М. П. Кривенко, Критерии значимости отбора признаков классификации, *Информ. и её примен.*, 2016, том 10, выпуск 3, 32–40

DOI: https://doi.org/10.14357/19922264160305

ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2016. Т. 10. Вып. 3. С. 32–40

# КРИТЕРИИ ЗНАЧИМОСТИ ОТБОРА ПРИЗНАКОВ КЛАССИФИКАЦИИ

М. П. Кривенко[1]

**Аннотация:** Рассмотрена задача отбора признаков классификации и вопросы оценивания качества получаемых решений. Среди различных методов отбора признаков внимание обращено на последовательные процедуры; мерой качества классификации выбрана вероятность правильной классификации. Для оценивания этой характеристики предложено использовать метод перепроверки и бутстреп-метод, а для исследования ряда выборочных значений — сравнительный анализ доверительных интервалов и критерии однородности биномиальных пропорций. При построении байесовского классификатора для данных применялась модель смеси нормальных распределений; ее параметры оценивались с помощью EM (expectation–maximization) алгоритма. В качестве эксперимента рассмотрена задача обоснованного выбора признаков классификации при прогнозировании типа мочевых камней в урологии. Показано, что сокращать совокупность анализируемых показателей можно не только без потери качества принимаемых решений, но и с повышением вероятности правильного прогноза типа камня.

**Ключевые слова:** селекция признаков; последовательная селекция вперед и назад; байесовская классификация данных; проверка однородности двоичных последовательностей; прогноз типа камня в урологии

**DOI:** 10.14357/19922264160305

## 1 Введение

Сокращение числа переменных (признаков) может привести к повышению эффективности классификации данных и к более глубокому пониманию их природы. Соответствующие постановки задач и применяемые методы востребованы на практике в весьма разнообразных областях распознавания образов и машинного обучения: категоризации текстов, дистанционном сканировании, обнаружении наркотиков, маркетинге, обработке речи, распознавании рукописных символов, медицине и т. д. Особый интерес вызывают задачи, когда данные оказываются высокоразмерными, а объем обучающей выборки относительно мал. Получить предварительное представление о соответствующей проблеме можно из разд. 10 [1], примерами обстоятельных обзоров могут служить работы [2, 3].

Причины, по которым приходится корректировать совокупность анализируемых признаков, могут заключаться в следующем:

- повышение эффективности обучения и применения классификатора путем сокращения вычислительных затрат (уменьшение времени обработки и освобождение памяти от хранения ненужных атрибутов);
- снижение стоимости последующего сбора данных за счет измерения только тех переменных, которые имеют отношение к распознаванию;
- создание предпосылок для повышения качества классификатора;
- упрощение описания классификатора благодаря более ясному пониманию природы данных и структуры модели;
- предоставление возможности специалистам в предметной области прояснить суть протекающих процессов.

## 2 Задачи и методы селекции признаков

Сокращение числа признаков может осуществляться либо путем отбора (селекции, выбора подмножества исходных переменных), либо путем извлечения (формирования, определения линейного или нелинейного преобразования совокупности исходных переменных для получения меньшего набора). Далее речь пойдет только о методах отбора, так как упрощение модели данных и повышение качества классификации ставятся в данной работе во главу угла.

Постановка задачи построения методов селекции подразумевает оптимизацию некоторой целевой функции. В зависимости от того, как связаны задачи предварительной обработки состава анализируемых признаков и задачи классификации дан-

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, mkrivenko@ipiran.ru

ных, обычно выделяются две основные категории методов:

(1) не связанные со свойствами классификатора, их обычно называют методами фильтрации;

(2) учитывающие характеристики последующей классификации данных, к ним относятся настраивающие классификатор методы (НКл-методы) и формирующие классификатор методы (ФКл-методы).

Понятно, что от второй группы методов следует ждать лучшие по сравнению с методами фильтрации результаты. Настраивающие классификатор методы (wrapper methods) построены так, что подмножества признаков оцениваются с учетом прогностической точности алгоритма классификации, т. е. подход зависит от классификатора. Вычислительно они более требовательны, чем методы фильтрации. Формирующие классификатор методы (embedded methods) отличаются тем, что поиск оптимального набора признаков встроен в структуру классификатора, а не реализуется отдельно от построения классификатора. Подход зависит от классификатора, и его можно рассматривать как поиск решения в объединенном пространстве признаков и моделей классификатора.

Обычно многие признаки оказываются неинформативными (не способствующими классификации) или избыточными (дублирующими, существенно связанными между собой признаками). Классификация может стать более продуктивной и эффективной только при использовании релевантных и неизбыточных признаков [4].

В процессе итерационного отбора признаков есть два ключевых шага — оценивание и генерирование подмножества признаков, суть которых состоит в следующем:

– оценивание определяет качество некоторого подмножества признаков; НКл- и ФКл-методы, как правило, используют меры, основанные на производительности классификатора; методы фильтрации — меры, основанные на свойствах данных;

– генерирование является средством создания некоторого подмножества признаков; процедура может основываться на простом ранжировании отдельных признаков или заключаться в коррекции текущего состава подмножества путем добавления или удаления элементов.

Для того чтобы выбрать подходящий набор признаков, необходимо средство измерения способности признака внести свой вклад в отделимость классов либо индивидуально, либо в контексте других уже выбранных признаков, т. е. необходимы средства измерения значимости и избыточности. Меры, учитывающие правила классификации, подразумевают использование выделенного набора признаков при формировании и применении классификатора, т. е. отличающиеся наборы признаков могут давать различные классификаторы. Широко используемым примером подобных мер может служить вероятность ошибок классификатора и оценка ее значения с помощью частоты появления ошибок классификатора, обученного с помощью выбранного набора признаков.

Существуют три основные категории алгоритмов поиска для выбора подмножества признаков: гарантированный, последовательный и случайный [2].

Гарантированный поиск обеспечивает нахождение оптимального (в смысле заданного критерия) подмножества признаков. Конечно, полный (исчерпывающий) поиск является гарантированным, но стратегия поиска необязательно должна быть исчерпывающей, чтобы стать гарантированной (например, метод ветвей и границ гарантирует получение решения).

Последовательный поиск: признаки добавляются или удаляются последовательно (последовательный вперед или назад отбор). Такие методы не являются оптимальными, но они просты в реализации и быстро дают результат.

Случайный поиск подразумевает подходы, которые используют случайные механизмы при реализации указанных выше способов. Встраивание случайности в методы отбора признаков может быть полезным в следующих случаях: набор возможных подмножеств признаков велик и не доступен для обработки; детерминированные алгоритмы оказываются подвержены попаданию в ловушку локальных экстремумов критерия отбора признаков; когда выгоды от получения хорошего решения значительно перевешивают возникающие затраты, т. е. стоит потратить время на привлечение случайного механизма, чтобы перепроверить полученные результаты. Варианты случайных процедур отбора признаков приведены в [5].

По ряду причин обычно приходится отказываться от гарантированного поиска в пользу субоптимальных методов (последовательный и случайный поиск): высокие вычислительные сложности получения оптимальных решений, не всегда выполняются условия того, чтобы неисчерпывающий поиск становился гарантированным (например, условие монотонности для метода ветвей и границ).

Последовательный поиск включает в первую очередь последовательный отбор вперед (sequential forward selection — SFS) и последовательный отбор

назад (sequential backward selection — SBS), а также различные обобщения этих двух методов.

Метод SFS итерационно добавляет по одному новые признаки к уже сформированному набору, руководствуясь качеством получаемого решения. Основным недостатком этого метода является то, что он не включает в себя механизм для удаления ранее добавленных признаков из уже сформированного на очередном шаге набора признаков, хотя добавления могут сделать ненужными признаки, ранее уже вошедшие в формируемый набор. Метод SBS отличается от SFS тем, что признаки не добавляются, а удаляются.

В данной работе в качестве классификатора рассматривается байесовская решающая процедура, причем модель данных — смесь нормальных распределений. В качестве показателя эффективности классификации выступает вероятность правильной классификации $P_c$. Оценка этого показателя есть частота правильной классификации при многократном применении классификатора. Из-за высокой сложности постановок реальных задач (многомерность данных, применение итерационного EM-алгоритма для оценивания параметров смеси, ограниченность объема обучающих данных) при оценивании показателя приходится прибегать к методам управления обработкой выборки.

Для селекции признаков предлагается применять последовательный метод с настройкой классификатора для каждого апробируемого набора признаков. При этом использовался как метод SBS, так и SFS. Выбор именно селекции и последовательных методов отбора объясняется рядом причин:

- простота реализации,
- на начальном, поисковом, этапе исследований порождаются информативные наборы признаков, обеспечивающие возможность интерпретации полученных решений, в частности с позиций качества классификатора появляется возможность высказать предположения об эффективной размерности пространства признаков, выделить наименее и наиболее информативные признаки, увериться или усомниться относительно сложившихся в предметной области моделей данных.

## 3 Критерии значимости

Последовательные методы отбора включают генерирование варианта набора признаков и оценивание для него вероятности правильной классификации. Таким образом, при использовании определенного метода управления обработкой выборки получаем бинарную последовательность — последовательность испытаний с двумя исходами («успех», когда классификатор не ошибся, «неуспех» в противном случае). При применении бутстреп-метода речь идет о последовательности испытаний Бернулли, тогда оценка вероятности правильной классификации есть случайная величина, имеющая биномиальное распределение. Если же привлекается метод перепроверки, то использование биномиального распределения даст некоторое приближение для распределения оценки вероятности правильной классификации, позволяющее получить первое представление о реальном качестве анализируемого набора признаков. Для обоих методов в ходе их реализации возникают последовательности наблюденных значений, которые могут стать источником непараметрических оценок для требуемых величин.

Дадим постановки возможных задач анализа, возникающих в связи с анализом результатов отбора признаков. Имеется $k$ независимых биномиальных популяций $X_1, \ldots, X_k$, распределения которых

$$X_i \sim \mathrm{Binomial}\,(n_i, \pi_i)\ , \quad i = 1, \ldots, k\,,$$

где $n_1, \ldots, n_k$ известны и, возможно, различны, а вероятности успеха $0 \leq \pi_1, \ldots, \pi_k \leq 1$ неизвестны. Со значением $i$ связаны следующие действия: выбор определенного набора признаков, построение соответствующего классификатора на основе смеси распределений, оценивание по $n_i$ наблюденным значениям вероятности правильной классификации $\pi_i$.

Для определенного значения $i$ оценка максимального правдоподобия вероятности успеха есть $\hat{\pi}_i = X_i/n_i$. При больших значениях $n_i$ и условиях, что ни $\pi_i$, ни $1 - \pi_i$ не являются малыми величинами, $\hat{\pi}_i$ имеет приблизительно нормальное распределение. Более практичной в случае сравнительного анализа результатов отбора признаков оказывается интервальная оценка для $\pi_i$. Используемый для этих целей доверительный интервал может быть двухсторонний или односторонний, а также принимать различную форму [6]. Наиболее распространенным при построении доверительного интервала является использование нормальной аппроксимации. При этом в случае одностороннего интервала из-за асимметрии истинного распределения ошибка может оказаться существенной, в [7] дается простое и эффективное решение этой проблемы.

Пусть классификатор оперирует с $M$ классами, вероятности появления которых равны $p_1, \ldots, p_M$. На практике нелишней оказывается проверка того, отличается ли построенная классификация от

действий «наугад», т. е. отнесения некоторого объекта к определенному классу случайным образом и только в соответствии со значениями $p_1, \ldots, p_M$. Дело в том, что классификация «наугад» дает $P_c = \sum_{j=1}^{M} p_j^2$; эта величина не меньше и может приближаться к 1 при возрастании разброса значений $p_1, \ldots, p_M$. Таким образом, наблюденное кажущееся большим значение $P_c$ может и не говорить о достоинствах принятого классификатора. Простым способом контроля значимости результатов, получаемых в ходе селекции признаков, является графический анализ зависимости односторонних доверительных интервалов от размерности признакового пространства с указанием уровня, отвечающего классификации «наугад»: выход нижней границы доверительного интервала за этот уровень говорит о том, что соответствующий вариант набора признаков вообще не информативен.

Более общий подход в анализе совокупности результатов отбора признаков заключается в построении и использовании тестов однородности биномиальных пропорций. Нулевая гипотеза об однородности проверяется против альтернативы весьма общего вида:

$$H_0 : \pi_1 = \cdots = \pi_k = \pi_0$$

против

$$H_1 : \pi_i \neq \pi_j$$

для некоторых $i \neq j$, где $\pi_0$ обычно неизвестно.

Достаточно много работ посвящено разработке подобных тестов, особенно для $k = 2$, что соответствует случаю $2 \times 2$ таблицы сопряженности (например, в [8] рассматриваются 22 различных теста). При анализе результатов селекции признаков речь идет о сравнении двух определенных наборов признаков с возможностью ответить на вопрос, какой из них более информативен. Интересно, что в этом случае существует исчерпывающее решение соответствующей проблемы — равномерно наиболее мощный несмещенный (РНМН) критерий, основанный на статистике числа успехов в одной популяции при условии конкретного суммарного числа успехов в обеих популяциях [9, разд. 4.5]. Но сказать, что этот критерий широко известен и активно используется, нельзя. Причина проста: он основывается на гипергеометрическом распределении и трудоемок в применении.

Автор данной статьи занимался вопросами точных вычислений для гипергеометрического распределения и его аппроксимации с помощью биномиального, пуассоновского, нормального и бета-распределений [10], что позволило обеспечить корректное использование РНМН-критерия для сравнения двух биномиальных совокупностей.

В более общем случае $k > 2$ имеется также достаточно много критериев, построенных на различных принципах и способах практической реализации (см., например, [11, 12]); в первую очередь речь идет о следующих критериях: точные; стандартные Пирсона (Pearson) и Вилкса (Wilks); Потхоффа–Витингхилла (Potthoff–Whittinghill); Сю (Xu); Пауля и Денга (Paul and Deng), а также модификации отдельных из них. Сравнительный анализ основных процедур из приведенного перечня был проведен в [12] методом моделирования с учетом ситуаций, когда присутствуют так называемые разреженные данные (sparse data): некоторые $n_i$ малы или некоторые из $\pi_i$ близки к 0 или 1. Этот анализ позволил сформулировать следующие выводы:

– обнаружены случаи разреженных данных, для которых стандартные тесты, тест Потхоффа–Витингхилла и его модификация, тест Пауля и Денга выполняются неадекватно;

– тесты точные и Сю обладают адекватными характеристиками в любых рассмотренных условиях моделирования, при этом точные методы в целом схожи между собой, но некоторые различия возникли между точными тестами и тестом Сю.

Данные выводы основаны на результатах моделирования, поэтому всегда остается место для сомнений. Его можно рассеять только при эмпирическом сравнительном анализе критериев в рамках конкретных ограничений.

# 4 Эксперименты

Совместно со специалистами НИИ урологии и интервенционной радиологии им. Н. А. Лопаткина — филиал ФГБУ «НМИРЦ» Минздрава России С. А. Голованова и А. В. Сивкова проводились исследования возможности прогнозировать химический состав мочевых камней у пациентов с уролитиазом по метаболическим показателям мочи и сыворотки крови.

Задача прогнозирования состава камня по набору показателей была сформулирована как задача обучаемой классификации типов камней, распадающаяся на следующие отдельные элементы.

1. Предполагается, что исследователем задана классификация типов камней по составу. Она может включать, в принципе, произвольное число классов $M$ и должна быть четкой, т. е. любой камень по составу должен соответствовать только одному классу.

2. Для отдельных классов принимается вероятностное описание входящих в него наборов показателей. Например, если $u$ — вектор значений всех показателей, то считается известным распределение $f_i(u)$ для каждого $j$-го класса. Кроме этого, предполагаются заданными вероятности $p_j$ появления классов, $j = 1, \ldots, M$. В качестве классификатора рассматривался байесовский классификатор с единичной функцией потерь.
3. Для реализации описанной схемы вместо $p_j$ и $f_j(u)$ подставляются их оценки $p_j^*$ и $f_j^*(u)$. Для $p_j^*$ это не что иное, как преваленс (prevalence) — доля субъектов в популяции, которые имеют камень из $i$-го класса. В качестве $f_j^*(u)$ предлагается применять модель смеси нормальных распределений и EM-алгоритм для оценивания параметров этой модели.

Для решения задачи прогнозирования исходные данные о составе мочевых камней и показателях состояния пациентов представлялись как таблица «объект–признак», где объекты — пациенты, а признаки включали группу признаков, характеризующих химический состав камня пациента, а также группу метаболических и антропологических признаков пациента.

В группу признаков, характеризующих химический состав камня пациента, входили такие минеральные компоненты мочевых камней, как вевеллит (WH), веделлит (WD), мочевая кислота безводная (UA), мочевая кислота дигидрат (UADH), аммония урат (AMUR), даллит, или карбонатапатит (Dh), брушит (BRU), струвит (STRU), цистин (CYS). Здесь и далее в подобных ситуациях в скобках приводятся общепринятые обозначения показателей.

В данной работе рассматривалась классификация типов камней по составу, описанная в табл. 1. Задание порога в 50% отражает привычное представление о классификации на основе доминирующего значения того или иного компонента.

В группе метаболических признаков пациента были представлены биохимические показатели сыворотки крови — общий кальций (Ca_ser), мочевая кислота (UA_ser), фосфаты (P_ser); биохимические показатели суточной экскреции с мочой кальция (Ca_ur), мочевой кислоты (UA_ur), фосфатов (P_ur).

Учитывая важное влияние концентраций этих веществ в конечной моче на степень перенасыщенности мочи и, следовательно, на ее литогенный потенциал, рассчитывали также концентрационные показатели (в мМол/л), такие как концентрация в моче кальция (CaUrC), мочевой кислоты (UaUrC), фосфатов (PhUrC). Помимо этого учитывали такие физико-химические показатели мочи, как удельный вес (Spec_Grav), pH мочи (pH), суточный объем мочи (Diuresis) и антропологические показатели пациента — рост (Height) и вес (Weight). Таким образом, в задаче обучаемой классификации использовались 14 показателей: Ca_ser, UA_ser, P_ser, Ca_ur, UA_ur, P_ur, CaUrC, UaUrC, PhUrC, Spec_Grav, pH, Diuresis, Height, Weight. Полученный классификатор использовался для прогноза типа камня по представляемым данным о показателях. Для того чтобы охарактеризовать качество классификатора, использовалась вероятность правильной классификации. Чем выше значение вероятности правильной классификации, тем выше точность прогноза. Анализ полученных результатов позволил сделать вывод, что прогноз возможен, но надо быть готовым к не очень высоким результатам при увеличении числа классов (в частности, от $M = 2$ к 4). Источником повышения эффективности прогноза может стать уточнение набора показателей, на основе которых строится классификация.

Оценивание вероятности правильной классификации можно проводить одним из следующих способов:

– повторно используя обучающую выборку как для построения классификатора, так и для получения $P_c^*$ (T-оценка, от слова thrifty);
– привлекая метод перепроверки (CV-оценка, от термина cross-validation);
– используя бутстреп-метод (B-оценка, от термина bootstrap).

В данной работе применялся классический вариант метода перепроверки: исключим из исходной

**Таблица 1** Классификация типов камней по составу

| Названия типа классификации и камней | Число классов | Правила классификации: номер класса, условие |
|---|---|---|
| Смешанная общего вида: оксалатные, уратные, фосфатные, пр. | 4 | 1, если WH + WD $>$ 50%<br>2, если UA + UADH + Amur $>$ 50%<br>3, если Dh + BRU + STRU $>$ 50%<br>4 иначе |

совокупности данных $i$-й объект; для оставшегося набора данных построим байесовский классификатор и применим его к исключенному объекту; далее, вспомнив номер класса, к которому в действительности принадлежит $i$-й объект, сравним его с тем, который дала байесовская процедура, и получим ответ, совершила она ошибку или нет; подобные действия повторим для каждого значения $i$. В результате будет получена CV-оценка для $P_c$. Обычно считается, что она является более качественной, чем Т-оценка, обладающая «завышенным оптимизмом» (подгонка модели байесовского классификатора и оценка его качества происходит по одной и той же обучающей выборке).

Описание B-оценки начнем с модели байесовского классификатора. Предположим, что распределение данных, относящихся к $j$-му классу, есть смесь нормальных распределений

$$f_j(u) = \sum_{i=1}^{k} q_{ji} N\left(\mu_{ji}, \Sigma_{ji}\right) .$$

Тогда плотность распределения данных, подвергающихся байесовской классификации, представима в виде

$$g(u) = \sum_{j=1}^{M} p_j \sum_{i=1}^{k} q_{ji} N\left(u, \mu_{ji}, \Sigma_{ji}\right) .$$

Но реально $g(u)$ не задана, а имеется лишь обучающая выборка, с помощью которой можно получить оценку

$$g^*(u) = \sum_{j=1}^{M} p_j^* \sum_{i=1}^{k} q_{ji}^* N\left(u, \mu_{ji}^*, \Sigma_{ji}^*\right) .$$

Таким образом, становятся заданными все элементы теперь уже эмпирического байесовского классификатора, с помощью которого любое $x$ можно отнести к некоторому классу. Качество классификатора характеризуется с помощью вероятности правильной классификации $P_c$, которую посчитать аналитически затруднительно, а потому она будет оцениваться с помощью моделирования обучающей выборки из $g^*(u)$ — так называемой бутстреп-выборки $x^B$. В результате получаем алгоритм бутстреп-оценивания $P_c$, включающий следующие шаги: оценивание по обучающей выборке параметров смеси $g^*(u)$; формирование бутстреп-выборки $x^B$ из $g^*(u)$; классификация с помощью $g^*(u)$ данных из $x^B$ и подсчет числа случаев правильной классификации. Такая процедура позволяет для любого набора признаков дать оценку вероятности правильной классификации. Теперь для различных наборов признаков на основе биномиального распределения можно строить необходимые доверительные интервалы, проверять гипотезу об однородности результатов анализа классификаторов для ряда наборов признаков, сравнивать отдельные варианты наборов. Результаты отбора признаков с помощью метода SBS и B-оценки для $P_c$ представлены на рис. 1.

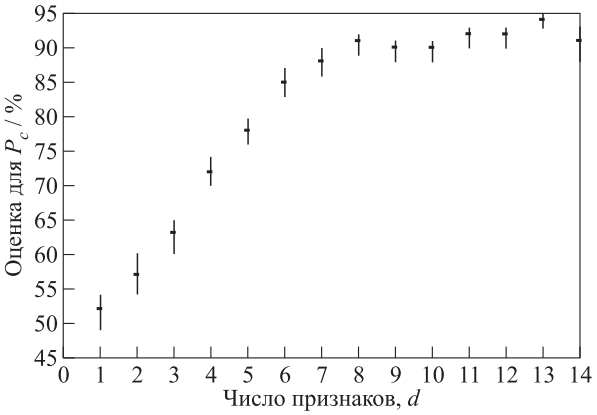


**Рис. 1** Зависимость эффективности селекции от числа признаков $d$ в случае SBS и B-оценки величины $P_c$ (для отдельного $d$ приведено значение оценки — горизонтальный штрих и 90%-ный доверительный интевал — вертикальный отрезок)

Результаты исследования наиболее интересных наборов признаков отражены в табл. 2. Для критерия Сю действенность предложенной им аппрок-

**Таблица 2** Выбор эффективной размерности признаков

| № | Набор значений $d$ | Критерий значимости | Критический уровень значимости | Принятие нулевой гипотезы об однородности при 5%-ном уровне значимости |
|---|---|---|---|---|
| 1 | 8, 9, 10, 11, 12, 13, 14 | Сю | 4,8% | Отвергается |
| 2 | 8, 9, 10, 11, 12, 14 | Сю | 61,1% | Принимается |
| 3 | 7, 8, 9, 10, 11, 12, 13, 14 | Сю | 0,0% | Отвергается |
| 4 | 13, 14 | РНМН | 1,2% | Отвергается |
| 5 | 8, 13 | РНМН | 1,0% | Отвергается |

симации проверялась с помощью моделирования и полностью подтвердилась. Анализ табл. 2 позволяет сделать следующие практически важные выводы:

- значения размерности пространства признаков от 8 до 14 с точки зрения эффективности классификации практически эквивалентны (строки 1 и 2 табл. 2);
- снижение размерности до 7 дает значимые потери в эффективности классификации (строка 3 табл. 2);
- значение размерности 13 заслуживает пристального внимания, так как может привести к повышению эффективности классификации по сравнению с обычным использованием всех показателей (строка 4 табл. 2) или по сравнению с применением наиболее «экономичного» варианта $d = 8$ (строка 5 табл. 2).

Практически те же результаты дал и метод SFS.

Данная схема бутстреп-анализа позволяет ответить на вопрос, на что можно рассчитывать с точки зрения качества классификации при использовании байесовского классификатора на основе смеси нормальных распределений и данных, распределенных действительно как смесь нормальных распределений.

Метод перепроверки из-за зависимости отдельных экспериментов по оценке качества классификатора не позволяет полностью формально сопоставлять их результаты, но по сравнению с результатами бутстреп-анализа он дает возможность получить представление об адекватности принятой модели данных (рис. 2).

Из рис. 2 видно, что метод перепроверки дает существенно более осторожные результаты, хотя из

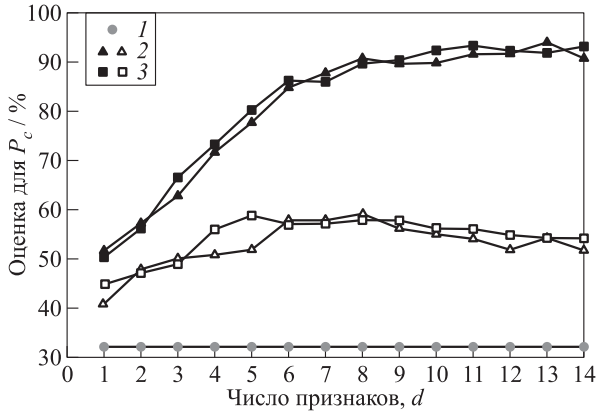


**Рис. 2** Зависимость эффективности селекции от числа признаков $d$ при классификации «наугад» (*1*) и с использованием байесовских классификаторов (*2* — SBS; *3* — SFS; черные значки — B-оценка для $P_c$; пустые значки — CV-оценка для $P_c$)

них опять же следует целесообразность применения для прогнозирования состава камня не всей, а лишь части совокупности показателей.

Снижение эффективности по методу перепроверки по сравнению с бутстреп-методом можно частично объяснять примененным вариантом CV-оценки и ограниченным объемом исходных данных. Для иллюстрации этого анализировалась зависимость эффективности классификации от объема обучающей выборки. Оказалось, что в рассматриваемом случае CV-оценка является нижней границей для истинного значения, а T- и B-оценки, близкие между собой, служат верхними границами. Их значения с ростом размера исходной выборки сходятся к истинной величине вероятности правильной классификации, но достигают этого значения при значительных размерах исходной выборки (более 2000 наблюденных значений).

## 5 Заключение

Существуют различные подходы к решению задачи отбора признаков, еще более богатым оказывается набор соответствующих методов. В прикладных областях к этому многообразию добавляется фактическая неоднозначность получающегося решения. Оно в принципе может оказаться формально единственным, но фактически значимо неотличимым от множества других. По этой причине возрастает роль критериев значимости полученных решений. Для нужд задачи классификации данных предлагается использовать такую интегральную характеристику эффективности, как вероятность правильной классификации. Для ее оценивания необходимо прибегать к анализу последовательности испытаний Бернулли, при этом полностью корректные результаты будут получены при использовании бутстреп-метода обработки исходной выборки.

Проведенные эксперименты в задаче прогнозирования химического состава мочевых камней создают предпосылки для повышения качества получаемых решений (сокращение перечня проводимых анализов может привести к росту точности прогноза), дают толчок специалистам в предметной области для прояснения сути протекающих процессов.

## Литература

1. *Webb A. R., Copsey K. D.* Statistical pattern recognition. — 3rd ed. — Chichester, U.K.: John Wiley & Sons, 2011. 616 p.

2. *Liu H., Yu L.* Toward integrating feature selection algorithms for classification and clustering // IEEE Trans. Knowl. Data Eng., 2005. Vol. 17. P. 491–502.
3. *Saeys Y., Inza I., Larrannaga P.* A review of feature selection techniques in bioinformatics // Bioinformatics, 2007. Vol. 23. P. 2507–2517.
4. *Yu L., Liu H.* Efficient feature selection via analysis of relevance and redundancy // J. Machine Learning Res., 2004. Vol. 5. P. 1205–1224.
5. *Stracuzzi D. J.* Randomized feature selection // Computational methods of feature selection. — Boca Raton, FL, USA: Chapman and Hall/CRC, 2007. P. 41–62.
6. *Dasgupta A., Zhang T.* Binomial and multinomial parameters, inference on // Encyclopedia of statistical sciences. — New York, NY, USA: John Wiley & Sons, 2006. P. 501–519.
7. *Hall P.* Improving the normal approximation when constructing one-sided confidence intervals for binomial or Poisson parameters // Biometrika, 1982. Vol. 69. P. 647–652.
8. *Upton G. J. G.* A comparison of alternative tests for the $2 \times 2$ comparative trial // J. Roy. Stat. Soc. A, 1982. Vol. 145. P. 86–105.
9. *Lehmann E. L., Romano J. P.* Testing statistical hypotheses. — 3rd ed. — New York, NY, USA: Springer, 2005. 784 p.
10. *Кривенко М. П.* Задачи выборочного контроля при досмотре лиц, багажа и транспорта // Обозрение прикладной и промышленной математики, 2011. Vol. 18. P. 125–126.
11. *Potthoff R. F.* Homogeneity, Potthoff–Whittighill tests of // Encyclopedia of statistical sciences. — New York, NY, USA: John Wiley & Sons, 2006. P. 3217–3220.
12. *Klein M., Linton P.* On a comparison of tests of homogeneity of binomial proportions. — Washington: Center for Statistical Research & Methodology Research and Methodology Directorate U.S. Census Bureau, 2013. https://www.census.gov/srd/papers/pdf/rrs2013-03.pdf.

*Поступила в редакцию 14.06.16*

# SIGNIFICANCE TESTS OF FEATURE SELECTION FOR CLASSIFICATION

M. P. Krivenko

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

**Abstract:** The paper considers the problem of feature selection for classification and issues related to the assessment of the quality of the solutions. Among the different methods of feature selection, attention is paid to sequential procedures; the probability of the correct classification is used to measure the quality of the classification. To evaluate this indicator, it is proposed to use cross-validation and the bootstrap method. At the same time, to investigate the set of sample values of probability of the correct classification, it is suggested to use comparative analysis of confidence intervals and the test for homogeneity of binomial proportions. While constructing Bayesian classifier as the data model mixture of normal distributions is adopted, the model parameters are estimated by the expectation–maximization algorithm. As an experiment, the paper considers the problem of well-thoughtout choice of classification characteristics when predicting the type of urinary stones in urology. It is demonstrated that the set of used features can be reduced not only without losing the quality of decisions, but also with increase of probability of correct prediction of the stone type.

**Keywords:** feature selection; sequential forward and backward selections; Bayes classification; test of homogeneity of binomial proportions; prediction of stone types in urology

**DOI:** 10.14357/19922264160305

## References

1. Webb, A. R., and K. D. Copsey. 2011. *Statistical pattern recognition*. 3rd ed. Chichester, U.K.: John Wiley & Sons. 616 p.
2. *Liu, H., and L. Yu.* 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.* 17:491–502.
3. Saeys, Y., I. Inza, and P. Larrannaga. 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics* 23:2507–2517.
4. Yu, L., and H. Liu. 2004. Efficient feature selection via analysis of relevance and redundancy. *J. Machine Learning Res.* 5:1205–1224.
5. Stracuzzi, D. J. 2007. Randomized feature selection. *Computational methods of feature selection*. Boca Raton, FL: Chapman and Hall/CRC. 41–62.
6. Dasgupta, A., and T. Zhang. 2006. Binomial and multinomial parameters, inference on. *Encyclopedia of statistical sciences*. New York, NY: John Wiley & Sons. 501–519.

7. Hall, P. 1982. Improving the normal approximation when constructing one-sided confidence intervals for binomial or Poisson parameters. *Biometrika* 69:647–652.
8. Upton, G. J. G. 1982. A comparison of alternative tests for the $2 \times 2$ comparative trial. *J. Roy. Stat. Soc. A* 145:86–105.
9. Lehmann, E. L., and J. P. Romano. 2005. *Testing statistical hypotheses*. 3rd ed. New York, NY: Springer. 784 p.
10. Krivenko, M. P. 2011. Zadachi vyborochnogo kontrolya pri dosmotre lits, bagazha i transporta [Tasks of sampling during the inspection of individuals, baggage and transport]. *Obozrenie prikladnoy i promyshlennoy matematiki* [Review of applied and industrial mathematics] 18:125–126.
11. Potthoff, R. F. 2006. Homogeneity, Potthoff–Whittighill tests of. *Encyclopedia of statistical sciences*. New York, NY: John Wiley & Sons. 3217–3220.
12. Klein, M., and P. Linton. 2013. On a comparison of tests of homogeneity of binomial proportions. Center for Statistical Research & Methodology Research and Methodology Directorate U.S. Census Bureau Washington. Available at: https://www.census.gov/srd/papers/pdf/rrs2013-03.pdf (accessed April 25, 2016).

*Received June 14, 2016*

## Contributors

**Krivenko Mikhail P.** (b. 1946) — Doctor of Science in technology, professor, leading scientist, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; mkrivenko@ipiran.ru