УДК 61:002.6:681.31:007

ПРОБЛЕМИ ВИЯВЛЕННЯ НОВИХ ЗНАНЬ ІЗ СХОВИЩ МЕДИЧНИХ ДАНИХ. ПЕРШЕ ПОВІДОМЛЕННЯ

О. П. Мінцер, С. В. Денисенко, Л. Ю. Бабінцева

Національна медична академія післядипломної освіти імені П. Л. Шупика

Розглянуті питання виявлення корисних і нових знань із інформації, яка постійно накопичується в медичних інформаційних системах. Проаналізовані підходи до виявлення нового знання в сховищах медичних даних. Підкреслюється думка про необхідність мінімізації втручання людини й аналізування даних по можливості більше автоматично. В першому повідомленні детальніше описані методи кластеризації даних.

Ключові слова: виявлення нового знання, виявлення та дослідження даних, кластеризація даних, розпізнавання образів, інтелектуальний аналіз даних.

ПРОБЛЕМЫ ИЗВЛЕЧЕНИЯ НОВЫХ ЗНАНИЙ ИЗ ХРАНИЛИЩ МЕДИЦИНСКИХ ДАННЫХ. ПЕРВОЕ СООБЩЕНИЕ

О. П. Минцер, С. В. Денисенко, Л. Ю. Бабинцева

Национальная медицинская академия последипломного образования имени П. Л. Шупика

Рассмотрены вопросы извлечения полезных и новых знаний из постоянно накапливающейся информации в медицинских информационных системах. Проанализированы подходы к обнаружению нового знания в хранилищах медицинских данных. Подчеркивается мысль о необходимости минимизации вмешательства человека и производстве анализа данных по возможности больше автоматически. В первом сообщении более подробно описаны методы кластеризации данных.

Ключевые слова: обнаружение нового знания, извлечение и исследование данных, кластеризация данных, распознавание образов, интеллектуальный анализ данных.

THE PROBLEM OF EXTRACTING NEW KNOWLEDGE FROM THE STORAGE OF MEDICAL DATA. FIRST ANNOUNCEMENT

O. P. Mintser, S. V. Denysenko, L. Yu. Babintseva

National Medical Academy of Post-Graduate Education named after P. L. Shupyk

The questions of extracting useful and new knowledge of the constantly accumulating data in medical information systems are considered. Approaches to the discovery of new knowledge in the storage of medical data are analyzed. There is emphasized the idea of the need to minimize human intervention and production data analysis as possible automatically. In the first report it is described in more detail the data clustering methods.

Key words: detection of new knowledge, data retrieval and research, data clustering, pattern recognition, data mining.

Вступ. Зрозуміло, що після застосування традиційних методів аналізу, будь то пов'язаного з аналізом перебігу хвороби та передбачуваного лікування або дослідження ефективності роботи медичного закладу, перед практичними лікарями встає завдання з подальшого покращення якості надання медичної допомоги. Для цього вкрай важливо використати накопичений досвід.

Проблема виявлення нового знання (ВНЗ) в сховищах медичних даних, що в англомовній літературі визначається так - knowledge discovery in databases (KDD) - і основного етапу цього процесу виявлення та дослідження даних (data mining), - одна з найважливіших у створенні сучасних інтелектуальних медичних інформаційних систем (МІС).

© О. П. Мінцер, С. В. Денисенко, Л. Ю. Бабінцева

Дивно, але статей про виявлення знань і, особливо, кластеризації інформації (як одного з основних інструментів, що використовуються для виявлення знань) відносно замало [1-3].

Основні визначення. Виявлення інформації є різновидом інформаційного оброблення матеріалу, при якому у зовні не пов'язаній інформації можна виявити логічні закономірності.

В сучасних інформаційних технологіях (ІТ) роль такої процедури, як виявлення інформації, все більше зростає через стрімке збільшення кількості неструктурованої інформації, зокрема, в Інтернеті. Ця інформація може бути зроблена більш структурованою за допомогою перетворення в реляційну форму або додаванням XML розмітки.

Два підходи - статистичний і нейромережевий - започаткували два досить різні за своїми методами та цілями класи систем інтелектуального аналізу баз даних. Можна сказати, що окремі статистичні методи є ніби знаряддями нижчого рівня, порівняно з методами виявлення знань (МВЗ). Методи виявлення знань користуються статистичними методами, комбінуючи їх у стандартних схемах рішення типових завдань.

Результати та їх обговорення. Загальна концепція МВЗ - мінімізувати втручання людини, зробити аналіз якомога більше автоматичним.

Більшість із існуючих методів виявлення знань використовують в якості окремих елементарних операцій класичні статистичні методи. Розглянемо їх детальніше.

Метод мультилінійної регресії з автоматичним вибором незалежних змінних дозволяє вибрати з великої кількості наявних незалежних параметрів найбільш важливі, результативно впливаючі на задану змінну. Фактично, цей метод у рамках деякої схеми застосування використовує стандартний метод лінійної регресії, тим самим дозволяючи значно менше апріорі знати про шукану модель.

Другий тип завдань - завдання опису наявних даних, виявлення в них залежностей із метою їх осмислення дослідником. Цей клас завдань включає: знаходження функціональних зв'язків між різними показниками та змінними у формі, що інтерпретується людиною; кластеризації симптомів, станів і захворювань; виділення "виключень" із зібраних даних, а також інтеграцію отриманих даних.

Якщо розглядати перший клас завдань, то зазвичай говорячи про функціональну залежність, мають на увазі залежності між безперервними числовими змінними.

Можна також розглядати залежності, що включають звичайні числові, булеві функції (типу "так/ні") і категоріальні змінні (нечислові параметри, скажімо, діагнози хвороб, що можуть бути закодовані).

При другому класі завдань є деякий набір описів об'єктів. Часто ці об'єкти не складають деякої єдиної маси, а природним чином розбиваються на будь-які групи. Наприклад, група пацієнтів, які страждають легеневими захворюваннями, дерматитами або розладами опорно-рухової системи. Необхідно ж виділити природні групи, або кластери, на основі наявної у базі даних інформації. Інтерпретація отриманої інформації може виявити чіткий медичний сенс, і нам було б цікаво зрозуміти, що означає таке групування точок і з чим воно пов'язане. Відмітимо також, що в проблемах дослідження організації медичної допомоги населенню знання розбиття всієї безлічі пацієнтів на деякі характерні групи може допомогти правильно організувати роботу медичного закладу.

У третьому класі завдань, що відноситься до опису виключень, виняткових ситуацій, записів (наприклад, окремих пацієнтів), що різко відрізняються чимнебудь від основної безлічі записів (групи хворих). Знання виключень може бути використане двояким чином. Можливо, ці записи утворюють собою випадковий збій, наприклад, помилки операторів, що вводили дані в комп'ютер. Подібну "шумову", випадкову складову має сенс виключити з подальших досліджень. З іншого боку, окремі, виняткові записи можуть представляти самостійний інтерес для дослідження, оскільки вони можуть вказувати на рідкісні, але важливі аномальні захворювання. Навіть сама ідентифікація цих записів, не говорячи про їх подальший аналіз і детальний розгляд, може виявитися дуже корисною для розуміння суті об'єктів, що вивчаються, або явищ.

Нарешті, останній, четвертий різновид завдань, що включається в даний клас інтелектуального аналізу даних, визначає підсумкову характеристику даних. Як приклад, розглянемо вибірку даних по пацієнтах не старше за тридцять років, що перенесли інфаркт міокарда. Якщо виявиться, що всі пацієнти, описані в цій вибірці, мають вагу не нижче 95 кг, це може бути важливим із точки зору розуміння логіки патологічного процесу, це практично нове знання. Отже, узагальнені дані - це знаходження будь-яких фактів, що вірні для всіх або майже всіх записів у вибірці даних, що вивчається, але які досить рідко зустрічалися.

Перший етап, що передує аналізу даних методами ВНЗ, полягає в приведенні даних до форми, придатної для застосування конкретних математичних

підходів, у тому числі побудова автоматичних рубрикаторів, класифікаторів симптомів, хвороб і так далі. Практично жодна з існуючих систем ВНЗ не може працювати безпосередньо з текстами. Відповідно, потрібні деякі похідні параметри, що відображають аналізований текст: частота зустрічання ключових слів, параметри, що характеризують поєднаність тих або інших слів у реченні тощо. Подібні завдання найменш автоматизовані в тому сенсі, що вибір системи цих параметрів або їх контроль (при автоматичному режимі) проводиться людиною.

Параметри, що описують наявну інформацію, зазвичай представляють у вигляді прямокутної таблиці, де кожен рядок є окремим випадком, об'єктом або станом об'єкта, що вивчається, а кожна колонка - параметри, властивості або ознаки всіх досліджуваних об'єктів. Рядки подібної таблиці в теорії ВНЗ прийнято називати записами, а колонки - полями. Практично всі наявні системи КОО працюють тільки з подібними прямокутними таблицями.

Отриману прямокутну таблицю необхідно заздалегідь обробити. Наступний етап - власне застосування методів ВНЗ. Сценарії цього застосування можуть бути найрізноманітнішими та включати складну комбінацію різних методів, особливо якщо використовувані методи дозволяють проаналізувати дані з різних точок зору. Власне цей етап дослідження і прийнято називати "розробкою даних".

Етапи виявлення нових знань. На першому етапі здійснюється верифікація та перевірка зібраних результатів. Він найчастіше здійснюється шляхом виділення дослідної та контрольної підгруп. Як правило, дослідна підгрупа - більшої кількості, контрольна - меншої. На більшій групі, застосовуючи ті або інші методи ВНЗ, отримуємо моделі залежності, а на меншій групі даних вони перевіряються. За різницею між двома підгрупами з точністю можна судити, наскільки адекватна, статистично значима побудована модель. Існує багато інших, складніших способів верифікації, такі як перехресна перевірка, метод "бутстреп" тощо, котрі дозволяють оцінити значущість моделей, що виводяться, без розбиття даних на дві групи.

Найважливішим, по суті, є другий етап - інтерпретація автоматично отриманих знань з метою їх застосування для прийняття рішень, додавання правил і залежностей у бази знань і так далі. Даний етап часто припускає використання методів, що знаходяться на стику технології ВНЗ і технології експертних систем. Від того, наскільки ефективним він буде, значною мірою залежить успіх вирішення поставленого завдання. Остаточне оцінювання цінності "нового знання" може бути проведено тільки після перевірки нового знання практикою.

Зазвичай, окрім статистичної значущості моделі оцінюється також її практична корисність, причому критерії оцінювання можуть бути не лише чисто медичні. Велику допомогу в оцінюванні корисності моделі можуть надати експертні оцінки графіки при застосуванні моделі (кількість позитивних відгуків на запропоновану моделлю зміну в курсі лікування, позитивний результат профілактичного заходу тощо). Важливою є також і економічна оцінка (вартість реалізації моделі й ефект від її реалізації).

Алгоритми виявлення знань. При дослідженні даних засобами data mining використовується велика кількість різних методів і їх різні комбінації. Найбільш важливі та часто використовувані стратегії включають: кластерізаційні підходи; методи пошуку асоціацій; побудову дерев рішень; аналіз із виборчою дією; методи побудови нейронних мереж; методи нечіткої логіки; генетичні алгоритми; регресійні методи; еволюційне програмування тощо.

Методи кластерного аналізу дозволяють розділити сукупність об' єктів, що вивчається, на групи "схожих" об'єктів, що називаються кластерами, рознести записи в різні групи або сегменти. Кластеризація в чомусь аналогічна класифікації, але відрізняється від неї тим, що для проведення аналізу не потрібно мати виділеної цільової змінної. Її зручно використовувати на початкових етапах дослідження, коли про дані мало що відомо. В більшості інших методів KDD дослідження починається, коли дані вже заздалегідь якось розкласифіковані, наприклад, на дані за якими перевіряється знайдена модель або для яких слід передбачити цільову змінну. Для етапу кластеризації характерна відсутність яких-небудь відмінностей як між змінними, так і між записами.

Існує велика кількість методів класифікації, які можна розділити на декілька груп. Наприклад, за способом завдання показника якості класифікації методи поділяються на евристичні й оптимізаційні. Евристичні алгоритми грунтовані на досвіді й інтуїції людини. Показник якості класифікації, що необхідно обернути в екстремум, в цих алгоритмах в явному виді не заданий. Евристичні алгоритми реалізують процедури, що мають раціональний сенс із точки зору логіки людини і що приводять у багатьох випадках до гарних результатів на практиці.

До *оптимізаційних алгоритмів* належать методи класифікації, у яких в явному виді заданий показник якості, що необхідно обернути в екстремум (мак-

симум або мінімум) за безліччю допустимого розбиття. На відміну від алгоритмів першої групи, розбиття, отримувані оптимізаційними алгоритмами класифікації є найкращими з точки зору обраного показника якості. Вибір конкретного показника залежить від специфіки й обмежень вирішуваного завдання, а також прийнятих пропозицій. Слід зазначити, що у багатьох випадках, в евристичних алгоритмах показник якості заданий у неявному виді, тому якщо вдасться його формалізувати та сформулювати в явному виді, алгоритми можуть стати оптимізаційними.

За способом об'єднання методи автоматичної класифікації поділяються на дивизимні, агломеративні й ітеративні.

Агломеративні методи послідовно об'єднують окремі об'єкти в групи (кластери), а *дивизимні методи* поділяють групи на окремі об'єкти. В свою чергу кожен метод класифікації як об'єднуючого, так і розділяючого типу може бути реалізований за допомогою різних алгоритмів. Слід зауважити, що як агломеративні, так і дивизимні алгоритми трудомісткі та їх складно використати для великих сукупностей. Крім того, результати роботи таких алгоритмів важко піддаються візуальному аналізу.

В кластерному аналізі існують також методи класифікації, які важко віднести до першої або до другої груп, - ітеративні методи - кластери формуються виходячи з умов розбиття, що задаються, котрі можуть бути змінені користувачем для досягнення бажаної якості. До ітеративних методів належать, наприклад, метод середніх, метод пошуку згущувань тощо. Ітеративні методи відносяться до швидкодіючих, що дозволяє використати їх для оброблення великих масивів початкової інформації. На відміну від агломеративних і дивизимних методів класифікації ітеративні алгоритми можуть призвести до створення пересічних кластерів, коли один об'єкт може одночасно належати декільком кластерам.

Популярніші в data mining дивизивні методи або методи розщеплювання, що безпосередньо розбивають усю сукупність записів на декілька кластерів. При цьому дуже важливими стають дослідження отриманих результатів.

Дослідження результатів, отриманих при кластерному аналізі. При вирішенні завдань кластерного аналізу доводитися зіткнутися з рядом проблем:

- кластерні об'єкти характеризуються великою кількістю чинників, що мають різні одиниці виміру та різні абсолютні розміри, буквально не порівнянні один із одним, і несуть різний обсяг інформації; спочатку невідомо число кластерів, на котре необхідно розбити вихідну сукупність елементів, і візуальні спостереження у багатовимірному випадку просто не приводять до успіху;

- неочевидні підходи до вибору метрик, що використовуються в якості міри відстані (міри близькості) між об'єктами;

- не існує єдиного алгоритму вибору цільової функції або методу об'єднання об'єктів у кластери.

Багато з методів кластеризації відрізняються між собою тим, що їхні алгоритми на кожному кроці обчисляють різноманітні функціонали якості розбивки. Необхідний кількісний критерій, слідуючи якому можна було б віддати перевагу одній розбивці, а не іншій.

Під найкращою розбивкою розуміють таку, при якій досягається екстремум (мінімум або максимум) обраного функціоналу якості. Вибір такого кількісного показника якості розбивки спирається часом на емпіричні розуміння. В якості таких функціоналів часто використовується "зважена" сума середкласових дисперсій відстаней, сума попарних середкласових відстаней між внутрішньокластерними елементами тощо. Об' єднуються за цим методом ті об' єкти, що дають мінімальне збільшення дисперсії.

Є декілька засобів порівняння різних ієрархічних агломеративних методів. За допомогою одного з них можна проаналізувати, як ці методи перетворять співвідношення між точками в багатовимірному просторі. Методи «стиснення» змінюють ці співвідношення, "зменшуючи" простір між будь-якими групами в даних. Коли чергова точка піддається опрацюванню таким методом, вона швидше за все буде приєднана до вже існуючої групи, а не послужить початком нового кластера. Методи розширення простору діють протилежно: по мірі проведення процедури кластеризації між вже створеними кластерами створюються нові, більш дрібні групи. Цей засіб групування також схильний до створення кластерів гіперсферичної форми та приблизно рівних розмірів. Методи Уорда та повних зв'язків є методами, що розширюють простір. І, нарешті, методи, що зберігають простір, такі, як метод середнього зв'язку, лишають без зміни властивості вихідного простору.

Щільність і локальність кластерів. Щільність розподілу точок (спостережень усередині кластера) - одна з важливих властивостей отриманих кластерів. Ця властивість дає нам можливість визначити наскільки даний кластер є компактним. Незважаючи на достатню очевидність цієї властивості, однозначного засобу обчислення такого показника (щільності) не існує. Найбільш вдалим показником,

що характеризує компактність, щільність "упаковування" багатовимірних спостережень у даному кластері є дисперсія відстані від центру кластера до окремих його точок. Чим менша дисперсія цієї відстані, тим більша щільність кластера. І навпаки, чим більша дисперсія відстані, тим більш розріджений даний кластер, і, отже, є точки, що знаходяться як поблизу центру кластера, так і достатньо віддалені від нього.

Основним показником розміру кластера є його радіус. Ця властивість найбільш повно відображає фактичний розмір кластера, якщо аналізований кластер має круглу форму та є гіперсферою у багатовимірному просторі. Проте, якщо кластери мають подовжені форми, то поняття радіуса або діаметра вже не дає реальніших розмірів кластеру.

Локальність кластера характеризує ступінь перекриття та взаємної далекості кластерів один від одного в багатовимірному просторі.

Як вже підкреслювалося, дуже поширеним є метод К-середніх. Ідея методу така. Задається К - число кластерів, на які спочатку розбиваються зібрані дані. Обирається К довільних початкових центрів - точок у просторі всіх змінних. При цьому не критично, які саме це будуть центри, процедура вибору вихідних точок відіб'ється, головним чином, тільки на часі розрахунку. Здійснюється ітераційний процес, що реалізовує одну і ту ж операцію з двох кроків. На першому кроці розбиваються всі дані на К-груп, найбільш близьких до одного з центрів. Мірою близькості може бути відстань у просторі всіх змінних (якщо змінним приписати геометричний сенс).

Підкреслимо, що останнім часом застосуванню методів ВНЗ у біології і медицині, зокрема, в розвитку методів додаткових репродуктивних технологій, у розшифровці макромолекул, і в створенні нових лікарських засобів (ЛЗ) приділяється велика увага.

Розпізнавання об 'єктів при ВНЗ. Під розпізнаванням зазвичай розуміється віднесення конкретного об'єкта (реалізації), представленого значеннями його властивостей (ознак), до одного з фіксованих переліків образів (класів) за певним вирішальним правилом відповідно до поставленої мети.

Розпізнавання може здійснюватися і для виміру значень ознак, проведення обчислень, що реалізовують вирішальне правило тощо. При цьому перелік образів, інформативних ознак і вирішальні правила або задаються ззовні системі, що розпізнає, або формуються самою системою. Допоміжна, але важлива функція систем, що розпізнають, - оцінка ризику втрат. Без цієї функції неможливо, наприклад, побудувати оптимальні вирішальні правила, обрати найбільш інформативну систему ознак, що використовуються при розпізнаванні тощо.

Введемо такі позначення:

– $2 \le S < \infty$ - множина розпізнаваних образів (класів), що називається іноді алфавітом;

- X - ознаковий (вибірковий) простір;

- N - розмірність ознакового простору (кількість ознак, що характеризують розпізнавані об'єкти);

– D_x - множина вирішальних правил, за якими здійснюється віднесення розпізнаваного об' єкта (реалізації) до того або іншого образу;

- R - ризик втрат при розпізнаванні.

Кількість розпізнаваних образів завжди кінцева та не може бути меншою двох. Перелік образів може задаватися розпізнаваній системі ззовні. В багатьох випадках система, що розпізнає, сама формує перелік розпізнаваних образів. Цей процес називають навчанням без учителя або самонавчанням.

Розмірність ознакового простору N зазвичай прагнуть зробити якомога меншою, оскільки при цьому скорочується кількість необхідних вимірів, спрощуються обчислення, формувальні та реалізовуючі вирішальні правила, підвищується статистична стійкість результатів розпізнавання. В той же час зменшення N взагалі кажучи, веде до зростання ризику втрат. Тому формування ознакового простору є компромісним завданням, що можна поділити на дві частини: формування початкового ознакового простору та мінімізація розмірності цього простору. В частині стосовно мінімізації розмірності існують формальні методи, алгоритми та програми. Що ж до початкового простору, то його формування поки що ґрунтується на досвіді й інтуїції. Теоретично обґрунтовані підходи до вирішення цього завдання в літературі нечисленні. Основна мета, що при цьому переслідується, - мінімізація ризику втрат.

Ризик втрат *Я* фактично є критерієм, за яким формується найбільш інформативний ознаковий простір і найбільш ефективні вирішальні правила. Й алфавіт, і ознаки, і вирішальні правила мають бути такими, щоб по можливості мінімізувати ризик втрат. Цей критерій (характеристика системи, що розпізнає) є складеним. У нього в загальному випадку входять втрати за помилки розпізнавання та витрати на виміри ознак розпізнаваних об'єктів. У частному, найширше використовуваному випадку, в якості ризику втрат фігурує середня вірогідність помилки розпізнавання або максимальна компонента матриці вірогідності помилок. На практиці, звичайно, йдеться не про вірогідність, а про їх вибіркові оцінки.

Продовження теоретичних міркувань та аналіз практичних результатів у наступних повідомленнях.

Висновки. 1. Більшість із існуючих методів виявлення знань використовують на першому етапі в якості основних класичні статистичні підходи.

2. Формування початкового простору ознак і станів до теперішнього часу грунтується на досвіді й інтуїції. Теоретично обгрунтовані підходи до вирішення цього завдання в літературі переважно будуються на мінімізації ризику втрат.

Література

1. Фролов Ю. В. Интеллектуальные системы и управленческие решения / Ю. В. Фролов. - М. : МГГО, 2000. - 294 с.
2. Berry M. J. A. Data Mining Techniques: For Marketing, Sales and Customer Relationship Management / M. J. A. Berry, G. S. Linoff. - 2nd ed. - USA : John Willey & Sons, Inc., 2004. - 643 p.
3. Bigus J. P. Data mining with Neural Networks: Solving Business Problems - From Application Development to Decision Support / J. P. Bigus. - New York : McGraw-Hill, 1996. - 174 p.