Интеллектуальные системы управления, анализ данных

© 2019 г. Е.К. КОРНОУШЕНКО, д-р техн. наук (ekorno@mail.ru)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва)

ПРОЦЕДУРА КЛАССИФИКАЦИИ ОБЪЕКТОВ С СЕМАНТИЧЕСКОЙ ИЕРАРХИЕЙ ПРИЗНАКОВ

Предложена процедура классификации объектов с иерархической структурой взаимоотношений (семантикой) признаков с учетом их модальностей. Понятия семантики признаков и их модальности объясняются перед описанием самой процедуры. Рассматривается трехуровневая модель с семантической иерархией признаков: «классифицируемые объекты – метапризнаки – признаки объектов». Метапризнаки интерпретируются как семантические обобщения относящихся к ним признаков объектов. Важным этапом предлагаемой процедуры является агрегирование признаков нижнего уровня с учетом их семантической связи с метапризнаками. Агрегирование приводит к существенному уменьшению размерности исходной задачи классификации, решаемой теперь в терминах значений функций агрегирования. В качестве примера используется выборка Dermatology из известного репозитория UCI Machine Learning. На этом примере показано, что несмотря на существенную несбалансированность выборки Dermatology результаты применения предлагаемой процедуры вполне сравнимы с лучшими результатами ряда известных алгоритмов классификации, полученными на этой выборке.

Ключевые слова: субпризнак, метапризнак, семантическая иерархия признаков, модальность признака, агрегирование, классификация.

DOI: 10.1134/S0005231019110084

1. Введение

Варианты постановок и решений задач классификации[1] существенным образом зависят от структуры множества признаков классифицируемых объектов[2] и иерархии классов, априори введенных на объектах исходной выборки. Структура множества признаков (как и сами признаки) может быть разной для разных постановок задачи классификации. Так, при допущении, что все признаки независимы и измеримы (в соответствующих шкалах), в качестве модели исходных данных может быть выбрана так называемая «плоская» (flat) модель [1]. Альтернативами «плоской» модели являются различ-

[1] В данной работе рассматривается только классификация «с учителем» (supervised classification), предусматривающая наличие обучающей выборки, объекты которой однозначно принадлежат тому или иному классу введенного на выборке разбиения.

[2] Под объектом может пониматься некая сущность (физический объект, наблюдение и т.п.), описываемая соответствующей совокупностью признаков из выбранного множества признаков. В зарубежных работах «носители» признаков называются по-разному: object, entity, instance и т.п. Далее для единообразия будет использоваться слово «объект».

ные иерархические модели, в которых каждый уровень иерархии характеризуется соответствующим распределением «признаков–классов» [1]. В данной статье предполагается, что каждый объект описывается соответствующей совокупностью признаков (которые могут принадлежать и другим объектам), а классов – небольшое число и они независимы.

В последние десятилетия в работах по искусственному интеллекту (особенно в работах по классификации документов, изображений и т.п.) стало широко употребляться понятие семантики данных [2], семантики признаков [3] и семантической иерархии признаков (СИП) [4]. Предпосылками для учета СИП при классификации объектов с признаками различной физической природы явились многочисленные примеры иерархии признаков от «простейших» понятий (например, желтый, красный и т.д.) до более сложных (например, цвет), которые «встраиваются» как промежуточные уровни в соответствующие иерархии [5, 6]. Как показано в работах [4, 7], при классификации объектов, допускающих построение СИП, учет СИП улучшает результаты классификации таких объектов по сравнению с использованием для их классификации «плоской» модели.

Следующим важным понятием в данной работе является понятие модальности признаков. Считается [8], что модальность признака определяется контекстом исследуемой ситуации, в которой под модальностью признаков может пониматься различие в смысловом раскрытии понятия того или иного признака. Так, основными модальностями при классификации изображений [5] являются такие модальности как текстура, цвет, наличие геометрических особенностей в изображении (углов, прямых линий и т.п.). В медицине, например, широко применяются так называемые мультимодальные исследования [6], в которых одновременно анализируются признаки с различными модальностями (изображения, видео, текст). При этом согласно [6] качество классификации состояния пациента намного лучше, чем в случае использования признаков только одной модальности.

В данной статье рассматривается процедура построения алгоритма классификации в случае, когда признаки объектов имеют разные модальности и допускают построение СИП. Верхний уровень модели с СИП составляют объекты – «носители» признаков, а нижний – сами «простейшие» признаки с теми или иными значениями. На промежуточном уровне модели располагаются так называемые метапризнаки, содержание каждого из них определяется соответствующим множеством относящихся к данному метапризнаку «простейших» признаков нижнего уровня. Суть предлагаемого подхода к классификации – обоснование перехода от большой совокупности «простейших» признаков к гораздо меньшей совокупности метапризнаков путем соответствующих агрегирований признаков нижнего уровня с учетом семантики взаимоотношений признаков среднего и нижнего уровней. Как указано в работах [1, 9, 10], специфика агрегирования признаков при переходе к метапризнакам в структурах с СИП состоит в том, чтобы при агрегировании сохранялась структура семантических связей между признаками нижнего и среднего уровней. В [11, 12] описываются требования, которым должна удовлетворять выбираемая функция агрегирования и сама процедура выбора. В частности, в такой функции должны учитываться совокупные эффекты, обусловленные

«горизонтальными» взаимовлияниями признаков на одном и том же уровне, и такие совокупности признаков должны входить в область определения функции агрегирования. Известные методы уменьшения размерности, такие как LLE (метод локального линейного вложения, local linear embedding), ISOMAP и другие методы, анализируемые в [13], широко используются в различных задачах. Однако, как указывается в [14], в подобных методах не учитываются семантические зависимости между переменными различных уровней и различных модальностей, все такие переменные «сливаются» в переменные пространств меньшей размерности.

Цель настоящей работы состоит в том, чтобы на примере простейшей трехуровневой структуры с СИП продемонстрировать основные этапы процедуры классификации объектов:

— обоснование вводимых элементов промежуточного слоя;

— формирование модели с СИП для каждого объекта;

— выбор функций агрегирования признаков нижнего уровня;

— решение исходной задачи классификации в терминах значений используемых функций агрегирования с учетом указанных выше требований на взаимоотношения переменных разных уровней.

В качестве иллюстративного примера рассматривается выборка Dermatology из известного репозитория UCI Machine Learning [15], описывающая диагностику 6 видов кожных заболеваний у 343 пациентов на основе 34 клинических и гистопатологических анализов. Показано, как по этой выборке строится модель с СИП и как вводимое в модель агрегирование уменьшает размерность исходной задачи классификации.

Работа построена следующим образом. Во введении приводятся необходимые определения, связанные с СИП, и кратко представлен круг работ по вопросам семантической иерархии и агрегирования. В разделе 2 подробнее раскрываются формальные понятия, входящие в описание модели с СИП. В разделе 3 обсуждается тип функций агрегирования, используемых для агрегирования признаков нижнего уровня. В разделе 4 представлена упрощенная версия алгоритма классификации из работы [16], адаптированная для решения задачи классификации в терминах значений используемых функций агрегирования, а в разделе 5 оценивается вычислительная сложность предложенной процедуры. В разделе 6 описывается применение этой процедуры к выборке Dermatology. Особенности данной процедуры, важные при рассмотрении практических задач, обсуждаются в разделе 7. В заключение делаются выводы о возможных применениях предложенного подхода.

2. Определение модели с СИП

Прежде всего, подробнее определим понятие модели с СИП. Пусть задана совокупность S из n объектов, которые необходимо классифицировать. Каждый из объектов описывется соответствующей совокупностью признаков, возможно, разных модальностей, называемой описанием объекта. Обозначим через N_0 мощность объединения X_0 всех признаков. Иерархическая модель H_i с СИП строится для каждого объекта O_i, $i \leq n$. На верхнем уровне модели H_i располагается сам объект O_i. Второй уровень модели содержит m метапри-

знаков $M_1, \ldots, M_m$, где m зависит от специфики рассматриваемой задачи (так, например, в приводимой ниже выборке Dermatology метапризнаками являются возможные виды заболеваний). Будем считать, что множество метапризнаков $\{M_1, \ldots, M_m\}$ сопоставляется каждому объекту из S. При этом если описание объекта O_i, $i \le n$ не содержит некоторого метапризнака M_u, $u \le m$, полагаем $M_u = 0$ в модели H_i. Каждому ненулевому метапризнаку M_j в модели H_i однозначно сопоставляется множество G_{ij} «простейших» признаков нижнего уровня (для простоты называемых субпризнаками). Все элементы множества G_{ij} суть значения признаков из описании объекта O_i. При этом каждый субпризнак из G_{ij} может входить в описания других объектов из S, и соответственно – в другие множества[3] $G_{is}, \ldots, G_{ik}$. Однозначность отнесения множества G_{ij} к метапризнаку M_j говорит о том, что все субпризнаки из G_{ij} имеют семантические связи с M_j, а некоторые субпризнаки из G_{ij} – и с другими метапризнаками.

Переменные разных уровней в модели H_i могут анализироваться «сверху-вниз» или «снизу-вверх». Для наглядности направление исследуемого соотношения указывается направлением соответствующей стрелки $\downarrow$ или $\uparrow$. Модель H_i может быть представлена следующим образом:

$$O_i \downarrow \{M_j\}_{j=1}^{m} \downarrow \{G_{ij}\}_{j=1}^{m} \downarrow \left\{ \{x_{ijk}\}_{k=1}^{N_j} \right\}_{j=1}^{m},$$

где x_{ijk} – элементы множества G_{ij} с мощностью N_j.

В плане дальнейшего перехода к классификации объектов из S рассмотрим, как распределяются метки (labels) классов на уровнях модели H_i. Допустим, что на исходной совокупности объектов S введено некоторое разбиение π на K ($K \le m$) блоков (классов), и пусть объект O_i принадлежит ν-му классу C_ν (т.е. имеет метку ν). Эта метка переносится и на субпризнаки из G_{ij}, входящие в описание объекта O_i. Заметим, что поскольку субпризнаки из G_{ij} могут входить в описания и других объектов из S, скажем, объекта O_r, входящего в блок C_μ разбиения π, таким субпризнакам кроме метки ν приписывается и метка μ, так что каждый субпризнак из G_{ij} может иметь, в принципе, несколько меток (такая совокупность меток называется в литературе bag of labels). Метки каждого субпризнака из G_{ij} переносятся и на те метапризнаки $M_j, \ldots, M_k$, с которыми данный субпризнак семантически связан. Этот факт учтем в виде соотношения

$$G_{ij} \uparrow \{M_k\}_{k=1}^{m_i} \uparrow \{O_p\}_{p=1}^{r} \uparrow \{\nu\}. \tag{1}$$

Здесь $\{O_p\}_{p=1}^{r}$ – совокупность объектов, в описания которых входят признаки, семантически связанные с $M_j, \ldots, M_k$, а $\{\nu\}$ – множество меток таких объектов, включающее метку ν объекта O_i. Аналогично строятся модели H_r для остальных объектов H_r, $r = 1, \ldots, n$ из S. Далее считаем выполненными следующие условия:

[3] Так, при постановке медицинского диагноза учитывается тот факт, что тот или иной симптом (субпризнак) может относиться к разным заболеваниям – см. приведенный ниже пример. На этапе классификации такие неоднозначности учитываются и усложняют выработку диагноза.

1. Все модели H_r, $r = 1, \ldots, n$ изначально отличаются только значениями субпризнаков нижнего уровня и структурой семантических связей субпризнаков с соответствующими метапризнаками (нулевые значения на втором и третьем уровнях допустимы).

2. Субпризнаки с неопределенными или неизвестными значениями не допустимы.

3. Каждый объект исходной совокупности S принадлежит однозначно к тому или иному классу разбиения $\pi = (C_1, \ldots, C_K)$, классы взаимно независимы и $K \le m$[4].

Ключевой момент предлагаемого подхода к классификации структур с СИП состоит в агрегировании субпризнаков в каждом из множеств G_{ij}, $i = 1, \ldots, n,\ j = 1, \ldots, m$ и переходе от пространства исходных признаков большой размерности N_0 к пространству значений функций агрегирования гораздо меньшей размерности m. Предварительная подготовка исходных данных к запуску процедуры классификации состоит в прохождении следующих этапов: а) нормализация субпризнаков в каждом множестве G_{ij}, $i = 1, \ldots, n$, $1 \le j \le m$ и выбор функций агрегирования; б) нахождение значений функций агрегирования для каждого множества G_{ij}; в) определение каждой модели P_i, $i = 1, \ldots, n$, являющейся модификацией соответствующей модели H_i в терминах найденных значений функций агрегирования.

3. Выбор функций агрегирования для субпризнаков

Этапы выбора функций агрегирования для субпризнаков:

а. Поскольку значения субпризнаков в каждом множестве G_{ij} могут изначально выбираться из разных шкал, перед агрегированием субпризнаки необходимо их нормализовать. Нормализация проводится по правилу $y_{ijk} = \frac{x_{ijk}}{\sum_{s=1}^{n} x_{sjk}}$, где каждая нормализованная величина y_{ijk} есть доля исходного значения субпризнака x_{ijk} из G_{ij} в сумме (по всем объектам совокупности S) значений этого субпризнака. Подчеркнем, что нормализующий множитель $\frac{1}{\sum_{s=1}^{n} x_{sjk}}$ для каждого признака $X_k \in X_0$, входящего (с разными значениями) в множества G_{ij}, $1 \le j \le m$, $i = 1, \ldots, n$, один и тот же. Распределение нормализованных значений $\{y_{ijk}\}$, $k = 1, \ldots, N$ на множестве G_{ij} не является вероятностным, поскольку сумма (по k) переменных $\{y_{ijk}\}$ в общем случае не равна единице.

б. Выбираемая функция агрегирования должна быть чувствительной к распределению и значениям ненулевых агрегируемых переменных. В случае вероятностного распределения таких значений можно использовать инверсную энтропию [17] (энтропию со знаком плюс). По аналогии с [17] функцию агрегирования переменных $_{ijk}\ y_{ijk}$ в множествах G_{ij} определим как

$$F_{ij}(y_{ij1}, \ldots, y_{ijN_j}) = \left(-\sum_{k=1}^{N_j} y_{ijk} \log_2(y_{ijk}) \,|\, y_{ijk} > 0 \right). \tag{2}$$

[4] Случай $K > m$ имеет особенности, требующие отдельного рассмотрения.

Для удобства отображение (2) назовем псевдоэнтропией. Отображение $F_{ij} : [0,1]^{N_j} \to R$ является нелинейным монотонным неубывающим отображением. Функция F_{ij} переводит вектор $(y_{ij1}, \ldots, y_{ijN_j})$ в соответствующее положительное число $w_{ij} \in R$, которое назовем весом метапризнака M_j в объекте O_i выборки S. Аналогичным образом найдем веса w_{ij} для всех метапризнаков M_j, $j = 1, \ldots, m$, $i = 1, \ldots, n$. Отличие модели P_r от модели H_i лишь в том, что элементы $\{M_1, \ldots, M_m\}$ второго уровня заменяются значениями соответствующих весов w_{rj}. Строку весов, сопоставляемых объекту O_i, обозначим как $W_i = \{w_{ij}\}$, $j = 1, \ldots, m$.

4. Процедура классификации объектов с использованием весов метапризнаков

4.1. Разбиение исходной совокупности S на обучающую и тестовую выборки

Классификация (supervised) объектов предусматривает разбиение исходной совокупности S на обучающую выборку (ОВ) и тестовую выборку (ТВ), которые содержат $n_{\text{ОВ}}$ и $n_{\text{ТВ}}$ объектов соответственно. При этом все объекты сохраняют те же метки, что и ранее в совокупности S. Другими словами, разбиение π разделяется на два подразбиения $\pi_{\text{ОВ}}$ и $\pi_{\text{ТВ}}$ с выполнением следующих важных требований:

Требования к формированию ОВ и ТВ.

а. Каждый класс подразбиения $\pi_{\text{ТВ}}$ должен иметь непустое пересечение с каким-либо классом подразбиения $\pi_{\text{ОВ}}$.

б. Модели для объектов O_i из ОВ и ТВ являются теми же, что и ранее построенные модели H_i для этих объектов из S. в. Метки объектов из ТВ не участвуют в процедуре построения алгоритма классификации, а используются лишь при определении точности классификации построенного алгоритма.

Задача классификации объектов с СИП формулируется следующим образом. Метки объектов из ОВ априори считаются известными. Метка ν объекта O_i из ОВ переносится на компоненты соответствующего этому объекту O_i вектора весов $W_i = \{w_{ij}\}$, $j = 1, \ldots, m$, так что соотношение (1) для объекта O_i выглядит так:

$$O_i \downarrow \{w_j\}_{j=1}^{m} \downarrow \nu.$$

Метки для объектов из ТВ определяются вхождением этих объектов в тот или иной блок подразбиения $\pi_{\text{ТВ}}$. Для удобства при рассмотрении моделей из ТВ все элементы снабжаются штрихами. При этом функция агрегирования субпризнаков в G'_{pj} определяется по аналогии с (2) как

$$F'_{ij}(y'_{ij1}, \ldots, y'_{ijN'_j}) = \left(-\sum_{t=1}^{N'_j} y'_{ijt} \log_2(y'_{ijt}) \,\middle|\, y'_{ijt} > 0 \right).$$

Поскольку значения субпризнаков в моделях объектов из ОВ и ТВ могут не совпадать, для корректного сравнения значений весов w_{ij} и w'_{ij}, получаемых с применением функций агрегирования F_{ij} и F'_{ij}, нормализующий множитель

$\frac{1}{\sum_{s=1}^{n} x_{sjk}}$ для субпризнаков из G'_{ij} должен быть тем же самым, что и при нормализации субпризнаков в G_{ij}.

4.2. Классификация объектов ТВ с использованием весов метапризнаков

Ниже с краткими комментариями приводятся основные этапы алгоритма из [16] с адаптацией к классификации объектов в терминах весов метапризнаков.

Пусть Z – тестовый объект, модель H_Z для которого характеризуется вектором весов метапризнаков $W'_Z = (w'_{Z1}, \ldots, w'_{Zm})$. Перечислим основные этапы адаптированного алгоритма классификации, обозначаемого для краткости как ААК.

А. *Определение понятия "допустимой близости" для весов метапризнаков.* Для каждого метапризнака M_j определяются его максимальный $m_{\max j}$ и минимальный $m_{\min j}$ веса по всем объектам из S. Разность этих значений делится на некоторое выбираемое число h (о выборе значения h будет сказано далее в п. Д). Обозначим $d_j = \frac{w_{\max j} - w_{\min j}}{h}$. Два значения w_{rj} и w'_{sj} назовем d_j-близкими, если модуль их разности не больше d_j. Для простоты будем считать, что h не зависит от j.

Б. *Построение матрицы весов для тестового объекта Z.* Пусть w'_{Zj} — вес метапризнака M_j в объекте Z. Совокупность весов метапризнака M_j в объектах из ОВ, таких что эти веса d_j-близки к значению w'_{Zj}, назовем d_{Zj}-окрестностью значения w'_{Zj}. Множество объектов из ОВ, образующих d_{Zj}-окрестность значения w'_{Zj}, обозначим как $U(w'_{Zj}, d_j)$, совокупность весов метапризнаков объектов из множества $U(w'_{Zj}, d_j)$ — как $V(w'_{Zj}, d_j)$, а совокупность меток объектов из $U(w'_{Zj}, d_j)$ — как $L(w'_{Zj}, d_j)$. Пусть $N_{Zj\nu}$ — число объектов из $U(w'_{Zj}, d_j)$, входящих в блок C_ν разбиения $\pi_{\text{ОВ}}$. Сопоставим значению w'_{Zj} число[5] $g_{Zj\nu} = \frac{N_{Zj\nu}}{|U(w'_{Zj}, d_j)||C_\nabla|}$, где $|*|$ обозначает мощность соответствующего множества. Найденное число $g_{Zj\nu}$ будем рассматривать как вес метки ν в множестве $L(w'_{Zj}, d_j)$. Подобным образом найдем множество $N_{Zj\mu}$ объектов из $U(w'_{Zj}, d_j)$, принадлежащих классу C_μ, $\mu \neq \nu$, и вычислим вес $g_{Zj\mu}$ каждой метки μ $(\mu = 1, \ldots, K)$. Сформируем столбец весов меток $G_{Zj} = (g_{Zj1}, \ldots, g_{ZjK})^T$ для веса w'_{Zj} метапризнака M_j в объекте Z. Заметим, что все координаты в G_{Zj} — неотрицательные и не большие единицы. Столбец весов G_{Zj} можно интерпретировать как локальный (по метапризнаку M_j) классификатор для объекта Z. Аналогичным образом найдем векторы-столбцы весов для ненулевых значений всех m метапризнаков объекта Z и сгруппируем эти столбцы в матрицу Q_Z размера $K \times m$, которую назовем матрицей весов меток метапризнаков в объекте Z.

В. *Определение метки для тестового объекта Z.* Интерпретируем матрицу Q_Z как совокупность локальных классификаторов (по метапризнакам объекта Z). В [18] говорится о том, что классифицирующая способность объединения локальных классификаторов может быть усилена путем их взвешивания (combining) с использованием какой-либо нелинейной монотонной

[5] В [16] объясняется структура коэффициента g_{Z_j}.

функции. В качестве такой функции взвешивания $\Phi_{Z\mu}$ по каждой строке $G_{Z\mu}$ с номером μ в матрице Q_Z возьмем функцию того же типа, что и F_{ij} (см. (2)), и применим ее к каждой строке $G_{Z\mu}$ матрицы Q_Z. Сформируем вектор $R_Z = E(G_{ZK})$, который назовем классифицирующим вектором для объекта Z. Номер координаты вектора R_Z с наибольшим значением считаем меткой, приписываемой тестовому объекту Z. Аналогичным образом производится классификация остальных объектов ТВ.

Г. *Параметры классификации.* Тестовый объект Z считается правильно классифицированным, если его метка, найденная в п. В, совпадает с меткой этого объекта в подразбиении $\pi_{\text{ТВ}}$. Точность классификации определяется как отношение правильно классифицированных объектов ТВ к общему количеству объектов ТВ. При анализе алгоритма классификации в первую очередь обращается внимание на значение точности классификации. Однако точность классификации является лишь одним из качеств алгоритма классификации. При наличии более двух классов с сильно различающейся мощностью основную "нагрузку" при классификации может брать на себя класс с наибольшей мощностью, и высокий процент правильной классификации может не означать хорошей классификации объектов в классах с небольшой мощностью, хотя во многих практических задачах именно такие классы представляют особый интерес. Подобный случай рассматривается в приводимом ниже примере (см. раздел 7). Классификация в случае сильно несбалансированных выборок представляет отдельное направление в классификации, в рамках этого направления известно много публикаций. В ААК предлагается использовать следующую меру $\rho_{\text{ТВ}}$ наполняемости блоков разбиения $\pi_{TB} = (C'_1, \ldots, C'_{K'})$: $\rho_{TB} = \sum_{\nu=1}^{K'} \frac{|C'_{0\nu}|}{|C'_\nu|}$, где $|C'_{0\nu}|$ – количество правильно классифицированных объектов в классе C'_ν. Значения показателя $\rho_{\text{ТВ}}$ определены в интервале $[0, K']$. Чувствительность меры $\rho_{\text{ТВ}}$ к наполнению того или иного класса обратно пропорциональна мощности наполняемого класса.

Д. *Выбор значения параметра* h существенным образом влияет на точность классификации в ААК, поэтому достижение «приемлемого» значения точности классификации определяет окончательное значение h. Это – типичный подход «с обратной связью» (wrapper approach). При условии, что выбор значения h не зависит от метапризнаков объектов, процедура выбора h может быть сведена к простейшему одномерному поиску [16]. При переходе к использованию весов метапризнаков исходная задача превращается в стандартную задачу классификации, которую можно решать в рамках «плоской» модели «объекты – метапризнаки» с применением любого алгоритма классификации. Однако в данной работе используется алгоритм из [16], поскольку он обладает рядом практически важных особенностей, обсуждаемых в разделе 7.

5. Оценка вычислительной сложности предложенного алгоритма

Описанная выше процедура классификации объектов с СИП состоит из двух последовательных частей: а) агрегирование входной информации на ОВ

и ТВ; б) решение задачи классификации с использованием весов метапризнаков. Пусть, как и ранее, N_0 – размерность пространства всех субпризнаков. При агрегировании рассматриваются значения каждого субпризнака по всем объектам ОВ и ТВ. Для простоты положим, что ОВ и ТВ имеют по n объектов, тогда вычислительная сложность этапа агрегирования оценивается как $O(2nN_0)$. ААК представляет собой сложный цикл, в котором для каждого объекта ТВ рассматриваются все объекты ОВ и для каждой пары объектов производится сравнение значений каждого из m соответствующих весов метапризнаков на предмет их «допустимой близости». Положим, что для определения «приемлемого» значения параметра h, определяющего значение $d(h)$-близости, полный цикл ААК повторяется q раз, $q \leq m$. Тогда вычислительная сложность этапа б) оценивается как $O(n^2mq)$ или, с учетом того, что $q \leq m$, – как $O(n^2m^2)$. Для оценки всей процедуры необходимо рассмотреть асимптотическое поведение функции $nN_0 + n^2m^2$. При выполнении условия $N_0 < nm^2$ с учетом замечаний, сделанных в [19], вычислительная сложность всей процедуры оценивается как $O(n^2m^2)$. Если же рассматривать задачу классификации в рамках «плоской» модели (без агрегирования) с использованием алгоритма из [16] , то вычислительная сложность алгоритма классификации оценивается как $O(n^2N_0^2)$ при условии, что $N_0 > qm$. Таким образом, агрегирование уменьшает вычислительную сложность задачи классификации в отношении порядка $\left(\frac{N_0}{m}\right)^2$.

6. Пример

В качестве примера, на котором демонстрируется эффективность предложенной процедуры классификации, использовалась выборка Dermatology из репозитория UCI Machine Learning [15]. В этой выборке представлены результаты 34 клинических анализов (наличие сыпи или покраснений на коже, температура тела и т.п.) и гистопатологических анализов (шелушение, соскабливания, биохимия и т.п.), проведенных над 343 пациентами с целью определения (классификации) у них тех или иных кожных заболеваний из 6 видов возможных заболеваний. Более подробная информация о связи тех или иных результатов анализов с конкретным видом заболевания, представляющая интерес для читателя, содержится в [20, 21][6]. Значения каждого из анализов у каждого пациента определялись в единой качественной шкале $(0, 1, 2, 3)$, где 0 означает отсутствие данного анализа у исследуемого пациента. Поскольку тот или иной анализ в общем случае не достаточен для однозначного указания вида заболевания, предварительно исследовалась «неоднозначность» каждого из анализов, и для каждого вида заболевания формировалось множество соответствующих анализов (без учета их значений), возможно связанных с этим заболеванием. Гистограмма количества таких связей для каждого из анализов представлена в [21]. По причинам, не относящимся к тематике данной статьи, из исходного множества анализов были удалены анализы с номерами 1, 2, 11, 13, 17, 18, 32, 34. Из оставшихся 26 анализов были выделены 6 совокупностей анализов, каждая из которых, возможно, связана

[6] В частности, в [20] более подробно указано, какие признаки и каких кожных заболеваний являются клиническими, а какие – гистопатологическими.

Таблица 1. Совокупности упорядоченных по «важности» анализов для соответствующих видов заболеваний

Номер вида заболевания	Номера анализов
1	20, 22, 21, 28, 16, 10, 9, 19, 24, 3, 26, 29, 6, 33, 12, 27
2	28, 20, 22, 5, 26, 21, 9, 24, 27, 16, 29, 6, 12, 25, 8, 33
3	33, 27, 29, 6, 12, 25, 8, 21, 14, 20, 22, 16, 9, 10, 4, 23
4	21, 9, 20, 22, 10, 28, 33, 27, 6, 12, 25, 8, 23, 29, 24, 4
5	15, 5, 14, 20, 10, 9, 22, 26, 28, 24, 27, 29, 6, 12, 25, 33
6	7, 31, 5, 22, 26, 21, 24, 30, 27, 29, 6, 12, 8, 15, 33,19

Таблица 2. Мощности классов разбиений $\pi_{\text{ОВ}}$ и $\pi_{\text{ТВ}}$

$\pi_{\text{ОВ}}$	55	31	30	25	24	6
$\pi_{\text{ТВ}}$	52	29	40	24	23	4
$\{C'_{0\nu}\}$	49	24	40	23	23	3

с соответствующим видом заболевания (см. табл. 1). Анализы в каждой из этих совокупностей упорядочивались по «важности» при отнесении их к указанному виду заболевания[7].

В терминах моделей с СИП все входящие в выборку Dermatology анализы интерпретируются как субпризнаки с разными значениями для разных пациентов, а виды заболеваний – как метапризнаки. При этом согласно табл. 1. тот или иной субпризнак может соответствовать нескольким метапризнакам. Множество пациентов в выборке Dermatology можно рассматривать как выборку S, в которой каждому пациенту соответствует определенная строка в выборке S. Состояние каждого пациента характеризуется конкретной совокупностью ненулевых анализов. По этой совокупности с привлечением консилиума врачей для каждого пациента был указан диагноз, т.е. конкретный вид заболевания из 6 возможных видов. В терминах задачи классификации такой диагноз интерпретируется как приписывание соответствующей метки каждому пациенту, а совокупность таких меток на множестве пациентов – как введение разбиения π на выборке S. При этом каждый пациент однозначно относится к некоторому классу C_ν разбиения π, т.е. характеризуется меткой ν. Таким образом, исходная информация вполне достаточна для применения описанной выше процедуры классификации.

По исходной выборке S сформируем ОВ и ТВ путем попеременного отнесения очередного пациента к ОВ или ТВ и определим разбиения $\pi_{\text{ОВ}}$ и $\pi_{\text{ТВ}}$ с соблюдением указанных в разделе 4.1 требований. Мощности классов этих разбиений приведены в первых двух строках табл. 2. Видим, что ОВ и ТВ являются несбалансированными выборками, поскольку число элементов самого крупного блока разбиения $\pi_{\text{ОВ}}$ и $\pi_{\text{ТВ}}$ превышает число элементов самого мелкого блока более чем в 9 раз. Точность классификации ААК, описанного в разделе 4.2 и примененного к объектам ТВ, составляет 94,7%. При этом число правильно классифицированных объектов в каждом блоке разбиения $\pi_{\text{ТВ}}$

[7] В силу агрегирования субпризнаков в моделях с СИП подобная упорядоченность далее не используется (см. раздел 7).

Таблица 3. Сравнительные оценки точности классификации некоторых алгоритмов на выборке Dermatology

Алгоритм классификации	kNN	«Наивный Байес»	ЛДА	Дерево решений	VFI5	ААК
Точность классификации (%)	97,2	98,3	96,1	98,3	96,2	94,7

представлено в третьей строке табл. 2, а определенный в разделе 4.2 коэффициент $\rho_{\text{ТВ}}$ заполнения блоков разбиения $\pi_{\text{ТВ}}$ равен 5,47, что говорит о хорошем качестве заполнения блоков разной мощности[8]. В табл. 3 приведены заимствованные из [21] данные о точности классификации на ТВ[9] ряда известных алгоритмов (в том числе алгоритма VFI5, описанного в [20], и ААК), в которых использовались 6 множеств признаков, приведенных в табл. 1. Сокращенные названия алгоритмов в табл. 3: kNN – некоторый алгоритм из семейства kNN-алгоритмов, базирующихся на понятии «ближайшей окрестности», а ЛДА – линейный дискриминантный анализ. Согласно табл. 3 точность классификации ААК несколько меньше, чем в приведенных алгоритмах, однако ААК обладает рядом положительных качеств при решении практических задач, эти качества обсуждаются в разделе 7.

7. Практически важные особенности ААК

Обсуждаемые ниже особенности ААК включают: 1) качество классификации на существенно несбалансированных выборках; 2) уменьшение размерности множества данных на этапе классификации; 3) ослабление контекстной зависимости модальностей субпризнаков в ААК.

1. Качество классификации ААК на существенно несбалансированных выборках. Если исходная выборка является существенно несбалансированной выборкой (как в приводимом выше примере), то высокая точность классификации в целом может не означать хорошей классификации объектов в классах с небольшой мощностью. Поясним эту мысль на приведенном выше примере. Здесь ААК обеспечил неплохое качество заполнения наименьшего блока C_6 разбиения $\pi_{\text{ТВ}}$, классифицировав правильно три объекта из четырех. Неплохое качество заполнения мелких блоков показал и исходный алгоритм автора [16] на ряде существенно несбалансированных выборок из репозитория UCI Machine Learning. Это свойство исходного алгоритма и ААК допускает следующее объяснение.

Общий подход к построению алгоритмов классификации (supervised) состоит в том, что каждому объекту из ОВ «навешивается» метка класса заданного разбиения $\pi_{\text{ОВ}}$, и далее этим меткам приписываются веса, вычисляемые тем или иным образом. Поскольку при каждом разбиении исходной выборки на ОВ и ТВ в каждой из них – конечное число объектов, каждой метке класса при каждом разбиении соответствует некоторое конечное множество

[8] Это же свойство алгоритма, описанного в [16], проявлялось и при рассмотрении ряда сильно несбалансированных выборок из репозитория UCI Machine Learning.

[9] Как сказано в [21], ОВ и ТВ составляли по 50% исходной выборки Dermatology.

объектов исходной выборки с точечными значениями признаков[10]. Поскольку точность классификации на конкретной выборке зависит от количества объектов в ОВ и ТВ, для обоснования результирующей точности классификации используют разные процедуры разбиений (butstrap, boosting и т.п.). Однако в таких процедурах каждой метке будет соответствовать другое, но конечное множество признаков с точечными значениями. Отличие алгоритма из работы [16] и ААК от известных алгоритмов состоит в том, что для всякого значения количественного или качественного признака из ТВ задается непрерывная окрестность «допустимо близких» к нему значений этого же признака из ОВ. (см. раздел 4.2, п. Д). При этом каждой метке будет соответствовать результирующее конечное множество интервалов значений этого признака. Поскольку проблема построения алгоритма с оптимальным значением точности классификации в классе обычно используемых выборок является, как известно, NP-полной, о превосходстве того или иного подхода к построению алгоритма можно судить лишь по полученным результатам на используемом множестве выборок.

В приводимом выше примере ААК позволяет вместо значений 26 качественных признаков рассматривать 6 весов метапризнаков как значений используемых функций агрегирования. Совокупность классов на ТВ, содержащей 171 объект, имеет вид (см. табл. 2): $C_1 = 52, \ldots, C_6 = 4$, т.е. число элементов в максимальном блоке превышает число элементов в минимальном блоке более чем в 13 раз). Точность классификации ААК на ТВ равна 94,7%. Заметим, что правильная классификация каждого из трех объектов блока $C6$ «забирает» $100(3:171) \approx 1{,}7\%$ от полной точности классификации. Согласно табл. 3 максимальная точность классификации двух известных алгоритмов классификации («Наивный Байес» и Дерево решений) на той же ТВ равна 98,3%. Отсюда следует, что такая точность классификации каждого из этих алгоритмов может быть достигнута и при полном «игнорировании» блока C_6 вместе с правильной классификацией объектов в остальных блоках ТВ, поскольку $100\% - 1{,}7\% \approx 98{,}3\%$ (это тем более справедливо для остальных алгоритмов из табл. 3). При этом согласно табл. 2 ААК правильно классифицирует 3 из 4 элементов в блоке C_6.

2. Уменьшение размерности множества данных на этапе классификации. В силу агрегирования субпризнаков в каждом из множеств, семантически связанных с соответствующими метапризнаками, размерность множества значений m функций агрегирования уменьшается в $\left(\frac{N_0}{m}\right)^2$ раз по сравнению с размерностью N_0 множества субпризнаков.

3. Ослабление контекстной зависимости модальностей субпризнаков в ААК. Во многих практических задачах (особенно в медицине) значения субпризнаков на ОВ и ТВ определяются приближенно (в качественных шкалах), что имеет место и в рассмотренном выше примере. В то же время известно, что такой популярный подход как построение дерева решений крайне чувствителен к возмущениям значений признаков из ОВ. Теоретический анализ качества алгоритмов классификации (supervised) в рамках статистического

[10] Особняком здесь стоит алгоритм VFI5, описанный в [20], в котором для каждой метки вычисляются интервалы значений каждого признака объектов из ОВ.

подхода представлен в работе [22] в предположении, что вероятностные распределения значений признаков статистически эквивалентны на ОВ и ТВ. Однако это предположение, как правило, не выполняется в практических задачах (в частности, в приводимом выше примере). В устройствах обработки информации с помехами для повышения помехоустойчивости давно применяется агрегирование, реализуемое на сумматорах различных видов. При этом, как правило, не обращается внимание на модальности агрегируемых сигналов.

Во многих практических задачах модальность признака зависит от его значений, при этом понятие модальности становится контекстно зависимым (context sensitive). Нормализация субпризнаков и их последующее агрегирование в каждом из множеств G_{ij} позволяет несколько уменьшить взаимовлияния субпризнаков в множествах G_{ij}. Поскольку и нормализация, и агрегирование проводятся для тех же субпризнаков в ТВ, распределения модальностей субпризнаков в соответствующих множествах G_{ij} и G'_{ij} могут не совпадать, как и значения соответствующих весов w_{ij} и w'_{ij}. Наличие меток разбиения $\pi_{\text{ОВ}}$ на субпризнаках из ОВ «удаляет» эту проблему, так как всем субпризнакам из множества G_{ij}, относящегося к объекту O_i с меткой ν, приписывается метка ν (как и всему вектору весов $W_i = \{w_{ij}\}$, $j = 1, \ldots, m$, и приписывание другой метки μ весу w'_{ij} зависит лишь от «допустимой близости» веса w'_{ij} к w_{ij}. Таким образом, наличие известного разбиения на ОВ является очень важной «подсказкой» при построении алгоритмов классификации для моделей с СИП.

И еще одно важное замечание относительно упорядочения анализов по «важности» в строках табл. 1. Можно заметить, что и в [20], где подробно описывается алгоритм VFI5, и тем более в ААК не используется такое упорядочение. Оно крайне важно лишь на этапе выработки исходного диагноза для каждого пациента, и становится ненужным на этапе классификации, когда метка для каждого пациента уже определена. Подобный процесс упорядочения анализов по «важности» довольно сложный и ответственный, тогда как наличие исходных меток на объектах выборки существенно упрощает применение алгоритмов классификации.

8. Заключение

Описанная в настоящей статье процедура классификации объектов с СИП предусматривает выполнение двух необходимых условий:

– функции агрегирования не нарушают исходных семантических связей субпризнаков с метапризнаками, к которым они относятся;

– нормализующие множители для субпризнаков из соответствующих множеств G_{ij} и G'_{ij} в ОВ и ТВ совпадают.

Несмотря на простоту иерархии моделей с СИП, практические особенности процедуры классификации объектов с СИП, приведенные в разделе 7, позволяют применять эту процедуру для проведения классификации во многих многомерных задачах, в которых понятия семантической связи и модальности переменных имеют существенное значение (в частности, в медицинских, финансовых и социальных задачах).

СПИСОК ЛИТЕРАТУРЫ

1. *Borges H.B., Silla C.N., Nievola J.C.* An evaluation of global-model hierarchical classification algorithms for hierarchical classification problems with single path of labels // Comp. Math. Appl. 2013. V. 66. P. 1991–2002.

   https://www.sciencedirect.com/science/. . . /S08981221130043.

2. *Liu H.* Towards Semantic Data Mining // www.ceur-ws.org/Vol-660/paper6.pdf

3. *Motik B., Maedche A., Volz R.* A Conceptual Modeling Approach for Semantics-Driven Enterprise Applications // Proc. Meaningful Internet Syst. 2002. P. 1082–1099. www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10

4. *Albaradei S., Wang Y.* Object Classification Using a Semantic Hierarchy // www.cs.umanitoba.ca/˜ywang/papers/isvc14_hierarchy.

5. *Fatimaezzahra M., Abdelaziz E., Mohamed S., Loubna B.* Towards Domain Ontology Creation Based on a Taxonomy Structure in Computer Vision // Int. J. Adv. Comput. Sci. Appl. (IJACSA). 2016. V. 7. No. 2. P. 28–43.

   https://thesai.org/Downloads/Volume7No2/Paper_38-Towards. . .

6. *Wang Y., Halper M., D. Wei D., Perl Y., Geller J.* Abstraction of complex concepts with a refined partial-area taxonomy of SNOMED // J. Biomed. Inform. 2012. V. 45. P. 15–42. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3313654

7. *Ciaramita M., Hofmann T., Johnson. M.* Hierarchical Semantic Classification: Word Sense Disambiguation with World Knowledge // https://pdfs.semanticscholar.org/faa4/a19f4edd1d97a09

8. *Deng W.-Y., Liu D., Dong Y.-Y.* Feature Selection and Classification for High-Dimensional Incomplete Multimodal Data // Math. Probl. Eng. 2018. V. 2018. Article ID 1583969. 9 pages. https://doi.org/10.1155/2018/1583969

9. *Fernandez M.J., Eastman C.M.* Basic Taxonomic Structures and Levels of Abstraction // Proc. 1st ASIS SIG/CR Classif. Res. Workshop. 1990. P. 59–70.

   https://journals.lib.washington.edu/index.php/acro/. . .

10. *Verma N., Mahajan D., Sellamanickam D., Nair V.* Learning Hierarchical Similarity Metrics // www.cs.toronto.edu/˜vnair/cvpr12.pdf

11. *Bettencourt L.M.A.* The Rules of Information Aggregation and Emergence of Collective Intelligent Behavior // onlinelibrary.wiley.com/doi/10.1111/j.1756-8765. . . /full

12. *Marichal J.-L.* Aggregation functions for decision making // https://arxiv.org>math

13. *Bengio Y., Paiement J.-F., Vincent P., Delalleau O., Le Roux N., Ouimet M.* Out-of-Sample Extensions for LLE, Isomap, MDS, Eigenmaps, and Spectral Clustering.

    https://papers.nips.cc/. . . /2461-out-of-sample-extensions-for-l

14. *Hua Y.* Cross-Modal Correlation Learning by Adaptive Hierarchical Semantic // www.ieeexplore.ieee.org/document/7422147/

15. Machine Learning Repository // archive.ics.uci.edu/ml/datasets.html

16. *Корноушенко Е.К.* Алгоритм классификации путем парного сравнения признаков // АиТ. 2017. № 11. С. 151–166.

    Kornoushenko E.K. Classification Algorithm Based on Pairwise Comparison of Features // Autom. Remote Control. 2017. V. 78. No. 11. P. 2062–2074.

17. *Magimai.-Doss M., Hakkani-Tür D., Cetin O., Shriberg E., Fung J., Mirghafori N.* Entropy-based C;assifier Combimation for Sentence Segmentation //

    www.citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1

18. *Воронцов К.В.* Комбинаторный подход к оценке качества обучаемых алгоритмов / Математические вопросы кибернетики. Под ред. О.Б. Лупанова. Т. 13. М.: Физматлит, 2004. С. 5–36.
19. *Zindros D.* A Gentle Introduction to Algorithm Complexity Analysis // www.discrete.gr/complexity/
20. *Govenir H.A., Demiroz G., Ilter N.* Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals // Artif. Intelligence Medicin. 1998. V. 13. P. 147–165.
21. *El-Baz A.H.* Filter Based Feature Selection for Automatic Detection of Erythemato-squamous Diseases // British J. Math. Comput. Sci. 2015. V. 9. No. 5. P. 394–406. www.journalrepository.org/. . . /El-Baz952015BJMCS17618.p. . .
22. *Schain M.* Machine Learning Algorithms and Robustness // Diss. Phd. Tel-Aviv. Univ. 2015. https://m.tau.ac.il/~mansour/students/Mariano_Schain_Phd.pdf

Статья представлена к публикации членом редколлегии О.П. Кузнецовым.

Поступила в редакцию 04.04.2018

После доработки 20.12.2018

Принята к публикации 07.02.2019