Г.М. Попова, И.Ф. Дятчина, Н.В. Мельникова

Нейро-статистическая модель классификации многомерных объектов медико-биологической природы

Аннотация. В работе представлена методика построения нейро-статистической модели, основанной на совместном использовании двух разных методов обработки и классификации многомерных объектов: дискриминантного анализа и нейронной сети. На примере анализа цитологических препаратов материала мокроты показана эффективность построения и использования такой модели для количественной дифференцированной оценки состояния препарата.

Ключевые слова: дискриминантная модель классификации объектов, нейронная модель классификации объектов, количественная оценка состояния препарата материала мокроты, рак легкого.

Введение

В настоящее время в медицине возросли требования к качеству диагностической информации, эффективности ее обработки и анализа. Известно, что лабораторная медицина обеспечивает значительную часть диагностической информации. Диагностика состояния препаратов (гематологических, цитологических, гистологических и мн. др.), на основании достоверности которой часто осуществляется верификация диагноза заболевания, выбирается адекватный курс лечения, является одной из наиболее трудоемких и важных задач в медицине.

Как правило, диагностика состояния препаратов «традиционно» выполняется врачом клинико-диагностической лаборатории на светооптическом уровне, путем идентификации изображений объектов - разнотипных клеточных структур и комплексов, встречаемых в препаратах. Для описания объектов обычно используется субъективное описание, что не всегда позволяет адекватно оценить состояние препарата. Поэтому предлагаемый в работе количественный метод оценки параметров клеточных структур и классификация их с дифференцированным счетом актуальны, так как позволят объективизировать диагностику препарата в целом.

Стремление к количественной оценке медико-биологических показателей определено, прежде всего, организацией научных исследований и практических работ, опирающихся на современные методы получения, обработки и анализа информации. Поэтому закономерным стало применение математических методов для проведения исследований, для создания экспертных систем, систем информационной поддержки принятия решений врачом.

Целью настоящего исследования является построение модели классификации множества многомерных объектов с дифференцированным счетом. На примере метрического анализа изображений клеточных структур цитологических препаратов материала мокроты показано, что данная модель позволяет распознавать объекты, классифицировать их с дифференцированным счетом и количественно оценивать состояние препарата в целом для диагностики бронхолегочной патологии (предраковых процессов и рака легкого)[1]. Количественная оценка состояния препарата поможет врачу принять решение по дальнейшему обследованию пациента и выбору метода терапии.

[1] Выбор данных препаратов обоснован актуальностью объективизации морфологической диагностики предраковых процессов и рака легкого, так как до 70% случаев рака легкого диагностируются по цитологическому материалу мокроты и материалу биопсий [1].

1. Система анализа изображений исследуемых объектов

В исследованиях использовалась многофункциональная компьютерная система анализа изображений разнотипных объектов «Морфолог», которая позволяет вводить изображения в компьютер с препарата, установленного на световой микроскоп с видеокамерой, а затем количественно оценивать как индивидуальные изображения объектов, так и состояние препарата в целом [2]. Программные средства системы ориентированы в основном на автоматизацию процессов поиска, обработки (включающей фильтрацию, сегментацию, выделение контура изображения объекта и его внутренних элементов, параметризацию всех выделенных элементов), распознавания и классификации объектов согласно выбранной модели. Все процедуры для оценки состояния препарата в целом выполняются автоматически, кроме сегментации, которая на начальном этапе анализа (при настройке порогов первого препарата) выполняется интерактивно, а затем – автоматически для данной серии препаратов, используя ранее установленные настройки [3]. Отметим, что программные средства системы позволяют анализировать весь препарат без пропусков и повторений одних и тех же объектов.

Анализ объектов, представленных в препаратах в виде изображений одноядерных и многоядерных клеточных структур, выполняется в системе на основе их структурно - параметрических моделей. Структурная модель объекта формируется в графическом виде, с учетом условно принятых псевдоцветов, соответствующих каждому элементу объекта [3]. Для распознавания и классификации объектов по их изображениям помимо основных признаков (цвета и структуры), определяющих тип объекта, формируются характерные, отличительные особенности каждого из них, путем морфологического анализа и исследования их геометрических, топологических и статистических параметров.

В работе для диагностики и оценки характера состояния бронхолегочной системы анализировались и идентифицировались 10 типов объектов, представленные в препаратах мокроты в виде изображений одноядерных клеточных структур, являющихся самыми многочисленными и информативными (Рис. 1): плоский эпителий (ПЭ), макрофаги (МФ), цилиндрический (бронхиальный) эпителий (ЦЭ), кубический эпителий (КЭ), метаплазированный эпителий (плоскоклеточная метаплазия) (М), метаплазированный эпителий с признаками дисплазии (атипическая плоскоклеточная метаплазия) (Д), плоскоклеточный рак

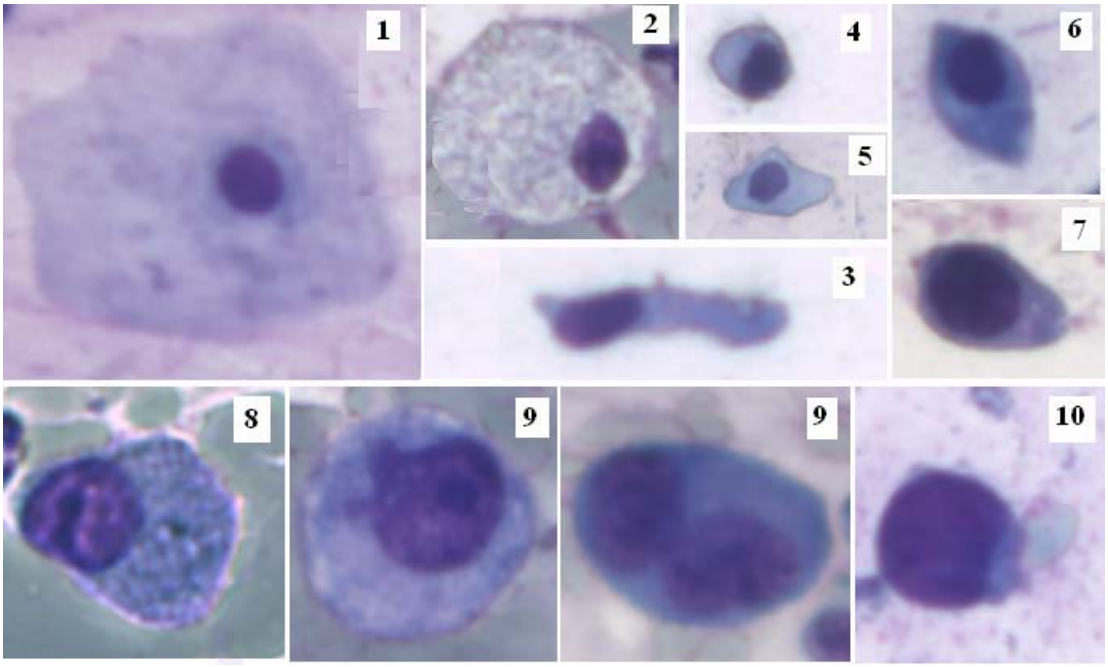

Рис. 1. Типы анализируемых клеточных структур

1 - плоский эпителий, 2 - макрофаг, 3 - цилиндрический эпителий, 4 - кубический эпителий, 5 - метаплазированный эпителий, 6 - метаплазированный эпителий с признаками дисплазии, 7 - плоскоклеточный рак, 8 - железистый рак, 9 - крупноклеточный рак, 10 - мелкоклеточный рак

(ПР), железистый рак (аденокарцинома) (ЖР), крупноклеточный рак (КР), мелкоклеточный рак (МР). Кроме перечисленных типов клеток, введен еще один тип клеток, названный «неопределенные клетки» - НК.

Для всех объектов характерна очень высокая вариабельность морфометрических параметров (размеров, форм, цветовых характеристик), что не позволило однозначно выделить значимые для классификации признаки и построить логическое классификационное дерево для их идентификации. Поэтому первоначально стояла задача выявить диагностически значимые признаки объектов и построить модель, которая могла бы быть использована для классификации многомерных объектов с дифференцированным счетом. А затем, построить диагностическую модель, которая, оценивая состояние всего препарата в целом, позволяла бы поддержать принятие решений врачом при диагностике заболеваний. При этом, естественно, модель должна работать максимально точно, быть статистически значима и удобна в работе.

Данная работа проводилась путем последовательного исследования двух разных статистических методов обработки и классификации многомерных объектов: дискриминантного анализа для формирования обучающей выборки - численных значений признаков объектов и нейронной сети, использующей эту выборку для обучения и классификации других аналогичных объектов с дифференцированным счетом. Принятая методика построения двух моделей достаточно разных по своим методам классификации направлена на сравнение их эффективности при решении данной задачи, а главное, на создание удобного инструмента поддержки принятия решений врачом при диагностике цитологических препаратов бронхолегочного материала.

2. Дискриминантная модель классификации многомерных объектов

Для исследований первоначально использовался многомерный статистический анализ - классификационный дискриминантный анализ с обучением.

Цель дискриминантного анализа состояла в следующем:

- выявить диагностически значимые признаки (морфометрические параметры) для идентификации клеточных структур;
- на основе анализа признаков объектов определить наиболее важные для процедуры дискриминации;
- разбить все выборочное пространство признаков на кластеры для получения групп наиболее близких объектов, а затем путем определения дискриминантных функций, позволяющих отнести каждый объект к одному из 10 типов объектов, сформировать обучающую выборку с учителем.

Для исследований было проанализировано более 50 верифицированных препаратов с бронхолегочной патологией, в банке изображений было собрано более 1500 разноименных типов клеточных структур, для обучения системы были использованы 1060 клеток.

Для статистической обработки данных использовался пакет прикладных программ Statistica 6. Результаты исследования обрабатывали с использованием критерия Стьюдента, теста Фишера. Достоверность статистических гипотез считалась доказанной при пороговых уровнях значимости $t < 0,05$ и $t < 0,01$.

Формально задача ставилась следующим образом. Имеется массив объектов, каждый из которых описывается фиксированным вектором числовых признаков, т.е. каждый i-ый исходный объект (клетка) в результате параметризации описывается k-мерным случайным вектором $X_i = (x_1, x_2, \ldots, x_k)$, где k количество числовых признаков объекта. Все одноименно идентифицируемые объекты образуют группы клеток (классы) с совокупностью координат вектора R_j, где $j = 1, 2,\ldots,10$.

Для первоначального обучения классификатора была взята случайная выборка X ($X_i \in X$), состоящая из разнотипных клеток, для которых на основании эмпирических заключений эксперта - цитолога была определена их принадлежность к одному из R_j заранее заданных классов. Далее обучение классификатора проводилось на основании сформированного случайного кластеризатора.

Для отбора наиболее значимых параметров был проведен пошаговый дискриминантный анализ, который позволил определить, имеются ли значимые различия между группами с точки зрения всех параметров. Для анализа качества различия R_j групп использовались критерии:

λ – статистика Уилкса, определяемая значимость (мощность) дискриминации после того, как соответствующая переменная вводится в модель; межгрупповая F – статистика, определяемая при включении каждой переменной в модель, учитывалось Fстат. > Fвкл (переменная с максимальным значением Fстат. будет включена в модель на первом шаге). В результате пошаговой процедуры для дискриминантного анализа из k = 22 параметров клеток были отобраны 17 значимых[2]. Для некоторых параметров (ELGц, ЦVц, ЦVя, CONTц) статистический анализ не выявил высокой диагностической значимости, однако они были сохранены, так как работали при построении дискриминантной функции. На Рис. 2 представлены графики средних значений наиболее значимых параметров клеток, из которых видно, что средние значения для групп отличаются, поэтому задача классификации объектов по сочетанию признаков разумна.

Обучающий набор из N исходных векторов (наблюдений) с известными признаками и ответами был представлен набором $\mathbf{T} = ((x_1,y_1), \ldots, (x_N,y_N)) \in (X \times Y)^N$, состоящим из пар (вектор признаков, ответ), где Y – любое пространство ответов, точками которого кодируются результаты распознавания. Для построения модели классификатора был использован метод обучения, максимизирующий оценку уверенности классификатора в ответах на обучающем наборе **T** с апостериорной вероятностью p, т.е. использовался метод наибольшего правдоподобия, который максимизирует различия между классами, но минимизирует дисперсию внутри классов [4, 5].

Значение p выбрано как «порог уверенности», что данная клетка принадлежит выбранному классу, если $p \geq 0,5$, то это может быть как правильно идентифицированная клетка, так и ошибочная, несовпадающая с тестом эксперта; клетки с $p < 0,5$ – это неопределенные клетки, они могут быть как ошибочными, так и правильными, но с вероятность $p \leq 0,499$ и ниже. Эти клетки практически лежат на пересечении классов (в зоне неопределенности), ранее их назвали неопределенными клетками, хотя, как увидим в дальнейшем, они как и все клетки в препарате определяются и подсчитываются.

В итоге обучения были определены линейные классификационные функции для каждого класса R_j:

$$S_j = c_j + w_{1j}*x_1 + w_{2j}*x_2 + \ldots + w_{kj}*x_k,,$$

где j – номер группы клеток, k – номер признака (параметра); c_j – константы j-ой группы клеток, w_{ij} - веса i-ого параметра при вычислении показателя классификации для j-ой группы клеток; x_i – i-й наблюдаемый параметр для соответствующей клетки, S_j – результат показателя классификационной функции. Функции классификации вычислялись для каждой совокупности R_j и поэтому могут быть использованы непосредственно для классификации клеточных структур (Табл. 1).

В работе применялся линейный дискриминантный анализ, основанный на правиле классификации Байеса [5], поэтому при вводе каждого нового наблюдения классификационные веса вновь пересчитывались для каждой совокупности клеток R_j. В результате анализируемый объект приписывается той группе, для которой классификационная функция имеет наибольшее значение S_j.

В Табл. 2 представлена матрица классификации, содержащая конечную информацию о количестве правильно и неправильно классифицированных клеток обучающего набора N в каждой группе R_j, с учетом достоверности этой классификации. Строки матрицы – исходные классы, столбцы – предсказанные классы. В первом столбце указан процент объектов T, которые были правильно классифицированы для каждой совокупности R_j клеток, полученными функциями S_j. Это и есть оценка дискриминирующих функций (апостериорной классификации).

[2] **Sц** – площадь цитоплазмы; **Sя** –площадь ядра (в случае многоядерности указывается суммарная площадь ядер); **Sя/Sц** – отношение площади ядра к площади цитоплазмы; **Рц**, **Ря** – периметры цитоплазмы и ядра; **Ря/Рц** – отношение периметра ядра к периметру цитоплазмы; **ROUo, ROUя** – округлости объекта и ядра; **TEXц** – текстура цитоплазмы (среднеквадратичное отклонение от среднего значения яркости, вычисленное по синей компоненте исходного изображения); **EXCя** – эксцентриситет ядра, который вычисляется как отношение расстояния между центром тяжести ядра и центром тяжести цитоплазмы к наибольшему размеру цитоплазмы; **ELGц** – элонгация клетки (отношение меньшей полуоси эллипса к большей); **CONTц**, **CONTя** – контрасты цитоплазмы и ядра, которые вычисляются как средняя разница в яркости между соседними точками с шагом в четыре пикселя; **ЦSц**, **ЦSя** - усредненные значения насыщенности цвета цитоплазмы и ядра; **ЦVц**, **ЦVя** - усредненное значение яркости цитоплазмы и ядра.

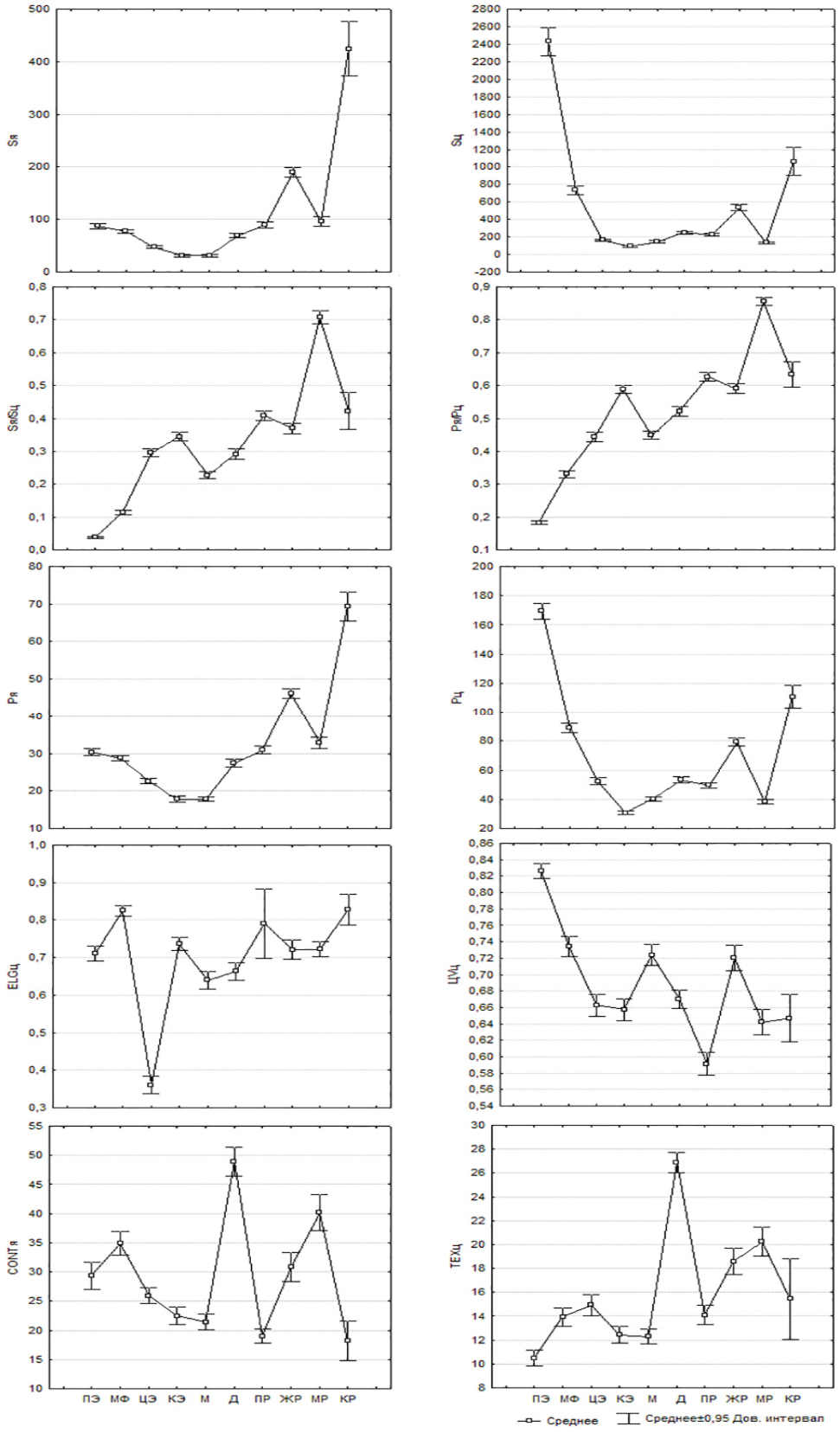

Рис.2. Графики средних значений параметров объектов

Табл. 1. Функции классификации ДМК

Перемен-ная	Функции классификации; группировка: Клетки СТАТ									
	ПЭ μ=,100	МФ μ=,100	ЦЭ μ=,100	КЭ μ=,100	М μ=,100	Д μ=,100	ПР μ=,100	ЖР μ=,100	МР μ=,100	КР μ=,100
Рц	7,684	6,410	5,656	6,137	5,762	5,983	6,087	6,147	6,500	6,317
Sя	0,110	0,019	-0,015	0,112	0,051	-0,007	-0,004	0,064	-0,005	0,604
Sя/Sц	-272,776	-271,694	-234,575	304,583	292,933	285,281	249,379	276,425	179,623	318,213
ROUц	218,596	201,118	103,352	152,699	143,193	154,581	158,527	151,258	184,512	150,600
TEXц	0,786	0,931	1,001	0,937	0,949	1,443	0,991	1,188	1,123	0,875
Ря	-9,471	-7,162	-6,392	-8,026	-7,070	-6,564	-6,784	-6,487	-7,326	-8,880
ЦSц	80,835	48,076	74,200	86,106	94,503	80,052	90,283	109,784	128,819	108,601
ЦVя	4,047	7,514	9,404	16,936	29,769	14,754	14,859	25,610	-1,668	-0,253
EXCя	37,219	74,394	91,547	62,249	43,427	47,594	57,180	71,801	51,693	68,284
Ря/Рц	870,008	779,390	746,653	870,417	788,795	812,306	809,001	820,336	830,716	865,314
CONTя	0,104	0,119	0,026	-0,058	-0,095	0,170	0,036	0,180	0,202	0,250
Sц	-0,213	-0,193	-0,170	-0,183	-0,173	-0,180	-0,183	-0,186	-0,192	-0,192
ROUя	208,527	236,460	265,836	228,544	235,997	239,676	245,074	252,936	226,902	226,527
ЦSя	56,790	59,363	55,847	35,102	50,764	58,382	54,886	27,469	0,713	28,865
CONTц	0,046	0,038	0,012	0,040	0,047	0,030	0,013	0,012	0,034	0,024
ЦVц	205,765	171,707	169,823	169,698	179,656	173,138	156,838	173,837	178,817	168,062
ELGц	-29,809	-28,869	-29,664	-29,096	-28,204	-30,922	-28,439	-30,281	-34,764	-28,626
Константа	-604,208	-496,008	-419,902	451,180	435,853	480,758	476,693	514,435	536,079	522,271

Табл. 2. Результаты классификации обучающей выборки ДМК

Группы клеток	Матрица классификации обучающей выборки ДМК										
	Процент правиль-ности	ПЭ μ=,100	МФ μ=,100	ЦЭ μ=,100	КЭ μ=,100	М μ=,100	Д μ=,100	ПР μ=,100	ЖР μ=,100	МР μ=,100	КР μ=,100
ПЭ	100,000	**115**	0	0	0	0	0	0	0	0	0
МФ	100,000	0	**115**	0	0	0	0	0	0	0	0
ЦЭ	100,000	0	0	**115**	0	0	0	0	0	0	0
КЭ	100,000	0	0	0	**115**	0	0	0	0	0	0
М	100,000	0	0	0	0	**115**	0	0	0	0	0
Д	99,130	0	0	0	0	**1**	**114**	0	0	0	0
ПР	99,138	0	0	0	0	0	0	**115**	**1**	0	0
ЖР	100,000	0	0	0	0	0	0	0	**116**	0	0
МР	100,000	0	0	0	0	0	0	0	0	**116**	0
КР	86,364	0	0	0	0	0	0	0	**3**	0	**19**
Всего	99,528	115	115	115	115	116	114	115	120	116	19

Примечание

Ввиду того, что априорные вероятности μ существенно влияют на точность классификации, для обучения были приняты априорные вероятности одинаковые для всех групп клеток.

Табл. 3. Статистика результатов классификации выборок в ДМК

Клетки	Контрольная выборка			Тестовая выборка			Выборка И - 100			Выборка И - 101		
	всего	правильно	ошибочно / НК	всего	правильно	ошибочно/НК	всего	правильно	ошибочно/НК	всего	правильно	ошибочно/НК
ПЭ	8	8		9	8	1 / 0	6	6		4	4	
МФ	4	4		2	2		1	1		1	1	
ЦЭ	6	6		6	6		4	4		2	2	
КЭ	22	20	1 / 1	17	17		21	19	2 / 0	27	24	3 / 0
М	37	37		26	26		41	41		30	30	
Д	32	32		52	49	2 / 1	8	7	1 / 0	9	8	0 / 1
ПР	35	35		1	1		5	5		22	22	
ЖР	7	7		0			0	0		28	27	1 / 0
МР	12	12		0			27	26	0 / 1	0	0	
КР	3	1	2 / 0	0			0	0		0	0	
Итог анализа	**166**	**162**	**3 / 1**	**113**	**109**	**3 / 1**	**114**	**110**	**3 / 1**	**123**	**118**	**4 / 1**
% правильных	**97,59**			**96,47**			**96,49**			**95,94**		

Примечание
Пустые ячейки в таблице означают отсутствие ошибок и неопределенных клеток: 0 / 0

Оценка качества обучения дискриминантной модели классификатора (ДМК) определялась ошибкой обучения на обучающей выборке, она составила 0,47%. Эффективность решения конкретной задачи определялась оценкой ДМК на контрольных выборках, т.е. по результатам классификации 4-х контрольных выборок, не участвующих в обучении и с проверкой правильности распознавания клеток экспертом. Результаты классификации клеток этих конкретных выборок представлены в Табл. 3. Данные этих выборок (отсюда и их названия) подбирались для контроля ниже рассмотренной нейронной структуры.

Из Табл. 2 и Табл. 3 видим, что построенная дискриминантная модель классификатора вполне применима для идентификации клеток мокроты, так как общий процент правильно классифицированных клеток из группы обучения составил - 99,53%, клеток из контрольных препаратов – 97,59%, 96,47%, 96,49%, 95,94%, соответственно, что является неплохим результатом.

Отметим, что дискриминантный анализ позволил довольно быстро сформировать выборку для обучения, причем ее легко пополнять новыми наблюдениями, при этом классификационные функции быстро меняются, качество модели улучшается. Иными словами, многофакторный дискриминантный анализ (ДА) – это удобный обучающийся статистический метод классификации и выделения признаков, позволяющих относить наблюдаемые объекты в реально наблюдаемые классы - R_j, т.е. для формирования обучающей выборки. Однако оценить эффективность этой модели на контрольных препаратах очень сложно, также как и использовать ДМК в качестве инструмента для диагностики состояния новых анализируемых препаратов, так как:

1. Система ДА при работе автоматически формирует только матрицу результатов классификации обучающей выборки с дифференцированным счетом объектов и не формирует результат классификации с дифференцированным счетом объектов, как контрольных препаратов, так и других новых анализируемых препаратов, что создает большие неудобства при использовании (результаты должен анализировать сам пользователь, исходя из апостериорных вероятностей). Поэтому дискриминантный классификатор удобен только при классификации небольших массивов с не более 10 новыми анализируемыми наблюдениями.

2. При идентификации клеток возможны неопределенные клетки, которые также автоматически не считаются в матрице результатов классификации, они должны оцениваться пользователем, который на основе апостериорной вероятности и, исходя из показателей принятых порогов, сам должен видоизменять матрицу

классификации. Все это требует время и большого внимания пользователя, как при анализе клеток, так и при их счете.

3. Многофакторный дискриминантный анализ классифицирует клетки согласно апостериорным вероятностям, но он как и большинство статистических методов, опирается на усредненные характеристики выборок, а не на генеральную выборку объектов, поэтому характеристики при исследовании не всегда являются корректными.

3. Нейросетевая модель классификации многомерных объектов

Известно, что нейронная сеть (НС) хотя и долго обучается, но в ней отсутствуют перечисленные недостатки, поэтому попытаемся на основе дискриминантной модели построить эквивалентную ей нейронную сеть и, если потребуется, то дообучить ее.

Стандартный способ обучения НС заключается в том, что сеть обучается на одном множестве наборов - многомерных наблюдений - Т, а на других множествах выполняется контроль и тестирование функционирования НС, т.е. проверяется НС на множествах, неиспользуемых для обучения. Чтобы гарантированно получить только информативные данные при обучении НС, необходимо выполнить перебор большого количества наборов данных и архитектур НС. Однако практически это реализовать трудно даже при наличии мощных и эффективных нейроимитаторов [6]. Поэтому в работе для анализа и выбора НС использовали обучающую выборку из базы данных дискриминантного классификатора, идентифицированные клетки которого были протестированы ранее экспертом. Это значительно облегчило выбор НС.

Для выбора варианта архитектуры сети использовался автоматический нейроконструктор (Мастер решений) из пакета Statistica Neural Networks, позволивший:

- автоматически использовать обученную выборку, полученную при построении ДМК, для обучения НС;

- выбрать из разных типов нейронных сетей лучшую сеть и архитектуру модели для классификации клеточных структур с производительностью (процентной долей правильно классифицированных наблюдений) не хуже ДМК, со средней ошибкой, получаемой на исходе последнего такта функционирования сети, равной наперед заданному значению $p_{теор} = 0,05$.

- получить графическое изображение схемы сконструированной НС; матрицы классификации обучающей, контрольной, тестовой выборок и других новых выборок, с дифференцированным счетом правильно, неправильно и неоднозначно определенных клеток.

На этапе моделирования исследовались и анализировались два типа нейронных сетей: радиальная базисная функция и трехслойный персептрон для классификации 10, ранее рассмотренных классов клеточных структур. НС рассматривались с 17 входами для ввода независимых переменных, представляющих собой вектора $X_i = (x_1, x_2, \ldots, x_k)$ с k классификационными признаками - параметрами клеточных структур, и одним выходом.

Обучение проводилось с использованием обучающей выборки из 1060 векторов (наблюдений) из файла данных ДМК, в режиме с кросспроверкой на контрольном множестве из 166 наборов наблюдений для оценки ошибки и 113 векторов тестового множества для сравнения альтернативных моделей, т.е. для независимого контроля качества сети. Контрольная выборка была сформирована из разных типов клеток, взятых с разных препаратов и не участвующих в обучении. Тестовая выборка была сформирована из клеток нового верифицированного препарата «Предраковое состояние». Для всех клеток были известны правильные ответы.

Для классификации задавались два доверительных уровня: порог принятия, равный 0,5 (это минимальное значение выхода, при котором наблюдение будет считаться принадлежащим выбранному классу) и порог отвержения, равный 0,5 (максимальное значение выхода, относящее измерение к отвергнутому классу). Ошибочные клетки это - клетки неправильно идентифицированные относительно теста эксперта с достоверностью $p \geq 0,5$. Неоднозначно определенные клетки (НК) также могут быть правильно или неправильно идентифицированы, но с $p < 0,5$.

Из 100 проанализированных НС, исходя из производительности сетей и ошибки на контрольной выборке, был выбран трехслойный персептрон. Затем из 150 моделей трехслойно-

го персептрона (с сохранением лучшей, соблюдая при этом баланс между ошибкой и сложностью сети) была выбрана одна стандартная модель трехслойного персептрона (МП 17:17-14-10:1) с 17 входами, 1 выходом и тремя промежуточным слоями с 17, 14 и 10 нейронными элементами, соответственно.

Обучение НС проводилось в два этапа: на первом этапе использовался метод обратного распространения ошибки со скоростью (с шагом) обучения 0,01 и длительностью 200 эпох, на втором проводилась оптимизация методом спуска по сопряженным градиентам длительностью 500 эпох. В качестве функции активации нейронных элементов на первом слое использовалась линейная функция, на втором – скрытом слое использовался гиперболический тангенс, на третьем слое – Softmax для минимизации перекрестной энтропии. Этот модуль позволил получить на выходе нейросети вероятности принадлежности входного объекта одному из не пересекающихся классов R_j. В этом случае для каждого произвольного вектора X_i на выходе выдаются некоторые значения апостериорной вероятности в интервале от 0 до 1. Значения этих вероятностей интерпретируются как меры уверенности принадлежности клеток к одному из классов R_j с указанием наилучшего приближения.

На Рис. 3 представлена архитектура выбранной модели нейронной сети для классификации клеток, а в Табл. 4 представлены подробные результаты этой модели. В первой строке таблицы представлена архитектура модели, в строке «обучение/элементы» дано краткое описание результатов используемых алгоритмов обучения: OP100, СГ20, СГ2b.

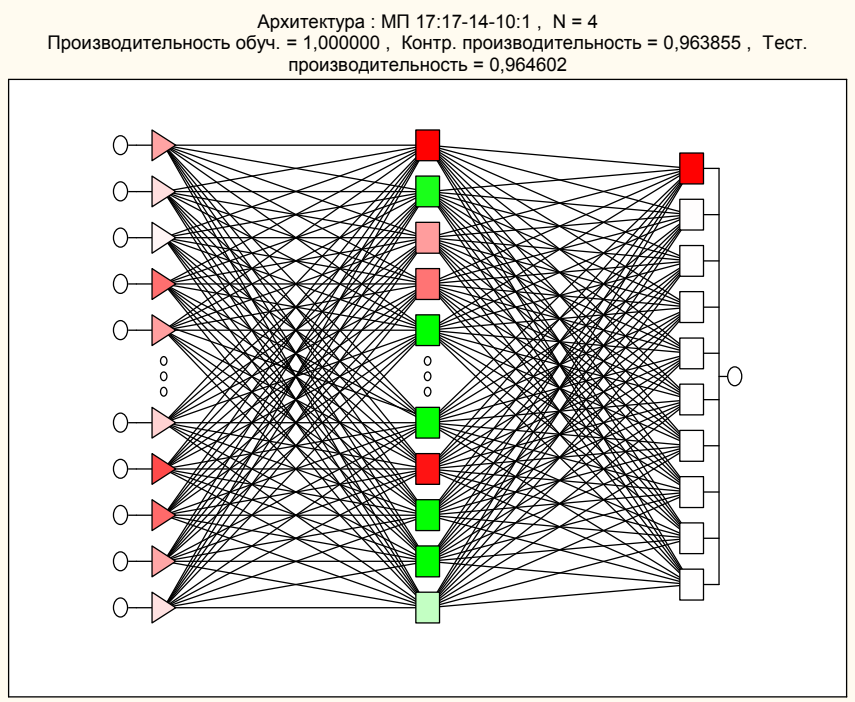

Рис. 3. Архитектура НМК

Табл. 4. Результаты модели НМК

Архитектура	МР 17:17-14-10-1
Производительность обучения	1,000
Контрольная производительность	0,9639
Тестовая производительность	0,9646
Ошибка обучения	0,0580
Контрольная ошибка	0,2918
Тестовая ошибка	0,2532
Обучение / элементы	OP100,СГ20,СГ2b
Входы	17
Скрытые (1)	14
Выход	1

Табл. 5. Статистика результатов классификации выборок в НМК

Клетки	Контрольная выборка			Тестовая выборка			Выборка И - 100			Выборка И - 101		
	всего	правильно	ошибочно/НК	всего	правильно	ошибочно/НК	всего	правильно	ошибочно/НК	всего	правильно	ошибочно/НК
ПЭ	8	8		9	9		6	6		4	4	
МФ	4	4		2	1	1 / 0	2	1	1 / 1	1	1	
ЦЭ	6	6		6	6		4	4		2	2	
КЭ	22	22		17	16	0 / 1	21	21		27	27	
М	37	36	1 / 0	26	26		41	41		30	30	
Д	32	32		52	52		8	8		9	9	
ПР	35	33	2 / 0	1	1		5	3	1 / 1	22	19	3 / 0
ЖР	7	7		0	0		0	0		28	27	1 / 0
МР	12	11	1 / 0	0	0		27	27		0	0	
КР	3	3		0	0		0	0		0	0	
Итог анализа	**166**	**162**	**4 / 0**	**113**	**111**	**1 / 1**	**114**	**111**	**2 / 1**	**123**	**119**	**4 / 0**
% правильных	**97,59**			**98,23**			**97,37**			**96,75**		

Примечание
Пустые ячейки в таблице означают отсутствие ошибок и неопределенных клеток: 0 / 0

При выборе модели учитывались: производительность НС, которая при классификации служит мерой правильно классифицированных наблюдений, как при обучении, так и при контроле и тестировании, и уровни ошибок на каждой выборке. Известно, что алгоритм итерационного обучения НС оптимизирует функцию ошибки (среднеквадратичную ошибку кросс-энтропии между наблюдаемыми и предсказанными выходами), корректируя веса сети таким образом, чтобы уменьшить суммарную ошибку, т.е. разность между желаемым (целевым) и реальным выходом. В Табл. 4 представлены уровни ошибок обучающей выборки, контрольной и тестовой, которые имеют большую значимость для алгоритма обучения.

Результаты анализов чувствительности выбранной нейронной сети к входным переменным при обучении и контроле подтвердили значимость 17 входных переменных для дифференциальной диагностики клеток, как и при дискриминантном анализе.

В Табл. 5 представлены результаты классификации выбранной модели, в которой приведены статистики о количестве правильно, неправильно и неклассифицированных наблюдений (нераспознанных клеток) по каждому классу R_j для контрольной, тестовой, а также двух игнорируемых при обучении выборок И-100 и И-101 для проверки обобщающих способностей нейронной сети. Параметры клеток для этих выборок получены с верифицированных новых препаратов «Мелкоклеточный рак», «Железистый рак», соответственно[3]. Клетки данных препаратов не принимали участие в обучении (численные значения их признаков поочередно пропускались через полученную модель классификатора). Эти выборки предназначены для контроля и оценки способности модели к обобщению, т.е. классифицировать новые препараты. В результате эффективность прогнозирования нейронной модели классификатора (НМК) на экспериментальных множествах составила 97,37% и 96,75%, что является очень неплохим результатом.

Однако следует отметить, что вероятность ошибок распознавания клеток присутствует, но анализ этих ошибок показал, что они в основном не выходят за рамки категорий (Приложение). Например, в обучающей выборке ДМК три клетки КР классифицированы как ЖР, в контрольной выборке НМК клетка МР классифицирована как ПР, М – как КЭ. Ошибки неизбежны из-за большой вариабельности параметров клеток, но ввиду того, что дифференцированный счет клеток в биологических препаратах выполняется на 100 и более клеток, то полученный процент ошибок невелик.

В Табл. 6 для сравнения представлены результаты эффективности прогнозирования двух классификаторов: нейронного и дискриминантного.

[3] Статистика для обучающей выборки не представлена ввиду 100% правильных ответов (Табл. 4).

Табл. 6. Результаты классификации анализируемых выборок в ДМК и в НМК

Итоги анализа	Матрицы классификации ДМК					Матрицы классификации НМК				
	Обучение	Контрольная	Тест	И-100	И-101	Обучение	Контрольная	Тест	И-100	И-101
Всего	1060	166	113	114	123	1060	166	113	114	123
Правильно	1055	162	109	110	118	1060	162	111	111	119
Ошибочно	5	3	3	3	4	0	4	1	2	4
Неизвестно	0	1	1	1	1	0	0	1	1	0
% Правильных	99,53	97,59	96,47	96,49	95,94	100,00	97,59	98,24	97,37	96,75
% Ошибочных	0,47	1,81	2,65	2,63	3,25	0,00	2,41	0,88	1,75	3,25
%Неизвестных	0,00	0,60	0,88	0,88	0,81	0,00	0,00	0,88	0,88	0,00

Процент правильно классифицированных клеток на одних и тех же выборках приблизительно одинаковый, но удобства получения результатов классификации у них разные.

Таким образом, видим, что разработанная нейро-статистическая модель классификатора способна к обобщению, она адекватно прогнозирует новые наблюдения с других препаратов. Причем статистика результатов классификации этих множеств наблюдений автоматически выводится в виде матриц не только с количеством правильно классифицированных клеток, но и неопределенных и ошибочных, что очень удобно при анализе биомедицинских препаратов. При этом достоверность классификации каждой выборки удобно представлена в таблицах предсказания каждого наблюдения с уровнем достоверности.

Следует также отметить, что используя сформированный обучающий набор с учетом контрольного и тестового наборов, «Мастер решений» уже после трех прогонов переменных из 150 НС автоматически нашел две лучшие сети, одна из которых, имея лучшие показатели к обобщению, была принята для рассмотрения.

Таким образом, имея результаты дифференцированного счета всех (распознанных и нераспознанных) клеток в анализируемых препаратах и методику количественной оценки препарата мокроты в целом (Приложение), можно предсказать бронхолегочную патологию, причем с определением риска предракового характера процесса (высокий, низкий, неопределенный) и злокачественного процесса с указанием типа рака.

Заключение

На основании настоящих исследований построена нейро-статистическая модель, позволяющая на примере диагностики состояния цитологических препаратов мокроты показать эффективность совместного использования двух разных методов обработки и классификации многомерных объектов: дискриминантного анализа, как относительно быстрого и удобного инструмента для формирования обучающей выборки числовых признаков объектов, и нейронной сети, позволяющей на этой обучающей выборке классифицировать объекты разных анализируемых препаратов с дифференцированным счетом.

Построенная нейро-статистическая модель классификатора позволяет автоматически тестировать разные препараты с заданными порогами доверительных уровней и выдавать матрицу результатов классификации не только с количеством правильно классифицированных клеток, но и неопределенных, что необходимо для достоверности анализа любого подобного биомедицинского препарата.

Особенность технологии построения нейро-статистической модели позволила сравнить два метода классификации клеточных структур по эффективности, оставляя право применения более удобной модели для визуального просмотра результатов. На наш взгляд, дискриминантная модель классификатора удобна, если требуется классифицировать небольшое число (менее 10) объектов, но если требуется классифицировать сто и более объектов с дифференцированным счетом, как например, в биологических препа-

ратах, то нейронная модель классификатора незаменима.

Разумеется, результаты построенной нейростатистической модели не могут заменить диагностическую оценку врача, так как модель - это искусственно создаваемая система, и она всегда упрощает и частично искажает оригинал. Но подобные модели необходимы в связи со сложностью процесса анализов не только цитологических, но и гематологических, гистологических и многих других биологических препаратов, и могут являться современными объективными инструментами для предсказания и поддержки принятия решений врачом по дальнейшему обследованию пациентов и выбору метода терапии.

Кроме того, организованные таким образом нейронные модели классификации, включенные в биомедицинские мониторинги, позволят следить не только за результатами исследования состояний препаратов до, во время, после терапии, но и за количественным соотношением присутствующих в них клеток, характеризующих заболевание. Оценка количественных соотношений клеток в динамике позволит прогнозировать течение заболевания и оценивать эффективность проводимой терапии.

Приложение

Количественная оценка препарата мокроты

Количественная оценка состояния всего препарата осуществляется на основе присутствия в препарате тех или иных типов клеток и их количественного соотношения. Для автоматизации процесса количественной оценки цитологического материала, и раннего выявления предраковых процессов и рака легкого было выделено шесть основных категорий различных вариантов состояний препаратов:

1. Неинформативный препарат.
2. Информативный препарат.
3. Доброкачественный характер процесса, когда в информативном препарате присутствуют клетки **ПЭ** или **КЭ** и их количество не менее 10 шт.
4. Злокачественный характер процесса, когда в препарате, наряду с другими клетками присутствуют раковые клетки и их суммарное количество не менее 10 шт. Желательно указывать тип рака, поскольку данные формы рака имеют разный прогноз и требуют дифференцированного подхода к выбору лечебных воздействий.
5. Предраковый характер процесса (ПХП), который в зависимости от количества присутствия в препарате клеток **М** и **Д** делится на три подкатегории:

- *ПХП высокого риска* развития злокачественного заболевания, когда в препарате присутствуют клетки рака, но их количество менее 10, и/или когда в препарате присутствуют клетки **М** и/или клетки **Д**, причем последних не менее 5 при отсутствии раковых клеток;
- *ПХП низкого риска* развития злокачественного заболевания, когда в препарате наряду с клетками **М** присутствуют клетки **Д**, но их менее 5 шт.;
- *ПХП неопределенного риска* развития злокачественного заболевания, когда в препарате кроме клеток **Д** и **М** имеются раковые клетки и в сумме их не менее 5. Врач может уточнить их количество в матрице классификации анализируемого препарата (Табл. 5).

6. Пограничный (неуточненный) характер процесса - во всех остальных случаях.

Обязательное присутствие клеток в данных категориях указано в Табл. 7. Более подробный алгоритм количественной оценки цитологического материала мокроты изложен в статье [7].

Согласно описанному правилу количественной диагностики биопрепаратов формируется бланк результатов анализируемого препарата с дифференцированным счетом клеток. В Табл. 7 представлен пример бланка результатов анализируемого тестового препарата.

Следует отметить, что алгоритм количественной диагностики состояния препарата построен так, чтобы избежать ложноотрицательных результатов и учесть все клетки, характеризующие предраковые процессы и рак легкого.

Табл. 7. Бланк результатов анализа препарата

Состояние препарата	Информативный препарат		Доброкачественный характер процесса		Предраковый характер процесса		Злокачественный характер процесса				Пограничный характер процесса		
Категории	**1**		**3**		**4**		**5**				**6**		
Клетки препарата	МФ	ЦЭ	ПЭ	КЭ	М	Д	ПР	ЖР	КР	МР	НК p < 0,5		
											6(3)	6(4)	6(5)*
Диф. счет клеток	**1**	**6**	**9**	**16**	**26**	**52**	**1**				**1**		
Сумма клеток по категориям	**7**		**25**		**78**		**1**				**1**		
Состояние препарата	**Предраковый характер процесса (высокий риск)**												

Примечание
НК - неопределённая клетка
*- 6(3), 6(4), 6(5) – показатели количества НК из категорий 3,4,5, у которых p < 0,5

Литература

1. Travis W. D., Brambilla E., Noguchi M. et al. Lung Adenocarcinoma Classification //Journal of Thoracic Oncology. 2011.v.6, № 2. p. 244-285
2. Попова Г.М., Степанов В.Н., Дружинин Ю.О., Дятчина И.Ф. Многофункциональный информационно – вычислительный комплекс анализа и диагностики изображений // Информационные технологии и вычислительные системы. 2010. № 4. С. 25 – 37.
3. Попова Г.М., Степанов В.Н., Дружинин Ю.О. Интерактивный метод обработки цветных изображений объектов биологической природы для получения их семантического описания // Системный анализ и управление в биомедицинских системах. 2009. Т. 8. № 3. С. 741-746.
4. Мерков А.Б. Распознавание образов: Введение в методы статистического обучения. М.: Едиториал УРСС, 2011. – 256 С.
5. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. М.: Финансы и статистика, 2000. – 352 С.
6. Нейронные сети под ред. Боровикова В.П. М.: Горячая линия–Телеком, 2008. - 392 С.
7. Попова Г.М., Дятчина И.Ф., Мельникова Н.В., Авилов О.Н. Компьютерная дифференциальная диагностика бронхолегочной патологии по цитологическому материалу мокроты // Системный анализ и управление в биомедицинских системах. 2012. Т.11. № 2. С. 325-336.

Попова Галина Михелевна. Ведущий научный сотрудник Института проблем управления им В.А. Трапезникова РАН. Окончила Московский энергетический институт в 1964 году. Кандидат технических наук. Область научных интересов: технологические средства (методы, модели, алгоритмы) информационных вычислительных систем, анализ и распознавание образов по их изображениям, организация систем биомедицинского мониторинга. E-mail: gmpopova@ipu.rssi.ru

Дятчина Ирина Федоровна. Научный сотрудник Института проблем управления им В. А. Трапезникова РАН. Окончила Московский институт радиотехники электроники и автоматики в 1982 году. Область научных интересов: обработка и анализ изображений, организация системы биомедицинского мониторинга.

Мельникова Надежда Васильевна. И.о. ведущий научный сотрудник лаборатории патоморфологии ФГБУ «РНЦРР» Минздрава России. Окончила ММСИ в 1993 году. Кандидат медицинских наук. Сертифицированный специалист по клинической лабораторной диагностике. Область научных интересов: цитологическая диагностика.
E-mail: n_melnikova@list.ru