операторів телекомунікаційних і технологічних систем та синтез тестів для їх відбору / Л.С. Сікора, Н.К. Лиса, Ю.Г. Міюшкович // Моделювання та інформаційні технології. - К. ІПМЕ. 2009. – Вип. 54. – С.190-195.
13.*Сікора Л.С., Лиса Н.К., Якимчук Б.Л.* Моделі оперативних експертних висновків при неповних даних про стан інтегрованих систем для формування образів ситуацій та управляючих рішень // Л.С. Сікора, Н.К. Лиса, Б.Л. Якимчук // ЗНП, Інститут проблем моделювання в енергетиці. - К. ІПМЕ. 2013. – Вип. 70. – С.177-192.
14. *Сікора Л.С.* Формування причинно – наслідкових зв'язків при оцінці динамічних термінальних ситуацій в потенційно – небезпечних об'єктах. / Л.С. Сікора, Б.Л.Якимчук, Т.Є.Рак // ЗНП, Інститут проблем моделювання в енергетиці. К.ІПМЕ ім.. Пухова – 2012. – Вип. 65. – С.107-125
15. *Сікора Л.С.* Термінальні та ситуаційні проблемні задачі інформаційного забезпечення опрацювання даних оператором від інформаційно – вимірювальних систем для АСУ-ТП складними об'єктами. / Л.С. Сікора, Н.К. Лиса, Б.Л. Якимчук, Р.С. Марцишин, Ю.Г. Міюшкович // Вісник НУ ,,ЛП,, ,,Інформаційні системи і мережі,.№783 – Львів. Вид. Львівської політехніки. 2014.- С.204-2016.

Поступила 26.09.2016р.

УДК 004.932.2:616-006.6

С.О. Вербовий, м.Тернопіль

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ПОШУКУ ЗАКОНОМІРНОСТЕЙ В БАЗАХ ДАНИХ БІОМЕДИЧНИХ ЗОБРАЖЕНЬ

Abstract. The paper presents the use of data mining to find patterns in a data base cytological and histological images. This will help make a classification features microscopic cytological and histological images and build association rules for the diagnosis of precancerous and cancerous conditions of the breast.

Актуальність

За даними канцер реєстру [1] смертність жінок від раку молочної залози посідає перше місце. Тому проблема діагностування злоякісних новоутворень на ранніх стадіях є актуальною.

Рання діагностика потребує точної і надійної методики постановки діагнозу. Діагностування базується на аналізі мікроскопічних зображень окремих клітин (цитологічні зображення) та тканин (гістологічні зображення).

На сьогодні широкого розповсюдження набули методи інтелектуального аналізу даних (ІАД), які застосовуються в різних областях [2].

Технології ІАД дають можливість виявляти шаблони правил. Правила в нашому випадку - це поєднання кількісних та якісних ознак мікрооб'єктів з

логічним висновком про стан окремих клітин або тканин. На основі сформованих правил експерт встановлює попередній діагноз. Програмні засоби ІАД дозволяють автоматично здійснювати пошук цих правил.

Тому актуальною задачею є застосування інтелектуального аналізу даних для пошуку закономірностей у базі даних цитологічних та гістологічних зображень.

Аналіз публікацій

В роботі [3] представлено удосконалений алгоритм Apriori, метою якого є зменшення розміру набору ознак мікрооб'єктів. Результатом його роботи є побудова асоціативних правил на основі ознак мікрооб'єктів. Інше дослідження [4] передбачає вивчення різних методик діагностування і прогнозування раку молочної залози, а також досліджує методи видобування даних. Точність діагностики на основі методів ІАД є прийнятною і допомагає медикам в прийнятті рішень на етапі ранньої діагностики. У статті [5] запропоновано алгоритм нечітких асоціативних правил для прогнозування ризиків виникнення раку молочної залози. У роботах [6,7] представлені різні підходи ІАД для підвищення точності діагностики і прогнозування раку молочної залози.

Постановка задачі

Метою даної роботи є застосування алгоритмів класифікації та пошуку асоціативних правил в базі даних гістологічних та цитологічних зображень передракових станів молочної залози.

Розв'язок задачі

Для розв'язку задачі класифікації ознак мікрооб'єктів використано алгоритм J48, який відноситься до алгоритмів дерева рішень. Цей алгоритм є відкритим вихідним кодом реалізації Java алгоритму C4.5 для інтелектуального аналізу даних інструменту Weka.

Нехай маємо множину прикладів *T*, заданої потужності |*T*|, де кожен елемент цієї множини описується *m* атрибутами. Крім цього задано класи *C1, C2... Ck.*

Завданням алгоритму є побудова ієрархічної класифікаційної моделі у вигляді дерева з множини прикладів *T*. Процес побудови дерева відбуватиметься зверху вниз. Спочатку створюється корінь дерева, потім нащадки кореня і т.д.

На першому кроці отримуємо порожнє дерево (є тільки корінь) і вихідну множину *T* (асоційовану з коренем). Потрібно розбити вихідну множину на підмножини. Це можна зробити, вибравши один з атрибутів в якості перевірки. Тоді в результаті розбиття виходять *n* (по числу значень атрибута) підмножин і, відповідно, створюються *n* нащадків кореня, кожному з яких поставлена у відповідність своя підмножина, отримана при розбитті множини *T*. Ця процедура рекурсивно застосовується до всіх підмножин (нащадків кореня).

Розглянемо детальніше критерій вибору атрибута, за яким проходить

розгалуження. Нехай ми маємо *m* можливих варіантів, з яких необхідно вибрати найбільш оптимальний. Деякі алгоритми виключають повторне використання атрибута при побудові дерева, але в даному випадку ми таких обмежень не накладаємо. Будь-який з атрибутів можна використовувати необмежену кількість разів при побудові дерева.

Нехай дано множину *X*, яка використовується для перевірки побудови дерева. Розбиття множини *T* з врахуванням множини *X* дасть нам підмножини *T1, T2 ... Tn*. Тоді множина *X* розбивається на підмножини *A1, A2 ... An*.

Позначимо через freq(Cj,S) - кількість прикладів з деякої множини *S*, що відносяться до одного і того ж класу *Cj*. Тоді ймовірність того, що випадково обраний приклад з множини *S* буде належати до класу *Cj* рівна:

$$\text{P} = \frac{\text{freq(Cj,S)}}{|\text{S}|}$$

Вираз оцінки середньої кількості інформації для визначення класу з множини *T*, представлено у вигляді:

$$Info(T) = -\sum_{j=1}^{k} \frac{freg(Cj,T)}{|T|} * \log_2 \frac{freg(Cj,T)}{|T|} \quad (1)$$

Після розбиття множини *T* з врахуванням множини *X* оцінка середньої кількості інформації рівна:

$$Info_x(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} * Info(T_i) \quad (2)$$

Тоді критерій для вибору атрибута задається виразом:

$$Gain(X) = Info(T) - Info_x(T) \quad (3)$$

Критерій (3) використовується для всіх атрибутів, причому вибираємо той атрибут, який максимізує цей вираз. Цей атрибут буде перевіркою в поточному вузлі дерева, а потім з цього атрибуту проводиться подальша побудова дерева.

Для пошуку асоціативних правил ознак мікрооб'єктів біомедичних зображень використано алгоритм Tertius.

Алгоритм Tertius використовує повний пошук із простору можливих правил зверху вниз згідно алгоритму A*. Алгоритм Tertius будує правила із значень пари атрибутів і ранжує їх відповідно до кількості підтвердженого правила в навчальній вибірці.

Правило складається з двох частин: тіла і голови. Тіло містить умови потрібні для формування правила, і може складатися з будь-якої кількості літералів. Голова містить подію, що підтверджують правила. На першому кроці навчання правило є порожньою множиною. З кожним наступним кроком правило уточняється шляхом додавання пари *атрибут-значення* в тому порядку, в якому вони з'являються в наборі даних. Після завершення, алгоритм підраховує, скільки разів правило підтверджено (тіло і голова правила істинні), і кількість випадків, коли правило дає помилковий результат (коли тіло вірно, але голова помилкова).

Якщо є *A* атрибутів з середнім значенням *V* і пошук правил здійснюється до *n* літералів, то кількість можливих правил рівна $(AV)^n$.

Для оцінки асоціативних правил використано два показники: підтримка та достовірність. Якщо в структурній одиниці даних зустрівся деякий набір елементів *X*, то на підставі цього можна зробити висновок про те, що інший набір елементів *Y* також має з'явитися в цій одиниці, тобто [8]:

$$X \Rightarrow Y \tag{4}$$

Правило $X \Rightarrow Y$ має підтримку (support) *s*, якщо *% s* транзакцій з множини *D*, містить множину $X \cup Y$, згідно виразу:

$$\text{supp}(X \Rightarrow Y) = \text{supp}(X \cup Y) = \frac{\text{count}(T : X \cup Y \subseteq T)}{\text{size}(D)} \cdot 100\% \tag{5}$$

Достовірність (confidence) правила показує, яка ймовірність того, що з множини *X* випливає *Y*. Правило $X \Rightarrow Y$ справедливе з достовірністю *c*, якщо *% c* транзакцій з множини *D*, що містять *X* та *Y*, згідно виразу:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \cdot 100\% \tag{6}$$

Експериментальні дослідження

Сьогодні на ринку програмних продуктів є декілька комерційних систем для інтелектуального аналізу даних, вартість яких коливається від $1000 до $10000 (PolyAnalyst, Weka, Orange Canvas, SVM^{light6}, Cognos і інші) [9,10].

Інструментом для інтелектуального аналізу даних обрано програмний засіб Weka.

В ході експериментальних досліджень було отримано базу даних цитологічних та гістологічних зображень молочної залози, яка містить більше 2000 зображень [11]. Вона вміщує такі нозологічні групи: проліферативна мастопатія, непроліферативна мастопатія, фіброаденома, папілярний рак та внутрішньопротоковий рак.

Опишемо основні кроки опрацювання вхідних даних в системі WEKA.

1. Формуємо файл формату .arff, який містить кількісні та якісні ознаки мікрооб'єктів. Структуру даного файлу приведено нижче:

relation byImages

@attribute Діагноз_Гістологія {'Інвазивний рак','Інфільтративний рак','Проліферативна мастопатія','Внутрішньопротоковий рак','Непроліферативна мастопатія','Фіброаденома (кістозна форма)','Листовидна фіброаденома','Фіброзно-кістозна мастопатія'}
@attribute 'Гістологічні характеристики_Властивість до апокринної секреції' {Висока,Низька,Відсутня,Помірна}
@attribute 'Гістологічні характеристики_Множинні вогнища проліферації' {'Епітеліальних клітин',Фібробластів}
@attribute 'Гістологічні характеристики_Некрози' {Поодинокі,'Не характерні',Множинні,Характерні,'Центральні некрози'}
@attribute 'Гістологічні характеристики_Підвищена проліферативна здатність внутрішньопротокового епітелію' {Наявна,Відсутня}

2. До сформованого файлу застосовуємо алгоритм класифікації J48. Результат роботи класифікаційної моделі навчальної та тестової вибірки представлено у таблиці 1.

Таблиця 1

Результат роботи класифікаційної моделі навчальної та тестової вибірки

Параметри	Навчальна вибірка	Тестова вибірка
Correctly Classified Instances	98.44 %	97.67 %
Incorrectly Classified Instances	1.56 %	2.33 %
Kappa statistic	0.98	0.96
Mean absolute error	0.0075	0.0097
Root mean squared error	0.06	0.07
Relative absolute error	4.58%	5.87 %
Root relative squared error	19.78 %	24.62 %
Total Number of Instances	257	257
Ignored Class Unknown Instances	2013	2013

3. Для пошуку асоціативних правил застосовуємо чотири алгоритми для отримання асоціативних правил, такі як Apriori (генерує набори кандидата і тестує, якщо вони є частими), Filtered Associator (виконує довільну асоціацію на вхідних даних, передану через довільний фільтр), Predective Apriori (виконує пошук зі збільшенням порогу підтримки для кращих *'N'* правил, що стосуються скоригованого значення достовірності на основі підтримки), Tertius (генерує і знаходить «цікаві» правила відповідно до міри їх підтвердження).

У таблиці 2 представлено порівняння експериментальних результатів згідно вибраних показників при знаходженні асоціативних правил цитологічних та гістологічних зображень. В якості показників було обрано підтримку, достовірність, кількість отриманих повних правил та час роботи алгоритму.

Таблиця 2

Порівняння результатів експерименту

		Показники			
		Підтримка	Достовірність	Кількість повних правил	Час виконання, сек.
Алгоритми	Apriori	0,15	0,9	10	0,45
	Filtered Associator	0,35	0,9	18	0,33
	Predective Apriori	0,92	1	33	18,25
	Tertius	0,98	1	54	25, 07

Проаналізуємо отримані результати. У таблиці 1 найбільш суттєві дані - це показники класифікації "Correctly Classified Instances" (97.67 % та "Incorrectly Classified Instances" (2.33 %).

Порівнюючи показник Correctly Classified Instances для тестового набору

97.67 % з цим же показником для навчального набору 98.44 %, можна констатувати, що точність моделі для двох різних наборів даних приблизно однакова. Це означає, що нові дані, які будуть використовуватися в цій моделі в майбутньому, не знизять точність її роботи.

З таблиці 2, відповідно до результатів експерименту слідує, що на великій кількості правил алгоритми Predective Apriori і Tertius працюють триваліше, але з більшою точністю. Тоді, коли результати Apriori і Filtered Associator безпосередньо залежать від кількості оброблених екземплярів даних [12].

В результаті дослідження було отримано такі приклади повних асоціативних правил передракових станів молочної залози на основі алгоритму Tertius:

/ 0,144536 0,000000 */ Клітина_Форма = Різна and Угрупування клітин або клітинні комплекси_Розташування = Багаточисельними округлими стурктурами ==> Діагноз_Гістологія = Проліферативна мастопатія*

*/*0,911736 0,001808 */ Діагноз_Цитологія = Папілярний рак ==> Угрупування клітин або клітинні комплекси_Розташування = Багаточисельними округлими стурктурами or Угрупування клітин або клітинні комплекси_Форма = Папілярні or Ядерце_Кількість = Одиничні дрібні ядерця*

Висновки

На основі кількісних та якісних ознак мікрооб'єктів гістологічних та цитологічних зображень проведено класифікацію за допомогою алгоритму J48. Результат роботи даного алгоритму показав близько 98% правильно класифіковних ознак мікрооб'єктів. Здійснено порівняння алгоритмів пошуку асоціативних правил і експериментально встановлено, що найкращим алгоритмом для пошуку асоціативних правил ознак мікрооб'єктів є алгоритм Tertius. На основі даного алгоритму отримано 78 повних правил діагностування ракових та передракових станів молочної залози.

1. Бюлетень національного канцер-реєстру України № 17, Київ – 2016.
2. *Барсегян А.А* Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP / А.А. Барсегян , М.С. Куприянов, В.В. Степаненко, И.И. Холод. 2-е изд., перераб. и доп. – СПБ.: БХВ-Петербург, 2007. – 384с.
3. *Ruijuan Hu* Medical Data Mining Based on Association Rules / Hu Ruijuan // Computer and Information Science. - Vol. 3, No. 4; November 2010, p. 104-108.
4. *Jaimini Majali* Data Mining Techniques For Diagnosis And Prognosis Of Breast Cancer / Jaimini Majali, Rishikesh Niranjan, Vinamra Phatak, Omkar Tadakhe // (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5) , 2014, 6487-6490.
5.*Lekha, A.* Fuzzy Association Rule Mining / Lekha, A., C.V. Srikrishna, Viji Vinod // Journal of Computer Science 2015, 11 (1): 71.74.
6. *Shweta Kharya* Using data mining techniques for diagnosis and prognosis of cancer disease / Shweta Kharya // International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012, p. 55-66.

7.*Jyotismita Talukdar* Detection of Breast Cancer using Data Mining Tool (WEKA) / Jyotismita Talukdar, Dr. Sanjib Kr. Kalita // International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015, p. 1124-1128.
8.*Дюк В.А*. Data Mining: учебный курс / В.А. Дюк, А.П. Самойленко // СПб.: Питер, 2001. - 368 с.
9. *Кречетов Н*. Продукты для интеллектуального анализа данных / Н. Кречетов // Рынок программныхсредств, N14-15_97, с. 32-39.
10. Data Mining, Web Mining, Text Mining, and Knowledge Discovery (http://www.kdnuggets.com).
11. *Березький О.М.* База даних цитологічних та гістологічних зображень ауто- та ксеногенних тканин / О.М. Березький, Г.М. Мельник, С.О. Вербовий, Т.В. Дацко // Науковий вісник національного лісотехнічного університету України: збірник науково-технічних праць. – Львів: РВВ НЛТУ України. – 2014. – Вип. 24.10. – С.338-345.
12. *Вербовий С.О.* Методи пошуку асоціативних правил в базі даних біомедичних зображень / С.О. Вербовий, В.С. Зубко // Сучасні комп'ютерні інформаційні технології: Матеріали VI Всеукраїнської школи-семінару молодих вчених і студентів ACIT'2016, 20-21 травня, 2016р.: матеріали. – Тернопіль: ТНЕУ, 2016. - С.61-63.

Поступила 10.10.2016р.

УДК 004.4

Ю.М. Батько, м.Тернопіль

АНАЛІЗ КОНТУРНИХ ОЗНАК МІКРООБ'ЄКТІВ НА ЦИФРОВИХ КОЛЬОРОВИХ БІОМЕДИЧНИХ ЗОБРАЖЕННЯХ

Abstract. The analysis of contour features of objects in biomedical images. We proposed criteria for assessing the importance of features based on its use in image processing systems. The estimation of contour and selected the most used features.

Актуальність

Зі стрімким розвитком інформаційних технологій та шкідливістю умов роботи все найбільшого більшої ваги набуває медична електронна апаратура діагностичного, терапевтичного та дослідницького призначення [1-2]. Все частіше при постановці діагнозу застосовуються сучасні медичних технології: електро-, магніто-, рентгенівська, ультразвукова, комп'ютерна томографія тощо. При проведенні комплексного дослідження медичного препарату в першу чергу проводиться дослідження його форми [3-5], що базується на аналізі його контурних ознак. Множина контурних ознак використовується для подальшої класифікації виділених мікрооб'єктів та визначення додаткових ознак мікрооб'єктів наприклад на основі аналізу за допомогою апарату нечітких множин. Медичні програмно-апаратні комплекси дозволяють перетворити рутинну технологію мікроскопічного

