

УДК 004.021

Б. І. ГОНЧАРЕНКО, студент;

О. Ю. ІЛЬІН, доктор техн. наук, професор;

А. Б. КОБА;

О. В. НЕГОДЕНКО, канд. техн. наук,

Державний університет телекомунікацій, Київ

УДОСКОНАЛЕННЯ МЕДИЧНОЇ ІНФОРМАЦІЙНОЇ СИСТЕМИ ДЛЯ МЕДИЧНИХ ЗАКЛАДІВ ЗА ДОПОМОГОЮ MACHINE LEARNING

Розглянуто використання машинного навчання для покращення та вдосконалення роботи медичних інформаційних систем України.

Більшість технічних вирішень медичних систем спрямовано на вдосконалення внутрішньої роботи закладу, але недостатньо приділено уваги взаємодії з пацієнтами. Запропоновано засоби машинного навчання для максимального спрощення запису клієнтом до медичного закладу.

Сформульовано задачу виявлення оптимального алгоритму для добору медичного спеціаліста, ґрунтуючись на симптомах клієнта без участі людини, а також удосконалення роботи такого алгоритму.

Досліджено процес збору та спеціальної підготовки набору даних для проведення машинного навчання машинної моделі. Проведено комплексну роботу з добору необхідних комбінацій даних, що уможливить отримання найкращого прогнозованого результату.

З огляду на особливість набору даних для навчання у процесі проведення дослідження використано підхід «один до багатьох» для багатокласових задач машинного навчання. Виявлено найбільш оптимальний алгоритм машинного навчання з учителем для прогнозування лікаря-спеціаліста, ґрунтуючись на інформацію щодо симптомів зазначеної клієнтом системи під час запису на прийом.

Запропонований метод було вдосконалено через недоліки у вигляді високої вибагливості стосовно використовуваних даних. Зміни роботи алгоритму стосуються оцінювання стану ваги поточного прикладу прогнозу, підвищуючи результативність роботи.

Удосконалений метод ґрунтується на використанні усереднених значень вагових критеріїв замість зберігання всієї колекції критеріїв ваги, що пришвидчує роботу алгоритму, а також робить його менш вибагливим до набору даних, використовуваних під час навчання моделі.

Ключові слова: медичні інформаційні системи; machine learning; perceptron; машинне навчання; алгоритм; метод; перцептрон з голосуванням.

ВСТУП

Медицина є однією з найважливіших сфер у кожній країні світу, зокрема і в Україні. Однією з можливостей підвищення рівня якості медицини є впровадження медичних інформаційних систем у медичні заклади. Хоча перші медичні системи було створено досить давно і останнім часом вони активно розвиваються, їм приділено не досить багато уваги. Більшість систем переходять до стану технологічного старіння, а нові ще не перейняли кращі практики функціонування.

Застосування машинного навчання в медичних інформаційних системах України до цього часу є доволі новою практикою і зазвичай для вдосконалення своєї роботи використовують напрацювання іноземних аналогів. Головною ідеєю є створення можливості он-лайн запису на прийом до медичного закладу, де пацієнт перелічує всі симптоми своєї хвороби і на їх основі прогнозується відповідний спеціаліст медичного закладу, який може виконати прийом.

Тому *метою статті* є покращення процесу прогнозування спеціальності лікаря, що ґрунтується на наборі симптомів пацієнта на основі аналізу та вдосконалення алгоритму машинного навчання з учителем.

Дослідження дасть можливість на основі вибраних симптомів клієнтом спрогнозувати найбільш ймовірного спеціаліста, проаналізувати поточний розклад приймальних годин медичного закладу і створити відповідний запис без втручання медичного працівника.

ОСНОВНА ЧАСТИНА

Основною ідеєю вдосконалення системи є перенесення ухвалення певних вирішень із людини на програмно навчену машину, яка матиме змогу дійти до не гірших вирішень. Після розглянутих аналогів медичних систем стає зрозуміло, що більшість вирішень спрямовано на організацію внутрішніх медичних процесів та автоматизацію роботи персоналу. Але перед виконанням стандартних процедур закладу стоїть потреба запису на прийом, коли клієнт ще ніяким чином не пов'язаний із клінікою. У такому разі запис до лікаря може відбуватися двома шляхами: або за допомогою працівника медичного закладу, або силами самого клієнта.

Перший варіант зводить кількість проблем при запису клієнта до мінімуму, але постає проблема персоналу. Клініка має забезпечити персонал, який відповідатиме за запис клієнта.

© Б. І. Гончаренко, О. Ю. Ільїн, А. Б. Коба, О. В. Негоденко, 2019

Другий варіант більш привабливий для клініки і відносно простий для клієнта. Для клієнта не складно заповнити відповідну он-лайн форму на сайті, що часто буває зручнішим, ніж телефонний дзвінок, і тим більш, ніж власне звертання до клініки. При всіх перевагах основою такого підходу є вирішення самого клієнта. У більшості відомих систем із он-лайн записом клієнт має змогу сам вибрати собі лікаря, до якого бажає піти на прийом, забронювати вільне місце в черзі. При цьому не факт, що клієнту потрібен саме вибраний ним лікар для розв'язання його проблеми. Також можливе бронювання більшої кількості вільних місць у черзі безвідповідальними користувачами, спричинюючи порушення в організованій роботі.

Для вирішення цих двох проблем краще за все підійде спеціально навчена машина, яка мала б змогу приймати рішення, не гірші за відповідального працівника медичного закладу, автоматично виявляти складність причини відвідування клієнта медичного закладу, призначати йому відповідного лікаря, записувати клієнта на вільне місце в розкладі вибраного лікаря. Таким чином, клієнтові буде достатньо вказати перелік своїх симптомів і на їх основі будуть прийняті відповідні рішення, які задовольнять потреби і медичного закладу, і клієнта.

Зазначений шаблон поведінки зменшить затрати клініки на оброблення запитів потенціальних пацієнтів, скорочуючи у такий спосіб перелік обов'язків медичного персоналу та унеможливаючи безглузде бронювання вільних місць у розкладі медичного закладу.

Для клієнта така концепція є більш привабливою, оскільки зводить до мінімуму ймовірність неправильного вибору лікаря. Люди не зобов'язані знати всі тонкощі захворювань і їхні симптоми, а тому, дістаючи інформацію з вільних ресурсів щодо симптомів, можуть приймати неправильні рішення.

Таким чином, розроблене вирішення дасть змогу мінімізувати часові та фінансові затрати для клініки, а також підвищити рівень обслуговування клієнтів.

Створення набору даних для моделі навчання

Для проведення будь-якого машинного навчання першим обов'язковим пунктом завжди є добір необхідного набору даних. Більшість уже розроблених алгоритмів машинного навчання теоретично можуть упоратись зі створенням працюючої моделі, але головною причиною невдач є набір даних, якого не вистачає для проведення навчання.

Процес машинного навчання поділяється на чотири основних етапи:

- 1) збір даних;
- 2) підготовка даних;
- 3) побудова моделі;
- 4) аналіз якості моделі.

Усі перелічені етапи виконуються послідовно і у разі невдалого навчання моделі повертаємось до першого пункту — збору даних.

Для задоволення поставленого завдання необхідно створити такий набір даних для навчання, який включатиме в себе перелік симптомів та спеціалістів медичного закладу. Існує майже шістьсот основних симптомів, які описують більшість поширених хвороб і покривають майже весь спектр захворювань у людини. Для подальшої побудови моделі необхідно зважати на всі основні симптоми, які можуть виникнути у людини під час захворювання. Варто зауважити, що напряму лікарі в жодному разі не пов'язані із симптомами, вони лікують безпосередньо хвороби. Але при визначеному наборі симптомів імовірність наявності певних хвороб більша, ніж наявність інших. В окремих випадках комбінація деяких симптомів зводить імовірність майже до нуля стосовно певних хвороб.

Тому в процесі побудови моделі беруть до уваги, що взаємозв'язки між лікарем та симптомами непрямі, але машину необхідно навчати так, щоб вона відшукувала правильні зв'язки, інколи навіть для тих випадків, які немає змоги заздалегідь передбачити.

Після отримання даних необхідно виконати основні процедури підготовки даних:

- видаляємо всі некоректні символи, розділові знаки, спеціальні символи, які не належать до літерних;
- виконуємо токенизацію тексту, розділивши його на окремі слова, наприклад певні скорочення;
- видаляємо нерелевантні слова, які можуть вплинути на якість даних;
- переводимо всі слова до одного нижнього регістру для того, щоб слова, наприклад, за типом «Біль», «БІЛЬ», «біль» вважались одним і тим самим словом, а не різними;
- розглядаємо можливість поєднання слів, написаних із помилками, або таких, що мають альтернативну назву;
- розглядаємо можливість проведення лематизації, тобто приведення різних форм до одного слова.

Після виконання всіх зазначених процедур відібрані дані можна вважати чистими і почати побудову моделі для машинного навчання.

Основним завданням у процесі створення набору даних є правильне зіставлення значень для прогнозування та значень, що виступають як описувальні характеристики.

У цій задачі немає змоги використовувати методи бінарної класифікації, адже розуміємо, що одна спеціалізація лікаря може покривати безліч симптомів, які також можуть перехрещуватися з іншими спеціалізаціями, тому спрощена структура матиме вигляд як «один до багатьох».

Приклад набору даних після підготовки за описаною структурою наведено в табл. 1.

Як уже було зазначено, можливе перехрещення спеціальності з окремими симптомами. Наприклад, візьмемо останній запис із табл. 1. Терапевт дійсно має відношення до симптому «біль у горлі». Цей симптом один із найпоширеніших і дуже часто виникає під час простуди або грипу. Звісно, у разі таких хвороб потрібно відвідувати терапевта. Але можливі ситуації, коли цей симптом виникає й у разі серйозних інфекційних захворювань, а отже, краще звернутись до інфекціоніста. Таким чином, з'являється перехрещення багатьох лікарів із одним симптомом і не завжди зрозуміло, яке співвідношення буде коректним у даній ситуації.

Іншою проблемою є наявність комбінації симптомів. Як правило, хвороби мають не один симптом, а кілька. Тому при ситуації, коли клієнт здійснює запис на прийом он-лайн, то він вибиратиме не один симптом із переліку, а набір симптомів.

Таким чином, машині необхідно аналізувати не одне конкретне значення симптому відповідно до спеціалізації лікаря, а набір симптомів. Згідно з отриманими даними змінимо наповнення набору даних для навчання. Змінений набір даних унаочнює табл. 2.

Для розв'язання описаних проблем було вирішено деталізувати набір даних із метою підвищення точності майбутнього алгоритму. Згідно з табл. 2 на кожну спеціалізацію може припадати деякий набір симптомів, що можуть відрізнятися за своєю кількістю. Також тут наявне перехрещення однакових симптомів, наприклад для терапевта і для невролога є симптом «головний біль». У невролога також є схожий симптом «нудота», який присутній і для гастроентеролога.

Але варто зазначити, що навіть у разі перехрещення певних симптомів, набір усіх інших симптомів загалом описують окремі хвороби. У першому випадку для терапевта перелічені основні симптоми звичайної простуди, а для невролога — основні симптоми, характерні в разі струсу мозку легкого або середнього рівня. Хоча в обох випадках зустрічається симптом головного болю, ми даємо зрозуміти машині, що головний біль у поєднанні з кашлем — це скоріше простуда, а головний біль із запамороченням — швидше проблема, пов'язана з травмою голови. Будуючи модель у такий спосіб, ми збільшуємо ймовірність правильного розрахунку порівняно з варіантом опису «один спеціаліст — один симптом».

Добір та вдосконалення алгоритму для розв'язання поставленого завдання

Описані набори даних насправді виникли в процесі постійного створення готових моделей і подальшого їх коригування відносно отриманих даних.

Оскільки використовується навчання з учителем, а як вхідні дані беруться ознаки, то було використано підхід опису ознаками.

Позначимо через X множину спеціалістів як прецедент предметної області, результат виміру характеристик якої зобразимо так:

$$f: X \rightarrow D_f, \tag{1}$$

де D_f — множина допустимих значень ознаки.

Для бінарної ознаки $D_f = \{0,1\}$.

Першою спробою було використання найбільш поширеного методу бінарної класифікації. Але від такого способу довелось відмовитись, адже результатами прогнозу є лише відповідь із двома показниками.

Таблиця 1

Формат даних для навчання формату текст–текст

Label	Value
Психотерапевт	Агресивність
Психотерапевт	Аменорея
Психотерапевт	Анорексія
Дерматолог	Висип
Дерматолог	Свербіж
Терапевт	Біль у горлі

Таблиця 2

Змінений формат даних для проведення навчання

Label	Value
Терапевт	Кашель, біль у горлі, нежить, чихання, головний біль
Гастроентеролог	Дискомфорт шлунку, біль у животі, нудота
Невролог	Головний біль, нудота, запаморочення, короткочасна втрата пам'яті
Дерматолог	Висип
Дерматолог	Свербіж
Терапевт	Біль в горлі

Якщо будувати модель на кожний окремий випадок пошуку лікаря за симптомами, то такий підхід міг би спрацювати. Але навчена модель має відповідати всьому масиву даних.

Тому наступним вирішенням було використання методів багатокласової класифікації, оскільки варіацій для симптомів більше двох. Як вже було зазначено, відношення спеціаліст–симптом має зв'язок «один до багатьох», що в машинному навчанні відповідає методології One-vs-All [1].

Найкращим варіантом у ситуації, що склалася з сформованим набором даних, буде використання математичної моделі перцептрону з голосуванням.

Перцептрон з голосуванням — це деяка математична модель «ванільного» перцептрону, яка виконується при навчанні з учителем та має свою систему голосів для оцінювання побудованої моделі [2]. Формула перцептрону з голосуванням має вигляд:

$$\gamma = \text{sign} \left(\sum_{k=1}^k c^{(k)} \text{sign}(\omega^{(k)} * x + b^{(k)}) \right). \quad (2)$$

Головою перевагою такого алгоритму є велика точність прогнозованих даних навіть за умови, що набір даних для навчання мав невелику кількість. Але основним недоліком алгоритму є надмірна вибагливість щодо правильності набору даних. За деяких умов алгоритм не міг закінчити своє навчання, оскільки не знаходив остаточного правильного розв'язку.

Наприклад, розглянемо набір даних із 10 000 прикладів. Припустимо, що після перших 100 прикладів перцептрон виявив хороший класифікатор. Тому він переходить до наступних 9 900 прикладів, не вносячи при цьому ніяких оновлень. Він досягає останнього 10 000 прикладу і робить помилку, руйнуючи весь вектор ваги, який добре відпрацював на 99,99% даних [3]. Тобто алгоритм завжди прямує до ідеального результату і якщо наданий йому набір даних не є доречним для завершення навчання, то результату не буде зовсім.

Для розв'язання цієї проблеми було введено рішення додавати певні вектори ваги, що «виживають» протягом довгого часу, аби зберегти більше інформації порівняно з векторами, які швидко скидаються.

Один із способів досягти цього — шлях голосування. Якщо перцептрон виживає протягом 30 прикладів, то він дістає 30 голосів. Якщо він вижив лише за два приклади, то він отримує лише 2 голоси. Ґрунтуючись на здобутій інформації, алгоритм перцептрону виглядатиме так:

$$\gamma = \text{sign} \left(\sum_{k=1}^k c^{(k)} (\omega^{(k)} * x + b^{(k)}) \right). \quad (3)$$

Після того, як дістали формулу, є змога описати базовий процес роботи алгоритму, зображеного на рисунку.

```

ω ← (o, o, ... o),   b ← o
v ← (o, o, ... o),   β ← o
c ← 1
for iter = 1 ... MaxIter do
  for all(x, y) ∈ D do
    if y(ω * x + b) ≤ o then
      ω ← ω + yx
      b ← b + y
      v ← v + yx
      β ← β + yc
    end if
  end for
  c ← c + 1
end for
return ω - 1/c v, b - 1/c β

```

Загальний принцип роботи вдосконаленого алгоритму

У першому рядку алгоритму відбувається ініціалізація ваги та біасу. У другому рядку — ініціалізація кешу для ваги та біасу. Після проведення всієї ітерації навчання результатами повертаються усереднені значення ваги та біасу, коли оригінальний алгоритм перцептрону з голосуванням повертає весь набір ваги для кожного проведеного прикладу.

Таким чином, розглядуваний алгоритм не буде виконувати пошук ідеального результату прогнозу, як у випадку перцептрону з голосуванням, і в разі виникнення помилки набуватиме значень вдалих ітерацій. При цьому алгоритм зберігатиме високий відсоток прогнозованих значень для поставленої задачі.

Висновок

У ході дослідження було проведено збір та спеціальну підготовку набору даних для проведення машинного навчання моделі. Було проведено кілька ітерацій добору необхідної комбінації даних, які під час навчання моделі давали найкращий результат при прогнозуванні результату.

Для розв'язання задачі було використано підхід «один до багатьох» через особливості підготовлених даних для машинного навчання. Описаний підхід дав найкращі результати для поставленого завдання.

Було виявлено найбільш оптимальний алгоритм машинного навчання з учителем для прогнозування лікаря-спеціаліста, ґрунтуючись на інформації симптомів зазначеної клієнтом системи, який записується на прийом. Вибраний метод демонстрував результати високої точності, але при цьому мав

недоліки у вигляді надмірної вибагливості стосовно використовуваних даних. Набір даних симптомів та спеціальностей лікарів досить обмежений і збільшити його задля отримання результату немає можливості. Тому було вирішено вносити зміни до алгоритму, які стосуються оцінювання стану ваги поточного прикладу, збільшивши у такий спосіб результативність роботи. Загалом, було здійснено аналіз використовуваного алгоритму машинного навчання та проведено оптимізацію щодо його роботи з метою покращення результатів прогнозування даних.

Список використаної літератури

1. **Karasikov M. E., Maximov Y. V.** Dimensionality reduction for multi-class learning problems reduced to multiple binary problems [Електронний ресурс]. URL: https://pdfs.semanticscholar.org/e54c/3435404710bf32afc_a0469e263621cbf6_95a.pdf (дата звернення 20.11.2019).
2. **Fayers M.** Elimination Tournaments Requiring a Fixed Number of Wins [Електронний ресурс]. URL: <https://neuralnet.info/chapter/perceptorns> (дата звернення 04.10.2019).
3. **A Course in Machine Learning** [Електронний ресурс]: [Веб-сайт]. URL: http://ciml.info/dl/v0_99/ciml-v0_99-ch04.pdf (дата звернення 09.10.2019).

Рецензент: доктор техн. наук, професор **В. В. Онищенко**, Державний університет телекомунікацій, Київ.

Б. И. Гончаренко, О. Ю. Ильин, А. Б. Коба, О. В. Негоденко СОВЕРШЕНСТВОВАНИЯ МЕДИЦИНСКОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ МЕДИЦИНСКИХ УЧРЕЖДЕНИЙ С ПОМОЩЬЮ MACHINE LEARNING

Рассмотрено использование машинного обучения с целью улучшить и усовершенствовать работу медицинских информационных систем Украины.

Большинство технических решений медицинских систем направлены на усовершенствование внутренней работы заведения и мало внимания уделяется взаимодействию с пациентами. Предложены средства машинного обучения для максимального упрощения записи клиентом в медицинское учреждение.

Сформулирована задача выявления оптимального алгоритма для подбора медицинского специалиста, основываясь на симптомах клиента без участия человека, а также усовершенствования работы такого алгоритма.

Исследован процесс сбора и специальной подготовки набора данных для проведения машинного обучения машинной модели. Проведена комплексная работа подбора необходимых комбинаций данных, которые давали лучший результат при его прогнозировании.

В связи с особенностью набора данных для обучения при проведении исследования использовался подход «один ко многим» для многоклассовых задач машинного обучения. Обнаружен наиболее оптимальный алгоритм машинного обучения с учителем для прогнозирования врача-специалиста, основываясь на информации относительно симптомов указанной клиентом системы во время записи на прием.

Предложенный метод был усовершенствован из-за недостатков в виде высокой требовательности относительно используемых данных. Изменения работы алгоритма касаются оценки состояния веса текущего примера прогноза, повышая результативность работы.

Усовершенствованный метод основывается на использовании усредненных значений весовых критериев вместо хранения всей коллекции критериев тяжести, что увеличивает скорость работы алгоритма, а также делает его менее требовательным к набору данных, используемых при обучении модели.

Ключевые слова: медицинские информационные системы; machine learning; перцептрон; машинное обучение; алгоритм; метод; перцептрон с голосованием.

B. I. Goncharenko, O. Y. Ilyin, A. B. Koba, O. V. Negodenko

IMPROVEMENT OF MEDICAL INFORMATION SYSTEM FOR MEDICAL INSTITUTIONS BY MACHINE LEARNING

This article is about using machine learning to improve and improve the performance of medical information systems in Ukraine.

Most medical systems technical solutions are designed to improve the internal work of the institution and little attention is paid to interacting with patients, so the work considers machine learning to maximize client access to the healthcare facility.

The article aims to identify the optimal algorithm for the selection of a medical specialist based on the client's symptoms without human involvement, as well as to improve the operation of such an algorithm.

The process of collecting and special training of the data set for machine learning of machine model is investigated. The complex work of selection of the necessary data combinations was carried out, which gave the best result in predicting the result.

The study used a one-to-many approach for many machine learning tasks, due to the particularity of the training dataset. The most optimal algorithm of machine learning with the teacher has been identified, for predicting specialist doctor based on the information of symptoms indicated by the client system, which is recorded for admission.

The method used has been refined, due to the disadvantages of high demanding data. Changes to the algorithm work to estimate the weight status of the current forecast example, increasing the performance.

The advanced method is based on the use of averaged values of weight criteria, instead of storing the entire collection of weight criteria, which increases the speed of the algorithm, and also makes it less demanding to the data set used in training the model.

Keywords: medical information systems; machine learning; perceptron; machine learning; algorithm; method; voting perceptron.