Модел. и анализ информ. систем. Т. **20**, № **3** (2013) 108–120

УДК 004.67

Применение нечеткой классификации для гибридных линейных методов прогнозирования

Таскин А. С., Миркес Е. М., Сиротинина Н. Ю.

Сибирский федеральный университет
660041, г. Красноярск, пр. Свободный, 79

e-mail: and0000@inbox.ru, mirkes@bk.ru, aasir@mail.ru

получена 14 января 2013

Ключевые слова: линейная регрессия, нечеткая классификация, гибридные методы прогнозирования

Статья посвящена проблеме прогнозирования для выборок с действительными признаками. Цель работы — оценить влияние порожденных бинарных признаков на точность прогнозирования линейной регрессии и гибридных линейных методов, основанных на кластеризации. Для этого исходный набор входных признаков выборки дополняется бинарными признаками, полученными из исходных посредством нечеткой классификации. Производится сравнительное тестирование рассматриваемых методов прогнозирования на исходной и полученной выборках. Результаты тестирования на трех различных базах данных показали, что для классической линейной регрессии использование порожденных признаков привело к существенному увеличению точности прогнозирования. Для линейной регрессии с кластеризацией методом k-means также наблюдалось увеличение точности прогноза, для линейной регрессии с кластеризацией методом knn — незначительное снижение, и неустойчивый результат — для двойной линейной регрессии.

Введение

Для решения задачи прогнозирования разработано и применяется множество методов, базирующихся на различных подходах: нейронные сети [1], деревья решений [2], линейная и нелинейная регрессия [3, 4] и др. Такое разнообразие объясняется тем, что ни один из существующих методов не является безусловно лучшим, каждый из них имеет свои преимущества в определенных областях.

Класс линейных методов прогнозирования является простым для понимания и вместе с тем эффективным. Известно эмпирическое наблюдение, согласно которому простые линейные методы чаще всего обеспечивают лучшую точность прогнозирования на вневыборочных данных по сравнению с нелинейными методами предположительно из-за общей устойчивости к неверной спецификации модели, к смещению и неточности при оценивании параметров модели, их структурным сдвигам и дрейфу [5].

Вместе с тем, точность линейных методов на всем множестве данных часто недостаточна. Повысить качество модели можно, используя предварительную обработку на основе анализа данных. В данной работе рассматриваются гибридные методы прогнозирования, основанные на линейной регрессии с предварительной кластеризацией данных, когда обучающее множество разбивается на подмножества (кластеры), для каждого из которых строится своя линейная регрессия [6, 7, 8, 9].

Было установлено, что дополнение исходного набора признаков бинарными признаками, полученными из исходных с использованием нечеткой классификации, позволяет повысить точность прогнозирования классической линейной регрессии [10].

В работе представлено экспериментальное исследование влияния порожденных бинарных признаков на точность прогнозирования линейной регрессии с предварительной кластеризацией для различных методов разбиения исходного множества на непересекающиеся подмножества. Было проведено исследование — как известных ранее (k-means [11, 12] и knn [13, 14]) методов разбиения, так и вновь разработанных (двойная линейная регрессия).

Постановка задачи

Цель работы — оценить влияние порожденных бинарных признаков и параметров нечеткой классификации (числа классов) на точность следующих методов прогнозирования:

- Классическая линейная регрессия;
- Линейная регрессия с кластеризацией методом knn;
- Линейная регрессия с кластеризацией методом k-means;
- Двойная линейная регрессия.

В качестве исходных данных имеется таблица A размерностью $N \times M$, где N — число объектов наблюдения, M — число признаков объектов. Строка таблицы является формализованным описанием некоторого объекта наблюдения. Каждый столбец таблицы — это некоторое свойство описываемых объектов. В данной работе рассматриваются задачи с признаками в непрерывной и номинальной шкалах.

Для простоты будем считать, что выборка A имеет один целевой признак. Примем обозначения: $x_1 \dots x_{M-1}$ — входные признаки выборки; y — целевой признак.

Исходная таблица данных A дополняется бинарными признаками[1]. Чтобы оценить влияние добавленных признаков на точность прогнозирования метода (алгоритма) α необходимо:

1. Построить модель данных $\mu(\alpha, A)$ по исходной выборке (или её части – обучающему множеству).
2. Используя полученную модель данных вычислить прогнозные значения целевого признака:
$$y_i = f(\mu, x_{1i}, \dots, x_{(M-1)i}),$$

[1]Процедура порождения бинарных признаков с использованием нечеткой классификации описывается в разделе «порождение бинарных признаков».

3. Экспериментально оценить точность прогнозирования при различном числе порожденных бинарных признаков. Точность прогнозирования оценивается по относительной ошибке прогнозирования:

$$e = \frac{\sum\limits_{i=1}^{N} \left| y_i - \bar{y}_i \right|}{\sum\limits_{k=1}^{N} \bar{y}_k}, \tag{1}$$

где $\bar{y}_i$ — истинное значение целевого признака i-го объекта, y_i — прогнозное значение целевого признака i-го объекта.

Классическая линейная регрессия

Линейная регрессия — это метод, позволяющий аппроксимировать зависимость между несколькими входными и одной выходной переменной линейной функцией зависимости. Модель линейной регрессии описывается гиперплоскостью.

Пусть j – номер вычисляемого признака. Запишем уравнение регрессии:

$$y = \sum_{\substack{i=1 \\ i \neq j}}^{M} b_i^j x_i + b_0^j, \tag{2}$$

где y — прогнозируемое значение, зависимая переменная, b_i^j — коэффициенты линейной регрессии, построенной для j-го признака, x_i — независимые переменные.

Коэффициенты уравнения линейной регрессии подбираются так, чтобы минимизировать целевую функцию, которая характеризует ошибку аппроксимации. В качестве таковой обычно используют сумму квадратов отклонения [4].

Линейная регрессия с предварительной кластеризацией

Для повышения точности прогнозирования методом линейной регрессии в ряде случаев применяется предварительная кластеризация данных.

В работе для кластеризации применяются известные методы классификации knn и k-means, а также разработанный метод двойной линейной регрессии.

Для методов k-means и knn исходное множество разбивается на подмножества (кластеры) на основе расположения точек в пространстве признаков. При этом исходят из предположения о том, что расположенные рядом в пространстве входных признаков объекты должны хорошо описываться линейной регрессией (лежат в окрестности одной гиперплоскости). В качестве меры близости между объектами (точками в пространстве признаков) используется евклидово расстояние:

$$dist(a_1, a_2) = \sqrt{\sum_i \left(a_1^i - a_2^i\right)^2}, \tag{3}$$

где a_j^i - значение i-го признака объекта a_j.

Следует учитывать, что для применения линейной регрессии необходимо, чтобы количество объектов кластера было больше количества признаков, характеризующих отдельный объект.

Порядок построения прогноза для всех методов следующий:

- определяется, к которому кластеру $\hat{c}$ относится целевой объект;
- по объектам кластера $\hat{c}$ строится линейная регрессия для целевого признака, по которой и определяется его прогнозное значение.

Кластеризация методом knn

Knn (k nearest neighbors, k ближайших соседей) — широко используемый алгоритм классификации. Его идея заключается в том, что целевой объект принадлежит к тому классу, представителей которого больше всего среди k ближайших соседей целевого объекта. Оптимальное значение k чаще всего подбирается. Для этого проводится дополнительное тестирование обучающего множества при различных значениях параметра k. Мерой близости для числовых данных чаще всего считают евклидово расстояние (3).

В данной работе для каждого целевого объекта формируется кластер с помощью алгоритма knn. Для этого в пространстве входных признаков вычисляется расстояние между целевым объектом и всеми объектами обучающей выборки и отбирается k его ближайших объектов-соседей. В связи с требованием превышения числа объектов над числом параметров в данной работе рассматривались k большие M.

Полученное таким образом подмножество объектов $\hat{c}$ — кластер — используется для вычисления прогноза целевого признака.

Кластеризация методом k-means

k-means – итерационный алгоритм кластеризации, основанный на разделении объектов согласно их расположению в пространстве входных признаков. При его использовании выполняется предварительная кластеризация всего обучающего множества по следующему алгоритму.

1. Задаются начальные ядра кластеров. Выбор их количества (если оно заранее неизвестно по условию задачи) и начального положения является довольно сложной задачей. Существуют различные способы её решения, например, многократное повторение разбиений со случайным выбором начальных ядер и модификация k-means++[15]. В данной работе использовался первый способ.

2. Выполняется формирование кластера таким образом, чтобы он состоял из ближайших к его ядру объектов.

$$c^i = \{a_j : dist(q^i, a_j) < dist(q^l, a_j)\}, \forall i \neq l, j,$$

где c^i — i-й кластер, a_j — j-й объект исходной выборки A, q^i — ядро i-го кластера.

3. Для каждого кластера вычисляется его ядро (центр масс) – вектор, элементами которого являются средние значения соответствующих входных признаков, вычисленных по всем объектам кластера:

$$q_s^i = \frac{1}{n^i} \sum_{j=1}^{n^i} c_{js}^i, i = 1 \dots k, s = 1 \dots (M-1),$$

где k — количество кластеров, $(M-1)$ — количество входных признаков, q^i — ядро i-го кластера, c^i_{js} — s-й признак j-го объекта i-го кластера, n^i — число объектов в i-м кластере.

При этом на каждой итерации происходит уменьшение суммарного уклонения, что обеспечивает сходимость алгоритма.

Шаги 2 и 3 повторяются до тех пор, пока состав кластеров не перестанет изменяться от итерации к итерации.

Считаем, что целевой объект относится к тому кластеру $\hat{c}$, расстояние до ядра которого является наименьшим.

Двойная линейная регрессия

В данной работе предлагается еще один способ кластеризации, названный методом двойной линейной регрессии, который разбивает обучающее множество на непересекающиеся подмножества. Его ключевое отличие от рассмотренных ранее методов состоит в том, что разбиение производится не на основе расположения точек в исходном пространстве, а на основе вычисления функции линейной регрессии.

Рассмотрим процедуру кластеризации:

1. Для целевого признака строится линейная регрессия по всему обучающему множеству (2).

2. Для каждого объекта a_i обучающего множества A посредством линейной регрессии вычисляется значение целевого признака y_i. Все объекты упорядочиваются по возрастанию вычисленного значения целевого признака:

$$A' = \{a_i : y_i \leq y_{i+1}, i = 1 \dots N\},$$

где y_i — вычисленное значение целевого признака объекта a_i.

3. Упорядоченное множество A' разбивается на k непересекающихся интервалов (кластеров)

$$A' = \bigcup_{i=1}^{k} c_i.$$

В данной работе значение k подбиралось эмпирически из диапазона от 2 до 10. Разбиение на число кластеров большее, чем 10, приводит к существенному увеличению вычислительной сложности эксперимента.

По каждому кластеру строится линейная регрессия для целевого признака. В результате данной процедуры линейная регрессия по всему обучающему множеству заменяется кусочно-линейной (ломаной) регрессией, построенной по отдельным кластерам.

Для оптимизации разбиения на кластеры границы интервалов подбираются таким образом, чтобы минимизировать суммарную ошибку отклонения истинных значений целевого признака объектов от прогнозных значений, вычисленных по полученным кластерам:

$$\sum_{i=1}^{k} \sum_{j} \left(y^i_j - \bar{y}^i_j\right)^2 \to min,$$

где k — количество кластеров, y_j^j — прогнозное значение целевого признака j-го объекта i-го кластера, $\bar{y}_j^i$ — истинное значение целевого признака j-го объекта i-го кластера.

В данной работе для поиска оптимального разбиения использовался полный перебор всех возможных положений границ интервалов.

Для учета возможного выхода за границы крайних (1-го и k-го) интервалов их границы расширялись:

$$y_1^1 = -\infty; y_n^k = +\infty;$$

где k — количество кластеров, y_j^i — прогнозное значение целевого признака j-го объекта i-го кластера, n — количество объектов в кластере.

Для построения прогноза определим, к какому интервалу (кластеру) относится целевой объект. Для этого вычислим промежуточное прогнозное значение целевого признака целевого объекта $\hat{y}$ с помощью линейной регрессии, построенной по всему обучающему множеству, и выберем тот кластер, в интервал прогнозных значений целевого признака которого попадает прогнозное значение целевого признака целевого объекта:

$$\hat{c} = \{c^i : y_1^i \leq \hat{y} < y_{n^i}^i\}, i = 1 \ldots k,$$

где k — количество кластеров, $\hat{y}$ — прогнозное значение целевого признака целевого объекта, y_j^i — прогнозное значение целевого признака j-го объекта i-го кластера, n^i — количество объектов в i-м кластере.

После этого вычисляется окончательное прогнозное значение целевого признака с использованием линейной регрессии, построенной по объектам выбранного кластера $\hat{c}$.

Порождение бинарных признаков

Было предложено добавить к набору непрерывных признаков дополнительно сформированные бинарные признаки. Для их формирования применялся подход, основанный на нечеткой классификации и нечетких множествах [16, 17]. Предполагается, что это позволит повысить точность прогнозирования методами, базирующимися на линейной регрессии.

Определим требования к нечеткой классификации.

Пусть необходимо построить k нечетких классов.

Пусть дан непрерывный признак $X = \{x_1, ..., x_n\}$, принимающий n значений. Построим нечеткий классификатор этого признака на k нечетких классов. Носителем i-го нечеткого класса будем называть интервал $W_i = [u_i, v_i)$ (рис. 1). Сами нечеткие множества будем обозначать так же, как их носитель. При этом объект со значением признака X равным x, принадлежит i-му нечеткому классу, если $x \in W_i$. Назовем нечеткие множества W_i и W_j соседними, если их номера отличаются на единицу.

Определим требования к нечетким классам.

1. Нечеткие классы должны покрывать все множество $X : X \subseteq \bigcup_{i=1}^{k} W_i$.

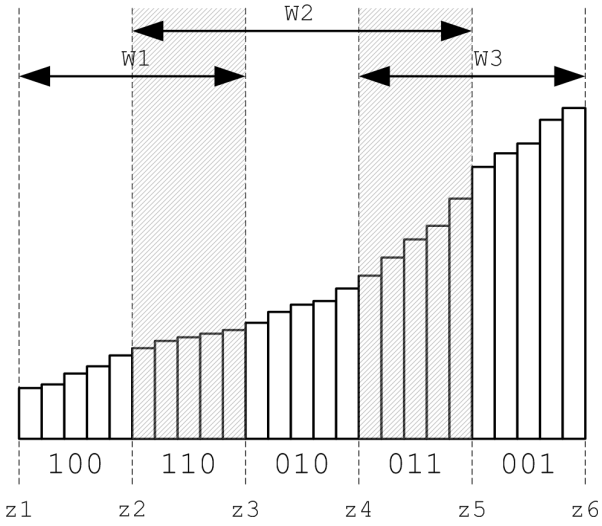

Рис. 1. Переход к бинарным признакам

2. Носители соседних нечетких классов должны пересекаться: $W_i \cap W_{i+1} \neq \emptyset$.

3. Носители не соседних нечетких классов не пересекаются $W_i \cap W_j \neq \emptyset, \forall (i,j) : |i-j| > 1$.

4. Носитель каждого нечеткого множества имеет фрагмент, не принадлежащий соседним множествам $(W_i \cap W_{i-1}) \cup (W_i \cap W_{i+1}) \neq W_i$.

5. Носители всех нечетких классов должны содержать одинаковое число точек множества X.

Для построения носителей нечетких множеств, удовлетворяющих предъявленным требованиям, проведем следующую процедуру.

1. Упорядочим значения признака X по возрастанию значений, получим множество $X = \{x_i, i = 1 \ldots n, x_i \leq x_{i+1}\}$.

2. Определим $2k$ точек z_i, разбивающих множество X на подмножества равной мощности. Мощность каждого подмножества будет равна

$$r = \frac{n}{2k-1}.$$

Значения точек z_i определим следующим образом:

$$z_1 = x_1; z_{2k} = x_n; z_i = \frac{x_{(i-1)r} + x_{(i-1)r+1}}{2}, i = 2, \ldots, 2k-1.$$

3. Дополним множество точек z_i двумя крайними точками, для учета возможного появления значений выходящих за границы интервала: $z_0 = 2z_1 - z_2$ и $z_{2k+1} = 2z_{2k} - z_{2k-1}$.

4. Определим границы интервала носителя каждого нечеткого множества по следующей формуле: $W_i = [u_i, v_i), u_i = z_{2i-2}, v_i = z_{2i+1}$.

По каждому входному признаку x_i формируются k бинарных признаков $\hat{x}_j^i, j = 1 \dots k$. Каждый бинарный признак $\hat{x}_j^i$ определяет список объектов, значения исходного признака x_i которых, попадает в соответствующий интервал W_j^i:

$$\hat{x}_j^i = \begin{cases} 1, & \text{если } x_i \in W_j^i; \\ 0, & \text{если } x_i \notin W_j^i. \end{cases}$$

Такой подход предполагает наличие во входных данных признаков, имеющих непрерывную шкалу.

Постановка вычислительного эксперимента

Тестирование проводилось на трех базах данных. База данных «Fried» [18] является искусственной выборкой. Входные признаки $(x_1 \dots x_{10})$ генерировались независимо друг от друга случайным образом из диапазона $(0, 1)$, а выходной — рассчитывался по формуле:

$$y = 10 \sin(\pi \cdot x_1 \cdot x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon,$$

где ε — шум с нормальным распределением из диапазона $(0, 1)$. База содержит 40768 объектов, 10 входных признаков и один выходной. Все признаки имеют непрерывные шкалы.

База данных «Statlog shuttle» предоставлена NASA [19] и содержит 58000 объектов и 10 признаков — 9 входных и один выходной. Выходной признак имеет номинальную дискретную шкалу. База данных разделена на обучающее множество (43500 объектов) и тестовое множество (13500) объектов.

База данных «Ископаемый уголь» предоставлена компанией Weatherford Laboratories. Она описывает зависимость химического состава образцов ископаемого угля от их физических характеристик. База данных содержит 6504 объекта и 12 признаков: 3 входных и 9 выходных.

Исходная выборка случайным образом разбивается на обучающую и тестовую выборку (для базы данных «Statlog shuttle» выборки были изначально разделены). В данной работе тестовое множество содержит 25% объектов выборки, а обучающее множество — 75%. По обучающему множеству строится модель данных, и с помощью этой модели рассчитываются прогнозы для каждого объекта тестового множества.

Тестируются все выходные признаки каждой базы данных. Выходные признаки тестируются независимо друг от друга.

Точность прогнозирования оценивается по относительной ошибке прогнозирования (1).

Исследуется зависимость точности прогноза от количества порождаемых бинарных признаков — количества нечетких классов. Каждый входной признак выборки порождает 3, 5, 7 или 9 бинарных признаков (при большем количестве порождаемых признаков существенно возрастают вычислительные затраты на построение модели данных). Действительные и бинарные признаки используются совместно.

Результаты вычислительного эксперимента

Результаты вычислительных экспериментов представлены в таблице 1.

Таблица 1. Относительная ошибка прогнозирования, %

Метод решения	L	Ископаемый уголь									Statlog shuttle	Fried
		Влажность	Летучее вещество	Связанный углерод	Зола	Водород	Углерод	Азот	Кислород	Сера		
ЛР	0	68,65	13,13	12,43	25,86	2,64	12,29	19,75	41,88	62,44	29,58	14,21
	3	51,08	10,75	11,05	23,94	2,01	10,11	17,26	31,97	57,81	17,96	10,27
	5	48,03	10,32	10,94	23,89	1,88	9,65	16,48	29,82	57,90	16,31	9,86
	7	44,50	9,78	**10,78**	**23,71**	1,74	**9,34**	**15,87**	**28,11**	**57,13**	12,50	9,80
	9	**44,27**	**9,69**	10,87	23,94	**1,73**	9,40	16,10	28,25	57,62	**11,34**	**9,73**
knn ЛР	0	45,94	10,21	**10,53**	**23,26**	1,66	9,51	**16,62**	30,04	**56,33**	0,58	**7,67**
	3	**40,53**	**9,54**	10,90	24,31	**1,50**	**9,36**	16,82	**27,82**	58,01	0,71	20,77
	5	44,67	10,68	12,16	27,13	1,65	10,32	18,61	30,28	63,53	**0,39**	55,70
	7	46,04	10,63	11,97	26,65	1,69	10,31	18,69	31,19	65,33	0,46	83,75
	9	48,69	11,08	13,25	29,60	1,82	11,34	19,55	32,56	67,37	0,48	108,41
k-means ЛР	0	59,37	12,04	11,12	24,29	2,19	10,87	18,50	36,32	56,11	4,02	9,96
	3	**39,61**	**9,31**	**10,31**	**23,02**	**1,44**	**8,82**	**15,68**	**26,61**	**54,72**	2,96	**9,01**
	5	40,99	9,64	10,62	24,09	1,49	9,24	16,47	27,42	56,38	3,04	9,06
	7	43,43	10,01	11,25	25,16	1,59	9,61	17,19	28,92	59,73	**1,05**	9,46
	9	45,85	10,70	11,83	26,63	1,76	10,58	17,92	30,67	62,35	1,21	9,99
Двойная ЛР	0	**37,23**	**9,00**	10,71	23,84	1,40	9,57	17,03	**25,75**	60,92	6,13	13,53
	3	37,51	9,49	10,39	**22,60**	1,40	9,04	15,80	25,82	**55,08**	2,64	9,70
	5	39,25	9,68	**10,38**	23,47	**1,38**	**9,01**	**15,76**	26,73	55,15	3,75	9,40
	7	41,36	9,52	10,95	23,50	1,44	9,69	16,14	27,04	56,03	**0,94**	9,39
	9	39,27	9,72	11,07	23,66	1,46	9,54	16,60	27,21	57,87	1,09	**9,28**

Результаты были получены при следующих параметрах методов:

- Для всех гибридных методов в каждом эксперименте подбиралось такое количество кластеров (количество объектов при кластеризации методом knn) k, при котором ошибка прогнозирования минимальна.
- Для всех гибридных методов минимальный размер кластера равен количеству признаков выборки M.
- Линейная регрессия (как отдельный метод и как этап гибридных):
  - Количество слагаемых уравнения регрессии (длина линейной регрессии) равно количеству входных признаков выборки $M - 1$.
- Двойная линейная регрессия
  - Величина шага при подборе границы кластеров: 10

Использованы обозначения: ЛР — классическая линейная регрессия; knn ЛР — линейная регрессия с кластеризацией методом knn; k-means ЛР — линейная регрессия с кластеризацией методом k-means; Двойная ЛР — двойная линейная регрессия; L — число порожденных бинарных признаков (нечетких классов).

Лучший результат для каждого эксперимента выделен жирным шрифтом.

Для удобства интерпретации результатов сведем их в таблицу (табл. 2).

Эффективность применения дополнительных бинарных признаков рассчитывалась как среднее от отношения ошибки, полученной при заданном числе нечетких классов, к ошибке, полученной без использования нечетких классов; чем меньше значение, тем лучше.

Таблица 2. Эффективность применения дополнительных бинарных признаков

Метод решения	L	Количество случаев, в которых лучший результат прогнозирования был получен при заданном количестве бинарных признаков	Средняя эффективность применения дополнительных бинарных признаков	Средняя эффективность применения дополнительных бинарных признаков по лучшим случаям
ЛР	0	0	1,00	–
	3	0	0,80	–
	5	0	0,77	–
	7	6	0,74	0,82
	9	5	0,73	0,62
knn ЛР	0	5	1,00	1,00
	3	5	1,15	0,93
	5	1	1,60	0,67
	7	0	1,95	–
	9	0	2,31	–
k-means ЛР	0	0	1,00	–
	3	10	0,82	0,82
	5	0	0,84	–
	7	1	0,84	0,26
	9	0	0,89	–
Двойная ЛР	0	3	1,00	1,00
	3	2	0,90	0,90
	5	4	0,93	0,96
	7	1	0,91	0,15
	9	1	0,91	0,69

Анализ таблицы 2 позволяет сделать следующие выводы.

1. Для линейной регрессии и линейной регрессии с кластеризацией методом k-means расширение пространства входных признаков за счет нечеткой классификации всегда дает положительный эффект: удается улучшить результат в среднем на 27% и 18% соответственно. Исходя из полученных результатов, для классической линейной регрессии следует использовать порядка 7 классов в нечеткой классификации, а для линейной регрессии с кластеризацией методом k-means следует использовать 3 класса.

2. При использовании классификации на основе метода knn расширение пространства входных признаков за счет нечеткой классификации малоэффективно. Более того, для данного метода применение нечеткой классификации для порождения дополнительных входных признаков в среднем приводит к ухудшению результатов.

3. Для метода двойной регрессии расширение пространства входных признаков дает самую смазанную картину. Причина этого достаточно проста. На этапе первичной регрессии, которая служит основанием для кластеризации, эффект от расширения множества входных параметров сказывается достаточно сильно. Вторичное использование того же приема при построении линейной регрессии внутри каждого класса приводит к дополнительному выигрышу. Для данного метода невозможно сформулировать рекомендации по числу используемых классов.

Заключение

Было исследовано влияние порожденных бинарных признаков на точность прогнозирования линейной регрессии и гибридных линейных методов, основанных на кластеризации. Для классической линейной регрессии использование таких признаков привело к существенному увеличению точности прогнозирования – в среднем на 27%. Для гибридных методов результат неоднозначный: повышение точности для линейной регрессии с кластеризацией методом k-means, снижение – для линейной регрессии с кластеризацией методом knn и переменный результат для двойной линейной регрессии.

Список литературы

1. Haykin S. Neural Networks and Learning Machines. New York: Prentice Hall, 2009.

2. Левитин А.В. Алгоритмы: введение в разработку и анализ. М.: Вильямс, 2006. (Levitin A.V. Algoritmy: vvedenie v razrabotku i analiz. Moskva.: Vilyams, 2006 [in Russian].)

3. Motulsky H., Christopoulos A. Fitting models to biological data using linear and non-linear regression. A practical guide to curve fitting. Oxford: UniversityPress, 2004.

4. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М.: Вильямс, 2007. (Draper N.R., Smith H. Applied regression analysis. New York: Wiley, 1998.)

5. Ицхоки О. Выбор модели и парадоксы прогнозирования // Квантиль. 2006. №1. С. 43–51. (English transl.: Itskhoki O. Model selection and paradoxes of prediction // Quantile. 2006. No. 1. P. 43–51.)

6. Tondel K., Indahl U., Gjuvsland A., Vik J. et al. Hierarchical Cluster-based Partial Least Squares Regression (HC-PLSR) is an efficient tool for metamodelling of nonlinear dynamic models // BMC Systems Biology, 2011. V. 5(90).

7. Camps-Valls G. et al. Cyclosporine concentration prediction using clustering and support vector regression methods // Electronic Letters. 2002. V. 38(12). P. 568–570.

8. Ari B., Guvenir H.A. Clustered linear regression // Knowledge-Based Systems. 2002. V. 15(3). P. 169–175.

9. Таскин А.С., Неволина С.С. Применение предварительной кластеризации при заполнении пробелов в таблицах данных // Сборник трудов VIII Всероссийской научно-практической конференции студентов, аспирантов и молодых ученых «Молодежь и современные информационные технологии», 2010. Ч. 1. С. 223–224. (Taskin A.S., Nevolina S.S. Primenenie predvaritelnoy klasterizatsii pri zapolnenii probelov v tablitsakh dannykh // Sbornik trudov VIII Vserossiyskoy nauchno-prakticheskoy konferentsii studentov, aspirantov i molodykh uchenykh «Molodezh i sovremennye informatsionnye tekhnologii», 2010. Part 1. P. 223–224 [in Russian].)

10. Таскин А.С., Миркес Е.М. Линейная регрессия с кластеризацией по признаку на данных с действительными величинами // Вестник СибГАУ. 2012. Вып. 3(43). С. 71–75. (Taskin A.S., Mirkes E.M. Lineynaya regressiya s klasterizatsiey po priznaku na dannykh s deystvitelnymi velichinami // Vestnik SibGAU. 2012. V. 3(43). P. 71–75 [in Russian].)

11. Gorban A.N., Zinovyev A.Y. Principal Graphs and Manifolds, Ch. 2 // Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI Global, Hershey, PA, USA, 2009.

12. Gan G., Ma C., Wu J. Data Clustering: Theory, Algorithms, and Applications. SIAM, Philadelphia, ASA, Alexandria, VA, 2007.

13. Abidin T., Perrizo W. SMART-TV: A Fast and Scalable Nearest Neighbor Based Classifier for Data Mining // Proceedings of ACM SAC-06, 2006. P. 536–540.

14. Zhang J., Mani I. kNN approach to unbalanced data distributions: A case study involving Information Extraction // Workshop on learning from imbalanced datasets II, ICML, 2003.

15. David Arthur, Sergei Vassilvitskii. k-means++: the advantages of careful seeding // Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07), 2007. P. 1027–1035.

16. Орлов А.И. Прикладная статистика. М.: Экзамен, 2004. (Orlov A.I. Prikladnaya statistika. Moskva: Ekzamen, 2004 [in Russian].)

17. Abonyi J., Feil J. Cluster Analysis for Data Mining and System identification. Basel: Birkhauser, 2007.

18. Bilkent University. Function Approximation Repository [Электронный ресурс]. URL: http://funapp.cs.bilkent.edu.tr.

19. UCI Machine Learning Repository [Электронный ресурс]. URL: http://archive.ics.uci.edu/ml.

Application of the Fuzzy Classification for Linear Hybrid Prediction Methods

Taskin A. S., Mirkes E. M., Sirotinina N. Y.

Siberian Federal University
79, Svobodny Prospect, Krasnoyarsk, 660041, Russia

Keywords: linear regression, fuzzy classification, hybrid prediction methods

The paper discusses the problem of forecasting for samples with real-valued attributes. The goal is to estimate the effect of generated binary attributes on forecasting accuracy for the linear regression and the hybrid methods based on clustering. The initial set of attributes is expanded by binary attributes which are derived from the initial set by fuzzy classification. A comparative testing of the discussed forecasting methods on the initial samples and the resulting ones is performed. The test results on three different databases showed that the use of generated attributes for the classical linear regression resulted in the significant increase of the forecasting accuracy. In case of the linear regression with the clustering based on k-means the increase of forecasting accuracy was also observed. In case of the linear regression with the clustering based on the knn–method we registered a slight decrease, and an unstable result was obtained for the double linear regression.

Сведения об авторах:
Таскин Андрей Сергеевич,
Сибирский федеральный университет, аспирант
Миркес Евгений Моисеевич,
Красноярский институт железнодорожного транспорта,
д-р техн. наук, профессор
Сиротинина Наталья Юрьевна,
Сибирский федеральный университет, канд. техн. наук, доцент