

## ДЕРЕВЬЯ КЛАССИФИКАЦИИ КАК МЕТОД ОПРЕДЕЛЕНИЯ ТЯЖЕСТИ БОЛЬНЫХ, СТРАДАЮЩИХ ХРОНИЧЕСКОЙ СЕРДЕЧНОЙ НЕДОСТАТОЧНОСТЬЮ

**А.А. Халафян** (к.т.н., доц. кафедры прикладной математики Кубанского государственного университета), **Л.Н. Елисеева** (д.м.н., зав. кафедрой факультетской терапии Кубанской государственной медицинской академии), **С.Г. Сафонова** (к.м.н., врач Центра медицинской профилактики департамента здравоохранения Краснодарского Края), **Савенко Д. В.** (аспирант Кубанского государственного университета).

Деревья классификации – это метод, позволяющий предсказывать принадлежность наблюдений или объектов к тому или иному классу зависимой категориальной переменной, измеренной в шкале наименований, в зависимости от соответствующих значений одной или нескольких независимых (предикторных) переменных.

В модуле *Деревья классификации* (*Classification Trees*) программы *STATISTICA* [1] с исчерпывающей полнотой реализованы методы построения бинарных деревьев классификации, основанных на ветвлении по одной переменной с использованием предикторных переменных различного типа.

Цель построения дерева классификации заключается в построении решающих правил, позволяющих классифицировать объекты. Способность деревьев классификации выполнять одномерное ветвление для анализа вклада отдельных переменных дает возможность работать с предикторными переменными различных типов – как с непрерывными, так и с категориальными. Деревья классификации могут быть, а иногда и бывают очень сложными. Однако использование специальных графических процедур, позволяет упростить интерпретацию результатов даже для очень сложных деревьев. Возможность графического представления результатов и простота интерпретации во многом объясняют большую популярность деревьев классификации в прикладных областях, в частности в медицине.

Следует заметить, что если в модуле *Дискриминантный анализ* предусмотрена возможность классификации программой нового наблюдения, в модуле *Деревья классификации* такой возможности нет. Пользователь вынужден по построенным программой решающим правилам самостоятельно отнести новое наблюдение к тому или иному классу. Также программа не определяет вероятностные характеристики принадлежности каждого наблюдения к классам и не указывает, в какой класс отнесено то или иное наблюдение.

Процесс вычисления (построения) дерева классификации состоит из четырех основных этапов:

- I. Выбор критерия точности прогноза;
- II. Выбор вариантов ветвления;
- III. Определение момента, когда дальнейшие ветвления следует прекратить;
- IV. Определение "подходящего размера" дерева.

Цель анализа с помощью деревьев классификации, в конечном счете, состоит в том, чтобы получить максимально точный прогноз. Наиболее точным прогнозом считается такой, который связан с наименьшей ценой. В большинстве методов цена – это просто доля неправильно классифицированных наблюдений. Поэтому, как правило, самый лучший прогноз – такой, который дает наименьший процент неправильных классификаций.

Первый этап анализа состоит в задании цены ошибок неправильной классификации. В модуле *Деревья классификации* предусмотрено два варианта задания цены неправильной классификации объектов, т.е. отнесения объекта, принадлежащего одному классу, к какому-то другому классу (рис.1). В первом варианте, в рамке *Цены ошибок классификации* выбирают опцию *Равные*. При этом цены назначаются одинаковые для всех классов, все внедиагональные элементы матрицы цен ошибок классификации (прогнозируемые классы – по строкам, наблюдаемые классы – по столбцам) полагаются равными 1. Во втором варианте выбирают опцию *Пользовательские*, для придания некоторым классам больших "весов", чем другим. Пользователь сам задает цены ошибок классификации.

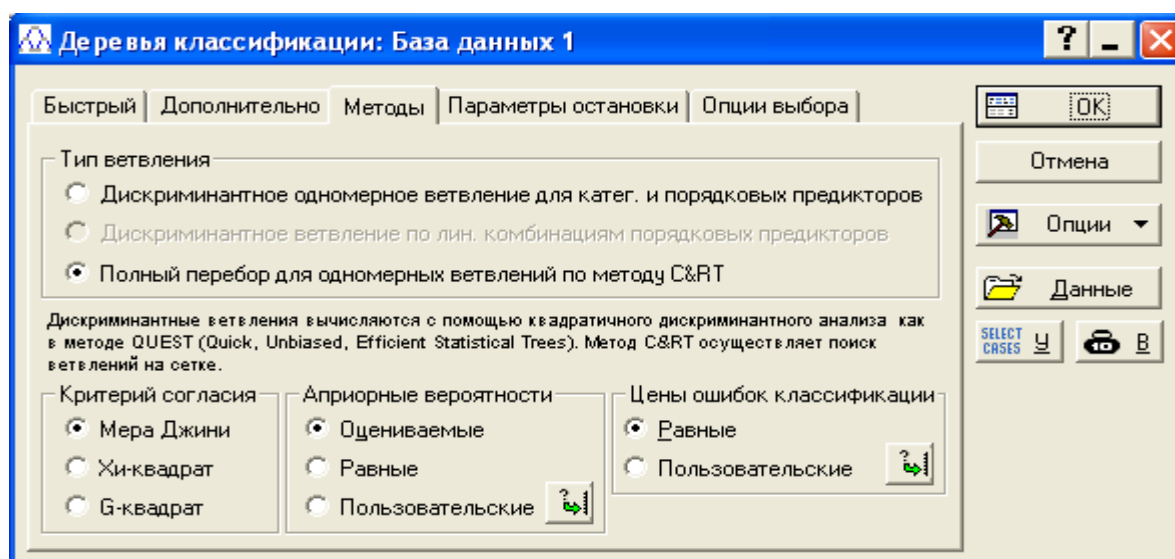


Рис.1

Априорные вероятности показывают, насколько мы, не зная ничего о значениях предикторных переменных, считаем вероятным, что объект будет принадлежать

определенному классу. Априорные вероятности выражают то, как, не располагая никакой априорной информацией о значениях предикторных переменных модели, мы оцениваем вероятность попадания объекта в тот или иной класс. Выбор априорных вероятностей, используемых для минимизации потерь, очень сильно влияет на результаты классификации. Если различия между исходными частотами в данной задаче не считаются существенными или если знаем заранее, что классы содержат примерно одинаковое количество наблюдений, то тогда можно взять *Равные* априорные вероятности. В случаях, когда исходные частоты связаны с размерами классов (например, при работе со случайной выборкой), следует в качестве оценок для априорных вероятностей взять относительные размеры классов в выборке. Для этого в рамке *Априорные вероятности* выбирают опцию *Оцениваемые*. Наконец, если (на основании данных предыдущих исследований) мы располагаем какой-то информацией об исходных частотах, то априорные вероятности нужно выбирать с учетом этой информации – опция *Пользовательские* (рис.1).

Второй этап анализа с помощью деревьев классификации заключается в том, чтобы выбрать способ ветвления по значениям переменных. В соответствии с иерархической природой деревьев классификации, такие ветвления производятся последовательно, начиная с корневой вершины, переходя к вершинам потомкам, пока дальнейшее ветвление не прекратится и "неразветвленные" вершины потомки окажутся терминальными. Терминальные вершины (или, как их иногда называют, листья) это узлы дерева, начиная с которых никакие решения больше не принимаются. На рисунках терминальные вершины показываются программой красными пунктирными линиями, а остальные – так называемые решающие вершины или вершины ветвления – сплошными черными линиями. Началом дерева считается самая верхняя решающая вершина, которую иногда также называют корнем дерева.

Различные методы выбора типа ветвления реализованы как опции в рамке *Тип ветвления* на вкладке *Методы* (рис.1). Первый метод *Дискриминантное одномерное ветвление* – можно использовать для категориальных, порядковых или смеси обоих типов предикторных переменных. Второй метод – *Дискриминантное многомерное ветвление по линейной комбинации* требует, чтобы предикторы были измерены как минимум в интервальной шкале. В обоих методах ветвления строятся с помощью квадратичного дискриминантного анализа. Третий тип ветвления – это *Полный перебор вариантов для одномерных ветвлений по методу C&RT*. Его, как и первый метод, можно использовать для категориальных, порядковых и интервальных или смеси трех типов предикторных переменных. В отличие от дискриминантных методов ветвления, в этом методе, для того

чтобы найти наилучший вариант ветвления, производится последовательный перебор всех возможных комбинаций уровней предикторных переменных. После выбора этой опции активируется рамка *Критерий согласия* (рис.2), где в качестве критерия согласия может быть выбрана одна из трех возможных мер: *мера Джини*, *Хи-квадрат* ( $\chi^2$ ), и *G-квадрат* ( $G^2$ ). Критерии согласия используются для выбора наилучшего из всех возможных вариантов ветвления.

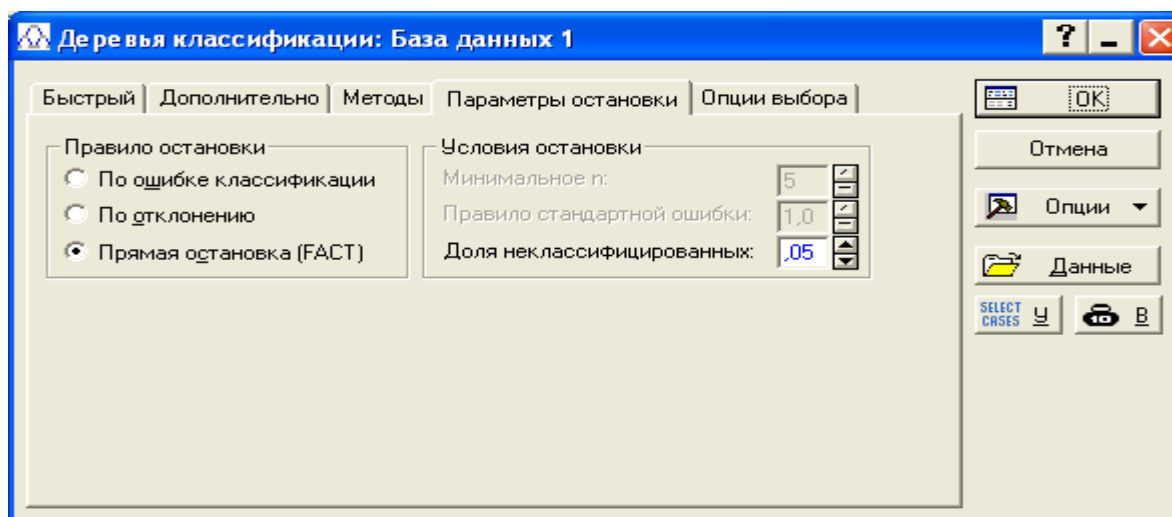


Рис.2

Третий этап анализа заключается в выборе момента, когда следует прекратить дальнейшие ветвления. Деревья классификации обладают тем свойством, что если не установлено ограничение на число ветвлений, то можно прийти к "чистой" классификации, когда каждая терминальная вершина содержит только один класс наблюдений (объектов). Но, как правило, данные о классификации зависимой переменной или уровни значений предикторных переменных содержат ошибки измерений или составляющую белого шума. Поэтому, было бы нереально пытаться продолжать сортировку до тех пор, пока каждая терминальная вершина не станет "чистой". В модуле на вкладке *Параметры остановки* в рамке *Правило остановки* реализованы три варианта остановки: *Отсечение по ошибке классификации*, *Отсечение по отклонению*, и *Прямая остановка по методу FACT*.

С определением момента, когда дальнейшие ветвления следует прекратить, непосредственно связан четвертый этап – определение "подходящих размеров" дерева. Очевидно, что чем больше размерность дерева классификации, тем точнее прогноз. Но сложнее интерпретация результатов и решающие правила, поэтому труднее пользователю сделать прогноз о принадлежности к классу нового наблюдения. Можно высказать ряд общих соображений о том, что следует считать "подходящими размерами" для дерева классификации. Дерево классификации должно быть достаточно сложным для того, чтобы учитывать имеющуюся информацию, и в то же время оно должно быть, как можно более простым для возможности интерпретировать результаты. Дерево должно уметь

использовать ту информацию, которая улучшает точность прогноза, и игнорировать ту информацию, которая прогноза не улучшает.

Одна из возможных стратегий выбора размера дерева состоит в том, чтобы наращивать его до нужного размера, который определяется самим пользователем на основе уже имеющихся данных, диагностических сообщений системы, выданных на предыдущих этапах анализа, или, на крайний случай, интуиции. По этой стратегии пользователь сам устанавливает размеры дерева, до которых оно может расти. В этом варианте в качестве правила остановки надо выбрать опцию *Прямая остановка Fact*, а затем задать при помощи опции *Доля неклассифицированных*, долю неправильно классифицированных, которая позволяет дереву расти до нужного размера.

В программе предусмотрено 3 способа оценки того, насколько удачно выбран пользователем размер дерева – 3 варианта кросс-проверки для построенного дерева классификации:

а). *Кросс-проверка на тестовой выборке* – наиболее предпочтительный вариант кросс-проверки. В этом варианте кросс-проверки дерево классификации строится по исходной – обучающей выборке, а его способность к прогнозированию проверяется путем предсказания классовой принадлежности элементов тестовой выборки. Если значение цены на тестовой выборке окажется больше, чем на обучающей выборке, то это свидетельствует о плохом результате кросс-проверки. Возможно, в этом случае следует поискать дерево другого размера, которое бы лучше выдерживало кросс-проверку. Цена – это доля неправильно классифицированных наблюдений при условии, что были использованы оцениваемые априорные вероятности, а цены ошибок классификации были взяты одинаковыми. Тестовая и обучающая выборки могут быть образованы из двух независимых наборов данных, или, если в нашем распоряжении имеется большая обучающая выборка, мы можем случайным образом отобрать часть (например, треть или половину) наблюдений и использовать ее в качестве тестовой выборки.

б). *V-кратная кросс-проверка*. Этот вид кросс-проверки целесообразно использовать в случаях, когда в нашем распоряжении нет отдельной тестовой выборки, а обучающее множество слишком мало для того, чтобы из него выделять тестовую выборку. Параметры кросс-проверки задаются на вкладке Опции выбора.

в). *Глобальная кросс-проверка*. В этом варианте производится заданное число итераций, причем всякий раз часть обучающей выборки (равная единице, деленной на заданное целое число) оставляется в стороне, а затем по очереди каждая из отложенных частей используется как тестовая выборка для кросс-проверки построенного дерева классификации. По умолчанию число итераций равно 3.

Вторая из возможных стратегий выбора "подходящего размера" для дерева – метод автоматического построения дерева Бримана (*Breiman*), который реализован *кросс-проверочным отсечением либо по минимальной цене-сложности*, либо по *минимальному отклонению-сложности*. Единственное различие между этими двумя опциями – способ измерения ошибки прогноза. При первой опции используется функция потерь, равная доли неправильно классифицированных объектов при оцениваемых априорных вероятностях и одинаковых ценах ошибок классификации. При второй опции используется мера, основанная на принципе максимума правдоподобия и называемая отклонением.

Для того чтобы в модуле *Дерева классификации* выполнить кросс–проверочное отсечение по минимальной цене–сложности, нужно выбрать опцию *По ошибке классификации* в качестве правила остановки. Кросс–проверочное отсечение по минимальному отклонению–сложности выполняется, если в качестве правила остановки выбрана опция *По отклонению*.

Рассмотрим применение метода *Дерева классификации* для оценки тяжести хронической сердечной недостаточности (ХСН), больных страдающих ишемической болезнью сердца [2].

Известен способ оценки тяжести хронической сердечной недостаточности (ХСН) в современной медицине с 2001г., основанный на применении классификации, предложенной Обществом Специалистов по Сердечной Недостаточности. При оценке стадий ХСН применяют понятия, например “адаптивное” и “дезадаптивное ремоделирование сердца и сосудов”, а также “финальная стадия ремоделирования органов”. За этими терминами стоит глубокий патофизиологический смысл, что приводит к неоднозначности классификации.

Альтернативным количественным способом градации ХСН может служить 6-мин. тест ходьбы, позволяющий по дальности расстояния, пройденного пациентом за 6 минут, установить функциональный класс (ФК) (табл.1), соответствующий степени тяжести хронической сердечной недостаточности [2].

Таблица 1. 6-минутный тест

ФК ХСН	расстояние расч. (м)
класс 4	0 – 200
класс 3	201 – 300
класс 2	301 – 400
класс 1	401 – 500
класс 0	более 500

Используемый 6-мин. тест приближен к повседневной физической активности пациентов, прост в исполнении, не требует специального оборудования, безопасен и хорошо воспроизводим. С его помощью можно судить не только о физической активности, толерантности пациентов с ХСН к физическим нагрузкам, но и контролировать эффективность проводимой терапии. Несмотря на простоту и доступность этого метода, результаты тестирования не могут в полном объеме оценить функциональное состояние больных, имеющих заболевания суставов, дефекты статики. Кроме того, при проведении 6-мин. теста ходьбы не учитываются индивидуальные психологические особенности личности больных, их волевые качества, а также нозологическая принадлежность заболевания сердечно-сосудистой системы, приведшая к ХСН. Указанные факты снижают эффективность и адекватность данного метода и приводят к неизбежным погрешностям при тестировании и, как следствие, к ошибкам в определении тяжести ХСН больных, а следовательно к неправильно выбранной тактике и стратегии лечения.

С учетом вышеизложенного, предложены другие альтернативные способы диагностики тяжести ХСН, которые используют методы математического моделирования. Эти методы позволяют более объективно подойти к оценке тяжести ХСН у больных ишемической болезнью сердца. Так, например, известен способ оценки начальных стадий тяжести сердечной недостаточности [3], заключающийся в измерении у больных ИБС стандартных параметров. Учитывают факторы, влияющие на сократимость миокарда и такие показатели, как 6-минутный тест, тредмил-тест, разность между текущими и должными показателями, частоты сердечных сокращений. Вычисляют показатель характеризующий индивидуальную гемодинамику. Рассчитывают диагностические индексы для каждого показателя. Тяжесть начальных стадий сердечной недостаточности вычисляют по сумме диагностических индексов. В соответствии с данным способом можно заболевания классифицировать как I-ю стадию самую легкую, II стадию, промежуточную между I и III стадиями. Проявлений у ХСН значительно больше и разнообразнее, а способ этого не учитывает.

Другой способ, использующий математические методы оценки тяжести сердечной недостаточности [4], состоит в определении у больного конечного диастолического размера левого желудочка, конечного диастолического размера левого предсердия, конечного систолического размера левого желудочка, диастолического артериального давления. Оценивают характер слабости и одышки. Каждый параметр приводят к балльной оценке. Полученные балльные значения параметров суммируют. Степень тяжести сердечной недостаточности или ее отсутствие определяют по общему суммарному

значению параметров в балах. Способ позволяет стандартизовать степень тяжести сердечной недостаточности, объективно оценить качество проводимой терапии. Основным недостатком способа является сложность математических расчетов и неоднозначность в балльной оценке параметров болезни.

Предложен способ оценки тяжести ХСН на основе дискриминантного анализа [2], метод реализован в виде программного продукта [5]. Недостатком метода является то, что необходимо выполнение определенных требований к данным, например переменные (признаки, характеризующие состояние больного) должны быть измерены минимум в интервальной шкале, с законом распределения, соответствующим нормальному закону. В то же время среди переменных, характеризующих состояние больного, присутствуют качественные (категориальные) переменные – слабость, одышка.

Следовательно, существующие современные подходы к оценке тяжести ХСН имеют свои недостатки, что сохраняет актуальность дальнейшего поиска способов ее диагностики. В [2] показано, что для классификации больных по степени тяжести ХСН достаточно воспользоваться такими показателями как конечный диастолический (КДР) и систолический (КСР) размеры левого предсердия (ЛП), диастолическое артериальное давление (АДД), клиническими проявлениями болезни – слабости, одышки. Для оценки слабости, одышки использовали порядковую шкалу: нет, слабая, умеренная, выраженная, сильно выраженная. В исследованиях участвовали 119 пациентов в возрасте от 39 до 75 лет с признаками ХСН. Измеряли диастолическое артериальное давление, при помощи эхо кардиографа – показатели центральной гемодинамики: конечный диастолический и конечный систолический размеры левого желудочка и левого предсердия. Каждый больной прошел 6-минутный тест. При обследовании осуществляли оценку клинических проявлений болезни – слабость, одышку.

При построении (поиске) дерева классификации подходящего размера с минимальной ошибкой классификации были реализованы все методы ветвления при различных комбинациях основных установок. Критериями качества дерева классификации были выбраны: количество ошибок классификации на обучающей выборке, соотношение и значение цен кросс-проверки и глобальной кросс-проверки. Количество терминальных вершин в качестве критерия не рассматривали, так как предполагалось реализовать процедуру классификации по условиям ветвления в виде программного продукта.

«Наилучшее» дерево удалось построить при следующих установках (рис.1, 2): метод ветвления – *Полный перебор вариантов для одномерных ветвлений по методу C&RT*, критерий согласия – *мера Джини*; априорные вероятности – *оцениваемые*; цены ошибок



классификации – *равные*; правило остановки – *прямая остановка (Fact)*; условия остановки – *доля неклассифицированных 0,05*.

К сожалению, ограниченность формата страницы не позволяет привести рисунок графа с надписями над узлами (вершинами) и ребрами дерева.

На рис.3 представлен фрагмент таблицы результатов, в которой выведена вся информация о графе дерева: номера узлов (вершин); номера дочерних вершин на левой и правой ветвях; исходное количество объектов в классах; предсказанные классы; условия ветвления. Например, из таблицы следует, что левые ветви содержит узлы под четными номерами 2, 4,..., правые – под нечетными номерами 3, 5,... Далее, из строки 1 таблицы следует, что в первой вершине все больные (класс 0 – 14 больных, класс 1 – 56 больных, класс 2 – 16 больных, класс 3 – 19 больных, класс 4 – 14 больных), классифицированы (предсказаны) как принадлежащие классу 1. Из вершины 1 выходят две ветви (правая и левая) с соответствующими вершинами 2 и 3. Условие разделения больных по вершинам 2 и 3 следующее: если значение переменной *одышка* = *слабая, нет*, то больные классифицируются как класс 1 (вершина 2, 73 больных ( $73 = 14 + 55 + 4$ )), в противном случае – как класс 3 (вершина 3, 46 больных). Из строки 2 следует, что по условию *одышка* – *нет*, 73 больных из вершины 2 классифицируются как класс 0 (вершина 4, 23 больных) и класс 1 (вершина 5, 50 больных). Из строки 3 следует, что по условию *КДР*  $\leq 5,65$ , больные (46 чел.) из вершины 3 ( $1+12+19+14$ ) классифицированы как класс 2 (вершина 6, 10 больных) и класс 3 (вершина 7, 36 больных). Таким образом, можно пройти по всем вершинам дерева.

Узел	Структура (База данных 1)										
	Левая ветвь	Правая ветвь	Класс 0	Класс 1	Класс 2	Класс 3	Класс 4	Предс. класс	Ветвл.по постоян.	Ветвл.по перемен.	Ветвл.по категор.
1	2	3	14	56	16	19	14	1		одышка	слабая
2	4	5	14	55	4	0	0	1		одышка	нет
3	6	7	0	1	12	19	14	3	-5,65	КДР	
4	8	9	12	11	0	0	0	0	-5,45	КДР	
5	10	11	2	44	4	0	0	1	-4,55	ЛП	
6	12	13	0	1	9	0	0	2	-85,	АДД	
7	14	15	0	0	3	19	14	3	-4,85	КСР	
8	16	17	12	2	0	0	0	0	-4,15	ЛП	
9			0	9	0	0	0	1			
10	18	19	2	44	2	0	0	1	-4,75	КДР	
11			0	0	2	0	0	2			
12			0	1	0	0	0	1			
13			0	0	9	0	0	2			
14	20	21	0	0	3	15	3	3		слабость	слабая
15	22	23	0	0	0	4	11	4	-4,35	ЛП	
16			11	0	0	0	0	0			
17	24	25	1	2	0	0	0	1	-4,75	КДР	
18	26	27	2	1	0	0	0	0	-3,9	ЛП	
19	28	29	0	43	2	0	0	1	-6,45	КДР	

Рис.4

Решающие правила также прописаны в таблице результатов *Условия ветвления*, на рис.5 приведен фрагмент таблицы. Для терминальных вершин дерева никакой информации об условиях ветвления не выдается.

Узел	Усл. ветвления (База данных 1) Усл. ветвления по всем верш.			
	Ветвл.по постоян.	Ветвл.по перемен.	Ветвл.по категор.	Ветвл.по категор.
1		одышка	слабая	нет
2		одышка	нет	
3	-5,65	КДР		
4	-5,45	КДР		
5	-4,55	ЛП		
6	-85,	АДД		
7	-4,85	КСР		
8	-4,15	ЛП		
9				
10	-4,75	КДР		

Рис.5

В процессе анализа программа строит различные деревья классификации и из них выбирает дерево подходящего размера. Если в качестве правила остановки было выбрано отсечение по ошибке классификации или по вариации (по отклонению), то в таблице результатов выводятся: *номера* терминальных вершин; *КП* – цена кросс-проверки; ее *стандартная ошибка*; *цена обучения*; *сложность* каждого из усеченных деревьев. Если в качестве *Правил остановки* была выбрана *Прямая остановка (FACT)*, то вся указанная информация выводится для выбранного дерева «подходящего размера». В таблице (рис.6) приведены результаты для дерева под номером 1, которое программа определила как

дерево «подходящего размера» и пометила звездочкой (*Деревья классификации – Результаты/Структура/Последовательность цен*).

Номер деревя	Посл. деревья (База данных 1) Последовательные деревья * - выбранное дерево				
	N терм. вершины	КП цена	Станд. ошибка	Обучения цена	Вершина сложност
*1	25	0,21008	0,04102	0,00	0,00

Рис.6

Можно построить график последовательности цен. Если было выбрано правило остановки отсечением по ошибке классификации или по вариации, на графике изображаются цена кросс-проверки и цена обучения для каждого дерева из последовательности усеченных деревьев. Если в качестве *Правил остановки* была выбрана *Прямая остановка (FACT)*, то эта информация выводится для выбранного дерева "подходящего размера".

Из таблицы на рис.6 видно, что при построении дерева классификации цена обучения равна 0, т.е. доля ошибочной классификации равна 0, а при *V*-кратной кросс-проверке средняя доля ошибочно классифицированных наблюдений достигла  $\approx 20\%$ .

Если выбрано *дискриминантное одномерное ветвление*, или *полный перебор* можно оценить значимость предикторов в процессе классификации. В таблице (рис.7) для каждой из анализируемых предикторных переменных устанавливается ранг ее значимости по 100-бальной шкале (*Деревья классификации – Результаты/Структура/Значимость предиктора*). Из данной таблицы следует, что наиболее значимым предиктором для классификации является переменная *КДР*, далее – *КСР*, наименее значимым – *слабость*. Это значит, что конечный диастолический (*КДР*) и систолический (*КСР*) размеры левого желудка играют ведущую роль при классификации больных методом *Деревья классификации*.

	Ранги
Перемен.	Ранг
слабость	50
одышка	67
КСР	91
КДР	100
ЛП	83
АДД	57

Рис.7

Если расположить предикторы в порядке убывания значимости, то получим последовательность *КДР, КСР, ЛП, одышка, АДД, слабость*.

На рис.8 представлена таблица сопряженности предсказанных и наблюдаемых (исходных) классов больных (*Анализ/Основные статистики/Таблицы сопряженности, флагов и заголовки/Двухвходовые таблицы*). Например, из таблицы следует, что 14 больных класса 1 были правильно классифицированы (предсказаны) как больные класса 1, из строки 2 следует, что 56 больных класса 2 были правильно классифицированы (предсказаны) как больные класса 2 и т.д., больные всех классов классифицированы верно.

Класс					
	Класс 0	Класс 1	Класс 2	Класс 3	Класс 4
0	14	0	0	0	0
1	0	56	0	0	0
2	0	0	16	0	0
3	0	0	0	19	0
4	0	0	0	0	14

Рис.8

Программа позволяет визуализировать ошибки классификации (*Деревья классификации – Результаты/Классификация/Ошибки классификации*). Как видно из табл. на рис.9 ошибки классификации отсутствуют. Построенное дерево классификации успешно классифицировало больных, представленных в исходной таблице данных – все больные классифицированы верно, но оказалась достаточно высокой цена кросс-проверки, которая составила  $\approx 21\%$ .

Класс	Ошибки классиф. на обуч. выборке Предсказ. (строки) и наблюд. Объем обуч. выб. N= 119				
	Класс 0	Класс 1	Класс 2	Класс 3	Класс 4
0		0	0	0	0
1	0		0	0	0
2	0	0		0	0
3	0	0	0		0
4	0	0	0	0	

Рис.9

Для того, чтобы проверить насколько успешным будет классификация больных по произвольной таблице данных, обратимся к процедуре глобальной кросс-проверки.

В информационном поле таблицы (рис.10) указана цена глобальной кросс-проверки (*Цена глоб. КП*) – 0,15966 и стандартное отклонение цены (ст.откл.цены) – 0,03358, которая характеризует степень разброса цены. Как видно из значения цены глобальной кросс-проверки, она значительно меньше цены кросс-проверки (*цена КП*  $\approx 0,21$ , см. рис.6) для выбранного дерева. При этом мала величина разброса цены глобальной кросс-проверки. Это значит, что процедуре "автоматического" отбора дерева удалось выбрать

дерево с ошибкой, близкой к минимальной, и ожидаемая доля неверно классифицированных больных по построенному дереву классификации составит  $\approx 15,6\%$ .

Класс	Матрица ошибок классиф. ГКП Предсказ. (строки) и наблюдаемые Цена глоб. КП = ,15966;				
	Класс 0	Класс 1	Класс 2	Класс 3	Класс 4
0		1	0	0	0
1	2		3	0	0
2	0	1		2	0
3	0	0	2		4
4	0	0	0	4	

Рис. 10

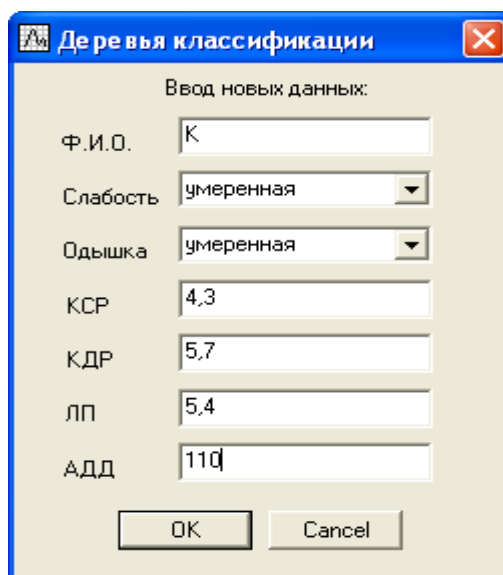
Недостатком построенного дерева является то, что высока цена КП, достоинством является то, что *цена глоб. КП* меньше цены кросс-проверки, а цена обучения равна 0.

Существенным недостатком модуля *Дерева классификации* является отсутствие информации о распределении наблюдений по классам в процессе классификации, т.е. неизвестно какое наблюдение в какой класс попало. Напомним, что такие возможности представлены в модулях *Кластерный анализ* и *Дискриминантный анализ*. Также не предусмотрена возможность классификации нового наблюдения – такая возможность есть в дискриминантном анализе. Поэтому, если дерево классификации небольшое (несложное), для того чтобы классифицировать наблюдение необходимо по условиям ветвления «вручную» определить класс наблюдения, либо, при сложном дереве написать программу, реализующую условия ветвления. Построенное дерево классификации для файла данных больных ХСН достаточно сложное и определить ФК ХСН «вручную» по условиям ветвления не представляется возможным, да и с точки зрения практической медицины даже при простом дереве классификации, такой способ классификации неприемлем. Поэтому по условиям ветвления была написана программа классификации больных ХСН. Программа имеет простой интерфейс, окно ввода данных открывается нажатием кнопки «Макрос ДК», установленной на панели инструментов программы

STATISTICA   Макрос ДК .

*Пример.* Больной К. предъявлял жалобы на одышку (умеренная), слабость (умеренная), периодические подъемы АД до 240/120 mm.Hg, давящими болями в области сердца при физической нагрузке. Результат 6-минутного теста – 250 м, что соответствует ФК ХСН 3. Данные инструментальных исследований: ЭХО-КС: КДР – 5,7 см, КСР – 4,3 см, ЛП – 5,4 см. Используемые в программе показатели больного, были введены в окно ввода данных (рис.11). Результат классификации приведен на рис. 12, из которого следует, что

программа определила принадлежность больного к классу 3, что соответствует классификации по 6-мин. тесту.



Дерева классификации

Ввод новых данных:

Ф.И.О. K

Слабость умеренная

Одышка умеренная

КСР 4.3

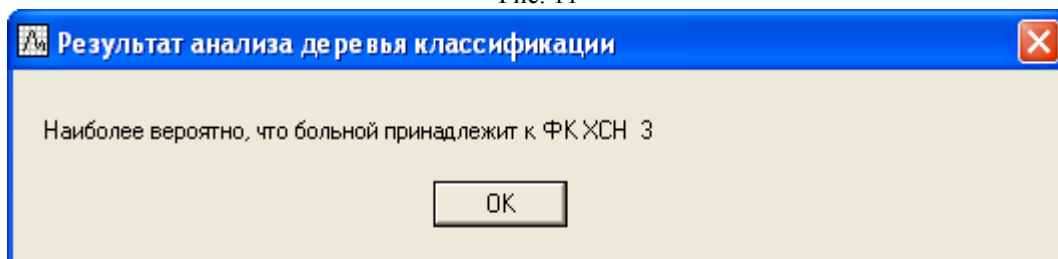
КДР 5.7

ЛП 5.4

АДД 110

OK Cancel

Рис. 11



Результат анализа дерева классификации

Наиболее вероятно, что больной принадлежит к ФК ХСН 3

OK

Рис. 12

Таким образом, в процессе реализации модуля *Дерева классификации* удалось построить граф дерева классификации. При помощи решающих правил данного дерева можно определить принадлежность больного к тому или иному функциональному классу. Граф дерева классификации дает наглядное графическое представление процедуры классификации. Важным, с точки зрения медицины, является ранжирование признаков, характеризующих состояние больных по их значимости в процедуре классификации. Так, из результатов ранжирования следует, что слабость является менее значимым признаком, характеризующим принадлежность больного к тому или иному ФК ХСН, чем одышка, а АДД является менее значимым признаком, характеризующим принадлежность больного к тому или иному ФК ХСН, чем КДР, КСР, ЛП.

Предлагаемый способ оценки тяжести сердечной недостаточности при использовании позволяет:

- повысить качество диагностической процедуры оценки тяжести сердечной недостаточности;
- снизить частоту ошибок при верификации степени тяжести хронической сердечной недостаточности;

– позволяет ускорить процесс определения тяжести ХСН в повседневной врачебной практике;

– выработать оптимальные для каждого больного стратегию и тактику лечения.

#### Литература

1. StatSoft.Inc.(2001). Электронный учебник по статистике. М.: StatSoft.WEB: <http://www.StatSoft.ru/home/textbook/default.htm>.
  2. Елисеева Л.Н., Халафян А.А., Сафонова С.Г. Применение методов классификационного анализа для определения функционального класса хронической сердечной недостаточности больных, страдающих ишемической болезнью сердца // Экологический вестник научных центров Черноморского экономического сотрудничества. 2005. № 3. С.78-93.
  3. Пат. РФ №2223030, МПК(7) А61В5/02. 4.04.2002. Способ оценки начальных стадий сердечной недостаточности / Елисеева Л.Н., Сафонова С.Г., Складенко В.Н.
  4. Пат. РФ №2266039, МПК(7) А61В5/02. 20.12.2005. Способ оценки тяжести сердечной недостаточности / Елисеева Л.Н., Сафонова С.Г., Лаптев В.Н.
  5. Свидетельство об официальной регистрации программы для ЭВМ № 20066101235. Федеральная служба по интеллектуальной собственности, патентам и товарным знакам. 10.01. 06. Определение функционального класса хронической сердечной недостаточности у больных, страдающих ишемической болезнью сердца. / Халафян А.А., Елисеева Л.Н., Сафонова С.Г., Фищенко Д.В.
- Успехи современного естествознания. 2006. № 11. С. 16–18.