Московский государственный университет имени М. В. Ломоносова

Факультет вычислительной математики и кибернетики

Кафедра математических методов прогнозирования

ЛЬВОВ Сергей Сергеевич

Разработка оптимальных процедур вычисления оценок, основанных на системах логических закономерностей

ДИПЛОМНАЯ РАБОТА

Научный руководитель:

д.ф-м.н., профессор

В. В. Рязанов

Москва, 2015

Содержание

1 Введение

Среди всех логических алгоритмов распознавания можно выделить класс алгоритмов вычисления оценок (частичной прецедентности). В своей основе они имеют следующую идею. Для распознаваемого объекта вычисляются оценки $\Gamma_j, j = \overline{1,l}$ близости к каждому из классов $K_j, j = \overline{1,l}$. В результате объект причисляется к тому классу, оценка за который максимальна. Обычно оценки вычисляются как доля некоторых предикатов (функций близости, логических закономерностей), выполненных на распознаваемом объекте, поэтому данные алгоритмы распознавания имеют и другое название — «алгоритмы голосования».

Одним из алгоритмов вычисления оценок является голосование по системам логических закономерностей классов. Под логическими закономерностями здесь понимаются предикаты, ограничивающие значения по некоторым признакам с помощью неравенств. В случае вещественных признаков логической закономерности соответствует простая геометрическая интерпретация: в некотором признаковом подпространстве имеется гиперпараллелепипед, содержащий максимальное число объектов обучения из заданного класса и только из него.

В данной работе рассматривается новая модель типа вычисления оценок, основанная на системах логических закономерностей. Сначала в результате анализа обучающей информации для каждого класса находится система логических закономерностей. Далее в процессе построения решающего правила (которое проводит классификацию произвольного нового объекта) находятся оптимальные весовые коэффициенты линейной формы от логических закономерностей. Весь процесс можно представить как последовательное применение двух известных алгоритмов. Сначала находятся логические закономерности классов, а затем — оптимальная линейная разделяющая гиперплоскость в некотором новом признаковом пространстве. Поэтому данный подход можно считать мультиалгоритмическим, в котором последовательно работают два алгоритма.

В настоящей работе проводится подробное исследование линейного решающего правила над системами логических закономерностей. Рассматриваются возникающие проблемы теоретического и практического характера, предлагаются способы предотвращения отказов от классификации (посредством введения оценок «антиблизости» и аппроксимации) и сокращения размерности нового признакового пространства.

1.1 Основные определения и обозначения

Рассматривается задача распознавания по прецедентам (классификации) в следующей постановке [1]. Пусть имеется множество $M = \{\mathbf{x}\}$ объектов произвольной природы. Каждому объекту $\mathbf{x} \in M$ ставится в соответствие числовой вектор $(x_1, x_2, \ldots, x_n), x_i \in \mathbb{R}, i = \overline{1,n}$ — его признаковое описание. В рамках настоящей задачи будем отождествлять объект с его признаковым описанием: $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. Множество M разбивается на l непересекающихся подмножеств K_i, называемых классами: $M = \bigcup_{i=1}^{l} K_i, K_i \cap K_j = \varnothing, i, j = \overline{1,l}, i \neq j$. Также дано некоторое подмножество объектов $X = \{\mathbf{x}_i\}_{i=1}^{m} \subseteq M$, называемое обучающей выборкой. Гарантируется, что обучающая выборка содержит хотя бы по одному представителю каждого класса. Требуется построить алгоритм A, относящий произвольный объект $\mathbf{x} \in M$ к одному из классов $K_i, i = 1, 2, \ldots, l$. Далее для простоты будем считать выражения вида $(a \leqslant b)$ и $(a \leqslant b \leqslant c)$ равными единице при их истинности и равными нулю в другом случае.

Приведем несколько определений, формализующих понятие логической закономерности [2].

Определение 1. *Логической закономерностью класса* $K_\lambda, \lambda = \overline{1,l}$ называется предикат вида

$$P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}) = \bigwedge_{i\in\Omega_1} (a_i \leqslant x_i) \bigwedge_{i\in\Omega_2} (x_i \leqslant b_i), \tag{1}$$

$$a_i \in \mathbb{R}, b_i \in \mathbb{R}, i = \overline{1,n},$$

$$\Omega_1 \subseteq \{1, 2, \ldots, n\}, \Omega_2 \subseteq \{1, 2, \ldots, n\}, |\Omega_1| = k_1, |\Omega_2| = k_2,$$

если выполнены условия:

1. $\exists \mathbf{x}_j \in X \cap K_t : P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}_j) = 1,$
2. $\forall \mathbf{x}_j \notin X \cap K_t : P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}_j) = 0,$
3. $F\left(P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x})\right) = \underset{\mathbf{a}^*,\mathbf{b}^*,\Omega_1^*,\Omega_2^*}{\operatorname{extr}} F\left(P^{\mathbf{a}^*,\mathbf{b}^*,\Omega_1^*,\Omega_2^*}(\mathbf{x})\right)$, где F — некоторый критерий качества предиката.

Иными словами, логическая закономерность класса K_λ должна покрывать хотя бы один объект обучающей выборки класса K_λ, не должна покрывать ни одного объекта обучающей выборки других классов, а также являться решением некоторой задачи локальной оптимизации по критерию F.

В дальнейшем для краткости будем пользоваться сокращением ЛЗ для обозначения логической закономерности некоторого класса.

Определение 2. Предикат вида (1), удовлетворяющий только первому и второму условию, называется *допустимым предикатом класса* K_λ.

Определение 3. Предикат вида (1), удовлетворяющий только первому и третьему условию, называется *частичной логической закономерностью класса* K_λ.

Определение 4. *Стандартным критерием качества* логической закономерности класса K_λ называется

$$F\left(P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x})\right) = |\{\mathbf{x}_j \in X \cap K_t \mid P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}_j) = 1\}|.$$

Таким образом, рассматривается задача локальной максимизации числа объектов обучающей выборки класса K_λ, которые покрываются ЛЗ.

Определение 5. Две ЛЗ $P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x})$ и $P^{\mathbf{a}',\mathbf{b}',\Omega_1',\Omega_2'}(\mathbf{x})$ класса K_λ называются *эквивалентными*, если из значения совпадают на всех объектах обучающей выборки:

$$P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}_j) = P^{\mathbf{a}',\mathbf{b}',\Omega_1',\Omega_2'}(\mathbf{x}_j), j = \overline{1,m}.$$

Определение 6. *Интервалом* логической закономерности $P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x})$ называется множество вида

$$N\left(P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}\right) = \{\mathbf{x} \in \mathbb{R}^n \mid a_i \leqslant x_i, i \in \Omega_1;\ x_i \leqslant b_i, i \in \Omega_2\}.$$

Определение 7. Логическая закономерность $P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x})$ называется *минимальной*, если не существует такая эквивалентная ей ЛЗ $P^{\mathbf{a}',\mathbf{b}',\Omega_1',\Omega_2'}(\mathbf{x})$, что

$$N\left(P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}\right) \subset N\left(P^{\mathbf{a}',\mathbf{b}',\Omega_1',\Omega_2'}\right).$$

Логические закономерности со стандартным критерием качества имеют следующую геометрическую интерпретацию: их интервалы представляют собой прямоугольные координатные гиперпараллелепипеды в признаковом пространстве $\mathbb{R}^n$, содержащие максимальное количество объектов обучающей выборки своего класса и не содержащие объектов обучающей выборки других классов. При том одна или обе границы по некоторым из признаков могут быть открытыми.

Очевидно, что для минимальных ЛЗ имеет место $\Omega_1 = \Omega_2 = \Omega$, а их интервалы «натянуты» на некоторые подмножества объектов обучающей выборки своего класса.

1.2 Алгоритм распознавания, основанный на голосовании по системам логических закономерностей

Пусть по обучающей выборке X для каждого класса K_λ найдены множества логических закономерностей $\mathbf{P}_\lambda = \{P_t(x)\}, \lambda = \overline{1,l}$. Распознавание, основанное на голосовании по системам логических закономерностей, осуществляется как стандартная процедура вычисления оценок. Оценка близоста нового объекта $\mathbf{x}$ к классу K_λ вычисляется по формуле

$$\Gamma_\lambda(\mathbf{x}) = \sum_{P_t \in \mathbf{P}_\lambda} \gamma_t P_t(\mathbf{x}), \tag{2}$$

где весовые коэффициенты $0 < \gamma_t < 1$ задаются одинаковыми для всех ЛЗ класса K_λ:

$$\gamma_t = \frac{1}{|\mathbf{P}_\lambda|}.$$

Далее объект $\mathbf{x}$ причисляется к тому классу K_λ, оценка $\Gamma_\lambda(\mathbf{x})$ за который максимальна [2].

Утверждение 1. Указанный алгоритм корректно классифицирует обучающую выборку X при условии, что каждый ее объект покрывается хотя бы одной логической закономерностью.

2 Модификация решающего правила

При вычислении оценок близости по формуле (2) в признаковом пространстве могут возникать большие зоны с нулевыми оценками за все классы. Для иллюстрации приведем простой пример (рис. 1). В задаче имеется два хорошо отделимых класса,

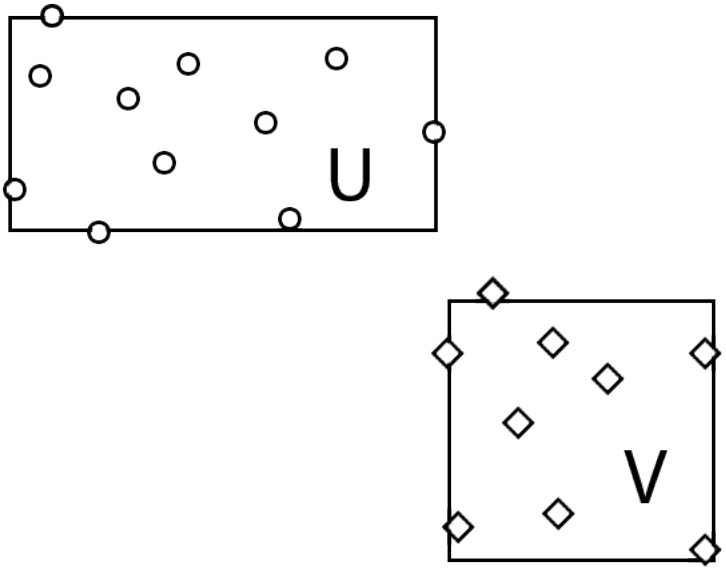


Рис. 1: Пример двумерной задачи с двумя классами

каждый из которых описывается одной минимальной ЛЗ. Интервалы закономерностей обозначены как U и V. При распознавании любого объекта, лежащего вне указанных интервалов, оценки за каждый из классов будут равны нулю, и произойдет отказ от классификации. Подобный эффект можно устранить путем введением оценок «антиблизости» — отрицаний логических закономерностей (таким образом, будет покрываться все признаковое пространство) или аппроксимацией ЛЗ гладкими потенциальными функциями.

Также представляется полезным найти способ задания весовых коэффициентов γ_t в формуле (2), позволяющий учитывать не только количество ЛЗ класса K_λ, но и специфику каждой отдельной логической закономерности.

2.1 Использование отрицаний дизъюнкций логических закономерностей

Здесь и далее рассматривается случай двух классов. Для вычисления оценок близости объекта $\mathbf{x}$ вместо (2) будем использовать выражение

$$\begin{aligned} f(\mathbf{x}) &= \sum_{i=1,2,\ldots,m(1)} \alpha_i^1 P_i^1(\mathbf{x}) + \alpha_0^1 \overline{\bigvee_{i=1,2,\ldots,m(2)} P_i^2(\mathbf{x})} - \\ &- \sum_{i=1,2,\ldots,m(2)} \alpha_i^2 P_i^2(\mathbf{x}) - \alpha_0^2 \overline{\bigvee_{i=1,2,\ldots,m(1)} P_i^1(\mathbf{x})}, \end{aligned} \tag{3}$$

где $P_i^1(\mathbf{x}), i = 1, 2, \ldots, m(1)$ и $P_i^2(\mathbf{x}), i = 1, 2, \ldots, m(2)$ — вычисленные ЛЗ первого и второго класса соответственно, $\alpha_i^1, \alpha_i^2, \alpha_0^1, \alpha_0^2$ — весовые коэффициенты. Объект $\mathbf{x}$ будем относить к первому классу, если $f(\mathbf{x}) > 0$, а при $f(\mathbf{x}) < 0$ — ко второму классу. При $f(\mathbf{x}) = 0$ происходит отказ от распознавания или случайная классификация.

Построение функции $f(\mathbf{x})$ можно рассматривать как последовательное решение двух задач:

1. Нахождение множеств логических закономерностей $P_i^1(\mathbf{x}), i = 1, 2, \ldots, m(1)$, $P_i^2(\mathbf{x}), i = 1, 2, \ldots, m(2)$ и переход к новому бинарному признаковому пространству размерности $m(1) + m(2) + 2$.

2. Вычисление весовых коэффициентов $\alpha_i^1, \alpha_i^2, \alpha_0^1, \alpha_0^2$ посредством построения в новом признаковом пространстве оптимальной разделяющей гиперплоскости, используя линейный метод, например, «метод опорных векторов» «линейный дискриминант Фишера» или «линейная машина» [3].

Выбор на втором шаге линейного классификатора обусловлен тем фактом, что в новом признаковом пространстве объектам обучающей выборки первого класса будут соответствовать векторы вида $\mathbf{y} = (\sigma_1, \sigma_2, \ldots, \sigma_{m(1)}, 1, 0, 0, \ldots, 0)$, а объектам второго — $\mathbf{z} = (0, 0, \ldots, 0, \theta_1, \theta_2, \ldots, \theta_{m(2)}, 1)$, и, значит, классы будут линейно разделимы в данном пространстве.

Отметим также, что из линейной разделимости классов в новом признаковом пространстве следует, что алгоритм корректно классифицирует объекты обучающей выборки.

2.2 Использование аппроксимаций логических закономерностей

Другим способом решения проблемы отказов является введение аппроксимаций логических закономерностей гладкими функциями. В качестве потенциальных функций могут использоваться любые гладкие функции одного аргумента, имеющие единственный максимум и стремящиеся к нулю при $x \to \pm\infty$, например, гауссианы. Однако, как было указано ранее, границы ЛЗ по некоторым признакам могут быть открытыми (в этом случае часть ограничений, накладываемых ЛЗ вырождаются в пороговые функции), поэтому наиболее удобным представляется использование сигмоид (логистических функций):

$$P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}) \approx \Phi^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}) = \prod_{i\in\Omega_1} \frac{1}{1+\exp(-\alpha_i(x_i-a_i))} \prod_{i\in\Omega_2} \frac{1}{1+\exp(\beta_i(x_i-b_i))},$$

$\alpha_i > 0, \beta_i > 0$. Таким образом, аппроксимация ЛЗ сводится к поиску коэффициентов α_i, β_i.

Упростим изложенную задачу следующим образом. Примем $\alpha_i = \beta_i \equiv \alpha$ и сведем поиск α к решению уравнения

$$\frac{1}{(1+\exp(\alpha\sigma))^k} = 1 - \sigma,$$

где $0 < \sigma < 1$ — параметр алгоритма, который задает «граничную область» отдельного сомножителя и «максимум» функции $\Phi^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x})$, $k = |\Omega_1| + |\Omega_2|$.

Окончательно аппроксимация $P^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x})$ вычисляется как

$$\Phi^{\mathbf{a},\mathbf{b},\Omega_1,\Omega_2}(\mathbf{x}) = \prod_{i\in\Omega_1} \frac{1}{1+\exp(-\alpha(x_i-a_i))} \prod_{i\in\Omega_2} \frac{1}{1+\exp(\alpha(x_i-b_i))}. \quad (4)$$

Простой пример одномерной аппроксимации с параметром $\alpha = 1$ приведен на рис. 2.

Оценки близости для объекта $\mathbf{x}$ вычисляются по формуле, аналогичной (3) (за исключением использования отрицаний дизъюнкций):

$$f(\mathbf{x}) = \sum_{i=1,2,\ldots,m(1)} \alpha_i^1 \Phi_i^1(\mathbf{x}) - \sum_{i=1,2,\ldots,m(2)} \alpha_i^2 \Phi_i^2(\mathbf{x}). \quad (5)$$

При $f(\mathbf{x}) > 0$ объект $\mathbf{x}$ будет относиться к первому классу, при $f(\mathbf{x}) < 0$ — ко второму, а при $f(\mathbf{x}) = 0$ произойдет отказ от распознавания или случайная классификация.

В данном случае построение функции $f(\mathbf{x})$ также можно рассматривать как выполнение двух последовательных шагов:

1. Нахождение множеств логических закономерностей $P_i^1(\mathbf{x}), i = 1, 2, \ldots, m(1)$, $P_i^2(\mathbf{x}), i = 1, 2, \ldots, m(2)$, вычисление их аппроксимаций $\Phi_i^1(\mathbf{x})$ и $\Phi_i^2(\mathbf{x})$ и переход к новому вещественному признаковому пространству размерности $m(1) + m(2)$.

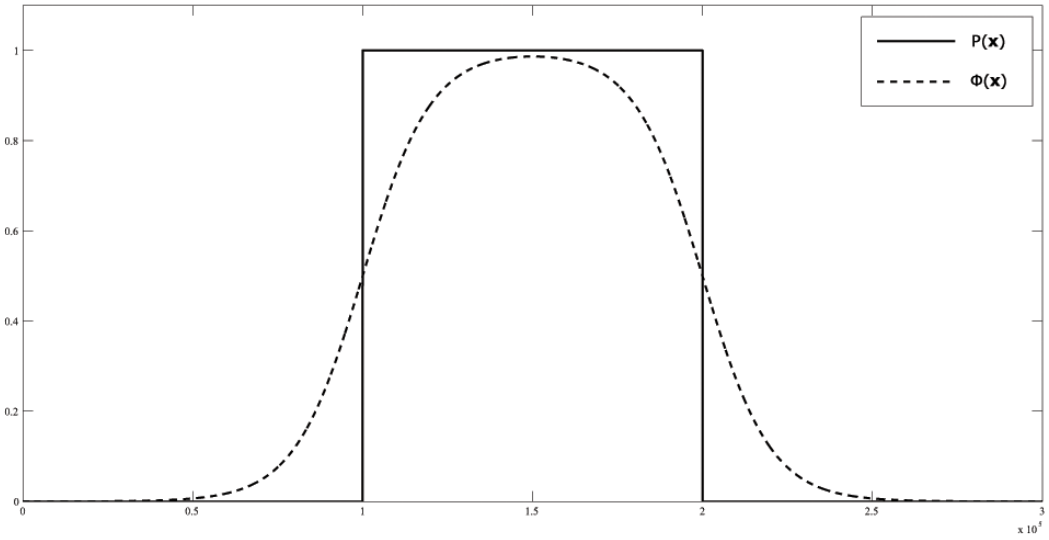


Рис. 2: Пример аппроксимации

2. Вычисление весовых коэффициентов α_i^1, α_i^2 посредством построения в новом признаковом пространстве оптимальной разделяющей гиперплоскости.

Отметим, что в описанном методе не утверждается о линейной разделимости классов в новом признаковом пространстве. Применение аппроксимаций также может нарушать корректность классификации обучающей выборки.

2.3 Случай многих классов

В описании алгоритмов предполагалось, что в рассматриваемых задачах присутствует только 2 класса. Случай, когда число классов больше двух, разрешается следующим образом. Непосредственно распознавание новых объектов происходит линейным классификатором в новом признаковом пространстве, поэтому здесь возможно применение стандартных стратегий сведения многоклассовой классификации к двухклассовой. Например, при использовании «метода опорных векторов» часто используется стратегия «один против всех».

3 Кластеризация логических закономерностей

Рассмотрим следующий пример (рис. 3): Здесь объекты первого класса изображе-

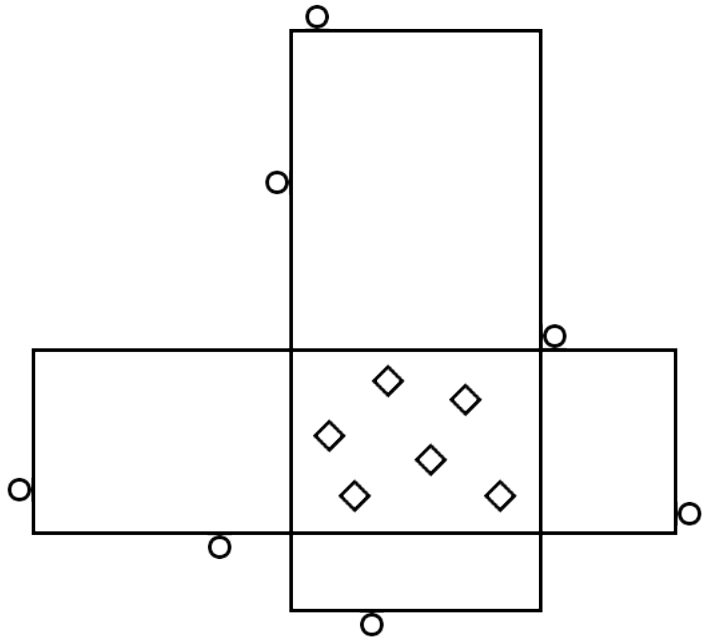


Рис. 3

ны ромбами, объекты второго — кружками. Показаны две логические закономерности первого класса, упирающиеся в объекты второго класса. По определению, указанные ЛЗ являются эквивалентными, т.е. неотличимыми на объектах обучающей выборки, поэтому имеет смысл исключить одну из закономерностей из системы.

Также было замечено, что при увеличении обучающей выборки (росте m) увеличивается и количество ЛЗ для каждого из классов, а сами ЛЗ «измельчаются», а значит, растет размерность нового признакового пространства. Этот факт обусловлен тем, что при больших объемах обучающей выборки становится трудно построить ЛЗ, охватывающие много объектов одного класса. Некоторые из полученных логических закономерностей эквивалентны (или почти эквивалентны), поэтому необходимо проводить «прореживание» среди близких друг к другу логических закономерностей. Для этого будем применять кластеризацию ЛЗ каждого из классов.

Каждой логической закономерности $P(\mathbf{x})$ можно поставить в соответствие бинарный вектор вида $\mathbf{p} = (p_1, p_2, \ldots, p_m)$, где $p_i = P(\mathbf{x}_i)$ — значение ЛЗ на i-м объекте обучающей выборки. На множествах $\mathbf{P}_\lambda = \{P_t(x)\}, \lambda = \overline{1,l}$ проведем кластеризацию методом минимизации дисперсионного критерия [4].

Обозначим кластеры как $L_i, i = \overline{1,k}$, а $n_i, i = \overline{1,k}$ — число объектов в каждом из кластеров. Разбиением объектов на кластеры назовем $L = \bigcup_{i=1}^{k} L_i, L_i \cap L_j = \varnothing, i, j = \overline{1,l}, i \neq j$ (по аналогии с разбиением объектов на классы). Задача кластеризации сводится к минимизации функционала J разброса объектов внутри каждого из кластеров:

$$\begin{aligned} J &= \sum_{i=1}^{k} J_i \to \min, \\ J_i &= \sum_{\mathbf{p} \in L_i} \|\mathbf{p} - \mathbf{m}_i\|^2, \end{aligned} \tag{6}$$

где $\mathbf{m}_i$ — центр кластера L_i:

$$\mathbf{m}_i = \frac{\sum_{\mathbf{p} \in L_i} \mathbf{p}}{n_i}.$$

Определение 8. *Соседним разбиением* называется разбиение, отличающееся от исходного переносом единственного объекта из одного кластера в другой.

Рассмотрим, как отличаются соседние разбиения. Пусть в разбиении L объект $\hat{\mathbf{p}}$ переносится из кластера L_i в L_j. Тогда в новом разбиении L' кластеры выглядят следующим образом:

$$\begin{aligned} L'_i &= L_i \setminus \{\hat{\mathbf{p}}\}, \\ L'_j &= L_j \cup \{\hat{\mathbf{p}}\}, \\ L'_\nu &= L_\nu, \nu \neq i, j. \end{aligned}$$

Новый центр кластера L'_j может быть вычислен по формуле

$$\begin{aligned} \mathbf{m}'_j = \frac{\sum_{\mathbf{p} \in L'_j} \mathbf{p}}{n'_j} = \frac{\sum_{\mathbf{p} \in L_j} \mathbf{p} + \hat{\mathbf{p}}}{n_j + 1} = \frac{\frac{\sum_{\mathbf{p} \in L_j} \mathbf{p} + \hat{\mathbf{p}}}{n_j}}{\frac{n_j + 1}{n_j}} = \\ = \frac{n_j}{n_j + 1} \mathbf{m}_j + \frac{\hat{\mathbf{p}}}{n_j + 1} = \mathbf{m}_j + \frac{\hat{\mathbf{p}} - \mathbf{m}_j}{n_j + 1}, \end{aligned}$$

а компонента J_j функционала J в формуле (6) может быть пересчитана как

$$\begin{aligned} J'_j = \sum_{\mathbf{p} \in L'_j} \|\mathbf{p} - \mathbf{m}'_j\|^2 = \sum_{\mathbf{p} \in L_j} \|\mathbf{p} - \mathbf{m}'_j\|^2 + \|\hat{\mathbf{p}} - \mathbf{m}'_j\|^2 = \\ = \sum_{\mathbf{p} \in L_j} \|\mathbf{p} - \mathbf{m}_j - \frac{\hat{\mathbf{p}} - \mathbf{m}_j}{n_j + 1}\|^2 + \frac{n_j^2}{(n_j + 1)^2} \|\hat{\mathbf{p}} - \mathbf{m}_j\|^2 = \\ = \sum_{\mathbf{p} \in L_j} \|\mathbf{p} - \mathbf{m}_j\|^2 + \sum_{\mathbf{p} \in L_j} \frac{\|\hat{\mathbf{p}} - \mathbf{m}_j\|^2}{(n_j + 1)^2} - 2 \sum_{\mathbf{p} \in L_j} \left\langle \mathbf{p} - \mathbf{m}_j, \frac{\hat{\mathbf{p}} - \mathbf{m}_j}{n_j + 1} \right\rangle + \frac{n_j^2}{(n_j + 1)^2} \|\hat{\mathbf{p}} - \mathbf{m}_j\|^2 = \\ = J_j + \frac{(n_j + n_j^2)\|\hat{\mathbf{p}} - \mathbf{m}_j\|^2}{(n_j + 1)^2} = J_j + \frac{n_j}{n_j + 1} \|\hat{\mathbf{p}} - \mathbf{m}_j\|^2. \end{aligned}$$

Аналогично можно получить

$$\mathbf{m}_i' = \mathbf{m}_i - \frac{\hat{\mathbf{p}} - \mathbf{m}_i}{n_i - 1},$$
$$J_i' = J_i - \frac{n_i}{n_i - 1}\|\hat{\mathbf{p}} - \mathbf{m}_i\|^2.$$

Таким образом, можно утверждать, что перенос объекта $\hat{\mathbf{p}}$ из L_i в L_j выгоден в том случае, если уменьшение J_i' больше увеличения J_j':

$$\frac{n_i}{n_i - 1}\|\hat{\mathbf{p}} - \mathbf{m}_i\|^2 > \frac{n_j}{n_j + 1}\|\hat{\mathbf{p}} - \mathbf{m}_j\|^2.$$

Итоговый метод кластеризации представлен алгоритмом 1.

Алгоритм 1. Кластеризация методом минимизации дисперсионного критерия

Вход: кластеризируемые объекты $\mathbf{p}_i$ класса K_λ, $i = \overline{1, m(\lambda)}$, $m(\lambda) = |\mathbf{P}_\lambda|$; число кластеров k; максимальное число итераций max_iter;

Выход: метки объектов $label_i \in \{1, 2, \ldots, k\}, i = \overline{1, m(\lambda)}$;

1: инициализировать кластеры;
2: вычислить центры $\mathbf{m}_i$ и значение минимизируемого функционала J;
3: $iter := 1$;
4: **пока** $iter < max\_iter$
5: для всех $\mathbf{p}_i$, $i = \overline{1, m(\lambda)}$ и для всех кластеров L_j, $j = \overline{1, k}$, $j \neq label_i$ вычислить изменение функционала J при перенесении $\mathbf{p}_i$ в L_j:
$\delta_{ij} = \frac{n_{label_i}}{n_{label_i} - 1}\|\hat{\mathbf{p}} - \mathbf{m}_{label_i}\|^2 - \frac{n_j}{n_j + 1}\|\hat{\mathbf{p}} - \mathbf{m}_j\|^2$;
6: $\delta := \max\limits_{i=\overline{1,m(\lambda)},\ j=\overline{1,k}} \delta_{ij}$;
7: **если** $\delta \leqslant 0$ **то**
8: останов;
9: **иначе**
10: $label_i := j$;
11: пересчитать $\mathbf{m}_{label_i}$, $\mathbf{m}_j$ и J;
12: $iter := iter + 1$;

Для инициализации кластеров предлагается использовать метод, описанный в [5]. Среди объектов необходимо выбрать k центроидов. Первый центроид выбирается случайным образом. Далее, для каждого объекта находится квадрат расстояния до ближайшего найденного центроида, и следующий центроид выбирается выбирается среди объектов с вероятностью, пропорциональной вычисленным квадратам расстояний. Этот

шаг повторяется до тех пор, пока не будут выбраны все k центроидов. Затем каждый объект приписывается к кластеру с ближайшим центроидом.

В каждом полученном кластере мы хотим оставить одну наиболее типичную логическую закономерность. Можно рассмотреть новый объект, соответствующий выборочному среднему всех объектов кластера и бинаризовать его по некоторому порогу. Для восстановление стандартного вида ЛЗ по ее бинарному описанию необходимо взять в качестве границ по признаку x_j минимальное и максимальное значение по всем объектам обучающей выборки $\mathbf{x}_i$, для которых $p_i = 1$. В этом случае, вообще говоря, полученный объект может не являться логической закономерностью, поэтому предлагается оставлять в кластере ту ЛЗ, чье бинарное описание ближе всего по Евклидовой метрике к центроиду кластера.

4 Вычислительные эксперименты

Алгоритмы перехода к новому признаковому пространству и кластеризации логических закономерностей были реализованы на языке C++. Входными данными являются файлы выборок, а также файлы исходных систем ЛЗ, сгенерированные с помощью системы РАСПОЗНАВАНИЕ [4]. Результатом вычислений являются те же выборки в новом признаковом пространстве, которые затем используются для проведения классификации в системе РАСПОЗНАВАНИЕ.

Предложенные в работе методы были протестированы на наборах модельных и реальных данных. Во всех примерах для визуализации использовалась проекция многомерных данных на плоскость обобщенных признаков [3][4].

4.1 Исходные данные и условия эксперимента

В качестве модельных задач использовались выборки из смеси многомерных нормальных распределений. Математические ожидания и матрицы ковариации выбирались таким образом, чтобы классы не являлись линейно отделимыми. Вся выборка разделялась на обучающую и контрольную в пропорции $4:1$. В дальнейшем конкретной «модельной задачей» является задача со следующими характеристиками: 2 класса, 3 признака, 240 объектов обучающей и 60 объектов контрольной выборки.

Практические данные представлены следующими задачами различных отраслей:

- диагностика рака груди (далее «breast») [6];
- прогнозирование утверждения выдачи кредитных карт (далее «credit») [7];
- сегментация изображений (далее «Image») [7];

Задача	Число классов	Число признаков	Число объектов обучающей выборки	Число объектов контрольной выборки
Модельная	2	3	240	60
«breast»	2	9	344	355
«credit»	2	15	342	348
«Image»	7	16	210	2100

Таблица 1: Характеристики задач

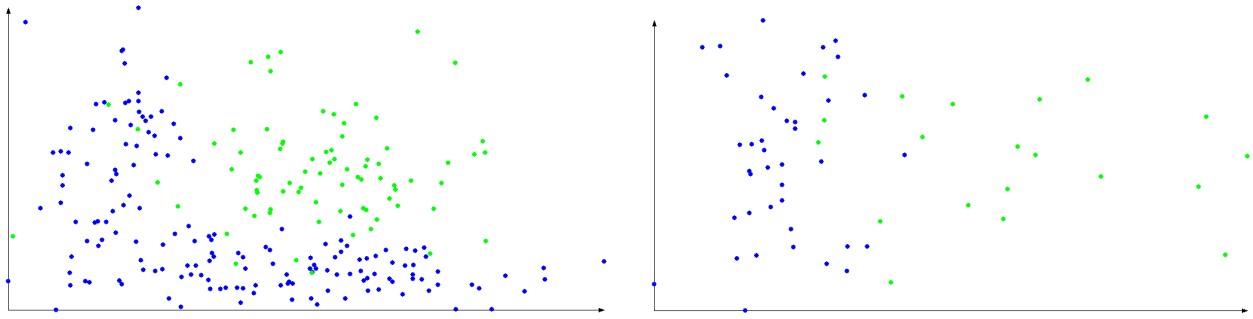
Характеристики данных приведены в таблице 1. Во всех реальных задачах пропуски данных заполнялись средним значением по признаку.

Существуют различные алгоритмы построения систем логических закономерностей [8]. Для экспериментов использовались два из них: первый позволяет находить только минимальные логические закономерности (далее «точный» метод), второй не ограничивается ими, но решает оптимизационную задачу приближенно, сводя ее к задаче поиска максимальной совместной подсистемы системы линейных неравенств (далее «приближенный» метод). Также «приближенный» метод допускает некоторое количество частичных закономерностей. Таким образом, по каждому из наборов данных строилось 4 новых набора: по «точным» и «приближенным» ЛЗ, с использованием отрицаний дизъюнкций ЛЗ или сигмоидных аппроксимаций. Параметр алгоритма аппроксимации: $\sigma = 0.9$.

Для сравнения результатов классификации в качестве линейного метода использовался «метод опорных векторов». При этом для настройки его параметров на обучении использовался скользящий контроль по 10 блокам.

4.2 Результаты эксперимента

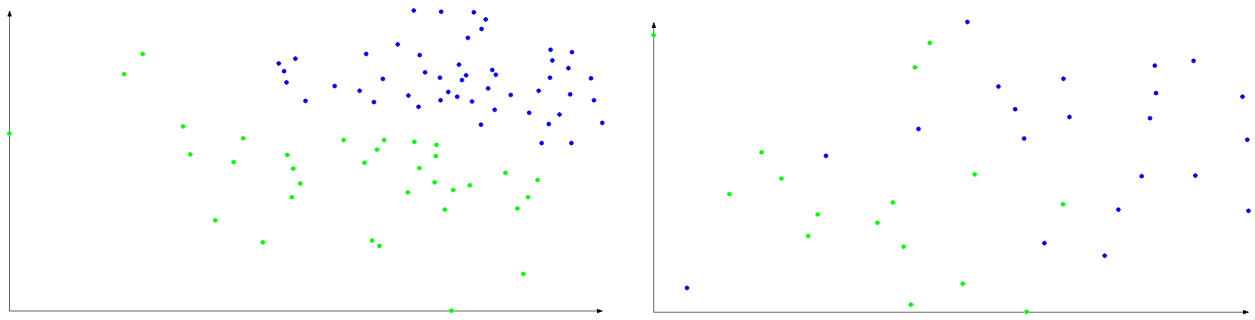
Некоторые визуализации обучающих и контрольных данных модельной задачи и задач «breast» и «credit» в исходном и новых признаковых пространствах приведены на рис 4–16. Визуализация задачи «Image» не приведена ввиду большого числа классов.



(a) Обучающая выборка

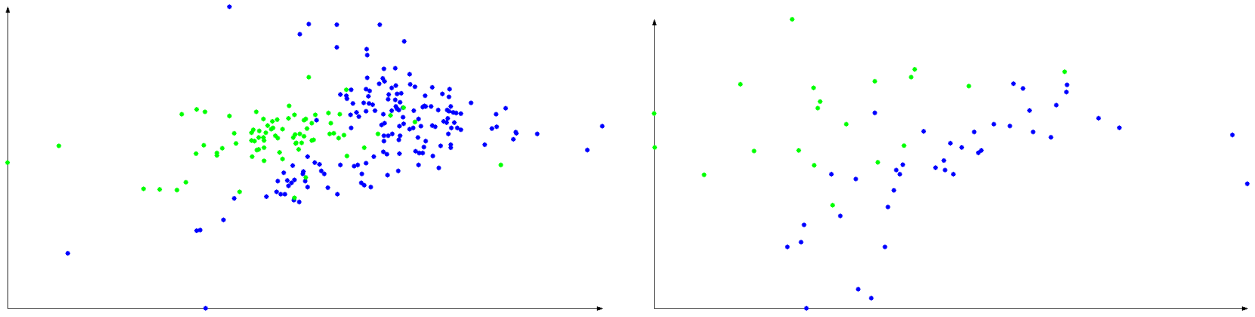
(b) Контрольная выборка

Рис. 4: Визуализации данных модельной задачи в исходном признаковом пространстве



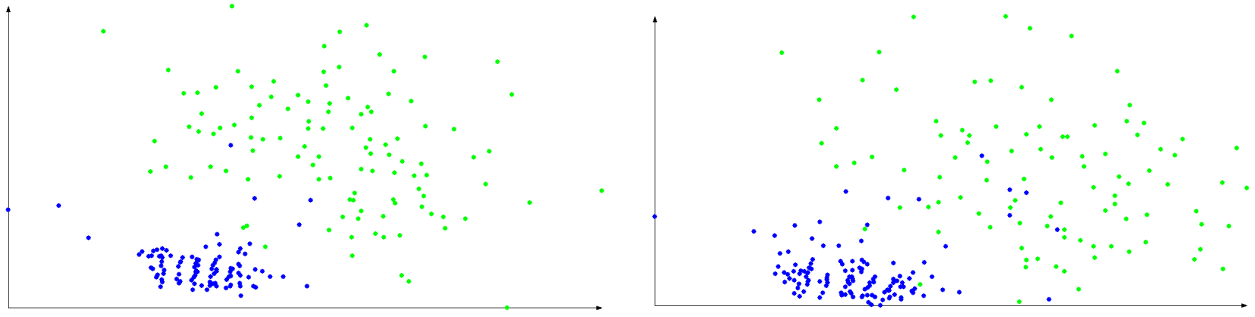
(a) Обучающая выборка (b) Контрольная выборка

Рис. 5: Визуализации данных модельной задачи в новом признаковом пространстве, «точные» ЛЗ и отрицания дизъюнкций



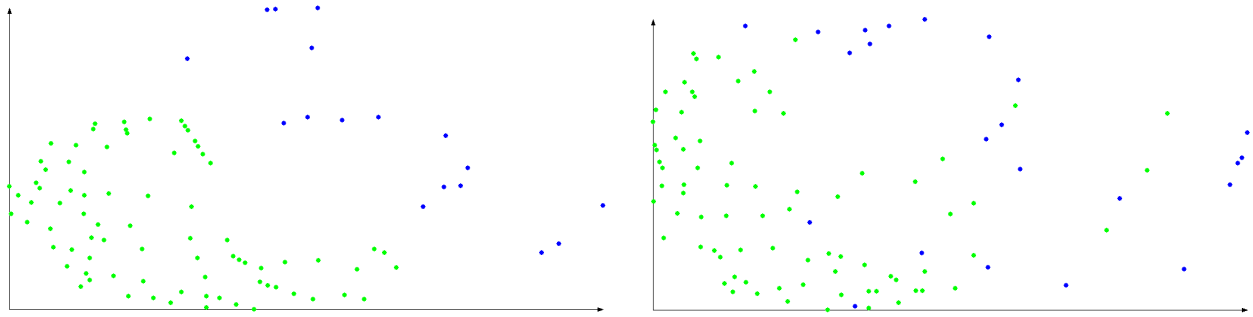
(a) Обучающая выборка (b) Контрольная выборка

Рис. 6: Визуализации данных модельной задачи в новом признаковом пространстве, «точные» ЛЗ с аппроксимацией



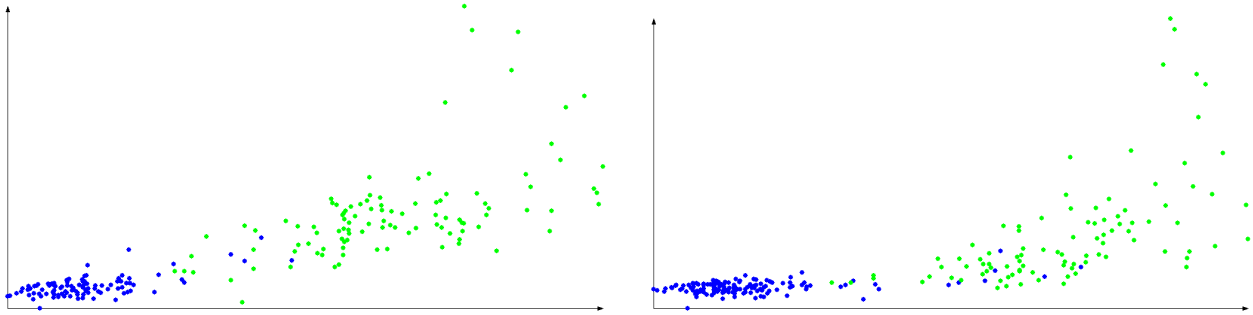
(a) Обучающая выборка (b) Контрольная выборка

Рис. 7: Визуализации данных задачи «breast» в исходном признаковом пространстве



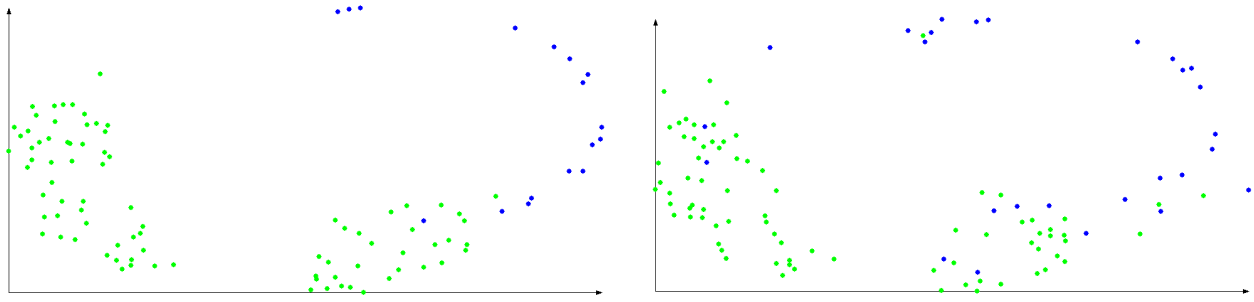
(a) Обучающая выборка (b) Контрольная выборка

Рис. 8: Визуализации данных задачи «breast» в новом признаковом пространстве, «точные» ЛЗ и отрицания дизъюнкций



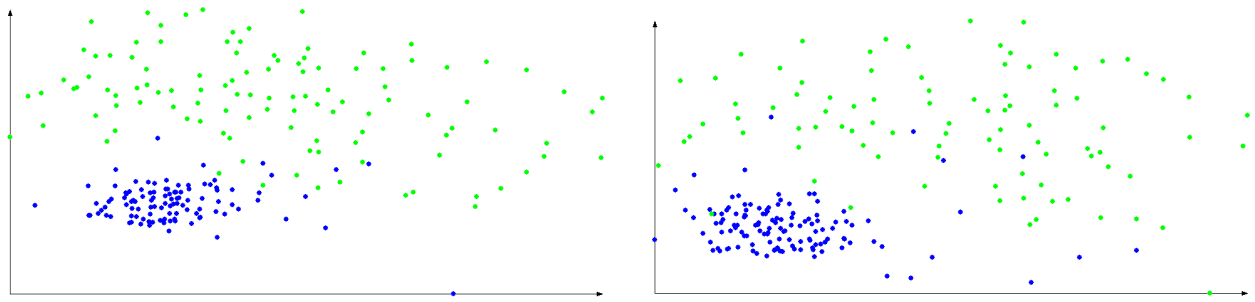
(a) Обучающая выборка (b) Контрольная выборка

Рис. 9: Визуализации данных задачи «breast» в новом признаковом пространстве, «точные» ЛЗ с аппроксимацией



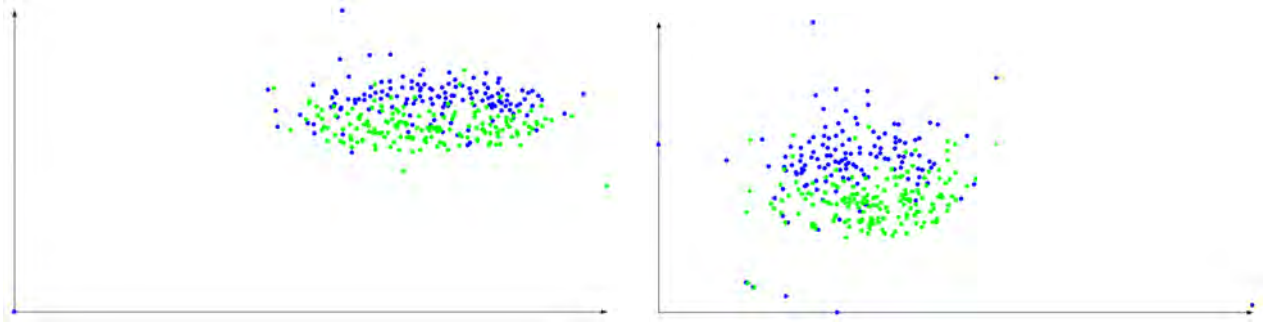
(a) Обучающая выборка (b) Контрольная выборка

Рис. 10: Визуализации данных задачи «breast» в новом признаковом пространстве, «приблизиженные» ЛЗ и отрицания дизъюнкций



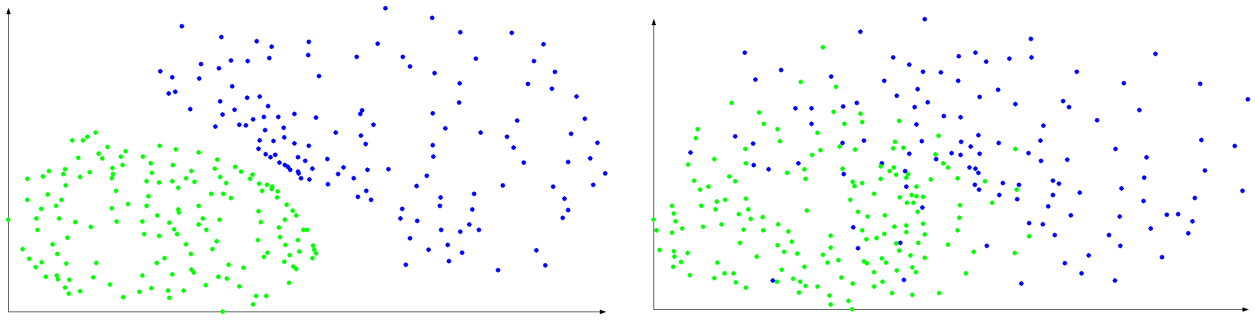
(a) Обучающая выборка (b) Контрольная выборка

Рис. 11: Визуализации данных задачи «breast» в новом признаковом пространстве, «приближенные» ЛЗ с аппроксимацией



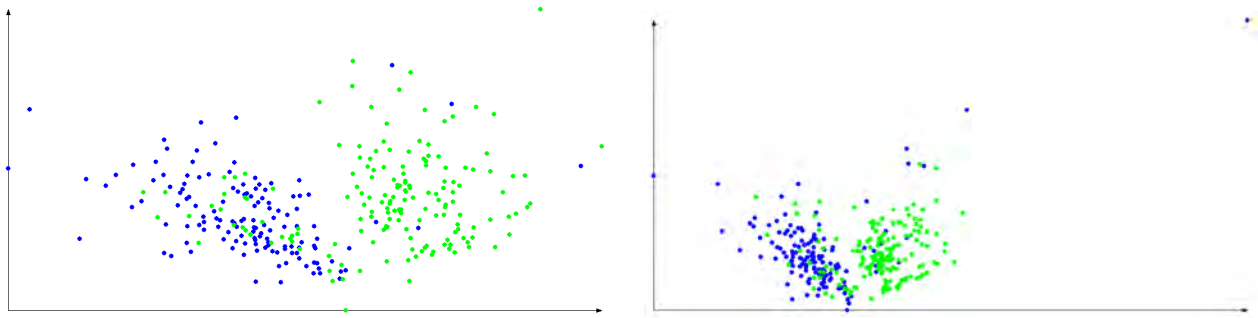
(a) Обучающая выборка (b) Контрольная выборка

Рис. 12: Визуализации данных задачи «credit» в исходном признаковом пространстве



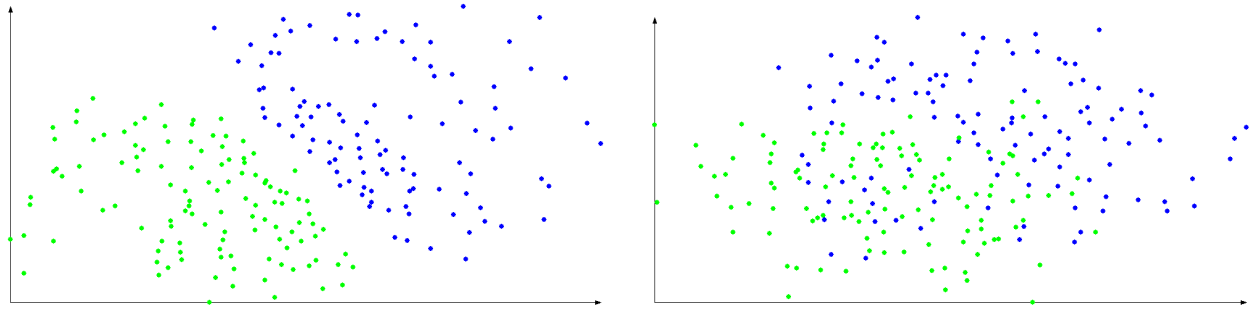
(a) Обучающая выборка (b) Контрольная выборка

Рис. 13: Визуализации данных задачи «credit» в новом признаковом пространстве, «точные» ЛЗ и отрицания дизъюнкций



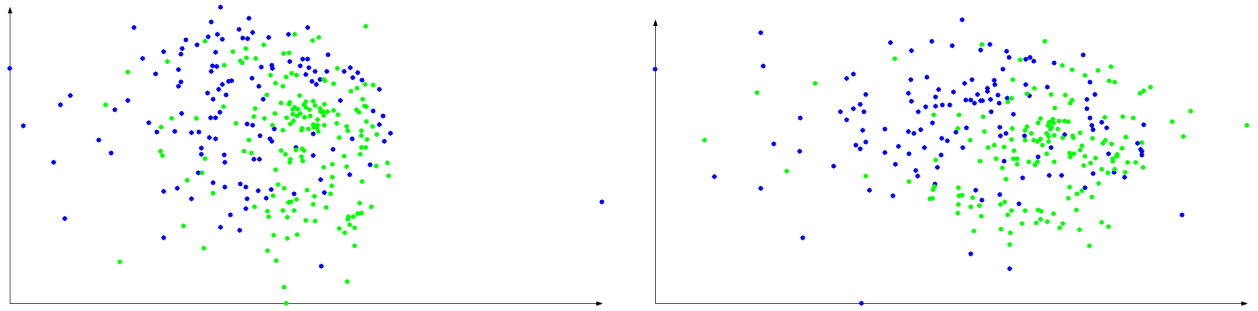
(a) Обучающая выборка (b) Контрольная выборка

Рис. 14: Визуализации данных задачи «credit» в новом признаковом пространстве, «точные» ЛЗ с аппроксимацией



(a) Обучающая выборка (b) Контрольная выборка

Рис. 15: Визуализации данных задачи «credit» в новом признаковом пространстве, «приближенные» ЛЗ и отрицания дизъюнкций



(a) Обучающая выборка (b) Контрольная выборка

Рис. 16: Визуализации данных задачи «credit» в новом признаковом пространстве, «приближенные» ЛЗ с аппроксимацией

Задача	Оригинальный метод	ЛЗ и отрицания дизъюнкций	Аппроксимация ЛЗ
Модельная	85% (3.3% отказов)	86.8%	88.3%
«breast»	94.6% (0.8% отказов)	95.8%	96.1%
«credit»	80.5% (4.3% отказов)	85.6%	84.5%
«Image»	68.8% (27.7% отказов)	73.8%	92% (0.6% отказов)

Таблица 2: Результаты тестов при использовании «точных» ЛЗ

Задача	Оригинальный метод	ЛЗ и отрицания дизъюнкций	Аппроксимация ЛЗ
Модельная	58.3% (25% отказов)	65%	80% (3.3% отказов)
«breast»	95.2% (1.1% отказов)	94.3%	94.6%
«credit»	76.1% (4.3% отказов)	82.2%	84.8%
«Image»	93.1%	93.8%	97.4% (0.4% отказов)

Таблица 3: Результаты тестов при использовании «приближенных» ЛЗ

Результаты тестов при использовании «точных» ЛЗ приведены в таблице 2, при использовании «приближенных» ЛЗ – в таблице 3. Указан процент верных классификаций контрольных данных и, если были отказы от распознавания, процент отказов.

В таблице 4 и на рис. 17 представлен пример кластеризации ЛЗ и отрицаний их дизъюнкций, сгенерированных «точным» методом для задачи «breast». На рисунке 18 представлен пример кластеризации аппроксимированных ЛЗ, сгенерированных «точным» методом для задачи «cluster». Под количеством кластеров подразумевается их суммарное количество по всем классам. Штриховой линией выделена размерность исходного признакового пространства.

Количество кластеров	Качество классификации
19	95.5%
18	95.5%
17	95.5%
16	94.9%
15	94.9%
14	94.9%
13	94.9%
12	94.9%
11	94.9%
10	94.9%
9	94.9%
8	94.9%
7	93.2%
6	93.2%
5	93.2%
4	89.8%
3	82.8%
2	79.4%

Таблица 4: Результаты кластеризации ЛЗ в задаче «breast»

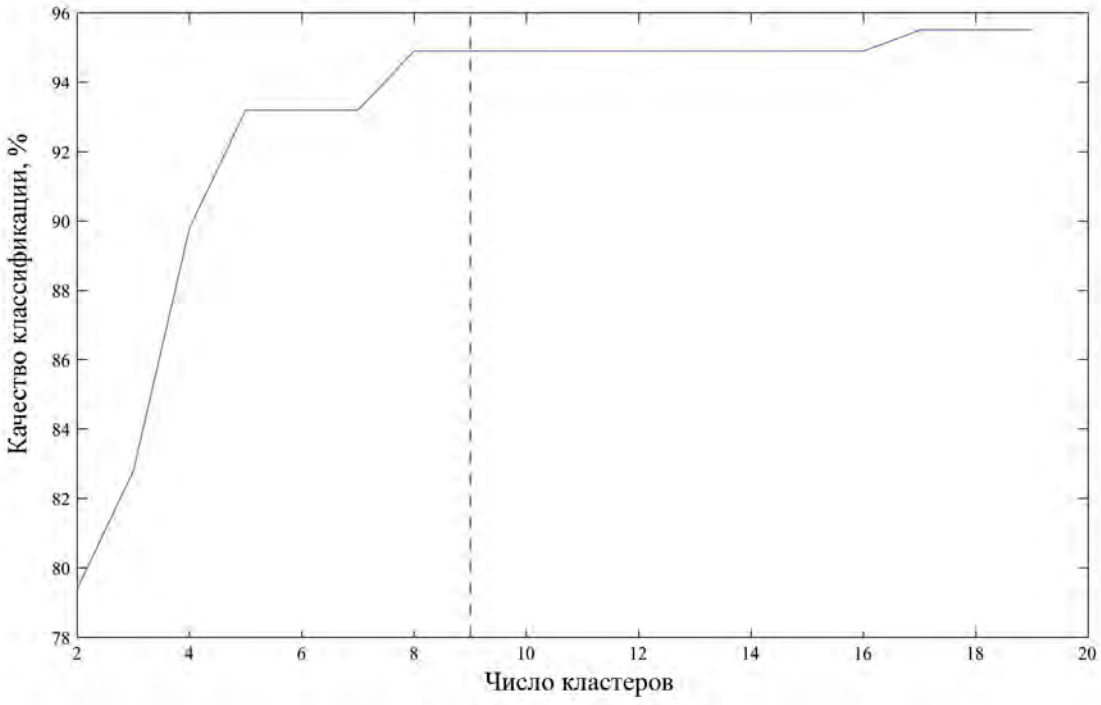


Рис. 17: Результаты кластеризации в задаче «breast»

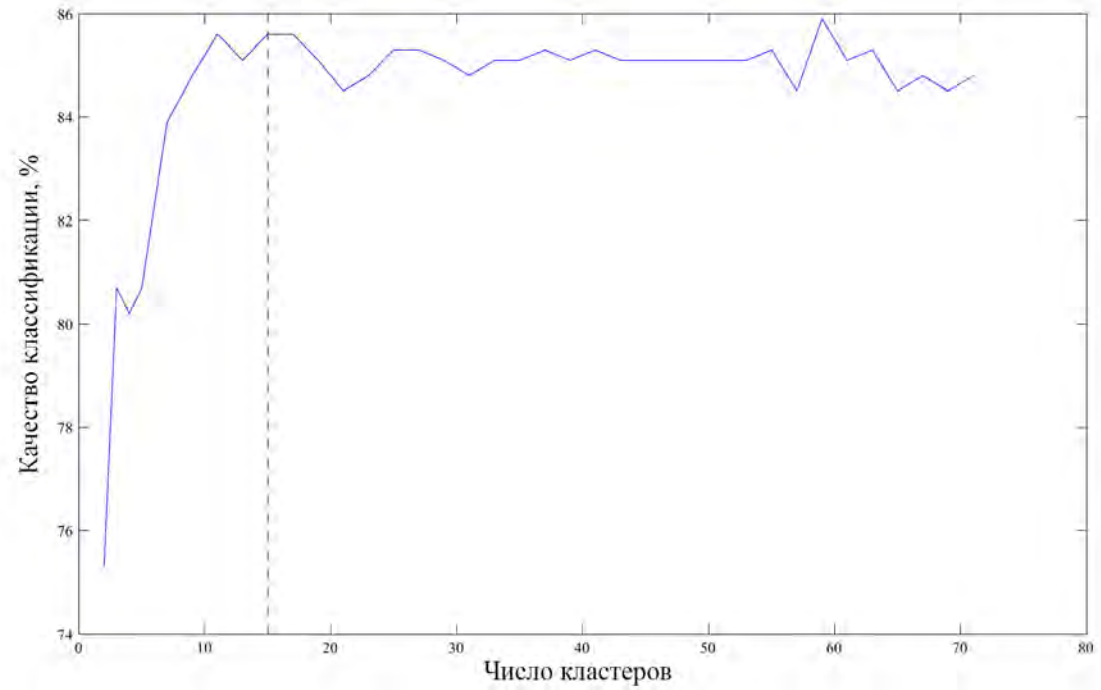


Рис. 18: Результаты кластеризации в задаче «credit»

4.3 Обсуждение и выводы

По визуализациям данных видно, что в новом признаковом пространстве, построенном по ЛЗ и отрицаниям их дизъюнкций, классы обучающей выборки действительно становятся линейно разделимыми, отделимость классов на контроле также улучшается. При использовании «приближенного» метода строгая отделимость может нарушаться из-за использования частичных закономерностей. При использовании аппроксимаций ЛЗ линейной разделимости классов не наблюдается.

По результатам исследований можно утверждать, что использование представленного в данной работе метода построения решающего правила (как с отрицаниями дизъюнкций ЛЗ, так и с аппроксимацией их сигмоидами) позволяет повысить качество классификации по сравнению с исходным оригинальным методом. Применение метода, использующего отрицания дизъюнкций ЛЗ, позволило полностью избавиться от отказов во всех рассмотренных задачах. Однако на модельных данных он показал неудовлетворительные результаты: процент правильно распознанных им объектов сравним с процентом верных классификаций оригинального метода плюс половиной отказов, т.е. таких же результатов можно было бы добиться, проведя случайную классификацию объектов, на которых произошел отказ. Метод, основанный на аппроксимации ЛЗ сигмоидами, почти на всех задачах показал лучшее качество распознавания, чем метод, использующий отрицания дизъюнкций ЛЗ и оригинальный метод, но не смог устранить все отказы от классификации. На задаче «breast» при использовании «приближенных» ЛЗ качество классификации предложенных методов хуже оригинального, что, впрочем, может быть связано с неточностью построения системы ЛЗ «приближенным» способом.

По результатам кластеризации ЛЗ можно заключить, что новое признаковое пространство может обладать излишней размерностью. Произведя предобработку систем логических закономерностей, удалось сократить размерность нового признакового пространства в рассматриваемых задачах до размерности исходного пространства признаков при минимальных потерях качества классификации.

5 Заключение

В рамках данной работы:

- разработан новый вид решающего правила и способ вычисления весовых коэффициентов для модели типа вычисления оценок, основанной на системах логических закономерностей;
- предложены методы устранения отказов от классификации с помощью введения оценок «антиблизости» и аппроксимации логических закономерностей;
- предложен способ сокращения размерности нового признакового пространства с помощью кластеризации логических закономерностей;
- для рассматриваемых методов проведены положительные эксперименты на реальных и модельных данных.

Список литературы

[1] *Журавлев Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, М.: Наука, вып. 33, 1978, с. 5–68.

[2] *Рязанов В. В.* Логические закономерности в задачах распознавания (параметрический подход) // Журнал вычислительной математики и математической физики, Т.47, №10, 2007, с. 1793–1808

[3] *Duda, R. O., Hart, P. E. and Stork, D. G.* Pattern Classication // John Wiley and Sons, 2nd edition, 2000

[4] *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* РАСПОЗНАВАНИЕ. Математические методы. Программная система. Практические применения // Изд-во «ФАЗИС», Москва, 2006.

[5] *Arthur, D. and Vassilvitskii, S.* k-means++: the advantages of careful seeding // Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2007, pp. 1027–1035.

[6] *Mangasarian, O. L. and Wolberg, W. H.* Cancer diagnosis via linear programming // SIAM News, Volume 23, Number 5, September 1990, pp. 1 - 18.

[7] *Bache, K. and Lichman, M.* UCI Machine Learning Repository [http://archive.ics.uci.edu/ml] // Irvine, CA: University of California, School of Information and Computer Science, 2013.

[8] *Ковшов Н. В., Моисеев В. Л., Рязанов В. В.* Алгоритмы поиска логических закономерностей в задачах распознавания // Журнал вычислительной математики и математической физики, М.: Наука, Т. 48, 2008, №2, стр. 329-344.