Выделение мультиграммных признаков в задачах классификации символьных последовательностей

Липатова Анна

Московский физико-технический институт
Факультет управления и прикладной математики
Научный руководитель: К. В. Воронцов

Группа 174, 2015

Цель исследования

Актуальность темы. Задача обработки и классификации символьных последовательностей является актуальной во многих сферах деятельности:

- медицина,
- биоинформатика и генетика,
- лингвистика.

Цель работы. Построить алгоритм классификации символьных последовательностей, максимизирующий значение функционала качества AUC(Area Under Curve).
Сравнить с ранее используемым методом классификации.

Задача классификации символьных последовательностей

Дано:
выборка $\mathfrak{D} = \{(x_i, y_i)\}_{i=1}^{p}$,
объекты x_i — символьные последовательности, ответы
$y_i \in \{X_m, X_o\}$ — классы объектов.
Требуется:

- построить алгоритм классификации $a(x) : \mathfrak{D} \to \{0, 1\}$, максимизирующий площадь под ROC-кривой AUC(Area Under Curve):

$$a = \arg \max_{a:D \to \{0,1\}} \{AUC(a, \mathfrak{D} \setminus T)\}.$$

- сравнить качество классификации при различных методах классификации

Наивный байесовский классификатор

Линейная модель классификации имеет вид:

$$a(x) = sign(\sum_{j=1}^{k} \gamma_j f_j(x) - \beta_m),$$

где γ_j — вес признака f_j, β_m — порог принятия решения для класса m.

Классификация с помощью признаков — частот n -грамм

n-грамма — последовательность из n букв, встречающихся в символьной последовательности $x_i \in \mathfrak{D}$.
Частота встречаемости $p_w(x_i)$ n -граммы $w = (w_0, \cdots, w_{n-1})$ в последовательности x_i:

$$r_w(x_i) = \sum_{r=1}^{N-n} \prod_{j=0}^{n-1} [s_{r+j} = w_j], \qquad p_w = \frac{r_w(x_i)}{N-n},$$

где s_j —j -й символ последовательности x_i.
Рассчитав значения частот встречаемости $p_1, \cdots, p_k$ для всевозможных n-грамм, получаем признаковое описание последовательности x_i.

Формулы для настройки весов признаков

$F_w(X_m)$ — среднее число вхождений n-граммы w в символьные последовательности объектов класса X_m,

$$F_w(X_m) = \frac{1}{|X_m|} \sum_{x_i \in X_m} p_w(S_{x_i}).$$

- $\gamma_w = 1$
- $\gamma_w = F_w(X_m)$
- $\gamma_w = F_w(X_m) - F_w(X_0)$
- $\gamma_w = ln(\frac{\tilde{F}_w(X_m)}{\tilde{F}_w(X_0)})$

Здесь

$$\tilde{F}_w(X_m) = \frac{1}{|X_m| + 1}(\sum_{S \in X_m} p_w(S)).$$

Настройка классификатора

Предположение

Каждый класс характеризуется своим набором n-грамм, называемым *диагностическим эталоном*.

Отбор n-грамм в диагностический эталон производится с помощью критерия информативности τ_w для данной n-граммы. Критерии информативности также можно варьировать:

- $\tau_w = F_w(X_m)$
- $\tau_w = F_w(X_m) - F_w(X_0)$
- $\tau_w = ln(\frac{\tilde{F}_w(X_m)}{\tilde{F}_w(X_0)})$
- $\tau_w = |ln(\frac{\tilde{F}_w(X_m)}{\tilde{F}_w(X_0)})|$

Классификация с помощью признаков — долей покрытия

Пусть в диагностический эталон $\mathcal{D}$ отобрано k n-грамм.
Покрытие последовательности x_i эталоном $\mathcal{D}$ — доля символов x_i, покрытых n-граммами эталона $\mathcal{D}$.
Доля покрытия θ — отношение мощности покрытия последовательности x_i к ее длине N.
Варьируя мощность $\mathcal{D}$, считаем доли покрытия $\theta_1, \cdots, \theta_k$ — новое признаковое описание для каждого объекта.

$$r_w(x_i) = \sum_{r=1}^{N-n} \prod_{j=0}^{n-1} [s_{r+j} = w_j]$$

$$\theta_j(x_i) = \frac{|\bigcap_{i=1}^{j} r_{w_i}(x_i)|}{N};$$

Настройка весов для новых признаков

Вместо частоты встречаемости n-граммы $F_j(X_m)$ и $F_j(X_0)$ используем усреднение $\hat{\theta}_j$ признака θ_j по символьным последовательностям объектов класса X_m и X_0 соответственно.

$$\hat{\theta}_j(X_m) = \frac{1}{|X_m|} \sum_{x_i \in X_m}^{n} \theta_j(x_i),$$

$$\hat{\theta}_j(X_0) = \frac{1}{|X_0|} \sum_{x_i \in X_0}^{n} \theta_j(x_i).$$

Можно использовать различные формулы для настройки весов γ_{θ_j} для новых признаков и, соответственно, различные критерии информативности τ_j.

Составной алгоритм

Вход: $\mathfrak{D} = \{(x_i, y_i)\}_{i=1}^{p}$ — генеральная выборка;
N — количество разбиений;
l —отношение мощностей обучающей и генеральной выборок;

Выход: $\hat{AUC}(k_1, k_2)$ — зависимость AUC от количества признаков двух типов в модели;

Идея

Добавить к диагностическому эталону дополнительные признаки - доли покрытия.

Цели эксперимента

- Сравнить рассмотренные методы классификации символьных последовательностей.
- Сравнить качество классификации при различных формулах весов.
- Оценить качество классификации используемых методов классификации.

Значения долей покрытия для объектов разных классов

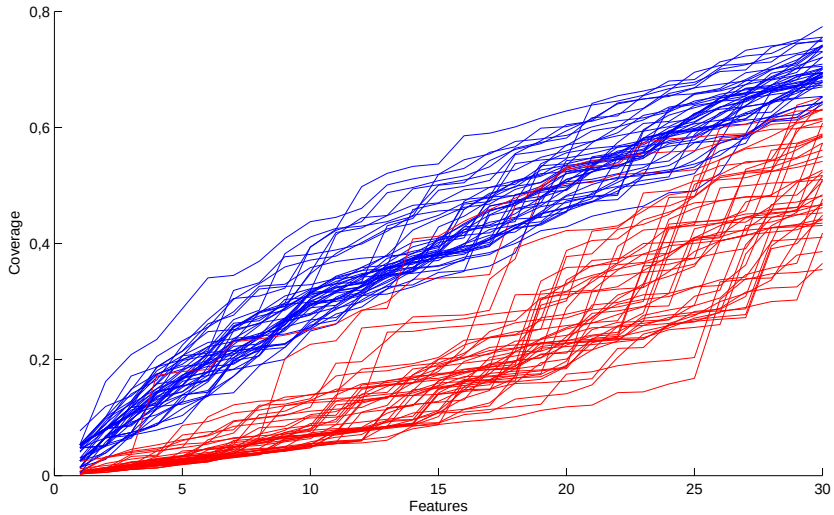


Рис.: Зависимость доли покрытия $\hat{\theta}_k$ от числа отобранных признаков k для больных ишемической болезнью сердца X_m (синяя кривая) и здоровых X_0 (красная кривая). N=200.

Вывод: доли покрытия можно использовать в качестве характерных признаков объектов класса больных.

Исследование близости признаков

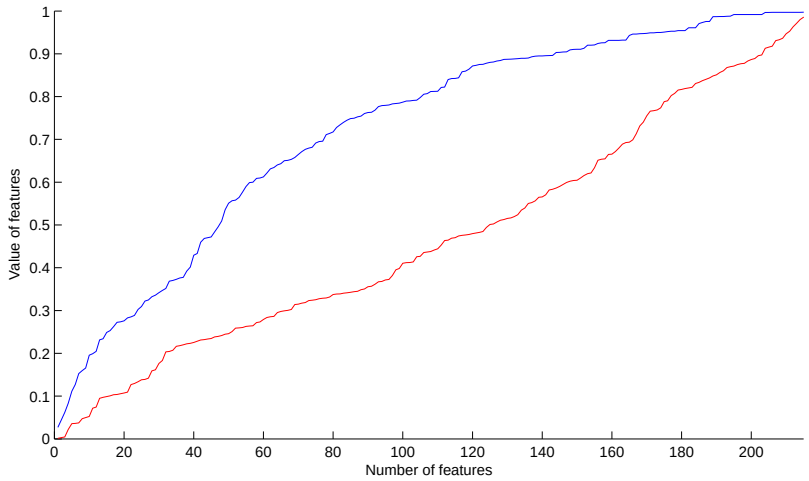
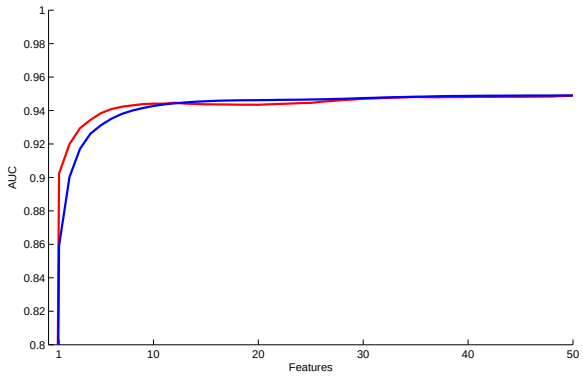


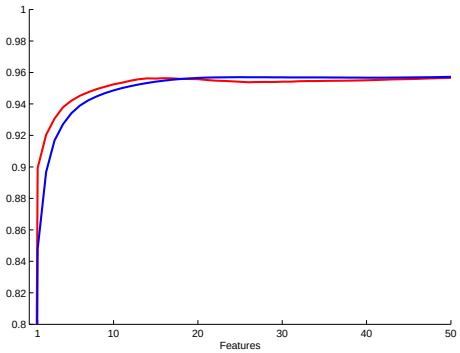
Рис.: Зависимость средней доли покрытия $\hat{\theta}_k$ (синяя кривая) и средней суммарной частоты встречаемости (красная кривая) от числа отобранных признаков k. (ИБС)

Вывод: признаки отличаются друг от друга. Покрытия учитывают возможное наложение триграмм.

Оценка качества классификации объектов по триграммам и долям покрытия



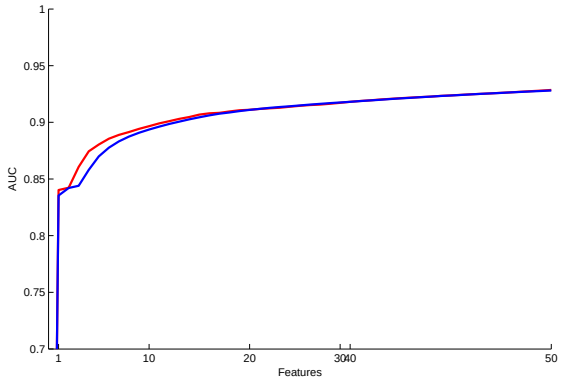
a) ИБС



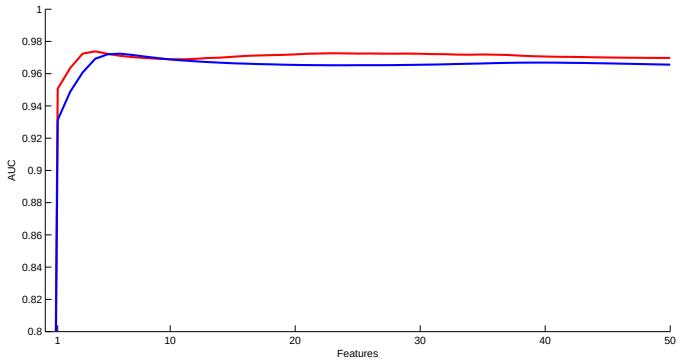
b) ЖДА

Рис.: Зависимость значения *AUC* при классификации с помощью триграмм (красная кривая) и долей покрытия (синяя кривая) от числа отобранных признаков k при логарифмической формуле весов. N=700.

Оценка качества классификации объектов по триграммам и долям покрытия



a) ДГПЖ



b) НГБК

Рис.: Зависимость значения *AUC* при классификации с помощью триграмм (красная кривая) и долей покрытия (синяя кривая) от числа отобранных признаков k при формуле весов $F_w(X_m) - F_w(X_0)$. N=700.

Оценка качества составного метода классификации

Болезнь	AUC (част.)	AUC (доли)	AUC (состав.)
ГБ	0,9589 (50)	0,9616 (50)	0,9595 (47,2)
ДГПЖ	0,9490 (50)	0,9489 (50)	0,9491 (9,45)
ДЖВП	0,9250 (50)	0,9244 (50)	0,9251 (12,41)
ЖДА	0,8761 (50)	0,8766 (50)	0,8766 (45,3)
ИБС	0,9581 (50)	0,9608 (50)	0,9583 (31,23)
МКБ	0,9257 (50)	0,9252 (50)	0,9256 (4,44)
НГБК	0,9777 (50)	0,9777 (50)	0,9782 (32,12)
РО	0,9491 (50)	0,9482 (50)	0,9489 (40,4)
СД	0,9572 (50)	0,9566 (50)	0,9572 (17,30)
ХГ1	0,9139 (50)	0,9152 (50)	0,9144 (3,43)
ХГ2	0,9331 (50)	0,9290 (50)	0,9340 (48,7)

Выводы

По результатам проведенного эксперимента можно сделать следующие выводы:

- можно использовать доли покрытия символьной последовательности в качестве признаков;
- целесообразно добавлять признаки-покрытия к набору информативных триграмм для повышения качества классификации;
- можно варьировать формулы весов признаков и критерии информативности.

Заключение

- предложен новый метод классификации символьных последовательностей, основанный на подсчете доли покрытия символьной последовательности набором наиболее информативных n-грамм;
- предложен метод, объединяющий два вышеописанных подхода к решению задачи классификации символьных последовательностей.
- произведено сравнение новых метода классификации с методом классификации символьных последовательностей с помощью подсчета частоты встречаемости n-грамм;