Методика формирования и определения параметров алгоритмов машинного обучения классификации электронных документов по значимости информации для должностных лиц организаций

И.Д. Королев, Д.В. Акинфиев

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

Аннотация: В статье рассмотрена методика формирования и определения параметров алгоритмов машинного обучения классификации электронных документов по значимости информации для должностных лиц организаций отличающаяся от известных динамическим формированием структуры и количества алгоритмов машинного обучения, за счет автоматизированного определения множеств структурных подразделений организации, множеств ключевых слов, отражающих задачи и функции структурных подразделений в процессе автоматизированного анализа Положения об организации, Положений структурных подразделений на основе теории распознавания образов.
Ключевые слова: лемматизация, распознавание образов, алгоритм машинного обучения, электронный документ, векторизация, формализованные документы.

В ходе расширенного заседания коллегии Министерства обороны Российской Федерации 21 декабря 2022 года, Верховный Главнокомандующий В.В. Путин акцентировал внимание на необходимости совершенствования системы управления и связи в целях обеспечения устойчивости и оперативности управления войсками в любых условиях, для чего требуется активнее использовать технологии искусственного интеллекта на всех уровнях принятия решений.

В настоящее время существует проблема оперативной и качественной обработки поступивших электронных документов в автоматизированные системы военного назначения.

Усовершенствование компьютерной техники и технологий в сочетании с увеличением роста подключенных электронных устройств, необходимостью обработки больших, слабо формализованных данных, привело к росту заинтересованности в разработке систем искусственного интеллекта и решений для автоматизации выполнения задач управления в условиях ограниченного времени для принятие решения [1,2]. Данный факт

привел к необходимости повышения оперативности обработки электронных документов на основе методов классификации документов, предполагающих применение алгоритмов машинного обучения. Машинное обучение, тесно связанное с интеллектуальным анализом данных, вычислительной статистикой и оптимизацией, имеет дело с изучением алгоритмов, способных обучаться и выполнять прогнозирование на основе данных.

Алгоритмы машинного обучения можно разделить на большие категории, такие, как обучение с учителем, обучение без учителя, обучение с подкреплением. В случае обучения с учителем, обучающие 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.

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

Вместе с тем, при разработке систем, реализующих алгоритмы машинного обучения, необходимо подбирать параметры работы алгоритма для каждого объекта функционирования и решаемых задач. Каждая организация Министерства обороны Российской Федерации (далее – организация), с точки зрения организационно-штатной структуры и

выполняемых задач, является уникальным объектом, вследствие чего возникает проблема подбора и обоснования параметров работы систем, реализующих алгоритмы машинного обучения, что сказывается на трудоемкости и временных затратах при проведении пуско-наладочных работ.

Методика формирования и определения параметров алгоритмов машинного обучения классификации электронных документов по значимости информации для должностных лиц организаций может применяться в автоматизированных системах организаций, осуществляющих обработку, передачу и хранение формализованных электронных документов.

Цель методики - повышение оперативности обработки электронных документов, поступающих в автоматизированные системы организаций в условиях ограниченного для принятия решения времени.

При построении алгоритма машинного обучения необходимо учитывать тот факт, что большинство из них осуществляют бинарную классификацию [3]. При применении подобных алгоритмов, для решения задачи распределения электронных документов необходимо строить алгоритм для каждого должностного лица организации, результатом работы которого будет решения об отнесения электронного документа к области деятельности должностного лица, а соответственно - степени значимости информации, содержащейся в документе. Формализуем среду алгоритмов машинного обучения.

Функционирование процесса обработки электронных документов в автоматизированных системах организаций происходит следующим образом:

после учета поступившего электронного документа, оператор выполняет классификацию и формирует решение о распределении документа на основе знания обязанностей должностных лиц организаций и функций, выполняемых структурными подразделениями. С точки зрения теории

исследования операций, оператор решает многокритериальную задачу, на основе введения критериев важности признаков, содержащихся в документе [4]. Под признаками понимаются ключевые слова, отражающие содержание документа. Оперативность и качество принятия решения зависят от компетенций и опыта оператора. В то же время каждая организация имеет Положение об организации, характеризующееся множеством:

$$O = \{R_1, R_2, ..., R_i\}, i = \overline{1,n}, \quad (1)$$

где R_i – i-ое структурное подразделение; n – общее количество структурных подразделений.

В свою очередь, каждое структурное подразделение характеризуется множеством ключевых слов, отражающих выполняемые задачи и функции:

$$R_i = \{x_1, x_2, ..., x_j\}, \ j = \overline{1,m}, \quad (2)$$

где x_j – j-ое ключевое слово; m – общее количество ключевых слов;

Под ключевым словом понимается слово или словосочетание, характерное для определенного структурного подразделения [5]. Множество ключевых слов выбирается, исходя из задач и функций, которые описаны в Положении о структурном подразделении и в Положении об организации.

Для работы алгоритма машинного обучения необходимо векторизовать множество ключевых слов для каждого структурного подразделения по выборке электронных документов данного подразделения. Векторизацию множества предлагается производить методом TF-IDF [6]:

$$TF_j = \frac{n_j}{N_j}, \quad (3)$$

$$IDF_j = \log \frac{p}{P}, \quad (4)$$

$$TFIDF = TF * IDF, \quad (5)$$

где n_j – сколько раз встречается ключевое слово в j-ом документе; N_j – общее количество слов в документе; p – количество документов выборки, в которых встречается ключевое слово; P – общее количество документов выборки.

При определении множества ключевых слов для структурного подразделения необходимо провести предварительную обработку текса, заключающуюся в откидывании стоп-слов (предлогов, союзов, местоимений и т.п.) и лемматизации ключевых слов [7]. Цель предварительной обработки - снижение размерности входного пространства признаков алгоритма машинного обучения.

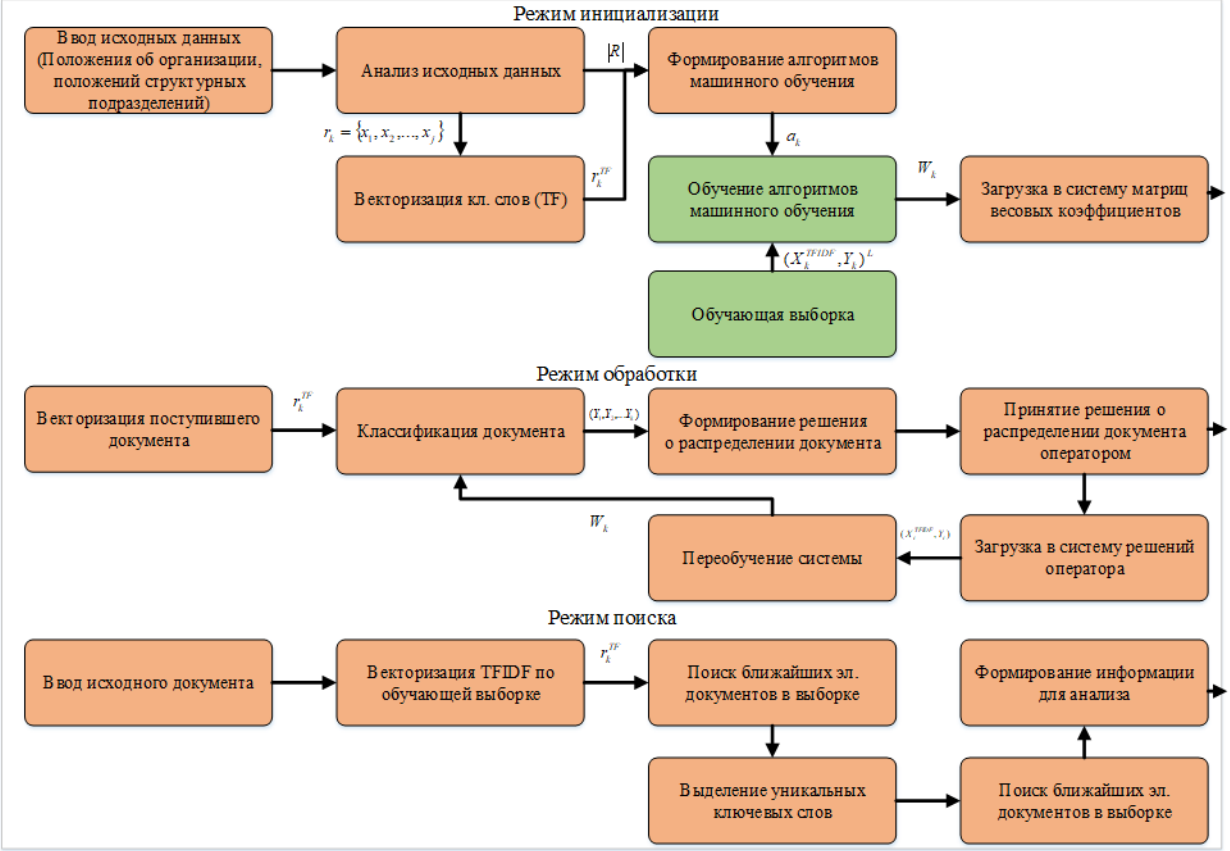

Рис. 1 – Структурно-функциональная модель методики формирования и определения параметров алгоритмов машинного обучения классификации электронных документов по значимости информации для должностных лиц организации.

На рисунке 1 представлена структурно-функциональная модель методики формирования и определения параметров алгоритмов машинного обучения классификации электронных документов по значимости информации для должностных лиц организации.

Применение методики предлагается производить в трех режимах: инициализации, в режиме обработки и режиме поиска.

В режиме инициализации, после ввода исходных данных (Положение об организации, Положения о структурных подразделениях), происходит их анализ, в ходе которого определяются параметры алгоритмов машинного обучения и их количество [8].

Мощность множества O будет отражать количество алгоритмов машинного обучения, которые необходимо построить для решения задачи классификации электронных документов по значимости информации для должностных лиц организации. В свою очередь, мощность множества R_i, будет отражать размерность пространства входных признаков для алгоритма машинного обучения.

В качестве алгоритма машинного обучения предлагается использовать двухслойную нейронную сеть прямого распространения с сигмоидной функцией активации, так как, в соответствии с теоремой Цыбенко, данная нейронная сеть может аппроксемировать любую непрерывную функцию многих переменных с любой точностью [9].

Тогда правило классификации электронного документа для каждого структурного подразделения выразится следующим образом:

$$U_h(X_k^{TFIDF}) = \sigma(\sum_{i=1}^{h} w_{hi} X_{ki}^{TFIDF}),\ h = \overline{1,|R|} \tag{6}$$

$$V_l = \sigma(\sum_{i=1}^{l} w_{li} U_h(X_{ki}^{TFIDF})),\ l = \overline{1,\frac{|R|}{2}} \tag{7}$$

$$Y_r = \begin{cases} 1, \textit{если } \sigma(V_l) \geq 0{,}7 \\ 0, \textit{если } \sigma(V_l) < 0{,}7 \end{cases} \quad (8)$$

Где $U_h(X_k^{TKIDF})$ – линейная функция 1-го слоя h-го ключевого слова; w_{hi} – весовой коэффициент h-го ключевого слова i-ой линейной функции; X_{ki}^{TFIDF} - признаковое описание h-го ключевого слова k-го электронного документа; σ - сигмоидная функция активации; V_l - линейная функция скрытого слоя; Y_r - функция единичного скачка выходного слоя для r-го структурного подразделения.

После формирования структуры алгоритмов машинного обучения для каждого структурного подразделения необходимо рассчитать весовые коэффициенты w_{hi} по обучающей выборке (X_i^{TFIDF}, Y_i) и построить для каждого алгоритма и для каждого слоя матрицу весовых коэффициентов:

$$W_k \begin{pmatrix} w_{11}^r & w_{12}^r & \ldots & w_h^r \\ w_{21}^r & w_{22}^r & \ldots & w_{2h}^r \\ \ldots & \ldots & \ldots & \ldots \\ w_{i1}^r & w_{i2}^r & \ldots & w_{ih}^r \end{pmatrix} \quad (9)$$

Обучение предлагается проводить методом обратного распространения ошибки [10]. После загрузки в систему матриц весовых коэффициентов методика используется в режиме нормального функционирования.

В режиме обработки происходит загрузка электронного документа, векторизация по правилу (5), классификация по правилам (6), (7), (8). Выходным значением будет вектор $Y(Y_1, Y_2, ..., Y_r)$ отражающий принадлежность документа к структурным подразделениям организации, а значит значимость для них информации, содержащейся в документе. После принятия решения оператором о распределении электронного документа, данное решение и векторизованный документ добавляются в обучающую выборку и происходит переобучение алгоритмов машинного обучения.

Новые матрицы весовых коэффициентов загружаются добавляются в алгоритмы машинного обучения и продолжается работа методики.

В режиме поиска происходит загрузка электронного документа, векторизация по правилу (5), поиск наиболее близких электронный документов, используя расчет евклидового расстояния. Оператором задается критерий поиска – количество электронных документов. Затем исключается формализованная часть, которая является пресечением множеств ключевых слов найденных электронных документов. Производится повторный поиск близких электронных документов (с минимальным евклидовым расстоянием).

Научная новизна заключается в том, что методика формирования и определения параметров алгоритмов машинного обучения классификации электронных документов по значимости информации для должностных лиц организаций отличается от известных динамическим формированием структуры и количества алгоритмов машинного обучения, за счет автоматизированного определения множеств структурных подразделений организации, множеств ключевых слов, отражающих задачи и функции структурных подразделений в процессе автоматизированного анализа Положения об организации, Положений структурных подразделений на основе теории распознавания образов.

Литература

1. Горлатов, Д. В. Машинное обучение прогнозных моделей на несбалансировнных данных по опасным астероидам // Инженерный вестник Дона. – 2023. – № 5. – URL: ivdon.ru/ru/magazine/archive/n5y2023/8394.

2. 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.

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

4. Вентцель Е.С. Исследование операций. – Москва: «Советское радио». – 1972. – 552 стр.

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

6. 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. — pp. 493—502.

7. 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.

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

9. Cybenko G.V. Approximation by Superpositions of a Sigmoidal function // Mathematics of Control Signals and Systems. – 1989. – T.2, №4. – pp. 303-314.

10. Rummelhart D. E., Hinton G. E., Williams R. J. Learning internal representations by error propagation // Vol. 1 of Computational models of cognition and perception, chap. 8. — Cambridge, MA: MIT Press, 1986. — Pp. 319–362.

References

1. Gorlatov, D. V. Inzhenernyj vestnik Dona, 2023, № 5. URL: ivdon.ru/ru/magazine/archive/n5y2023/8394.

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

3. Vapnik V.N., Chervonenkis A.Ja. Teorija raspoznavanija obrazov (statisticheskie problemy obuchenija) [Pattern recognition theory (statistical learning problems)]. Moskva, Proc. «Nauka», The main editorial office of the physical and mathematical literature, 1974. 416 p.

4. Ventcel' E.S. Issledovanie operacij [Operations Research]. Moskva, «Sovetskoe radio», 1972. 552 p.

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

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

7. 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.

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

9. Cybenko G.V. Mathematics of Control Signals and Systems, 1989. T.2, №4. Pp. 303-314.

10. Rummelhart D. E., Hinton G. E., Williams R. J. Vol. 1 of Computational models of cognition and perception, chap. 8. Cambridge, MA. MIT Press, 1986. Pp. 319–362.