Л.В. Уткин, Ю.А. Жук, И.А. Селиховкин

Модель классификации на основе неполной информации о признаках в виде их средних значений

Аннотация. Предлагается модель классификации при ограниченной информации в форме математических ожиданий признаков, основанная на минимаксной (минииминной) стратегии принятия решений. Дискриминантная функция вычисляется максимизацией (минимизацией) функционала риска, как меры ошибочной классификации, по множеству распределений вероятностей с границами, которые определяются информацией о признаках, и минимизацией по множеству параметров. Алгоритм сводится к решению параметрической задачи линейного программирования.

Ключевые слова: классификация, машинное обучение, линейное программирование, функционал риска, функция потерь, математическое ожидание, минимаксная стратегия.

Введение

Бинарная или двоичная классификация является одной из частных задач статистического машинного обучения, которая может рассматриваться как задача разделения некоторых объектов на два класса в соответствии с их свойствами или признаками [1]. Основной целью статистического обучения является прогнозирование ненаблюдаемых выходных значений y на основе наблюдаемых входных векторах $\mathbf{x}$. Для этого необходимо оценить некоторую функцию f, исходя из обучающей выборки, которая представляет собой множество пар $(\mathbf{x}, y)$. Поэтому задача бинарной классификации сводится к тому, чтобы отнести каждый образец или пример $\mathbf{x}$ к одному из двух классов при помощи дискриминантной функции f.

За последнее десятилетие было предложено огромное число методов для решения задачи классификации. Однако многие из них предполагают наличие большой обучающей выборки, состоящей при этом из точных данных или наличие известного типа функции распределения вероятностей признаков, т.е. основаны на достаточно жестких предположениях. В связи с этим, значительным образом ограничивается использование известных методов классификации в различных прикладных областях. Предлагаемые робастные методы классификации [6, 7] также используют дополнительную информацию, которая не всегда может быть получена.

Одной из наиболее часто встречающихся ситуаций является та, когда нам необходимо классифицировать объекты при наличии экстремально малого количества информации об их признаках. Иными словами мы можем вообще не наблюдать объекты, но должны их классифицировать на основе ограниченной информации. Предположим, что производится древесностружечная плита, качество или прочность которой зависит от некоторого набора параметров, например, от плотности, веса, типа наружного слоя, водостойкости и т.д. Если параметры не измерялись или не наблюдались перед классификацией, то достаточно сложно забраковать новые плиты или классифицировать их на два класса (брак и качественные), так как у нас нет обучающей выборки с измеренными параметрами. Однако если мы знаем, например, какое количество древесных частиц и связующего вещества использовалось при производстве N плит, то можно оценить средний вес одной плиты. В данной ситуации очень важной частью информации, которая может быть использована, является оценки экспертов. Они могут предоставить информацию или свои суждения о некоторых усредненных значениях признаков для каждого класса объектов, Для экспертов это не сложно, так как такая информация является наиболее понятной и простой. В то же время, для эксперта намного труднее дать оценки, например, дисперсии или других статистических характеристик.

В связи с этим мы ограничимся рассмотрением только информации о среднем и попытаемся на ее основе решить задачу двоичной классификации. Если признаки не являются ограниченными, то можно доказать, что только средние значения не позволяют построить модель классификации. Поэтому небольшое предположение, которое необходимо добавить для получения нетривиальных оценок, является задание конечных границ возможных значений каждого признака. Данное предположение выполняется в большинстве случаев, особенно в ситуациях, когда признаки шкалируются на основе конечных шкал.

В работе рассматривается решение задачи классификации при условиях, что известны математические ожидания или средние значения каждого признака, а также то, что значения признаков ограничены нижней и верхней границами. При этом, несмотря на существенную ограниченность исходной информации, она может быть успешно использована для построения модели классификации.

Основная идея построения новых моделей, учитывающих ограниченность исходной информации, состоит в том, что средние значения и ограничения на значения признаков образуют множество распределений вероятностей, ограниченное некоторыми нижней и верхней функциями распределения вероятностей (ФРВ). При этом границы множества ФРВ зависят от неизвестных параметров дискриминантной функции, которая в конечном итоге должна быть найдена для решения задачи классификации [10, 12]. Следует отметить, что рассматриваемое множество распределений не является параметрическим множеством, т.е. множеством распределений вероятностей одного типа. Это множество всех возможных распределений ограничено нижней и верхней границами, что является важной особенностью предлагаемого в работе подхода. Два распределения вероятностей выбираются из всего множества распределений. Они максимизируют и минимизируют функционал риска, являющегося показателем ошибки классификации. Другими словами, известная минимаксная стратегия принятия решений используется для решения классификационной задачи. Эта стратегия предполагает страховку против наихудшего случая [8]. Другая стратегия – миниминная или оптимистическая также рассмотрена в работе.

Вторая идея, лежащая в основе предлагаемых моделей, использует тот факт, что при наличии информации о средних значениях признаков и предположении, что дискриминантная функция линейна, можно просто выразить среднее значение функции f и ее границ через параметры классификации α. Поэтому, идея заключается в том, что множество распределений вероятностей строится не для каждого признака $\mathbf{x}^{(i)}$, а для функции f. В результате этого границы ФРВ и значений функции f становятся функциями параметров классификации.

Параметры классификации α дискриминантной функции в итоге вычисляются путем минимизации верхней границы функционала риска решением задачи линейного программирования.

Предлагаемый метод является развитием результатов, представленных в работе [13].

Стандартная задача классификации

Задача двоичной классификации сводится к оценке области значений независимой переменной $\mathbf{x}$, при которых каждый класс наблюдается наибольшей частотой. Имеется множество значений независимой переменной или вектора переменных $\mathbf{x}$, каждому из которых можно поставить в соответствие одно из двух значений зависимой переменной y, определяющей различные классы объектов, например $y=-1$ и $y=1$. В некоторых случаях используют значения $y=0$ и $y=1$. Предположим, что имеются эмпирические данные

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n) \in \mathbb{R}^n \times \{-1, +1\}.$$

Здесь $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ – некоторое непустое множество *образцов, примеров, входных переменных*; $y_1, \ldots, y_n$ *– значения классов*. Разобьем обучающее множество данных на два подмножества. Первое подмножество, обозначаемое Ψ_{-1}, соответствует точкам с $y=-1$, т.е.

$$\Psi_{-1}=\{(\mathbf{x}_i, y_i):\ y_i=-1,\ i=1,...,n\}.$$

Второе подмножество Ψ_{+1} соответствует точкам с $y=1$, т.е.

$$\Psi_{+1}=\{(\mathbf{x}_i, y_i):\ y_i=+1,\ i=1,...,n\}.$$

В дальнейшем будем предполагать, что подмножество Ψ_y содержит n_y элементов. При этом выполняется условие $n_{-1}+n_1=n$.

Задача классификации обычно характеризуется неизвестной ФРВ $F_0(\mathbf{x}, y)$ или функцией плотности вероятности (ФПВ) $p_0(\mathbf{x}, y)$, определенной на множестве $\mathbb{R}^n\times\{-1,+1\}$.

Для решения задачи классификации необходимо определить *решающую функцию* $g(\mathbf{x})$, которая наиболее точным образом прогнозирует значения классов y для любого образца $\mathbf{x}$, который может принадлежать, а может и не принадлежать обучающей выборке.

Одним из возможных подходов для решения этой задачи основан на использовании *дискриминантной функции* $f(\mathbf{x})$, знак которой определяет соответствующий класс: $g(\mathbf{x})=\mathbf{sgn}(f(\mathbf{x},\alpha))$. Дискриминантная функция $f(\mathbf{x})$ в свою очередь зависит от параметров $\alpha=(\alpha_0,\alpha_1,...,\alpha_m)$, $\alpha\in\Lambda$, которые определяются на основе обучающей выборки посредством некоторого алгоритма обучения. Поэтому в дальнейшем функция f будет записываться с параметрами, т.е. $f(\mathbf{x},\alpha)$. Кроме того, будем предполагать, что дискриминантная функция является линейной, т.е.

$$f(\mathbf{x},\alpha)=\alpha_0+\sum_{i=1}^{m}\alpha_i x_i,\ x_i\in\mathbf{x}.$$

Для сокращения записи, будем иногда использовать, также стандартное обозначение для поточечного произведения $f(\mathbf{x},\alpha)=\alpha_0+\langle\alpha\mathbf{x}\rangle$. Элемент с номером i вектора $\mathbf{x}_k$ будет обозначаться $x_i^{(k)}$.

Таким образом, задача классификации во многих случаях сводится к задаче определения параметров α дискриминантной функции. Одним из наиболее распространенных способов решения этой задачи является минимизация *функционала риска* [1]:

$$R(\alpha)=\int_{\mathbb{R}^n\times\{-1,1\}}\mathbf{1}\{g(\mathbf{x})\neq y\}\mathbf{d}F_0(\mathbf{x},y)\mathbf{d}y.$$

Здесь *функция потерь* $L(\mathbf{x},y)=\mathbf{1}\{g(\mathbf{x})\neq y\}$ принимает значение 1, если знак дискриминантной функции (соответствующего прогнозируемого класса) не совпадает с классом y. Как следует из определения функционала риска, он является мерой ошибки классификации. Минимизация функционала риска осуществляется по классу функций $f(\mathbf{x},\alpha)$, $\alpha\in\Lambda$ или просто по множеству параметров α при заданном виде дискриминантной функции. Другими словами, функция $f(\mathbf{x},\alpha_{\text{opt}})$ обеспечивает минимум функционала $R(\alpha)$, т.е. $R(\alpha_{\text{opt}})=\min_\alpha R(\alpha)$.

Задача классификации при множестве функций распределения вероятностей

Следует отметить, что функция потерь зависит от дискриминантной функции *f*. Заменяя ФРВ $F_0(\mathbf{x},y)$ вектора случайных величин $\mathbf{x}$ функцией $F(f,y)$ случайной величины *f*, перепишем функционал риска в следующем виде:

$$R(\alpha)=\int_{\mathbb{R}\times\{-1,1\}}\mathbf{1}\{\mathbf{sgn}(f)\neq y\}\mathbf{d}F(f,y)\mathbf{d}y.$$

Представим совместную ФРВ $F(f,y)$ в виде $F(f,y)=F(f\mid y)\cdot P(y)$. Здесь $P(y)$ является априорной вероятностью того, что случайно выбранный образец $\mathbf{x}$ принадлежит классу y. Тогда функционал риска можно записать с учетом двух значений величины y

$$R(\alpha)=P(-1)R_{-1}(\alpha)+P(1)R_{+1}(\alpha).$$

Здесь

$$R_{-1}(\alpha)=\int_{\mathbb{R}}\mathbf{1}\{g(f)\neq -1\}\mathbf{d}F(f\mid -1),$$

$$R_{+1}(\alpha)=\int_{\mathbb{R}}\mathbf{1}\{g(f)\neq 1\}\mathbf{d}F(f\mid 1).$$

Можно предположить, что известны некоторые границы для множества $\mathcal{F}(y)$ всех ФРВ $F(f\mid y)$ с точностью до α. Нижняя граница обозначена $\underline{F}(f\mid y)$ и верхняя граница – $\overline{F}(f\mid y)$. При этом точная ФРВ не известна. Отсюда можно записать:

$\mathcal{F}(y) =$
$= \{F(f \mid y) \mid \forall f \in \mathbb{R}, \underline{F}(f \mid y) \le F(f \mid y) \le \overline{F}(f \mid y)\}.$

Другими словами, имеется некоторое неизвестное истинное распределение вероятностей $F(f \mid y) \in \mathcal{F}(y)$ для каждого $y \in \{-1,+1\}$, но оно неизвестно, и вся информация о нем заключается в том, что оно принадлежит множеству $\mathcal{F}(y)$.

Поскольку вместо одной функции распределения $F(f \mid y)$ известно только некоторое множество распределений, то поиск параметров дискриминантной функции зависит от того, какая функция распределения из множества $\mathcal{F}(y)$ будет выбрана. В связи с этим будут использоваться две возможные стратегии выбора функции распределения, а значит две стратегии для определения параметров дискриминантной функции: *минимаксная* (пессимистическая) и *минимииная* (оптимистическая) *стратегии*. В соответствии с минимаксной стратегией выбирается такое распределение вероятностей из множества $\mathcal{F}(-1)$ и распределение вероятностей из множества $\mathcal{F}(+1)$, что показатели риска $R_{-1}(\alpha)$ и $R_{+1}(\alpha)$ достигают для каждого фиксированного значения α своего максимального значения. Другими словами, нам следует выбрать наихудшее распределение, приводящее к наибольшему значению функционала риска. Минимаксную стратегию можно рассматривать как некоторую страховку против наихудшей ситуации, так как эта стратегия минимизирует ожидаемые потери в наименее благоприятной ситуации [8]. В соответствии с минимииной стратегией выбирается такое распределение вероятностей из множества $\mathcal{F}(-1)$ и распределение вероятностей из множества $\mathcal{F}(+1)$, что показатели риска $R_{-1}(\alpha)$ и $R_{+1}(\alpha)$ достигают для каждого фиксированного значения α своего минимального значения.

Отметим, что оптимальные распределения вероятностей могут быть различными при различных значениях параметров α. Отсюда следует, что соответствующие оптимальные распределения вероятностей зависят от α.

Так как множества $\mathcal{F}(-1)$ и $\mathcal{F}(+1)$ получены независимо для $y=-1$ и $y=1$ соответственно, то

$$\overline{R}(\alpha) = \max_{F(f \mid y) \in \mathcal{F}(-1) \times \mathcal{R}(+1)} R(\alpha) =$$
$$= P(-1) \max_{F(f \mid -1) \in \mathcal{F}(-1)} R_{-1}(\alpha) + P(1) \max_{F(f \mid +1) \in \mathcal{F}(+1)} R_{+1}(\alpha).$$

Минимаксный функционал риска принимает вид: $\overline{R}(\alpha_{\text{opt}}) = \min_{\alpha} \overline{R}(\alpha)$.

Рассмотрим более подробно задачу оптимизации $\max_{F(f \mid -1) \in \mathcal{F}(-1)} R_{-1}(\alpha)$ и условие $\mathbf{1}\{g(f) \ne -1\}$. Функция потерь $L(f,-1)$ является возрастающей при $f \le 0$. Следовательно, верхняя граница функционала риска $R_{-1}(\alpha)$, т.е. максимум функционала $R_{-1}(\alpha)$ на множестве всех распределений из $\mathcal{F}(-1)$, достигается на распределении вероятностей $\underline{F}(f \mid -1)$. Отсюда можно записать:

$$\overline{R}_{-1}(\alpha) = \int_{\mathbb{R}} \mathbf{1}\{g(f) \ne -1\} \mathbf{d}\underline{F}(f \mid -1) = 1 - \underline{F}(0 \mid -1).$$

Аналогичным образом можно рассмотреть задачу $\max_{F(f \mid +1) \in \mathcal{F}(+1)} R_{+1}(\alpha)$. Функция потерь $L(f,+1)$ в этом случае является убывающей. Поэтому верхняя граница функционала $R_{+1}(\alpha)$ достигается на распределении вероятностей $\overline{F}(f \mid 1)$. Отсюда следует, что

$$\overline{R}_{+1}(\alpha) = \int_{\mathbb{R}} \mathbf{1}\{g(f) \ne 1\} \mathbf{d}\overline{F}(f \mid 1) = \overline{F}(0 \mid 1).$$

Полученные выражения дают возможность записать верхнюю границу суммарного функционала риска $R(\alpha)$ в следующем виде:

$$\overline{R}(\alpha) = P(1)\overline{F}(0 \mid 1) + P(-1) - P(-1)\underline{F}(0 \mid -1). \quad (1)$$

Задача оптимизации для вычисления оптимальных значений параметров α дискриминантной функции записывается как

$$\overline{R}(\alpha_{\text{opt}}) = \min_{\alpha} \left(P(1)\overline{F}(0 \mid 1) + P(-1) - P(-1)\underline{F}(0 \mid -1) \right) =$$
$$= \max_{\alpha} \left(P(-1)\underline{F}(0 \mid -1) - P(1)\overline{F}(0 \mid 1) - P(-1) \right) = \max_{\alpha} Q(\alpha).$$

Здесь для упрощения дальнейших расчетов введен функционал качества классификации $Q(\alpha) = -\overline{R}(\alpha)$. Теперь мы имеем две задачи. Первая заключается в определении граничных ФРВ $\overline{F}(f=0 \mid y=1)$ и $\underline{F}(f=0 \mid y=-1)$. Вторая задача сводится к определению вероятностей $P(-1)$ и $P(1)$.

Аналогичным образом рассмотрим оптимистическую стратегию. Задачу минимизации функционала риска представляем в виде суммы двух задач минимизации маргинальных функционалов риска:

$$\underline{R}(\alpha) = \min_{F(f\mid y)\in\mathcal{F}(-1)\times\mathcal{R}(+1)} R(\alpha) =$$
$$= P(-1)\min_{F(f\mid -1)\in\mathcal{F}(-1)} R_{-1}(\alpha) + P(1)\min_{F(f\mid +1)\in\mathcal{F}(+1)} R_{+1}(\alpha).$$

Рассмотрим детально задачу $\min_{F(f\mid -1)\in\mathcal{F}(-1)} R_{-1}(\alpha)$ и условие $\mathbf{1}\{g(\mathbf{x}) \neq -1\}$. Так как функция потерь $L(f,-1)$ является возрастающей, то нижняя граница функционала $R_{-1}(\alpha)$, т.е. минимум функционала $R_{-1}(\alpha)$ на множестве всех распределений вероятностей $\mathcal{F}(-1)$ достигается на распределении $\overline{F}(f,-1)$. Отсюда

$$\underline{R}_{-1}(\alpha) = \int_{\mathbb{R}} \mathbf{1}\{g(f) \neq -1\} \mathrm{d}\overline{F}(f,-1) = 1 - \overline{F}(0,-1).$$

Аналогичным образом определим нижнюю границу функционала $R_{+1}(\alpha)$

$$\underline{R}_{+1}(\alpha) = \int_{\mathbb{R}} \mathbf{1}\{g(f) \neq 1\} \mathrm{d}\underline{F}(f,1) = \underline{F}(0,1).$$

Полученные выражения дают возможность записать нижнюю границу суммарного функционала риска $R(\alpha)$ в следующем виде:

$$\underline{R}(\alpha) = P(-1) + P(-1)\overline{F}(0\mid -1) + P(1)\underline{F}(0\mid 1).$$

Множества распределений, построенные на основе среднего значения и границ случайных величин

Предположим, что имеется информация о границах $X \in [a,b]$ и о среднем значении M ($a < M < b$) случайной величины *X*. Верхняя граница ФРВ $\overline{F}(x)$ может быть получена с использованием принципа продолжения [2, 5, 15], который в данном случае представляется в виде следующей задачи линейного программирования:

$$\overline{F}(x) = \min_{v,w}\left(v + wM\right),$$

при ограничениях $v, w \in \mathbb{R}$, $v + wz \geq \mathbf{1}\{z \leq x\}$, $\forall z \in [a,b]$.

Отсюда получаем

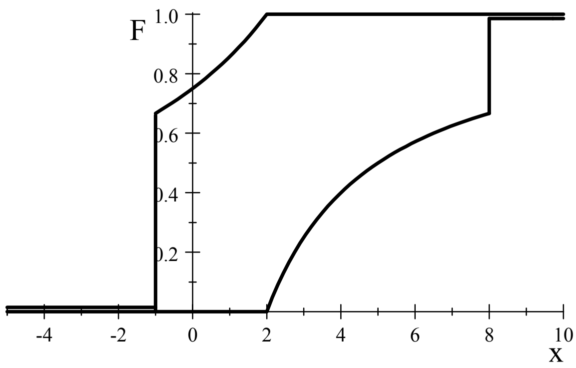

Рис.1. Нижние и верхние ФРВ

$$\overline{F}(x) = \begin{cases} 0, & x < a, \\ \min\left(1, \frac{b-M}{b-x}\right), & a \leq x < b, \\ 1, & x \geq b. \end{cases}$$

Нижняя граница ФРВ $\underline{F}(x)$ может быть получена также при помощи принципа продолжения

$$\underline{F}(x) = \begin{cases} 0, & x < a, \\ \max\left(0, \frac{x-M}{x-a}\right), & a \leq x < b, \\ 1, & x \geq b. \end{cases}$$

Нижняя и верхняя границы ФРВ показаны на Рис.1, где $M = 2$, $a = -1$, $b = 8$. Результирующие границы являются наилучшими в том смысле, что при имеющихся исходных данных невозможно найти более узкий интервал распределений.

Функционал риска и множества ФРВ

Нижняя и верхняя границы дискриминантной функции зависят от параметров α и от граничных значений всех признаков. Поэтому наша цель – найти эту зависимость для того, чтобы записать функционал риска (1). Прежде всего, предположим, что вектор входных переменных $\mathbf{x} = (x_1,...,x_m)$ ограничен, т.е. каждый его элемент удовлетворяет условиям:

$$\underline{x}_k \leq x_k \leq \overline{x}_k,\ k = 1,...,m. \tag{2}$$

Здесь $\underline{x}_k$ и $\overline{x}_k$ – нижняя и верхняя границы для k - го элемента вектора $\mathbf{x}$.

Обозначим M_y средние значения (математические ожидания) функции f , соответствующие каждому классу y. Первые моменты M_{-1} и M_1 тогда могут быть получены как:

$$M_{-1} = \alpha_0 + \sum_{k=1}^{m} \alpha_k v_k,\ M_1 = \alpha_0 + \sum_{k=1}^{m} \alpha_k w_k.$$

Здесь v_k и w_k – средние значения k - го элемента входного вектора из множеств Ψ_- и Ψ_+ соответственно, т.е.

$$v_k = \frac{1}{n_-} \sum_{i\ :\ y_i=-1} x_k^{(i)},\ w_k = \frac{1}{n_+} \sum_{i\ :\ y_i=1} x_k^{(i)}.$$

Для сокращения записи здесь и ниже предполагается, что $v_0 = 1$ и $w_0 = 1$. Из условия (1) следует, что верхняя граница для функционала риска

$$\overline{R}(\alpha) = -P(-1) \cdot \max\left(0, \frac{f - M_{-1}}{f - a}\right) + \\ + P(1) \cdot \min\left(1, \frac{b - M_1}{b - f}\right) - P(-1).$$

Можно заметить из приведенного выражения, что слагаемое $-P(-1)$ в некоторых случаях может быть удалено из выражения, так как оно не зависит от параметров α.

Предположим, что выполняется условие $a < M_{-1} < M_1 < b$. Тогда достаточно просто увидеть, что верхняя граница функционала риска равна:

$$\overline{R}(\alpha) = \\ = \begin{cases} 0, & f < a, \\ P(1)\frac{b-M_1}{b-f} + P(-1), & a \le f < M_{-1}, \\ -P(-1)\frac{f-M_{-1}}{f-a} + P(1)\frac{b-M_1}{b-f} + P(-1), & M_{-1} \le f < M_1, \\ -P(-1)\frac{f-M_{-1}}{f-a} + P(1) + P(-1), & M_1 \le f < b, \\ 0, & f \ge b. \end{cases}$$

Случаи $0 < a$ и $0 \ge b$ не рассматриваются, так как должно выполняться условие $a \le f \le b$. Поэтому, принимая во внимание необходимое условие $f = 0$, получаем:

$$\overline{R}(\alpha) = \\ = \begin{cases} 1 - P(1)M_1 / b, & a \le 0 < M_{-1}, \\ 1 - P(-1)M_{-1} / a - P(1)M_1 / b, & M_{-1} \le 0 < M_1, \\ 1 - P(-1)M_{-1} / a, & M_1 \le 0 < b. \end{cases}$$

Ниже будем использовать функционал $Q(\alpha)$ вместо $\overline{R}(\alpha)$, который необходимо максимизировать по α. Рассмотрим три случая соотношений между M_{-1}, M_1, 0. Каждый случай соответствует своему множеству параметров α.

Случай 1. $M_{-1} \le 0 < M_1$. Для того чтобы найти оптимальные значения параметров α, определим граничные условия, т.е. определим a и b. Заметим, что последнее выражение для верхней границы функционала риска оперирует границами a и b функции f, но не входных переменных. Следовательно, нельзя в явном виде записать эти границы. В то же время, например, нижняя граница определяется из решения задачи линейного программирования

$$a_- = \min_{x_k} \left\{ \alpha_0 + \sum_{k=1}^{m} \alpha_k x_k \right\}$$

при ограничениях (2).

Данная задача имеет m переменных и $2m$ ограничений. Отсюда следует, что m ограничений являются равенствами, т.е. следует искать решение задачи на комбинациях точек $\underline{x}_k$ и $\overline{x}_k$. Тогда граница a удовлетворяет множеству ограничений

$$a \le \alpha_0 + \sum_{k=1}^{m} \alpha_k x_k^*,\ \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\}.$$

Другими словами, имеем 2^m таких неравенств, что одно из них равенство. Так как граница a является отрицательной, то для максимизации $1 - \overline{R}(\alpha)$ необходимо принимать $|a|$ как можно меньше. Отсюда значение a должно быть выбрано как можно большим, но этот рост ограничивается приведенными выше линейными ограничениями.

Аналогичным образом определяется граница b, которая также неизвестна, но она находится из задачи линейного программирования

$$b = \max_{x_k} \left\{ \alpha_0 + \sum_{k=1}^{m} \alpha_k x_k \right\}$$

при ограничениях (2).

Отсюда граница b удовлетворяет множеству ограничений

$$b \ge \alpha_0 + \sum_{k=1}^{m} \alpha_k x_k^*,\ \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\}.$$

Заметим, что граница b является положительной. Поэтому для увеличения $1 - \overline{R}(\alpha)$, значение b необходимо брать как можно меньше, но это уменьшение ограничено приведенными выше линейными ограничениями. Задача оптимизации для вычисления оптимального вектора параметров α принимает вид:

$$Q_1(\alpha) = \max_{\alpha}\left(P(-1)\frac{\sum_{k=0}^{m}\alpha_k v_k}{a} + P(1)\frac{\sum_{k=0}^{m}\alpha_k w_k}{b} \right)$$

при ограничениях

$$a \le \alpha_0 + \sum_{k=1}^{m}\alpha_k x_k^*,\ \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\},$$

$$b \ge \alpha_0 + \sum_{k=1}^{m}\alpha_k x_k^*,\ \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\},$$

$$\sum_{k=0}^{m}\alpha_k v_k \le 0,\ \sum_{k=0}^{m}\alpha_k w_k \ge 0.$$

Здесь два последних неравенства соответствуют условиям $M_{-1} \le 0$ и $M_1 \ge 0$.

Введем новые переменные оптимизации: $q_k = \alpha_k / a$, $k = 0,...,m$, $q = (q_1,...,q_m)$. Более того, введем такой параметр r, что $r = a/b$. Тогда можно переписать задачу оптимизации

$$Q_1(q,r) = \max_{q}\left(P(-1)\sum_{k=0}^{m}q_k v_k + P(1)\,r\sum_{k=0}^{m}q_k w_k \right) =$$

$$= \max_{q}\sum_{k=0}^{m}q_k\left(P(-1)v_k + r\,P(1)w_k\right)$$

при ограничениях

$$\begin{aligned} &1 \ge q_0 + \sum_{k=1}^{m}q_k x_k^*,\ \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\},\\ &1 \ge r q_0 + r\sum_{k=1}^{m}q_k x_k^*,\ \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\},\\ &\sum_{k=0}^{m}q_k v_k \ge 0,\ \sum_{k=0}^{m}q_k w_k \le 0. \end{aligned} \tag{3}$$

Так как r принимает только отрицательные значения ($a \le 0$ и $b \ge 0$), то можно переписать ограничения (3) следующим образом:

$$\frac{1}{r} \le q_0 + \sum_{k=1}^{m}q_k x_k^* \le 1,\ \forall x_k^* \in \{\underline{x}_k, \overline{x}_k\}.$$

Полученная оптимизационная задача является задачей параметрического линейного программирования с параметром $r \le 0$, который есть отношение нижней и верхней границ функции f. Изменяя параметр r и решая задачу линейного программирования для каждого r, найдем оптимальные параметры $q_0,...,q_m$, соответствующие наибольшему значению функционала $Q_1(q,r)$ для всех r. Целевую функцию можно также записать в виде:

$$Q_1(q,r) = P(-1)M_{-1}/a + r\cdot P(1)M_1/a.$$

Поскольку $M_1/a \le 0$, то $Q_1(q,r)$ возрастает с уменьшением параметра r. Следовательно, необходимо параметр r уменьшить. Однако, уменьшая r, увеличиваем $1/r$ и область допустимых решений задачи оптимизации уменьшается. Поэтому существует компромиссная точка r, обеспечивающая максимум целевой функции.

Заметим, что из полученного оптимального решения $q_0,...,q_m$ невозможно найти оптимальные значения параметров $\alpha_0,...,\alpha_m$, так как значение границы не может быть найдено. Однако нет необходимости искать оптимальный вектор α. Поскольку задача классификации является решенной, если найдена такая линейная функция $f(\mathbf{x},\alpha)$, что ее знак определяет принадлежность объекта определенному классу. Но в случае дискриминантной функции ее знак не изменится, если умножить функцию на некоторое положительное число. Отсюда следует, что все линейные функции $f(\mathbf{x},\alpha)$ с параметрами $\alpha_k = q_k a$, $k = 0,...,m$, $a > 0$, могут рассматриваться в качестве дискриминантных функций. Например, если предположить, что $\alpha_0 = 1$, то $a = 1/q_0$, и можно вычислить каждый параметр $\alpha_k = q_k / q_0$.

Случай 2. $a \le 0 < M_{-1}$. Обозначим соответствующий функционал $Q_2(\alpha) = P(1)M_1/b$. Введем новые переменные оптимизации $q_k = \alpha_k / b$, $k = 0,...,m$. Тогда получаем задачу линейного программирования

$$Q_2(q_{\mathrm{opt}}) = P(1)\cdot\max_{q}\sum_{k=0}^{m}q_k w_k$$

при ограничениях

$$1 \ge q_0 + \sum_{k=1}^{m}q_k x_k^*,\ x_k^* \in \{\underline{x}_k, \overline{x}_k\},\ \ \sum_{k=0}^{m}q_k v_k \le 0.$$

Случай 3. $M_1 \le 0 < b$. Обозначим соответствующий функционал $Q_3(q) = P(-1)M_{-1}/a$.

Введем новые переменные оптимизации $q_k = \alpha_k / a$, $k = 0,...,m$. Тогда получаем задачу линейного программирования

$$Q_3(q_{\mathrm{opt}}) = P(-1) \cdot \max_q \sum_{k=0}^{m} q_k v_k$$

при ограничениях

$$1 \geq q_0 + \sum_{k=1}^{m} q_k x_k^*,\ x_k^* \in \{\underline{x}_k, \overline{x}_k\},\ \sum_{k=0}^{m} q_k w_k \leq 0.$$

Окончательно оптимальный вектор q_{opt} определяется из условия

$$q_{\mathrm{opt}} = $$
$$= \arg\max\left\{\max_q \max_r Q_1(q,r), \max_q Q_2(q), \max_q Q_3(q)\right\}.$$

Одним из простейших способов оценки априорных вероятностей $P(y)$ при $y=-1,1$ является непосредственное вычисление с использованием количества точек во множествах Ψ_{-1} и Ψ_{+1}, т.е.

$$P(-1) = n_{-1}/n,\ P(1) = n_{+1}/n = 1 - P(-1).$$

В частности, если нет вообще никакой информации о значениях n_{-1} и n_{+1}, то априорные вероятности классов могут приниматься 0.5.

Оптимистическая стратегия

Нижнюю границу для функционала риска можно записать аналогичным образом

$$\underline{R}(\alpha) = P(-1) - P(-1)\cdot \max\left(0, \frac{M_{-1}}{a}\right) +$$
$$+ P(1)\cdot \min\left(1, \frac{b - M_{+1}}{b}\right)$$

или

$$\underline{R}(\alpha) = \begin{cases} P(1)M_{+1}/a, & a < M_{+1} \leq 0, \\ 0, & M_{-1} \leq 0 < M_{+1}, \\ P(-1)M_{-1}/b, & 0 < M_{-1} \leq b. \end{cases}$$

Можно заметить из приведенного выражения, что нижняя граница функционала риска равна 0, когда $M_{-1} \leq 0 < M_{+1}$. Отсюда следует, что непустое выпуклое множество параметров классификации α, удовлетворяющих этому условию, в рамках оптимистической стратегии обеспечивает множество решений. Действительно, условия для средних значений

$$\sum_{k=0}^{m} \alpha_k v_k \leq 0,\ \sum_{k=0}^{m} \alpha_k w_k \geq 0$$

имеют непустое множество решений α, т.е. можно всегда найти вектор α, удовлетворяющий $M_{-1} \leq 0 < M_{+1}$ для произвольных v_k, w_k, $k = 0,...,m$. Мы не рассматриваем случай, когда все средние значения признаков различных классов совпадают, т.е. $v_k = w_k$, $k = 0,...,m$. Примем $\alpha_0 = 0$. Тогда каждое неравенство образует плоскость, проходящую через начало координат $(0,...,0)$. Отсюда следует, что плоскости пересекаются и отрезают сектор с общими векторами α.

Так как наименьшее значение функционала риска равно 0 и условие $M_{-1} \leq 0 < M_{+1}$ всегда выполняется, можно окончательно записать, что $\underline{R}(\alpha) = 0$ при данной ограниченной информации. Таким образом, оптимистическая стратегия не имеет смысла без дополнительной информации.

Априорные вероятности классов

Одним из простейших способов оценки априорных вероятностей $P(y)$ при $y=-1,1$ является непосредственное вычисление с использованием количества точек во множествах Ψ_{-1} и Ψ_{+1}, т.е.

$$P(-1) = n_{-1}/n,\ P(1) = n_{+1}/n = 1 - P(-1).$$

В частности, если нет вообще никакой информации о значениях n_{-1} и n_{+1}, то априорные вероятности классов могут приниматься как 0.5.

Однако такой путь может привести к некорректным значениям вероятностей, когда объем обучающей выборки n мал. Для принятия более осторожного решения предлагается использовать обобщенную модель Дирихле [14] или в рассматриваемом частном случае двух классов или бинарной классификации предлагается использовать обобщенную модель бета-распределения. Согласно ей, нижняя и верхняя границы вероятности $P(y)$ могут быть записаны в следующем виде:

$$\underline{P}(y) = \frac{n_y}{n+s},\ \overline{P}(y) = \frac{n_y + s}{n+s},\ y = -1,1.$$

Здесь s – параметр распределения Дирихле. Согласно работе [14], параметр s целесообразно брать со значением 1 или 2. Данные выражения выводятся из следующих вероятностей:

$$P(-1)=\frac{\gamma s+n_{-1}}{n+s},\; P(1)=\frac{(1-\gamma)s+n_{+1}}{n+s}$$

путем их минимизации и максимизации по всем значениям параметра $\gamma\in(0,1)$.

Используя выражения для априорных вероятностей, запишем задачу оптимизации для минимаксной стратегии (Случай 1)

$$Q(\alpha_{\text{opt}})=\max_{\alpha}\max_{0\le\gamma\le 1}\Big((\gamma s+n_{-1})\underline{F}(0\mid -1)-$$
$$-\big((1-\gamma)s+n_{+1}\big)\overline{F}(0\mid 1)\Big),$$

при ограничениях $f(\mathbf{x},\alpha)=0$.

Перепишем целевую функцию следующим образом:

$$Q(\alpha)=\gamma s\left(\underline{F}(0\mid -1)+\overline{F}(0\mid 1)\right)+$$
$$+n_{-1}\underline{F}(0\mid -1)-n_{+1}\overline{F}(0\mid 1)-s\overline{F}(0\mid 1).$$

Можно увидеть из приведенного выражения, что максимум функционала $Q(\alpha)$ по множеству значений γ достигается в точке $\gamma=1$ и равен:

$$Q(\alpha_{\text{opt}})=\max_{\alpha}\left((s+n_{-1})\underline{F}(0\mid -1)-n_{+1}\overline{F}(0\mid 1)\right).$$

В итоге задача линейного программирования для вычисления оптимальных параметров α_{opt} имеет целевую функцию

$$Q(\alpha,r)=\sum_{k=0}^{m}q_k\left((s+n_{-1})v_k+rn_{+1}w_k\right).$$

Аналогичные выражения можно получить для случаев 2 и 3.

В предлагаемой модели классификации используется двойная размытость. Первая заключается в замене одного определенного типа распределений вероятностей множеством распределений $\mathcal{F}$, вторая – в использовании обобщенной модели Дирихле для размытия априорных вероятностей классов.

Числовой пример

Рассмотрим пример для иллюстрации предлагаемого подхода. Предположим, что необходимо классифицировать дефектные бруски из древесины. Имеются три признака, характеризующих брак древесины: ширина x_1 и длина x_2 трещины, цвет x_3. Ширина и длина трещины ограничены размерами бруска. В частности, ширина находится в интервале от 0 до 0.01 м. ($\underline{x}_1=0$, $\overline{x}_1=0.01$), длина − в интервале от 0 до 0.5 м. ($\underline{x}_2=0$, $\overline{x}_2=0.5$). Цвет измеряется по шкале целых чисел и находится в пределах от 16 до 64 ($\underline{x}_3=16$, $\overline{x}_3=64$). Будем предполагать, что априорные вероятности классов равны 0.5. Эксперты дали оценки средних значений признаков для бракованной древесины $(y=1)$ и для качественной древесины $(y=-1)$:

$$v_1=0.002,\; v_2=0.1,\; v_3=38,$$
$$w_1=0.005,\; w_2=0.3,\; w_3=54.$$

Наибольшее значение 0.834 целевой функции $Q_1(q,r)$ достигается при $r=-0.7$ в точке $q_0=1$, $q_1=0$, $q_2=-4.857$, $q_3=0$ ($q_{\text{opt}}=(1,0,-4.857,0)$).

Во втором случае, когда $a\le 0<M_{-1}$, задача имеет оптимальное решение: $q_0=-19/13$, $q_1=0$, $q_2=0$, $q_3=1/26$. Отсюда $Q_2(q_{\text{opt}})=0.615$.

В третьем случае, когда $M_1\le 0<b$, задача имеет оптимальное решение $q_0=1$, $q_1=0$, $q_2=-10/3$, $q_3=0$. Отсюда $Q_3(q_{\text{opt}})=0.666$.

В результате получена дискриминантная функция $f(\mathbf{x},\alpha)=1-4.857x_2$, так как $Q_1(q_{\text{opt}},0.7)\ge Q_3(q_{\text{opt}})\ge Q_2(q_{\text{opt}})$. Из полученных результатов следует, что второй признак дает наибольшее значение функционала риска. Если удалить всю информацию, касающуюся признаков x_1 и x_3, то получим ту же самую дискриминантную функцию. Это свойство рассмотренного алгоритма определяет то, что второй признак обеспечивает наибольшую возможную ошибку классификации и затем ищет такую дискриминантную функцию, которая минимизирует эту ошибку.

Заключение

В работе была рассмотрена задача классификации при экстремально ограниченной информации в форме условных математических ожиданий признаков. Ее решение основано на пессимистической (минимаксной) и оптимистической (миниминной) стратегиях принятия решений. Однако осторожная стратегия, как линейная комбинация пессимистической и оптимистической стратегий с некоторым заранее определенным коэффициентом осторожности, также может быть изучена предлагаемым методом [5].

Очевидно, что рассмотренные стратегии принятия решений являются слишком «экстремальными» с практической точки зрения и их использование в реальных задачах может привести к некорректным результатам. В связи с этим в работе не приводится детальное тестирование предлагаемых моделей на реальных объектах и реальных статистических данных. Основная цель работы заключалась в демонстрации и детализации общей идеи того, как можно работать с неполной информацией. Реализация этой идеи для решения конкретных прикладных задач с учетом их специфики и имеющейся дополнительной информации может привести к более «взвешенному» решению.

Предложенный метод достаточно просто обобщается на случай интервальных данных. Нижняя и верхняя границы ФРВ соответствующих множеств распределений при интервальных средних значениях признаков полностью определяются границами интервалов. Еще одним направлением дальнейших исследований могут являться комбинированные методы классификации, когда некоторые признаки описываются достаточной статистической информацией, а часть - их средними значениями. Эти методы предполагают применение унифицированного представления различных типов исходной информации и его использование в стандартных известных методах классификации.

В работе были рассмотрены только линейные дискриминантные функции. Однако, заменяя признаки их некоторыми функциями, можно решать аналогичным образом и задачи с нелинейными дискриминантными функциями. Интересным направлением дальнейших исследований, развивающим построение нелинейных моделей классификации, является модификация метода опорных векторов [3, 9] с учетом неполной исходной информации. Эта задача решается добавлением к полученным в работе функционалам риска специального члена регуляризации [4]. Такая модификация для интервальных данных с использованием минимаксной стратегии была предложена в работе [11].

Литература

1. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. М.: Наука, 1974. – 416 с.
2. Кузнецов, В.П. Интервальные статистические модели. М.: Радио и связь. 1991.
3. Мерков А.Б. Распознавание образов: Введение в методы статистического обучения. М.: Едиториал УРСС, 2011. – 256 с.
4. Тихонов А.Н., Арсенин В.Я. Методы решения некорректных задач. М.: Наука, 1974. – 223 с.
5. Уткин Л.В. Анализ риска и принятие решений при неполной информации. СПб.: Наука, 2007. 404 с.
6. Alaiz-Rodrguez R., Guerrero-Curieses A., Cid-Sueiro J. Minimax regret classifier for imprecise class distributions // J. Mach. Learn. Res., V.8, P.103-130, 2007.
7. Lanckriet G.R., Ghaoui L.E., Bhattacharyya C., Jordan M.I. A robust minimax approach to classification // The Journal of Machine Learning Research, V.3, P. 555-582, 2003.
8. Robert C. The Bayesian Choice. Springer, New York. 1994.
9. Scholkopf B., Smola A. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, 2002.
10. Utkin L.V. Regression analysis using the imprecise Bayesian normal model // International Journal of Data Analysis Techniques and Strategies, 2:356–372. 2010.
11. Utkin L.V., Coolen F.P.W. Interval-valued regression and classification models in the framework of machine learning // Proc. of the Seventh Int. Symposium on Imprecise Probabilities: Theories and Applications, ISIPTA'11, SIPTA, 2011, P. 371-380.
12. Utkin L.V. Coolen F. On reliability growth models using Kolmogorov-Smirnov bounds // International Journal of Performability Engineering, V. 7, P. 5–19, 2011.
13. Utkin L.V., Zhuk Y.A. A machine learning algorithm for classification under extremely scarce information // International Journal of Data Analysis Techniques and Strategies, 2012. (в печати).
14. Walley P. Inferences from multinomial data: Learning about a bag of marbles // Journal of the Royal Statistical Society, Series B, V. 58, P. 3–57, 1996.
15. Walley P. Statistical Reasoning with Imprecise Probabilities. Chapman and Hall, London. 1991.

Уткин Лев Владимирович. Проректор по научной работе и заведующий кафедрой управления и автоматизации производственных процессов Санкт-Петербургского государственного лесотехнического университета. Окончил Ленинградский электротехнический институт им. В.И. Ульянова (Ленина) в 1986 году. Доктор технических наук, профессор. Автор 230 печатных работ и шести монографий. Область научных интересов: теория неточных вероятностей (imprecise probability theory), интервальные статистические модели, теория принятия решений, теория надежности, машинное обучение. E-mail: lev.utkin@gmail.com

Жук Юлия Александровна. Доцент кафедры информационных систем и технологий Санкт-Петербургского государственного лесотехнического университета, Окончила Белорусский государственный университет в 1997 году. Кандидат педагогических наук. Автор 26 печатных работ. Область научных интересов: интеллектуальные технологии, мультимедиатехнологии, психология восприятия видеоинформации, биоинформатика. E-mail: zhuk_yua@mail.ru

Селиховкин Иван Андреевич. Аспирант кафедры управления и автоматизации производственных процессов Санкт-Петербургского государственного лесотехнического университета. Окончил. Санкт-Петербургскую государственную педиатрическую медицинскую академию в 2004 году. Автор 2 печатных работ. Область научных интересов: медицинские информационные системы, интеллектуальные технологии, машинное обучение. E-mail: ivan.selihovkin@gmail.com