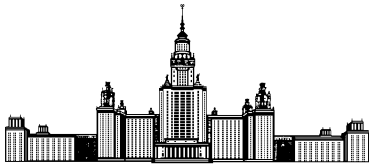
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 517 ГРУППЫ

«Исследование задачи кластеризации логических закономерностей, представленных булевыми векторами»

Выполнил:

студент 5 курса 517 группы

Новиков Максим Сергеевич

Научный руководитель:

д.ф-м.н., профессор

Рязанов Владимир Васильевич

Москва 2014

Содержание

1 Введение

Рассматривается стандартная задача распознавания по прецедентам. Считаем, что существует множество $M = \{x\} = \cup_{i=1}^{l} K_i$ допустимых объектов , разбитое на конечное число подмножеств (классов). Метку класса объекта $x \in X$ будем обозначать переменной $y = y(x) \in \{1, 2, \ldots, l\}$. Будем считать, что $x = \{x_1, x_2, \ldots, x_n) \in R^n$ и описываются значениями числовых признаков. Дана выборка $\{\mathbf{X_m}, \mathbf{Y_m}\} = \{y_i, x_i; i = 1, 2, \ldots, m\}$, содержащая объекты каждого класса. Требуется по данной выборке (обучающей выборке) отнести произвольный объект $x \in X$ к одному из классов. Данная задача обычно решается в два этапа. Сначала по обучающей выборке строится алгоритм классификации A (решается задача обучения). На втором этапе уже применяется найденный алгоритм A. В настоящее время существует много различных подходов для классификации объектов. Рассмотрим комбинаторно-логический подход для решения задачи классификации. Данные алгоритмы известны также как алгоритмы, основанные на принципе частичной прецедентности или (в более широком плане) алгоритмы вычисления оценок.

В данном подходе сначала в результате анализа обучающей информации находятся связи между признаками или значениями признаков и принадлежностью объектов классам (опорные множества, тупиковые тесты, представительные наборы, логические закономерности и т.п.). Определяются функции близости пар объектов (объекта обучающей выборки и объекта классифицируемого), принимающие значения 1 или 0 в зависимости от выполнения или невыполнения некоторой связи. В итоге строится алгоритм классификации, вычисляющий оценку объекта классификации («степень близости») за каждый класс. Объект относится в тот класс, за который он имеет максимальную оценку (при наличии нескольких максимальных оценок происходит отказ от распознавания). Мы будем рассматривать одну модель классификации данного типа: алгоритмы, основанные на голосовании по системам логических закономерностей классов (ЛЗК).

Здесь в результате анализа обучающей информации для каждого класса K_t вычисляются множества логических закономерностей класса: $\mathbf{P_t} = \{P_j(x)\}$. Это специальные предикаты , являющиеся конъюнкциями характеристических функций интервалов значений некоторых признаков, равные 1 на части объектов соответству-

ющего класса, равные 0 для всех обучающих объектов из других классов и удовлетворяющие некоторым условиям оптимальности. Второе условие может быть ослаблено. Логическая закономерность класса может быть равной 1 на небольшом числе объектов других классов. В этом случае говорят о «частичных логических закономерностях класса». Пусть для каждого класса K_t вычислено множество $\mathbf{P_t}$. При классификации произвольного $x \in X$ вычисляются оценки объекта за классы вида $\Gamma_t(x) = \sum_{P_j \in \mathbf{P_t}} \gamma_j P_j(x)$, где $\gamma_j \geq 0$ - весовые коэффициенты, соответствующие ЛЗК. Классификация по его оценкам проводится обычным образом. Достоинством данного подхода является не только возможность классификации новых объектов, но и то, что множества ЛЗК имеют самостоятельный интерес. Каждая ЛЗ является простым и наглядным свойством некоторого класса. Однако для заданной обучающей выборки их число может быть велико. Кроме того, приближенные методы могут вычислять «почти равные» или «вырожденные» ЛЗ. Таким образом, для каждого найденного множества $\mathbf{P_t}$ возникает необходимость его обработки, вычисления практически небольшого числа ЛЗК, которые нам давали бы представление о классах и были бы "наглядны". Один из таких подходов основан на построении кратчайших логических описаний классов. Пусть обучающая выборка непротиворечива. Тогда для каждого класса K_t находится система логических закономерностей, дизъюнкция которых может быть рассмотрена как характеристическая функция класса. Задача построения кратчайшего логического описания класса состоит в поиске минимального подмножества в $\mathbf{P_t}$, дизъюнкция предикатов которого равна 1 на объектах K_t.

Предлагается другой подход к анализу множеств ЛЗК, основанный на кластеризации ЛЗК. Наша задача состоит в построении для каждого класса небольшого числа «информативных и существенно различных» предикатов, которые можно рассматривать как (по крайней мере) частичные логические закономерности классов. Пусть каждой логической закономерности класса поставлен в соответствие элемент некоторого эвклидова пространства H. Пусть это будет выборка $Z = z_i, i = 1, 2, \ldots, h$. Пусть для каждого элемента H у нас есть способ вычисления некоторой частичной логической закономерности класса. Тогда можно провести кластеризацию выборки Z методом минимизации дисперсионного критерия на произвольное число k кластеров, выбрать в каждом кластере выборочный средний элемент и вычислить по

выборочным средним k частичных логических закономерностей класса (эталонных элементов). Данные вычисления можно провести для $k = 1, 2, \ldots, k_0$, оценить полученные частичные логические закономерности класса и оставить из них те, которые имеют высокую информативность.

В настоящей работе каждой логической закономерности P_j класса K_t, поставлен в соответствие булевский вектор $z_j = (z_{j1}, z_{j2}, \ldots, z_{jm})$, длина которого равна числу обучающих объектов в выборке, а единицы соответствуют тем объектам x, для которых $P_j(x) = 1$. Создан метод кластеризации выборки Z, учитывающий веса каждого объекта z_i. Предложены методы вычисления эталонных элементов как нахождение решений некоторых оптимизационных задач. Проведено численное сравнение данного подхода с методом построения кратчайших покрытий. Приведены результаты сравнений для нескольких практических задач.

2 Основные определения задачи поиска логических закономерностей классов

Будем рассматривать стандартную задачу распознавания по прецедентам. Приведем рассматриваемую постановку задачи классификации. Пусть существует множество $M = \{x\} = \cup_{i=1}^{l} K_i$ допустимых объектов , разбитое на конечное число подмножеств (классов). Метку класса объекта $x \in X$ будем обозначать переменной $y = y(x) \in \{1, 2, \ldots, l\}$. Будем считать, что $x = \{x_1, x_2, \ldots, x_n) \in R^n$ и описываются значениями числовых признаков. Дана выборка $\{y_i, x_i; i = 1, 2, \ldots, m\}$, содержащая объекты каждого класса. Требуется по данной выборке (обучающей выборке) отнести произвольный объект $x \in M$ к одному из классов.

Далее будем использовать также обозначения: $\hat{K}_i = X \cap K_i, i = 1, 2, \ldots, l$, $X = \{x_i, i = 1, 2, \ldots, m\}$. Приведем основные определения метода распознавания, основанного на голосовании по множествам логических закономерностей классов

Определение 1. *Предикат* $P^{\Omega_1, c_1, \Omega_2, c_2}(x) = \wedge_{j \in \Omega_1} c_{1j} \leq x_j \wedge_{j \in \Omega_2} x_j \leq c_{2j}$, $\Omega_1, \Omega_2 \subset \{1, 2. \ldots, n\}, c_{ij} \in R, c^{\lambda} = (c_{\lambda 1}, c_{\lambda 2}, \ldots, c_{\lambda n}), \lambda = 1, 2$ *назовем логической закономерностью класса* K_t*, если выполнены условия:*

1. $\exists x \in \hat{K}_t$: $P^{\Omega_1, c_1, \Omega_2, c_2}(x) = 1$

2. $\forall x \notin \hat{K}_t$*:* $P^{\Omega_1,c_1,\Omega_2,c_2}(x) = 0$

3. $P^{\Omega_1,c_1,\Omega_2,c_2}(x)$ *- локальный экстремум критерия качества предиката* Φ*.*

Предикат, удовлетворяющий только первым двум ограничениям, называется допустимым предикатом рассматриваемого класса. Предикат, удовлетворяющий только первому и третьему ограничениям, называется частичной логической закономерностью (ЛЗ) класса

Определение 2. *Стандартным критерием качества ЛЗК класса* K_t *будем называть следующий критерий:* $F(P^{\Omega_1,c_1,\Omega_2,c_2}(x)) = |\{x_i : x_i \in \hat{K}_t P^{\Omega_1,c_1,\Omega_2,c_2}(x_i) = 1\}|$

Определение 3. *Множество* $N(P^{\Omega_1,c_1,\Omega_2,c_2}(x)) = \{x \in R^n : c_{1j} \leq x_j, j \in \Omega_1, x_j \leq c_{2j}, j \in \Omega_2,\}$ *будем называть интервалом предиката* $P^{\Omega_1,c_1,\Omega_2,c_2}(x)$ *по аналогии с интервалами элементарных конъюнкций в алгебре логики.*

Определение 4. *Предикаты* $P^{\Omega_1,c_1,\Omega_2,c_2}(x)$ *и* $P^{\Omega_3,c_3,\Omega_4,c_4}(x)$ *называются эквивалентными, если* $P^{\Omega_1,c_1,\Omega_2,c_2}(x_t) = P^{\Omega_3,c_3,\Omega_4,c_4}(x_t), t = 1, 2, \ldots, m$

Определение 5. *Интервалы* $N(P^{\Omega_1,c_1,\Omega_2,c_2}(x))$ *и* $N(P^{\Omega_3,c_3,\Omega_4,c_4}(x))$ *называются эквивалентными, если* $N(P^{\Omega_1,c_1,\Omega_2,c_2}(x)) \cap X = N(P^{\Omega_3,c_3,\Omega_4,c_4}(x)) \cap X$

Логические закономерности со стандартным критерием качества имеют простую геометрическую интерпретацию: по данным обучающей выборки требуется найти прямоугольный координатный гиперпараллелипипед, лежащий в некотором признаковом подпространстве, содержащий максимальное число эталонов из класса K_t и только класса K_t . Границы параллелипипеда по различным признакам могут быть конечными и бесконечными, открытыми или закрытыми.

3 Обработка множеств логических закономерностей классов

Пусть для класса K_t найдена система логических закономерностей $\mathbf{P_t} = \{P_j(x)\}$

3.1 Подход 1 - построение логических описаний классов

Определение 6. *Функция* $D_t(x) = \vee_{j=1,2,\ldots,|\mathbf{P_t}|} P_j(x)$ *называется логическим описанием класса* K_t.

Определение 7. *Функция* $D_t^{sh}(x) = \vee_{P_j \in \mathbf{P'_t} \subset \mathbf{P_t}} P_j(x)$ *называется кратчайшим логическим описанием класса* K_t, *если* $D_t^{sh}(x)$ *эквивалентна* $D_t(x)$ *и* $|\mathbf{P'_t}|$ *минимально.*

3.2 Подход 2 - кластеризация логических закономерностей

Предлагается следующий подход на основе кластеризации логических закономерностей. Каждой ЛЗ $P^{\Omega_1,c_1,\Omega_2,c_2}(x)$ можно поставить в соответствие булевский вектор $\mathbf{z} = (z_1, z_2, \ldots, z_h)$, где

$$z_i = \begin{cases} 1 & P^{\Omega_1,c_1,\Omega_2,c_2}(x_i) = 1 \\ 0 & \text{Иначе} \end{cases}$$

Таким образом, множеству из N_t ЛЗ класса соответствует множество из N_t булевских векторов $\{\mathbf{z_i}, i = 1, 2, \ldots, N_t\}$. Предлагается следующий алгоритм построения множеств эталонных предикатов для каждого из классов.

Алгоритм 1 Алгоритм кластеризации логических закономерностей

Для каждого класса K_t вычисляется множество ЛЗК и формируется свое множество $\mathbf{P_t}$

Фиксируется натуральное $k = 1, 2, \ldots, k_0$ и осуществляется кластеризация множества $\mathbf{P_t}$ модификацией метода минимизации дисперсионного критерия, учитывающего вес β_j каждой ЛЗК $P_j(x)$

По каждому кластеру для выборочного среднего кластера (вектора $(m) = (m_1, m_2, \ldots, m_h)$, $0 \leq m_i \leq 1$, $i = 1, 2, \ldots, h$) вычисляется оптимальный бинарный вектор , по которому вычисляется предикат, который интерпретируется как частичная логическая закономерность класса K_t и оценивается различными способами.

4 Кластеризация логических закономерностей класса

Пусть дана выборка бинарных векторов $Z = \{\mathbf{z_i}, i = 1, 2, \ldots, N\}\mathbf{z_i} \in H$, соответствующих множеству ЛЗ $\mathbf{P_t} = \{P_j\}$ класса K_t и каждый объект имеет вес β, $\beta \geq 0$. Пусть фиксировано требуемое число кластеров k. Предлагается следующая модификация алгоритма минимизации дисперсионного критерия кластеризации, для выборки $\{\beta_i, \mathbf{z_i}\}$, $\beta_i \geq 0$ - заданные веса объектов.

Задача 1.

$$J(K) = \sum_{i=1}^{k} \sum_{j:\mathbf{z_j} \in K_i} \beta_j ||\mathbf{z_j} - \mathbf{m_i}||^2 \to min$$

где

$$K = \cup_{\alpha=1}^{k} K_\alpha, K_i \cap K_j = \emptyset, i \neq j$$

- разбиение выборки на k множеств, а

$$\mathbf{m_i} = \frac{\sum_{\mathbf{z_j} \in K_i} \beta_j \mathbf{z_j}}{\sum_{\mathbf{z_j} \in K_i} \beta_j}$$

Для простоты записи будет писать $\sum \beta$ вместо $\sum_{z_j \in K_j} \beta_j$ и $\sum \beta z$ вместо $\sum_{z_j \in K_j} \beta_j z_j$ пусть объект $\hat{z}$ переносится из кластера K_i в кластер K_j, тогда $K_i \to K_i^* = K_i \backslash \{\hat{z}\}, K_j \to K_j^* = K_j \cup \{\hat{z}\}$

Тогда новый центр кластера i может быть вычислен по формуле

$$\mathbf{m_i^*} = \frac{(\sum \beta z - \hat{\beta}\hat{z})/\sum \beta}{(\sum \beta - \hat{\beta})/\sum \beta} = \mathbf{m_i} \frac{\sum \beta + \hat{\beta} - \hat{\beta}}{\sum \beta - \hat{\beta}} - \frac{\hat{\beta}\hat{z}}{\sum \beta - \hat{\beta}} = \mathbf{m_i} + \mathbf{m_i} \frac{\hat{\beta}}{\sum \beta - \hat{\beta}} - \frac{\hat{\beta}\hat{z}}{\sum \beta - \hat{\beta}} =$$

$$= \mathbf{m_i} - \frac{\hat{\beta}(\hat{z} - \mathbf{m_i})}{\sum \beta - \hat{\beta}}$$

Таким образом,

$$\mathbf{m_i^*} = \mathbf{m_i} - \frac{\hat{\beta}(\hat{z} - \mathbf{m_i})}{\sum \beta - \hat{\beta}}$$

Рассмотрим теперь изменение компоненты функционала качества кластеризации, соответствующей i-му кластеру.

$$J_i(K_i^*) = \sum \beta ||z - \mathbf{m_i^*}||^2 - \hat{\beta}||\hat{z} - \mathbf{m_i^*}||^2 = \sum \beta (z - \mathbf{m_i} + \frac{\hat{\beta}(\hat{z} - \mathbf{m_i})}{\sum \beta - \hat{\beta}}, z - \mathbf{m_i} + \frac{\hat{\beta}(\hat{z} - \mathbf{m_i})}{\sum \beta - \hat{\beta}}) - ||\hat{z} - \mathbf{m_i^*}||^2 =$$

$$= \sum \beta||z-\mathbf{m_i}||^2+2\sum \beta(z-\mathbf{m_i}, \frac{\hat{\beta}(\hat{z}-\mathbf{m_i})}{\sum\beta-\hat{\beta}})+\sum \beta\frac{\hat{\beta}^2||z-\mathbf{m_i}||^2}{(\sum\beta-\hat{\beta})^2}-\hat{\beta}||(\hat{z}-\mathbf{m_i})+\frac{\hat{\beta}(\hat{z}-\mathbf{m_i})}{\sum\beta-\hat{\beta}}||^2 =$$

$$= J_i(K_i)+\sum \beta\frac{\hat{\beta}^2||\hat{z}-\mathbf{m_i}||^2}{(\sum\beta-\hat{\beta})^2}-\frac{\hat{\beta}||\hat{z}-\mathbf{m_i}||^2(\sum\beta)^2}{(\sum\beta-\hat{\beta})^2} = J_i(K-i)+\frac{||\hat{z}-\mathbf{m_i}||^2}{(\sum\beta-\hat{\beta})^2}(\sum\beta\hat{\beta}^2-\hat{\beta}(\sum\beta)^2) =$$

$$= J_i(K_i) - \frac{\sum\beta\hat{\beta}}{\sum\beta-\hat{\beta}}||\hat{z}-\mathbf{m_i}||^2$$

Таким образом,

$$J_i(K_i^*) = J_i(K_i) - \frac{\sum\beta\hat{\beta}}{\sum\beta-\hat{\beta}}||\hat{z}-\mathbf{m_i}||^2$$

Проведя аналогичные выкладки для K_j получим

$$\mathbf{m_j^*} = \mathbf{m_j} + \frac{\hat{\beta}(\hat{z}-\mathbf{m_j})}{\sum\beta+\hat{\beta}}$$

$$J_j(K_j^*) = J_j(K_j) + \frac{\sum\beta\hat{\beta}}{\sum\beta+\hat{\beta}}||\hat{z}-\mathbf{m_j}||^2$$

Такми образовм, условием локального минимума при минимизации функционала

$$J(K) = \sum_{j=1}^{l}\sum_{z_i\in K_j}\beta_i||z_i-\mathbf{m_j}||^2$$

является выполнение неравенств

$$\frac{\sum\beta\hat{\beta}}{\sum\beta+\hat{\beta}}||\hat{z}-\mathbf{m_j}||^2 \geq \frac{\sum\beta\hat{\beta}}{\sum\beta-\hat{\beta}}||\hat{z}-\mathbf{m_i}||^2$$

Для всех пар кластеров K_i и K_j для всех объектов $\hat{z}\in K_i$

5 Вычисление множеств эталонных элементов

Пусть для набора ЛЗ $\mathbf{P_t}$ класса K_t решена задача 1, то есть вычислено

$$K = \cup_{\alpha=1}^{k}K_\alpha, K_i\cap K_j = \emptyset, i\neq j$$

- разбиение выборки на k множеств и выборочные средние

$$\mathbf{m_i} = \frac{\sum_{\mathbf{z_j}\in K_i}\beta_j\mathbf{z_j}}{\sum_{\mathbf{z_j}\in K_i}\beta_j}$$

. Требуется построить эталонные предикаты (эталонные ЧЛЗ), оптимальные по некоторому критерию, соответсвующие центрам кластеров $\mathbf{m_i} = (m_{i1}, m_{i2}, \ldots, m_i h), 0 \leq m_{ij} \leq 1, j = 1, 2, \ldots, h$

Данная задача решается в два этапа - на первом этапе производится бинаризация выборочных средних $\mathbf{m_i}$ по порогам θ (для каждого центра кластера используется собственный порог бинаризации), то есть, вектору $\mathbf{m_i}$ ставится в соответствие булев вектор $\mathbf{b_i} = (b_{i1}, b_{i2}, \ldots, b_{ih})$

$$b_{ij} = \begin{cases} 1 & m_{ij} \geq \theta_i \\ 0 & \text{Иначе} \end{cases}$$

На втором этапе каждому вектору $\mathbf{b_i}$ ставится в соответствие предикат по правилу

$$P_i(\mathbf{x}) = \wedge_{j=1}^{n} min_{\gamma:\mathbf{b}_\gamma=1} x_{\gamma j} \leq \mathbf{x_j} \leq max_{\gamma:\mathbf{b}_\gamma=1} x_{\gamma j} \quad (1)$$

Полученный таким образом предикат, вообще говоря, не обязательно является чистой логической закономерностью своего класса. Также нетрудно видеть, что каждому вещественному центру кластера $\mathbf{m_i}$ соответствует набор эталонных предикатов $P_i(\mathbf{x})$, соотвествующих разным значениям порога бинаризации θ_i. Варьируя параметр θ_i для каждого центра кластера $\mathbf{m_i}$, получим эталонные предикаты, оптимальные по некоторому критерию.

Нетрудно видеть, что при построении эталонной частичной логической закономерности по формуле 1 для вычисления значения полученного предиката будут использоваться все n признаков. Модифицируем формулу 1 таким образом, чтобы в построенную логическую закономерность вошли только признаки, разделение по которым действительно приводит к разделению объектов обучающей выборки.

Определим множество Ω_j информативных признаков следующим образом: признак номер i входит в Ω_j тогда и только тогда, когда сущетствеют объекты обучающей выборки $x' \in X$ и $x'' \in X$, такие что значения предиката $\tilde{P}_i(x) = min_{\gamma:\mathbf{b}_\gamma=1} x_{\gamma i} \leq \mathbf{x_i} \leq max_{\gamma:\mathbf{b}_\gamma=1} x_{\gamma j}$ $\tilde{P}_i(x') \neq \tilde{P}_i(x'')$.

Модифицируем правило построения эталонных закономерностей следующим образом:

$$P_i(\mathbf{x}) = \wedge_{j\in\Omega_j} min_{\gamma:\mathbf{b}_\gamma=1} x_{\gamma j} \leq \mathbf{x_j} \leq max_{\gamma:\mathbf{b}_\gamma=1} x_{\gamma j} \quad (2)$$

В силу определения множества Ω_j, эталонные закономерности, построенные по формулам 1 и 2 эквивалентны, однако, для интерпритации результатов кластеризации предпочтительнее использование более коротких описаний.

Варьируя параметр θ_i для каждого центра кластера $\mathbf{m_i}$, получим эталонные предикаты, оптимальные по некоторому критерию. Критерии выбора порога бинаризации приведены в следующем разделе.

6 Критерии выбора порога бинаризации на основе критериев сравнения логических закономерностей

При построении эталонных закономерностей по заданной кластеризации

$$K = \cup_{\alpha=1}^{k} K_\alpha, K_i \cap K_j = \emptyset, i \neq j$$

(заданному разбиению выборки на k множеств) и заданным выборочным средним

$$\mathbf{m_i} = \frac{\sum_{\mathbf{z_j}\in K_i} \beta_j \mathbf{z_j}}{\sum_{\mathbf{z_j}\in K_i} \beta_j}$$

производится бинаризация выборочных средних $\mathbf{m_i}$ по порогам θ (для каждого центра кластера используется собственный порог бинаризации), то есть, вектору $\mathbf{m_i}$ ставится в соответствие булев вектор $\mathbf{b_i} = (b_{i1}, b_{i2}, \ldots, b_{ih})$

$$b_{ij} = \begin{cases} 1 & m_{ij} \geq \theta_i \\ 0 & \text{Иначе} \end{cases}$$

Для выбора порогов бинаризации θ_i предлагается использовать критерии качества частичных логических закономерностей классов.

Для сравнения между собой логических закономерностей были предложены различные критерии, наиболее популярные из которых приведены ниже. В основе всех этих критериев лежит число покрытых объектов класса K_t и других классов, что

позволяет естественным образом использвовать их для оценки результатов кластеризации и выбора порогов бинаризации без построения этаклонных закономерностей, получив не только метод выбора порогов, но и упростив вычисления.

Для упрощения записи в этом и только этом разделе дипломной работы приняты следующие обозначения: p - число покрытых объектов того же класса, что и закономерность, n - число объектов других классов, покрытых закономерностью, P - всего объектов данного класса в выборке, N - всего объектов других классов в выборке.

6.1 Энтропийный критерий IGain

$$IGain(p,n) = h(\frac{p}{l}) - \frac{p+n}{l}h(\frac{p}{p+n}) - \frac{l-p-n}{l}h(\frac{P-p}{l-p-n}) \to max$$

, где

$$h(q) = -q\log_2(q) - (1-q)\log_2(1-q)$$

На рисунке 1a показано поведение критерия для выборки из 1000 элементов.

6.2 Простой критерий p-n

Один из самых простых критериев для срвнения логических закономерностей

$$p - n \to max$$

На рисунке 1c показано измененеие значения критерия в зависмости от p и n от 0 до 1000 (для выборки из 1000 элементов).

Обратите внимание, что для значений $p + n > 1000$ значения критерия не вычислялись ввиду того, что общее число объектов, покрытых закономерностью (число покрытых объектов как своего, так и других классов) не может превосходить размера выборки.

6.3 Нормированный критерий p-n

Очевидным недостатком простого критерия $p - n$ является тот факт, что он не учитывает несбалансированность выборки. Рассмотрим модификацию критерия $p - n$, учитывающую различие в количестве объектов разных классов:

$$\frac{p}{P} - \frac{n}{N} \to max$$

На рисунке 1b показано измененеие значения критерия в зависмости от p и n от 0 до 500 (для выборки из 1000 элементов).

Обратите внимание, что для значений $p + n > 1000$ значения критерия не вычислялись ввиду того, что общее число объектов, покрытых закономерностью (число покрытых объектов как своего, так и других классов) не может превосходить размера выборки.

6.4 Критерий бустинга

Предложенный William W. Cohen и Yoram Singer в 1999 [1] простой и эффективный критерий для сравнения логических закономерностей. Определен как

$$p^{\frac{1}{2}} - n^{\frac{1}{2}} \to max$$

На рисунке 1f показано измененеие значения критерия в зависмости от p и n от 0 до 500 (для выборки из 1000 элементов).

Обратите внимание, что для значений $p + n > 1000$ значения критерия не вычислялись ввиду того, что общее число объектов, покрытых закономерностью (число покрытых объектов как своего, так и других классов) не может превосходить размера выборки.

6.5 Нормированный критерий бустинга

Как и простой критерий $p - n$, критерий бустинга для сравнения логических закономерностей не учитывает несбалансированность выборки. Рассмотрим модифицированный критерий бустинга:

$$(\frac{p}{P})^{\frac{1}{2}} - (\frac{n}{N})^{\frac{1}{2}} \to max$$

На рисунке 1d показано измененеие значения критерия в зависмости от p и n от 0 до 500 (для выборки из 1000 элементов).

Обратите внимание, что для значений $p + n > 1000$ значения критерия не вычислялись ввиду того, что общее число объектов, покрытых закономерностью (число покрытых объектов как своего, так и других классов) не может превосходить размера выборки.

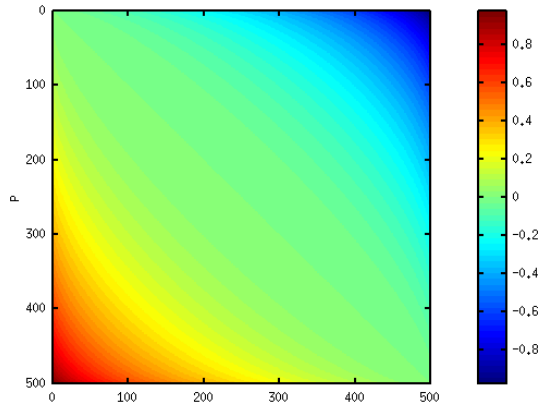
7 Сравнение эталонных закономерностей, полученных в результате кластеризации и кратчайших описаний классов

Пусть решена задача кластеризации логических закономерностей класса K_t,то есть найдены эталонные предикаты $P_i^*(x)$. Пусть также найдено кратчайшее описание $D_t^{sh}(x) = \vee_{P_j^{sh} \in \mathbf{P'_t} \subset \mathbf{P_t}} P_j^{sh}(x)$ класса K_t. Рассмотрим вопрос о сходстве предикатов $P_j^{sh}(x)$, входящих в кратчайшее описание класса K_t с эталонными предикатами $P_i^*(x)$.

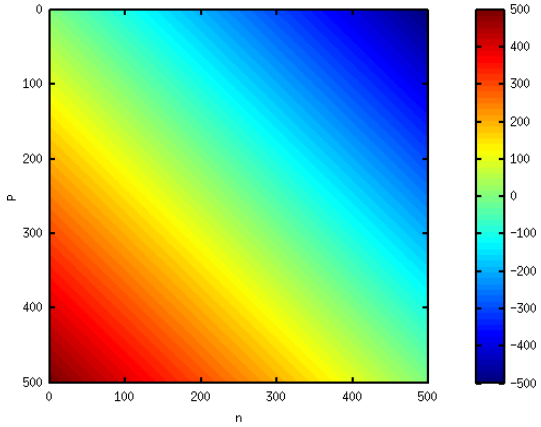
Определим расстояние между предикатами $\rho(P_i^*, P_j^{sh})$. Поставим предикатам P_i и P_j^{sh} в соответсвие векторы $\mathbf{b_i}$ и $\mathbf{b_j}$. $\mathbf{b_i} = (b_{i1}, b_{i2}, \ldots, b_{ih})$, где

$$b_{ij} = \begin{cases} 1 & P_i(x_j) = 1 \\ 0 & \text{Иначе} \end{cases}$$

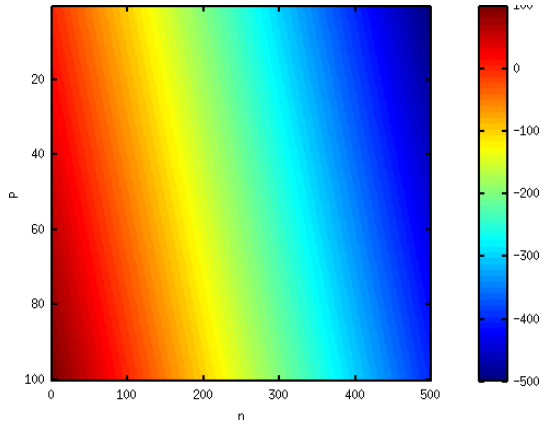
Теперь расстояние $\rho(P_i^*, P_j^{sh})$ можно вычислить как $\rho(P_i^*, P_j^{sh}) = L_2(\mathbf{b_i}, \mathbf{b_j}) = \sqrt{\sum_{k=1}^{h} (\mathbf{b_{ik}} - \mathbf{b_{jk}})^2}$ Таким образом, для каждого эталонного предиката P_i^* может быть найден соответствующийему предикат из кратчайшего описания класса $P_i^{sh*} = min_j \rho(P_i^*, P_j^{sh})$. Среднее расстояние по всем эталонным предикатам $\rho(\{P_1^*, P_2^*, \ldots, P_k^*\}, D^{sh}) = \frac{\sum_{i=1}^{k} \rho(P_i^*, P_i^{sh*})}{k}$ является мерой различия описания класса, полученного в результате кластеризации логических закономерностй класса (ЛЗК) K_t на k кластеров с кратчайшим описанием класса.



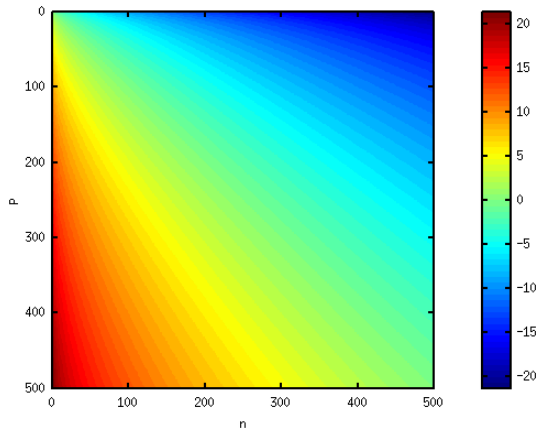
(a) Критерий IGain для выборки из 1000 элементов



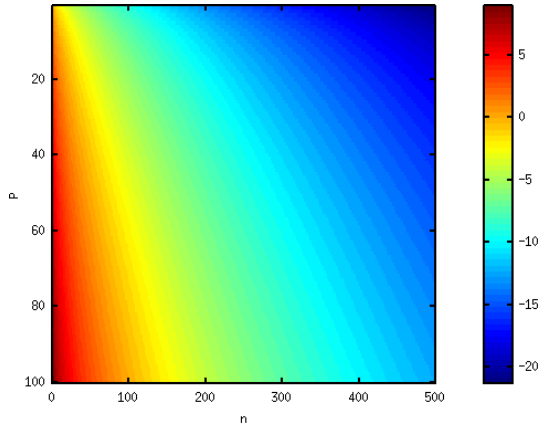
(b) Критерий p-n для выборки из 1000 элементов



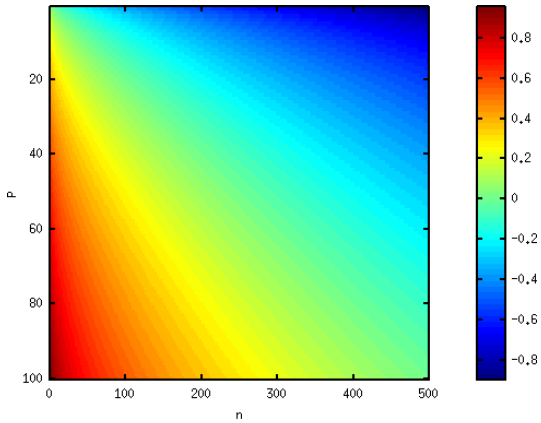
(c) Критерий p-n для выборки из несбалансированной выборки из 100 и 500 элементов в каждом классе



(d) Критерий бустинга для выборки из 1000 элементов



(e) Критерий бустинга для несбалансированной выборки из 100 и 500 элементов



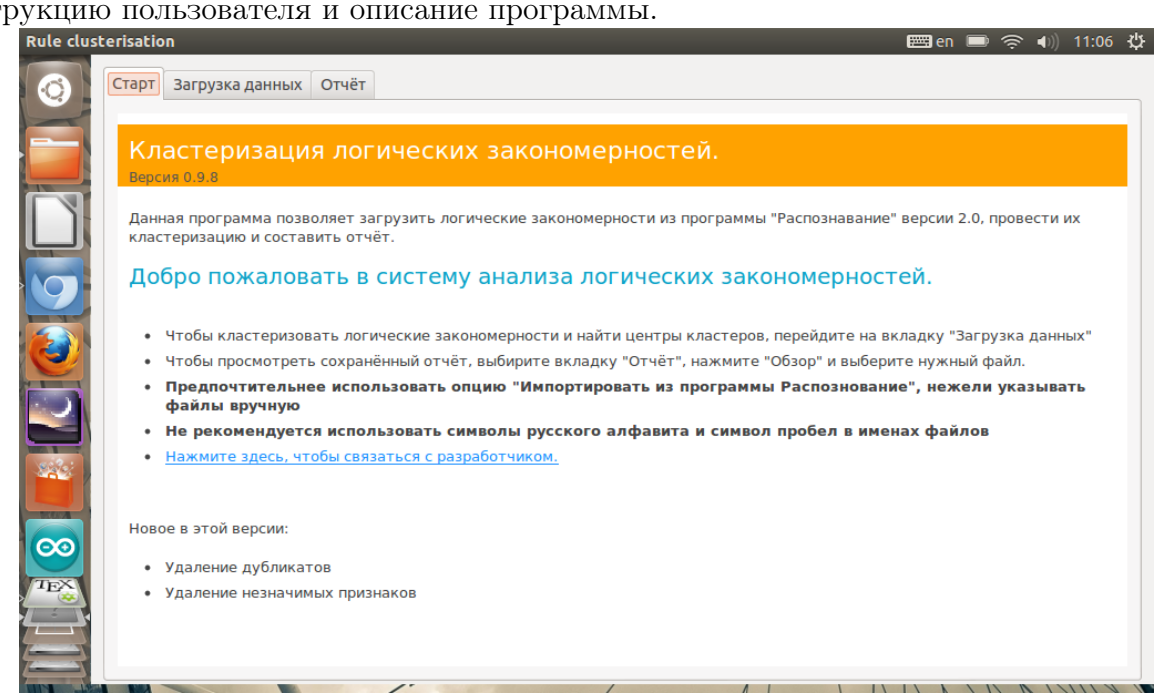
(f) Нормированный критерий бустинга для несбалансированной выборки из 100 и 500 элементов

Рис. 1: Значения критериев при фиксированных P и N в зависимости от p и n

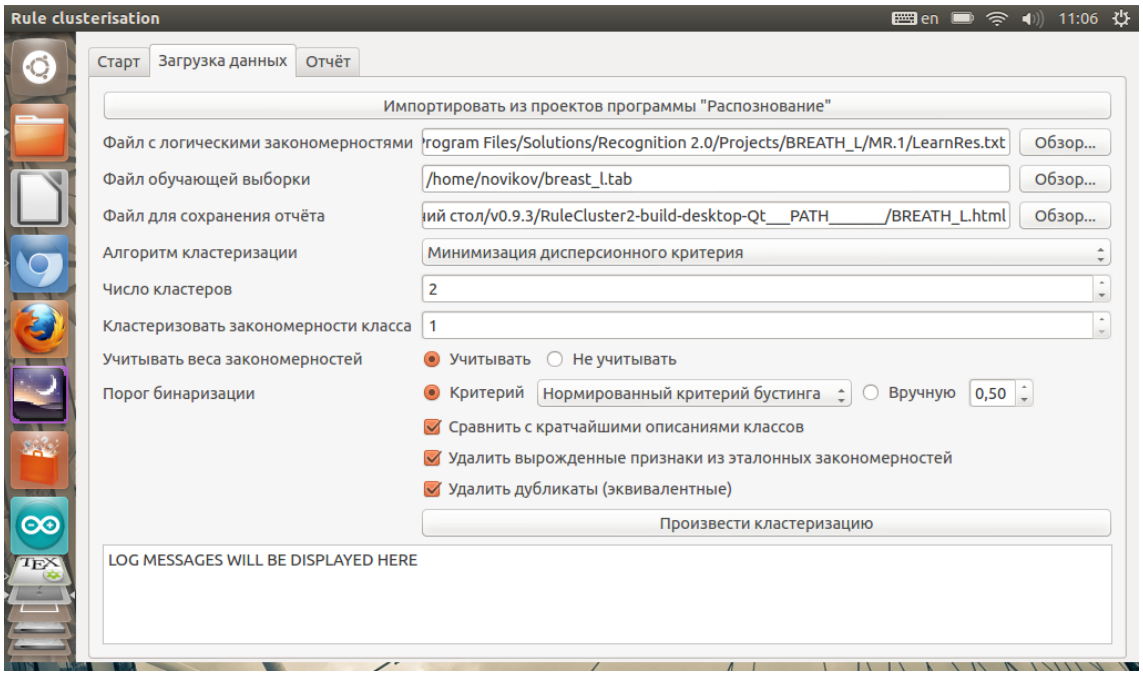
8 Описание реализованной программы

В ходе дипломной работы была разработана программа на языке C++ с графическим интерфейсом пользователя, реализующая предложенный подход к исследованияю множества $\mathbf{P_t}$ логических закономерностей класса K_t. В программе реализован алгоритм кластеризации логических закономерностей, все рассмотренные в данной работе критерии выбора порога бинаризации, а также предобработка, позволяющая удалить эквивалентные закономерности из множества $\mathbf{P_t}$.

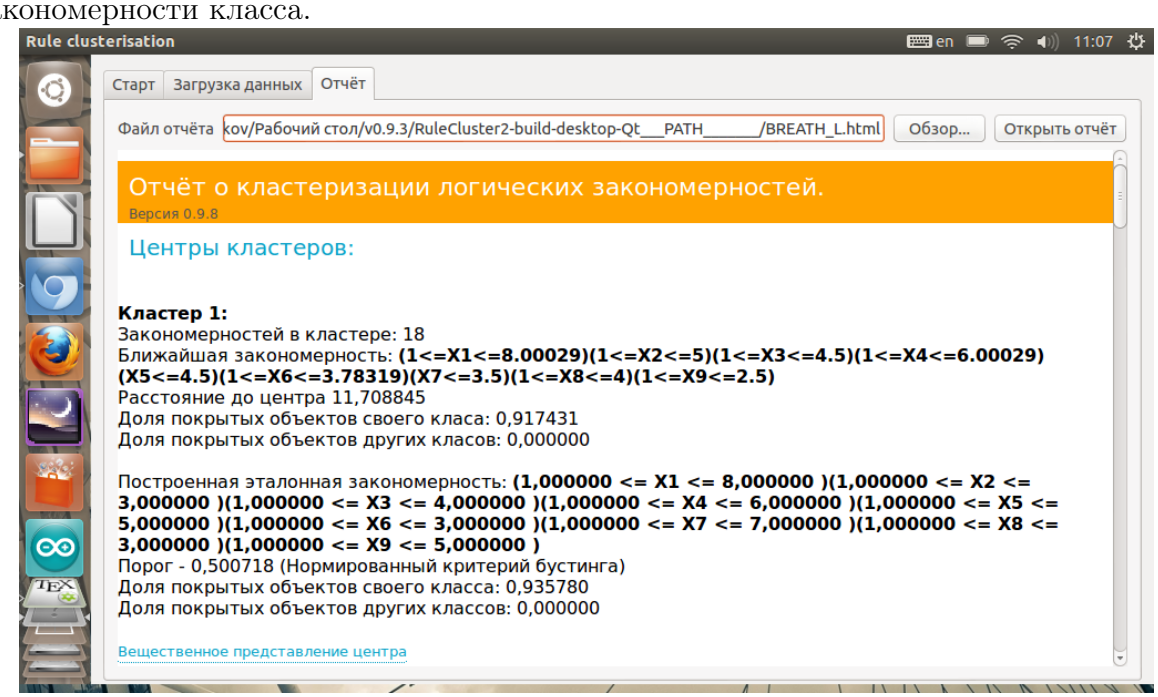
Программа состоит из трёх основных экранов и диалога импорта из проектов системы "Распознавание". Первый экран - вспомогательный, содержит краткую инструкцию пользователя и описание программы.



На втором экране указываются необходимые для работы файлы обучающей выборки и список логических закономерностей, а также устанавливаются параметры метода. Расположение файлов указывается либо с ипользованием импорта из программы "Распознавание либо вручную.



На третьем экране выводится отчёт о результатах кластеризации. Отчёт содержит полное описание каждого кластера, ближайщую к центру закономерность, результаты сравнения с кратчайшим описанием класса и бинарное представление каждой закономерности класса.



9 Вычислительные эксперименты

Применим предложенный подход на основе кластеризации логических закономерностей к некоторым практическим задачам и оценим различия результатов описания класса, полученного в результате кластеризации логических закономерностй класса (ЛЗК) K_t на k кластеров с кратчайшим описанием класса. Для поиска множества логических закономерностей класса K_t исполуется метод поиска логических закономерностей, реализованный в виде компонента системы "Распознавание"[4].

Предложенный подход был реализован как в виде отдельной программы на языке C++ с возможностью импорта из программы "Распознавание так и в виде набора функций на языке MATLAB.

9.1 Задача диагностики рака груди

Рассматривается задача диагностики рака груди. Обучающая выборка состоит из 218 наблюдений здоровых людей и 126 наблюдений поциентов с раком. Каждый поциент описывается значениями 9 вещественных признаков, однако, не для каждого поциента известнывсе признаки.

В результате поиска логических закономерностей методом, реализованным в системе "Распознавание"было найдено 27 закономерностей первого класса и 26 закономерностей 2 класса. Произведём кластеризацию указанных наборов закономерностей на различное число кластеров и оценим получившиеся результаты.

Сравним также закономерности, являющиеся эталонноми по различным критериям.

Кластеризация на 1 кластер

Кластеризация на 1 кластер имеет простую интерпритацию - таким образом находится средняя закономерность множества $\mathbf{P_t}$. Варьируя пороги, получим законо-

мерность, оптимальную по некоторому критерию.

Критерий	Класс 1	Класс 2
	Покрыто класса 1 | класса 2	Покрыто класса 2 | класса 1
IGain	98.1651% 2.3810%	89.6825% 3.6697%
Критерий бустинга	92.2018% 0.00000%	67.4603% 0.00000%
Критерий p-n	98.1651% 2.3810%	96.0317% 7.3394%
Бустинга, нормированный	92.2018% 0.00000%	67.4603% 0.00000%
p-n, нормированный	98.1651% 2.3810%	96.0317% 7.3394%

Как видно из таблицы 9.1, для задачи диагностики рака качественные информативные закономерности получаются как взвешенное среднее всех закономерностей класса. Такми образом, получено интерпритируемое человеком описание каждого класса состоящее всего из одной закономерности на класс.

Проведем сравнение эталонной закононерности, найденной в результате кластеризации закономерностей каждого класса на один кластер с использованием энтропийного критерия IGain и кратчайшего описания класса.

Класс 2 Кратчайшее описание:

$(5.5 \leq X1)(2.14141 \leq X6)(3.5 \leq X7)$

$V(3.5 \leq X3)(2.5 \leq X9)$

$V(1.37478 \leq X1)(2.74978 \leq X3)(X4 \leq 5)(5 \leq X6)$

$V(6.5 \leq X4)$

$V(3.87489 \leq X3)(X5 \leq 6.75)(5.25 \leq X6)$

$V(1.9997 \leq X3)(4.5 \leq X7)$

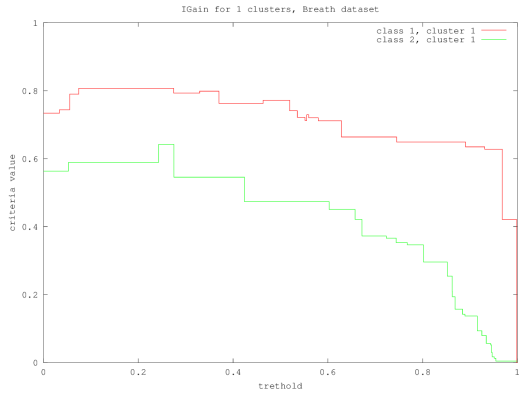
Эталонная закономерность кластера 1:

$(3 \leq X1 \leq 10)(3 \leq X2 \leq 10)(2 \leq X3 \leq 10)(2 \leq X5 \leq 10)(3 \leq X7 \leq 10)$

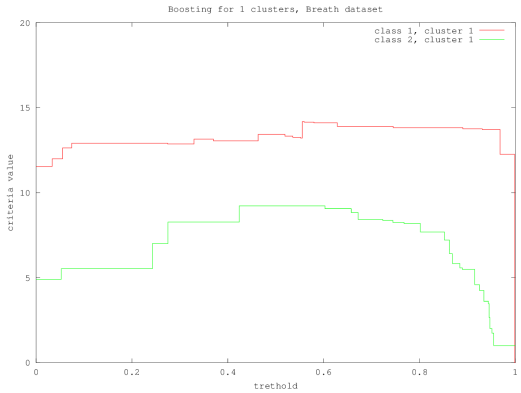
Ближайшей закономерностью из кратчайшего описания класса является закономерность $(1.9997 \leq X3)(4.5 \leq X7)$. нормированное расстояние до эталонной закономерности: 0.037791.

9.2 Задача классификации изображений

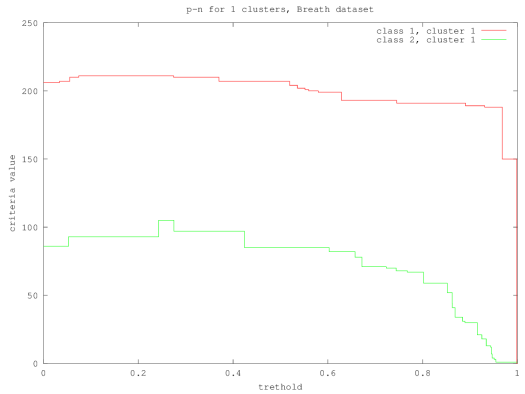
Рассматривается задача классификации изображений на 7 классов. Каждый класс задан 30 своими представителями, описанными 16 вещественными признаками.



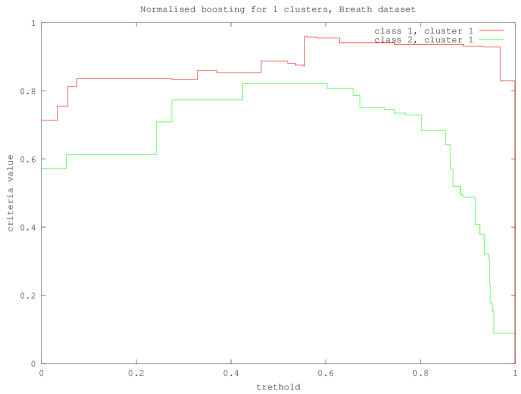
(a) Энтропийный критерий IGain при кластеризации на 1 кластер



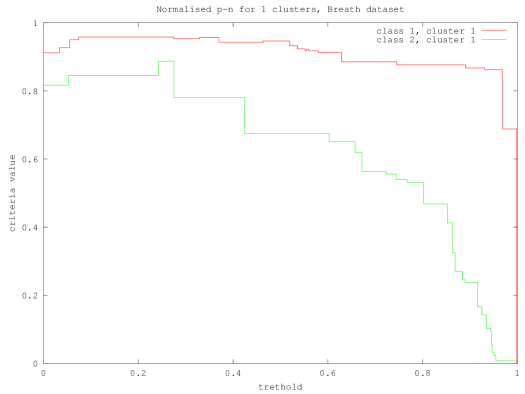
(b) Критерий бустинга при кластеризации на 1 кластер



(c) Критерий p-n при кластеризации на 1 кластер



(d) Нормированный критерий бустинга при кластеризации на 1 кластер



(e) Нормированный критерий p-n при кластеризации на 1 кластер

Рис. 2: Значения критериев в зависимости от порога бинаризации для задачи диагностики рака при кластеризации на 1 кластер

Кластеризация на 1 кластер

Кластеризация на 1 кластер имеет простую интерпритацию - таким образом находится средняя закономерность множества $\mathbf{P_t}$. Варьируя пороги, получим закономерность, оптимальную по некоторому критерию.

Класс 1

Критерий	Класс 1	Другие классы
IGain	73.3333%	0.00000%
Критерий бустинга	73.3333%	0.00000%
Критерий p-n	73.3333%	0.00000%
Бустинга, нормированный	73.3333%	0.00000%
p-n, нормированный	73.3333%	0.00000%

Класс 2

Критерий	Класс 2	Другие классы
IGain	86.6667%	0.00000%
Критерий бустинга	86.6667%	0.00000%
Критерий p-n	86.6667%	0.00000%
Бустинга, нормированный	86.6667%	0.00000%
p-n, нормированный	86.6667%	0.00000%

Класс 3

Критерий	Класс 3	Другие классы
IGain	60.0000%	4.4444%
Критерий бустинга	46.6667%	0.5556%
Критерий p-n	46.6667%	0.5556%
Бустинга, нормированный	46.6667%	0.5556%
p-n, нормированный	73.3333%	11.6667%

Класс 4

Критерий	Класс 4	Другие классы
IGain	56.6667%	0.0000%
Критерий бустинга	56.6667%	0.0000%
Критерий p-n	60.0000%	0.5556%
Бустинга, нормированный	56.6667%	0.0000%
p-n, нормированный	83.3333%	15.0000%

Класс 5

Критерий	Класс 5	Другие классы
IGain	90.0000%	2.2222%
Критерий бустинга	26.6667%	0.0000%
Критерий p-n	43.3333%	2.7778%
Бустинга, нормированный	26.6667%	0.0000%
p-n, нормированный	90.0000%	2.2222%

Класс 6

Критерий	Класс 6	Другие классы
IGain	80.0000%	0.0000%
Критерий бустинга	80.0000%	0.0000%
Критерий p-n	80.0000%	0.0000%
Бустинга, нормированный	80.0000%	0.0000%
p-n, нормированный	80.0000%	0.0000%

Класс 7

Критерий	Класс 7	Другие классы
IGain	90.0000%	7.7778%
Критерий бустинга	90.0000%	7.7778%
Критерий p-n	90.0000%	7.7778%
Бустинга, нормированный	90.0000%	7.7778%
p-n, нормированный	90.0000%	7.7778%

Рассмотрим классы с простой и сложной внутриклассовой структурой. Такими классами являются классы с номерами 1 и 3. В кратчайщем описании класса 1 присутствует только 1 закономерность, в описании класса 3 - 4 закономерности. Произ-

ведем кластеризацию закономерностей класса 3 на 4 кластера и сравним полученные результаты с кратчайшим описанием класса.

Кратчайшее описание класса 4:

$(X2 <= 138.25)(X9 <= 33.5022)(X11 <= -5.44445)(X12 <= 24.1954)(X15 <= 0.903517)(-2.16796 <= X16)$

$V(11.9942 <= X2 <= 123)(0.777778 <= X8 <= 27.2258)(-13.5568 <= X11 <= -1.94444)(-13.0011 <= X13 <= -5.05556)$

$V(X2 <= 147.19)(0 <= X3 <= 1.86462)(X11 <= 0.000144495)(-10.6182 <= X13)(-2.08265 <= X16)$

$V(38.2444 <= X1)(0 <= X3 <= 2.09036)(X11 <= 0.00028899)(-10.9586 <= X13 <= 0.000275094)(0 <= X15 <= 0.783438)(-2.18814 <= X16 <= 3.89851e-005)$

Предикаты, входящие в кратчайшее описание покрывают 0.36, 0.74,0.43 и 0.53 объектов своего классо соответственно.

Рассмотрим полученные центры кластеров. Первой закономерности соответсвует эталонная закономерность, покрывающая 70.0000% объектов своего класса и 11.1111% объектов других классов.Нормированное расстояние: 0,076190.

Второй закономерности соответствуют 2 эталонные закономерности, покрвыющие 83.3333% и 46.6667% своего класса и 17.7778% и 9.4444% других классов соответственно. Расстояние: 0.25 и 0.09 соответственно.

Третьей закономерности соответсвует эталонная закономерность, покрывающая 70.0000% объектов своего класса и 25.0000% объектов других классов.Нормированное расстояние: 0,123810.

Как видно из проведенных вычислительных экспериментов, не всем закономерностям, входящим в кратчайшее описание, соответствует центр кластера логических закономерностей при кластеризации на то же число кластеров, что и число предикатов P_j^{sh}, входящих в кратчейщее описание класса K_t.

Как видно, во многих случаях, выбор критерия не оказывает существенного влияния на результирующие характеристики центров кластеров, однако,по мере ухудшения характеристик центров кластеров, различия становятся всё более заметны.

10 Заключение

Была рассмотрена стандартная задача распознавания по прецедентам и предложен подход к анализу множества логичеких закономерностей $\mathbf{P_t}$ класса K_t. Предложена модификация метода кластеризации, основанном на минимизации дисперсионного критерия, учитывающая заданные веса $\gamma_i, 0 \leq \gamma_i \leq 1$ объектов x_i. Метод кластеризации, основанный на минимизации дисперсионного критерия является частым случаем предложенной модификации, когда веса всех объектов одинаковы и равны 1. Предложен метод построения эталонных закономерностей, полученных в результате кластеризации множества $\mathbf{P_t}$, являющихся локально-оптимальными частичными логическими закономерностями класса K_t. Был предложен метод сравнения эталонных предикатов и кратчайших описаний D^{sh} класса K_t. Проведенные вычислительные эксперименты показали возможность практического применения предложненного подхода в различных задах. Существенным отличием предложенного подхода от построения кратчайших описаний классов $D^{sh}(x)$ является то, что предложенная процедура расширяет набор $\mathbf{P_t}$ локально-оптимальными эталонными закономерностями, в то время как кратчайшие описания классов строятся как подмножество $\mathbf{P_t}$.

Дальнейшими исследованиями в рамках предложенного подхода могут стать вопросы об использовании критериев для автоматического выбора количества кластеров, использование различных методов кластеризации и метрик.

Список литературы

[1] William W. Cohen and Yoram Singer *Simple, Fast, and Effective Rule Learner*, AAAI/IAAI 1999: 335-342.

[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] 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, N 2, стр. 329-344.

[4] Журавлёв Ю. И., Рязанов В. В., Сенько, О. В. *«Распознавание». Математические методы. Программная система. Практические применения*, М., Фазис, 2006

[5] Журавлев Ю.И., Никифоров В.В. *Алгоритмы распознавания, основанные на вычислении оценок* Кибернетика. 1971. №3. С. 1-11.