

Обработка множеств логических закономерностей с помощью
дисперсионного критерия

Выпускная квалификационная работа

Выполнил: Лисяной А. Е.

Руководитель: д.ф.-м.н., проф. Рязанов В. В.

Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования
МГУ имени М. В. Ломоносова

28 мая 2015

Задачи выпускной квалификационной работы

- 1 Разработка метода кластеризации логических закономерностей
- 2 Построение множеств логических закономерностей небольшой мощности
- 3 Сравнение с существующими методами на прикладных задачах

Задача классификации

Определение задачи классификации

$X \in \mathbb{R}^D$ — пространство объектов

$Y = \{1, \dots, M\}$ — конечное множество имен классов

$X^l = (x_i, y_i)_{i=1}^l$ — обучающая выборка

Построить $a: X \rightarrow Y$, аппроксимирующий $y^*(x_i) = y_i$ на X .

Алгоритмы решения задачи классификации

- Метод логистической регрессии
- Метод опорных векторов
- Решающие деревья
- Нейронные сети
- Логические алгоритмы классификации

Определение логической закономерности

Пусть каждый объект выборки $x \in X^l$ имеет размерность D и пусть $\Omega \subseteq \{1, 2, \dots, D\}$. Предикат

$$\varphi(x) = P^{\Omega, c_1, c_2}(x) = \bigwedge_{j \in \Omega} P^{c_1^j, c_2^j}(f_j(x))$$

называется логической закономерностью класса K , если:

- ❶ $\exists x \in K: \varphi(x) = 1$
- ❷ $\forall x \notin K: \varphi(x) = 0$
- ❸ $\varphi(x)$ максимизирует некоторый критерий качества Φ .

Построенное множество логических закономерностей:

- может содержать большое количество правил
- может содержать похожие правила

Это приводит к тому, что:

- логические закономерности сложно интерпретировать
- по похожим правилам плохо проводить классификацию

Задача кластеризации множества логических закономерностей

- По исходному множеству логических закономерностей построить множество меньшей мощности, тем самым упростив задачу интерпретации полученных правил.
- Построенное множество должно иметь качество классификации, сравнимое с исходным множеством.

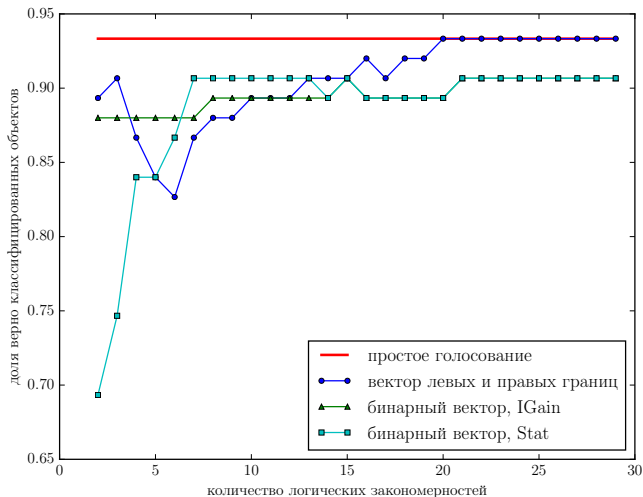
Алгоритм:

- ① Для каждого из t правил составить признаковое описание
 - Вектор левых и правых границ
 - Бинаризованное описание правил
- ② Кластеризовать на $k \leq t$ кластеров, найти их центры

$$S^* = \arg \min_S \sum_{i=1}^k \sum_{z_j \in S_i} \|z_j - \mu_i\|^2$$

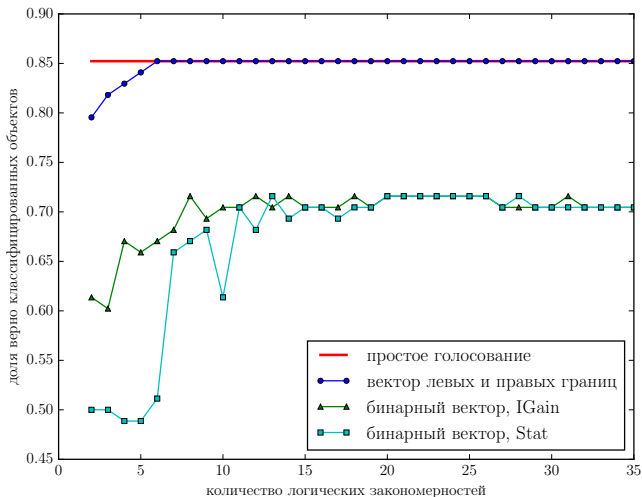
- ③ По центрам кластеров построить k новых правил
 - Выбрать центры кластеров в качестве новых правил
 - Центры кластеров + критерий качества \rightarrow новые правила

Эксперименты и сравнение на прикладных задачах



Выборка «Ирисы Фишера». Метод простого голосования.

Эксперименты и сравнение на прикладных задачах



Выборка «Вино». Метод простого голосования.

- Создан метод обработки множеств логических закономерностей с помощью кластеризации на основе дисперсионного критерия.
- Проведено сравнение метода обработки, использующего вектор левых и правых границ, и метода обработки, использующего бинаризованное описание логических закономерностей.
- Экспериментально показано, что удастся получить обработанное множество логических закономерностей с меньшим числом правил и сравнимым с исходным множеством качеством классификации.

Критерий Stat

p_1, \dots, p_K — верно выделяемые объекты

P_1, \dots, P_K — выделяемые объекты

$$\text{Stat} = -\ln \frac{C_{P_1}^{p_1} \dots C_{P_K}^{p_K}}{C_l^{p_1 + \dots + p_K}}$$

Критерий IGain

P, N — объекты из K и не из K

p, n — объекты из K и не из K , выделенные правилом

$$H(p, q) = -p \log_2(p) - q \log_2(q)$$

$$\begin{aligned} \text{IGain} = & H\left(\frac{P}{P+N}, \frac{N}{P+N}\right) - \frac{p+n}{P+N} H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) \\ & - \frac{P+N-p-n}{P+N} H\left(\frac{P-p}{P+N-p-n}, \frac{N-n}{P+N-p-n}\right) \end{aligned}$$