

Н. Г. Загоруйко, И. А. Борисова, В. В. Дюбанов, О. А. Кутненко

**ФУНКЦИИ КОНКУРЕНТНОГО СХОДСТВА В АЛГОРИТМАХ
РАСПОЗНАВАНИЯ КОМБИНИРОВАННОГО ТИПА***

Для решения комбинированных задач распознавания образов предлагаются методы, основанные на функции конкурентного сходства (FRiS-функции). С ее помощью можно оценивать сходство между объектами, строить решающие правила, оценивать компактность образов и информативность признаков. Приводятся примеры использования предлагаемых методов для решения задач распознавания и прогнозирования.

Ключевые слова: функция конкурентного сходства, распознавание, компактность, информативность.

В задачах распознавании для принятия решения о принадлежности контрольного объекта z к образу A недостаточно знать расстояние $R(z, A)$. Нужно знать также расстояние $R(z, B)$ до конкурирующего образа B и определить, что $R(z, A)$ меньше $R(z, B)$. В методе k ближайших соседей (kNN) эти расстояния сравниваются в шкале порядка. Но оценками сходства можно распорядиться более эффективно, если измерять сходство в более сильной абсолютной шкале. Сформулируем следующие требования, которым должна удовлетворять мера $F(a, b|z)$ сходства объекта z с объектом a в конкуренции с объектом b .

Свойство локальности. Мера сходства должна зависеть не от характера распределения всего множества объектов, а от особенностей распределения объектов в окрестности объекта z .

Свойство нормированности. Если оценивается мера сходства объекта z с объектом a , и ближайшим соседом z является объект b , $b \neq a$, то при совпадении объектов z и a мера $F(a, b|z)$ должна иметь максимальное значение, равное $+1$, а при совпадении z с b – минимальное значение, равное -1 . Во всех остальных случаях мера конкурентного сходства принимает значения от -1 до 1 . При одинаковых расстояниях $R(z, a)$ и $R(z, b)$ объект z в равной степени будет похожим на объекты a и b , и $F(a, b|z) = F(b, a|z) = 0$.

Свойство инвариантности. Значения $F_{a/b}(z)$ и $F_{b/a}(z)$ должны сохраняться при сдвиге начала и повороте координат, а также при умножении всех координат на одно и то же положительное число.

Предлагаемая нами функция конкурентного сходства FRiS (от Function of Rival Similarity) [1] удовлетворяет всем этим требованиям:

$$F(a, b|z) = (r_2 - r_1)/(r_2 + r_1), \quad (1)$$

где r_1 r_2 – расстояния $R(z, a)$ и $R(z, b)$ соответственно. Опыт показал, что использование этой меры сходства позволяет создавать эффективные алгоритмы решения задач распознавания, легко объединяемые в алгоритмы комбинированного типа. Опишем некоторые из этих алгоритмов.

Оценка компактности и информативности. Практически все алгоритмы распознавания основаны

на использовании гипотезы компактности, имеющиеся определения которой не позволяют найти ее количественной оценки. Мы предлагаем использовать для этой цели FRiS-функцию. Для каждого объекта a_i образа A , $i = 1, 2, \dots, M_A$ найдем оценки сходства с ним всех M_A объектов a_j этого образа и оценки отличия от a_i всех M_B объектов b_s конкурирующего образа B . При определении сходства объектов a_j с a_i конкурентом будет служить объект b_j образа B , который является ближайшим соседом объекта a_j . Для оценки отличия от a_i объекта b_s образа B нужно найти объект b_q , который является ближайшим соседом объекта b_s . Расстояние от b_s до b_q принимается равным r_1 , от b_s до a_i – равным r_2 , по ним вычисляется мера сходства объекта b_s с объектом b_q в конкуренции с объектом a_i , и эта величина принимается в качестве меры отличия b_s от a_i .

Просуммировав оценки сходства объекта a_i со всеми своими объектами и оценки отличия от всех чужих объектов и разделив полученную сумму на величину $(M_A + M_B)$, мы получим среднюю оценку C_i его сходства со своими и отличия от чужих объектов. Повторив эти процедуры для всех объектов образа A , мы можем найти оценку компактности C_A этого образа:

$$C_A = (1/M_A) \sum_{i=1}^{M_A} C_i. \quad (2)$$

Общая оценка C FRiS-компактности K образов может быть получена путем геометрического усреднения оценок C_i :

$$C = \sqrt[K]{\prod_{i=1}^K C_i}. \quad (3)$$

Описанная мера компактности тем больше, чем выше плотность объектов внутри образов и дальше образы отстоят друг от друга. Таким же свойством обладает и мера, предложенная Фишером для оценки информативности признаков при нормальном распределении образов. Вполне естественным является использование FRiS-компактности в качестве критерия информативности признаков при произвольных распределениях образов. Она используется в этом качестве в алгоритме GRAD [2].

* Работа выполнена при финансовой поддержке РФФИ, проект № 08-01-00040.

Наши эксперименты с этим критерием показали его существенное преимущество по сравнению с широко используемым критерием минимума ошибок при распознавании тестовой выборки методами Cross Validation или One Leave Out.

Выбор эталонов (алгоритм FRiS-Stolp). Для распознавания контрольного объекта используются меры его сходства с эталонными представителями («столпами») каждого образа. Алгоритм FRiS-Stolp [1] выбирает эталоны, которые обладают высокими значениями двух свойств: «обороноспособности» – высокая степень сходства с ним других объектов этого образа позволяет надежно распознавать свои объекты; «толерантности» – низкая степень сходства с ним объектов других образов предотвращает распознавание чужих объектов в качестве своих.

Алгоритм FRiS-Stolp состоит из следующих шагов.

1. Оценивается объект a_i в роли столпа. Для этого, как и в предыдущем пункте, вычисляем сходство с ним всех M_A объектов образа A и несходство с ним (отличие) всех M_B объектов образа B . Находим среднее значение этих величин C'_i .

2. После того, как все объекты образа A побывают в роли столпа, выбирается тот объект a_i , который набрал наибольшую величину C'_i . Он объявляется столпом первого кластера. В состав этого кластера включаются m_1 объектов образа A , сходство которых со столпом превышает заданный порог F^* , например, $F^* = 0$. Среднее значение сходства m_1 объектов со своим столпом (C_1) служит мерой компактности полученного кластера.

3. Если $m_1 < M_A$, то для остальных объектов повторяются шаги 1–2. В итоге получаем список k_A столпов образа A с указанием величин $C_j, j = 1, 2, \dots, k_A$, накопленных каждым столпом.

4. Повторяем шаги 1–4 для объектов образа B и получаем список k_B столпов этого образа и оценки их компактности $C_q, q = 1, 2, \dots, k_B$.

5. Если образов больше, чем два, то при определении компактности каждого t -го образа, $t = 1, 2, \dots, K$, будем считать его образом A , а объекты всех остальных образов объединим в образ B .

Алгоритм FRiS-Stolp при нормальных распределениях, в первую очередь, выбирает столпы, расположенные в районе математического ожидания. Если распределения полимодальны и образы линейно неразделимы, то столпы будут стоять в центрах мод. В процессе распознавания, решение принимается в пользу того образа, на столп которого контрольный объект похож больше всего, а значение функции сходства объекта с выбранным образом позволяет судить о достоверности принятого решения.

Построение решающих правил с одновременным выбором признаков. Рассмотрим следующие алгоритмы.

Алгоритм AdDel [2]. Пусть из N признаков выбирается наиболее информативная подсистема, состоящая из n признаков. Хорошо известны два жадных полиномиальных алгоритма решения этой задачи. Ал-

горитм Deletion (Del) начинает работать со всеми N признаками и последовательно на каждом шаге вычеркивает из системы наименее полезный признак, пока в системе не останется n признаков. Алгоритм Addition (Ad), напротив, вначале находит самый информативный признак и на каждом из n шагов добавляет в систему тот признак, участие которого в системе делает ее наиболее информативной. Оба описанных алгоритма дают оптимальное решение на каждом шаге, но это не обеспечивает глобального оптимального решения. В алгоритме AdDel методом Ad набирается некоторое количество n_1 информативных признаков, затем n_2 из них ($n_2 < n_1$) исключается методом Del. Такое чередование алгоритмов Ad и Del продолжается до достижения заданного количества признаков n . Наблюдения показывают, что по мере увеличения числа признаков компактность вначале растет, потом рост прекращается и начинается ее снижение за счет добавления малоинформативных, шумящих признаков. Перегиб кривой качества позволяет автоматически выбрать *оптимальное количество признаков*. Это очень важное свойство алгоритмов семейства AdDel, которым не обладают другие алгоритмы выбора информативных признаков.

Алгоритм GRAD [2]. Добавлять и исключать признаки можно как по одному, так и группами (гранулами), состоящими из нескольких признаков. В наших экспериментах было обнаружено, что если все признаки упорядочить по убыванию их индивидуальной информативности, то в составе наиболее информативных пар преобладают признаки с малыми порядковыми номерами. В алгоритме GRAD (GRanulated AdDel) алгоритм AdDel работает на множестве G наиболее информативных гранул, состоящих из w признаков, $w = 1, 2, 3$. Гранулы мощности 1 – это признаки, занимающие первые m мест по индивидуальной информативности. Из них методом полного перебора выбираются гранулы мощности 2 и 3. Затем весь список из m самых информативных гранул подается на вход алгоритма AdDel. Сравнительные эксперименты показали, что алгоритм GRAD по качеству получаемых решений значительно лучше алгоритма AdDel.

Алгоритм FRiS-GRAD [2]. Этот алгоритм одновременно выбирает информативные признаки и в пространстве этих признаков строит решающее правило. Признаки выбираются алгоритмом GRAD, а решающие правила – алгоритмом FRiS. Эффективность этого алгоритма комбинированного типа подтверждена опытом решения большого количества реальных задач.

Примеры приложений. Рассмотрим примеры приложений.

Распознавание двух видов лейкемии (ALL и AML). Задача распознавания двух типов лейкемии интересна тем, что в литературе представлены результаты ее решения разными группами исследователей. Это дает возможность сравнить наши результаты с лучшими прежними результатами. Анализируемые данные представлены матрицей векторов экспрессии генов,

полученных с помощью биочипов для пациентов с двумя типами лейкемии – ALL и AML [3]. Обучающая выборка содержит 38 объектов, тестовая выборка – 34 объекта. Исходное количество признаков (генов) $N = 7\,129$.

Рассмотрим результаты решения этой задачи, описанные в работе [3]. Информативное подмножество признаков выбиралось методом RFE (разновидностью алгоритма Deletion), решающие правила основаны на методе SVM. Были найдены наилучшие подсистемы, размерность которых кратна степени числа 2: 4 096, 2 048, ..., 4, 2 и 1. По двум лучшим признакам, которые можно выбрать по результатам обучения, правильно распознано 30 объектов, по 4 лучшим признакам – 31, по 128 признакам – 33. Нами на тех же данных получены следующие результаты. Информативное подмножество признаков выбиралось с помощью алгоритма FRiS-GRAD. Информативность признаков оценивалась по критерию FRiS-компактности. Из 7 129 признаков было выбрано 18, из которых программа FRiS-Stolp построила 30 вариантов решающих правил. В состав каждого правила входит с разными весами от 4 до 6 признаков. Первые 27 правил дают результат 34 из 34.

Различие между приведенными результатами могут зависеть как от метода выбора признаков, так и от типа решающих правил. Для сравнения решающих правил SVM и FRiS был проведен такой эксперимент. В подпространстве двух признаков (генов 803 и 4 846), выбранных методом FRE, по правилу SVM было получено 30 правильных ответов, а FRiS-методом – 33. По лучшему одному гену (4 846), выбранному методом FRE, результат SVM равен 27, а результат FRiS – 30. Отсюда можно сделать вывод, что как метод выбора признаков, так и решающие правила, основанные на FRiS-функции, обладают достаточно высокими конкурентными качествами.

Прогнозирование покупательского спроса. Достаточно успешно оказалось применение FRiS-функции при решении задачи на международном конкурсе Data Mining Cup 2009. Задача состояла в предсказании значений переменных, измеренных в абсолютной шкале, и заключалась в следующем. Анализировались данные о том, сколько книг того или иного жанра было продано в разных магазинах в течение года. Эти данные представляли собой очень разрежен-

ную таблицу (84 % клеток таблицы имели значение 0), в которой M -строками (объектами) являлись магазины ($M = 4\,812$), а N -столбцами (признаками) – жанры книг ($N = 1\,864$). На пересечении строк и столбцов указывалось количество книг данного жанра, проданных в течение года в том или ином магазине. Последние 8 признаков являлись целевыми. Таблица была разделена по горизонтали на два слоя. В первом (обучающем) слое содержалось $M_o = 2\,394$ магазина. Для 2 418 контрольных магазинов требовалось предсказать, сколько и каких из 8 жанров книг было продано в каждом из них. В конкурсе изъявили желание участвовать 618 команд из 164 организаций 42 стран. Эту задачу решила 321 команда и прислала свои результаты. Наша команда, используя алгоритм FRiS-Pro [4], заняла 4 место. Полученные результаты подтверждают возможность использования FRiS-функции в алгоритмах решения задач прогнозирования количественных переменных.

Таким образом, использование относительной меры сходства, учитывающей конкурентную обстановку, позволяет строить эффективные методы решения всех основных задач Data Mining, в том числе задач распознавания образов. Методы инвариантны к количеству образов, характеру их распределений и обусловленности обучающей выборки (соотношению между M и N).

Библиографические ссылки

1. Methods of Recognition Based on the Function of Rival Similarity / N. G. Zagoruiko [et al.] // Pattern Recognition and Image Analysis. 2008. Vol. 18. P. 1–6.
2. Attribute selection through decision rules construction (algorithm FRiS-GRAD) / N. G. Zagoruiko [et al.] // Pattern Recognition and Image Analysis : New Information Technologies : Proc. of 9th Intern Conf. Nizhni Novgorod. 2008. Vol. 2. P. 335–338.
3. Gene Selection for Cancer Classification using Support Vector Machines / I. Guyon [et al.] // Machine Learning. 2002. Vol. 46 (1–3). P. 389–395.
4. Дюбанов В. В. Использование FRiS-функции в алгоритмах предсказания количественных переменных (Алгоритм FRiS-Pro) // Знания-Онтологии-Теории (ЗОИТ-09) : тр. Всерос. конф. Новосибирск, 2009. Т. 2. С. 258–260.

N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov, O. A. Kutnenko

FUNCTIONS OF RIVAL SIMILARITY IN ALGORITHMS OF RECOGNITION OF COMBINED TYPE

For decision of combined tasks of pattern recognition we offer methods based on function of rival similarity (FRiS-function). With the help of the function we can estimate similarity between objects, build decision rules, estimate compactness of patterns and informativeness of attributes. Examples of use of the offered methods for decision of tasks of recognition and forecasting are given in the article.

Keywords: function of rival similarity, recognition, compactness, informativeness.

© Загоруйко Н. Г., Борисова И. А., Дюбанов В. В., Кутненко О. А., 2010