Автоматика и телемеханика, № 3, 2019

Стохастические системы

© 2019 г. В.В. ЗЕНКОВ, канд. техн. наук (zenkov-v@yandex.ru)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва)

ОЦЕНКА АПОСТЕРИОРНОЙ ВЕРОЯТНОСТИ КЛАССА ПО СЕРИИ ДИСКРИМИНАНТНЫХ ФУНКЦИЙ АНДЕРСОНА

Предложен способ оценки по обучающей выборке с учителем апостериорных вероятностей принадлежности объекта к классам по совокупности известных значений его признаков. Эти вероятности являются исчерпывающей информацией для решения задачи классификации. Они позволяют решать задачу классификации с использованием различных критериев (максимум вероятности, минимум средней стоимости ошибки и пр.), задаваемых, как правило, субъективно. Метод решения задачи основан на построении серии аппроксимаций специальных дискриминантных функций, в нулевых точках которых апостериорные вероятности классов заданы при их построении. Эвристический алгоритм аппроксимации используемой дискриминантной функции в окрестности нулевых значений является частью метода. В методе отсутствует необходимость в дополнительных надстройках, например типа калибратора Платта для метода опорных векторов. Приведены модельные примеры применения метода и пример из медицинской диагностики с реальными данными.

Ключевые слова: машинное обучение, классификация, дискриминантная функция Андерсона, аппроксимация дискриминантной функции, оценка апостериорной вероятности класса, калибратор Платта, ROC–анализ.

DOI: 10.1134/S000523101903005X

1. Введение

Задача классификации приобретает бо́льшую привлекательность, если в заданной точке пространства признаков объектов определять апостериорные вероятности (АпоВ) принадлежности точки (объекта классификации) к классам. АпоВ классов относятся к условным вероятностям по определению. Они отличаются от безусловных (априорных) вероятностей тем, что вероятности классов при классификации объекта находятся при условии, что известна совокупность значений признаков, характеризующих объект. Это обстоятельство позволяет называть эти условные вероятности классов в точке апостериорными вероятностями. Зная оценки АпоВ классов, специалисту легче принять и объяснить решение об отнесении объекта к одному из классов или продолжить обследование объекта дальше. В частности, подсчитывая на основе АпоВ классов в заданной точке средние стоимости ошибок от отнесения ее в тот или иной класс, выбирая класс с минимальными средними потерями, получаем оптимальное по Байесу решение задачи классификации при заданных стоимостях потерь от ошибок классификации [1].

Задание стоимостей ошибок классификации является определенной проблемой, носящей зачастую субъективный характер. Решение задачи классификации в два этапа отделяет объективную часть — определение АпоВ классов в точках — от субъективной части, где для собственно классификации используются заданные стоимости потерь от ошибок. Задачу можно решать также одновременно и для нескольких вариантов стоимостей потерь от ошибок для всестороннего анализа принимаемого решения.

В применении к задаче классификации объектов (пациент, состояние производственного процесса и т.п.) в стохастической постановке качественные состояния объекта называются классами, которые надо устанавливать или прогнозировать на основе вектора признаков[1] — результатов обследования объекта. Классы и признаки рассматриваются как зависимые случайные величины. Для каждого класса имеет место свое условное распределение вероятностей признаков класса. Каждому значению вектора признаков соответствуют свои АпоВ классов.

Если известны условные распределения вероятностей признаков классов и априорные вероятности классов, то по формуле Томаса Байеса, богослова и математика середины XVIII в., опубликовавшего лишь одну работу по математике, вычисляются до сих пор АпоВ классов в заданной точке пространства признаков. Если они не известны, но есть обучающая выборка с учителем, состоящая из строк наблюдений, включающих вектор-строки признаков и номера классов, которым они принадлежали при проведении эксперимента, то можно построить условные распределения вероятностей признаков классов и по формуле Байеса затем вычислять АпоВ классов [1].

Кроме этого традиционного способа, требующего немалого объема обучающей выборки и большого количества оцениваемых параметров, для решения задачи классификации с оценкой АпоВ классов применяются методы, использующие либо заданные формы зависимости АпоВ классов от признаков (например, логистическая регрессия [3, 4]), либо комбинированные способы, сочетающие в себе построение границ между классами в пространстве признаков с помощью дискриминантных функций (ДФ), например метод опорных векторов с добавлением, например калибратора Платта [5] для пересчета отступов от границ классов в АпоВ классов. Есть также и методы (например наивный байесовский, решающих деревьев), которых касаться не будем, оценивающие непосредственно сами АпоВ классов, для улучшения которых используется изотоническая регрессия [6].

В данной работе предложен иной, по мнению автора простой и естественный, метод оценивания АпоВ классов, основанный на ДФ специального вида, называемой далее ДФ Андерсона. Метод использует ее тождественную связь с АпоВ классов и со стоимостями ошибок классификации. Метод не требует применения искусственных приспособлений для пересчета отступов от границ в АпоВ классов, как это делается в калибраторе Платта.

Определение ДФ Андерсона непосредственно вытекает из решения многоклассовой байесовой задачи классификации, предложенного Теодором Ан-

[1] По В.Ю. Кнеллеру [2] термин “признак” есть качественная характеристика, но в машинном обучении он прочно обосновался как количественная величина.

дерсоном в середине прошлого века [1]. Оптимальное по Байесу решение задачи классификации получается в результате попарного сравнения средних потерь от отнесения точки в один из двух классов. ДФ, по знаку которой из двух классов будет отбрасываться не перспективный для дальнейших сравнений класс, получим, взяв разность средних потерь для сравниваемой пары классов. Это и есть ДФ Андерсона. Она, как будет показано далее, имеет свойства, удобные для получения методов оценивания АпоВ классов по обучающей выборке с учителем.

Один метод основан на построении по обучающей выборке с учителем путем подбора стоимостей ошибок классификации такой аппроксимации ДФ Андерсона (АДФ), которая проходит через заданную точку [7], т.е. принимающую нулевое значение в этой точке. По подобранным стоимостям ошибок классификации вычисляется АпоВ класса. Достоинством этого метода построения индивидуальной АДФ для каждой заданной точки является то, что для решения задачи используется вся обучающая выборка. Если из обучающей выборки удаляются точки или добавляются, то сам алгоритм не изменяется. Метод удобен как при малом объеме выборки, так и в нестационарном случае. Недостатком является то, что он сравнительно медленно работает.

Второй метод предлагается в данной работе. В нем по обучающей выборке строится серия АДФ в заданном количестве с заданными на границах величинами АпоВ классов. Для оценки в дальнейшем АпоВ классов в заданной точке по серии АДФ используется метод линейной интер- и экстраполяции, в котором расстояния до двух соседних границ, между которыми располагается точка, измеряется модулями значений этих АДФ в заданной точке. Оценки АпоВ классов в точке могут находиться и методом ядерного сглаживания (потенциальных функций), но уже по всей серии АДФ.

Точность решения задачи зависит от эффективности аппроксимации ДФ Андерсона в окрестности ее нулевых значений. В приведенных далее примерах использовался эвристический алгоритм построения АДФ по обучающей выборке в окрестности нулевых значений [8].

2. ДФ Андерсона, ее свойства и связь с АпоВ классов

ДФ Андерсона $f_{rs}(x)$, разделяющая классы r и s в d-мерном евклидовом пространстве признаков, $x \in R^d$, для АпоВ классов и заданных стоимостей ошибок классификации имеет вид

$$f_{rs}(x) \equiv G_r(x) - G_s(x) \equiv \mathrm{M}_{k|x}(C_{rk} - C_{sk}), \tag{1}$$

где $G_r(x) = \sum_k C_{rk}p(k|x)$, $G_s(x) = \sum_k C_{sk}p(k|x)$ – средние потери по k в точке x, если точку отнести к классу r или, соответственно, к классу s; C_{ij} – стоимость ошибки, когда точка из класса j ошибочно относится в класс i, $C_{ij} \geq 0$, $C_{ii} = 0$; $p(k|x)$ – АпоВ класса k в точке x, $p(k|x) = P_k p(x|k)/p(x)$, $p(x) = \sum_k P_k p(x|k)$, P_k – априорные вероятности классов, $p(x|k)$ – условные распределения признаков классов; k – номера классов от 1 до K, K – количество классов; $M_{k|x}(\cdot)$ – математическое ожидание по k в точке x. Таким образом, в точке x случайная по k дискретная величина $C_{rk|x} - C_{sk|x} = f_{rs}(x) + \varepsilon_{rsk|x}$ имеет распределение АпоВ $p(k|x)$,

$\sum_k p(k|x) = 1$, среднее $f_{rs}(x)$ и случайное отклонение от него $\varepsilon_{rsk|x}$. Дискретная случайная величина $\varepsilon_{rsk|x}$ имеет нулевое среднее и распределение $p(k|x)$.

Если $f_{rs}(x) \leq 0$, то точка x относится в класс r и класс s исключается из дальнейшего процесса сравнения, иначе — в класс s и класс r исключается. Так обеспечивается минимум средней стоимости ошибок классификации — критерий решающего правила.

В случае двух классов, $K = 2$, имеем $C_{1k|x} - C_{2k|x} = f_{12}(x) + \varepsilon_{12k|x}$. Дискретная случайная величина $\varepsilon_{12k|x}$ в точке x принимает два значения: $-C_{21} - f_{12}(x)$ с вероятностью $p(1|x)$ и $C_{12} - f_{12}(x)$ с вероятностью $1 - p(1|x)$, с нулевым средним и дисперсией $(C_{12} + C_{21})^2 p(1|x)(1 - p(1|x))$.

2.1. Свойство ДФ Андерсона

Утверждение. ДФ Андерсона есть ограниченная функция регрессии.

Доказательство следует из (1), если ограничены стоимости ошибок классификации.

Следствие 1. В случае $K = 2$ имеем $-C_{21} < f_{12}(x) < C_{12}$.

Следствие 2. Чтобы аппроксимировать ДФ Андерсона как функцию регрессии, следует преобразовать обучающую выборку с учителем задачи классификации в выборку задачи регрессионного анализа, заменив номера классов в выборке следующим образом: первый класс на $-C_{21}$, а второй класс на C_{12}.

Следствие 3. Факт регрессионной зависимости ДФ Андерсона от признаков позволяет, в частности, выполнять и отбор признаков, используемых для решения задачи аппроксимации, по коэффициентам корреляции признаков со столбцом, в котором номера классов заменены стоимостями ошибок классификации. Учитывать при этом необходимо и коэффициенты корреляции признаков между собой [9].

2.2. Связь ДФ Андерсона с АпоВ классов

Утверждение. При $K = 2$ для АпоВ первого класса и ДФ Андерсона, полученной для заданных C_{12} и C_{21}, имеет место тождество

$$p(1|x) \equiv (C_{12} - f_{12}(x))/(C_{12} + C_{21}). \tag{2}$$

Доказательство. Из определения (1) для случая $K = 2$ имеет место

$$f_{12}(x) \equiv C_{12}p(2|x) - C_{21}p(1|x).$$

Используя $p(2|x) \equiv 1 - p(1|x)$, получим (2).

Следствие 1. Из (2) следует, что в точках на границе классов, где $f_{12}(x) = 0$, АпоВ первого класса (обозначим ее через p^*) и стоимости ошибок, для которых задана ДФ Андерсона, связаны соотношениями

$$p^* = C_{12}/(C_{12} + C_{21}), \quad 1 - p^* = C_{21}/(C_{12} + C_{21}), \quad C_{12}/C_{21} = p^*/(1 - p^*), \tag{3}$$

а в точках, относимых в первый класс, $f_{12}(x) < 0$, из (2) следует, что $p(1|x) > p^*$.

С л е д с т в и е 2. На p^* влияет лишь C_{12}/C_{21} при $C_{21} > 0$. Если для ДФ Андерсона задавать стоимости ошибок при условии $C_{12} + C_{21} = 1$, что не приведет к потере общности, то тождество (2) не только предельно упростится, но и стоимости ошибок обретут удобный для интерпретации смысл, а именно, $C_{12} = p^*$, $C_{21} = 1 - p^*$.

В ДФ Андерсона в случае двух классов для подчеркивания ее зависимости от стоимостей ошибок классификации введем p^* в качестве параметра:

$$p(1|x) \equiv p^* - f_{12}(x, p^*). \tag{4}$$

2.3. Условия неразличимости классов

В пространстве признаков в области первого класса по определению $f_{12}(x, p^*) < 0$ и из (4) следует, что $p(1|x) > p^*$ в этой области. Если $\min_x f_{12}(x, p^*) > 0$, то классы становятся неразличимыми при заданной стоимости ошибок, так как везде $p(1|x) < p^*$. Область первого класса в пространстве признаков исчезает и все точки надо относить во второй класс. Условия неразличимости классов в задаче с двумя классами для заданного p^* имеют вид

$$\left(\min_x f_{12}\left(x, p^*\right) > 0\right) \vee \left(\max_x f_{12}\left(x, p^*\right) < 0\right), \tag{5}$$

где второе условие ликвидирует область второго класса в пространстве признаков и все точки нужно относить в первый класс.

3. Постановка и идея решения задачи

3.1. Цель и базовый принцип решения задачи

Основываясь на тождестве (4) и на следствии 2 тождества (2), надо построить серию аппроксимаций ДФ Андерсона с заданными АпоВ в точках на границах двух классов, чтобы по полученной серии аппроксимаций оценивать АпоВ классов в заданной точке пространства признаков классов. Способ оценивания АпоВ классов должен иметь возможность использования его и в случае нескольких классов, $K > 2$.

Этот метод принципиально отличен от способов решения задачи классификации, во-первых, основанных на построении условных распределений признаков классов $p(x|k)$ с последующим использованием формулы Байеса для оценки АпоВ классов в заданной точке пространства признаков $p(k|x)$, во-вторых, от логистической регрессии, использующей аппроксимацию АпоВ класса в виде сигмоидной функции, и в-третьих, от методов, когда после построения границ между классами, например методом опорных векторов, АпоВ классов находятся по величинам отступов от этих границ с помощью надстроек типа калибратора Платта.

3.2. Исходная информация

Исходной информацией для решения задачи является обучающая выборка с учителем в виде матрицы, состоящей из N линейно независимых строк. Строка n состоит из вектора признаков x_n и номера класса k_n, которому он принадлежал в эксперименте n.

3.3. Частные требования к решению

Аппроксимировать ДФ Андерсона в области нулевых значений следует методом [8], в котором аппроксимация задается в виде $\lambda'\varphi(x)$, где $\varphi(x)$ – заданная вектор-функция от вектора признаков x, первая компонента ее является единицей, λ – вектор коэффициентов, который получается в процессе аппроксимации. В примерах в качестве $\varphi(x)$ использованы линейная вектор-функция и полином второго порядка. К известным методам выбора класса функций [4] добавим возможность отбора признаков, применяемых в регрессионном анализе, по коэффициентам корреляции [9].

Преобразование обучающей выборки в выборку регрессионного анализа следует в случае двух классов осуществлять по правилу: номер первого класса заменять на $-C_{21}$, а номер второго класса – на C_{12}.

В случае нескольких классов, $K > 2$, задачу нужно решать путем сведения к $K-1$ задачам с двумя классами по принципу один класс против всех остальных классов. АпоВ одного из классов следует находить из условия равенства единице суммы АпоВ всех классов.

4. Способ решения задачи

Задача решается в два этапа: этап обучения и этап использования. Этап обучения заканчивается построением серии АДФ. В случае $K > 2$ нужно построить $K-1$ серий АДФ. На этапе использования вычисляются оценки АпоВ классов в заданных точках по серии или по сериям АДФ тем или иным способом. Способов оценивания может быть несколько. Далее описаны два способа.

4.1. Случай $K > 2$

Случай нескольких классов сводится к $K-1$ случаям с двумя классами путем рассмотрения точек других классов, кроме одного (далее — первого), в качестве точек другого (второго) класса. При этом для АДФ, отделяющей один класс от остальных, может потребоваться более сложный вид функции, чем при отделении одного класса от другого.

Манипулировать стоимостями ошибок будем следующим образом. Стоимости ошибок между классами, кроме первого, приравняем нулю, $C_{rs} = 0$, $r > 1$, $s > 1$. Обозначим через g класс, объединяющий все, кроме первого, классы. Обозначим стоимость ошибки от отнесения точки из g в первый класс через $C_{1g} = C > 0$. Стоимость ошибки от отнесения точки из первого класса в класс g примем $C_{g1} = 1 - C$. Тогда средние потери от отнесения точки в первый класс будут (1) : $G_1(x) = C\,(1 - p\,(1|x))$. Средние потери от отнесения точки в класс g будут: $G_g(x) = (1 - C)\,p\,(1|x)$. ДФ Андерсона, оптимально по Байесу отделяющая первый класс от остальных, $f(x, C) = G_1(x) - G_g(x) = C - p(1|x)$, что совпадает с (4). $C = p^*$ есть АпоВ первого класса на границе между первым классом и остальными классами.

4.2. Построение серии АДФ на этапе обучения

После отбора информативных признаков классов и выбора вида АДФ, методы которых не рассматриваем, серия АДФ строится по обучающей выборке методом [8]. Суть метода при $K = 2$ в том, что неизвестная ДФ Андерсона является функцией регрессии и ее надо аппроксимировать в окрестности нулевых значений. В обучающей выборке номера классов k заменяются на $C_{1k} - C_{2k}$, т.е. при $k = 1$ заменяются на $-C_{21}$, а при $k = 2$ заменяются на C_{12}. Далее запускается итерационный процесс, на первом шаге которого методом наименьших квадратов находится первая АДФ. На последующих шагах АДФ строятся взвешенным методом наименьших квадратов с заданной весовой функцией, принимающей максимальные значения в окрестности нулевых значений предыдущей АДФ. При этом минимизируется по λ_j путем решения системы линейных уравнений критерий аппроксимации

$$Q(\lambda_j) = N^{-1} \sum_n \left[C_{1k|n} - C_{2k|n} - \lambda_j' \varphi(x_n)\right]^2 \exp(-W_j|\lambda_{j-1}' \varphi(x_n)|), \tag{6}$$

где λ_{j-1} – решение, полученное на предыдущем шаге, N – количество строк в обучающей выборке, n – номер строки, $k \,|\, n$ – номер класса в строке n, $\varphi(x_n)$ – заданная вектор-функция от признаков в точке x_n , W_j – заданный коэффициент весовой функции на шаге j, $W_j > 0$, j – номер итерации, $j = 2, \ldots, J$, J – заданное количество шагов. Номер класса $k = \{1, 2\}$.

В качестве лучшего значения λ по итерациям выбирается вектор с наименьшими потерями

$$G(N, \lambda_j) = N^{-1}(C_{12}N_{12}(\lambda_j) + C_{21}N_{21}(\lambda_j)), \tag{7}$$

где N_{rs} – количество точек выборки из класса s, ошибочно отнесенных решающим правилом в класс r, N – количество точек в выборке. Процесс заканчивается после выполнения, например, заданного количества шагов. Лучшей принимается та АДФ в итерации, которая делит классы с наименьшей средней по выборке стоимостью ошибок (7).

Построение серии АДФ начинаем с большого значения $p^*_{\max}$, например 0,95. Задаем шаг уменьшения p^*, например 0,1, и нижнее значение $p^*_{\min}$, например 0,05. И шаг, и пределы в процессе построения серии, как правило, могут изменяться вследствие (5).

Шаг алгоритма 1. Вычисляем $C_{12} = p^*$ и $C_{21} = 1 - p^*$. Строим АДФ (6), (7), заменив в обучающей выборке номера классов k на $C_{1k} - C_{2k}$, получая таким образом выборку для АДФ.

Если точек, отнесенных в первый класс по полученной АДФ, “почти” не окажется, уменьшаем p^* на заданный шаг и повторяем пункт 1 при меньшем значении p^*. Величину “почти” здесь и в конце процесса задает оператор исходя из объема выборки.

Запоминаем первую удачную АДФ серии, соответствующее ей значение $p^*_{\max} = p^*_1$ и количество точек, попавших в первый класс. Переходим к шагу 2.

Шаг алгоритма 2. Задаем для p^* меньшее значение. Вычисляем C_{12} и C_{21} для заданного p^*, находим АДФ (6), (7) по выборке, заменив в ней номера классов k на $C_{1k} - C_{2k}$. Если количество точек, отнесенных в первый класс, больше, чем в предыдущем удачном случае[2], но меньше, чем количество то-

[2] Область первого класса в пространстве признаков, соответствующая меньшей АпоВ p^*, включает в себя область, соответствующую большей p^*. Естественнее считать, что чем больше область, тем больше точек выборки первого класса должно попасть в нее.

чек в выборке, то, запомнив очередную удачную АДФ, значение p^* и количество точек, попавших в первый класс, повторяем шаг 2.

Иначе шаг 2 повторяется без запоминания удачной АДФ с целью проверки следующего, меньшего, значения p^* в качестве кандидата для очередной АДФ.

Если "почти" все точки выборки попали в первый класс или достигнуто нижнее значение $p^*_{\min}$, то процесс создания серии АДФ прекращаем, запомнив этот шаг как последний J, запомнив значение $p^*_{\min} = p^*_J$ и количество точек, попавших в первый класс.

Если предполагается использовать для оценки качества метода классификации ROC-анализ [4], как это сделано далее в последнем примере, то при получении серии АДФ для вычисления показателей ROC-анализа следует так же фиксировать количества точек выборки, классифицированных правильно и неправильно для каждой АДФ в серии.

5. Оценка АпоВ класса по серии АДФ

Пусть в задаче $K = 2$ имеется серия из АДФ в количестве J, $F_j(x, p^*_j)$, $j = 1 \div J$, в точках нулевых значений которых АпоВ первого класса равны p^*_j по построению. Для оценки АпоВ в заданных точках можно использовать различные методы. Рассмотрим два метода: метод линейной интерполяции и экстраполяции и метод потенциальных функций [10] (ядерного сглаживания).

5.1. Экстраполяция и линейная интерполяция

Для оценки АпоВ первого класса в заданной точке x нужно взять первую АДФ из упорядоченной по убыванию p^* серии, для которой $F_j(x, p^*_j) \le 0$.

Экстраполяция. Если это первая АДФ в серии, то точка относится к первому классу, $F_1(x, p^*_{\max}) \le 0$ и АпоВ $> p^*_{\max}$. Для получения оценки АпоВ в этой точке следует выполнить нижеописанную процедуру экстраполяции в большую сторону, где $p^*_{\max} <$ АпоВ < 1, и закончить процесс.

Если в процессе перебора АДФ от первой и до последней включительно все АДФ $F_j(x, p^*_j) > 0$, т.е. точка относится к первому классу с АпоВ $< p^*_{\min}$, то для получения оценки АпоВ в этой точке следует выполнить нижеописанную процедуру экстраполяции в меньшую сторону, $0 <$ АпоВ $< p^*_{\min}$ и закончить процесс.

Для точек, расположенных до первой АДФ, $F_1(x, p^*_{\max}) \le 0$, экстраполируем по (4):

$$p(1|x) = \min\left\{1,\ p^*_{\max} - F_1(x, p^*_{\max})\right\}. \tag{8}$$

Для точек за последней АДФ, $F_J(x, p^*_{\min}) > 0$, экстраполируем по (4):

$$p(1|x) = \max\left\{0, p^*_{\min} - F_J(x, p^*_{\min})\right\}. \tag{9}$$

Линейная интерполяция. Если в процессе перебора АДФ встретится первая АДФ $F_-(x, p^*_-) \le 0$ (отмечена индексом минус) после предыдущей, в которой $F_+(x, p_+^*) > 0$ (отмечена индексом плюс), то с учетом знаков АДФ оценкой АпоВ в точке x будет

$$p(1|x) = \frac{(p_-^* - F_-(x, p_-^*))F_+(x, p_+^*) - (p_+^* - F_+(x, p_+^*))F_-(x, p_-^*)}{F_+(x, p_+^*) - F_-(x, p_+^*)}. \tag{10}$$

В качестве расстояния точки до границы использован модуль АДФ в заданной точке.

5.2. Сглаживание

Имея серию АДФ $F_j(x, p_j^*)$, $j = 1 \div J$, с известными на границах классов величинами p_j^* логично вычислять оценку АпоВ в заданной точке по близким к ней АДФ с большими весами. В качестве меры близости к АДФ будем использовать модуль значения АДФ в заданной точке. Чтобы близкие АДФ оказывали большее влияние на оценку, используется ограниченная функция, убывающая по мере удаления от нулевых значений АДФ. Из многих видов таких функций выберем, например, функцию

$$w_j(x) = \exp(-W|F_j(x, p_j^*)|), \quad j = 1 \div J, \tag{11}$$

где W – параметр, задающий сглаживающие свойства метода, $W > 0$.

Взвешенной оценкой АпоВ в точке x с помощью серии АДФ и (11) будет

$$p(1|x) = \sum_j p_j^* w_j(x) / \sum_j w_j(x). \tag{12}$$

По сравнению с интерполяционным методом сглаживание зависит от выбора вида функции и параметра сглаживания.

6. Примеры

6.1. Нормальные условные распределения вероятностей признаков классов

Приведем модельный пример расчета АпоВ разными способами в случае разделения двух классов в двумерном пространстве признаков при нормальных условных распределениях вероятностей признаков классов. Вектор средних первого класса — $m_1 = (0, 0)$, ковариационная матрица — $S_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Для второго класса: $m_2 = (0, 2)$, $S_2 = \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix}$. Априорная вероятность первого класса $P_1 = 0{,}6$, второго — $P_2 = 0{,}4$. По исходным данным и по сгенерированной выборке в 500 точек вычислены АпоВ способами:

1) точным — по известным исходным данным с использования формулы Байеса;

2) выборочным — по выборочным оценкам параметров распределений с использованием формулы Байеса;

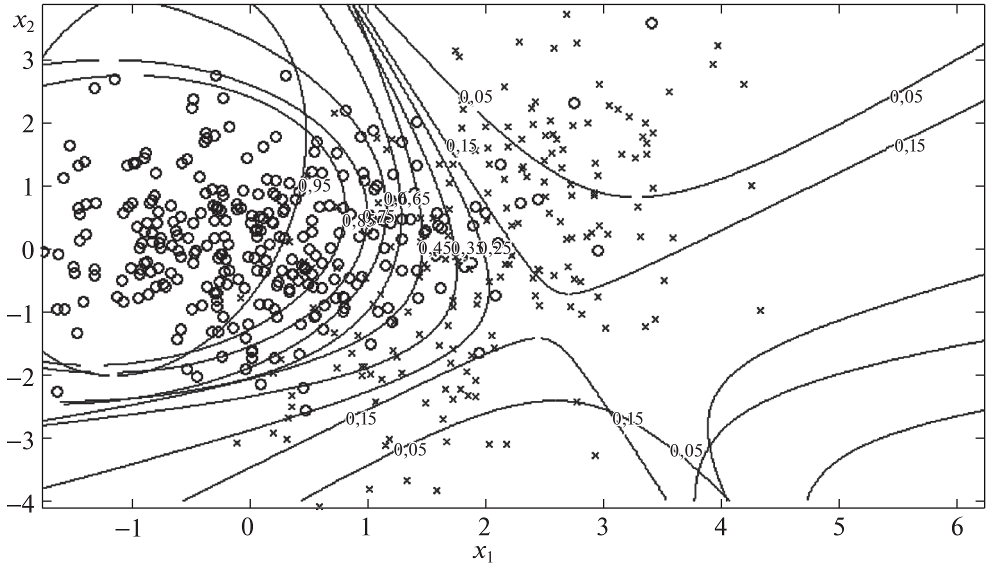

Рис. 1.

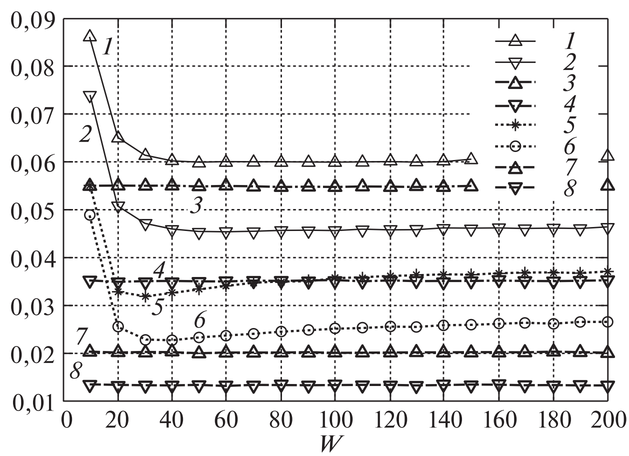

Рис. 2.

3) интер-экстраполяционным по сериям АДФ. АДФ — полином второго порядка;

4) методом сглаживания (11)–(12) по той же серии АДФ.

На рис. 1 представлены АДФ, построенные в диапазоне от 0,95 до 0,05 АпоВ первого класса методом [8]. Кружками изображены точки первого класса, крестиками — второго класса, сплошными линиями — границы $F_j(x, p_j^*) = 0$. На рис. 2 изображены среднемодульные (СМО) и среднеквадратичные (СКО) ошибки методов оценки АпоВ в зависимости от W (11). Пересечения границ на рис. 1 в области малого количества точек можно при необходимости избежать путем уменьшения количества АДФ в серии или увеличения объема выборки.

Пояснения к легенде на рис. 2: *1, 2* — СКО и СМО расхождения между АпоВ, полученными ядерным сглаживанием, и АпоВ первого класса, полученными по исходным данным; *3, 4* — СКО и СМО расхождения между АпоВ по исходным данным и АпоВ, полученными путем интер- и экстраполяции; *5, 6* — расхождение между АпоВ, полученными ядерным сглаживанием и АпоВ, полученными путем интер- и экстраполяции; *7, 8* — СКО и СМО расхождения между АпоВ по выборочным оценкам параметров распределений и АпоВ по исходным данным. Графики *1, 2* и *5, 6* зависят от W.

По выборкам обычно нельзя сделать надежных сравнительных выводов. В данном случае можно лишь заметить, что результаты довольно близки друг другу. Метод сглаживания уступил интер-экстраполяционному, но картина может измениться в пользу сглаживания, если более удачно выбрать вид функции ядра и ее параметры.

Интер-экстраполяционный метод был хуже метода оценки АпоВ с выборочными оценками параметров распределений. Но следует подчеркнуть, что он не связан с предположениями о законах распределений, если строить АДФ по методике [8]. Для него важен лишь вид ДФ Андерсона в окрестности их нулевых значений.

СКО и СМО оценки АпоВ первого класса интер-экстраполяционным методом и методом оценки их по формуле Байеса по исходным данным составили 0,053 и 0,035. На примере с выборкой в 5000 точек, (графики не приведены) эти показатели снизились соответственно до 0,021 и 0,016.

6.2. Диагностика заболевания

Рассмотрим пример решения конкурсной задачи диагностики заболевания [11]. Конкурсная задача не преследовала практических целей. Ее целью являлось сравнение способов классификации при большом количестве признаков и сравнительно небольшом объеме обучающей выборки. Наименование болезни, характеристики признаков, методы сбора данных в данном случае не известны. Обучающая выборка — матрица 252×217. В 252 строках первые 99 строк соответствуют здоровым людям, остальные 153 строки — больным. Первый столбец содержит номер класса: 0 — здоровые, 1 — больные. Остальные столбцы соответствуют признакам. В роли признаков использованы частоты триграмм исходных анализов.

Цель примера — показать, что простые методы отбора признаков для аппроксимации ДФ Андерсона по методике [8] в сочетании с предложенным методом оценивания АпоВ классов способны дать обнадеживающие результаты уже на начальном этапе решения задачи.

ДФ Андерсона является функцией регрессии, из набора 216 признаков отобраны наиболее коррелированные с искомой величиной — первым столбцом — признаки. Это три признака с номерами столбцов в обучающей матрице 22, 104, 115 имеют коэффициенты корреляции 0,60, 0,64, 0,64. Корреляция их между собой: 0,52, 0,48 и 0,66.

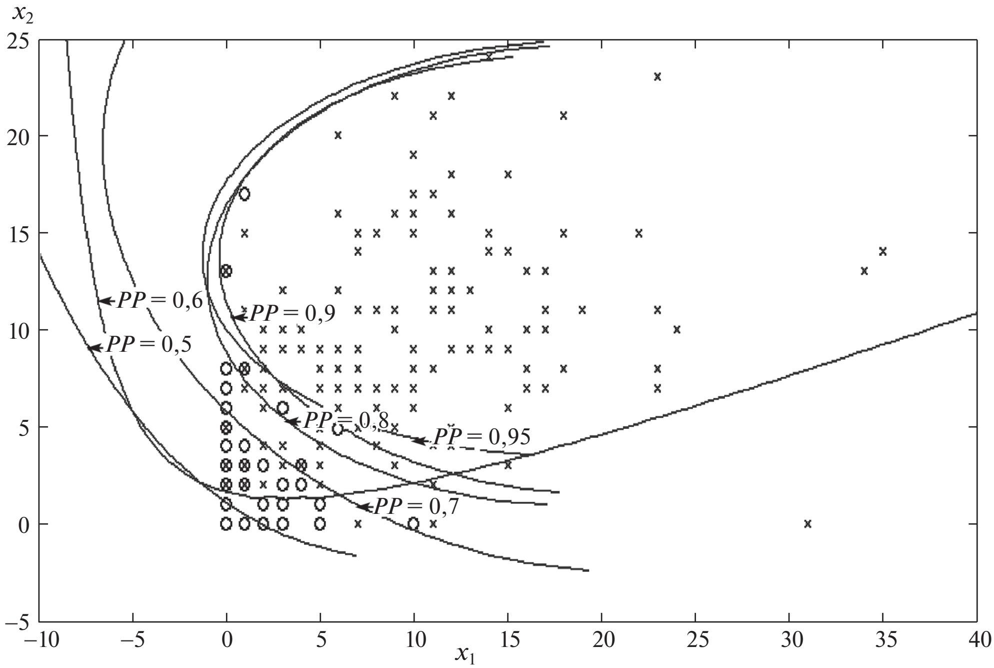

Рис. 3.

Аппроксимировав ДФ Андерсона полиномом второго порядка методом [8], получили по АДФ среднюю по выборке вероятность ошибки диагностики 4,8%. АДФ:

$$
\begin{aligned}
(13)\qquad f_{12}(x) = {}&-0{,}97 - 0{,}00093x_1^2 - 0{,}013x_2^2 - 0{,}0045x_3^2 + 0{,}0067x_1x_2 + \\
&+ 0{,}062x_1x_3 + 0{,}011x_2x_3 + 0{,}053x_1 + 0{,}19x_2 + 0{,}13x_3,
\end{aligned}
$$

где x_1, x_2, x_3 – признаки (вышеуказанные столбцы).

Если $f_{12}(x) < 0$, то точку следует отнести в класс здоровых, иначе — в класс больных.

При аппроксимации ДФ линейной функцией вероятность ошибки 6,8%. АДФ:

$$
(14)\qquad f_{12}(x) = -0{,}97 + 0{,}17x_1 + 0{,}11x_2 + 0{,}14x_3.
$$

АДФ (13), (14) получены для равных стоимостей ошибок.

Для сравнения получены решения задачи с помощью методов quadratic classify и linear classify МАТЛАБ [12], в которых используются предположения о нормальных условных законах распределения признаков классов с разными и одинаковыми ковариационными матрицами и применяются выборочные оценки параметров распределений. Величины ошибок классификации для quadratic АДФ и linear АДФ составили 10,3% и 9,9% соответственно. На рис. 3 изображены проекции на плоскость x_1x_2 точек выборки и участков некоторых из границ АДФ типа (13). Проекции точек здоровых пациентов

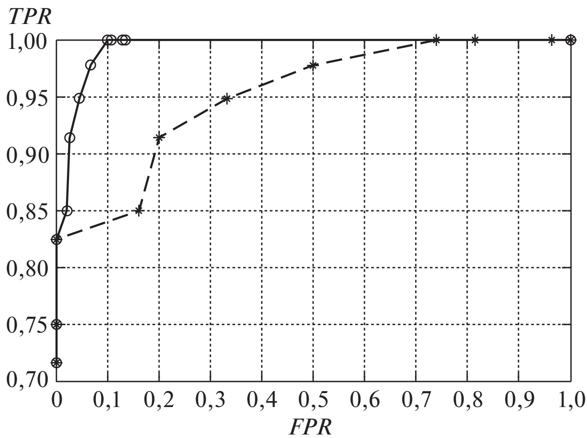

Рис. 4.

отмечены кружками, больных — крестиками. АпоВ класса, $PP = p^*$, на границах указаны для больных.

Относительные положения проекций границ на плоскость зависят от уровня фиксации x_3.

Расхождения по СМО и СКО между АпоВ, полученными по сериям полиномиальных АДФ типа (13) и линейных типа (14) (рисунок не представлен), составили 0,07 и 0,11.

Качество самого метода классификации, а не одной АДФ, иногда оценивают с помощью ROC–анализа [4]. Для построения ROC — кривой для серии АДФ, созданных одним методом, вычислялись показатели: TPR — доля правильных положительных классификаций и FPR — доля ошибочных положительных классификаций. Положительными считались заболевшие, отрицательными — здоровые. Изменяемым параметром метода являлась величина p^*. Полученные TPR и FPR для ряда значений p^* наносились на график, рис. 4. Классический график должен содержать значения FPR от нуля до единицы. В данном случае при изменении p^* в пределах [0,02; 0,95] показатель FPR изменялся только в пределах [0; 0,14] и TPR соответственно в пределах [0,717; 1], выходя на свой естественный предел, равный единице. Поэтому строились два графика: один растягивался по оси FPR добавлением точки (1, 1), сплошная линия на рис. 4, второй — изменением масштаба FPR до пределов от 0 до 1.

В качестве интегрального показателя качества метода является AUC — площадь между графиком и осью FPR. Для АДФ — полинома второго порядка, рис. 4, показатели AUC для двух графиков равны 0,99 и 0,95. Для линейного АДФ, график не представлен, показатели AUC равны 0,98 и 0,95. Работа выполнена в среде МАТЛАБ [12].

7. Заключение

1. Предложен способ оценки апостериорных вероятностей классов в заданных точках пространства признаков классов путем построения по обучающей

выборке с учителем серии аппроксимаций дискриминантных функций Андерсона, по которым путем интер- и экстраполяции, ядерного сглаживания или иным методом вычисляются оценки апостериорных вероятностей класса.

2. Указан способ сведения задачи для нескольких классов к задачам с двумя классами.

3. Приведен модельный пример для двух классов с нормальными законами условных распределений вероятностей двумерных признаков классов на обучающей выборке с учителем в 500 точек для сравнения методов вычисления апостериорных вероятностей класса в точках выборки: точного — по известным нормальным законам распределений; выборочного — по выборочной оценке параметров нормальных распределений; предложенного — по серии аппроксимаций дискриминантных функций с использованием линейной интерполяции и метода ядерного сглаживания при различных значениях параметра сглаживания.

4. В качестве примера с реальными данными выполнен расчет апостериорных вероятностей наличия у 252 пациентов некоторого заболевания по данным конкурсной задачи — выборки, содержащей 216 признаков, опубликованной в интернете. Значение показателя качества метода AUC = 0,95, полученное по ROC — кривой, при трех признаках, отобранных по коэффициентам корреляции с искомой величиной, свидетельствует о перспективности метода.

Результаты оказались лучше методов МАТЛАБ quadratic classify и linear classify.

СПИСОК ЛИТЕРАТУРЫ

1. *Anderson T.W.* An Introduction to Multivariate Statistical Analysis. Third Edition. John Wiley & Sons, 2003. 721 p.
2. *Кнеллер В.Ю.* Об определении и специфике автоматического контроля // АиТ. 1962. № 4. С.509–518.
3. *Мерков А.Б.* Распознавание образов. Введение в методы статистического обучения. Едиториал УРСС, 2011. 256 с.
4. *Воронцов К.В.* Машинное обучение (курс лекций),

   http://www.machinelearning.ru/wiki/index.php?title=Заглавная_страница)
5. *Platt J.* Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods (PDF) // Advances Large Margin Classifiers. 1999. V. 10. No. 3. P. 61–74.
6. *Nuculescu-Mizil A., Caruana R.* Predicting Good Probabilities with Supervised Learning. Department of Computer Science. Cornell University. Ithaca NY 14853. P. 1–8.

   http://datascienceassn.org/sites/default/files/Predicting%20good%20probabilities

   %20with%20supervised%20learning.pdf
7. *Зенков В.В.* Оценка апостериорной вероятности класса в точке по аппроксимации одной дискриминантной функции // АиТ. 2018. № 9. С. 46–59.

   Zenkov V.V. Estimating the Probability of a Class at a Point by the Approximation of one Discriminant Function // Autom. Remote Control. 2018. V. 79. No. 9. P. 1580–1590.

8. *Зенков В.В.* Использование взвешенного метода наименьших квадратов при аппроксимации дискриминантной функции цилиндрической поверхностью в задачах классификации // АиТ. 2017. № 9. С. 145–158.

   Zenkov V.V. Using Weighted Least Squares to Approximate the Discriminant Function with a Cylindrical Surface in Classification Problem // Autom. Remote Control. 2017. V. 78. No. 9. P. 1662–1673.

9. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ, 3-е изд.: Пер. с англ. М.: Изд. дом "Вильямс", 2007. 912 с.

10. *Айзерман М.А., Браверман Э.М., Розоноэр Л.И.* Метод потенциальных функций в теории обучения машин М.: Наука, 1970. 384 с.

11. Данные для задания на ТМШ 2014.

    http://www.machinelearning.ru/wiki/images/e/e1/School-VI-2014-task-3.rar

12. *Ануфриев И.Е., Смирнов А.Б., Смирнова Е.Н.* MATLAB 7 СПб.: БХВ-Петербург, 2005. 1104 с.

Статья представлена к публикации членом редколлегии Б.М. Миллером.

Поступила в редакцию 21.12.2017

После доработки 25.09.2018

Принята к публикации 08.11.2018