Анализ методов добычи данных для классификации многомерных объектов медико-биологической природы

Г.М. Попова, В.Н. Степанов

Аннотация. В работе представлены анализ и сравнение различных методов классификации многомерных объектов медико-биологической природы, применяемых в области «добычи данных», а именно деревья принятия решений, метод опорных векторов и гибридные нейронные сети. Эффективность использования этих методов рассматривалась на примере анализа изображений клеточных структур цитологических препаратов материала мокроты.

Ключевые слова: добыча данных, метод опорных векторов, деревья принятия решений, нейро-нечеткие сети, классификация многомерных объектов, цитологические препараты.

Введение

Для анализа и обработки больших объемов информации в разных областях знаний применяют совокупность методов, объединенных под термином «добыча данных» (data mining) [1, 2]. Data Mining - это процесс обнаружения в «сырых» данных ранее неизвестных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных областях деятельности человека. В общем случае эти методы предназначены как для обнаружения закономерностей в массиве данных, так и для их анализа при классификации и принятии решений. Классификация является одной из основных задач, решаемых методами «добычи данных», в которых исходные данные могут быть ненадежными и слабо формализованными.

Более 25 лет с успехом применяются методы «добычи данных» для решения задач двух типов: описательного и предсказательного. К описательному типу относятся задачи, требующие обнаружения и описания скрытых новых правил в данных, такие как кластерный анализ, поиск ассоциативных паттернов. Задачи предсказательного типа требуют обработки и анализа данных, которые не участвуют в формировании обучающей выборки. Это задачи классификации, которые и рассматриваются в данной работе.

В области анализа многомерных объектов естественной природы существует много работ. Например, для моделирования диагностики онкологических заболеваний молочной железы была построена и промоделирована адаптивная нейро-нечеткая система [3]. Для прогнозирования поведенческих реакций у животных использовали нейро-нечеткую систему ANFIS [4], для поддержки принятия решения дифференциальной диагностики (по степени тяжести) бронхиальной астмы - факторный и дискриминантный анализ [5].

В работе [6] рассмотрен нейросетевой метод классификации многомерных объектов медико-биологической природы, основанный на обучающей выборке, сформированной путем дискриминантного анализа. Показаны неплохие результаты при использовании полученной нейро-статистической модели для распознавания и классификации с дифференцированным

счетом разных типов многомерных объектов - изображений клеточных структур цитологических препаратов материала мокроты, для предсказания новых наблюдений, выработки выводов и принятия решений. Однако искусственные нейронные сети (НС) весьма неудобны из-за их непрозрачности, обученная сеть – «черный ящик» для пользователя. Кроме того, обучение НС происходит обычно достаточно долго, и при этом нет уверенности, что сеть нашла глобальный минимум, а ввод априорной информации (знаний экспертов) для ускорения процесса обучения нейронной сети отсутствует.

В данной работе проводится анализ и сравнение разных методов классификации данных объектов медико-биологической природы на основе деревьев принятия решений, опорных векторов и гибридных нейро-нечетких сетей с целью определения наиболее эффективной (точность, скорость обучения, удобство использования) модели применительно для распознавания и классификации больших объемов выборок новых объектов с дифференцированным счетом. Для выбора модели многомерного классификатора использовались исходные данные: обучающая выборка, тестовая и др. обобщающие выборки, с согласия авторов, из работы [6]. Данные выборки анализировались, параметризовались и идентифицировались в компьютерной системе анализа изображений разнотипных объектов «Морфолог» [7].

Для диагностики состояния препаратов использовались 10 типов (классов) биообъектов, представленных в виде изображений одноядерных клеточных структур: плоский эпителий (**ПЭ**), макрофаги (**МФ**), цилиндрический эпителий (**ЦЭ**), кубический эпителий (**КЭ**), метаплазированный эпителий (**МЭ**), дисплазия (**Д**), плоскоклеточный рак (**ПР**), железистый рак (аденокарцинома - **ЖР**), крупноклеточный рак (**КР**), мелкоклеточный рак (**МР**).

В процессе исследования было проанализировано более 50 верифицированных препаратов с бронхолегочной патологией. В банке изображений было собрано более 1500 разных типов клеточных структур, для обучения были использованы 1060 клеток. Для анализа и идентификации этих клеточных структур были отобраны 17 входных значимых параметров [1] [6].

Для первоначального обучения классификатора была взята случайная выборка, состоящая из разнотипных клеток, для которых на основании эмпирических заключений эксперта - цитолога была определена их принадлежность к одному из заранее заданных 10 классов. Выборка К-101 из 166 наборов наблюдений была выбрана в качестве тестовой. Она была сформирована из разных типов клеток, взятых с разных препаратов, не участвующих в обучении. Выборка Т-110 из 113 векторов наблюдений была сформирована из клеток нового верифицированного препарата «Предраковое состояние", выборка И-100 из 114 векторов - с препарата «Мелкоклеточный рак».

1. Деревья принятия решений

На данный момент существует ряд алгоритмов автоматизированной классификации на основе деревьев принятия решений. Эти алгоритмы нашли широкое применение в области «добычи данных». Известно, что деревья принятия решений [2] отлично справляются с задачами классификации и позволяют визуально рассматривать многомерные данные с разных точек зрения. Пользователь может видеть, что происходит с моделью при добавлении новых переменных или, наоборот, при удалении одной или нескольких переменных. Такая визуализация позволяет уменьшать или увеличивать степень

[1] **Sц** – площадь цитоплазмы; **Sя** –площадь ядра; **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; **ТЕХц** – 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 –** усредненные значения яркости цвета цитоплазмы и ядра.

2 Дерево принятия решения представляет собой логическое дерево, в узлах которого стоят условия перехода по одной из двух ветвей, а в листьях – целевые классы. Если узловое условие – истина, осуществляется переход по левой ветви, если ложь – то по правой.

детальности модели, наблюдать изменения классификационного дерева при изменении порога отсечения малозначимых ветвей. Возможность подавления излишних деталей, как правило, очень важна для понимания моделей.

Для решения задачи прогнозирования принадлежности биообъектов к определенному классу использовали деревья классификации системы Statistica 6. Для первоначального рассмотрения были выбраны два дерева классификации с методами одномерного ветвления: дискриминантное многомерное ветвление по линейным комбинациям переменных и ветвление по методу C&RT (Classification And Regression Trees - деревья классификации и регрессии) [8].

Основным требованием, предъявляемым к классификаторам, является высокая обобщающая способность, т.е. обученная модель должна выдавать достаточно точные предсказания на новых (не входящих в обучающую выборку) наблюдениях. Наиболее точный прогноз считается такой, который связан с наименьшей ценой. Цена в нашем случае – это доля неправильно классифицированных наблюдений, так как априорные вероятности и цена ошибок неправильной классификации объектов были приняты равными. В медицинской практике допустимой считается ошибка, которая не превышает 5%-ный уровень.

Деревья классификации строились по исходной – обучающей выборке, а их способность к прогнозированию проверялась путем предсказания классовой принадлежности элементов тестовой выборки К-101. Исследования классификации объектов по 10 классам проводили с различными комбинациями параметров. Комбинации параметров, показавшие лучшие результаты, представлены в Табл. 1.

Здесь минимум n определяет момент, когда прекращается выбор ветвлений и начинается отсечение, т.е. пока терминальные вершины не будут содержать неправильно классифицированных наблюдений более чем заданное n. Правило стандартной ошибки используется для выбора усеченного дерева с наименьшим числом терминальных вершин и наименьшей ценой кросс-проверки. Кратность кросс-проверки v - это число случайных выборок, которые формируются из обучающей выборки и используются как тестовая выборка для кросс-проверки. Начальное значение датчика случайных чисел порождает заданное число случайных выборок из обучающей выборки. Критерий согласия G^2 – это мера максимума правдоподобия χ^2, используемая для выбора наилучшего ветвления, при котором максимально уменьшается значение выбранного критерия согласия.

Из Табл.1 видно, что лучшую точность, которая составила 83,7% на тестовой выборке, показало дерево классификации с методом «дискриминантное многомерное ветвление по линейным комбинациям переменных», но она

Табл. 1. Результаты классификации биообъектов с разными методами ветвления

Параметры	**Дискриминантное ветвление по линейным комбинациям переменных**		**Ветвление по методу C&RT**
Правила остановки ветвления	По ошибке классификации	Прямая остановка FACT	Прямая остановка FACT
Условия остановки	Минимум n= 5 Правило стандартной ошибки: 0,5	Доля неклассифицированных: 0,04	Критерий согласия -G^2 Доля неклассифицированных: 0,03
Кросс-проверка а тестовые выборки	Кратность кросс-проверки v = 5	Кратность кросс-проверки v = 5	Кратность кросс-проверки v = 10
Результаты	**Количество терминальных вершин**		
	26	40	74
Точность обучающей выборки (1060 наблюдений)	Ошибочно – 17 Правильно – 1043 (98,4%)	Ошибочно – 6 Правильно – 1054 (99,4%)	Ошибочно –11 Правильно – 1049 (99 %)
Точность тестовой выборки (166 наблюдений)	Ошибочно – 27 Правильно –139 (83,7%)	Ошибочно – 29 Правильно – 137 (82,5%)	Ошибочно – 56 Правильно – 110 (66,5 %)

невелика. Значение цены на тестовой выборке К-101 во всех методах оказалось больше, чем на обучающей выборке, что свидетельствует о плохом результате кросс-проверки. Лучшее дерево другого размера улучшало цену обучающего дерева, но цена тестовой проверки оставалось примерно одинаковой и довольно большой.

Исходя из полученных результатов, можно сказать, что данные деревья не приспособлены к классификации больших объемов новых многопараметрических наблюдений, как это требуется при классификации биообъектов, причем с дифференцированным их счетом.

Существует модификация алгоритма деревьев классификации и регрессии – это алгоритм «случайный лес» (random forest) [9]. Данный алгоритм обладает ключевыми преимуществами в виде приспособленности к работе с большим пространством признаков, а также хорошо распараллеливается (что могло бы позволить в дальнейшем ускорить его работу с применением технологий GPGPU). В результате действия этого алгоритма получается не одно дерево, а множество (в данном случае 60), «лес» деревьев. На каждой итерации происходит случайная выборка переменных, после чего на ней запускается процесс построения дерева по алгоритму C&RT. Итерации повторяются достаточно большое число раз, причем 2/3 обучающей выборки используется для обучения, а 1/3 - для проверки результата. На исследуемой выборке данный алгоритм показал неудовлетворительный результат, ошибка классификации доходила до 30%. Это объясняется тем, что данный алгоритм рассчитан на выборки с большим количеством переменных, так как для каждого дерева используется сокращенный случайный набор переменных.

Воспользуемся еще одним алгоритмом машинного обучения - «градиентные растущие деревья» (Tree Net) [10], который заключается в построении последовательности очень простых предсказывающих деревьев, ограниченных по количеству узлов. Каждое последующее простое дерево строится с учетом прогноза всех предшествующих деревьев, как независимо сформированной случайной выборки. На каждой итерации определяется наилучшее простое разбиение данных, вычисляется отклонение полученных данных от ожидаемых средних (остатка для каждого разбиения). Обучение ведется пока отклонение не перестанет уменьшаться. Так как каждое последующее простое дерево строится только по случайно сформированной подвыборке данных, то внесение такой случайности в анализ служит мощным средством предохранения от переобучения, поскольку все последующие деревья строятся по разным выборкам.

Оценка эффективности алгоритма «градиентные растущие деревья» для предсказания классов биообъектов проводилась на основе вариантов проверки достоверности модели. В варианте 1 количество наблюдений задает пользователь напрямую: обучающую выборку задали равной 1060, тестовую – 166, т.е. как и в предыдущих деревьях. В варианте 2 общий объем выборки приняли равным 1226 (1060обучение + 166 тест), а долю тестовой выборки, которая формируется случайным образом, задали равной 0,1.

Комбинация параметров, задаваемых пользователем для получения лучших результатов точности предсказаний, двух экспериментальных моделей классификаторов, следующая: для ограничения сложности каждого дерева задали минимальное число дочерних узлов равным **1**, максимальное число вершин (сложность каждого дерева) - **3**, минимальное число наблюдений (число неклассифицированных объектов) - **3**. Долю подвыборки и начальное значение датчика случайных чисел использовали, как рекомендовано по умолчанию, 0,5 и 1 соответственно. Цена ошибок классификации и априорные вероятности классов были приняты равными. На Рис.1 представлен график обучения растущих деревьев для варианта 1. Здесь отклонение тестовой выборки стабилизировалось на 174 итерации, при максимальном размере дерева – 3.

Результаты классификации обучающего набора из 1060 наблюдений по 10 классам содержали одну ошибку, причем она не существенна с практической точки зрения: клетка из одного типа рака (ПР) перешла в другой тип (МР). Тестовая выборка из 166 наблюдений имела 25 ошибочно классифицированных объектов, что составило 84,94% правильных предсказаний.

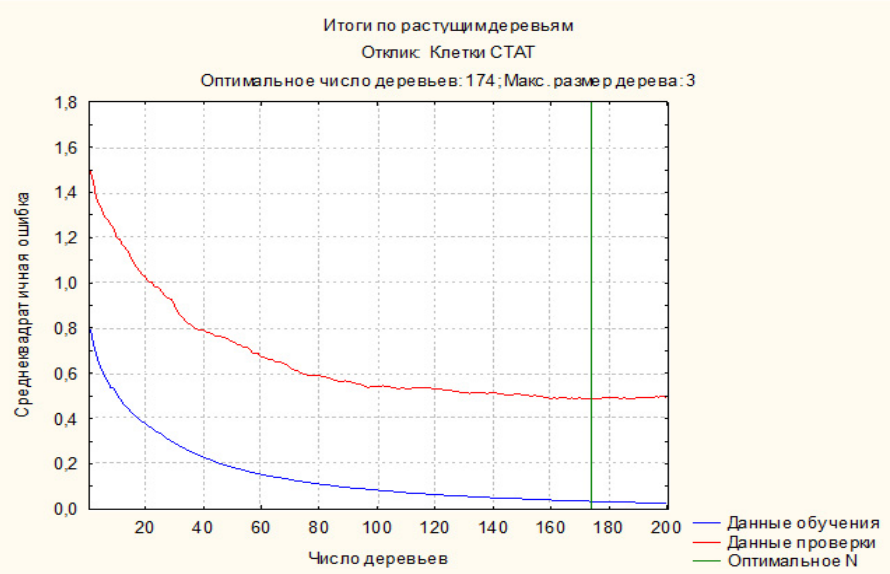

Рис. 1. График обучения алгоритма «растущие деревья» варианта 1

Поскольку программный пакет Statistica в данном случае позволяет анализировать новые, не участвующие в обучении выборки, было дополнительно проклассифицировано два препарата: Т-110, И-100. В Табл. 2 даны результаты классификации всех выборок, включая и последние препараты, у которых общий показатель правильно классифицированных клеток практически одинаков – 92%.

На Рис. 2 представлен график обучения выборки варианта 2, где отклонение тестовой выборки из 124 наблюдений стабилизировалось на 184 итерации, при максимальном размере дерева – 3.

Результаты классификации обучающего набора из **1102** наблюдений показали 100%-ную точность, тестовая выборка из **124** наблюдений имела 6 ошибочно классифицированных биообъектов, что составило 95,16% правильных предсказаний.

Табл. 2. Результаты классификации выборок методом «растущие деревья» варианта 1

Клетки	**Выборка обучения**			**Выборка К-101**			**Выборка Т-110**			**Выборка И-100**		
	всего	правильно	ошибочно	всего	правильно	ошибочно	всего	правильно	ошибочно	всего	правильно	ошибочно
ПЭ	115	115	0	8	7	1	9	7	2	6	4	2
МФ	115	115	0	4	3	1	2	2	0	2	2	0
ЦЭ	115	115	0	6	6	0	6	6	0	4	4	0
КЭ	115	115	0	22	16	6	17	16	1	21	21	0
М	115	115	0	37	32	5	26	25	1	41	38	3
Д	115	115	0	32	29	3	52	47	5	8	7	1
ПР	116	115	1	34	30	4	1	1	0	5	4	1
ЖР	116	116	0	8	5	3	0	0	0	0	0	0
МР	116	116	0	12	10	2	0	0	0	27	25	2
КР	22	22	0	3	3	0	0	0	0	0	0	0
Итог анализа	**1060**	**1059**	**1**	**166**	**141**	**25**	**113**	**104**	**9**	**114**	**105**	**9**
Точность, %	**99,9**			**84,94**			**92,04**			**92,11**		

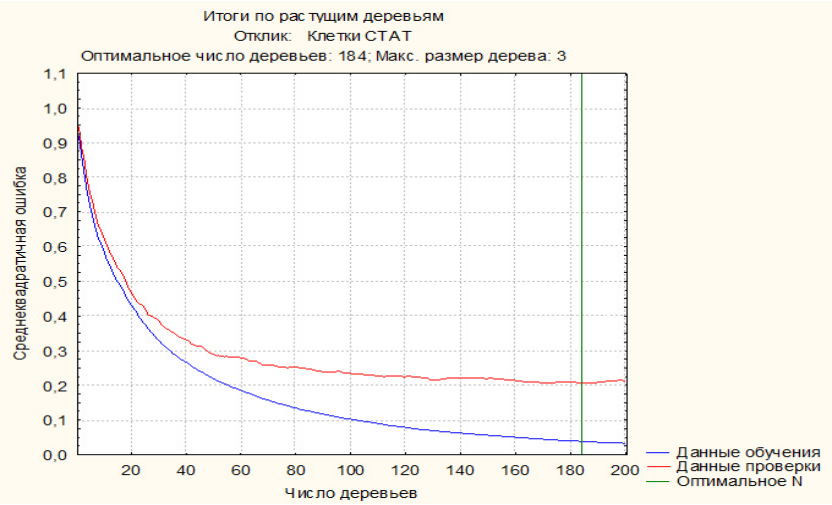

Рис. 2. График обучения алгоритма «растущие деревья» варианта 2

В Табл. 3 представлены результаты классификации всех выборок, включая и препараты Т-110 и И-100, у которых точность классификации клеток по классам составила 95,58 и 94,75% соответственно. Эти результаты являются весьма неплохими, что действительно доказывает, что «растущие деревья» могут хорошо обобщать новые наблюдения, т.е. давать прогнозы с высокой степенью достоверности.

Исходя из результатов, представленных в Табл. 2 и Табл. 3 видим, что точность классификации выше при формировании обучающей и тестовой выборки случайным образом. Кроме того, анализ ошибок показал, что многие из них возникли при пороге вероятности предсказания p = 0,4 и менее. Эти объекты правильнее отнести к неопределенным объектам. При этом если бы значение p – «порог уверенности», что данный объект принадлежит выбранному классу, принять равным 0,5 или 0,6, то это позволило бы результаты представлять в виде матрицы классификации с дифференцированным счетом клеток и с учетом неопределенных, которые всегда можно визуально проанализировать пользователем.

Кроме того, из Табл. 3 видим, что раковые клетки практически точно прогнозируются, что очень важно для данных исследований. Препарат И-100 со 100%-ной точностью определяет

Табл. 3. Результаты классификации выборок методом «растущие деревья» варианта 2

Клетки	**Выборка обучения**			**Выборка тест доля 0,1**			**Выборка Т-110**			**Выборка И-100**		
	всего	правильно	ошибочно	всего	правильно	ошибочно	всего	правильно	ошибочно	всего	правильно	ошибочно
ПЭ	112	112		11	11		9	9		6	6	
МФ	106	106		13	12	1	2	2		2	2	
ЦЭ	108	108		13	13		6	6		4	4	
КЭ	120	120		17	16	1	17	16	1	21	19	2
М	140	140		12	11	1	26	24	2	41	38	3
Д	133	133		14	14		52	50	2	8	7	1
ПР	132	132		18	15	3	1	1		5	5	
ЖР	112	112		12	12		0	0		0	0	
МР	116	116		12	12		0	0		27	27	
КР	23	23		2	2		0	0		0	0	
Итог анализа	**1102**	**1102**		**124**	**118**	**6**	**113**	**108**	**5**	**114**	**108**	**6**
Точность, %		**100**			**95,16**			**95,58**			**94,75**	

состояние препарата - мелкоклеточный рак. Предраковое состояние препарата Т-110 определяют клетки дисплазии, точность предсказания которых составила 94,23%. Это большая точность при большом их количестве (52) в препарате (для регистрации предракового состояния препарата достаточно 5 клеток дисплазии [6]). Поэтому отметим, что модель, построенная с помощью алгоритма «растущие деревья», могла бы быть использована уже в настоящее время биологами в качестве инструмента для диагностики состояния препаратов. Однако отсутствие в программном пакете Statistica вывода матрицы классификации новых выборок затрудняет использование данных деревьев, так как дифференцированный счет по вероятностям предсказаний очень утомителен для пользователя и требует много времени (в препарате возможно 100-200 и более клеток). Для анализа состояния любых биопрепаратов (цитологических, гематологических, гистологических и др.) всегда требуется дифференцируемый счет клеток по классам.

2. Классификация многомерных объектов методом опорные вектора

Рассмотрим еще один метод классификации, применяемый в системах добычи данных – метод опорных векторов. Он подразумевает нахождение оптимальных гиперплоскостей, разделяющих объекты на классы в пространстве признаков. Под оптимальностью гиперплоскости подразумевается максимизация расстояния до границ классов. Алгоритм линейной классификации можно записать в следующем виде:

$$a(x) = sign(\sum_{i=1}^{n} \lambda_i c_i x_i \times x - b),$$

где $a(x)$-принадлежность объекта x к одному из двух классов $(1,-1)$, n-размерность пространства признаков, $\lambda = (\lambda_1,\ldots,\lambda_n)$ - вектор двойственных переменных, c_i - класс переменной x_i, b-расстояние до начала координат [11, 12].

На практике, выборка крайне редко может быть разделена на классы линейно, что приводит к ошибкам при использовании линейного классификатора. Есть два способа решения этой проблемы: фильтрация и удаление исходных данных, мешающих линейной разделимости, а также применение нелинейного классификатора. Нелинейный классификатор заменяет скалярное произведение нелинейной функцией, которая называется «ядром» [13,14]. В работе был проведен анализ, как линейного классификатора, так и нелинейного с разными типами ядер (полиномиальное, радиальная базисная функция, сигмоидное), из которых наилучший результат был получен при использовании радиальной базисной функции:

$$k(x,x') = \exp(-\gamma \mid x - x'|^2), \gamma > 0,$$

где параметр γ влияет на точность обучения: большие значения γ приводят к переобучению, при малых значениях γ ядро постепенно сводится к постоянной функции.

Как и в «растущих деревьях» оценим эффективность использования метода опорных векторов в двух вариантах. В варианте 1 деревья классификации строились по исходной – обучающей выборке из 1060 наблюдений, а их способность к прогнозированию проверялась путем предсказания классовой принадлежности элементов тестовой выборки К-101, равной 166 наблюдениям. В варианте 2 общий объем выборки 1226 наблюдений, обучающая и тестовая выборки формировались случайным образом: 90% - обучающая выборка (1094) и 10% - тестовая выборка (132 объекта).

Исследования классификации объектов по 10 классам проводили с различными комбинациями параметров. Комбинации параметров, показавшие лучшие результаты, и сами результаты представлены в Табл. 4 и Табл. 5.

Здесь С – коэффициент регуляризации, который позволяет регулировать отношение между максимизацией ширины, разделяющей полосы, и минимизацией суммарной ошибки. Штраф определяет баланс между ошибками, совершаемыми в процессе обучения, и установкой жестких границ; высокие значения способствуют созданию более точной модели. Величина порога позволяет отбрасывать точки, у которых вероятность принадлежности точки определенному классу меньше, чем заданная величина порога.

Табл. 4. Результаты классификации объектов методом опорных векторов (вариант1)

Тип классификатора	**Линейный**				**Нелинейный** Ядро: радиально-базисная функция			
Тип классификации Сложность Gamma параметр	C – MOB C = 0,085		Ню – MOB Ню = 0,045		C – MOB C = 0,110 γ = 0,2		Ню – MOB Ню = 0,038 γ = 0,055	
Кросс-проверка Сетка поиска:	V = 10 0,01 / 3,0 / 0,005		V = 10 0,01 / 0,2 / 0,005		V = 10 0,01 / 2,0 / 0,01		V = 10 0,01 / 0,2 / 0,001	
Порог точности:	0,001		0,0001		0,001		0,0001	
Штраф:	1000		1000		1000		1000	
N опорных векторов	260 (0)		339 (3)		300 (0)		325(2)	
Итоги анализа	обучения	теста	обучения	теста	обучения	теста	обучения	теста
Всего	1060	166	1060	166	1060	166	1060	166
Правильно	1057	153	1056	155	1056	156	1058	153
Ошибочно	3	13	4	11	4	10	2	13
Точность, %	99,7	**92,2**	99,6	**93,4**	99,7	**94,0**	99,8	**92,8**

Табл. 5. Результаты классификации объектов методом опорных векторов (вариант 2)

Тип классификатора	**Линейный**				**Нелинейный** Ядро: радиально-базисная функция			
Тип классификации Сложность Gamma параметр	C – MOB C = 0,2 Сл. число=1000		Ню – MOB Ню = 0,03 Сл. число=1000		C – MOB C = 0,02 γ = 0,6		Ню – MOB Ню = 0,032 γ = 0,06	
Кросс-проверка Сетка поиска	V = 10 0,01 / 2,0 / 0,01		V = 10 0,01 / 0,2 / 0,01		V = 10 0,01 / 2,0 / 0,01		V = 10 0,01 / 0,2 / 0,001	
Порог точности	0,001		0,0001		0,001		0,001	
Штраф	1000		1000		2000		1000	
N опорных векторов	266 (0)		312 (0)		366 (0)		330(0)	
Итоги анализа	обучения	теста	обучения	теста	обучения	теста	обучения	теста
Всего	1094	132	1094	132	1094	132	1094	132
Правильно	1092	126	1093	128	1092	129	1092	129
Ошибочно	2	6	1	4	2	3	2	3
Точность, %	99,82	**95,45**	99,91	**96,97**	99,82	**97,73**	99,82	**97,73**

Исходя из Табл. 4, можно сказать, что точность рассмотренных классификаторов на тестовых выборках практически одинакова, во всех случаях их ошибка превысила 5% -ный уровень, принятый за допустимый в медицинской практике. Что же касается варианта 2 (Табл. 5), то точность классификации методом опорных векторов на тестовых выборках практически соизмерима с точностью НС [6]. Данный метод значительно выигрывает по времени обучения, но проигрывает по удобству его использования для предсказания классов биообъектов новых выборок.

Таким образом, видим, что использовать деревья для классификации биообъектов не всегда целесообразно, ввиду значительной ошибки, которая возникает в основном из-за большой вариабельности клеток, особенно при патологии, что приводит к пересечению классов в пространстве признаков. Действительно, при работе с природными объектами возникают ситуации, когда элементы выборки не поддаются классификации с использованием четких алгоритмов, особенно, когда объекты патологически изменены – типа клеточных структур материала мокроты, которые имеют почти одинаковые признаки, хотя относятся к разным классам. Кроме того, при компьютерной морфометрии биообъектов такие параметры, как размер, форма, цвет, плотность и т.д., могут быть определены с недостаточной точностью. Это происходит потому, что не всегда возможна качественная автоматическая сегментация, отсюда не точное выделение контура объекта и

вычисление его параметров, т.е. информация не полностью надежна, противоречива особенно при плохой подготовке препарата. Неопределенность, вызванная ошибками эксперимента, частично устраняется в системе «Морфолог». А вот природная флуктуация клеточных и тканевых структур, особенно при патологии, явно неустранима. Поэтому для классификации биообъектов, когда исходные данные являются ненадежными, слабо формализуемыми попробуем использовать нечеткие гибридные нейронные сети.

3. Нечеткие гибридные нейронные сети

Нейро-нечеткие сети призваны объединить достоинства и скомпенсировать недостатки систем нечеткого вывода и нейронных сетей. Выводы делаются на основе аппарата нечеткой логики, а соответствующие функции принадлежности объектов к определенному классу подстраиваются с использованием алгоритмов обучения нейронных сетей. Такие гибридные нейронные системы используют не только априорную информацию, но могут приобретать новые знания и что, не менее важно, для пользователя являются логически прозрачными. Это позволяет применять их в задачах управления и принятия решений в условиях неопределенности.

Гибридная сеть – это нейронная сеть с четкими входами, весами и функцией активации, с нечетким выводом и с логическими операциями, замененными t-нормами и t-конормами. Т-норма и t-конорма представляют собой нечеткие аналоги операций конъюнкции и дизъюнкции [4, 15-18].

В данной работе рассматривается нашедшая широкое применение гибридная сеть ANFIS (Adaptive Neuro-Fuzzy Inference System – Адаптивная сеть нечеткого вывода) из пакета Fuzzy Logic Toolbox системы MATLAB. Это - пятислойная сеть с N входами (в нашем случае 17 – по количеству переменных) и одним выходом. Она плохо приспособлена для задач классификации, так как у нее единственный нечеткий выход. Соответственно, напрямую сеть ANFIS можно использовать только в том случае, если классы линейно зависимы друг от друга (например «маленький», «средний», «большой»), или количество классов равно двум.

В рассматриваемой задаче классификации многомерных биообъектов количество классов равно десяти, и они не зависимы друг от друга. Поэтому предлагается разбить эту задачу на подзадачи, отделив каждый класс от остальных. Получится десять подзадач классификации с одним выходом, для решения которых можно применять гибридную сеть ANFIS. В этом случае проводилось обучение десяти гибридных сетей. На Рис. 3 показан пример автоматически сгенерированных правил для сети класса «плоский эпителий (ПЭ)», аналогично генерируются правила и для других классов. Здесь столбцы – переменные,

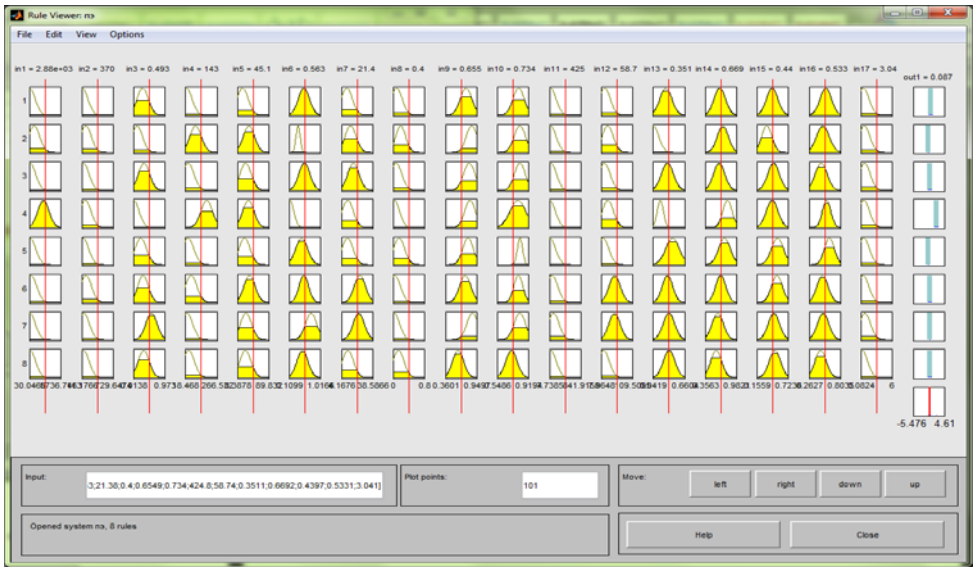

Рис. 3. Сгенерированные правила для каждой переменной сети класса «ПЭ»

а строки – правила. Для каждой переменной система сгенерировала 8 правил.

В сети ANFIS входы представляют собой вещественные числа в диапазоне 0-1, поэтому необходимо было произвести нормализацию входных данных. Также, лингвистические наименования классов были преобразованы в логические, а именно: 1 - объект принадлежит данному классу, 0 – не принадлежит. Таким образом, каждая сеть из 10 выполняет отдельно нечеткую классификацию на два класса.

На Рис. 4 показан график изменения ошибки обучения сгенерированной сети класса «ПЭ», где видно, что ошибка обучения перестает уменьшаться после 20 итерации. Ошибка при тестировании обученной сети класса «ПЭ» составила 0,076 (Рис. 5).

В Табл. 6 приведен фрагмент результирующей таблицы по тестовой выборке по 10 сетям. В первом столбце находится результат классификации объектов экспертом, а в последнем – результат классификации совокупностью гибридных сетей. В остальных столбцах – вывод каждой сети по каждому объекту. Поскольку вывод нечеткий, значения могут быть любые, поэтому примем, что класс со значением вывода, максимально близкий к единице, признается выбранным для данного объекта потому, что в каждой обучающей выборке признак принадлежности к текущему классу выбран «1», а не принадлежности - «0».

Преимущество выбранного подхода с множеством отдельных сетей заключается в том, что появляется возможность проанализировать

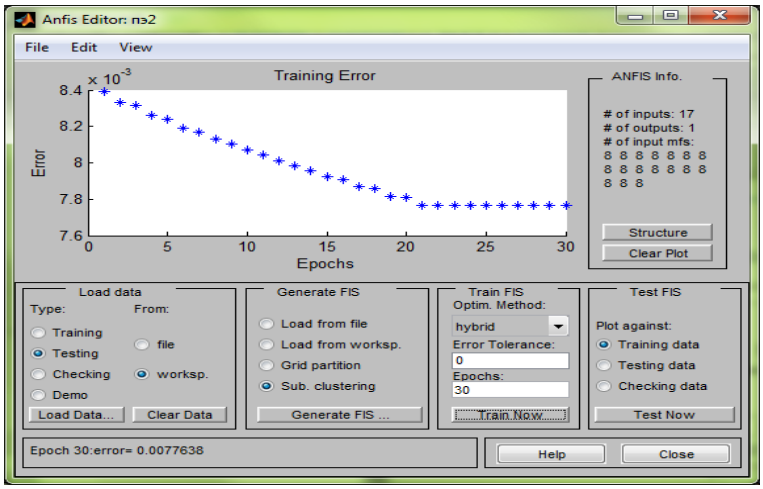

Рис. 4. График изменения ошибки обучения

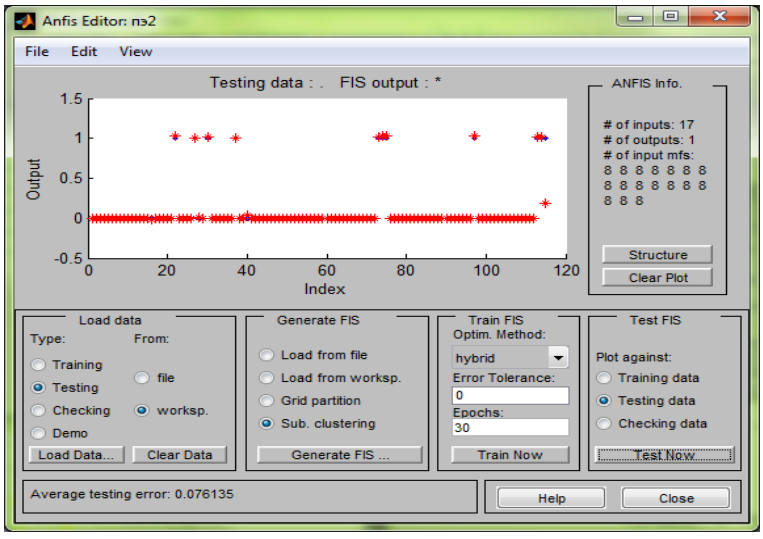

Рис. 5. Результат тестирования сети для класса клеток ПЭ

Табл. 6. Фрагмент результирующей таблицы классификации тестовой выборки

Эксперт	ПЭ	МФ	ЦЭ	КЭ	М	Д	ПР	ЖР	МР	КР	Выход
Д	0,0001	-0,0004	-0,0013	0,3895	0,0169	0,8446	-0,017	-0,023	-0,0731	0,0307	Д
ПЭ	0,9914	-0,0041	-0,0051	0,0027	0,0004	0,0126	-0,005	0,0014	0,0063	0,0325	ПЭ
М	-0,0173	0,0036	1,1296	0,0384	0,4811	0,01	-0,022	-0,176	0,008	-0,0128	ЦЭ
ЦЭ	0,0025	0,0041	1,0248	0,0179	-0,0311	0,0256	0,1167	0,1313	0,0275	-0,0194	ЦЭ

Табл.7. Фрагмент преобразованной таблицы классификации тестовой выборки

Эксперт	ПЭ	МФ	ЦЭ	КЭ	М	Д	ПР	ЖР	МР	КР	Выход
Д	15,54	15,53	15,52	**25,45**	15,28	**100,0**	15,28	15,19	14,48	16,03	Д
ПЭ	**100,0**	0,86	0,86	0,86	0,86	0,87	0,86	0,86	0,87	0,89	ПЭ
М	12,74	13,01	**100,0**	13,48	**24,98**	13,09	12,68	11,02	13,06	12,80	ЦЭ
ЦЭ	2,49	2,49	**100,0**	2,53	2,41	2,55	2,81	2,85	2,55	2,43	ЦЭ

Табл.8. Результаты классификации с использованием гибридной нейро-нечеткой сети

Итоги анализа	**Нейро-нечеткая сеть**	
	Обучение	Тест
Всего	1060	166
Правильно	1060	153
Ошибочно	0	13
% правильных	100	92,17
% ошибочных	0	7,83

точность классификации и выдать эксперту неопределенные объекты для коррекции. Для удобства анализа преобразуем результирующую таблицу к виду, представленному в Табл. 7. Здесь выходные значения сети нормализованы по максимуму близости к единице. Наиболее близкое к единице значение принято за 100%. Порог неопределенности выбирается эмпирически, в данном случае 20%. В таблице жирным шрифтом выделены значения, превышающие этот порог. Видно, что в первой строке возникла неопределенность, хотя вывод сети и правильный (Д), вероятность (КЭ) тоже следует учесть. В третьей строке вывод сети ошибочный (ЦЭ), но опять возникла неопределенность и правильный вывод (М) находится на втором месте по вероятности (24,98%).

При отсутствии эксперта и необходимости полностью автоматического вывода, объекты с неоднозначной классификацией признаются нераспознанными и не учитываются при дальнейшей диагностике. Результат классификации анализируемых выборок представлен в Табл.8. Точность классификации тестовой выборки данным методом практически сравнима с методом опорных векторов, но удобство интерпретации результатов и возможность оценки относительной вероятности принадлежности объекта к каждому классу делает этот метод интересным для дальнейших исследований.

Заключение

На основании проведенных экспериментальных исследований и анализа разных методов (деревья принятия решений, метод опорных векторов, гибридные нейро-нечеткие сети, а также дискриминантный анализ и нейронная сеть, изложенные в [6]) построения эффективной модели для распознавания и классификации с дифференцированным счетом многомерных объектов медико-биологической природы можно отметить, что

- для формирования общей выборки биообъектов с учителем - параметрических наблюдений, рекомендуется использовать дискрими-

нантный анализ, как наиболее быстродействующий и удобный инструмент не только для классификации биообъектов, но и для выявления диагностически значимых признаков и процедуры дискриминации клеточных структур;

- формировать обучающую и тестовую выборки эффективней случайным образом. Этот метод более быстродействующий и позволяет анализировать и распознавать объекты с вероятностью правильного принятия решений. Точность данного метода позволяет прогнозировать объекты с ошибкой допустимой в медицинской практике (5%);

- метод "градиентные растущие деревья" - быстродействующий, он позволяет классифицировать объекты с вероятностью правильного принятия решений, но ввиду отсутствия в алгоритме автоматического формирователя матрицы классификации новых объектов с дифференцированным счетом, рекомендуется его использовать только для предсказания одиночных или небольших наборов выборок;

- использовать на сегодняшний день метод опорных векторов в качестве классификатора было бы возможно, так как его точность на тестовых выборках практически соизмерима с точностью НС (их тестовая ошибка не превысила 5%-ный уровень). Однако реализация метода в программном пакете Statistica требует доработки для возможности автоматически прогнозировать принадлежность новых биообъектов к определенному классу, с препаратов не участвующих в обучении, причем с дифференцированным счетом и учетом неопределенных клеток;

- гибридные нейро-нечеткие сети, в предложенном варианте с разбиением на подсети по числу классов, позволяют гибко настраивать порог неопределенности и ранжировать неопределенные объекты по вероятности принадлежности к каждому классу; этот метод интересен для дальнейших исследований.

Поэтому, на наш взгляд, только нейронные сети можно, на сегодняшний день, использовать в качестве модели для распознавания и классификации множества многомерных объектов медико-биологической природы. Хотя, как уже говорилось, НС может долго обучаться и при этом не всегда возможно попадание в глобальный минимум, но она имеет лучшие показатели к обобщению с учетом дифференцированного счета объектов по классам, включая неопределенные клетки.

Литература

1. Дюк В.А., Самойленко А.П. Data Mining: учебный курс. СПб.: Питер, 2001. - 368 С.
2. Дюк В.А. Обработка данных на ПК в примерах. СПб.: Питер, 1997. - 368 С.
3. Безруков Н.С., Еремин Е.Л. Построение и моделирование адаптивной нейро-нечеткой системы в задаче медицинской диагностики // Медицинская информатика. 2005. № 2(10). С. 36 – 46.
4. Пермяков А.А., Юдицкий А.Д. Применение нейронечеткой системы ANFIS в анализе поведенческих показателей у животных в тесте «открытое поле» // Исследования в области естественных наук. Октябрь 2013. № 10. (электронный ресурс http://science.nauka.ru/2013/10/5995).
5. Данилова Ю.С., Коровин Е.Н. Поддержка принятия решения дифференциальной диагностики бронхиальной астмы на основе факторного и дискриминантного анализа // Системный анализ и управление в биомедицинских системах. 2013. Т.12. № 3. С. 856-861.
6. Попова Г.М., Дятчина И.Ф., Мельникова Н.В. Нейростатистическая модель классификации многомерных объектов медико-биологической природы // Искусственный интеллект и принятие решений. 2014. № 1. С. 66-78.
7. 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.
8. Breiman L., Friedman J., Stone C. J., Olshen R.A. Classification and Regression Trees. Taylor & Francis, 1984. 59 P.
9. Breiman L., Forests Random. Machine Learning, 2001. Vol. 45, No.1. P. 5-32.
10. Дружков П. Н., Половинкин А. Н. Программная реализация градиентного бустинга деревьев решений//Вестник Нижегородского государственного университета им Н. И. Лобачевского. 2011. №1. С. 193-200.
11. Вапник В. Н. Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979. -448 С.
12. Christopher M. Bishop. Pattern recognition and machine learning, Springer, 2006. -738 P.
13. Boser B., Guyon I., Vapnik V. In Fifth Annual Workshop on Computational Learning Theory, p. 144-152, Pittsburgh, ACM. 1992.
14. Cortes C., Vapnik V. Support-Vector Networks. Machine Learning, 1995. Vol. 20, No. 3. P. 273-297.
15. Штовба С.Д. Проектирование нечетких систем средствами MATLAB. М.: Горячая линия. Телеком, 2007. – 288 С.
16. Круглов В.В., Борисов В.В. Гибридные нейронные сети. Смоленск: Русич, 2001. 224 С.

17. Кореневский Н.А. Проектирование нечетких решающих сетей, настраиваемых по структуре данных для задач медицинской диагностики // Системный анализ и управление в биомедицинских системах. 2005. Т.4. № 1. С. 12-20.

18. Кореневский Н.А. Проектирование систем принятия решений на нечетких сетевых моделях в задачах медицинской диагностики и прогнозирования // Вестник новых медицинских технологий. 2006. Т.13, №2. С. 25-31.

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

Степанов Василий Николаевич. Старший научный сотрудник Института проблем управления им. В.А. Трапезникова РАН. Окончил Московский государственный институт радиотехники электроники и автоматики (технический университет) в 2001 году. Кандидат технических наук. Область научных интересов: анализ и обработка изображений, распознавание образов.E-mail: vnstepanov@yandex.ru