Алгоритмы классификации за минимальное число шагов[1]

А. Т. Вахитов
Санкт-Петербургский государственный университет
О. А. Граничина
Российский государственный педагогический университет[2]

Рассматривается проблема распознавания: определение класса, к которому принадлежит некоторый объект, за минимальное в среднем число измерений признаков. Определяется постановка задачи и формулируется рандомизированный алгоритм сортировки признаков по их значимости для определения. Обосновывается практическая значимость алгоритма, на примере определителя биологических видов.

1. Введение

Когда феноменов набирается достаточно много, "вдруг обнаруживается", что некоторые из них имеют нечто общее, но отличное от других - феномены как-то группируются, приходится выделять некоторые общие и отличные признаки и связи - и мы поневоле переходим к этапу классификации (систематизации)... В адекватном переводе с латыни классификация (classis - группа, facio - делаю) - "группирование".

В.Д. Ермак Классификация?.. Типология... Идентификация !..

Задача информационного обеспечения процесса классификации является классической в области искусственного интеллекта. Рассмотрим ситуацию, когда система классификации уже задана, и необходимо уметь эффективно определять класс, к которому принадлежит некоторый объект, идентифицировать его, для того чтобы выявить интересующие нас свойства (определить диагноз больного, вид неисправности механизма, биологический вид растения или животного).

Если измерения признаков объекта, используемых при его идентификации, требуют каких-то затрат для человека (временных, физических, и т. п.), то становится актуальным минимизировать число необходимых измерений. Остановимся подробнее на проблеме

[1]Работа выполнена при поддержке Российского фонда фундаментальных исследований (грант №05-07-90179).
[2]©А. Т. Вахитов, О. А. Граничина, 2006

определения порядка измерения признаков, который обеспечит минимальность числа измерений в некотором смысле. В качестве примеров различных подходов, диктующих свои методы решения, приведем [1–3]. Определителем далее будет называться некоторая процедура, решающая описанную задачу идентификации.

В [1] излагаются различные эвристические подходы, использовавшиеся в задачах биологического определения. Выдвигается собственная методика, для применения которой необходим сравнительно большой объем экспертных знаний. Исследования Свиридова по моделированию процесса определения говорит в пользу предлагаемой им методики как достаточно надежной.

Авторы [2] решают задачу классификации за минимальное число шагов, не учитывая априорного распределения вероятностей появления объектов различных классов на входе определителя, — классы считаются равновероятными. Приводится математически точное оптимальное решение задачи. К сожалению, на практике часто встречаются задачи достаточно объемные, чтобы данный метод не мог их решить, в силу его экспоненциальной сложности.

Автор [3] основывается на постановке задачи, сходной с описанной в [2], однако классам сопоставляются определенные вероятности. Ставится задача поиска оптимального дерева решений для данного набора классов с сопоставленными им вероятностями и признаков с конечными наборами значений. Она решается экспоненциальным от числа признаков алгоритмом.

Разрабатываемая авторами в рамках сотрудничества с ЗИН РАН система [4] является политомическим определителем, работает интерактивно, предоставляя пользователю на каждом шаге возможность выбора признака из упорядоченного набора. Признаки упорядочиваются по значимости для определения (которую можно называть также релевантностью).

В этой работе уделим основное внимание постановке задачи, предложим наиболее адекватную на взгляд авторов функцию стоимости, опишем известные методы поиска оптимального (или субоптимального) решения.

2. Основные предположения

Пусть задано некоторое конечное множество A объектов, B классов, C признаков. Каждому признаку $c \in C$ поставим в соот-

ветствие конечное *множество состояний* s_c либо интервал значений $(0;1)$.

Определим вероятностное пространство на множестве объектов A. Оно порождает вероятности, заданные на классах B.

Далее будут рассмотрены две задачи оценивания, которые должен решать одновременно *идеальный* [4] определитель.

1. Задача определения неизвестных вероятностей p_{b_i}, $b_i \in B$. Требуется получить вектор оценок $\hat{\theta} = (\hat{p}_1, \ldots, \hat{p}_{|B|})^\tau$. На их основе следует делать выводы о предпочтениях между классами объектов. Учитывать вероятности $\hat{\theta}$ при сортировке признаков необходимо, это обеспечивает оптимизацию среднего числа измерений при работе пользователя с определителем.

Собирая статистику обращений и определяемых объектов, построим вектор $\hat{\theta}$ и будем обновлять его после каждого успешного завершения процесса идентификации. Такой подход, конечно, делает систему сильно зависимой от преимущественного способа ее использования. Но, поскольку предлагаемый далее алгоритм идентификации только использует данный вектор, можно воспользоваться вместо такой прямой статистики сглаженным или вообще равномерным распределением, если результаты получаются неудовлетворительными.

2. Задача выявления зависимостей между классом определяемого объекта и признаками, которые следует использовать при его идентификации за минимальное число шагов.

Шагом идентификации будем называть факт получения определителем значения или набора состояний некоторого признака (измерения y_i). Рассмотрим последовательность $\alpha = (c_i^\alpha)_{i=1}^m$, $m = |C|$, шагов, где c_i^α-признак, определяемый на шаге i. Идентификация считается законченной, когда полученный набор измерений (признаков) $S = \{y_i\}_{i=1}^{s_\alpha}$ однозначно задает некоторый класс, $s_\alpha \leq m$.

Сопоставим паре (b_i, c_j), $b_i \in B$, $c_j \in C$ класса и признака некий вес w_{ij}. Пусть веса для одного признака в сумме по всем классам дают 1. Смысл w_{ij} заключается в оценке, насколько хорош признак c_j для отделения объекта класса b_i от остальных классов из B. Обозначим как $\chi_{c_i}(s_j)$ множество классов, удовлетворяющих значению s_j признака c_i. Пусть $w_{ij} = \hat{P}\{s|b_i \notin \chi_c(s)\}$ - оценка вероятности того, что класс b_i окажется неподходящим после определения значения s некоторого признака c_j.

Введем показатель вариативности признака $H(c_i)$ — меру раз-

нообразия распределения множества его значений. Редукционные формулы [1] являются хорошими, эмпирически оправданными показателями вариативности. Рассмотрим упрощенную формулу Лобанова [5] для диагностической ценности признака

$$H^{(1)}(c) = \sum_{i=1}^{n_i} |\chi_c(s_i)|^{-2}$$

и характеристику диагностической ценности на основе энтропии Шеннона [1]:

$$H^{(2)}(c) = -\sum_{i=1}^{n_i} |\chi_c(s_i)| \ln |\chi_c(s_i)|.$$

Функция $H^{(1)}(c)$ дает необоснованно большое значение в простом случае, когда для всех классов ответами являются все состояния признака. Вторая формула также страдает от этого недостатка. Корректируя формулы, можно $|\chi_c(s_i)|$ заменить на

$$\sum_{s,t=1}^{|B|} |P\{b_s|s_i\} - P\{b_t|s_i\}|,$$

так как важно не абсолютное значение вероятности того или иного ответа, а распределение вероятностей получения ответов при условии принадлежности объекта определенному классу. Поэтому можно рассматривать энтропию распределений значений признака в рамках одного класса, затем рассчитывать среднюю энтропию по всем классам и считать эту величину вариативностью:

$$H^{(3)}(c_i) = \frac{1}{|B|} \sum_{b_i \in B} \sum_{s_j \in c_i} -P\{s_j|b_i\} \ln P\{s_j|b_i\}.$$

Так или иначе, расчет вариативности достаточно производить однократно, оперируя всеми классами множества B, так как в ходе определения при исключении классов из множества тех, которые могут содержать определяемый объект, будут изменяться соотношения весов w_{ij} между классами этого множества, и этого должно быть достаточно для выявления подходящих и неподходящих признаков для определения объекта данного класса.

Назовем *селективностью* признака c величину

$$S(c) = \frac{1}{H(c)} \sum_{i=1}^{|B|} w_{ij} P\{b_i\},$$

где $P\{b_i\}$ обозначает вероятность появления объекта из класса b_i на входе определителя. Будем сортировать признаки по убыванию их селективности, суммируя по текущему набору классов.

Число измерений признаков в ходе процесса идентификации является параметром, требующим оптимизации. Но зависимость этой величины от набора весов хотя и безусловно присутствует, но не является легко формализуемой. В связи с этим используем другой показатель, по смыслу ей аналогичный, но лучше укладывающийся в модель вычислений:

$$y_n = \sum_{c_j \in C_n^+} (1 - w_{i_n j}) + \sum_{c_i \in C_n^-} w_{i_n j}.$$

Здесь i_n — номер класса, к которому, как оказалось, принадлежит определенный объект, C_n^+ — множество признаков, оказавшихся полезными в процессе идентификации, и C_n^- — множество признаков, оказавшихся бесполезными. Разбиение $\{C_n^+, C_n^-\}$ образуется простым рандомизированным алгоритмом:

• в начале $C_n^+ = C$; $C_n^- = \emptyset$,

• повторять в цикле до тех пор, пока набор измерений признаков C_n^+ одноначно задает класс i_n,

•случайно выбрать $c \in C$ и переместить его в C_n^-.

3. Алгоритм

Поскольку приведенная выше модель содержит существенно многомерное множество параметров (w_{ij}), актуальным становится применение рандомизированных алгоритмов стохастической аппроксимации. Лучше всего в данном случае использовать рандомизированный алгоритм стохастической аппроксимации с двумя измерениями из [6]

$$\hat{W}_{i_n} = \hat{W}_{i_{n-1}} - \frac{\alpha_n}{\beta_n} \Delta_n (y_{2n} - y_{2n-1}), \tag{1}$$

где $\hat{W}_{i_n}$ — текущий набор оценок $(w_{i_n j})$, $\{\alpha_n\}$ и $\{\beta_n\}$ — некоторые последовательности положительных чисел, стремящиеся к нулю. При этом первое измерение y_{2n-1} будет получаться в результате взаимодействия с пользователем, второе же y_{2n} — в результате случайного эксперимента с рандомизированным выбором признаков $\hat{W}_{i_{n-1}} + \beta_n \Delta_n$.

Аддитивная помеха в измерениях v_n (см. [6]) может интерпретироваться как особенность пользователя, отвечающего на вопросы по признакам, — сложность понимания тех или иных признаков и сложность считывания признаков для объекта, который он идентифицирует. Вводя соответствующие метрики качества, можно получить и численную величину v_n. В этих случаях перечисленные свойства вполне можно считать независимыми или слабокоррелированными с порядком представления признаков и, следовательно, с выбранным вектором весов (w_{ij}).

Так как в модели наблюдений будет использоваться и измерение, использующее автоматический выбор признаков, для него тоже имеется помеха v_n. Особенности интеллекта, определяющие выбор признака для определения и существенно зависящие от смысла, заложенного в его формулировку, позволяют сделать более верный выбор признака. Таким образом этот недостаток автоматического определения и будет интерпретироваться как помеха v_n, опять же мало связанная с упорядочением признаков.

4. Пример использования

Пусть имеется 2 класса, 7 объектов, 2 признака. Рассмотрим исходные данные, необходимые для расчетов по изложенной выше методике. Пусть $\hat{\theta} = (1/3\ 1/3\ 1/3)^\tau$ — распределение вероятностей появления классов на входе определителя на основе имеющейся статистики. Рассмотрим матрицы m_1 и m_2, определяющие вероятности иметь некий ответ на один из двух признаков для объектов всех трех классов. У обоих признаков имеется три возможных ответа. Строки матриц соответствуют классам.

$$m_1 = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1/2 & 1/2 \\ 1 & 0 & 0 \end{pmatrix}, \quad m_2 = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 1/2 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

Из приведенных данных видно, что при использовании первым признака 2 можно определить один из объектов класса 3 за один шаг, в то время как при использовании первым признака 1 все объекты определяются за 2 шага.

Используем как показатель вариативности функцию $H^{(3)}$. При расчетах получаем, что первый признак имеет вариативность $h_1 = 1,79 < h_2 = 2,08$, и при равных весах w_{ij} он будет стоять в списке первым. Хотелось бы, чтобы первым был признак 2, так как иногда он способен сделать путь идентификации короче.

Изначально веса (w_{ij}) распределены равномерно между признаками для каждого класса (строки соответствуют признакам, столбцы — классам):

$$W = \begin{pmatrix} 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 \end{pmatrix}.$$

Используется алгоритм (1), где первое измерение делается "человеком" по первому признаку списка, второе же измерение делается по случайно выбранному из списка признаку. Последовательности выбраны как $\alpha_n = \frac{1}{5\sqrt{n+0.5}}$, $\beta_n = \frac{1}{4\sqrt{n}}$.

После 7 запусков алгоритма на имеющейся базе данных из 7 объектов матрица весов изменяется следующим образом:

$$W = \begin{pmatrix} 0.01 & 0.06 & 0.01 \\ 0.99 & 0.94 & 0.99 \end{pmatrix}.$$

При такой матрице весов признак 2 становится первым.

5. Заключение

Рассмотренный алгоритм включает в себя как частный случай расчет по редукционным формулам. В то же время он выделяет предпочтения по использованию признаков при определении объекта того или иного класса, что обеспечивает более интеллектуальное поведение определителя и позволяет за разумное время осуществлять сортировку признаков в случае существенно больших объемов данных, чем при непосредственном пересчете редукционных формул на каждом шаге. Приведенный выше пример демонстрирует применимость алгоритма в упрощенном, но показательном случае.

Однако, алгоритм не решает проблему выявления взаимной корреляции признаков. Действительно, если имеются два дизъюнктных набора признаков, являющихся минимальными путями для

определения некоторого класса, то алгоритм выделит признаки этих обоих подмножеств как важные, однако не сможет подсказать пользователю, что если он воспользовался одним из путей, то признаки из другого пути выбирать не следует: они по-прежнему будут вверху списка, так как имеют высокие веса.

Таким образом авторы видят возможность для дальнейшего улучшения приведенного здесь алгоритма.

Список литературы

[1] *Свиридов А. В.* Ключи в биологической систематике: теория и практика. М.:Издательство Московского Университета. 1994. 224 с.

[2] *Valdes-Perez R., Pericliev V., Pereira F.* Concise, Intelligible, and Approximate Profiling of Multiple Classes.

[3] *Владимирович А. Г.* Субоптимальный алгоритм распознавания образов в дискретном случае // Стохастическая оптимизация в информатике. под ред. О. Н. Граничина. СПб.: Изд-во СПбГУ. 2005. С. 8–16.

[4] *Лобанов А. Л., Кирейчук А. Г., Смирнов И. С., Граничин О. Н., Вахитов А. Т., Дианов М. Б.* К реализации идеального интерактивного определителя биологических объектов в Интернете // Труды Всероссийской научной конференции "Научный сервис в сети Интернет: технологии параллельного программирования". Новороссийск. 18-23 сентября 2006 г. Изд-во МГУ. 2006. С. 202–204.

[5] *Лобанов А. Л.* Принципы построения определителей насекомых с использованием электронных вычислительных машин. Автореферат диссертации на соискание ученой степени канд. биол. наук. Л.: ЗИН АН СССР. 1983. 19 с.

[6] *Граничин О. Н., Поляк Б. Т.* Рандомизированные алгоритмы оценивания и оптимизации при почти произвольных помехах. М.: Наука. 2003. 291 с.