ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

2013 Управление, вычислительная техника и информатика № 3(24)

УДК 519.246

В.М. Неделько

НЕКОТОРЫЕ ВОПРОСЫ ОЦЕНИВАНИЯ КАЧЕСТВА МЕТОДОВ ПОСТРОЕНИЯ РЕШАЮЩИХ ФУНКЦИЙ

Работа посвящена проблеме оценивания качества методов построения решающих функций в задачах классификации. Исследуется возможность использования подхода, основанного на задании эталонного набора тестовых задач, а также связь данной проблемы с проблемой оценивания риска в задаче классификации, в частности нахождения распределений, при которых погрешность оценки риска максимальна. Предлагается метод использования полученных ранее результатов о максимальном смещении эмпирического риска для гистограммного классификатора для получения эмпирических оценок риска.

Ключевые слова: *распознавание образов, машинное обучение, решающая функция, вероятность ошибочной классификации, эмпирический риск.*

Решение задач построения решающих функций (интеллектуальный анализ данных, машинное обучение, статистические решения, распознавание образов, классификация с учителем) остаётся в настоящее время слабоформализованной областью, в которой качество получаемых решений существенно зависит от опыта и интуиции исследователя. Такое положение вызвано тем, что существует большое число различных методов решения, подходящих для одних и тех же задач, и в то же время практически отсутствуют формализованные рекомендации по выбору метода для заданной задачи.

Рассмотрим следующую общую схему решения прикладной задачи анализа данных:

1. Определение класса задач, к которому принадлежит задача, требующая решения.
2. Установление методов, подходящих для решения данного класса задач. Методы следует упорядочить по перспективности.
3. Применение метода, выглядящего наиболее перспективным.
4. Оценивание качества полученного решения.
5. Если полученное качество неудовлетворительно, выбираем следующий метод и возвращаемся к 3.

Как можно заметить, ни один из приведённых шагов до настоящего времени не систематизирован в достаточной степени. Более того, существуют противоположные точки зрения на то, как их следует проводить.

Проанализируем каждый из этапов.

Задачи анализа данных можно классифицировать по разным признакам.

Наиболее очевидная классификация по «техническим» характеристикам: число переменных, типы переменных, объём данных и т.п. Понятно, что подобные особенности задачи в некоторой мере определяю выбор метода, но получаемые классы задач и соответственно классы подходящих методов остаются слишком широкими.

Более плодотворной выглядит классификация по наличию и виду априорной информации о модели данных. Если есть основания считать данные выборкой из распределения заданного параметрического семейства, то естественно выбирать метод, специализированный для данного случая. Вместе с тем, в большинстве прикладных задач нет оснований предполагать определённый класс распределений и параметрические методы неприменимы.

В настоящей работе будем основываться на подходе, предложенном в системе «Полигон» в 80-х годах [1]. Идея подхода заключается в следующем принципе формирования набора тестовых задач: для каждого алгоритма подбираются задачи, которые этим алгоритмом эффективно решаются (на которые этот метод ориентирован), и все методы тестируются на объединённом наборе таких задач.

Подход основан на естественном постулате, что существование метода оправдано тогда и только тогда, когда существуют задачи, на которых данный метод работает лучше других, либо если метод является универсальным, т.е. на широком классе задач работает достаточно хорошо, хотя, возможно, хуже специализированных методов. Чтобы обосновать метод при таком подходе как раз нужно подобрать задачи, на которые он ориентирован, и исследовать, не окажется ли на этих задачах лучшим некоторый другой метод.

Для следующего шага требуется упорядочить методы по ожидаемому качеству их работы на заданном классе задач. Данный пункт требует объёмных исследований: оценить эффективность каждого метода на каждом классе задач. Кроме того, выбор критерия качества метода в данном случае не очевиден, поскольку на разных задачах одного класса эффективность одного и того же метода разная.

В качестве показательного примера можно привести задачу распознавания двух образов в предположении нормальности распределений характеристик. Эта классическая задача рассматривается в числе первых в любом учебнике по дискриминантному анализу. Однако даже для этого класса задач неизвестен лучший метод решения (понятие лучшего метода тоже строго не введено, но здесь достаточно неформального представления). Действительно, если объём выборки относительно велик, мы можем с достаточной точностью оценить все параметры распределений и построить квадратичную разделяющую поверхность, которая будет вполне приемлемым решением. Однако известно, что если объём выборки мал, то лучше строить линейную разделяющую функцию, даже если нет оснований предполагать матрицы ковариаций в действительности равными. При этом нет убедительной аргументации, какой метод лучше использовать: дискриминант Фишера, метод опорных векторов или другие. Кроме того, при переходе от квадратичной к линейной функции происходит резкий скачок сложности решения, хотя интуитивно понятно, что «хороший» метод должен позволять плавно регулировать сложность решения в зависимости от объёма выборки.

Применение выбранного метода представляется наиболее понятным шагом, хотя и здесь могут быть затруднения, связанные с наличием у алгоритмов настраиваемых параметров.

Оценивание качества полученного решения – проблема, активно исследуемая в течение более полувека [2–6]. При всём разнообразии и богатстве полученных результатов для практического применения чаще всего выбирается метод скользящего контроля. Однако ряд исследований свидетельствует, что качество решения по обучающей выборке можно оценивать точнее, чем методом скользящего контроля.

Наконец, последний шаг – критерий останова – также неочевиден. До каких пор стоит пробовать новые методы в расчёте получить решение лучшего качества и когда следует остановиться – однозначных ответов не существует.

В данной работе обсуждается ряд идей и результатов, касающихся рассмотренных проблем и их взаимосвязи.

1. Постановка задачи

Для введения основных понятий рассмотрим общую постановку задачи построения решающих функций.

Пусть X – пространство значений переменных, используемых для прогноза, а Y – пространство значений прогнозируемых переменных, и пусть C – множество всех вероятностных мер на заданной σ-алгебре подмножеств множества $D = X \times Y$. При каждом $c \in C$ имеем вероятностное пространство $\langle D, B, \mathsf{P}_c \rangle$, где B – σ-алгебра, P_c – вероятностная мера. Параметр c будем называть стратегией природы.

Решающей функцией назовем соответствие $\lambda : X \to Y$.

Качество принятого решения оценивается заданной функцией потерь: $L : Y^2 \to [0, \infty)$.

Под риском будем понимать средние потери

$$R(c, \lambda) = \int_D L(y, \lambda(x)) \, \mathsf{P}_c(dx).$$

В данной работе будем рассматривать задачу классификации, когда $Y = \{1, 2\}$, и функцию потерь в виде индикатора ошибочной классификации. В этом случае риск есть вероятность ошибочной классификации.

Пусть $V_N = \left\{ \left(x^i, y^i \right) \in D \;\middle|\; i = \overline{1, N} \right\}$ – случайная независимая выборка из распределения P_c, $V \in D^N$. В большинстве случаев объём выборки N будет фиксированным, поэтому этот параметр в обозначении выборки обычно будем опускать.

Эмпирический риск определим как средние потери на выборке:

$$\tilde{R}(v, \lambda) = \frac{1}{N} \sum_{i=1}^{N} L\left(y^i, \lambda\left(x^i \right) \right).$$

Пусть $Q : D^N \to \Lambda$ – алгоритм построения решающих функций, а $\lambda_{Q,V} \in \Lambda$ – функция, построенная по выборке V алгоритмом Q, Λ – заданный класс решающих функций.

Для фиксированного метода Q определён средний риск $F(c) = \mathsf{E} R\left(c, \lambda_{Q,V} \right)$.

Метод $\tilde{Q}$, минимизирующий эмпирический риск, есть $\lambda_{\tilde{Q},V} = \arg\min_{\lambda \in \Lambda} \tilde{R}(V, \lambda)$.

2. Полигон тестовых задач

В 80-х годах была разработана, насколько известно автору, первая в СССР программная система для сравнения алгоритмов распознавания [1]. Данная система называлась «Полигон» и позволяла сравнивать качество решений, получаемых разными алгоритмами, на наборе тестовых задач.

Важной идеей, заложенной в системе, было то, что тестовый набор формировался из задач, на решение которых были ориентированы исследуемые алгоритмы. Таким образом, можно было оценивать, как алгоритм работает на «своих» и на «чужих» задачах.

В настоящее время существует и активно развивается новый проект [7] по разработке системы сравнения методов классификации на репозитории задач. Особенностью данного проекта можно назвать ориентацию на использование в качестве тестовых широкого набора реальных задач распознавания образов, собранных из открытых источников (в частности, известного репозитория UCI), а также задач, присланных непосредственно их постановщиками. Другая особенность – вычисление большого числа разнообразных (как известных ранее, так и оригинальных) характеристик для оценивания качества решений. В этом проекте под задачей понимается таблица данных с указанием целевой переменной (и, опционально, других параметров задачи).

Использование реальных задач в качестве тестовых является, безусловно, достоинством этого подхода, но в то же время обуславливает ряд ограничений, в частности затрудняет статистический анализ результатов и ограничивает в наборе исследуемых характеристик. Ограничения связаны с тем, что в статистической постановке таблица данных является лишь частным набором прецедентов, в то время как для детального анализа нужна полная (вероятностная) модель. Кроме того, используемая в этом подходе оценка скользящего экзамена имеет относительно большую дисперсию, поэтому любой метод может на некоторых задачах случайно получить хорошие оценки, которые не означают, что построенное решение действительно хорошее.

Таким образом, не уменьшая важности тестирования методов на реальных задачах, следует отметить актуальность разработки методов исследования алгоритмов на синтетических данных.

Первым нетривиальным моментом в этом направлении является выбор тестовой единицы (т.е. объекта, на котором тестируется метод).

Основным требованием к такому выбору выдвинем возможность введения понятия оптимального на данной тестовой единице метода классификации, так чтобы это понятие было содержательным.

Понятно, что для одной выборки даже понятие оптимального решающего правила не является содержательным.

На первый взгляд, напрашивается в качестве тестовой единицы использовать распределение. Однако при заданном распределении определено понятие оптимального (байесовского) решающего правила, но понятие оптимального метода обучения будет вырожденным: оптимальным будет, очевидно, метод, который независимо от обучающей выборки будет давать в качестве результата байесовское решающее правило.

Понятие оптимального метода обучения становится содержательным, только если рассматривать в качестве тестовой единицы целый класс распределений.

Заметим, что при проведении, например, конкурсов по анализу данных, наилучший метод определяется не по классу распределений и даже не по распределению, а по частной выборке. Но это возможно только потому, что есть скрытая информация (тестовая часть выборки), которую разработчики методов не знают. Однако в нашем случае метод обучения – это всего лишь некоторое отображение множества выборок во множество решений, и непонятно как требование «метод

не должен знать ответов для тестовой выборки» задать в виде формальных ограничений на класс отображений из D^N в Λ.

Итак, под тестовой единицей будем понимать класс распределений.

Естественным критерием качества метода на заданном распределении является средний риск. Чтобы определить качество метода на классе распределений, нужно решить проблему многокритериальности: на разных распределениях метод имеет разное качество. Поскольку вводить какую-либо меру на распределениях нет оснований, усреднять качество нет возможности, то напрашивается использовать минимаксный подход. Применить последний непосредственно также не получится, так как полученная величина будет отражать только качество методов на «худшем» распределении, что неинформативно.

Один из вариантов решения данной проблемы – это приписать каждому распределению некоторую величину $b(c)$ – назовём её базовым уровнем риска.

Тогда в качестве содержательной меры качества метода на классе распределений $\overline{C} \subseteq C$ можно использовать, например, величину $\sup\limits_{c \in \overline{C}} \left(F(c) - b(c) \right)$.

Таким образом, в роли тестовой единицы выступает пара $\left\langle \overline{C}, b(\cdot) \right\rangle$.

Класс распределений может конструироваться на основе класса решающих правил, которыми оперирует рассматриваемый метод. Например, для методов, основанных на решающих деревьях, будет естественным рассмотреть классы кусочно-постоянных распределений, с областями постоянства из разбиений, задаваемых случайными деревьями.

3. Оценивание риска

Одним из важнейших направлений в области машинного обучения являются исследования по проблеме переобучения, которая заключается в том, что при ограниченном объёме выборки относительно сложные методы построения решающих функций проигрывают по качеству более простым [3]. Для выбора адекватной сложности метода требуется оценивать качество решения (вероятность ошибочной классификации, или риск), не используя при этом контрольную выборку.

Существуют точечные и интервальные оценки риска. К первым относятся эмпирический риск, оценка скользящего экзамена, bootstrap и др. Качество точечных оценок естественно характеризовать средним квадратом отклонения от оцениваемой величины. Однако такая характеристика позволяет сравнивать оценки друг с другом, выбирая лучшую, но не даёт достаточной информации, к какой мере можно полагаться на полученное численное значение оценки риска в конкретной задаче. Последнее требует построения интервальных оценок. Наиболее известной интервальной оценкой риска является оценка Вапника – Червоненкиса [2].

В настоящей работе развивается подход, основанный на явном нахождении распределений, при которых погрешность оценок максимальна. Это, однако, не означает, что мы будем ориентироваться на «худший случай» с точки зрения ожидаемого качества классификации, поскольку большая погрешность оценки риска не означает большое значение риска, и распределения, при которых погрешность оценки максимальна, ни в коей мере не являются «плохими» Скорее наоборот, на действительно «плохих» распределениях можно очень точно оценить вероятность ошибочной классификации.

Заметим, что оценки Вапника – Червоненкиса являются чрезмерно пессимистичными именно из-за того, что они «ориентированы на худший случай». Но здесь играет роль не то, что предполагается «худшее» распределение [8], а, в частности, то, что допускаются любые методы классификации, в том числе методы, у которых классификаторы максимально различны (не учитывается эффект «сходства классификаторов» [9]).

Как будет показано далее, распределение, доставляющее максимальное смещение эмпирического риска для гистограммного классификатора (а смещение этой оценки даёт основной вклад в её погрешность) является вполне типичным.

Смещением эмпирического риска является величина $S(c) = F(c) - \tilde{F}(c)$, где $F(c) = \mathsf{E}\, R\left(c, \lambda_{Q,V}\right)$, $\tilde{F}(c) = \mathsf{E}\, \tilde{R}\left(c, \lambda_{Q,V}\right)$.

Задача нахождения максимального смещения эмпирического риска состоит в вычислении

$$\hat{S}\left(\tilde{F}_0\right) = \max_{c:\tilde{F}(c)=\tilde{F}_0} S(c). \tag{1}$$

Здесь нас интересует не безусловно максимальное значение смещения, а максимальное смещение при условии заданного значения ожидаемого эмпирического риска. Такая постановка объясняется тем, что на практике нас интересует значение риска при известном (полученном на исследуемой таблице данных) значении эмпирического риска.

Очевидно, что

$$\hat{S}\left(\tilde{F}_0\right) = \hat{F}\left(\tilde{F}_0\right) - \tilde{F}_0, \text{ где } \hat{F}\left(\tilde{F}_0\right) = \max_{c:\tilde{F}(c)=\tilde{F}_0} F(c),$$

и

$$\hat{F}^{-1}(F_0) = \breve{F}(F_0), \text{ где } \breve{F}(F_0) = \min_{c:F(c)=F_0} \tilde{F}(c).$$

Последнее соотношение полезно тем, что для функции $\breve{F}(F_0)$ для рассмотренной ниже задачи получено простое приближённое выражение.

Рассмотрим метод, называемый «гистограммный классификатор», который удобен для исследования тем, что позволяет делать аналитические выкладки [5], а также простотой метода классификации и отсутствием в методе каких-либо характеристик, учёт которых мог бы уточнить оценки [9, 10].

Пусть X дискретно, то есть $X = \{1, \ldots, k\}$. Тогда вероятностная мера $c \in C$ задается набором вероятностей

$$c = \left\{ \varsigma_j^{\omega} = \mathsf{P}(x = j, y = \omega) \;\middle|\; j = \overline{1,k},\ \omega = \overline{1,2} \right\}.$$

Обозначим

$$\alpha_j = \mathsf{P}(x = i) = \varsigma_j^1 + \varsigma_j^2,\ p_j = \mathsf{P}(y = 1/x = i),\ q_j = 1 - p_j,\ c_j = (\alpha_j, p_j).$$

Будем рассматривать алгоритм $\tilde{Q}$, который минимизирует эмпирический риск независимо в каждой точке x пространства X, т.е. приписывает образ с наибольшей выборочной частотой в этой точке и принимает равновероятно значения 1 и 2 при равенстве частот. Метод $\tilde{Q}$ называется гистограммным классификатором.

Пусть $N \geq k$. Оказывается [9], что распределение, на котором достигается значение смещения, не более чем на $1/k$ отличающееся от максимального, имеет следующий вид:

$$\alpha_j = \alpha', \quad p_j = p', \quad j = 1,...,k-1, \ \alpha_k = 1 - \tfrac{k-1}{N}, \quad p_k = 0.$$

Иными словами, «наихудшее» распределение является равномерным, за исключением одной «ячейки», куда помещается «излишек» вероятности.

Значения α' и p' определяются следующим образом. При $\tilde{F}_0 \le \tilde{F}_T$ имеем $\alpha' = \frac{1}{N}$, а p' вычисляется на основе $\tilde{F}_0$. При $\tilde{F}_0 \ge \tilde{F}_T$ имеем $p' = 0{,}5$, а α' вычисляется на основе $\tilde{F}_0$. Здесь $\tilde{F}_T$ – математическое ожидание эмпирического риска при $\alpha' = \frac{1}{N}$, $p' = 0{,}5$.

На рис. 1 левая диаграмма показывает максимальное смещение эмпирического риска в зависимости от относительного объёма выборки $M = \frac{N}{k}$.

Представляет интерес сравнение полученных для гистограммного классификатора точных оценок смещения эмпирического риска со сложностными оценками Вапника–Червоненкиса. Такое сравнение показывает, что для гистограммного классификатора оценки Вапника–Червоненкиса завышают смещение эмпирического риска в 2 – 3 раза [8].

Оказывается, что для функции максимального смещения выполняется приближённое соотношение

$$\breve{F}(F_0) \approx F_0\left(1 - \frac{\theta}{\sqrt{1+4F_0M}}\right), \qquad F_0M \ge \frac{1}{2},$$

где $\theta = (1-2\Theta)\cdot\sqrt{3}$, а $\Theta \approx 0{,}163$.

Хотя последнее выражение получено для гистограммного классификатора, практика показывает, что оно может использоваться как оценка смещения эмпирического риска и в других задачах. Для этого нужно в качестве M использовать

$$M^* = \frac{1}{2}\left(\left(\frac{\theta}{1-2\tilde{F}^*}\right)^2 - 1\right),$$

где $\tilde{F}^*$ – среднее значение эмпирического риска, полученное при статистическом моделировании на распределениях с пересекающимися классами (при «нулевой» гипотезе).

4. Численные эксперименты

В случае непрерывного пространства переменных можно задать распределения, «похожие» на распределение, доставляющее максимальное смещение эмпирического риска для гистограммного классификатора.

Проанализируем распределения, доставляющие максимальное смещение эмпирического риска. Эти распределения можно получить, максимизируя ожидаемый риск при фиксированном $\tilde{F}_0$ либо минимизируя ожидание эмпирического риска при фиксированном среднем риске F_0.

Рассмотрим, как меняется экстремальное распределение при изменении F_0.

При $F_0 = 0{,}5$ минимум $\tilde{F}$ достигается на равномерном распределении в D, т.е. когда все $\alpha_j = \frac{1}{k}$, $p_j = 0{,}5$.

При уменьшении F_0 все p_j остаются равными 0,5, за исключением одного, например последнего, которое становится равным 0 или, что эквивалентно, 1. При этом вероятность перераспределяется в эту «ячейку», т.е. α_j увеличивается соответственно уменьшению F_0.

При дальнейшем уменьшении F_0 перераспределение вероятности продолжается до тех пор, пока α' не уменьшится до $\frac{1}{N}$. После этого α' не меняется, а начинает меняться p'.

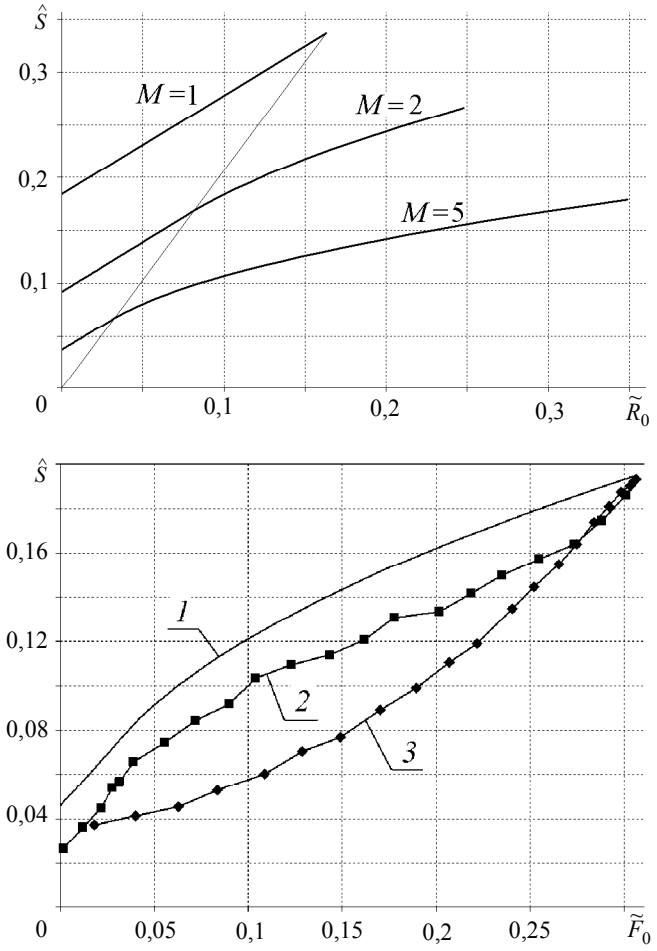

Рис. 1. Максимальное смещение эмпирического риска.
Оценка смещения эмпирического риска методом статистического моделирования

Легко заметить, что полученные распределения характеризуются тем, что пространство X оказывается разбитым на две подобласти: в одной из которых байесовский уровень ошибки нулевой, а в другой – значительный. Подобную особенность распределения легко обеспечить в непрерывном случае.

Пусть $X = [0,1]^n$ – n-мерный гиперкуб, на котором задано равномерное распределение.

Чтобы полностью определить меру P, осталось задать $g(x) = \mathsf{P}(y = 1/x)$ – условную вероятность первого класса при попадании в точку x. Будем задавать $g(x)$ в виде

$$g(x) = \begin{cases} g_1, & x_j < \delta,\ j = 1,...,n, \\ g_2, & \text{иначе}, \end{cases} \qquad \delta = \vartheta^{\frac{1}{m}}.$$

Иными словами, $g(x)$ является кусочно-постоянной, первая область постоянства есть гиперкуб объёма ϑ, вторая – дополнение внутреннего гиперкуба до единичного.

Первое семейство распределений (назовём его моделью A) зададим следующим образом: параметр ϑ положим равным некоторой константе ϑ_0, $g_2 \equiv 1$, а параметр g_1 изменяется от 0 до 0,5; далее полагаем $g_1 \equiv 0,5$, а ϑ изменяем от ϑ_0 до 0. Заметим, что данное семейство задано в некотором смысле «по подобию» распределений, доставляющих максимум смещения эмпирического риска для гистограммного классификатора.

Для сравнения будем рассматривать также другое семейство распределений (назовём его моделью B), которое задаётся следующим образом: $\vartheta \equiv 0,5$, $g_1 = g'$, $g_1 = 1 - g'$, где g' задаёт байесовский уровень ошибки и изменяется от 0 до 0,5.

На правой диаграмме рис. 1 приведены результаты статистического моделирования для метода направленного построения деревьев решений. Обозначения: *1* – оценка $\hat{S}(\cdot)$ при $M = 4$, *2* – моделирование на стратегии A с параметром $\vartheta_0 = 0,83$, 3 – моделирование на стратегии B.

Значение параметра сложности M можно подобрать с помощью моделирования на равномерном распределении.

Заключение

Таким образом, рассмотрена проблема создания полигона тестовых задач для исследования качества методов построения решающих функций.

Предложено в качестве тестовых единиц в таком полигоне использовать специальным образом подобранные классы распределений. Классы распределений подбираются так, чтобы статистическое моделирование на них по возможности полно отражало особенности тестируемого метода обучения. Это, в частности, означает, что класс распределений должен являться параметрическим семейством, один из параметров которого есть наименьшее достигаемое значение риска в заданном классе решающих правил.

Другим важным параметром является величина смещения эмпирического риска.

Установлено, что для гистограммного классификатора «наихудшее» распределение (для которого смещение эмпирического риска максимально) является смесью равномерного (по X) распределения и распределения, сосредоточенного в одной точке.

Подобного вида распределения легко могут быть заданы и в непрерывном пространстве переменных. Результаты статистического моделирования для задачи

классификации с помощью деревьев решений позволяют предположить, что на таких распределениях будет достигаться максимальное смещение эмпирического риска и для других (помимо гистограммного) методов классификации.

ЛИТЕРАТУРА

1. *Лбов Г.С., Старцева Н.Г.* Сравнение алгоритмов распознавания с помощью программной системы «Полигон» // Анализ данных и знаний в экспертных системах. Новосибирск. Вычислительные системы. 1990. Вып.134. С. 56–66.
2. *Вапник В.Н., Червоненкис А.Я.* Теория распознавания образов. М.: Наука, 1974. 415 с.
3. *Лбов Г.С.* Выбор эффективной системы зависимых признаков // Вычислительные системы. Новосибирск., 1965. Вып. 19. С. 21–34.
4. *Лбов Г.С., Старцева Н.Г.* Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск: Институт математики СО РАН, 1999. 211 с.
5. *Braga-Neto U. and Dougherty E.R.* Exact performance of error estimators for discrete classifiers // Pattern Recognition, Elsevier Ltd. 2005. V. 38. No. 11. P. 1799–1814.
6. *Langford J.* Quantitatively tight sample complexity bounds. Carnegie Mellon Thesis. 2002. http://citeseer.ist.psu.edu/langford02quantitatively.html. 130 p.
7. *Воронцов К.В., Ивахненко А.А., Инякин А.С. и др.* «Полигон» – распределённая система для эмпирического анализа задач и алгоритмов классификации // Всерос. конференция «Математические методы распознавания образов-14». М.: МАКС Пресс, 2009. С. 503–506.
8. *Неделько В.М.* О точности интервальных оценок вероятности ошибочной классификации, основанных на эмпирическом риске // Всерос. конференция «Математические методы распознавания образов-14». М.: МАКС Пресс, 2009. С. 56–59.
9. *Неделько В.М.* Точные и эмпирические оценки вероятности ошибочной классификации. // Научный вестник НГТУ. Новосибирск: Изд-во НГТУ, 2011. № 1(42). С. 3–16.
10. *Nedelko V.M.* Estimating a quality of decision function by empirical risk // LNAI 2734. Machine Learning and Data Mining in Pattern Recognition. Third International Conference, MLDM 2003, Leipzig. Proceedings. Springer-Verlag. P. 182–187.
11. *Лемешко Б.Ю., Лемешко С.Б., Постовалов С.Н.* Сравнительный анализ мощности критериев согласия при близких альтернативах. II. Проверка сложных гипотез // Сибирский журнал индустриальной математики. 2008. Т.11. № 4(36). С. 78–93.
12. *Неделько В.М.* Оптимизация оценки вероятности ошибочной классификации в дискретном случае // Classification, Forecasting, Data Mining. Int. Book Series «Information Science and Computing», No. 8. Supplement to the Int. J. «Information Technologies and Knowledge», V. 3. ITA, FOI ITHEA, Sofia, 2009. P. 47–54.

Неделько Виктор Михайлович
Институт математики СО РАН (г. Новосибирск)
E-mail: nedelko@math.nsc.ru

Поступила в редакцию 15 мая 2012 г.

Nedelko Victor M. (Institute of Mathematics SB RAS. Novosibirsk). **Some aspects of estimating a quality of decision functions construction methods.**

Keywords: pattern recognition, machine learning, decision function, misclassification probability, empirical risk.

The paper is devoted to a problem of estimating a quality of decision functions construction methods in classification (pattern recognition) task. An approach based on constructing some special set of testing tasks is investigated. As the testing tasks the distributions delivering the maximal bias of empirical risk are, in particular, used. The statistical modeling performed allows to evaluate an applicability of the results obtained.