

Комлева Наталія Олегівна
кандидат технічних наук,
доцент кафедри системного програмного забезпечення,
Одеський національний політехнічний університет

Комлевої Олександр Миколайович
старший викладач кафедри клінічної імунології,
генетики та медичної біології,
Одеський національний медичний університет

МЕТОДИКА ВИДІЛЕННЯ ІСТОТНИХ ОЗНАК КЛАСІВ У ЗАДАЧАХ МЕДИЧНОГО ДІАГНОСТУВАННЯ

METHOD OF SELECTION OF ESSENTIAL ATTRIBUTES CLASSES FOR THE MEDICAL DIAGNOSIS

Комлевая Наталья Олеговна
кандидат технических наук,
доцент кафедры системного программного обеспечения
Комлево́й Александр Николаевич
старший преподаватель кафедры клинической иммунологии,
генетики и медицинской биологии,
Одесский национальный медицинский университет
Komleva N.O.

Doctor of Philosophy (PhD),
Odessa National Polytechnic University
Komlevoý O.M.
PhD student, the First Medical Faculty,
Odessa National Medical University

Анотація: У статті розглянуті проблеми вирішення завдань медичного діагностування з використанням автоматизованого підходу. З використанням онтологічного підходу сформована база знань з метаданими про захворювання дихальної системи людини. Докладно розглянута методика, що дозволяє виділяти суттєві для класифікації ознаки при діагностуванні стану бронхо-легеневої системи. Побудовано дерево класифікації, що містить 39 непересічних класів, що характеризують стан дихальної системи у хворих і здорових обстежуваних. Наведено практичні результати класифікації з використанням спеціалізованої програми для статистичного аналізу STATISTICA.

Ключові слова: класифікація, медичне діагностування, онтологія, ознаки класів.

Summary: The article deals with the problem of solving medical diagnosis using an automated approach. The knowledge base metadata on the human respiratory system diseases is formed with the use of the ontological approach. The technique that allows to allocate significant for the classification of signs in the diagnosis of the state of broncho-pulmonary system is considered in detail. The classification tree containing 39 non-overlapping classes, characterizing the state of the respiratory system in patients and healthy humans is constructed. The practical results of classification using specialized applications for statistical analysis STATISTICA are shown.

Key words: classification, medical diagnostics, ontology, features classes.

Вступ

Під класифікацією в найбільш загальному сенсі розуміється однорівневий або багаторівневий логічний поділ обсягу поняття. Згідно іншому формулюванню, класифікація є розподілом досліджуваних об'єктів за класами (групами), що задовольнить двом умовам: 1) перетин будь-яких двох класів класифікації або порожнім, або дорівнює одному з цих класів; 2) об'єднання усіх виділених класів дорівнює множині, що класифікується. Клас тут розуміється традиційно, в теоретико-множинному сенсі як сукупність об'єктів, що виділяються серед множини інших за подібністю (збігом) в якихось ознаках. Друге визначення не накладає особливих обмежень на спосіб побудови класифікацій. Вони можуть будуватися як зверху, через операцію логічного поділу, так і знизу, шляхом об'єднання об'єктів класифікації в класи послідовно зростаючого обсягу.

Вирішення багатьох завдань медичного діагностування прямо пов'язане з класифікацією станів досліджуваних об'єктів. При цьому розробка ефективної технології обробки медичних даних, що дозволяє виконувати автоматизоване діагностування стану пацієнта, залишається актуальним завданням.

Метою даної роботи є формування методики виділення істотних ознак непересічних класів, що дозволяють вирішувати задачу класифікації стану дихальної системи людини з використанням статистичних методів. Проведений аналіз базується на дослідженні наборів медичних даних, одержуваних у результаті спеціалізованих обстежень пацієнтів, які входять в комплекс заходів для проведення неінвазивного пульмонологічного діагностування. Отримані результати використовуються для оцінки стану дихальної системи обстеженого пацієнта [1].

Основна частина

Об'єктом діагностування є стан дихальної системи людини; при цьому аналіз відповідної медичної літератури показав, що оцінку стану гомеостазу дихальної системи людини доцільно проводити на підставі аналізу конденсату вологи видихуваного їм повітря. Даний метод зарекомендував себе як перспективний і неінвазивний; він легко піддається формалізації та автоматизації [2].

З використанням онтологічного підходу сформована і формалізована база знань, в якій використані метадані про клінічний перебіг та діагностику захворювань дихальної системи. Знання щодо опису захворювань, які підлягають формалізації, витягали з різних літературних джерел, потім вони формувалися на основі принципів синтезу – таких як об'єднання і доповнення, і були формалізовані [3]. Згідно онтологічної моделі, формальні описи були представлені сукупністю якісних і кількісних характеристик, отриманих в результаті використання лабораторних та інструментальних методів діагностики; для кожної характеристики були визначені діапазон значень і варіант норми.

Як матеріал для досліджень взято дані по різним групам пацієнтів, кожен з яких пройшов пульмонологічне обстеження. Проміжні результати обстеження кожного пацієнта представлені вектором з 32 ознак, які характеризують стан дихальної системи. Значення ознак є кількісними величинами, вимірними на безперервній шкалі. Угрупування пацієнтів проведена апріорно на підставі медичних рекомендацій, заснованих на стандартних

діагностичних методах [4]; число пацієнтів в обстежуваних групах (ОГ) різне.

Класифікатор ОГ К містить повною мірою умови проведення експерименту і ті дані про обстежуваного, які дозволяють віднести його до тієї чи іншої класифікаційної групи. Шаблон класифікатора має наступний вигляд:

$$K = \langle k1 \ k2 \ k3 \ k4 \ k5 \ k6 \ k7 \ k8 \rangle,$$

де k1 – код апріорно відомого захворювання або групи захворювань (k1=0 – здоровий);

k2 – стать (0 – чоловіча, 1 – жіноча);

k3 – наявність спортивного способу життя (0 – не займається спортом, 1 – займається);

k4 – вік (0 – до 14 років, 1 – с 14 до 25 років, 2 – с 25 до 40 років, 3 – с 40 до 65 років, 4 – старше 65 років);

k5 – куріння (0 – не курить, 1 – курить);

k6 – сезон (0 – літо, 1 – осінь, 2 – зима, 3 – весна);

k7 – час доби (0 – ранок, 1 – день, 2 – вечір);

k8 – наявність фізичного навантаження (0 – до навантаження, 1 – після навантаження).

При проведенні різних досліджень деякі поля класифікатора К можуть не враховуватися. Наприклад, при обстеженні групи студентів, що не палять, в ранковий час взимку необхідно встановити такі значення полів класифікатора: k4 = 1, k5 = 0, k6 = 2, k7 = 0. Значеннями інших полів класифікатора можна знехтувати.

Використання можливих значень полів шаблону класифікатора К дозволило побудувати дерево класифікації (табл. 1).

Таблиця 1

Вузол	Предиктор (незалежна змінна)	Тип предиктора	Потужність	Категорії	Підлеглі вузли
K	Вік	Альтернативний	2	Дорослі Діти	K1 K2
K1	Заняття спортом	Альтернативний	2	Ні Так	K11 K12
K11	Стан здоров'я	Альтернативний	2	Здоровий Хворий	K111 K112
K111	Паління	Альтернативний	2	Палить Не палить	K1111 K1112
K1111	Фізичне навантаження	Альтернативний	2	Перед Після	K11111 K11112
K11111	Стать	Альтернативний	2	Чоловічий Жіночий	K111111 K111112
K111111	Сезон	Номінальний	3	Осінь Зима Весна	K1111111* K1111112* K1111113*
K111112	Сезон	Номінальний	3	Осінь Зима Весна	K1111121* K1111122* K1111123*
K11112	Стать	Альтернативний	2	Чоловічий Жіночий	K111121 K111122
K111121	Сезон	Номінальний	3	Осінь Зима Весна	K1111211* K1111212* K1111213*
K111122	Сезон	Номінальний	3	Осінь Зима Весна	K1111221* K1111222* K1111223*

K1112	Фізичне навантаження	Порядковий	2	Перед	K11121
				Після	K11122
K11121	Стать	Альтернативний	2	Чоловічий	K111211
				Жіночий	K111212
K111211	Сезон	Номинальний	3	Осінь	K1112111*
				Зима	K1112112*
				Весна	K1112113*
K111212	Сезон	Номинальний	3	Осінь	K1112121*
				Зима	K1112122*
				Весна	K1112123*
K11122	Стать	Альтернативний	2	Чоловічий	K111221
				Жіночий	K111222
K111221	Сезон	Номинальний	3	Осінь	K1112211*
				Зима	K1112212*
				Весна	K1112213*
K111222	Сезон	Номинальний	3	Осінь	K1112221*
				Зима	K1112222*
				Весна	K1112223*
K112	Захворювання	Номинальний	2	ХОЗЛ	K1121
				Туберкульоз перед початком лікування	K1122*
K1121	Час обстеження	Порядковий	3	Перед початком лікування	K11211*
				Під час лікування	K11212*
				Після лікування	K11213*
K12	Вид спорту	Номинальний	3	Самбо	K121
				Баскетбол	K122
				Волейбол	K123
K121	Навантаження на тренуванні	Порядковий	2	Перед	K1211*
				Після	K1212*
K122	Навантаження на тренуванні	Порядковий	2	Перед	K1221*
				Після	K1222*
K123	Навантаження на тренуванні	Порядковий	2	Перед	K1231*
				Після	K1232*
K2	Стан здоров'я	Альтернативний	2	Здоровий	K21*
				Хворий	K22
K22	Захворювання	Номинальний	3	Бронхіт	K221
				Пневмонія перед початком лікування	K222*
				Вегето-судинна дистонія перед початком лікування	K223*
K221	Час обстеження	Порядковий	2	Перед початком лікування	K2211*

Примітка. Символом «*» позначені діагностичні класи – кінцеві листові вузли дерева класифікації.

Для проведення статистичного аналізу даних використаний дискримінантний аналіз, що дозволяє вирішувати завдання ідентифікації та класифікації об'єктів по заданому набору характерних ознак [5].

Наведемо основні положення для вирішення задачі класифікації, для цього введемо такі позначення: g – кількість груп, p – кількість змінних, q – кількість обраних змінних, X_{ijk} – значення змінної i у спостереженні k в групі j , f_{jk} – вага спостереження k в групі j (для незваженої ситуації $f_{jk}=1$), m_j – кількість спостережень в групі j , n_j – сума ваг спостережень в групі j (для незваженої ситуації n_j – кількість спостережень в групі j), n – загальна сума ваг (для

незваженої ситуації n – загальна кількість спостережень). Середнє значення для змінної i в групі j визначається за формулою:

$$\bar{X}_{ij} = \left(\sum_{k=1}^{m_j} f_{jk} X_{ijk} \right) / n_j.$$

Для внутрішньогрупової матриці розсіювання спостережуваних змінних від середніх значення елемента матриці W в i -тому рядку і j -том стовпці визначається за формулою:

$$w_{il} = \sum_{j=1}^g \sum_{k=1}^{m_j} f_{jk} X_{ijk} X_{ilk} - \sum_{j=1}^g \left(\sum_{k=1}^{m_j} f_{jk} X_{ijk} \right) \left(\sum_{k=1}^{m_j} f_{jk} X_{ilk} \right) / n_j,$$

де $i, l = 1, \dots, p$.

При виборі змінних для аналізу вони включаються в аналіз в тому порядку, в якому записані. Змінна додається в аналіз, якщо при цьому жодна інша змінна не має допуск (TOL) менше

встановленої межі (за замовчуванням – 0.001). Під час вибору змінних на кожному кроці (при кожному додаванні нової змінної в аналіз) розраховується матриця W^* . Якщо перші q змінних були обрані для аналізу, матрицю W можна розділити наступним

чином: $W = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}$, де W_{11} має розмірність

$q \times q$. Тоді матрицю W^* можна визначити як

$$W^* = \begin{pmatrix} -W_{11}^{-1} & W_{11}^{-1}W_{12} \\ W_{21}W_{11}^{-1} & W_{22} - W_{21}W_{11}^{-1}W_{12} \end{pmatrix} = \begin{pmatrix} W_{11}^* & W_{12}^* \\ W_{21}^* & W_{22}^* \end{pmatrix}.$$

Допуск i -тій змінної визначається за формулою:

$$TOL_i = \begin{cases} 0, & \text{якщо } w_{ii} = 0 \\ w_{ii}^* / w_{ii}, & \text{якщо змінна } i \text{ не включена в аналіз та } w_{ii} \neq 0 \\ -1/(w_{ii}^* w_{ii}), & \text{якщо змінна } i \text{ включена в аналіз та } w_{ii} \neq 0 \end{cases}.$$

Коли обраний набір з q змінних, класифікаційні функції можна обчислити, використовуючи такі формули:

для коефіцієнтів:

$$b_{ij} = (n - g) \sum_{l=1}^q w_{il}^* \overline{X_{lj}}, \quad i = 1, 2, \dots, q, \quad j = 1, 2, \dots, g$$

для константи:

$$a_j = \ln p_j - \frac{1}{2} \sum_{i=1}^q b_{ij} \overline{X_{ij}}, \quad j = 1, 2, \dots, g, \quad \text{де } p_j -$$

апріорна ймовірність групи j , яку можна визначити за

$$\text{формулою: } p_j = \frac{n_j}{n}.$$

Таким чином, з усієї множини ознак були обрані ті, які є істотними для проведення класифікації. Для розрізнення різних класів використовуються різні ознаки. Кількість істотних ознак коливається від 8 до 17 у порівнянні з первинною їх кількістю – 32.

Однак дисперсійний аналіз дозволив перевірити лише гіпотезу про відсутність відмінностей між порівнюваними групами по досліджуваним ознаками; з його допомогою неможливо дізнатися, які саме групи розрізняються між собою. Для з'ясування цього були використані методи множинних порівнянь, що є частиною, так званого, апостеріорного аналізу (Post-hoc analysis). Механізм їх роботи полягає в проведенні попарних порівнянь середніх значень всіх груп, включених до дисперсійного аналізу [6].

З набору тестів post-hoc аналізу обрані LSD (перевірка критерію найменшої значущої різниці) і Newman-Keuls test (критерій Ньюмана-Кеулса). Згідно тесту LSD, на першому етапі методу за допомогою критерію Фішера була перевірена гіпотеза про рівність математичних очікувань розподілів, з яких отримано вибірки значень ознак. Попередньо перевірено, що ці вибірки є нормально розподіленими і однойменні вибірки мають однакові дисперсії. Якщо гіпотеза приймалася, то метод зупинявся, інакше переходили до наступного етапу. На другому етапі методу вибірки були впорядковані за зростанням вибіркового середнього і далі поетапно перевірені гіпотези рівності середніх сусідніх вибірок за допомогою критерію Стюдента. В якості оцінки дисперсії використовувалося внутрішньогрупове середнє. Якщо гіпотеза про рівність математичних очікувань приймалася, то відповідні вибірки об'єднувалися у одну групу.

Щоб визначити, чи є суттєва різниця між двома вибірками з рівними обсягами, метод Ньюмана-Кеулса використовує наступну формулу:

$$q = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{MSE}{n}}},$$

де q представляє значення стьюдентизованого діапазону, \overline{X}_A і \overline{X}_B є відповідно найбільшим і найменшим середнім значенням двох вибірок, MSE – дисперсія помилки, яка береться з таблиці ANOVA, а n – обсяг вибірок.

Обчислення значення q потім порівнювали з критичним значенням q , узятим з таблиці розподілу q . Якщо обчислене значення q дорівнює або більше, ніж критичне значення q , то нульова гіпотеза (про рівність середніх двох вибірок) може бути відхилена.

При різних обсягах вибірок використана наступна формула для методу Ньюмана-Кеулса:

$$q = \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{MSE}{2} \left(\frac{1}{n_A} + \frac{1}{n_B} \right)}},$$

де N_A та N_B – обсяги вибірок.

Обробка даних була проведена з використанням двох методів дискримінантного аналізу з пакету STATISTICA: стандартного і покрокового (включення і виключення). В якості змінної, що групує, вибирався діагноз пацієнта, в якості незалежних змінних – значення 32 ознак. Були побудовані функції класифікації – лінійні функції, які обчислювалися для кожного діагностичного класу і були використані для класифікації станів дихальної системи. На підставі функцій будувалася матриця класифікації, що містила інформацію про кількість та відсоток коректно класифікованих пацієнтів у кожній групі. Для більшої зручності користувалися квадратами відстані Махаланобіса, що показували, на яку величину стан кожного пацієнта відстоїть від центроїда групи. При діагностуванні стану дихальної системи нового пацієнта його відносили до тієї діагностичної групи, до якої він найближче розташований.

Висновки

Таким чином, у статті була представлена методика, що дозволяє з множини ознак діагностичних класів вибрати ті, які дозволяють проводити якісну діагностику стану дихальної системи людини. Використання методики дало можливість зменшити кількість аналізованих діагностичних параметрів більш ніж вдвічі. За результатами класифікації 278

об'єктів, для яких був апріорно відомий діагноз, з використанням описаної вище методики, в 266 випадках було отримано істинний результат, що показує високу ступінь кореляції апріорної і апостеріорної класифікаційної інформації.

Список літератури:

1. Yu. I. Bazhora, A. N. Komlevo, M. M. Chesnokova, A. Nalazek, W. Zukow. Respiratory system estimation at the healthy children and children with bronchitis with the use of laser correlative spectroscopy // Journal of Health Sciences. – 2013. – Vol. 3. No. 7. – P. 135-150.
2. Комлевая Н.О., Комлевой А.Н., Чернега К.С. Проектирование специализированной компьютерной системы для проведения пульмонологического диагностирования. – Научный журнал "Проблеми програмування". – Киев, 2014. – № 2 – 3. – С. 253 – 262.
3. Кавицкая В.С., Любченко В.В., Лысюк А.В. Модель представления знаний для систем управления знаниями. – Праці Одеського політехнічного університету – Одеса, 2013. – Вип. 3. – С. 167 – 172.
4. Комлевая Н.О., Комлевой А.Н. Разработка информационной модели диагностирования состояния дыхательной системы. – Холодильна техніка і технологія. – Одеса, 2011. – Вип. 2(130). – С. 75 – 79.
5. Комлевая Н.О. Построение системы диагностических признаков с использованием метода дискриминантного анализа в офтальмологических исследованиях. – Радіоелектронні і комп'ютерні системи. – Харків «ХАІ», 2010. – Вип. 6 (47). – С. 250 – 253.
6. Комлевая Н.О., Комлевой А.Н., Тимченко Б.И. Сравнительный анализ двух подходов при решении задачи классификации. – Наукотехнічний журнал "Радіоелектронні і комп'ютерні системи". – Харьков, 2014. – № 6(70). – С. 115 – 119.

Пизинцали Людмила Викторовна

*Кандидат технических наук, доцент кафедры судоремонта,
Одесский национальный морской университет*

ПОДГОТОВКА ПРОЕКТА УТИЛИЗАЦИОННОГО ПРЕДПРИЯТИЯ ПІДГОТОВКА ПРОЕКТУ УТИЛІЗАЦІЙНОГО ПІД

ПРИЄМСТВА

PREPARATION OF A DRAFT RECYCLING FACILITY Пизинцали Людмила Викторовна

Кандидат технічних наук, доцент кафедри судноремонту,

Одеський національний морський університет

Pizintsali L.V.

Ph.D., assistant professor of ship repair,

Odessa National Maritime University

Аннотация. В статье рассмотрены вопросы экологии и утилизации – основного принципа устойчивого развития государства. Рассмотрены этапы проекта утилизации судна, приведены обязательные требования к «утилизатору». Сделан акцент на то, что при разборке судна, основным критерием должно быть применение энергосберегающих, безотходных и малоотходных технологических решений; на каждое конкретное судно должна быть разработана программа утилизации, удовлетворяющая всем требованиям экологического и природоохранного законодательства. Поднимается вопрос о возможности введения утилизационного сбора или предоставления украинским судовладельцам утилизационного гранта

Ключевые слова: утилизация, судно, отходы, утилизационное предприятие, окружающая среда, Украина, проек утилизации, стандарты ISO 30 000

Анотація. У статті розглянуті питання екології та утилізації - основного принципу сталого розвитку держави. Розглянуті етапи проекту утилізації судна, наведено обов'язкові вимоги до «утилізатору». Зроблено акцент на те, що при розбиранні судна, основним критерієм має бути застосування енергозберігаючих, безвідходних та маловідходних технологічних рішень; на кожне конкретне судно має бути розроблена програма утилізації, що задовольняє всім вимогам екологічного та природоохоронного законодавства. Піднімається питання про можливість введення утилізаційного збору або надання українським судновласникам утилізаційного гранту

Ключові слова: утилізація, судно, відходи, утилізаційного підприємство, навколишнє середовище, Україна, проек утилізації, стандарти ISO 30000

Summary: The article deals with environmental issues and waste - the basic principle of sustainable development of the state. The stages of the draft ship recycling, given the mandatory requirements to "the exchanger." It emphasizes the fact that the dismantling of the vessel, the main criterion should be the use of energy-saving, low-waste and non-waste technology solutions; for each particular vessel should be designed Bat recycling program that meets all the requirements of the environmental and nature protection legislation. It raises questions about the possibility of the recycling or collection of Ukrainian shipowners scrappage grant

Key words: recycling, ship waste recycling facility, the environment, Ukraine, recycling projects, the ISO standards 30000