Прогнозирование на основе отбора предикторов из большого числа сильно коррелированных переменных

А.Ю. Тимофеева[1], Ю.А. Мезенцев[1]

[1]Новосибирский государственный технический университет, Карла Маркса 20, Новосибирск, Россия, 630073

Аннотация. Исследуется возможность использования метода отбора признаков на основе корреляций для выбора оптимального подмножества из набора сильно коррелированных предикторов. Такие задачи возникают, например, при предсказании временных рядов экономических показателей на основе регрессионных моделей с большим числом возможных опережающих индикаторов с разными лагами. Жадные алгоритмы (прямого отбора и обратного исключения) в таких случаях дают сбои. Для поиска глобального оптимума проблема отбора признаков сформулирована как задача смешанного целочисленного программирования. Для ее решения используется метод бинарных отсечений и ветвлений. Результаты вычислительных экспериментов показали преимущество использования метода бинарных отсечений и ветвлений по сравнению с алгоритмами эвристического поиска. На реальном примере подбора опережающих индикаторов роста индекса потребительских цен показана приемлемость использования метода отбора на основе корреляций.

1. Введение

Аналитика больших данных включает задачу отбора предикторов для построения прогнозных моделей [1]. Во многих практических приложениях потенциальные предикторы сильно коррелируют. Примером может служить задача прогнозирования временных рядов на основе опережающих индикаторов [2], где показатели, взятые с лагами, сильно взаимосвязаны.

Для отбора предикторов в таких условиях быстрые и хорошо масштабируемые одномерные методы не подходят. Они оценивают признаки индивидуально, поэтому итоговый набор включает множество избыточных сильно коррелированных признаков.

Многомерные методы учитывают эти взаимосвязи и пытаются исключить не только нерелевантные (не влияющие на отклик), но и избыточные предикторы. Чаще всего отбор предикторов осуществляется совместно с построением прогнозных моделей с привлечением встроенных методов, таких как регрессия LASSO [2, 3]. Она дает разреженное решение, включающее только существенные признаки, которое, однако, очень чувствительно к параметру регуляризации.

Кроме этого в задачах прогнозирования временных рядов нередко используется шаговая регрессия [4], относящаяся к так называемым методам «обертки». Они выбирают оптимальное подмножество признаков из всех возможных вариантов совместно с оцениванием модели. В общем случае это задача характеризуется экспоненциальной сложностью по числу признаков. На практике для ее решения прибегают к жадным алгоритмам поиска [5]. Однако они не гарантируют достижения глобального оптимума.

Наконец, еще один класс методов отбора признаков – методы фильтрации. Из многомерных методов хорошо известен подход на основе корреляций [6]. Этот подход предложен для решения задач классификации. Его применимость к отбору признаков в условиях их сильной корреляции, в частности при подборе опережающих индикаторов, слабо изучена. Именно этот пробел мы и пытаемся восполнить в нашей статье.

2. Метод отбора предикторов на основе корреляций

Метод CFS (Correlation-based Feature Selection) предложен в работе [6]. Признаки выбираются так, чтобы обеспечить наибольшую корреляцию с откликом и наименьшую взаимосвязь между самими признаками. Тем самым решается следующая оптимизационная задача:

$$\frac{\sum_{i \in S_k} R_i}{\sqrt{k + 2\sum_{i,j \in S_k, i \neq j} r_{ij}}} \to \max_{S_k}, \quad (1)$$

где R_i – абсолютное значение коэффициента корреляции между i-м признаком и откликом, r_{ij} – абсолютное значение коэффициента корреляции между i-м и j-м признаком, S_k – подмножество из k признаков. При анализе временных рядов экономических показателей чаще всего как отклик, так и предикторы измеряются в количественных шкалах, поэтому допустимо использование обычного коэффициента корреляции Пирсона.

Представим задачу (1) как задачу нелинейного целочисленного программирования:

$$\frac{\sum_{i-1}^{n} R_i^2 x_i + 2\sum_{i \neq j} R_i R_j x_i x_j}{\sum_{i-1}^{n} x_i + 2\sum_{i \neq j} r_{ij} x_i x_j} \to \max_{x_1,\ldots,x_n}, \quad (2)$$

где $x_i \in \{0,1\}, i = 1,\ldots,n$, n – число признаков. Если $x_i = 1$, то i-й признак входит в оптимальный набор, иначе $x_i = 0$.

Задача (2) является дробно-полиномиальной. Следуя [7], перейдем к полиномиальной задаче. Для этого введем непрерывную переменную u. Кроме того перейдем от задачи максимизации к минимизации. Тем самым исходная задача сводится к следующей:

$$-\sum_{i-1}^{n} R_i^2 x_i u - 2\sum_{i \neq j} R_i R_j x_i x_j u \to \min_{x_1,\ldots,x_n,u}$$

$$\sum_{i-1}^{n} x_i u + 2\sum_{i \neq j} r_{ij} x_i x_j u = 1,$$

$$u > 0,\ x_i \in \{0,1\}.$$

В соответствии с техникой, предложенной в [8], для линеаризации слагаемых вида $x_i u$, $x_i x_j u$ вводятся переменные z_i, $i = 1,\ldots,n$, v_{ij}, $i = 1,\ldots,n$, $j = 1,\ldots,n$, $i \neq j$. Тогда получаем линейную задачу смешанного целочисленного программирования вида

$$-\sum_{i-1}^{n} R_i^2 z_i - 2\sum_{i \neq j} R_i R_j v_{ij} \to \min_{\substack{x_1,\ldots,x_n,u,\\ z_1,\ldots,z_n,v_{11},\ldots,v_{nn}}} \quad (3)$$

$$z_i \geq 0,\ v_{ij} \geq 0,\ u > 0,\ x_i \in \{0,1\},$$

$$\sum_{i-1}^{n} z_i + 2\sum_{i \neq j} r_{ij} v_{ij} = 1,$$

$$M(x_i - 1) + u \leq z_i \leq M(1 - x_i) + u,\ z_i \leq M x_i,$$

$$M(x_i + x_j - 2) + u \leq v_{ij} \leq M(2 - x_i - x_j) + u,\ v_{ij} \leq M x_i, v_{ij} \leq M x_j,$$

где M – большое число.

Тем самым исходная нелинейная задача (1) сводится к задаче линейного программирования (ЛП) большой размерности. К n бинарным переменным добавляется еще $(n^2+n+2)/2$ непрерывных переменных. А число ограничений составляет $(2n^2+n+2)$.

3. Метод бинарных отсечений и ветвлений

Метод бинарных отсечений и ветвлений (МБОВ) изначально был разработан для решения задач ЛП с булевыми переменными [9, 10] и затем распространен на случай общей задачи линейного программирования со смешанными переменными (milp) [11]. Любая подобная задача, частным случаем которой является (3), представима в виде:

$$\gamma(x) = c^{1T}x + c^{2T}y + const \to \max, \tag{4}$$

$$A^1x + A^2y \le b, \overline{0} \le x \le \overline{1},\ y \ge \overline{\overline{0}}, \tag{5}$$

$$x \in I^2_{n^1}. \tag{6}$$

(4)-(6) – milp с булевыми переменными x и непрерывными y. Здесь и далее I^m_n множество целочисленных векторов размерности n, принимающих значения из диапазона от 0 до $m-1$. Условия (6) определяют принадлежность компонент решения x одной из вершин единичного гиперкуба размерности n^1, $c^1, x, \overline{0}, \overline{1}$ - векторы той же размерности, $\overline{0}$ - нулевой, $\overline{1}$ - вектор, состоящий из единиц, $const$ - некоторая константа и $c^1 \ge \overline{0}$. Векторы $c^2, y, \overline{\overline{0}}$ имеют размерность n^2. Фактически к (4)-(6) может быть компактно сведена любая milp и значительная часть задач mip, см., например [10]. Пусть x^0, y^0 решение ослабленной задачи (4)-(5), $[\cdot]$ - целая часть числа, $\beta_0 = \hat{\alpha}^T x^0$, где $\hat{\alpha}_j \in \{0,\ 1\}$, $j = \overline{1,n}$. Любое неравенство, вида:

$$\hat{\alpha}^T x \le \hat{\beta}_0,\ \hat{\alpha}_j \in \{0,\ 1\},\ j = \overline{1,n},\ \hat{\beta}_0 = [\beta_0],\ \beta_0 = \hat{\alpha}^T x^0, \tag{7}$$

будем именовать бинарным отсечением (БО) для задачи (4)-(5).

Если x^0 - часть решения ослабленной задачи (4)-(5) x^0, y^0, то

$$\varsigma^T x \le \phi_0,\ \phi_0 = \varsigma^T x^0, \tag{8}$$

(8) может являться порождающим неравенством при условиях $\varsigma_j = \sum_{i \in I^B} \lambda_i a_{ij}$, $\lambda_i \ge 0$, где $a_{ij}, i \in I^B$ - коэффициенты базисной части A^1, и λ_i - веса базисных ограничений. В частности, если λ_i - двойственные оценки ограничений (5), то $\varsigma_j = c_j$,. $j = \overline{1,n}$. Определим дополняющую систему БО к системе ограничений (5)

$$\mathrm{A}^{1D}x \le \beta, \tag{9}$$

где $\mathrm{A}^{1D} = \left\| \hat{\alpha}^i_j \right\|_{m^D \times n^1}$, $\hat{\alpha}^i_j \in \{0,\ 1\}$, $j = \overline{1,n}$ - матрица коэффициентов дополняющей системы, и вектор правых частей отсечений β определяется в соответствии с (7).

Можно предложить несколько способов идентификации правильности БО [9, 10]. В частности, в МБОВ используется следующий признак. Определим:

$$\hat{\alpha}^T x \ge \beta(x^0) + 1, \tag{10}$$

где $\beta(x^0) = \left[\hat{\alpha}^T x^0\right]$, и x^0 - часть оптимального решения x^0, y^0 задачи (4),(5),(9).

Если задача (4),(5),(10) имеет решение, то отсечение $\hat{\alpha}^T x \le \beta(x^0)$ неправильное. Напротив, если (4),(5),(10) решения не имеет из-за противоречивости условий (5),(10), то $\hat{\alpha}^T x \le \beta(x^0)$ - правильное БО. На применении этого признака основана процедура синтеза БО, названная *выбором на множестве ближайших отсечений* [9, 11]. Опишем названную процедуру.

Определим неравенство-следствие базисной системы (5),(9), перестановкой упорядочив ς по убыванию (обозначим $\vec{\varsigma}$). Рассмотрим совокупность из n^1 векторов, размерности n^1: $\hat{\alpha}^1 = (1,0,...,0)$, ..., $\hat{\alpha}^j = (1,1,...,1,0,0,...,0)$ (j начальных единиц), ..., $\hat{\alpha}^{n^1} = (1,1,...,1)$.Каждому $\hat{\alpha}^j$ поставим в соответствие величину $cs(\hat{\alpha}^j) = \dfrac{\vec{\varsigma}^T \hat{\alpha}^j}{|\varsigma|_2 |\hat{\alpha}^j|_2}$, $j = \overline{1,n^1}$.

Дискретная функция $cs(\hat{\alpha}^j)$ имеет строгий максимум и однозначно определяет приоритет каждого из альтернативных отсечений с коэффициентами $\hat{\alpha}^j$. Добавление всей совокупности таких БО в (9) с решением (4),(5),(10) позволяет выявить противоречивые условия при их наличии, идентифицируя правильные отсечения. Далее, при наличии правильных БО, выбирается единственное из них с максимальным значением $cs(\hat{\alpha}^j)$. При отсутствии правильных отсечений, выбирается БО, соответствующее максимуму $cs(\hat{\alpha}^j)$, $j = \overline{1,n^1}$.

Другой важнейшей характеристикой БО является мера радикальности. Данная величина характеризует глубину отсечения рассматриваемого типа. Для БО $\hat{\alpha}^T x \le \hat{\beta}_0$, $\hat{\alpha}_j \in \{0,1\}$, $j = \overline{1,n}$, $\hat{\beta}_0 = [\beta_0]$, $\beta_0 = \hat{\alpha}^T x^0$. Под радикальностью r понимается число вершин единичного гиперкуба, отсекаемых БО в предположении, что отсечение правильное. В общем случае $\hat{a}^T x \le b$, $b \in I_1^k, x \in I_{n^1}^2, \hat{a} \in I_{n^1}^2$, $k = \sum_{j=1}^{n^1} \hat{a}_j, 1 \le k \le n^1$ или $\sum_{j=1}^{n^1} \hat{a}_j x_j \le b, x_j \in I_1^2, j = \overline{1,n^1}$, $\hat{a}_j \in \{0,1\}$ - коэффициенты отсечения.

При произвольном $b \in I_1^k$, $k = \sum_{j=1}^{n^1} \hat{a}_j, 1 \le k \le n^1$ и $C_k^l = \dfrac{k!}{l!(k-l)!}$. определим:

$$r_k^b = 2^{n^1-k} \sum_{l=b+1}^{k} C_k^l \to \max \qquad (11)$$

В (11) учитываются вершины единичного гиперкуба, лежащие «выше» уровня $\hat{a}^T x \le b$, т.е. принадлежащие гиперплоскостям $\hat{a}^T x = l$ с правыми частями ($b+1, b+2, ..., k$). Сама же гиперплоскость $\hat{a}^T x = b$ содержит $2^{n^1-k} C_k^b$ вершин. *Максимально радикальное БО* определяется в соответствии с (11) и $\hat{\alpha}_j = 1, j = \overline{1,n^1}$ с возможным исключением компоненты минимального относительно $\vec{\varsigma}$ порядка, если сумма коэффициентов левой части БО $\hat{a}^T \tilde{x} \le b$ без такого исключения оказывается целым числом.

Независимо от того, какая мера, близость к порождаемому неравенству, или радикальность рассматривается в качестве приоритетной, алгоритм МБОВ выглядит следующим образом.

1. Пусть получено решение исходной релаксированной задачи (4),(5),(7) x^0, y^0 и $\gamma(x^0, y^0)$. Если x^0 целые, останов алгоритма. В противном случае:

2. На шаге $t(1,2,...)$ выбираем вершину для зондирования с максимальной оценкой $\gamma(x^q, y^q)$, $q \in (1,2,...,t-1)$. Если список вершин пуст, задача не имеет целочисленного решения. Останов алгоритма. Если вершина с максимальной оценкой $\gamma(x^q, y^q)$ содержит целочисленные x^q, решение (x^q, y^q) является оптимальным. Останов алгоритма. Иначе:

3. Образуем два новых кандидата, для каждого из которых дополняем текущую матрицу A^{1D} для шага q БО (7) и (10) в соответствии с процедурами выбора отсечения *(по величине* $cs(\hat{\alpha}^j)$,

либо по радикальности отсечения (8)): $(\hat{\alpha}^{(t+1)})^T x \le \beta(x^q)$ и $(\hat{\alpha}^{(t+1)})^T x \ge \beta(x^q)+1$, соответственно.

4. Решаем пару альтернативных подзадач с отсечениями $(\hat{\alpha}^{(t+1)})^T x \le \beta(x^q)$ и $(\hat{\alpha}^{(t+1)})^T x \ge \beta(x^q)+1$.

5. Запоминаем компоненты их решения $\underline{x}^{t+1}$, x^{t+1} и оценки $\gamma(\underline{x}^{t+1}, \underline{y}^{t+1})$, $\gamma(x^{t+1}, y^{t+1})$, добавляя в список вершин дерева. Если какой - либо кандидат не имеет решения, вычеркиваем его из списка вершин.

6. Увеличиваем номер шага ($t := t+1$), переходим к п. 2. ■

4. Результаты вычислительных экспериментов

Применимость метода CFS для отбора признаков в условиях их сильной корреляции исследована на следующем модельном примере.

Релевантные признаки $x_1^{(m)}, x_2^{(m)}, x_3^{(m)}, x_5^{(m)}, x_6^{(m)}$ моделировались как независимые случайные величины со стандартным нормальным распределением. Релевантные признаки $x_4^{(m)}, x_7^{(m)}$ находились из соотношений:

$$x_4^{(m)} = x_3^{(m)} + e_1, x_7^{(m)} = x_5^{(m)} + x_6^{(m)} + e_2,$$

где e_1, e_2 – независимые случайные величины со стандартным нормальным распределением.

Отклик определялся как

$$y = \sum_{i=1}^{7} x_i^{(m)} + e_3$$

где e_3 – независимая от e_1, e_2 и от $x_i^{(m)}$ нормально распределенная случайная величина с нулевым математическим ожиданием и стандартным отклонением, равным 0,1.

Избыточные признаки моделировались как коррелирующие с основными предикторами:

$$x_i^{(r)} = x_i^{(m)} + \varepsilon_i,\ i = 1,...,7,$$
$$x_i^{(r)} = x_{i-7}^{(m)} + \varepsilon_i,\ i = 8,...,14.$$

Нерелевантные предикторы моделировать как случайный шум: $\xi_1,...,\xi_5$ – независимые случайные величины со стандартным нормальным распределением. Кроме того в число таких шумовых кандидатов включены $\varepsilon_1,...,\varepsilon_4$.

Для каждой случайной величины моделировались выборки объемом 1000. В итоге в набор признаков, из которых осуществлялся отбор, включены

- релевантные предикторы, на основе которых рассчитывается отклик, $x_1^{(m)},...,x_7^{(m)}$;
- избыточные признаки, коррелирующие с релевантными, $x_1^{(r)},...,x_{14}^{(r)}$;
- нерелевантные признаки $\xi_1,...,\xi_5$, $\varepsilon_1,...,\varepsilon_4$.

Всего 30 признаков. Модельные эксперименты повторялись 1000 раз.

Задача смешанного целочисленного программирования содержала 496 переменных и 1832 ограничения.

В качестве альтернативных методов решения задачи привлекались жадные алгоритмы: прямого отбора и обратного исключения [12]. Использовалась их реализация в среде R в виде функций forward.search и backward.search из пакета FSelector. Значимость подмножества признаков задавалась в виде целевой функции из (1).

В таблице 1 представлены результаты вычислительных экспериментов – доля случаев, в которых каждый из признаков вошел в набор. Нерелевантные признаки не были включены в набор ни в одном эксперименте. Использованы обозначения: прямого отбора – forward, обратного исключения – backward, бинарных отсечений и ветвлений – binary cut-and-branch.

Таблица 1. Результаты отбора предикторов.

	Метод				Метод		
Признак	forward	backward	binary cut-and-branch	Признак	forward	backward	binary cut-and-branch
$x_1^{(m)}$	0,219	0,955	0,927	$x_4^{(r)}$	0,049	0,938	0,613
$x_2^{(m)}$	0,216	0,945	0,929	$x_5^{(r)}$	0,006	0,716	0,024
$x_3^{(m)}$	0,258	0,991	0,962	$x_6^{(r)}$	0,013	0,742	0,036
$x_4^{(m)}$	0,954	0,996	0,98	$x_7^{(r)}$	0,217	0,997	0,915
$x_5^{(m)}$	0,208	0,993	0,977	$x_8^{(r)}$	0	0,046	0,006
$x_6^{(m)}$	0,204	0,995	0,969	$x_9^{(r)}$	0	0,061	0,003
$x_7^{(m)}$	1	1	0,999	$x_{10}^{(r)}$	0	0,069	0,005
$x_1^{(r)}$	0,011	0,462	0,057	$x_{11}^{(r)}$	0,012	0,37	0,053
$x_2^{(r)}$	0,013	0,504	0,066	$x_{12}^{(r)}$	0,001	0,065	0,002
$x_3^{(r)}$	0,012	0,642	0,053	$x_{13}^{(r)}$	0	0,048	0,001

Из таблицы 1 видно, что метод прямого отбора чаще всего не включает избыточные предикторы. Однако в итоговое подмножество практически не попадает и существенную часть релевантных признаков. Стабильно в набор попадают только $x_4^{(m)}, x_7^{(m)}$, которые моделировались как коррелирующие с остальными существенными предикторами. Это сказывается на значениях целевой функции, которые далеки от оптимальных. Графически они представлены на рисунке 1 в виде ящиков с усами. По графику можно судить о том, что разброс значений целевой функции в проведенных экспериментах очень велик, а среднее значение намного меньше, чем достигаемое методами обратного исключения и бинарных отсечений и ветвлений.

В свою очередь методы обратного исключения и бинарных отсечений и ветвлений включают в набор предикторов все релевантные признаки. Но метод обратного исключения часто оставляет в подмножестве слишком много избыточных предикторов. По сравнению с этим для результатов применения МБОВ такие случая относительно редки. Часто включатся только несколько избыточных признаков $x_4^{(r)}, x_7^{(r)}$. Это объясняется тем, что они коррелируют с релевантными предикторами $x_4^{(m)}, x_7^{(m)}$, которые, в свою очередь, взаимосвязаны с остальными существенными предикторами. Тем не менее, несмотря на эту проблему, метод бинарных отсечений и ветвлений обеспечивает лучшие значения целевой функции по сравнению с методом обратного исключения (рис. 1).

По всей видимости, случай многосторонней коррелированности признаков сложен для обработки методом CFS. Это является проблемой самого метода, а не алгоритмов оптимизации, что может сказаться на его применимости для решения практических задач.

5. Отбор опережающих индикаторов роста индекса потребительских цен

Проверим применимость метода на основе корреляций на реальном примере. Для этого была выбрана задача прогнозирования индекса потребительских цен. Из экономических исследований [13] известно, что одним из опережающих индикаторов изменения цен в ходе экономических циклов является индекс цен на промышленные материалы. Среди данных, предоставляемых единой межведомственной информационно-статистической системой [14], доступна ежемесячная динамика индекса цен приобретения машин и оборудования инвестиционного назначения. Индексы сгруппированы по видам экономической деятельности и территориям РФ. Далее в качестве территории выбрана РФ в целом.

В качестве тестового временного отрезка выбран период с июня по ноябрь 2010 года, соответствующий резкому росту базового индекса цен в процентах к соответствующему

периоду предыдущего года. Обучающий набор данных взят с января 2006 по май 2010 года. На этом наборе построена модель ARIMA. Но она закономерно предсказывает продолжающееся падение цен, что продемонстрировано на рисунке 2. Проверим, удастся ли улучшить этот прогноз путем использования опережающих индикаторов.

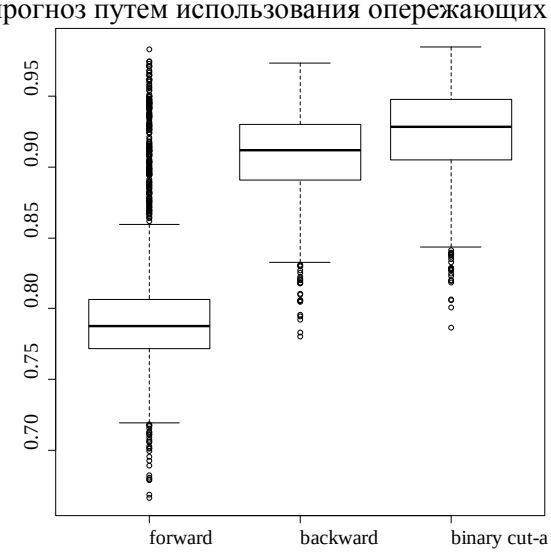

Рисунок 1. Ящики с усами для значений целевой функции, полученных в вычислительных экспериментах.

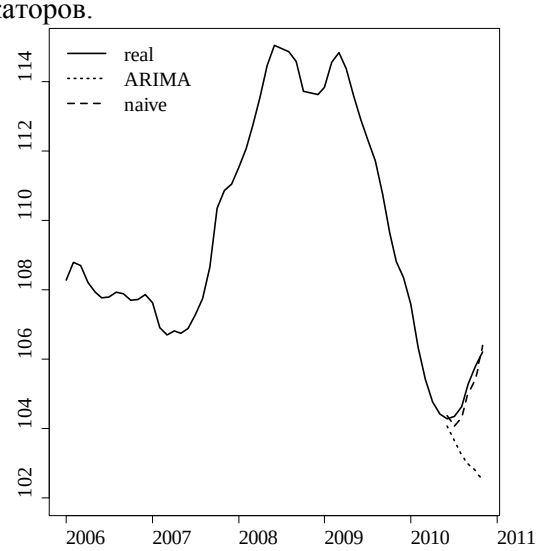

Рисунок 2. Фактические и прогнозные значения базового индекса потребительских цен, в % к соответствующему периоду предыдущего года.

За временной период, выбранный для обучения модели, доступны данные об индексах цен приобретения машин и оборудования инвестиционного назначения по 92 видам экономической деятельности, в том числе по 16 разделам ОКВЭД и суммарный индекс по всем видам деятельности. Данные взяты в процентах к соответствующему периоду предыдущего года.

Сначала использовался наивный подход, предполагающий, что лучшим предиктором является суммарный индекс по всем видам деятельности. С помощью подхода CFS подобраны только оптимальные лаги. В качестве кандидатов рассматривались лаги от 0 до -6. В результате выбраны лаги -4 и -6. Такой подход привел к довольно низкому значению целевой функции F^* (табл. 2).

Таблица 2. Результаты решения оптимизационной задачи.

Метод	n	n^*	F^*
наивный	7	2	0.857957
прямого отбора	644	4	0.954952
обратного исключения	644	39	0.963451
прямого отбора	119	4	0.942901
обратного исключения	119	18	0.950590
бинарных отсечений и ветвлений	119	7	0.954994

В качестве потенциальных предикторов для применения жадных алгоритмов сначала рассматривались индексы цен приобретения машин и оборудования инвестиционного назначения по видам экономической деятельности, взятые с лагами от 0 до -6. Тем самым общее число предикторов n составило 644. Отобранное оптимальное их количество n^* представлено в таблице 2.

Далее число предикторов было сужено до индексов по разделам и суммарного индекса, то есть 17 индексов. С учетом возможных лагов общее число признаков составило 119. В таблице 2 приведены оптимальные значения целевых функций F^* и числа предикторов n^*, полученные с помощью методов прямого отбора, обратного исключения и бинарных отсечений и ветвлений.

В обоих случаях метод прямого отбора оставляет очень мало предикторов. При этом значения целевой функции самые низкие. В то время как метод обратного исключения отбирает много переменных, которые, безусловно, являются избыточными. Наконец, метод бинарных отсечений и ветвлений обеспечивает наилучшее значение целевой функции.

Для того чтобы проверить, как полученные результаты сказываются на качестве прогнозов на тестовых данных, по обучающему набору построены регрессионные модели с включением отобранных переменных с выбранными лагами. Для этого использовался пакет dynlm статистической среды R.

Для того чтобы построить прогноз на шесть месяцев вперед нужно иметь в распоряжении будущие значения предикторов. Считалось, что фактические их значения доступны только до мая 2010 года. Если для прогнозирования были необходимы более поздние значения, то строился прогноз временного ряда предиктора на основе ARIMA-модели. Для автоматического подбора структуры ARIMA-модели использовалась функция auto.arima пакета forecast, реализованная в среде R. Для прогнозирования использовалась функция forecast из того же пакета. Прогнозирование по результатам отбора предикторов методом обратного исключения не осуществлялось, поскольку число признаков (39 и 18) явно избыточно для построения модели по 47 месяцам (с учетом самого раннего лага -6).

В таблице 3 представлены отклонения фактических значения индекса потребительских цен от его прогнозов. Полужирным начертанием выделены наименьшие по модулю расхождения. Оказалось, что наивный подход в целом дает хороший результат. Графически это изображено на рисунке 2.

Значит, подход на основе корреляций может быть рекомендован для выбора оптимальных лагов объясняющих переменных в модели временного ряда. Ситуация с одновременным отбором индексов цен приобретения машин и оборудования инвестиционного назначения по видам экономической деятельности и их лагов более сложная. Временные ряды индексов очень близки (некоторые даже почти идентичны). Отсюда очень высокая корреляция между переменными. В то же время лаговые значения индексов тоже сильно коррелируют. Это напоминает случай многосторонней коррелированности признаков из модельного примера. Как было выявлено выше, в такой модели достижение оптимума в задаче (1) не гарантирует, что будут отобраны только релевантные признаки. В структуре решения допускается некоторая доля избыточных предикторов.

Таблица 3. Результаты прогнозирования индекса потребительских цен.

Метод	июнь	июль	август	сентябрь	октябрь	ноябрь
ARIMA	0,22	0,68	1,39	2,31	2,98	3,73
наивный	**-0,07**	0,28	0,30	0,27	**0,36**	**-0,18**
прямого отбора, 92 индекса	-1,6	-0,9	-0,67	**0,18**	0,71	0,6
прямого отбора, 17 индексов	-0,08	0,53	0,37	0,84	1,06	0,69
бинарных отсечений и ветвлений, 17 индексов	-0,86	**0,13**	**0,29**	0,93	1,22	1,13

Видимо, эта проблема имеет место и при отборе индексов как опережающих индикаторов. В результате оптимальное решение, полученное с помощью метода бинарных отсечений и ветвлений, обеспечивает лучшее качество прогноза только на среднесрочную перспективу (2-3 месяца). В долгосрочном же плане на 5-6 месяцев вперед прогнозы получаются хуже, чем при отборе только оптимальных лагов для суммарного индекса.

Этот эффект хорошо заметен при сравнении результатов отбора методами прямого поиска и бинарных отсечений и ветвлений по 17 индексам. Результаты прямого отбора, дающие меньшее число предикторов, лучше прогнозируют на долгосрочную перспективу. Это можно

объяснить эффектом переобучения, поскольку с помощью метода бинарных отсечений и ветвлений отобрано большее число предикторов, часть из которых может быть избыточной.

6. Выводы и рекомендации

Таким образом, показана применимость одного из методов фильтрации на основе корреляций для отбора предикторов в задаче прогнозирования динамики на основе опережающих индикаторов. Построение прогнозных моделей производится в условиях сильной корреляции между потенциальными предикторами. Из вычислительных экспериментов стало ясно, что в таких условиях жадные алгоритмы при оптимизации эвристики не дают удовлетворительного результата. Метод прямого отбора не включает многие релевантные признаки, метод обратного исключения оставляет в наборе много избыточных предикторов. Метод бинарных отсечений и ветвлений даёт оптимальный результат. Однако метод CFS не идеален: оптимальное значение эвристики не гарантирует, что будут отобраны только все релевантные предикторы, возможно, включение и избыточных признаков. Это происходит, если релевантные признаки коррелируют и между собой, и с избыточными признаками. Чтобы этого избежать, рекомендуется при формировании исходного множества потенциальных предикторов заранее исключать дублирующиеся показатели и индикаторы с очень схожей динамикой.

7. Литература

[1] Bolón-Canedo, V. Recent advances and emerging challenges of feature selection in the context of big data / V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos // Knowledge-Based Systems. – 2015. – Vol. 86. – P. 33-45.

[2] Sagaert, Y.R. Tactical sales forecasting using a very large set of macroeconomic indicators / Y.R. Sagaert, E.H. Aghezzaf, N. Kourentzes, B. Desmet // European Journal of Operational Research. – 2018. – Vol. 264(2). – P. 558-569.

[3] Tibshirani, R. Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society. Series B (Methodological). – 1996. – Vol. 58(1) – P. 267-288.

[4] Fite, J.T. Forecasting freight demand using economic indices / J.T. Fite, G. Don Taylor, J.S. Usher, J.R. English, J.N. Roberts // International Journal of Physical Distribution & Logistics Management. – 2002. – Vol. 32. – № 4. – P. 299-308.

[5] Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. – М.: ДМК Пресс, 2015. – 400 с.

[6] Hall, M.A. Correlation-based feature selection for machine learning. PhD thesis // Hamilton: University of Waikato, 1999.

[7] Nguyen, H. Optimizing a class of feature selection measures / H. Nguyen, K. Franke, S. Petrovic // NIPS Workshop on Discrete Optimization in Machine Learning: Submodularity, Sparsity & Polyhedra (DISCML), 2009.

[8] Chang, C.-T. On the polynomial mixed 0-1 fractional programming problems // European Journal of Operational Research. – 2001. – Vol. 131(1). – P. 224-227.

[9] Мезенцев, Ю.А. Метод бинарных отсечений и ветвлений целочисленного программирования // Доклады академии наук высшей школы РФ. – 2011. – Т. 1, № 16. – С. 12-25.

[10] Mezentsev, Y.A. Binary Cut-and-Branch Method for Solving Linear Programming Problems with Boolean Variables // CEUR Workshop Proceedings. – 2016. – Vol. 1623. – P. 72-85.

[11] Mezentsev, Y. Binary cut-and-branch method for solving mixed integer programming problems Constructive Nonsmooth Analysis and Related Topics (Dedicated to the Memory of V.F. Demyanov) // CNSA, 2017. DOI: 10.1109/cnsa.2017.7973989.

[12] Sutter, J.M. Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection / J.M. Sutter, J.H. Kalivas // Microchemical journal. – 1993. – Vol. 47(1-2). – P. 60-66.

[13] Klein, P.A. The leading indicator approach to economic forecasting – retrospect and prospect / P.A. Klein, G.H. Moore // Journal of forecasting. – 1983. – Vol. 2(2). – P. 119-135.

[14] ЕМИСС [Электронный ресурс]. – Режим доступа: https://fedstat.ru/ (13.11.2018).

Благодарности
Работа поддержана грантом Министерства образования и науки РФ в рамках проектной части государственного задания, проект № 2.2327.2017/4.6 «Интеграция моделей представления знаний на основе интеллектуального анализа больших данных для поддержки принятия решений в области программной инженерии».

Forecasting using predictor selection from a large set of highly correlated variables

A.Yu. Timofeeva[1], Yu.A. Mezentsev[1]

[1]Novosibirsk State Technical University, Karl Marx Ave. 20, Novosibirsk, Russia, 630073

Abstract. The potential of Correlation-based Feature Selection has been explored in selecting an optimal subset from a set of highly correlated predictors. This problem occurs, for example, in time series forecasting of economic indicators using regression models on multiple lags of a large number of candidate leading indicators. Greedy algorithms (forward selection and backward elimination) in such cases fail. To obtain the globally optimal solution, the feature selection problem is formulated as a mixed integer programming problem. To solve it, we use the binary cut-and-branch method. The results of simulation studies demonstrate the advantage of using the binary cut-and-branch method in comparison with heuristic search algorithms. The real example of the selection of leading indicators of consumer price index growth shows the acceptability of using the Correlation-based Feature Selection method.