УДК 519.95

ИСПОЛЬЗОВАНИЕ ФУНКЦИИ КОНКУРЕНТНОГО СХОДСТВА ДЛЯ ПРОГНОЗИРОВАНИЯ КОЛИЧЕСТВЕННЫХ ПЕРЕМЕННЫХ

В. В. Дюбанов

Учреждение Российской академии наук
Институт математики им. С. Л. Соболева Сибирского отделения РАН,
630090, г. Новосибирск, просп. Академика Коптюга, 4
E-mail: vladimir.dyubanov@gmail.com

Предлагается метод использования функции конкурентного сходства (FRiS-функции) при решении задачи прогнозирования переменных, измеренных в сильных (количественных) шкалах. Описывается алгоритм прогнозирования FRiS-Pro и опыт его применения при решении задачи прогнозирования целевого признака, измеренного в абсолютной шкале (покупательского спроса) на международном конкурсе "Data Mining Cup 2009".

Ключевые слова: прогнозирование, ближайшие соседи, функция конкурентного сходства.

Введение. Функция конкурентного сходства (FRiS-функция (Function of Rival Similarity)) была введена для решения задач распознавания образов, в которых требуется предсказывать значение целевого признака, измеряемого в номинальной шкале [1, 2]. Кратко напомним суть FRiS-функции. В задачах распознавания обычно используются меры сходства, которые носят абсолютный характер и применяются для оценок сходства между контрольным объектом z и эталонами каждого образа в отдельности. Но легко убедиться, что восприятие человеком сходства носит относительный характер. Чтобы ответить на вопросы типа «близко — далеко?», «похож — не похож?», нужно знать ответ на вопрос «по сравнению с чем?». Для оценки сходства объекта z с эталоном первого образа необходимо определить не только расстояние r_1 до него, но и расстояние r_2 до того образа, который является его ближайшим конкурентом. Мера сходства является функцией от этих двух расстояний и может измеряться в шкале порядка, как это делается в методе ближайшего соседа (kNN). Но оказалось, что знанием величин r_1 и r_2 можно распорядиться более эффективно, если оценивать сходство не в шкале порядка, а в более сильной абсолютной шкале. Для этого можно использовать функцию F, которая называется FRiS-функцией:

$$F(z, 1 \mid 2) = (r_2 - r_1)/(r_2 + r_1).$$

Здесь величина $F(z, 1 \mid 2)$ характеризует сходство объекта z с эталоном первого образа в конкуренции с эталоном второго образа. Значения этой функции меняются в пределах от $+1$ до -1. При расстояниях $r_1 = r_2$ значения $F(z, 1 \mid 2) = F(z, 2 \mid 1) = 0$, что указывает на границу между образами.

Полезность применения FRiS-функции в задачах распознавания образов показана в ряде работ [1–3]. Целью данного исследования является демонстрация применения FRiS-функции при решении задачи прогнозирования целевого признака, измеренного в абсолютной шкале, т. е. хорошо известной задачи, решаемой методами регрессионного анализа. Разработанный для её решения алгоритм FRiS-Pro относится к классу методов заполнения пробелов в таблицах данных. В литературе этот класс известен под названиями "Inserting" и "Imputation" [4]. Методы заполнения пробелов, основанные на использовании закономерных связей между всеми элементами таблицы данных, называются «глобальными». Их описание можно найти в [4, 5]. Данная работа основана на «локальном» подходе,

при котором оценивается взаимосвязь между строками и столбцами таблицы в ограниченной окрестности заполняемого элемента [6].

В отличие от задач распознавания в задачах заполнения пробелов конкурирующие классы не задаются, здесь потребовалось создание виртуальных конкурирующих классов. Алгоритм FRiS-Pro был успешно применён при решении задачи прогнозирования на международном конкурсе "Data Mining Cup 2009".

Описание задачи прогнозирования. 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 представлены таблицей [7], в которой M строк (объектов) являются магазинами ($M = 4812$), а N столбцов (признаков) — жанрами книг ($N = 1864$). На пересечении строк и столбцов указывалось количество книг данного жанра, проданных за год в том или ином магазине. Количество таких продаж колебалось от 0 до 2300. Таблица сильно разрежена: 84 % клеток таблицы содержат нулевые значения. Последние восемь признаков целевые.

Таблица разделена по горизонтали на две части. В первой (обучающей) для $M_0 = 2394$ магазина указаны значения как описывающих, так и целевых признаков. Во второй (контрольной) для $M_k = 2418$ магазинов содержится информация только об $N - 8$ описывающих признаках. Требуется определить (угадать), сколько и каких книг из восьми жанров продано в каждом из контрольных магазинов, т. е. предсказать значения целевых признаков, измеренные в абсолютной шкале в 19344 ячейках матрицы размером 2418×8.

Решение задачи. Для прогнозирования каждой пустой ячейки матрицы можно использовать локальный метод заполнения пробелов, аналогичный методу ZET [5]. Основная идея алгоритма ZET состоит в следующем. Пусть клетка (i, j) с пробелом находится на пересечении j-го столбца и i-й строки. Выбирается подмножество из k «компетентных» строк, являющихся ближайшими соседями i-й целевой строки, и n «компетентных» столбцов, наиболее сильно коррелированных с j-м целевым столбцом. Строки и столбцы этой «компетентной подматрицы» используются для прогнозирования значения пробела с опорой на известный постулат естественной классификации: «Объекты, похожие по $(n-1)$-му признаку, обычно похожи и по n-му признаку». Каждый столбец и каждая строка подматрицы вырабатывают свои варианты значения пробела (подсказки). В отличие от алгоритма ZET веса подсказок в алгоритме FRiS-Pro оцениваются по величине функции конкурентного сходства. Кроме того, существенные отличия имеются и в стратегии выбора параметров алгоритма.

При решении данной задачи обучение и распознавание делалось для каждого целевого признака в отдельности. Таким образом, задачу можно рассматривать как состоящую из восьми независимых задач. Для предсказания значений j-го целевого признака используется своё подмножество из n наиболее информативных описывающих признаков, которое выбирается с помощью алгоритма направленного перебора GRAD [5]. При определении величины b_{ij}, расположенной на пересечении i-й строки и j-го признака, используется своё подмножество из k строк обучающей выборки, наиболее похожих на i-ю строку по этим описывающим признакам (k ближайших соседей). Значения параметров n и k, как и других параметров алгоритма, определяются в процессе обучения. Каждому v-му сочетанию значений параметров P_v и оптимальному для него подмножеству признаков X_v соответствует своё решающее правило D_v, которое вырабатывает свой вариант b_{ij}^v предсказываемой величины b_{ij}. Окончательное решение получается в результате взвешенного усреднения значений b_{ij}^v, предсказанных коллективом наиболее компетентных решающих правил.

Каждое отдельное решающее правило представляет собой пару $\langle P, X\rangle$, где X — некоторое множество описывающих признаков, а P — фиксированный набор возможных значений следующих параметров:

1) N_1 — способ нормировки значений строки (отсутствие нормировки, нормировка по максимальному или по среднему значению);

2) N_2 — способ нормировки значений описывающего столбца (отсутствие нормировки, нормировка по максимальному значению);

3) L — вид метрики ($L1$ или $L2$);

4) k — число соседей ($k = 1, 2, \ldots, 10$);

5) O — способ округления результата предсказания до целого числа (в бóльшую или в меньшую сторону).

При заданном v-м варианте параметров P_v, $v = 1, 2, \ldots, V$, и выбранном подмножестве признаков X_v прогнозирование варианта b_{ij}^v по решающему правилу $D_v = \langle P_v, X_v\rangle$ осуществляется таким образом:

$$b_{ij}^v = \Big[\Big(\sum_{t=1}^{k} F_t b_{tj}\Big)\Big/k\Big],$$

где F_t — весовой коэффициент t-й строки, $t = 1, 2, \ldots, k$; b_{tj} — j-й элемент t-й строки обучающей выборки.

В качестве весовых коэффициентов F_t для каждой из k строк, наиболее похожих на i-ю строку, используются значения FRiS-функции, рассчитываемые следующим образом:

1. Среди объектов обучающей выборки находятся $2k$ объектов (строк), которые являются ближайшими соседями анализируемой i-й строки в пространстве X_v описывающих признаков. Из них формируются два виртуальных класса — «свой» и «чужой». В свой класс включаются первые k ближайших соседей, а в чужой класс — следующие по порядку k ближайших соседей.

2. Расстояния от i-го объекта до каждого t-го из k своих ближайших соседей в отдельности играют роль расстояний r_{1t}, а за расстояние r_2 до класса-конкурента принимается среднее расстояние от i-го объекта до k объектов чужого класса. По этим расстояниям вычисляется значение FRiS-функции:

$$F_t = (r_2 - r_{1t})/(r_2 + r_{1t}).$$

Полученное значение F_t принимается в качестве веса своего t-го ближайшего соседа.

С учётом изложенного алгоритм FRiS-Pro построения решающих правил и принятия решений можно представить следующими процедурами:

1. Фиксируется v-й вариант сочетания значений параметров P_v, $v = 1, 2, \ldots, 360$.

2. Методом направленного перебора GRAD выбирается наилучший вариант подмножества признаков X_v.

3. Компетентность (предсказательная способность) правила $D_v = \langle P_v, X_v\rangle$ оценивается количеством правильных решений, полученных этим правилом на объектах обучающей выборки в режиме скользящего экзамена Cross Validation (CV) [8].

4. При том же варианте значений параметров P_v алгоритм GRAD запускается q раз, начиная с разных стартовых признаков. В результате находится q разных подпространств признаков и, следовательно, q разных решающих правил.

5. Процедуры 1–4 повторяются для всех 360 вариантов вектора параметров, в итоге генерируется $q \times 360$ решающих правил $\langle P, X\rangle$.

6. Из этих правил в ансамбль правил, по которым принимается коллективное решение, с помощью алгоритма GRAD выбирается наилучшее подмножество правил в количестве

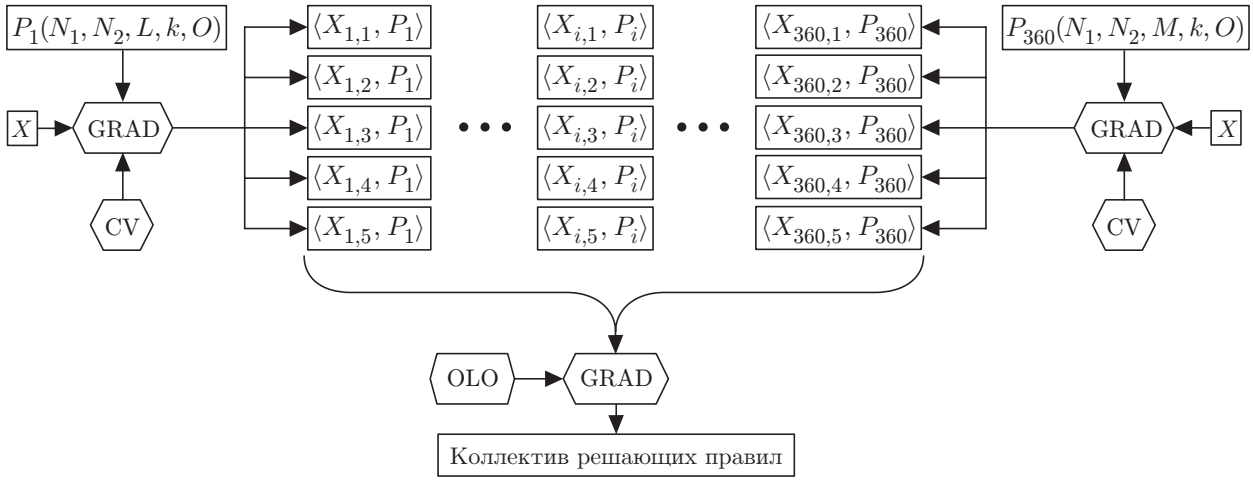

Схема работы алгоритма прогнозирования FRiS-Pro для случая $q = 5$

от 10 до 20. Некоторые из них могут оказаться в составе подмножества более 1 раза. Качество каждого такого ансамбля (предлагаемого GRAD) оценивается в режиме скользящего экзамена типа One Leave Out (OLO). Общее решение принимается путём взвешенного усреднения результатов, полученных правилами, входящими в ансамбль. При этом веса правил равняются величине их компетентности, помноженной на число вхождений правила в коллектив. Это один из видов коллективного решающего правила, обсуждаемых сейчас в литературе [9].

Описанный алгоритм представлен в виде схемы на рисунке.

Решение конкурсной задачи. Конкурсы "Data Mining Cup" проводятся ежегодно накануне Международной конференции по анализу данных и машинному обучению (KDD/ML) [7]. В 2009 г. в конкурсе изъявили желание участвовать 618 команд из 164 организаций 42 стран: 231 команда решила эту задачу и прислала результаты; 49 команд преодолели порог приемлемых результатов, установленный организаторами. Качество решения оценивалось суммой модулей разностей между фактическими и предсказанными значениями в каждой из 19344 ячеек исходной таблицы данных. Результаты первых де-

Рейтинг	Команда	Очки	Рейтинг	Команда	Очки
1	Uni Karlsruhe TH_II	17260	16	TU Graz	23626
2	TU Dortmund	17912	18	Uni Weimar_I	23796
3	TU Dresden	18163	19	Zhejiang University	23952
4	Novosibirsk State University	18353	20	University Laval	24884
5	University Karlsruhe TH_I	18763	24	University of Southampton	25694
6	FH Brandenburg_I	19814	25	Telkom Institute of Techn.	25829
7	FH Brandenburg_II	20140	26	University of Central Florida	26254
8	Hochschule Anhalt	20767	32	Indian Institute of Technology	28517
9	University of Hamburg	21064	34	Anna University Coimbatore	28670
10	KTH Royal Institute of Technol.	21195	38	Technical University of Kosice	32841
11	RWTH Aachen_I	21780	39	University of Edinburgh	45096
14	Budapest University of Technol.	23277	48	Warsaw School of Economics	77551
15	Isfahan University of Technol.	23488	49	FH Hannover	1938612

сяти и некоторых других команд приведены в таблице. Как видно, среднее отличие прогнозных значений для каждой ячейки от истинного у разных команд колебалось от 0,89 до 100,22. У пяти команд эти ошибки были меньше единицы. Команда Новосибирского университета сделала 0,95 ошибки на ячейку и заняла четвёртое место.

Заключение. Полученные результаты подтверждают возможность использования FRiS-функции не только в алгоритмах распознавания, связанных с предсказанием номинальных признаков, но и в алгоритмах прогнозирования количественных переменных. Высокие конкурентные качества предложенного алгоритма FRiS-Pro продемонстрированы результатами его сравнения с сотнями других алгоритмов, использованных участниками конкурса "Data Mining Cup 2009".

СПИСОК ЛИТЕРАТУРЫ

1. **Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.** Methods of recognition based on the function of rival similarity // Pattern Recogn. Image Analys. 2008. **18**, N 1. P. 1–6.
2. **Загоруйко Н. Г.** Интеллектуальный анализ данных, основанный на функции конкурентного сходства // Автометрия. 2008. **44**, № 3. С. 31–40.
3. **Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А.** Сходство и компактность // Тр. 14-й Всерос. конф. «Математические методы распознавания образов» (ММРО-14). Суздаль, 2009. С. 89–92.
4. **Литтл Р. Дж. А., Рубин Д. Б.** Статистический анализ данных с пропусками. М.: Финансы и статистика, 1991. 336 с.
5. **Лапко А. В., Лапко В. А.** Анализ непараметрических алгоритмов распознавания образов в условиях пропуска данных // Автометрия. 2008. **44**, № 3. С. 65–74.
6. **Загоруйко Н. Г.** Прикладные методы анализа данных и знаний. Новосибирск: Изд-во ИМ СО РАН, 1999. 270 с.
7. **Обзор** результатов конкурса "Data Mining Cup 2009".
URL: http://www.data-mining-cup.de/en/review/dmc-2009/ (дата обращения: 25.08.2010).
8. **Загоруйко Н. Г., Кутненко О. А.** Алгоритм GRAD для выбора информативного подпространства признаков // Вычислительные системы «Анализ структурных закономерностей». Новосибирск: Изд-во ИМ СО РАН, 2005. Вып. 174. С. 3–12.
9. **Kittler J., Hatef M., Duin R. P. W., Matas J.** On combining classifiers // IEEE PAMI. 1998. **20**, N 3. P. 226–239.

Поступила в редакцию 28 января 2010 г.