УДК 519.7.007

ПРОСТОЙ АЛГОРИТМ НОМИНАЛЬНОЙ КЛАССИФИКАЦИИ ПО КАЧЕСТВЕННЫМ ПРИЗНАКАМ

Е.К. Корноушенко

Предложен альтернативный подход к классификации по качественным признакам, отличающийся от известных подходов тем, что вместо сравнения кортежа значений признаков тестового объекта с аналогичными кортежами значений признаков объектов обучающей выборки производится независимое попарное сравнение каждой пары значений соответствующих кортежей признаков сравниваемых объектов. Это позволяет сформировать матрицу весов признаков для каждого тестового объекта, более детальную, чем ближайшая окрестность тестового объекта. В рамках данного подхода предложен простой алгоритм классификации, обладающий рядом важных особенностей в плане интерпретации результатов классификации. На примере несбалансированной выборки из известного репозитория UCI проверено качество алгоритма. Показано, что алгоритм обеспечивает хорошую точность классификации объектов «малых» классов.

Ключевые слова: классификация, ближайшая окрестность, метка класса, взвешенное голосование, взвешивание признаков, матрица весов признаков.

ВВЕДЕНИЕ

Классификация — одно из направлений в обширной области интеллектуального анализа данных (*data mining*). Различают три вида классификации:

— классификация «с учителем» (*supervised classification*), когда исходные данные, отнесенные к разным категориям, называемым классами, разбиваются на две выборки — на обучающей выборке производится обучение (настройка) применяемого алгоритма, а на контрольной — собственно классификация и оценка его качества;

— классификация «без учителя» (*unsupervised classification*), когда на исходных данных формируются классы с помощью различных процедур группировки данных (как правило, процедур кластеризации);

— гибридная классификация, когда сгруппированные данные разбиваются на обучающую и контрольную выборки.

Систематизация подходов, применяемых в классификации «с учителем», требует отдельного рассмотрения, укажем лишь основные из известных подходов. Это — метрический подход, базирующийся на использовании той или иной метрики при сравнении классифицируемого объекта с объектами обучающей выборки и введении понятия «ближайшей окрестности» тестового объекта (см. например, работы [1—6]), теоретико-информационный подход, базирующийся на изменении энтропии вероятностных распределений при учете тех или иных признаков классифицируемых объектов (к нему относятся, например, алгоритмы с использованием деревьев решений [7—10], алгоритмы ассоциативной классификации [11—13]), подход с формированием недоопределенных множеств (*rough sets*) [14—17]), а также многокритериальные методы принятия решений (см. например, работы [18—20]).

В настоящей работе рассматривается альтернативный подход к классификации с использованием понятия «ближайшей окрестности» тестового объекта. Суть предлагаемого подхода состоит в том, что вместо обычного сравнения в целом кортежа значений признаков тестового объекта с кортежами значений признаков объектов обучающей

выборки рассматривается попарное независимое сравнение соответствующих значений признаков из сравниваемых кортежей. В работе рассматривается случай номинальной классификации по качественным признакам, в этом случае считаем, что значения качественных признаков суть их наименования[1] [21]. При этом если качественный признак является ординальным, каждое из его значений будет рассматриваться как номинальный признак. Для каждого номинального признака тестового объекта формируется множество из объектов обучающей выборки с тем же значением (именем) сравниваемого признака, и каждому из таких объектов приписывается определенный вес. После аналогичного рассмотрения значений всех признаков формируется матрица весов тестового объекта (а не вектор весов, как в известных алгоритмах классификации). Показано, что в результате предложенных процедур алгоритм классификации обладает рядом свойств, выгодно отличающих его от известных алгоритмов данного направления.

Применение предлагаемого алгоритма показано на примере выборки из известного репозитория UCI (*UCI Machine Learning Repository*) [22], описывающей результаты психологического эксперимента и характеризующейся тем, что классы наблюдений сильно различаются по мощности, т. е. образуют несбалансированную выборку. В рамках предлагаемого алгоритма классификации предложены три варианта взвешивания объектов обучающей выборки и показано, как выбор того или другого варианта влияет на качество классификации объектов тестовой выборки.

Данная статья построена следующим образом. В рамках подхода, базирующегося на понятии «ближайшей окрестности», в § 1 приводятся краткие сведения по эволюции процедур этого подхода (переход от невзвешенного голосования к тому или иному виду взвешенного голосования, к взвешиванию признаков объектов, входящих в ближайшую окрестность тестового объекта, использование различных метрик при сравнении объектов). В § 2 описываются этапы предлагаемого алгоритма номинальной классификации по качественным признакам. В § 3 показаны особенности и отличия процедур, применяемых на том или ином этапе алгоритма, от аналогичных процедур, упомянутых в § 1. В § 4 показано применение предлагаемого алгоритма к конкретной выборке и обсуждаются полученные результаты.

[1] Хотя в данном случае понятие «признак» и «значение признака» совпадают, будем пользоваться понятием значения признака как более «привычным».

1. КЛАССИФИКАЦИЯ НА ОСНОВЕ ПОНЯТИЯ «БЛИЖАЙШЕЙ ОКРЕСТНОСТИ»

1.1. Исходные данные

Здесь исходной информацией для построения алгоритмов классификации служит конечная выборка объектов. Для наглядности изложения в качестве объектов будем рассматривать физические объекты той или иной природы (объекты недвижимости, продукты, лекарства и др.) и считать, что каждый объект исходной выборки характеризуется некоторой совокупностью признаков из заданного набора количественных и/или качественных признаков. Количественные признаки могут принимать любые действительные значения внутри соответствующих интервалов. А качественные признаки подразделяются на номинальные, ординальные, интервальные и признаки отношений (*ratio features*). Каждый из видов качественных признака характеризуется соответствующей шкалой, отображающей значения этого признака и их взаимное расположение. Самой «слабой» шкалой является шкала наименований [21] для номинальных признаков, когда каждый используемый в выборке номинальный признак кодируется каким-либо образом (числом, символом, словом и т. п.). Всякий ординальный признак имеет несколько значений, и на шкале значений ординального признака определено отношение порядка. Другие виды качественных признаков здесь не рассматриваются. Признаки называются независимыми, если изменения значений одного из признаков не влияют на значения других признаков, в противном случае признаки зависимы.

Принципиальный момент в классификации «с учителем» — введение на исходной выборке разбиения объектов на классы. Понятие класса нельзя определить априори до окончания формирования исходной выборки. Вид вводимого разбиения зависит от пользователя: разбиение может состоять из двух блоков, когда пользователь руководствуется двумя оценками помещаемых в выборку объектов (объект соответствует выбранному критерию или нет), либо из нескольких блоков при использовании нескольких оценок. В итоге каждому объекту исходной выборки однозначно приписывается номер (*метка, label*) того класса, к которому принадлежит данный объект. Далее исходная выборка разбивается на обучающую и тестовую выборки (соответственно ОВ и ТВ). Каждый класс может содержать объекты из ОВ и ТВ. Если некоторый класс содержит объекты из ОВ и ТВ, то метки таких объектов совпадают, причем множества объектов из ОВ и ТВ с одной и той же меткой могут существенно различаться по мощ-

ности. Метки объектов ТВ не участвуют в процедурах классификации, а служат лишь для анализа качества алгоритма классификации, в частности, для проверки его точности и наполняемости классов в ТВ (об этом см. ниже). Суть классификации состоит в том, чтобы «вложить» каждый объект ТВ в некоторый класс введенного разбиения таким образом, при котором номер этого класса как результат процедуры классификации совпадает с исходной меткой этого объекта в ТВ, и оценить точность классификации.

1.2. Процедура сравнивания объектов исходной выборки

Введем понятие кортежа признаков. Поскольку в набор выбранных признаков для классификации объектов исходной выборки могут входить качественные признаки (с закодированными значениями, см. выше), совокупность значений признаков (количественных и качественных), характеризующих конкретный объект, будем называть *кортежом* признаков этого объекта. Для сравнения кортежей признаков двух объектов (обязательный этап, присутствующий в алгоритмах данного направления) необходима их предварительная нормализация, т. е. отображение значений соответствующих признаков в интервал [0, 1]. При этом если a_i — значение количественного признака a, то его нормализованное значение можно определить, в частности, по формуле минимаксной нормализации: $a_{i\text{норм}} = \dfrac{a_i - \min(a_i)}{\max(a_i) - \min(a_i)}$. При сравнении качественных признаков с закодированными значениями учитывается лишь совпадение (или несовпадение) кодов этих значений[2]. При совпадении кодов разности этих значений приписывается 0, а при несовпадении — 1. В качестве мер отличия двух кортежей чаще применяются евклидово расстояние и расстояние Махалонобиса (здесь не рассматривается).

1.3. Суть *kNN*-алгоритма и его модификации

Идея использования понятия «ближайшей окрестности» довольно проста (см., например, работу [1]). Кортеж признаков тестового объекта Z сравнивается с кортежем признаков каждого объекта ОВ и выделяется заданное число (скажем, k) ближайших (по расстоянию) к Z объектов из ОВ, которые образуют *ближайшую окрестность* $NN_k(Z)$ (*Nearest Neighborhood*) объекта Z. Совокупность меток объектов из окрестности $NN_k(Z)$ обозначим как $L_k(Z)$. Класс, к которому будет отнесен объект Z, выбирается путем невзвешенного голосования, т. е. из множества $L_k(Z)$ выбирается метка с наибольшим числом вхождений в $L_k(Z)$. На такой простой идее базируется известный *kNN*-алгоритм [1, 2] и его более сложные модификации, использующие тот или иной алгоритм взвешиваний как меток, так и признаков объектов [3—6].

1.3.1. Взвешенное голосование [3—5, 23, 24]

Для повышения точности *kNN*-алгоритма были предложены различные методы взвешивания меток из множества $L_k(Z)$, причем веса меток определялись тем или иным образом через расстояние (как правило, евклидово, обозначаемое далее как d_i^{NN}) от объекта x_i^{NN} из окрестности $NN_k(Z)$ до тестового объекта Z. Одной из первых в этом направлении была работа [3]: расстояния d_i^{NN} упорядочивались в возрастающем порядке и вес i-го объекта из окрестности $NN_k(Z)$ определялся как $w_i = \dfrac{d_k^{NN} - d_i^{NN}}{d_k^{NN} - d_1^{NN}}$, $i = 1, \ldots, k$, причем $w_i = 1$ при $d_k^{NN} = d_1^{NN}$. Веса, относящиеся к одной и той же метке, суммировались, и в качестве искомой метки выбиралась метка с наибольшим суммарным весом. В случае $k = 1$ этот метод совпадал с невзвешенным голосованием. В работах [4, 5] предложена дуальная структура веса w_i:

$$w_i = \frac{d_k^{NN} - d_i^{NN}}{d_k^{NN} - d_1^{NN}} \cdot \frac{1}{i}, \quad i = 1, \ldots, k,$$

и говорится, что при этом распределение совокупностей меток во множестве $L_k(Z)$ становится более «контрастным» и менее зависящим от значений признаков объекта Z. Более сложные процедуры взвешенного голосования, требующие отдельных комментариев, предложены в работах [23, 24].

1.3.2. Взвешивание признаков объектов [6, 25—28]

При более глубоком анализе свойств «ближайшей окрестности» выяснились некоторые обстоятельства (влияющие на качество *kNN*-алгоритма и, прежде всего, на его чувствительность к помехам), обусловленные тем фактом, что при классифика-

[2] Поскольку в данной работе вопросы, связанные с упорядочением значений ординального признака, не рассматриваются, каждое из его значений интерпретируется как номинальное значение.

ции все признаки рассматривались как равноправные, хотя по структуре той или иной задачи одни признаки могут быть более значимыми[3] для правильной классификации, нежели другие. Таким образом, помимо проблемы взвешивания меток классов возникает проблема взвешивания признаков, когда более важным признакам присваивается больший вес. Исследование проблемы взвешивания признаков началось с работы [6] и продолжается в многочисленных работах (см., например, работы [25—28]). В рамках классификации «с учителем» и использования понятия «ближайшей окрестности» входными переменными служат признаки объектов ОВ и тестового объекта. Из совокупности методов взвешивания признаков следует выделить семейство алгоритмов *Relief* [25—27] как наиболее проработанных и применяемых при построении регрессионных моделей и в задачах классификации. В основе алгоритмов семейства *Relief* лежит неявное предположение о том, что дисперсия значений признаков, принадлежащих объектам из одного и того же класса, должна быть меньше аналогичной дисперсии значений признаков объектов из разных классов, причем такая разница в значениях дисперсии должна быть большей для более значимых признаков [25]. На каждой итерации для выбранного объекта x определяются две ближайших окрестности: одна (обозначаемая как $NH(x)$) из объектов того же класса, что и класс объекта x, а другая — $NM^{(i)}(x)$ — из объектов, принадлежащих классам, отличным от класса объекта x. Пусть $x^{(i)}$ — значение i-го признака в кортеже признаков объекта x, а $NH^{(i)}(x)$ и $NM^{(i)}(x)$ — значение i-го признака в объектах из совокупностей $NH(x)$ и $NM(x)$. Вес w_i i-го признака объекта x вычисляется по правилу [25]:

$$w_i' = w_i + |x^{(i)} - NM^{(i)}(x)| - |x^{(i)} - NH^{(i)}(x)|.$$

Таким образом, наличие объектов в окрестности $NH(x)$ приводит к уменьшению веса w_i i-го признака, тогда как наличие в окрестности $NM(x)$ объектов приводит к увеличению веса w_i i-го признака. При этом расстояние между двумя объектами определяется как сумма весов соответствующих признаков (манхэттенское расстояние). Первоначально алгоритм *Relief* был разработан для случая бинарной классификации. Обобщение этого подхода на случай большего числа классов (*ReliefF*) дано в работах [25—27]. Суть алгоритма *ReliefF* [26] из семейства *Relief* состоит в итеративной оценке весов признаков с тем, чтобы усилить различия в свойствах ближайших окрестностей каждого из объектов ОВ.

2. ПРОСТОЙ АЛГОРИТМ НОМИНАЛЬНОЙ КЛАССИФИКАЦИИ ПО КАЧЕСТВЕННЫМ ПРИЗНАКАМ

Исходная выборка — такая же, как в п. 1.1, с учетом того, что длина ОВ равна S, а длина ТВ равна T и каждый объект исходной выборки имеет m качественных признаков (номинальных или ординальных). На множестве объектов исходной выборки определено разбиение $\pi = (C_1, ..., C_K)$ на классы, и каждый объект ОВ и ТВ принадлежит к какому-либо одному из K классов разбиения π. Обозначим через $\pi_{ОВ}$ ($\pi_{ТВ}$) подразбиение разбиения π, блоки которого содержат только объекты из ОВ (ТВ). Метки объектов ТВ не участвуют в процедурах классификации, а служат лишь для анализа качества применяемого алгоритма классификации, в частности, для проверки его точности и заполняемости классов подразбиения $\pi_{ТВ}$ (об этом см. ниже) в процессе классификации объектов ТВ.

Для корректной работы предлагаемого ниже алгоритма номинальной классификации (далее просто Алгоритма) вводятся следующие ограничения.

1. Поскольку все признаки объектов ОВ и ТВ в Алгоритме рассматриваются как номинальные, код каждого признака в объектах ТВ должен совпадать с кодом этого признака, присутствующего в кортежах некоторых объектов ОВ.

2. Никакой класс разбиения π не может целиком содержаться в ТВ.

3. В объектах ОВ н ТВ не допустимы неопределенные значения признаков и меток классов. Факт отсутствия данного признака[4] в рассматриваемом объекте обозначается символом 0.

Основная особенность Алгоритма в том, что вместо сравнения кортежа признаков тестового объекта и объектов из ОВ попарно сравниваются коды каждого из признаков тестового объекта с кодами соответствующих признаков объектов из ОВ. Результатом сравнения двух кодов может быть лишь их совпадение (согласно ограничению 1) или несовпадение (при отсутствии данного (значения) признака в кортеже признаков сравниваемого объекта). Это позволяет для каждого признака A тестового объекта Z определить совокупность $[Z]_A$ из объектов ОВ, кортежи которых содержат признак A

[3] Значимость признака определяется в процессе его дальнейшего использования при построении регрессионных моделей и/или алгоритмов классификации.

[4] Данное замечание касается лишь ординальных признаков, имеющих несколько значений, интерпретируемых как независимые номинальные подпризнаки (со своими подкодами) кода исходного признака.

Таблица 1

Общий вид матрицы весов M_Z для тестового объекта Z

	A_1	A_2	...	A_m
C_1	g_{11}	g_{21}	...	g_{m1}
C_2	g_{12}	g_{22}	...	g_{m2}
...	...	...	...	...
C_K	g_{1K}	g_{2K}	...	g_{mK}

(а при ординальном A — код его соответствующего значения). Пусть $M_{A\nu}$ — число объектов из совокупности $[Z]_A$ таких, что эти объекты входят в блок C_ν разбиения $\pi_{\text{ОВ}}$ (т. е. имеют метку ν). Сопоставим совокупности этих объектов число $g_{A0} = M_{A\nu}/|[Z]_A||C_\nu|$, где $|[Z]_A|$ и $|C_\nu|$ — соответственно мощности множеств $[Z]_A$ и класса C_ν разбиения $\pi_{\text{ОВ}}$, $1 \le \nu \le K$. Отношение $M_{A\nu}/|[Z]_A|$ есть, по существу, показатель «доверия» (*confidence*[5]) к утверждению, что всякий объект из ОВ с признаком A относится к классу[6] C_ν. Обратим внимание на то, что показатели доверия взвешиваются с помощью множителей $1/|C_\nu|$, где ν — метка класса, к которому относится тот или иной показатель «доверия»[7]. Найденное число $g_{A\nu}$ будем рассматривать как вес метки ν в множестве меток объектов из множества $[Z]_A$. Подобным образом найдем множество $M_{A\mu}$ для объектов из множества $[Z]_A$, принадлежащих другим классам C_μ, $\mu \ne \nu$, из разбиения $\pi_{\text{ОВ}}$, и вычислим вес $g_{A\mu}$ каждой метки μ. Сформируем K-вектор весов $G_A = (g_{A1}, g_{A2}, ..., g_{AK})^T$ для признака A объекта Z. Заметим, что все координаты в векторе G_A — не отрицательные и не больше единицы. Вектор весов G_A можно интерпретировать как локальный (по признаку A) классификатор для объекта Z.

Аналогичным образом найдем столбцы весов для всех m признаков объекта Z и сгруппируем эти столбцы в матрицу M_Z размера $K \times m$, которую назовем *матрицей весов* объекта Z. Структура матрицы M_Z показана в табл. 1, где $A_1, A_2, ..., A_m$ — признаки объектов, а $C_1, ..., C_K$ — блоки разбиения $\pi_{\text{ОВ}}$.

Номер строки матрицы M_Z с наибольшей суммой весов будем считать меткой класса, к которому отнесем объект Z. На основе матрицы весов M_Z можно получить три варианта голосования в зависимости от определения элементов g_{ij}, $i = 1, ..., m$, $j = 1, ..., K$.

Вариант 1. Если $g_{A\nu} = M_{A\nu}$ — обычное невзвешенное голосование.

Вариант 2. Если $g_{A\nu} = M_{A\nu}/|[Z]_A|$ — голосование со взвешиванием показателей «доверия».

Вариант 3. Если $g_{A\nu} = M_{A\nu}/|[Z]_A||C_\nu|$ — голосование со взвешиванием меток из множества $M_{A\nu}$, скорректированных с учетом показателей «доверия» и мощностей блоков разбиения $\pi_{\text{ОВ}}$. Данный вариант используется с применением к строкам матрицы M_Z преобразования (1) (см. ниже).

Рассмотрим матрицу M_Z как совокупность локальных (по признакам объекта Z) классификаторов. С учетом сказанного в работе [31] относительно усиления классифицирующей способности совокупности локальных классификаторов путем введения монотонного нелинейного отображения преобразуем строки $M_{Z\mu}$ матрицы M_Z с помощью монотонной нелинейной функции

$$E(M_{Z\mu}) = \Big(\sum_{A \in A_Z} g_{A\mu} \log_2(g_{A\mu}) \big| g_{A\mu} > 0 \Big), \quad \mu = 1, ..., K, \qquad (1)$$

где A_Z — кортеж признаков объекта Z.

Поскольку значения $g_{A\mu}$ не являются в общем случае вероятностными величинами, выражение (1) назовем в соответствии с работой [31] *квазиэнтропией* строки $M_{Z\mu}$ матрицы M_Z. Сформируем вектор $H_Z = (E(M_{Z1}), ..., E(M_{ZK}))$, который назовем *классифицирующим вектором* для объекта Z. Номер координаты вектора H_Z с наибольшим значением принимается как искомая метка класса разбиения π, приписываемая тестовому объекту Z. Аналогичным образом производится классификация остальных объектов ТВ (а также классификация объектов без меток, удовлетворяющих ограничениям 1—3).

Считается, что тестовый объект Z правильно классифицирован, если приписанная ему метка класса разбиения π совпадает с исходной меткой этого объекта в разбиении $\pi_{\text{ТВ}}$. Точность всякого алгоритма классификации определяется как отно-

[5] Понятие *confidence* употребляется во многих алгоритмах классификации (см., например, работы [29, 30]), где подчеркивается важность использования данного понятия для повышения точности классификации.

[6] Другими словами, это означает степень «доверия» тому, что класс C_ν помимо некоторых объектов из ОВ, входящих в множество $[Z]_A$, содержит и тестовый объект из ТВ.

[7] Как показывают многочисленные эксперименты, подобное взвешивание показателей «доверия» способствует лучшему наполнению блоков разбиения при классификации Алгоритма (см. далее Пример, § 4).

шение числа правильно классифицированных объектов к длине ТВ.

Однако точность служит лишь одним из качеств алгоритма классификации. При наличии классов с сильно различающейся мощностью («несбалансированная» выборка) основную «нагрузку» при классификации может брать на себя класс с наибольшей мощностью, и высокий процент правильной классификации может не означать хорошей классификации объектов из классов с небольшой мощностью, хотя во многих практических задачах именно такие классы представляют особый интерес. Проблема классификации в случае несбалансированных выборок детально рассмотрена в работе [32]. В ряде публикаций приводятся оценки степени несбалансированности выборки. В случае нескольких классов известны попытки учесть подобную ситуацию путем введения той или иной меры «наполнения» всех классов при классификации объектов ТВ (см., например, работу [33]). В этом плане практически целесообразным может быть использование так называемого *коэффициента Криппендорфа* (см., например, работу [34]). В данном Алгоритме предлагается использовать меру $\mu(\text{ТВ})$ наполняемости классов подразбиения $\pi_{\text{ТВ}}$ при классификации объектов ТВ:

$$\mu(\text{ТВ}) = \frac{1}{T}\sum_{\nu=1}^{K}\frac{\left|\hat{C}_\nu(\text{ТВ})\right|}{\left|C_\nu(\text{ТВ})\right|},$$

где T — число объектов ТВ, $|C_\nu(\text{ТВ})|$ — число объектов ТВ в классе C_ν разбиения $\pi_{\text{ТВ}}$, $|\hat{C}_\nu(\text{ТВ})|$ — число правильно классифицированных объектов ТВ в классе C_ν. Значения показателя $\mu(\text{ТВ})$ определены в интервале [0, 1]. Использование меры $\mu(\text{ТВ})$ для оценки качества классификации конкретной выборки показано на примере в § 4. Детальный анализ связи меры $\mu(\text{ТВ})$ со структурой исходной выборки и особенностями алгоритма классификации выходит за рамки настоящей работы.

3. ОСОБЕННОСТИ АЛГОРИТМА КЛАССИФИКАЦИИ

• Поскольку вместо сравнения в целом кортежа признаков тестового объекта Z с кортежами признаков объектов ОВ каждый из номинальных признаков сравниваемых объектов рассматривается по отдельности и независимо, то появляется возможность обойтись без нормализации кортежей сравниваемых объектов. Нормализованная разность двух кортежей с номинальными признаками является двоичным вектором (с координатами (0, 1)), и евклидова норма разности инвариантна к перестановкам нулей и единиц в этой разности. Это означает, что в «ближайшую окрестность» объекта Z попадут объекты, дающие при сравнении с объектом Z не изменяющееся число нулей и единиц в соответствующих разностях. Число таких объектов растет с увеличением длины ОВ и ухудшает точность и качество применяемого варианта *kNN*-алгоритма.

• В матрице весов M_Z каждого объекта Z из ТВ подробно раскрыта структура множества весов, приписываемых объектам из множества, по признакам и по классам. Используемые в Алгоритме процедуры взвешивания признаков принципиально отличаются от аналогичных процедур, описанных в пп. 1.3.1 и 1.3.2. Информация, содержащаяся в матрице M_Z, намного «богаче» той, которая содержится в «ближайшей окрестности» каждого объекта Z из ТВ в *kNN*-алгоритмах. Матрицы M_Z позволяют исследовать влияние того или иного признака на процесс классификации объектов ТВ. Подобная информация может оказаться весьма полезной при анализе влияния признаков на процесс классификации в практических задачах. А тот факт, что в предложенном алгоритме признаки рассматриваются и сравниваются по отдельности, открывает возможности для выбора самых «экзотических» свойств для определения признаков при классификации объектов.

• Вычислительную сложность Алгоритма можно оценить как $O(mST)$. Заметим, что структура Алгоритма содержит лишь циклы, на которых определены соответствующие арифметические операции, управляемые условными операторами. Это позволяет реализовать Алгоритм в формате Excel, без привлечения более сложных компьютерных средств, что делает Алгоритм доступным для широкого круга пользователей.

4. ПРИМЕР: НОМИНАЛЬНАЯ КЛАССИФИКАЦИЯ ПРИ НЕСБАЛАНСИРОВАННОЙ ВЫБОРКЕ

Для иллюстрации качества Алгоритма была взята выборка с названием «Balanced Scale» из репозитория UCI [22]. Выборка, содержащая результаты психологического эксперимента, включает в себя 625 наблюдений, каждое из которых содержит четыре качественных признака с пятью номинальными значениями (1—5). Все наблюдения разбиты на три класса, два из которых содержат по 288 элементов, а третий — 49 элементов, так что выборка несбалансированная. Перед применением к этой выборке предлагаемого Алгоритма из нее выделялись ОВ и ТВ путем попеременного отнесения текущих наблюдений то к ОВ, то к ТВ. Таким обра-

Таблица 2

Качество Алгоритма при разных вариантах определения элементов g_{ij} матрицы M_Z

Варианты определения элементов матрицы M_Z	Наполненность блоков разбиения $\pi_{ТВ}$	μ(ТВ)	Точность классификации, %
Исходное разбиение $\pi_{ТВ}$	$\pi_{ТВ}$ = (4, 154, 154)	—	—
Вариант 1: $g_{Av} = M_{Av}$	$\hat{\pi}_{ТВ}$ = (0, 154, 154)	0,6666	98,7
Вариант 2: $g_{Av} = M_{Av}/\|[Z]_A\|$			
Вариант 3: $g_{Av} = M_{Av}/\|[Z]_A\|C_v\|$ и выполнено преобразование (1)	$\hat{\pi}_{ТВ}$ = (4, 128, 128)	0,8874	83,3

зом, ОВ и ТВ содержали по 312 наблюдений (последнее наблюдение не использовалось), причем в разбиении $\pi_{ОВ}$ класс C_1 имеет 44 элемента, а классы C_2 и C_3 — по 134 элемента. На ТВ в разбиении $\pi_{ТВ}$ класс C_1 имеет 4 элемента, а классы C_2 и C_3 — по 154 элемента, так что ТВ является существенно несбалансированной выборкой.

При применении к этой выборке Алгоритма были рассмотрены приведенные выше три варианта определения элементов g_{ij} матрицы весов M_Z. Качество Алгоритма для каждого из этих вариантов показано в табл. 2. В ней наглядно видно, как при невзвешенном голосовании заполняются «большие» классы разбиения $\pi_{ТВ}$, обеспечивая высокую точность классификации, и абсолютно «игнорируется» «малый» класс C_1. Применение отображения (1) к строкам матрицы M_Z снижает несколько точность классификации из-за полного наполнения «малого» класса C_1. Видно, как чувствительна мера μ(ТВ) к наполненности «малых» классов. Проблема наполнения «малых» классов при классификации требует более глубокого исследования.

ЗАКЛЮЧЕНИЕ

Предложенный простой алгоритм номинальной классификации по качественным признакам обладает рядом указанных выше особенностей, которые позволяют получить дополнительную информацию о процессе классификации, упрощают его применение и делают его доступным для широкого круга пользователей при решении практических задач. Тем не менее, основная тяжесть решаемых проблем с использованием качественных признаков сопряжена с выбором свойств, составляющих признаки классифицируемых объектов. Собственно классификация является заключительным этапом при формировании иерархической структуры базисных понятий «объект» → «признак» → «значение признака» → «класс» → «классификация». Для формирования такой иерархии при решении исследуемой проблемы требуется привлечение хороших специалистов по интеллектуальному анализу данных.

ЛИТЕРАТУРА

1. *Царьков С.А.* Алгоритм ближайшего соседа. — URL: https://basegroup.ru/community/articles/knn (дата обращения 26.10.2016).
2. *Thirumuruganathan S.A.* Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm / May 17, 2010. — URL: http://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed-introduction-to-k-nearest-neighbor-knn-algorithm/ (дата обращения 26.10.2016).
3. *Dudani S.A.* The Distance-Weighted k-Nearest-Neighbor Rule. Syst. Man Cybern // IEEE Trans. — 1976. — Vol. 6, N 4. — P. 325—327. — URL: refhub.elsevier.com/S0031-3203(14)00282-9/sbref14 (дата обращения 26.10.2016).
4. *Gou J., Xiong T., Kuang Y.* A Novel Weighted Voting for K-Nearest Neighbor Rule // J. Comput. — 2011. — Vol. 6, N 5. — P. 833—840. — URL: https://www.researchgate.net/publication/220405196_A_ (дата обращения 26.10.2016).
5. *Zavred J.* An Empirical Re-examination of Weighted Voting for KNN. — URL: citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.36.573 (дата обращения 26.10.2016).
6. *Kira K., Rendell L.A.* The Feature Selection Problem: Traditional Methods and New Algorithm. — Proc. AAAI-1992. — URL: www.aaai.org/Papers/AAAI/1992/AAAI92-020.pdf (дата обращения 26.10.2016).
7. *Kohavi R., Quinlan R.* Decision Tree Discovery. — URL: ai.stanford.edu/~ronnyk/treesHB.pdf (дата обращения 26.10.2016).
8. *Ratanamahatana C., Gunopulos D.* Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection — URL: www.cs.ucr.edu/~ratana/DCAP02.pdf (дата обращения 26.10.2016).
9. *Jiang Su J., Zhang H.* A Fast Decision Tree Learning Algorithm. — URL: www.cs.unb.ca/~hZhang/publications/AAAI06.pdf (дата обращения 26.10.2016).
10. *Ibrahim P.S., Chandran K.R., Kanthasamy C.J.K.* LACI: LaZy Associative Classification Using Information Gain // IACSIT Int. J. of Engineering and Technology. — 2012. — Vol. 4, N 1. — URL: www.ijetch.org/papers/309-T828.pdf (дата обращения 26.10.2016).
11. *Association* Analysis: Basic Concepts and Algorithms. — URL: www.users.cs.umn.edu/~kumar/dmbook/ch6.pdf (дата обращения 26.10.2016).

12. *Shekhawat P.B., Dhande S.S.* A Classification Technique using Associative Classification // Int. Journal of Computer Applications. — 2011. — Vol. 20. — N 5. — URL: www.ijcaonline.org/volume20/number5/pxc3873268.pdf (дата обращения 26.10.2016).
13. *Shfna T.T., Jayasudha J.S.* A Survey of Different Associative Algorihms // Asian Journal of Computer Science and Information Technology. — 2013. — Vol. 3, N 6. — P. 88—93. — URL: (citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.183.1158) (дата обращения 26.10.2016).
14. *Bal M.* Rough Sets Theory as Symbolic Data Mining Method: An Application on Complete Decision Table // Inf. Sci. Lett. 2013. — Vol. 2, N 1. — P. 35—47. — URL: www.naturalspublishing.com/files/published/92srlr5h... (дата обращения 26.10.2016).
15. *Slowinski R., Greco S., Matarazzo B.* Rough Set and Rule-based Multicriteria Decision Aiding. — URL: www.scielo.br/scielophp?script=sci_arttext&pid=S... (дата обращения 26.10.2016).
16. *Yao Y.* Three-Way Decision: An Interpretation of Rules in Rough Set Theory. — URL: www2.cs.uregina.ca/~yyao/PAPERS/rskt2009.pdf (дата обращения 26.10.2016).
17. *Grzymala-Busse J.W.* A Comparison of Three Strategies to Rule Induction from Data with Numerical Attributes // Electronic Notes in Theoretical Computer Science. — March 2003. — Vol. 82, iss. 4. — P. 132—140. — URL: www.sciencedirect.com/science/article/pii/S1571066104... (дата обращения 26.10.2016).
18. *Filho A.T.B.* A Novel Approach Based on Multiple Criteria Decision Aiding Methods to Cope with Classification Problems. — URL: www.sapili.org/livros/en/cp136471.pdf (дата обращения 26.10.2016).
19. *Руа Б.* Проблемы и методы принятия решений в задачах с многими целевыми функциями. — Вопросы анализа и процедуры принятия решений. — М.: Мир, 1976. — С. 20—57.
20. *Evseyeva I.* Solving Classification Problems with Multicriteria Decision Aiding Approaches. — URL: jyx.jyu.fi/dspace/bitstream/handle/123456789/... (дата обращения 26.10.2016).
21. *Измерительные* шкалы. — URL: www.e-educ.ru/tsisa20.html. (дата обращения 26.10.2016).
22. *Machine* Learning Repository. — URL: http://archive.ics.uci.edu/ml/ (дата обращения 26.10.2016).
23. *Gou J., Du L., Zhang Y, Xiong T.* A New distance-weighted k-nearest neighbor classifier // Journal Inf. Comput. Sci. — 2012. — Vol. 9, N 6. — P. 1429—1436.
24. *Geler Z., Kurbalija V., Radovanović M., Ivanovic M.* Comparison of different weighting schemes for the *kNN* classifier on time-series data. — URL: in print. (DOI: 10.1007/s10115-015-0881-0) link.springer.com/article/10.1007/s10115-015-0881-0 (дата обращения 26.10.2016).
25. *Kononenko I.* Estimating attributes: analysis and extensions of Relief — URL: link.springer.com/.../10.1007 %2F3-540-57868... (дата обращения 26.10.2016).
26. *Robnik-Sikonja M., Kononenko I.* Theoretical and Empirical Analysis of ReliefF and RReliefF. — URL: lkm.fri.uni-lj.si/rmarko/papers/robnik03-mlj.pdf (дата обращения 26.10.2016).
27. *Scherf M., Brauer W.* Feature Selection by Means of a Feature Weighting Approach. — URL: citeseer.ist.psu.edu/viewdoc/summary? (дата обращения 26.10.2016) doi=10.1.1.43.1821.
28. *Guvenir H.A., Akkus A.* Weighted k-Nearest Neighbor Classification on Feature Projections. — URL: www.cs.bilkent.edu.tr/tech-reports/1997/BU-CEIS-9719.pdf (дата обращения 26.10.2016).
29. *Lei J.* Classification with Confidence // Biometrika — 2014. — P. 1—5. — URL: www.stat.cmu.edu/~jinglei/conf_class_R2.pdf (дата обращения 26.10.2016).
30. *Zaragoza H., d'Alché-Buc F.* Confidence Measures for Neural Network Classifiers // Proc. 7-tgh Conf. IPMU. 1998. — URL: research.microsoft.com/pubs/66924/hugo*Z*_ipmu98.pdf (дата обращения 26.10.2016).
31. *Воронцов К.В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // Математические вопросы кибернетики / под ред. О.Б. Лупанова. — М., 2004. — Вып. 13. — С. 5—36.
32. *López V., Fernandez A., Garcia S.,* et al. An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics // Information Sciences. — 2013. — N 250. — P. 113—141. — URL: sci2s.ugr.es/imbalanced (дата обращения 26.10.2016).
33. *Jurman G., Riccadonna S., Furlanello C.* A Comparison of MCC and CEN Error Measures in Multi-Class Prediction // PLoS One. — 2012. — Vol. 7. — N 8. — URL: www.ncbi.nlm.nih.gov/pmc/articles/PMC3414515/ (дата обращения 26.10.2016).
34. *Krippendorff K.* Computing Krippendorff's Alpha-Reliability. — URL: www.asc.upenn.edu/usr/krippendorff/mwebreliability4.pdf (дата обращения 26.10.2016).

Статья представлена к публикации членом редколлегии А.А. Дорофеюком.

Корноушенко Евгений Константинович — д-р техн. наук, гл. науч. сотрудник, Институт проблем управления им. В.А. Трапезникова РАН, г. Москва,
✉ ekorno@mail.ru.

Содержание сборника «Управление большими системами», 2016, вып. 64

Пьяных А.И. Многошаговая модель биржевых торгов с элементами переговоров и счетным множеством состояний

Резчиков А.Ф., Кушников В.А., Иващенко В.А. и др. Анализ и прогнозирование характеристик безопасности авиационных транспортных систем на основе уравнений системной динамики

Емельянова Ю.П. Стабилизация нелинейных систем Форназини — Маркезини

Сазонов В.В., Скобелев П.О., Лада А.Н. и др. Применение мультиагентных технологий в транспортной задаче с временными окнами и несколькими пунктами погрузки

Мелентьев В.А. «Реберное» масштабирование гиперкубических вычислительных систем

Алгазин Г.И., Алгазина Д.Г. Информационное равновесие в модели динамики коллективного поведения на конкурентном рынке

Федянин Д.Н., Чхартишвили А.Г. Консенсус в социальной сети со сложными узлами

Чернявский А.Л., Дорофеюк Ю.А., Мандель А.С. и др. Анализ процесса госпитализации пациентов в крупной клинике методами коллективной многовариантной экспертизы

Тексты статей в свободном доступе на сайте http://ubs.mtas.ru/