№ 6/1(37) 2021

ISSN 2181-8460

HISOBLASH VA AMALIY MATEMATIKA MUAMMOLARI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

MUHAMMAD AL-XORAZMIY NOMIDAGI
TOSHKENT AXBOROT TEXNOLOGIYALARI
UNIVERSITETI

AXBOROT-KOMMUNIKATSIYA
TEXNOLOGIYALARI
ILMIY-INNOVATSION MARKAZI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

Спецвыпуск № 6/1(37) 2021

Журнал основан в 2015 году.
Издается 6 раз в год.

Учредитель:
Научно–инновационный центр информационно-коммуникационных технологий.

Журнал зарегистрирован в Узбекском Агентстве по печати и информации.
Регистрационное свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

Адрес редакции:
100125, г. Ташкент, м-в. Буз-2, 17А.
Тел.: +(99871) 231-92-45.
E-mail: info@pvpm.uz.
Сайт: www.pvpm.uz.

Дизайн и компьютерная вёрстка:
Шарипов Х. Д.

Отпечатано в типографии НИЦ ИКТ.
Подписано в печать 24.12.2021 г.
Формат 60x84 1/8. Заказ №9.
Тираж 100 экз.

Содержание

Алоев Р.Д., Дадабаев С.У., Улашев А.Э., Ботиров И.Б.
Неявная противопоточная разностная схема для линейных гиперболических 2-х уравнений с младшими членами 5

Файзиев А.А., Тургунов А.М., Мамадалиев Х.А.
Статистический анализ и прогнозирование динамики урожайности хлопчатника Наманганской области Республика Узбекистан 21

Фаязов К.С., Абдуллаева З.Ш.
Внутренняя краевая задача для системы уравнений в частных производных смешанного типа 29

Нармурадов Ч.Б., Гуломкодиров К.А., Холмурзаева Н.А.
О сходимости итерационной схемы переменных направлений с оптимальными итерационными параметрами 44

Шадиметов Х.М., Гуломов О.Х.
Об одном новом методе построения составных оптимальных квадратурных формул 55

Джаббаров О.Р.
Математическое моделирование процессов диффузии с переменным коэффициентом демпфирования с использованием ВКБ решений 64

Равшанов Н., Шадманов И.У.
Математическая модель и эффективный численный алгоритм для исследования процессов тепло-влагопереноса в неоднородных пористых средах . . . 75

Саидов У.М.
Моделирование ионообменной фильтрации жидкости с учетом процессов кольматации и суффозии 90

Равшанов Н., Юсупов М., Каршиев Д.К., Аминов С.
Об одном подходе численного решения нелинейных интегро – дифференциальных уравнений описивающий вынужденного колебания вязкоупругих тел 103

Игнатьев Н.А., Рахимова М.А., Лолаев М.Я.
Особенности отбора информативных наборов признаков на данных с пропусками 113

Адылова Ф.Т, Давронов Р.Р.
Представления структуры лекарств на основе BERT: сравнение токенизаторов 123

УДК 519.95

ОСОБЕННОСТИ ОТБОРА ИНФОРМАТИВНЫХ НАБОРОВ ПРИЗНАКОВ НА ДАННЫХ С ПРОПУСКАМИ

Игнатьев Н.А., Рахимова М.А., Лолаев М.Я.

*n_ignatev@rambler.ru

Национальный университет Узбекистана имени Мирзо Улугбека,
100174, Узбекистан, г. Ташкент, ул. Университетская 4.

Рассматриваются отбор информативных наборов разнотипных признаков с учётом наличия пропусков в данных при описании объектов классов. Для сокращения комбинаторной сложности алгоритмов предлагаются использовать новые эвристики с учётом неизмеренных значений (пропусков) в описании объектов. Условием отбора является свойство инвариантности данных к масштабам их измерений. Инвариантность достигается за счёт применения методов разбиения значений признаков на непересекающиеся интервалы. Разбиение на интервалы используются в двух способах предобработки данных, проводимой с целью унификации шкал измерений признаков. По результатам предобработки формируются последовательности, упорядоченные по отношению устойчивости признаков или по отношению межклассового различия по значениям пар признаков. Первое отношение выбирается для числа классов, равному 2, второе – при числе больше или равному 3. При вычислении устойчивости используется значения функции принадлежности к классам. Исследуется изменение порядка следования признаков в последовательностях в зависимости от процента пропусков в данных. В качестве показателей эффективности рекомендуется рассматривать изменение рангов признаков или дисперсию их устойчивости в упорядоченных последовательностях при разных долях пропусков в данных.

Ключевые слова: тип закономерности, последовательность, устойчивость признака, пропуски в данных.

Цитирование: *Игнатьев Н.А., Рахимова М.А., Лолаев М.Я.* Особенности отбора информативных наборов признаков на данных с пропусками // Проблемы вычислительной и прикладной математики. – 2021. – № 6/1(37). – С. 113-122.

1 Введение

Информативные наборы признаков являются источником нового знания для слабо формализованных предметных областей. Наиболее распространённой формой для представления и анализа данных в информационных моделях являются таблицы «объект–свойство». В качестве предмета исследования в данной работе рассматриваются таблицы «объект–свойство», содержащие пропуски в данных. Как правило, задачи отбора информативных признаков является NP – полными, для сокращения комбинаторной сложности алгоритмов используются различные эвристики. Одной из целей создания новых эвристик является разработка и реализация алгоритмов с учётом неизмеренных значений (пропусков) в описании объектов.

Разработка эффективных методов заполнения пропусков основывается на вычислении регрессионных зависимостей между данными [1, 2]. Потребность в этих методах очевидна на примере использования моделей алгоритмов распознавания. Условием реализации значительной части алгоритмов является отсутствие пропусков в данных.

Отбор информативных признаков связан с изменением отношений между объектами внутри классов и между классами [3]. Размытость отношений между объектами при реализации вычислительных алгоритмов, связанная с ростом числа признаков, получила название «проклятие размерности» [4]. В качестве средства для анализа различных структур отношений в [5, 6] предлагалось использовать меру компактности классов и выборки в целом.

Значения мер компактности классов положены в основу ряда критериев для формирования латентного признака. Объясняется это тем, что выбор нового признакового пространства и отбор информативных латентных признаков реализуется параллельно друг с другом. Примером этому служат методы, использующие правила иерархической агломеративной группировки для формирования описания объектов [7]. Правила содержат условия перехода в пространство меньшей размерности через линейное и нелинейное отображение описаний объектов на числовую ось.

Наличие пропусков значений ряда признаков в описании объекта не всегда является причиной отказа от принятия решения по нему. Например, при постановке диагноза и назначении курса лечения пациенту в медицине. У специалистов может отсутствовать (находится в ремонте) уникальное оборудование для измерений или нет набора химических реагентов для проведения экспресс анализа группы крови. Поиск альтернативного варианта для решения проблемы через использование знаний из экспертных систем (ЭС) рассматривается как тупиковый. Сами ЭС являются замкнутыми и не могут выдать пользователю и малой доли того знания, которое можно получить при анализе многообразия отношений между объектами из слабо структурированных предметных областей.

В работе [8] были выделены 3 группы методов анализа неполных данных. Реализация двух групп (отбрасывание и заполнение пропусков) рассчитана на последующее применение обычных процедур. Отдельную группу образуют методы по разработке оригинальных процедур для совокупных вычислений по измеренным и пропущенным данным. Потребность в такой разработке для отбора информативных наборов признаков связана с тем, что удаление объектов с неполными данными приведёт к значительному снижению объёма выборки.

Актуальной проблемой для исследования является оценка эффективности принимаемых решений с учётом неизмеренных данных. Для восстановления пропусков в данных в [2] предложены гибридные модели экстремального градиентного бустинга. Утверждается, что гибридные методы на данных об осадках превзошли по качеству предсказания простые модели классификации и регрессии.

Для достижения инвариантности количественных признаков к масштабам измерений применяются методы разбиения их значений на непересекающиеся интервалы [9]. Границы интервалов используются для нормирования данных и определения градаций признака в номинальной шкале измерений. Никаких предположений о природе среды при реализации алгоритмов методов не делается.

Отбор информативных признаков можно свести к поиску последовательности как одной из 5 стандартных типов закономерностей [10] в интеллектуальном анализе данных (ИАД). Наличие пропусков существенно ограничивает число анализируемых комбинаций признаков для формирования последовательностей. Например, невозможно оценить структуру отношений между объектами классов для отбора информативных признаков с использованием мер расстояния по определяемой метрике $\rho(x, y)$.

Предлагаемые в работе процедуры реализуют способы формирования последовательностей признаков на основе 2-х отношений. Одно из отношений определяет порядок по значению устойчивости количественного или номинального признака, другое – значение межклассового различия по паре признаков. При вычислении устойчивости используется значения функции принадлежности к классам. В качестве показателей эффективности рекомендуется рассматривать изменение рангов признаков или дисперсию их устойчивости в упорядоченных последовательностях при разных долях пропусков в данных. Следствием из малой дисперсии значений устойчивости является малая вариабельность значений рангов последовательностей.

Проблему совместимости шкал измерений предлагается решать через:

– преобразование значений количественных признаков в градации номинальных с использованием интервальных методов;
– вычисление устойчивости признаков через значения функций принадлежности объектов к классам.

2 Постановка задачи

Пусть задано множество объектов $E_0 = \{S_1, ..., S_m\}$, содержащее представителей $l (l \geqslant 2)$ непересекающихся классов $K_1, ..., K_l$. Описание объектов производится с помощью набора из n разнотипных признаков $X(n) = (x_1, ..., x_n)$, $\delta (\delta < n)$ из которых измеряются в номинальной, $n - \delta$ в интервальной шкалах. Допускаются наличие пропусков в данных. Считается, что на множестве измеренных значений признаков из $X(n)$ определены неотрицательные меры $\mu(x_i)$ для числа классов $l = 2$ и $\eta(x_i, x_j)$ для числа классов $l \geqslant 2$, $i, j \in \{1, ..., n\}$, $i \neq j$.

Требуется:

– построить последовательности признаков, упорядоченных по значениям $\mu(x_i)$ для числа классов $l = 2$ и $\eta(x_i, x_j)$ для числа классов $l \geqslant 2$;
– оценить разброс значений $\mu(x_i)$ в зависимости от числа измеренных значений;
– оценить изменение порядка следования по $\eta(x_i, x_j)$ с учётом доли измеренных пар значений по $\{(x_i, x_j)\} \subset X(n)$.

3 Формирование последовательности признаков по мере $\mu(x_i)$

Пусть для значений количественного признака $x_c \in X(n)$ в описании объектов E_0 построена упорядоченная по неубыванию последовательность

$$r_1, ..., r_g, ..., r_\sigma, 2 < \sigma \leqslant m. \tag{1}$$

При разбиении (1) на непересекающиеся интервалы их число считается неизвестным. Определено условие разбиения, что в границах каждого интервала частота встречаемости значений признака из описаний объектов класса K_t больше чем в K_{3-t}, $t = 1, 2$.

Критерий для разбиения (1) на множество из p_c $(p_c \geqslant 2)$ непересекающихся интервалов $\left\{[r_u; r_v]^i\right\}$, $1 \leqslant u$, $u \leqslant v \leqslant m$, $i \in \{1, ..., p_c\}$ был предложен в [5]. Значения данных в границах интервала $[r_u; r_v]^i$ могут использоваться методами ИАД как градация номинального признака. Считается, что множество чисел, идентифицирующих p_c градаций номинального признака, всегда можно взаимно-однозначно отобразить в множество $\{1, ..., p_c\}$.

Пусть $d_{tc}(u, v)$, $d_{3-t}(u, v)$ – количество представителей классов K_t и K_{3-t} в интервале $[r_u; r_v]^i$, $i \in \{1, ..., p_c\}$. Для рекурсивной процедуры выбора значений r_u, r_v

используется критерий

$$\left| \frac{d_{tc}(u,v)}{T_{tc}} - \frac{d_{3-t,c}(u,v)}{T_{3-t,c}} \right| \to \max \tag{2}$$

где $T_{tc}, T_{3-t,c}$ – количество значений признака $x_c \in X(n)$ без пропусков у объектов E_0 соответственно из классов K_t и K_{3-t}. Естественным условием для реализации (2) является:

– число различных значений признака больше или равно 2;
– значения $T_{tc} > 0$, $T_{3-t,c} > 0$

Границы первого интервала $[r_u; r_v]^1$ на последовательности (1) вычисляются по максимуму критерия (2). Аналогичным образом определяются границы для $[r_u; r_v]^q$, $q > 1$ на значениях (1) не вошедших в $[r_u; r_v]^1, ..., [r_u; r_v]^{q-1}$. Критерием останова процедуры служит покрытие всех значений (1) непересекающимися интервалами.

Количество непересекающихся интервалов значений признака $x_c \in X(n)$ по (2) не является постоянным на разных выборках из генеральной совокупности. По этой причине есть интерес использования других показателей напрямую не связанных с количеством интервалов.

В зависимости от шкалы измерений признака $x_c \in X(n)$ через $d_{tc}(\gamma)(d_{3-t,c}(\gamma))$, $t = 1, 2$ будем обозначать число значений объектов в границах интервала $[r_u; r_v]^\gamma$ или объектов, описываемых градацией $\gamma \in \{1, ..., p_c\}$ из класса $K_t(K_{3-t})$. Значение функции принадлежности $f_c(\gamma)$ к классу K_1 вычисляется как

$$f_c(\gamma) = \frac{d_{1c}(\gamma)/T_{1c}}{d_{1c}(\gamma)/T_{1c} + d_{2c}(\gamma)/T_{2c}}. \tag{3}$$

Важной характеристикой для анализа данных, определяемой с помощью значений функции принадлежности (3), является устойчивость признака. Устойчивость $\mu(c) = \mu(x_c)$ признака $x_c \in X(n)$ по множеству значений градаций $\gamma \in \{1, ..., p_c\}$ вычисляется как

$$\mu(c) = \frac{1}{T_{1c} + T_{2c}} \sum_{r=1}^{m} \begin{cases} f_c(\gamma), x_{rc} = \gamma & f_c(\gamma) > 0.5, \\ 1 - f_c(\gamma), x_{rc} = \gamma & f_c(\gamma) < 0.5, \\ 0, x_{rc} = \gamma & f_c(\gamma) = 0.5 \quad x_{rc} = @, \end{cases} \tag{4}$$

где @ является символом пропуска в данных.

Показателем качества разбиения на интервалы является устойчивость [11], множество допустимых значений которой принадлежит [0;1]. При максимальном значении устойчивости $\mu(c) = 1$ каждый интервал содержит представителей (значения объектов) класса K_1 или K_2.

Целесообразность формирования последовательности признаков по мере $\mu(c)$ основывается на предположении, что оценка (математическое ожидание) множества значений признака по (4) на обучающих выборках из генеральной совокупности является несмещённой [12]. Доказательством несмещённости оценки могут служить результаты вычислительного эксперимента на выборках со случайным распределением пропусков в описании объектов.

Для эксперимента были использованы данные [13], представленные 820 объектами, описываемые 22 признаками, 17 из которых были количественными, 5 – номинальными. В класс K_1 вошли представители возрастной группы 20–24 лет, в K_2 –группы 40–44 лет. Рассматривались варианты с количеством пропусков в данных

Таблица 1 Результаты экспериментов по значениям устойчивости (4)

Признак	Мат. ожидание	Дисперсия	Признак	Мат. ожидание	Дисперсия
x_1	0.5575	0.00005889	x_{12}	0.6983	0.00000227
x_2	0.5504	0.00000138	x_{13}	0.6934	0.00001101
x_3	0.5511	0.00000783	x_{14}	0.5823	0.00000045
x_4	0.6245	0.00004025	x_{15}	0.6615	0.00000010
x_5	0.5392	0.00000754	x_{16}	0.5463	0.00001168
x_6	0.5433	0.00000643	x_{17}	0.5193	0.00000222
x_7	0.5008	0.00000038	x_{18}	0.5268	0.00000207
x_8	0.5029	0.00000002	x_{19}	0.6020	0.00000537
x_9	0.5582	0.00000841	x_{20}	0.6353	0.00001281
x_{10}	0.6108	0.00001963	x_{21}	0.6890	0.00000002
x_{11}	0.6440	0.00000067	x_{22}	0.6011	0.00000001

по каждому признаку от 0.0% до 35.0%. Результаты экспериментов по исследованию влияния пропусков в данных на изменение значений устойчивости (4) приводятся в табл.1.

Из анализа показателей дисперсии по табл. 1 следует вывод об малой изменчивости значений устойчивости (4) при наличии пропусков в данных.

Обозначим через

$$\omega_1(t_1), ..., \omega_g(t_g), ..., \omega_n(t_n), t_i \in \{1, ..., n\} \tag{5}$$

последовательность значений математических ожиданий признаков, упорядоченных в порядке невозрастания. Индекс элементов в (5) интерпретируется как ранг признака. Определим информативность набора из k признаков без пропусков ($k < n$) относительно (5) в описании произвольного допустимого объекта S.

Пусть $D(S,k)$ – сумма рангов k измеренных значений признаков объекта S. Значение суммы находится между $k(1+k)/2 \leqslant D(S,k) \leqslant k(2n-k+1)/2$. Для оценки информативности набора из k признаков объекта S предлагается использовать меру

$$\Omega\left(S,k\right) = 1 - \left(\frac{D\left(S,k\right) - \alpha}{\beta - \alpha}\right)\left(1 - \frac{k}{n}\right), \tag{6}$$

где $\alpha = k(1+k)/2$, $\beta = k(2n-k+1)/2$. Применение меры (6) рекомендуется при разведочном анализе данных.

4 Рекурсивный метод формирования последовательности признаков по мере $\eta(x_i, x_j)$

Особености вычисления меры $\eta(x_i, x_j)$ заключаются в предобработке данных для формирования матрицы различий между классами по парам в общем–то разнотипных признаков. Для преобразования к единной шкале измерений используется разбиение значений измеренных количественных признаков на непересекающиеся интервалы. Необходимым условием такого разбиения является: число интервалов должно быть равным числу классов. Для выполнения достаточного условия требуется чтобы не сушествовало интервала, содержащего все значения из двух и более классов.

Пусть $\pi_1, ..., \pi_p, ..., \pi_{l+1}$, $(\pi_i < \pi_{i+1})$– границы непересекающихся интервалов $[\pi_1; \pi_2]$, $(\pi_2; \pi_3], ..., (\pi_l; \pi_{l+1}]$ признака $x_a \in X(n)$ в описании объектов класса K_i, $i = 1, ..., l$

и u_i^p – количество измеренных значений признака объектов класса K_i в $[\pi_p;\pi_{p+1}]$ при $p=1$ и $(\pi_p;\pi_{p+1}]$ при $p>1$. С учётом наличия пропусков $\sum\limits_{p=1}^{l} u_i^p = q_i \leqslant |K_i|$, $\sum\limits_{i=1}^{l} q_i = q \leqslant m$.

Значения границ интервалов $\pi_1,...,\pi_p,...,\pi_{l+1}$ определяются по критерию

$$\left(\frac{\sum\limits_{p=1}^{l}\sum\limits_{i=1}^{l}(u_i^p-1)u_i^p}{\sum\limits_{i=1}^{l} q_i(q_i-1)}\right)\left(\frac{\sum\limits_{p=1}^{l}\sum\limits_{i=1}^{l} u_i^p\left(q-q_i-\sum\limits_{j=1}^{l}u_j^p+u_i^p\right)}{\sum\limits_{i=1}^{l} q_i(q-q_i)}\right)\to \max_{\pi_1<\pi_2<...<\pi_{l+1}} \tag{7}$$

Выражение в левых скобках (7) представляет внутриклассовое сходство, в правых – межклассовое различие. Множество допустимых значений критерия (0;1]. Если в каждом интервале содержатся все измеренные значения признака объектов одного класса, то критерий равен 1.

Разбиение на интервалы по (7) применяется при выборе градаций для значений количественного признака в номинальной шкале измерений. Преобразование признаков в номинальную шкалу измерений по (7) рассматривается как один из этапов предобработки данных для поиска информативных признаков. Выбор такого способа преобразования связан с инвариантностью значений (7) от масштабов измерений.

Пусть $Y(n)=(y_1,...,y_n)$ – набор признаков для описания объектов E_0 в номинальной шкале измерений, полученный с учётом предобработки данных по (7). Будем считать, что при числе градаций номинального признака $y_c \in Y(n)$, равном $p_c (p_c \geqslant 2)$, его (признака) значениями являются $1,...,p_c$. Для этого необходимо произвести замену исходных значений градаций из набора $\alpha(c)=(\alpha_1,...,\alpha_{p_c})$ их индексы. Тогда представление объектов E_0 по $Y(n)$ будет иметь вид целочисленной таблицы $Z=(z_{ij})_{m\times n}$, в которой $z_{ij}=0$ интерпретируется как код пропуска.

При формировании последовательности для отбора информативных признаков по описаниям объектов с пропусками в данных предлагается использовать значения из таблицы Z. Элементы матрицы парных различий признаков $B=\{b_{ij}\}_{n\times n}$ между классами по множеству пар объектов $(S_a=(z_{a1},...,z_{an}), S_b=(z_{b1},...,z_{bn})$, вычисляются (как мера $\eta(x_i,x_j)$) с помощью функций

$$g(a,b,i,j)=\begin{cases}2, & z_{ai}\neq z_{bi} \quad and \quad z_{aj}\neq z_{bj},\\ 1, & z_{ai}=z_{bi} \quad or \quad z_{aj}=z_{bj},\\ 0, & z_{ai}=z_{bi} \quad and \quad z_{aj}=z_{bj};\end{cases}$$

$$\alpha(a,b,i,j)=\begin{cases}0, & S_a,S_b\in K_d \quad or \quad z_{aj}\times z_{bj}=0, d=1,...,l\\ 1, & S_a\in K_d, S_b\in CK_d,\end{cases}$$

как

$$b_{ij}=\begin{cases}\dfrac{\sum\limits_{a=1}^{m}\sum\limits_{b=1}^{m}\alpha(a,b,i,j)g(a,b,i,j)}{2\sum\limits_{p=1}^{l}\left|\left\{S_d\in K_p\middle|z_{di}z_{dj}\neq 0\right\}\right|\left|\left\{S_d\in CK_p\middle|z_{di}z_{dj}\neq 0\right\}\right|}, i\neq j,\\ 0, i=j.\end{cases} \tag{8}$$

Процесс формирования последовательности признаков

$$y_{i_1}, y_{i_2}, ..., y_{i_n} \tag{9}$$

реализуется рекурсивной процедурой по матрице B, начиная с пары, имеющей максимальное значение b_{ij} по (8).

Отметим особенности формирования упорядоченной последовательности признаков с учетом пропусков в данных на основе попарного различия между классами. Наличие пропусков увеличивает вероятность события, при котором множество пар признаков по классам или выборки в целом для вычисления (8) будет пустым. Предлагается исследовать влияния количества пропусков в данных на процесс формирования последовательности (9) следующим образом.

Изначально считается, что в описании объектов E_0 неизмеренных значений признаков нет. Генерация пропусков по определяемому распределению на E_0 производится по фиксированному признаку $x_c \in X(n)$. Если $c \in l$, то необходима проверка условия достаточности разбиения значений признака на непересекающиеся интервалы по (7).

При невыполнении условия достаточности признак удаляется из дальнейшего анализа. Элементы $\{b_{ij}\}_{i,j\in\{1,\dots,n\}}$ рассматривается как математическое ожидание от значений (8) при генераци и пропусков по $x_c \in X(n)$. Исследуется дисперсия значений $\{b_{ij}\}_{i,j\in\{1,\dots,n\}}$и порядок следования признаков в (9) в зависимости от доли (процента) пропусков. Эксперимент проводился на данных по коммуникации пакетов [14] из 1075 объектов, разделённых на 4 непересекающихся класса. Из исходного описания объектов 21 признаком были удалены 5 как несоответствующие условию достаточности использования критерия (7). Для генерации пропусков от 5% до 45% по равномерному распределению использовалась библиотека Numpy [15] языка программирования Python. Порядок следования признаков по (9) при разных процентах пропусков и соответствующие парам признаков значения (8) приводятся в табл.2.

Таблица 2 Порядок следования признаков по (9) с учётом пропусков в данных

0%		5%		10%		15 %		20%	
(x_4, x_5)	0,7651	(x_5, x_6)	0.7591	(x_4, x_6)	0.7589	(x_5, x_6)	0.7574	(x_5, x_6)	0.7579
(x_1, x_6)	0,7492	(x_2, x_4)	0.7529	(x_2, x_5)	0.7530	(x_2, x_4)	0.7545	(x_2, x_4)	0.7537
(x_2, x_{13})	0.7231	(x_1, x_{13})	0.7223	(x_1, x_{13})	0.7220	(x_1, x_{13})	0.7211	(x_1, x_{13})	0.7204
(x_{12}, x_{15})	0.7036	(x_{12}, x_{15})	0.7046	(x_{12}, x_{15})	0.7048	(x_{12}, x_{15})	0.7048	(x_{12}, x_{15})	0.7042
(x_7, x_{16})	0.6243	(x_9, x_{16})	0.6231	(x_9, x_{16})	0.6231	(x_9, x_{16})	0.6224	(x_9, x_{16})	0.6220
(x_9, x_{11})	0.6079	(x_7, x_{11})	0.6036	(x_7, x_{11})	0.6012	(x_7, x_{11})	0.6032	(x_7, x_{11})	0.6035
(x_3, x_8)	0.5784	(x_3, x_8)	0.5855	(x_3, x_8)	0.5867	(x_3, x_8)	0.5863	(x_3, x_8)	0.5855
(x_{10}, x_{14})	0.5081	(x_{10}, x_{14})	0.5081	(x_{10}, x_{14})	0.5091	(x_{10}, x_{14})	0.5093	(x_{10}, x_{14})	0.5111
25%		30%		35 %		40%		45%	
(x_5, x_6)	0.7576	(x_5, x_6)	0.7574	(x_5, x_6)	0.7562	(x_4, x_5)	0.7546	(x_5, x_6)	0.7552
(x_2, x_4)	0.7532	(x_2, x_4)	0.7517	(x_2, x_4)	0.7514	(x_2, x_6)	0.7517	(x_2, x_4)	0.7508
(x_1, x_{13})	0.7205	(x_1, x_{13})	0.7194	(x_1, x_{13})	0.7188	(x_1, x_{13})	0.7179	(x_1, x_{13})	0.7172
(x_{12}, x_{15})	0.7043	(x_{12}, x_{15})	0.7044	(x_{12}, x_{15})	0.7046	(x_{12}, x_{15})	0.7041	(x_12, x_{15})	0.7035
(x_9, x_{16})	0.6214	(x_9, x_{16})	0.6217	(x_9, x_{16})	0.6207	(x_9, x_{16})	0.6218	(x_9, x_{16})	0.6227
(x_7, x_{11})	0.6027	(x_7, x_{11})	0.6013	(x_7, x_{11})	0.6024	(x_7, x_{11})	0.6039	(x_7, x_{11})	0.6034
(x_3, x_8)	0.5859	(x_3, x_8)	0.5858	(x_3, x_8)	0.5837	(x_3, x_8)	0.5850	(x_3, x_8)	0.5822
(x_{10}, x_{14})	0.5121	(x_{10}, x_{14})	0.5126	(x_{10}, x_{14})	0.5127	(x_{10}, x_{14})	0.5152	(x_{10}, x_{14})	0.5152

Результаты из табл.2 показывают наличие изменения порядка следования признаков по (9) в зависимости от процента пропусков. Другая зависимость выражается через дисперсию показателей (8). Пример такой зависимости показан на рис.1.

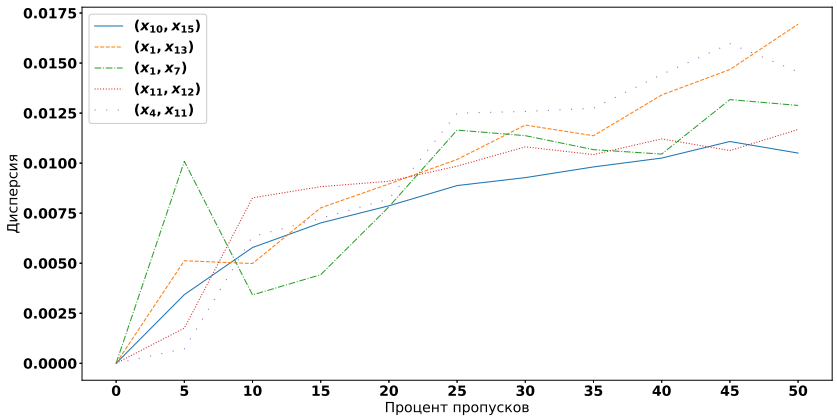

Рис. 1 Дисперсия признаков по (8) в зависимости от процента пропусков

С ростом процента пропусков по парам признаков (см. рис.1) увеличивается значение дисперсии.

5 Заключение

Методы заполнения пропусков или удаления объектов с пропусками не имеют практического применения для отбора информативных наборов признаков. В статье показана возможность построения последовательностей для отбора информативных разнотипных признаков на данных с пропусками. Инвариантность количественных признаков к масштабам их измерений достигается при использовании двух интервальных методов. За счёт свойства инвариантности возрастает практическая значимость применения последовательностей признаков при поиске скрытых закономерностей в данных из слабо структурированных предметных областей.

Литература

[1] *Россиев А. А.* Моделирование данных при помощи кривых для восстановления пробелов в таблицах. // Методы нейроинформатики, / Под. ред. А.Н. Горбаня. – Красноярск: КГТУ, 1998.

[2] *Горшенин А. К., Мартынов О. П.* Гибридные модели экстремального градиентного бустинга для восстановления пропущенных значений в данных об осадках // // Информатика и её применения, 2019. Т. 13. №3. С. 34–40. doi: `http://dx.doi.org/10.14357/19922264190306`.

[3] *Ignatyev N. A., Zguralskaya E. N., Markovtseva M. V.* Nonlinear transformation of signs and the search for patterns in the data of patients with chronic lymphocytic leukemia Proceedings of the VI International conference Information Technology and Nanotechnology, 2011. Samara, Russia, May 26-29, 2020. Session Data Science P. 333-336.

[4] *Гудфеллоу Я., Бенджио И., Курвилль А.* 2017. Глубокое обучение Пер. с англ. – М.: ДМК Пресс `https://library.kre.dp.ua/Books` С. 652.

[5] *Ignatyev N. A.* Structure Choice for Relations between Objects in Metric Classification Algorithms // Pattern Recognition and Image Analysis, 2018, 28(4). С. 695–702. doi: `http://dx.doi.org/10.1134/S1054661818040132`

[6] *Игнатьев Н. А., Лолаев М. Я.* 2021. Анализ соответствия структур отношений объектов классов на многообразиях их описаний, Информационные технологии, Москва, № 1 (27), 18-24 б

[7] *Saidov D. Y.* Data Visualization and its Proof by Compactness Criterion of Objects of Classes // International Journal of Intelligent Systems and Applications, 2017. Vol.9 №8. С. 51–58. doi: http://dx.doi.org/10.5815/ijisa.2017.08.06

[8] *Кривенко М. П.* Обучаемая классификация неполных клинических данных // Информатика и её применения, 2017. Т. 11. №3. С. 27–33. doi: http://dx.doi.org/10.14357/19922264170303.

[9] *Игнатьев Н. А.* Вычисление обобщённых показателей и интеллектуальный анализ данных // Автоматика и телемеханика, 2011. №5. С. 183–190.

[10] *Дюк В. А.* 2001. Data Mining – интеллектуальный анализ данных http://www.olap.ru/basic/dm2.asp С. 368.

[11] *Згуральская Е. Н.* 2018. Устойчивость разбиения данных на интервалы в задачах распознавания и поиск скрытых закономерностей, Известия Самарского науч. центра Рос. акад. наук. 20, № 4(3). – С. 451-455.

[12] *Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика. Классификация и снижение размерности. // Москва. Финансы и статистика. 1989. 607 с.

[13] The South Korean Open Government Data portal. URL: https://www.data.go.kr.

[14] UCI repository of machine learning databases [Electronic resource]. URL: http://archive.ics.uci.edu, 03.12.2018

[15] Инструменты численных вычислений [Electronic resource]. URL: https://numpy.org/.

Поступила в редакцию 06.10.2021

UDC 519.95

SPECIFICS OF SELECTING INFORMATIVE FEATURE SETS BASED ON DATA WITH MISSINGS

* ***Ignat'ev N.A., Rakhimova M.A., Lolaev M.Y.***

* n_ignatev@rambler.ru

National University of Uzbekistan named after Mirzo Ulugbek,

4 Universitetskaya str., Tashkent, 100174 Uzbekistan.

Informative feature sets are a source of new knowledge for weakly formalized subject areas. The absence of measured values for a number of features in the description of an object is not always a reason for refusing to make a decision on it. Decision-making problems are associated with changing relationships between objects within classes and between classes. The selection of informative sets of different types of features is considered, taking into account the presence of gaps in the data when describing class objects. To reduce the combinatorial complexity of algorithms, it is proposed to use new heuristics taking into account unmeasured values (gaps) in the description of objects. The selection condition is the property of data invariance to the scale of their measurements. Invariance is achieved through the use of methods for dividing feature values into disjoint intervals. Partitioning into intervals is used in two methods of data preprocessing, carried out in order to unify the scales of measurements of features. According to the results of

preprocessing, sequences are formed, ordered by the relation of stability of features or by the ratio of interclass differences in the values of pairs of features. The first relation is chosen for the number of classes equal to 2, the second - if the number is greater than or equal to 3. When calculating stability, the values of the class membership function are used. The change in the order of the features in the sequences is investigated depending on the percentage of gaps in the data. As performance indicators, it is recommended to consider the change in the ranks of features or the variance of their stability in ordered sequences with different fractions of gaps in the data. The results of computational experiments confirm the truth of the statement about the small variability of the stability values on samples from the general population.

Keywords: pattern type, subsequence, feature stability, missing values in dataset.

Citation: Ignat'ev N.A., Rakhimova M.A., Lolaev M.Y. 2021. Specifics of selecting informative feature sets based on data with missings. *Problems of Computational and Applied Mathematics.* 6/1(37): 113-122.