КЛАССИФИКАЦИЯ ОБЪЕКТОВ ЛЮМИНЕСЦЕНТНЫХ ИЗОБРАЖЕНИЙ

Е. В. Лисица, Н. Н. Яцков, В. В. Апанасович

Белорусский государственный университет
Минск, Беларусь
E-mail: ylisitsa@gmail.com

Исследуются методы классификации с использованием смоделированных данных на устойчивость алгоритмов классификации к размеру обучающей выборки и уровня шума.

Ключевые слова: методы классификации, анализ данных, моделирование, люминесцентная микроскопия.

Введение

Методы обработки и анализа данных люминесцентной микроскопии получили широкое распространение в цитологических методах диагностики [1; 2; 3]. Однако анализ люминесцентных изображений, получаемых в результате исследования, проводится только интегрально, что не позволяет достигнуть необходимой точности исследования. Внедрение автоматических методов анализа усложняется различными особенностями проведения эксперимента, такими как метод изготовления образца, способ регистрации изображения, методы обработки изображения.

Подобное разнообразие независимых факторов, проявляющихся одновременно в одном эксперименте, не позволяет отобрать наиболее устойчивые алгоритмы к определенным, самым часто-встречаемым воздействиям. Однако данная проблема может быть успешно решена, используя имитационную модель. Имитационное моделирование позволяет создавать упрощенные наборы данных, максимально приближенные к экспериментальным результатам, и нигилировать наименее важные эксприментальные признаки, усилить влияние наиболее интересующих факторов. Второе преимущество в использовании имитационной модели – объективная заранее известная информация, независимая от экспертов, о структуре данных. Использование модели позволяет задать точное количество клеток и их тип, а также их параметры. Третий положительный аргумент – компьютерная имитация позволяет исследовать условия, которые не достижимы в эксперименте.

Методы статистической обработки информации можно использовать в цитологии для решения задачи классификации, когда разбиение объектов на группы проходит с учителем (тренеровочной заранее известной выборкой экспериментальных данных), и кластеризации (обучение без учителя). Параметрические методы классификации используют параметрические семейства зависимостей (разделяющей поверхности в распознавании образов, плотности распределения вероятности, модели объекта) и свойства объектов [12, 13]. Методы дискриминантного анализа успешно используются для изучения биомаркеров выраженности генов [14] и классификации маммограмм [15]. Применение наивного байесовского классификатора характерно для решения задачи степени ракового заболевания [16]. Преимущества использования искусственных нейронных сетей (ИНС) для решения задачи статистической об-

работки данных перечислены ниже: а) решение задач при неизвестных закономерностях; б) устойчивость к шумам во входных данных; в) приспосабливаемость к изменениям окружающей среды; г) потенциальное сверхвысокое быстродействие [13].

Теоретическое описание

По результатам теоретических исследований для практической реализации отобраны следующие методы параметрической и непараметрической классификации: алгоритмы линейного и квадратичного дискриминантного анализа, наивного байесовский классификатора, метод *к*-средних, крата и слой Кохонена:

1) Дискриминантный анализ (ДА). В ДА рассматривается предположение о нормальном распределении признаков объектов. В линейном ДА ищется линейная комбинация признаков, позволяющая построить классифицирующее правило для объектов. Квадратичный ДА работает аналогично, с той лишь разницей, что предварительно производится преобразование пространства и в новом пространстве ищется линейная комбинация признаков. В исходном же пространстве эта комбинация будет нелинейной (квадратичной) [17,18].

2) Наивный байесовский классификатор (НБК). НБК использует теорему Байеса, применяя «наивное» предположение о независимости признаков. Далее в качестве плотностей распределения подставляются предполагаемые распределения классов и производится оценка параметров этих распределений по методу максимального правдоподобия, т. е. ищутся параметры, максимизирующие вероятность появления наблюдаемых данных [19].

3) Метод *k* ближайших соседей – простейший метрический классификатор, основанный на оценивании сходства объектов [20]. Пусть задана обучающая выборка $T = \{(n_1,y_1), \ldots ,(n_m, y_m)\}$. Допустим, что на множестве объектов задана функция расстояния $\rho(n, n')$. Эта функция должна быть достаточно точной моделью сходства объектов. Чем больше значение этой функции, тем менее схожими являются два объекта *n* и *n'*.

Для произвольного объекта *u* расположим объекты обучающей выборки n_i в порядке возрастания расстояний до *u*: $\rho(u, n_{1;u}) \le \rho(u, n_{2;u}) \le \ldots \le \rho(u, n_{m;u})$, где через $n_{i;u}$ обозначается тот объект обучающей выборки, который является *i*-м соседом объекта *u*. Аналогичное обозначение введем и для ответа на *i*-м соседе: $y_{i;u}$. Таким образом, произвольный объект *u* порождает свою перенумерацию выборки. В этом случае принадлежность объекта к классу определяется по формуле

$$a(u) = \arg\max_{y \in Y} \sum_{i=1}^{m} \left[n_{i;u} = y \right] \omega(i,u),$$

где $\omega(i, u)$ — заданная весовая функция, которая оценивает степень важности *i*-го соседа для классификации объекта *u* [20].

4) Искусственная нейронная сеть – это математическая модель, построенная по принципу организации и функционирования биологических нейронных сетей, т. е. ИНС представляют собой соединенную систему простых процессов [13]. В работе исследованы следующие виды нейронных сетей: 1. Слой Кохонена – представляет собой один слой адаптивных линейных сумматоров, работающих по принципу WTA – Winner Takes All, или «Победитель забирает все». 2. Карта Кохонена.

Постановка эксперимента

Для исследования устойчивости методов классификации был сгенерирован набор из пяти независимых изображений с различным количеством объектов-клеток на каждом из них.

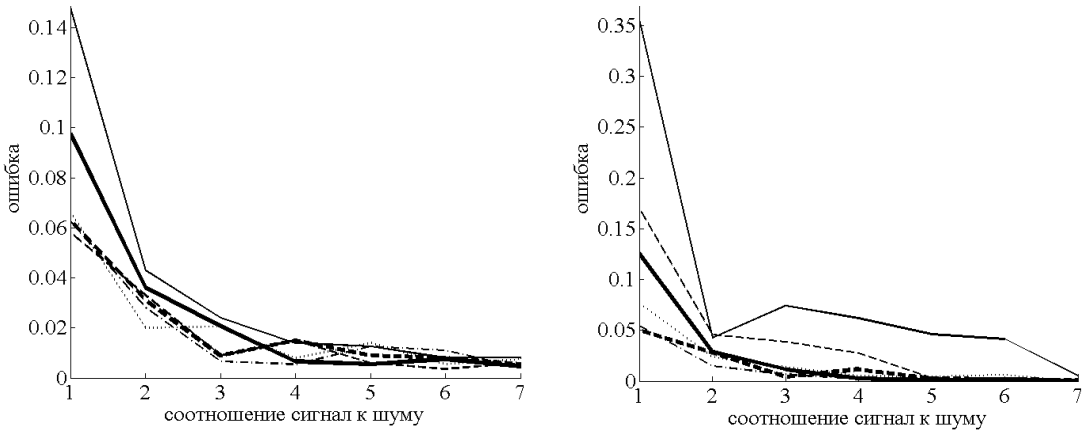
Для смоделированных изображений с помощью методов сегментации были построены маски ядер и цитоплазм. С каждого ядра были сняты характеристики, подробно описанные в работе [17]. Для изучения воздействия шума к данным добавлялся белый шум, со среднеквадратичным отклонением. [7]: формула (2). Соотношение сигнала к шуму варьировалось от 0,9 до 10. Для оценки качества работы методов использовались матрицы ошибок (1, 2 рода и общая ошибка) для различных уровней шума и различных размеров тренировочных выборок, где ошибка первого рода характеризует, что «здоровая» клетка была отнесена к классу больных клеток, а ошибка второго рода показывает, когда больные клетки были проклассифицированы как здоровые [21].

Результаты

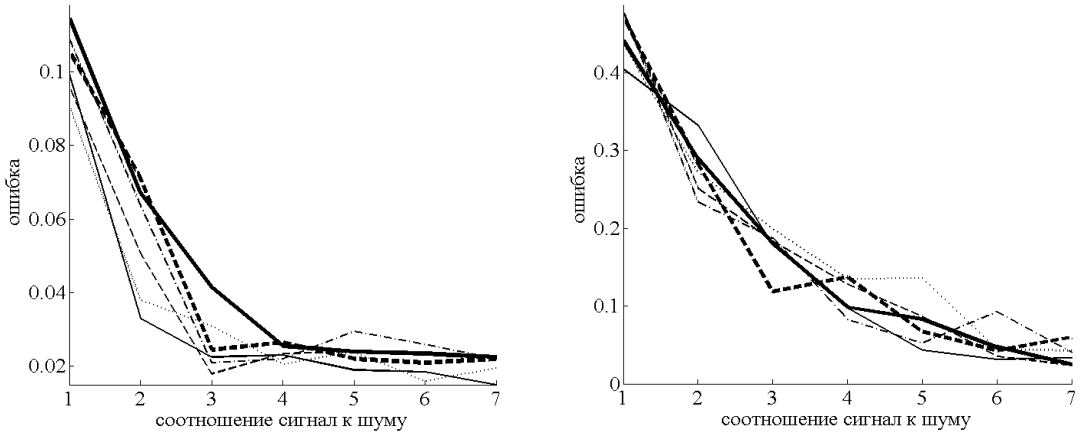
Проведен сравнительный анализ методов классификации на смоделированных данных. Результаты анализа представлены на рис. 1. Исследована устойчивость методов классификации к размеру обучающей выборки при различном уровне шума. На рис. 1*а* показана общая ошибка классификации для линейного дискриминантного классификатора.

В методе линейного дискриминантного анализа качество классификации улучшается при уменьшении шума и увеличении объема размера обучающей выборки. Наибольшая общая ошибка составляет около 20 %. При высоком уровне шума общая ошибка классификации для эталонной выборки 300 элементов в два-три раза меньше, чем для выборки из 150 элементов.

На рис. 1, *б* показана общая ошибка классификации смоделированных данных для квадратичного дискриминантного классификатора. Для квадратичного классификатора характерна высокая ошибка классификации при размере обучающей выборки до 100 элементов. На большом и среднем размере обучающей выборки классификатор показывает схожие результаты. Точность классификации повышается при увеличении размера обучающей выборки и увеличении соотношения сигнал-шум.



а *б*



в *г*

Рис. 1. – Общая ошибка классификации при анализе смоделированных изображений: *а*) линейный ДА; *б*) квадратичный ДА; *в*) наивный байесовский классификатор; *г*) *k* ближайших соседей: Черная линия – размер эталонной выборки 50, пунктир – размер эталонной выборки 100, штрих – размер эталонной выборки 150, штрих-пунктир – размер эталонной выборки 200, жирная сплошная – размер эталонной выборки 250, жирный пунктир – размер эталонной выборки 300.

На средней и большой обучающей выборке точность классификации квадратичным и линейным классификатором сопоставима.

На рис. 1, *в* представлены расчеты общей ошибки классификации для наивного байесовского классификатора в результате анализа смоделированных изображений. В отличие от дискриминантного анализа эффективность алгоритма байесовской классификации не зависит от размера обучающей выборки. Точность классификации увеличивается при увеличении соотношения сигнал-шум. Качество классификации сопоставимо с методами дискриминантного анализа.

На рис. 1, *г* показана общая ошибка классификации смоделированных данных для метода *k* ближайших соседей. Метод *k* ближайших соседей демонстрирует наихудшие результаты классификации. При соотношении сигнал-шум 1 ошибка составляет около 50 %. Наихудшие результаты получены для самоорганизующихся карт и слоя Кохонена, где ошибка классификации составляет около 50 % даже при высоком соотношении сигнал-шум. Общая ошибка классификации при максимальном размере обучающей выборки представлена в табл. 1. Наилучшие результаты классификации получены для наивного байесовского классификатора и линейного дискриминантного классификатора.

Таблица 1

Сводная таблица сравнительного анализа методов классификации (ЛДА – линейный дискриминантный анализ, КДА – квадратичный дискриминантный анализ, НБК – наивный байесовский классификатор, *k*ББ – метод *k* ближайших соседей)

	Уровень сигнала										
SNR	0,9	1	2	3	4	5	6	7	8	9	10
Метод	Общая ошибка классификации (в %)										
ЛДА	13	7	3	1	1	1	0,1	0,1	0	0	0,1
КДА	14	5	3	0,4	1	0,3	0,2	0,1	0	0	0
НБК	16	11	7	2	3	2	2	2	2	2	2
*k*ББ	50	48	28	12	14	7	4	6	3	3	2

Выводы

В результате сравнительного анализа на примере смоделированных изображений установлено, что для классификации объектов наилучшими методами являются: линейный и квадратичный дискриминантный анализ, наивный байесовский классификатор. При этом наивный байесовский классификатор можно использовать на малых размерах обучающей выборки.

Библиографические ссылки

1. *Hoheisel JD.* Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet. 2006. №7(3). С. 200–210.
2. *Абламейко С. В., Лагуновский Д. М.* Обработка изображений: технология, методы, применение. Минск : Амалфея, 2000. С. 304.
3. Материалы международной научно-практической конференции «Информационные технологии, электронные приборы и системы» (ITEDS'2010). 6–7 апреля 2010 г. Минск : БГУ. 2010.
4. The Stanford Microarray Database: implementation of new analysis tools and open source release of software / J. Demeter [et al] // Nucleic Acids Res. 2007. С. D766–770.
5. Molecular Biology of the Cell. / B. Alberts [et al] Garland Science. 2007. С. 1392.
6. *Феофанов А. В.* Спектральная лазерная сканирующая конфокальная микроскопия в биологических исследованиях. Успехи биологической химии. 2007. № 47. С. 371–410.
7. *Bushberg J. T., Leidholdt E. M., Boone J. M.* The essential physics of medical imaging. Philadelphia: Lippineott Williams & Wilkins, 2001.
8. *Гонсалес Р., Вудс Р., Эддинс С.* Обработка изображений в среде MATLAB. М.: Техносфера, 2006.
9. Spatial quantitative analysis of fluorescently labeled nuclear structures: Problems, methods, pitfalls / O. Ronneberger [et al] // Chromosome Research. 2008. № 16. С. 523–562.
10. *Lehmussola A., Ruusuvuori P., Selinummi J., Huttunen H., Yli-Harja O.* Computational framework for simulating fluorescense microscope images with cell populations. 2007. IEEE transactions on medical imaging. P. 1010–1016.
11. *Novikov E., Barillot E.* An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments. BMC Bioinformatics. 2005. № 6. P. 293.
12. *Deng H., Runger G., Tuv E.* Bias of importance measures for multi-valued attributes and solutions. // Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN). 2011. P. 293–300.
13. *Liu Z., Bensmail H., Tan M.* Efficient feature selection and multiclass classification with integrated instance and model based learning // Evol Bioinform Online 2012. V. 8. P. 197–205.
14. *Han, H., Li X.-L.* Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery // BMC Bioinformatics 2011. V. 12(Suppl 1).
15. *Costa D. D., Campos L. F., Barros A. K.* Classification of breast tissue in mammograms using efficient coding // Biomed Eng Online 2011.
16. *Chopra P., Lee J., Kang J., Lee S.* Improving cancer classification accuracy using gene pairs // PLoS One–2010 V. 5(12).
17. *Ким Дж. О.* Математическая статистика: факторный, дискриминантный и кластерный анализ. М. : Финансы и статистика, 1989.
18. *Ahdesmäki M., Strimmer K.* Feature selection in omics prediction problems using cat scores and false nondiscovery rate control // Annals of Applied Statistics, 2010 V. 4 (1). P. 503–519.
19. *Chopra, P., Lee J., Kang J., Lee S.* Improving cancer classification accuracy using gene pairs // PLoS One. 2010 V. 5(12).
20. *Vassilvitskii S.* K-means: Algorithms, Analyses, Experiments. University S.: Stanford University, 2007.
21. *Лисица Е. В., Яцков Н. Н., Апанасович Т. В., Апанасович В. В.* Применение искусственных нейронных сетей для классификации раковых клеток // Информационные технологии и системы 2012 (ИТС 2012). Материалы международной научной конференции, БГУИР. Минск, Беларусь, 24 октября 2012 г. С. 254–255.