УДК 616.24-006

Вопросы онкологии, 2017. Том 63, № 2

А.А. Барчук[2], М.Д. Подольский[1], А.М. Беляев[2,5], И.Ю. Коцюба[1], Н.Ф. Гусарова[1], В.А. Трофимов[1], П.Д. Виноградов[1], В.С. Гайдуков[1], В.И. Кузнецов[3], В.М. Мерабишвили[2], А.С. Барчук[2,5], А.В. Атрощенко[4], М.В. Харитонов[4], А.В. Нефедова[2], Ю.И. Комаров[2], А.И. Арсеньев[2,5], С.В. Канаев [2], С.А.Тараканов[1]

Автоматизированная диагностика в популяционном скрининге рака легкого

[1]ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»,
[2]ФГБУ «НИИ онкологии им. Н.Н. Петрова» Минздрава России,
[3]ООО «Конструкторское бюро современных технологий Санкт-Петербургского Государственного Университета ИТМО»,
[4]ГБУЗ «Онкологический диспансер Московского района»,
[5]ГБОУ ВПО «СЗГМУ им. И.И. Мечникова» Минздрава России,
Санкт-Петербург

В современной онкологии врачи постоянно сталкиваются с необходимостью обработки большого потока гетерогенных данных диагностических исследований. Возможные ошибки в определении характера и степени распространения опухолевого процесса неизбежно снижают эффективность лечения и повышают неоправданные затраты на него. Для уменьшения нагрузки на врачей в настоящее время разрабатываются различные компьютеризированные решения, основанные на методах, или алгоритмах, машинного обучения. Была осуществлена попытка оценить эффективность тринадцати методов машинного обучения в задачах классификации образцов патологической ткани при злокачественных процессах органов грудной полости на основе уровней экспрессии генов. Для предварительного исследования был выбран доступный и открытый набор данных молекулярно-генетического состава групп опухолей двух типов: аденокарциномы легкого и мезотелиомы. Эффективность методов машинного обучения оценивалась по коэффициенту корреляции Мэтьюса и площадью под характеристической (ROC) кривой. Наилучшую эффективность продемонстрировали два метода: байесовская логистическая регрессия и дискриминационный полиномиальный наивный байесовский классификатор. Все методы были достаточно эффективны при автоматической дискриминации двух типов опухолей, а результаты подтверждали применимость методов машинного обучения при решении задач морфологической классификации опухолей. В дальнейшем будет проведен аналогичный анализ диагностической ценности методов для других злокачественных новообразований, дифференциальный морфологический диагноз при которых более сложен. Использование данных методик возможно и при других диагностических исследованиях, в том числе для анализа изображений компьютерной томографии при дифференциальной диагностике узлов легкого.

Ключевые слова: автоматизированная диагностика, рак легких, ROC кривая, большие данные, классификаторы, машинное обучение

Введение

Рак легкого остается одной из главных проблем противораковой борьбы. Ежегодно в мире раком легкого заболевает не менее 1,5-2,0 миллионов человек [10]. В структуре онкологической заболеваемости мужского населения многих стран рак легкого занимает 1 место [10].

Ежегодно в России регистрируется не менее 60 тыс. (60351–2015 г.) новых случаев рака легкого. За последние 10 лет зарегистрирован рост абсолютных чисел на 4,7%. Стандартизованные показатели заболеваемости раком легкого среди мужского населения снизились на 16,2%, среди женского возросли на 11,2%. Уровни грубых и стандартизованных показателей заболеваемости мужчин и женщин в среднем по России и Северо-Западном Федеральном округе близки (табл. 1) [1,2].

Очевидно, что отдалённые результаты лечения онкологических заболеваний непосредственно зависят от своевременности диагностики и начала реализации противоопухолевых мероприятий. При этом от адекватного и точного определения морфологического типа злокачественного новообразования (ЗН) напрямую зависят эффективность лечения, предиктивные и прогностические факторы. Например, при местнораспространенной злокачественной мезотелиоме

Таблица 1. Динамика абсолютных чисел, «грубых» и стандартизованных показателей заболеваемости населения России раком легкого [1]

	2005	2010	2014	2015
Мужчины				
Абс. числа	47884	46407	46224	48139
На 100 тыс. мужского населения «грубый» показатель	72,75	70,70	68,29	70,97
Стандартизованный показатель (мировой стандарт)	57,62	53,97	48,78	49,88
Женщины				
Абс. числа	9746	10578	11461	12212
На 100 тыс. женского населения «грубый» показатель	12,80	13,87	14,61	15,54
Стандартизованный показатель (мировой стандарт)	6,68	7,13	7,30	7,77

плевры (ЗМП) оптимальным является алгоритм мультимодального лечения с экстраплевральной пневмонэктомией и химиолучевой терапией, в то время как в лечении первичной аденокарциномы легкого с поражением плевры возможны варианты с применением таргетной терапии [27].

На современном этапе морфологический диагноз все больше опирается на молекулярно-генетический состав опухоли, определяемый широким спектром методик [23]. В частности, используется гибридизация на олигонуклеотидном микрочипе [6, 17] с последующей алгоритмической обработкой, что позволяет оценить уровень экспрессии генов в исследуемом образце. Поскольку изменения профилей экспрессии в сравниваемых образцах, как правило, невелики, данный метод в чистом виде может значительно усложнить дифференциальную диагностику опухолей. Для решения этой проблемы может использоваться двухстадийная алгоритмическая процедура – 1) выделение сравнительно небольшой группы генов, наиболее репрезентативных для определения морфологического типа опухоли, и 2) собственно классификация с использованием именно этой группы [17]. Но в последние годы, на фоне эволюции вычислительной техники появилась возможность использования алгоритмов, выполняющих одновременную статистическую обработку уровней экспрессии всех генов образца. Такой подход носит название Machine Learning (ML, «машинное обучение»). В исследовании, опубликованном в 2006 г. J.A. Cruz, D.S. Wishart [13], показано, что алгоритмы ML могут быть использованы для существенного (15-25%) улучшения точности постановки первичного диагноза, случаев рецидива и установки посмертного диагноза при аутопсии.

Для того, чтобы алгоритмы ML начали предугадывать тип ЗН, их необходимо подготовить (натренировать) с помощью обучающего набора данных: 1) предоставить алгоритму открытые данные образцов с заранее известными и верифицированными различными методами диагнозами; 2) проверить результат на опытном образце с неизвестными данными; 3) уточнить объем ошибок; 4) разработать способы предотвращения ошибок.

Целью настоящего исследования является экспериментальная проверка эффективности тринадцати алгоритмов машинного обучения (ML) при решении задач определения морфологического типа опухоли на основе данных об уровне экспрессии генов.

Материалы и методы

Основная задача исследования было изучить точность постановки дифференциального диагноза мезотелиомы плевры или аденокарциномы легкого при использовании алгоритмов ML на основе экспрессии генов. Исследование эффективности тринадцати алгоритмов ML было проведено на основе имеющейся в открытом доступе базы данных экспрессии генов образцов опухолей с известным диагнозом [34] от Harvard Medical School и Brigham and Women's Hospital, Бостон, Массачусетс, США. Образцы были собраны и заморожены во время хирургических операций пациентов в период 1993 — 2001 гг. База данных содержит результаты оценки экспрессии генов 181 образцов — среди них 31 образец злокачественной мезотелиомы плевры и 150 образцов аденокарциномы легкого. Образцы разбиты на два набора: обучающий (по шестнадцать образцов каждого типа опухоли) и тестовый (оставшиеся 149 образцов). Каждый образец описан уровнями экспрессии 12 533 генов. Подробные данные о методике определения экспрессии генов представлены в оригинальных работах создателей базы данных [6, 17].

Отбор производился среди следующих алгоритмов ML: 1) Байесовская логистическая регрессия (Bayesian logistic regression) [15]; 2) дерево решений (Decision trees) [30]; 3) дискриминационный полиномиальный наивный байесовский классификатор (Discriminative Multinomial Naive Bayes classifier) [3]; 4) конъюнктивный метод (Conjunctive rule) [21]; 5) линейная логистическая регрессионная модель (Logistic regression model trees) [22]; 6) метод ближайших соседей (k-Nearest Neighbors) [24]; 7) метод локального взвешивания (Locally weighted learning) [20]; 8) перцептрон (Voted perceptron – лат. восприятие) [16]; 9) правило повторяющейся инкрементной обрезки ветвей для компенсации погрешности (RIPPER rule) [12]; 10) Решающий стамп (Decision Stump) [19]; 11) сеть радиально-базисных функций (Radial basis function network) [25]; 12) случайные распределения (Random forests) [9]; 13) усиление (Adaboost M1) [32].

В результате для каждого метода машинного обучения строится характеристическая (ROC) кривая, которая показывает зависимость доли верных положительных классификаций (чувствительность) от доли ложных положительных классификаций (специфичность).

Как один из критериев оценки эффективности методов машинного обучения был выбран показатель AUC (площадь под ROC кривой). Ввиду того, что диапазон допустимых значений по обеим осям лежит на замкнутом интервале: [0; 1], то максимальным значением площади под графиком ROC кривой может быть 1, что соответствует идеальному классифицированию. Чем выше показатель AUC, тем качественнее классификатор. При этом значение меньше 0,5 является неприемлемым, а результат является случайным. В данной работе проверялась статистическая гипотеза об отличии полученного значения от 0,5 путем оценки 95% доверительного интервала значения AUC каждого метода.

Вторым критерием оценки является коэффициент корреляции Мэтьюса (Matthews Correlation Coefficient —

MCC). Данный показатель, основываясь на количестве истинных положительных, ложноположительных, истинных отрицательных и ложно-отрицательных заключений, показывает уровень возможностей дифференциальной диагностики выбранного метода по шкале от «-1» до «1»; где «1» соответствует идеальному методу, «0» – случайному значению, а «-1» − постоянным ошибкам [4].

Результаты

В целом все методы машинного обучения показали значения AUC больше 0,5 (таб. 1, рис. 1). Набор данных был обработан с помощью пакета машинного обучения Weka (версия 3.6.12 stable) [18] для языка программирования Java (версия SE 8 Update 40). Пакет машинного обучения, в котором представлены реализации тринадцати исследуемых алгоритмов, и библиотека для визуализации графиков были интегрированы в конечную программу с помощью приложения, написанного с использованием системы сборки «gradle» (версия 2.3 stable). Авторами статьи было разработано приложение, которое считывает информацию из набора данных, анализирует его с использованием выбранных алгоритмов и выводит ROC кривые на общих графиках.

Указанные значения AUC и MCC получены с учетом 95% доверительного интервала, значение которого обусловлено непараметрической оценкой для использованного в качестве несмещенного показателя вероятности ошибки

Таблица 2. Характеристики методов машинного обучения

№ п/п	Название метода	AUC [CI 95%]	MCC [CI 95%]
1	Байесовская логистическая регрессия	0,931 [0,924–0,938]	0,923 [0,918–0,928]
2	Дискриминационный полиномиальный наивный байесовский классификатор	0,993 [0,988–0,998]	0,882 [0,857–0,907]
3	Усиление	0,940 [0,931–0,949]	0,844 [0,801–0,887]
4	Линейная логистическая регрессионная модель	0,991 [0,981–1,001]	0,840 [0,788–0,892]
5	Перцептрон	0,923 [0,919–0,927]	0,801 [0,771–0,831]
6	Сеть радиально-базисных функций	0,944 [0,937–0,951]	0,803 [0,776–0,830]
7	Метод локального взвешивания	0,860 [0,809–0,911]	0,681 [0,609–0,753]
8	Дерево решений	0,860 [0,710–1,010]	0,674 [0,641–0,707]
9	Метод ближайших соседей	0,721 [0,642–0,800]	0,662 [0,643–0,681]
10	Решающий стамп	0,823 [0,771–0,875]	0,651 [0,634–0,668]
11	Правило повторяющейся инкрементной обрезки ветвей для компенсации погрешности	0,822 [0,643–1,001]	0,653 [0,611–0,695]
12	Конъюнктивный метод	0,790 [0,680–0,900]	0,623 [0,601–0,645]
13	Случайные распределения	0,994 [0,985–1,003]	0,551 [0,459–0,643]

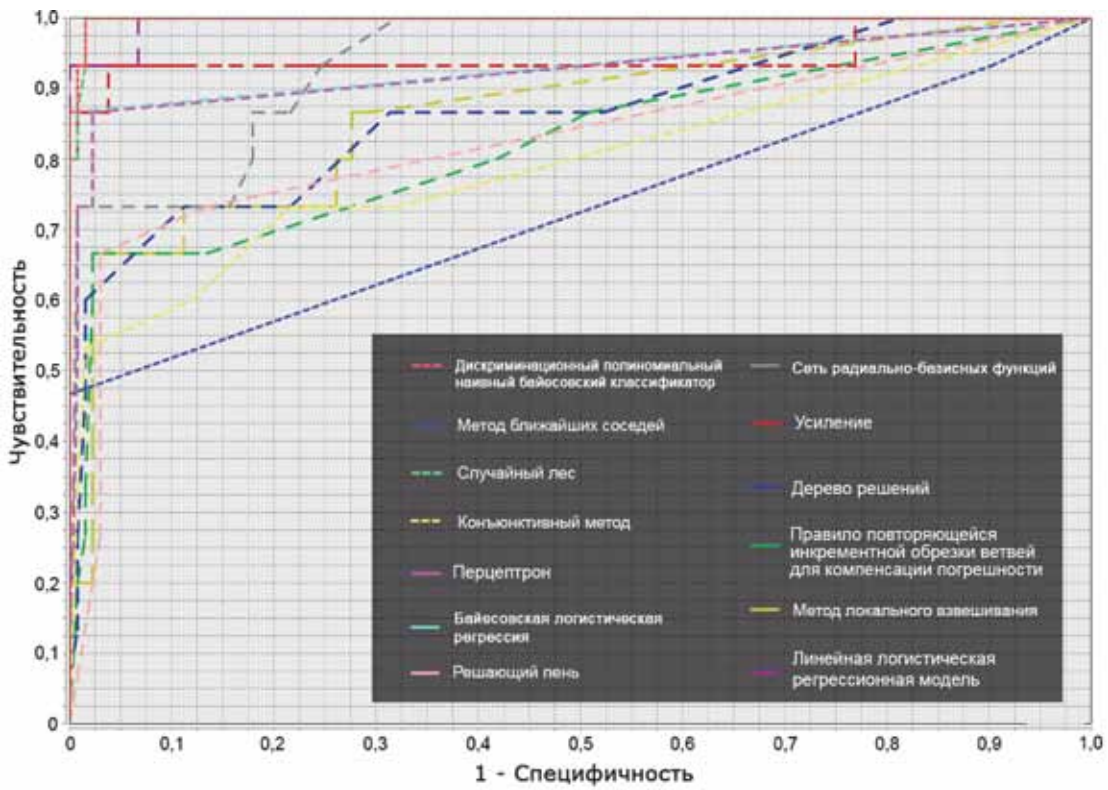

Рис. 1. ROC кривые для тринадцати методов машинного обучения

скользящего контроля, с осуществлением 20 операций разбиения. Среди тринадцати исследованных методов машинного обучения наилучшие результаты показали Байесовская логистическая регрессия (AUC 0,931 и MCC 0,923) и дискриминационный полиномиальный наивный байесовский классификатор (AUC 0,993 и MCC 0,882). За ними по шкале эффективности расположились с аналогичными показателями MCC линейная логистическая регрессионная модель (AUC 0,991 и MCC 0,840) и усиление (AUC 0,940 и MCC 0,844).

Обсуждение

В нашей работе для оценки качества дифференциальной диагностики мы использования площадь по характеристической кривой (AUC) и коэффициент корреляции Мэтьюса (MCC). Результаты, показавшие высокую эффективность Байесовской логистической регрессии, согласуются с результатами исследования R.G. Ramani и S.G. Jacob (2013г.), в котором Байесовские классификаторы продемонстрировали наилучшую точность при классификации мелкоклеточного и немелкоклеточного рака легких [29]. В предыдущем исследовании авторов настоящей статьи оценивалась другие алгоритмы классификации [28], при этом «метод опорных векторов» [14] показал наилучшую эффективность (AUC 0,99 и MCC 0,97) на том же самом наборе данных.

Очевидно, эффективность методов машинного обучения во многом зависит от конкретных обучающих и обучаемых выборок данных. При этом для дифференцировки данных используются и другие методы математической обработки, однако эффективность их несколько ниже. Например, статистическая обработка с использованием алгоритмов линейного дискриминантного анализа информации об экспрессии генов позволила добиться диагностической точность лишь в 43–70%. Некоторые авторы [17] вместо анализа коэффициентов AUC и MCC в качестве критерия оценки используют отношение числа верно классифицированных образцов к общему числу исследованных образцов. Данный подход заведомо несет в себе статистическую неточность, так как не позволяет оценить число ложноположительных результатов на фоне большого числа верно диагностированных заболеваний (истиноположительных результатов).

Описанные в данной работе методы могут быть использованы при обработке большого объема данных и выборок, описывающих различные характеристики заболевания, в частности, в условиях популяционного скрининга онкологических заболеваний [5, 33]. В частности, целесообразность использования методов машинного обучения, как вспомогательных инструментов, облегчающих процесс принятия диагностических решений в настоящее время изучается нами при анализе изображений полученных при компьютерной томографии, эндоскопических изображений и данных анализа выдыхаемого воздуха [26, 31].

Анализ литературы последних лет также демонстрирует планомерное увеличение количества публикаций, посвященных анализу различных медицинских статистических показателей и изображений с помощью компьютеризованных решений [7, 8], что подтверждает общую тенденцию применимости методов машинного обучения в задачах дифференциальной диагностики онкологических заболеваний при наличии большого объема данных. Несомненно это поможет оптимизировать трудовые затраты, уменьшить нагрузку на узких специалистов, снизить стоимость организации программ скрининга онкологических заболеваний и повысить их качество.

Настоящая работа поддержана Минобрнауки России в рамках проекта RFMEFI57814X0008

ЛИТЕРАТУРА

1. Злокачественные новообразования в России в 2015 году (заболеваемость и смертность) / Под ред. А.Д. Каприна, В.В. Старинского, Г.В. Петровой. – М.: МНИОИ им. П.А. Герцена- филиал ФГБУ «НМИРЦ» Минздрава России, 2017. – 250 с.
2. Злокачественные новообразования в Санкт-Петербурге и других административных территориях Северо-Западного федерального округа России (заболеваемость, смертность, контингенты, выживаемость, больных). Экспресс-информация. Второй выпуск / под ред. А.М. Беляева, Г.М. Манихаса, В.М. Мерабишвили. – СПб.: Т8 Издательские технологии, 2016. – 208 с.
3. Aref A., Tran T. Using ensemble of Bayesian classifying algorithms for medical systematic reviews // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). – 2014. – Vol. 8436 LNAI. – P. 263-268.
4. Baldi P., Brunak S., Chauvin Y. et al. Assessing the accuracy of prediction algorithms for classification: an overview // Bioinformatics. – 2000. – Vol. 16. – № 5. – P. 412–424.
5. Barchuk A.A., Podolsky M.D., Gaidukov V.S. et al. Intelligent distributed system of population cancer screening // Voprosy onkologii. – 2015. – Vol. 61. – № 4. – P. 517–522.
6. Bhattacharjee A., Richards W.G., Staunton J. et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses // PNAS. – 2001. – Vol. 98. – № 24. – P. 13790–13795.

7. Brady S.M., Highnam R., Irving B., Schnabel J.A. Oncological image analysis // Medical Image Analysis. – 2016. – Vol. 33. – P. 7–12.
8. De Bruijne M. Machine learning approaches in medical image analysis: From detection to diagnosis // Medical Image Analysis. – 2016. – Vol. 33. – P. 94–97.
9. Cai Z., Xu D., Zhang Q. et al. Classification of lung cancer using ensemble-based feature selection and machine learning methods // Mol. BioSyst. – 2015. – Vol. 11. – № 3. – P. 791–800.
10. Cancer incidence in Five Continents Vol. X / Ed. D. Forman. F. Btay, D.H. Brewster, C. Gombe Mbalawa, B. Kohler, M. Pineros, E. Steliarova-Foucher, R. Swaminathan and J. Ferlay. IARC Scientific Publication №164. – Lyon, 2014. – 1365 p.
11. Cheng P., Cheng Y., Li Y. et al. Comparison of the Gene Expression Profiles Between Smokers With and Without Lung Cancer Using RNA-Seq // Asian Pacific Journal of Cancer Prevention. – 2012. – Vol. 13. – № 8. – P. 3605–3609.
12. Cohen W.W. Fast Effective Rule Induction // Proceedings of the Twelfth International Conference on Machine Learning. California. – 1995. – P. 115–123.
13. Cruz J.A., Wishart D.S. Applications of machine learning in cancer prediction and prognosis // Cancer Inform. – 2006. – Vol. 2. – P. 59-77.
14. Devi A.V., Devaraj D., Venkatesulu M. Gene expression data classification using Support Vector Machine and mutual information-based gene selection // Procedia Computer Science. – 2014. – Vol. 47. – P. 13–21.
15. Dumouchel W. Multivariate bayesian logistic regression for analysis of clinical study safety issues // Statistical Science. – 2012. – Vol. 27. – № 3. – P. 319–339.
16. Freund Y., Schapire R.E. Large margin classification using the perceptron algorithm // Machine Learning. – 1999. – Vol. 37. – № 3. – P. 277–296.
17. Gordon G.J., Jensen R.V., Hsiao L.-L. et al. Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma // Cancer Res. – 2002. – Vol. 62. – № 17. – P. 4963-4967.
18. Hall M., Frank E., Holmes G. et al. The WEKA data mining software: an update // SIGKDD Explorations. – 2009. – Vol. 11. – № 1. – P. 10–18.
19. Hosseinzadeh F., Ebrahimi M., Goliaei B., Shamabadi N. Classification of lung cancer tumors based on structural and physicochemical properties of proteins by bioinformatics models // PLoS ONE. – 2012. – Vol. 7. – № 7.
20. Jiang L., Cai Z., Zhang H., Wang D. Naive Bayes text classifiers: A locally weighted learning approach // Journal of Experimental and Theoretical Artificial Intelligence. – 2013. – Vol. 25. – № 2. – P. 273–286.
21. Kohli R., Krishnamurti R., Jedidi K. Subset-conjunctive rules for breast cancer diagnosis // Discrete Applied Mathematics. – 2006. – Vol. 154. – № 7. – P. 1100–1112.
22. Landwehr N., Hall M., Frank E. Logistic model trees // Machine Learning. – 2005. – Vol. 59. – № 1–2. – P. 161–205.
23. Liu M., Pan H., Zhang F. et al. Screening of Differentially Expressed Genes among Various TNM Stages of Lung Adenocarcinoma by Genomewide Gene Expression Profile Analysis // Asian Pacific Journal of Cancer Prevention. – 2013. – Vol. 14. – № 11. – P. 6281–6286.
24. Murphy K., Van G., Schilham A.M.R. et al. A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification // Medical Image Analysis. – 2009. – Vol. 13. – № 5. – P. 757–770.
25. Naveen N., Ravi V., Rao C.R. Rule extraction from differential evolution trained radial basis function network using genetic algorithms. – 2009. – P. 152–157.
26. Orozco H.M., Villegas O.O.V., Sánchez V.G.C. et al. Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine // BioMedical Engineering OnLine. – 2015. – Vol. 14. – № 1. – P. 9.
27. Pass H.I. Malignant Pleural Mesothelioma: Surgical Roles and Novel Therapies // Clinical Lung Cancer. – 2001. – Vol. 3. – № 2. – P. 102–117.
28. Podolsky M.D., Barchuk A.A., Kuznetcov V.I. et al. Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels // Asian Pacific Journal of Cancer Prevention. – 2016. – Vol. 17. – № 2. – P. 835–838.
29. Ramani R.G., Jacob S.G. Improved Classification of Lung Cancer Tumors Based on Structural and Physicochemical Properties of Proteins Using Data Mining Models // PLOS ONE. – 2013. – Vol. 8. – № 3. – P. e58772.
30. Rouhi R., Jafari M. Classification of benign and malignant breast tumors based on hybrid level set segmentation // Expert Systems with Applications. – 2016. – Vol. 46. – P. 45-59.
31. Sun T., Wang J., Li X. et al. Comparative evaluation of support vector machines for computer aided diagnosis of lung cancer in CT based on a multi-dimensional data set // Computer Methods and Programs in Biomedicine. – 2013. – Vol. 111. – № 2. – P. 519-524.
32. Wang C.-W., Yu C.-P. Automated morphological classification of lung cancer subtypes using H&E tissue images // Machine Vision and Applications. – 2013. – Vol. 24. – № 7. – P. 1383–1391.
33. Yoo C., Ramirez L., Liuzzi J. Big data analysis using modern statistical and machine learning methods in medicine // International Neurourology Journal. – 2014. – Vol. 18. – № 2. – P. 50–57.
34. Data Repository — Lung Cancer [Electronic resource]. URL: http://datam.i2r.a-star.edu.sg/datasets/krbd/LungCancer/LungCancer-Harvard2.html (accessed: 23.09.2016).

Поступила в редакцию 16.12.2016 г.

A.A. Barchuk[2], M.D. Podolsky[1], A.M. Belyaev[2,5], I.Yu. Kotsyuba[1], N.F. Gusarova[1], V.A. Trofimov[1], P.D. Vinogradov[1], V.S. Gaidukov[1], V.I. Kuznetsov[3], V.M. Merabishvili[2], A.S. Barchuk[2,5], A.V. Atroshchenko[4], M.V. Kharitonov[4], A.V. Nefedova[2], Yu.I. Komarov [2], A.I. Arseniev[2,5], S.V. Kanaev[2], S.A. Tarakanov[1]

Automated diagnosis in a population-based screening for lung cancer

[1]Saint Petersburg National Research University of Information Technologies, Mechanics and Optics
[2]N.N. Petrov Research Institute of Oncology
[3]Limited liability company «Saint-Petersburg State University ITMO Design Bureau of Modern Technologies»
[4]Oncology Dispensary of the Moscow District
[5]I.I. Mechnikov North-West State Medical University St. Petersburg

Oncologists nowadays are faced with big amount of heterogeneous medical data of diagnostic studies. Possible errors in determining the nature and extent of spread the tumor process will inevitably reduce the effectiveness of treatment and increase the unnecessary costs to it. To reduce the burden on clinicians, various computer-aided solutions based on machine learning algorithms are being developed. We made an attempt to evaluate effectiveness of thirteen machine learning algorithms in the tasks of classification of pathologic tissue samples in cancerous thorax based on gene expression levels. For a preliminary study we used open data set of molecular genetics composition of lung adenocarcinoma and pleural mesothelioma. Effectiveness of machine learning algorithms was evaluated by Matthews correlation coefficient and Area Under ROC Curve. Best results were showed by two methods: Bayesian logistic regression and Discriminative Multinomial Naive Bayes classifier. Nevertheless, all of the methods were effective at automatic discrimination of two types of cancer. That proves machine learning algorithms are applicable in lung cancer classification. In the future studies it will be carried out a similar analysis of the diagnostic value of methods for other malignancies with more complex differential morphological diagnosis. Similar methods can be applied to other diagnostic studies including computerized tomography image analysis in the differential diagnosis of lung nodules.

Key words: computer-aided diagnosis, lung cancer, ROC curve, big data, classifiers, machine learning