 176 стр.

2. Головко В. А. Нейронные сети: обучение, организация и применение. — М.: ИПРЖР, 2001 – 257 стр.

3. Федутинов, К. А. Машинное обучение в задачах поддержки принятия решений при управлении охраной природы // Инженерный вестник Дона. – 2021. – № 9. – URL: ivdon.ru/ru/magazine/archive/n9y2021/7186.

4. Van der Maaten L.J.P., Hinton G.E. Visualizing Data Using t-SNE // Journal of Machine Learning Research. – 2008. P. 2579-2605.

5. Jones K. S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation — MCB University: MCB University Press, 2004. — Vol. 60, no. 5. — P. 493—502.

6. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов (статистические проблемы обучения) – Москва: Издательство «Наука», Главная редакция физико-математической литературы – 1974 – 416 стр.

7. Королев И.Д., Акинфиев Д.В. Методика формирования и определения параметров алгоритмов машинного обучения классификации электронных документов по значимости информации для должностных лиц организаций // Инженерный вестник Дона. – 2023. – № 9. – URL: ivdon.ru/ru/magazine/archive/n9y2023/8675.

8. Загоруйко Н. Г., Ёлкина В. Н., Лбов Г. С. Алгоритмы обнаружения эмпирических закономерностей. – Новосибирск: Наука – 1985. – 110 стр.

9. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989. – 608 стр.

10. Simard P. Y., Steinkraus D., Platt J. C. Best practices for convolutional neural networks applied to visual document analysis // Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. — 2003. — Aug. — pp. 958–963.

References

1. Mandel' I. D. Klasternyj analiz [Cluster analysis]. Moskva, Finansy i Statistika, 1988. 176 p.

2. Golovko V. A. Nejronnye seti: obuchenie, organizacija i primenenie [Neural networks: training, organization and application]. Moskva. IPRZhR, 2001. 257 p.

3. Fedutinov, K. A. Inzhenernyj vestnik Dona, 2021, № 9. URL: ivdon.ru/ru/magazine/archive/n9y2021/7186.

4. Van der Maaten L.J.P., Hinton G.E. Journal of Machine Learning Research, 2008. pp. 2579-2605.

5. Jones K. S. A Journal of Documentation, MCB University, MCB University Press, 2004. Vol. 60, no. 5. pp. 493-502.

6. Vapnik V.N., Chervonenkis A.Ja. Teorija raspoznavanija obrazov (statisticheskie problemy obuchenija) [Pattern recognition theory (statistical learning problems)]. Moskva, «Nauka, 1974. 416 p.

7. Korolev I.D., Akinfiev D.V. Inzhenernyj vestnik Dona, 2023, № 9. URL: ivdon.ru/ru/magazine/archive/n9y2023/8675.

8. Zagorujko N. G., Yolkina V. N., Lbov G. S. Algoritmy obnaruzheniya empiricheskih zakonomernostej [Algorithms for detecting empirical patterns]. Novosibirsk, Nauka, 1985. 110 p.

9. Ajvazjan S. A., Buhshtaber V. M., Enjukov I. S., Meshalkin L. D. Prikladnaja statistika: klassifikacija i snizhenie razmernosti [Applied statistics: classification and dimensionality reduction]. Moskva. Finansy i statistika, 1989. 608 p.

10. Simard P. Y., Steinkraus D., Platt J. C. Seventh International Conference on Document Analysis and Recognition, 2003, Proceedings, 2003, pp. 958-963.