Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»
при Вычислительном центре им. А. А. Дородницына РАН

Липатова Анна Николаевна

Выделение мультиграммных признаков в задачах классификации символьных последовательностей

010900 — Прикладные математика и физика

Бакалаврская диссертация

Научный руководитель:
ст.н.с ВЦ РАН, д.ф.-м.н.
Воронцов Константин Вячеславович

Москва, 2015 г.

Содержание

Аннотация

Рассматривается задача классификации символьных последовательностей путем выделения мультиграммных признаков. В качестве базового алгоритма используется наивный байесовский классификатор. Сравниваются методы классификации символьных последовательностей путем отбора информативных признаков, которые рассчитываются на основе частот встречаемости n-грамм с учетом их пересечений и без. Проведены эксперименты, сравнивающие эти подходы к решению задачи классификации. Предложен алгоритм классификации, основанный на объединении методов классификации символьных последовательностей двумя вышеописанными способами.

Введение

Актуальность темы. Задача обработки и классификации символьных последовательностей [1] является актуальной во многих сферах деятельности: медицина — диагностика заболеваний [2], [3], биоинформатика и генетика — обработка и классификация генетических данных [4], лингвистика и обработка текстов — классификация и идентификация авторов текста [5].

Цель работы. Применение различных алгоритмов классификации, таких, как наивный байесовский классификатор, логистическая регрессия, алгоритм Random Forest для диагностики болезней по данным электрокардиограмм дает неплохие результаты. Использование наивного байесовского классификатора позволяет производить отбор признаков — выбирать наиболее информативные признаки для каждого класса. Использование различных принципов построения признаковых описаний влияет на качество диагностики. Целью данной работы является расширение набора информативных признаков в задаче классификации символьных последовательностей для повышения качества классификации.

Научная новизна.

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

Практическая ценность. Разработан программный модуль, который

- позволяет решать задачу классификации символьных последовательностей любым из рассматриваемых способов;

- позволяет выбрать оптимальную составную модель, содержащую различные признаки - доли покрытия и частоту встречаемости n-грамм, подобрав оптимальное число используемых признаков каждого типа;
- позволяет оценить качество классификации;
- визуализирует результаты.

1 Постановка задачи

Дана генеральная выборка $\mathfrak{D} = \{(x_i, y_i)\}_{i=1}^{p}$ состоящая из p пар объект-метка класса. Каждый объект принадлежит одному из двух классов: $y_i \in \{X_m, X_0\}$. Объектами являются символьные последовательности S конечной длины в конечном алфавите. По последовательности S длины N для объектов генеральной выборки строится признаковое описание

$$f(x) = (f_1(x), \ldots, f_t(x)).$$

В данной работе признаковое описание объекта строится по последовательностям из n букв, встречающихся в символьной последовательности S — n -граммам. Признак $f_w(x)$ принимает значение, равное частоте $p_w(S)$ n -граммы $w = (w_0, \cdots, w_{n-1})$ в последовательности S. Частота n -граммы w определяется как отношение её числа вхождений $r_w(S)$ в последовательность S к общему числу n -грамм в последовательности S, равному $N - n$:

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

где s_j —j -й символ последовательности S.

Задача классификации состоит в том, чтобы по выборке прецедентов двух классов построить алгоритм классификации $a(x) : \mathfrak{D} \to \{0, 1\}$, максимизирующий площадь под ROC-кривой AUC(Area Under Curve):

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

где T — обучающая выборка, $T \subset \mathfrak{D}$. Настрока классификатора производится по обучающей выборке T. Выбор AUC в качестве характеристики классификации связан с тем, что данная величина не зависит от соотношений цен ошибок первого и второго рода.

В данной работе по построенному признаковому описанию объекта (частоте встречаемости n -грамм) и соответствующей символьной последовательности S строится новое признаковое описание — покрытие символьной последовательности S набором информативных признаков, позволяющее улучшить качество классификации. Задача данной работы —расширить множество информативных признаков и улучшить качество классификации $a(x)$.

2 Описание алгоритма

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.

Хорошие результаты в задачах классификации дает наивный байесовский классификатор:

$$a(x) = [\ln \frac{\pi_m(S)}{\pi_0(S)} \geqslant \beta_m],$$

где $\pi_m(S)$ — модель плотности распределения класса y_m, β_m — порог принятия решений, зависит от соотношенния потерь от ошибок на объектах класса y_m и y_0.

Наивный байесовский классификатор действует в предположении, что все n-граммы в символьной последовательности S появляются независимо друг от друга и появления одной и той же n-граммы в символьной последовательности S независимы. Будем предполагать, что частоты n-грамм $p_w(S)$ в символьной последовательности S в каждом классе y_m являются независимыми случайными величинами $p_w(S)$. Тогда число появлений $r_w(S)$ n-граммы w в символьной последовательности S описывается распределением Пуассона, а многомерная плотность распределения $\pi_m(S)$ представляется в виде произведения одномерных плотностей:

$$\pi_m(S) = \prod_{w \in S} \frac{\lambda_{mw}^{r_w(S)}}{r_w(S)} exp(-\lambda_{mw}).$$

Несмещенная оценка $\lambda_{mw} = (N-3)F_w(X_m)$ параметра распределения Пуассона λ_{mw} совпадает со средним числом вхождений n-граммы w в символьные последовательности, соответствующие прецедентам класса y_m. Подставив эти оценки в плотности $\pi_m(S)$, а затем эти плотности в формулу классификатора, получим формулы для значения весов классификатора:

$$\gamma_w = log \frac{F_w(X_m)}{F_w(X_0)},$$

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

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

где $F_w(X_0)$ — среднее число вхождений триграммы w в символьные последовательности объектов класса X_0,

$$F_w(X_0) = \frac{1}{|X_0|} \sum_{x_i \in y_0} p_w(S_{x_i}).$$

Также можно использовать другие формулы для настройки весов признаков:

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

Здесь $\tilde{F}_w(X_m)$ — регуляризованная оценка частоты встречаемости n -граммы в символьных последовательностях класса X_j:

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

$$DF_w(X_m) = \frac{2F_w(X_m) - F_w^{max} - F_w^{min}}{F_w^{max} - F_w^{min}},$$

где

$$F_w^{min} = \min_{k=1,\ldots,K} F_w(X_m^k), \qquad F_w^{max} = \max_{k=1,\ldots,K} F_w(X_m^k),$$

а X_m^k — k-я выборка, получаемая случайными перестановки элементов символьных последовательностей класса X_m.

Каждый класс характеризуется своим набором n-грамм, называемым *диагностическим эталоном*. Отбор n-грамм в диагностический эталон производится с помощью критерия информативности для данной n-граммы. Критерии информативности, как и формулы подсчета весов, можно варьировать:

- $\tau_w = F_w(X_m)$
- $\tau_w = F_w(X_m)[w \notin T_0]$
- $\tau_w = F_w(X_m) - F_w(X_0)$
- $\tau_w = ln(\frac{\tilde{F}_w(X_m)}{\tilde{F}_w(X_0)})$
- $\tau_w = |ln(\frac{\tilde{F}_w(X_m)}{\tilde{F}_w(X_0)})|$

- $\tau_w = DF_w(X_m)$.

В диагностический эталон отбираются k признаков —n-грамм с наибольшими значениями выбранного критерия информативности. Использование шумовых n-грамм ухудшает качество классификации. Идеей предлагаемого в данной работе алгоритма является добавить к набору отобранных информативных признаков дополнительные признаки — доли покрытия символьной последовательности n-граммами диагностического эталона.

2.1 Генерация новых признаков

С помощью различных критериев информативности можно осуществлять отбор n-грамм в диагностический эталон $\mathcal{D}$. Мощность диагностического эталона можно варьировать. Пусть в диагностический эталон $\mathcal{D}$ отобрано k n-грамм.

Покрытием символьной последовательности S диагностическим эталоном $\mathcal{D}$ будем называть объединение всевозможно расположенных n-грамм из диагностического эталона $\mathcal{D}$, содержащихся в символьной последовательности S.

Долей покрытия θ символьной последовательности S диагностическим эталоном $\mathcal{D}$ назовем отношение мощности покрытия символьной последовательности S к длине N символьной последовательности S.

Предполагается, что если диагностический эталон $\mathcal{D}$ покрывает большую часть символьной последовательности S, значит, n-граммы диагностического эталона $\mathcal{D}$ часто встречаются в символьной последовательности S и объект принадлежит классу y_m. Варьируя размер k диагностического эталона $\mathcal{D}$, можно получить различные значения долей покрытия γ_k символьной последовательности S диагностическим эталоном $\mathcal{D}$. Посчитав таким образом доли покрытия $\theta_1, \cdots, \theta_k$ для каждого объекта и считая эти значения новыми признаками, получим новое признаковое описание для каждого объекта.

2.2 Настройка линейного классификатора для новой модели

Для настройки весов линейного классификатора $a(x)$ с k новыми признаками — долями покрытия θ_j можно пользоваться формулами (2), используя вместо частоты встречаемости n-граммы $F_j(X_m)$ и $F_j(X_0)$ усреднение $\hat{\theta}_j$ признака θ_j по символьным

последовательностям объектов класса X_m и X_0 соответственно.

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

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

Таким образом, можно использовать различные формулы для настройки весов для новых признаков и, соответственно различные критерии информативности:
Формулы для подсчета весов для признаков — долей покрытия:

- $\gamma_{\theta_j} = \hat{\theta}_j(X_m)$
- $\gamma_{\theta_j} = \hat{\theta}_j(X_m) - \hat{\theta}_j(X_0)$
- $\gamma_{\theta_j} = \ln(\frac{\hat{\theta}_j(X_m)}{\hat{\theta}_j(X_0)})$
- $\gamma_{\theta_j} = \ln(\frac{\hat{\theta}_j(X_m)}{\hat{\theta}_j(X_0)})$
- $\gamma_{\theta_j} = D\hat{\theta}_j(X_m)$

Формулы для критериев инфоматичности для признаков — долей покрытия:

- $\tau_{\theta_j} = \hat{\theta}_j(X_m)$
- $\tau_{\theta_j} = \hat{\theta}_j(X_m)[j \notin T_0]$
- $\tau_{\theta_j} = \hat{\theta}_j(X_m) - \hat{\theta}_j(X_0)$
- $\tau_{\theta_j} = ln(\frac{\hat{\theta}_j(X_m)}{\hat{\theta}_j(X_0)})$
- $\tau_{\theta_j} = |ln(\frac{\hat{\theta}_j(X_m)}{\hat{\theta}_j(X_0)})|$
- $\tau_{\theta_j} = D\hat{\theta}_j(X_m)$

Таким образом, можно производить отбор новых полученных признаков для добавления к диагностическому эталону $\mathcal{D}$ класса X_m.

2.3 Описание алгоритма

На вход алгоритма подается выборка $\mathcal{D} = \{(S_i, y_i)\}_{i=1}^{p}$— множество символьных последовательностей S_i фиксированной длины и ответов y_i — принадлежность данной сивольной последовательности S_i к одному из классов X_m или X_0.

Для каждой символьной последовательности S выборки $\mathfrak{D}$ вычисляются частоты встречаемости n-грамм.

Далее множество $\mathfrak{D}$ разбивается на две подвыборки: обучающую выборку $\mathfrak{T}$ и контрольную выборку $\mathfrak{D} \setminus T$.

Затем, для каждого разбиения настройка весов классификатора (2) производится по обучающей выборке T.

Далее с помощью выбранного критерия информативности (2) отбирается k_1 наиболее информативных признаков-частот встречаемости n-грамм и k_2 наиболее информативных n-грамм и для всех символьных последовательностей множества $\mathfrak{D}$ вычисляются значения нового признака — доли покрытия θ_{k_2} символьной последовательности S отобранными k_2 n-граммами.

Получаем для каждой символьной последовательности S расширенное признаковое описание: k_1 отобранных признаков — частот n-грамм и долю покрытия $\hat{\theta_{k_2}}$ символьной последовательности S отобранными k_2 n-граммами.

Варьируя параметры k_1 и k_2 и сравнивая значения AUC выбираем алгоритм классификации, максимизирующий AUC на контрольной выборке $\mathfrak{D} \setminus T$.

Для того, чтобы результаты не зависели от конкретного разбиения выборки $\mathfrak{D}$ на обучение и контроль, производим L разбиений и усредняем полученные значения AUC по всем разбиениям. Таким образом, алгоритм можно представить в виде псевдокода:

Составной метод классификации символьных последовательностей

Вход: $\mathfrak{D} = \{(x_i, y_i)\}_{i=1}^{p}$ — генеральная выборка;

K — максимальное количество признаков в модели;

N — количество разбиений генеральной выборки на обучающую и контрольную;

l —отношение мощностей обучающей и генеральной выборок;

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

для каждой символьной последовательности $S \in \{x_1, \ldots, x_p\}$ рассчитать частоты p_w;

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

для всех $i = 1, \ldots, N$

  разбить выборку $\mathfrak{D}$ на обучающую $\mathfrak{T}$ и контрольную $\mathfrak{D} \setminus \mathfrak{T}$;

  по выборке $\mathfrak{T}$ рассчитать τ_w и γ_w для признаков p_w;

  отсортировать p_w по убыванию τ_w;

  рассчитать значения θ_j для объектов $\mathfrak{D}$ и значения γ_{θ_j} и τ_{θ_j} для признаков θ_j;

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

  отсортировать признаки θ_j по убыванию τ_{θ_j};

  для всех $k_1 = 0, \ldots, K$, $k_2 = 0, \ldots, K$

    отобрать k_1 первых признаков $p_{w_1}, \ldots, p_{w_{k_1}}$
    и k_2 первых признаков $\theta_1, \ldots, \theta_{k_2}$;

    рассчитать значение AUC для построенной модели на контрольной выборке $\mathfrak{D} \setminus \mathfrak{T}$;

усредняем значение $AUC(k_1, k_2)$ по всем разбиениям:
$\hat{AUC}(k_1, k_2) = \frac{\sum_{i=1}^{N} AUC(k_1, k_2)}{N}$;

3 Вычислительный эксперимент

В экспериментах используются данные электрокардиограмм различных пациентов, обработанных при помощи технологии информационного анализа [2], [7] кардиосигналов. Технология информационного анализа электрокардиосигналов основана на преобразовании каждой электрокардиограммы сначала в последовательность интервалов и амплитуд кардиоциклов, а затем - в символьную последовательность фиксированной длины, называемую *кодограммой*. В качестве признаков в нижеописанных экспериментах рассматриваются триграммы.

3.1 Значения долей покрытия для объектов разных классов

Пусть в диагностический эталон $\mathcal{D}$ отобрано k наиболее информативных триграмм. Варьируя k от 1 до K и считая для каждого k доли покрытия $\hat{\theta}_k$ получим K различных признаков $(\hat{\theta}_1, \cdots, \hat{\theta}_K)$. Настроив веса для классификатора с полученными признаками-покрытиями по формулам (2.2) и упорядочив полученные признаки согласно критерию информативности (2.2) построим зависимость средней доли покрытия объектов классов X_0 и X_m. На рис. 1 изображена зависимость средней величины доли покрытия $\hat{\theta}_k$ от числа k отобранных признаков для больных с диагнозом ишемическая болезнь сердца и здоровых. Для того, чтобы результаты не зависели от выбранных кодограмм каждого класса, выбирается N кодограмм каждого класса и полученные значения усредняются. Получаем, что средняя доля покрытия объектов класса больных значимо выше средней доли покрытия объектов класса здоровых, что позволяет предположить, что доли покрытия можно использовать в качестве характерных признаков объектов класса больных.

3.2 Схожесть признаков

Сравним доли покрытия для разных мощностей диагностического эталона с суммарной частотой триграмм, входящий в диагностический эталон. Если значения различаются - значит, признаки различны. Такое происходит, т. к. рассмотрение покрытий в качестве признаков позволяет учитывать возможные наложения триграмм друг на друга. На графике изображена зависимость средней доли покрытия $\hat{\theta}_k$ (синяя кривая) и средней суммарной частоты встречаемости (красная кривая) от числа отобранных признаков k для больных ишемической болезнью сердца. Значит,

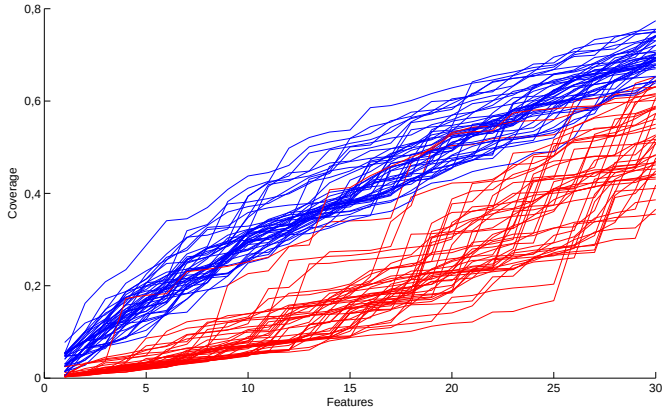


Рис. 1: На графике изображена зависимость величины доли покрытия $\hat{\theta}_k$ от числа отобранных признаков k для классов больных ишемической болезнью сердца X_m (синяя кривая) и здоровых X_0 (красная кривая). N=150.

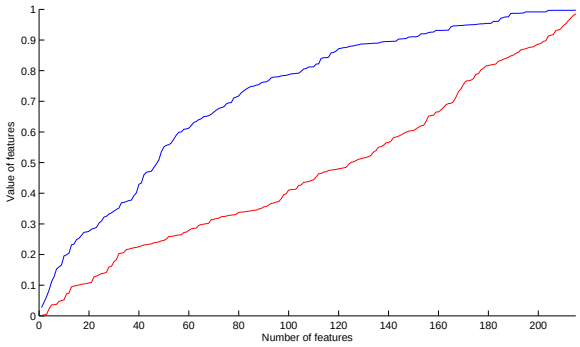


Рис. 2: Зависимость средней доли покрытия $\hat{\theta}_k$ (синяя кривая) и средней суммарной частоты встречаемости (красная кривая) от числа отобранных признаков k. (ИБС)

признаки действительно отличаются.

3.3 Оценка качества классификации объектов по триграммам

В данном эксперименте будем рассматривать модель классификатора, содержащую в качестве признаков только триграммы диагностического эталона. Рассмотрим зависимость значения AUC от числа триграмм, входящих в диагностический эталон. Веса классификатора настраиваются по обучающей выборке T, а значение AUC считается на контрольной выборке $\mathfrak{D} \setminus T$. Для того, чтобы результаты не зави-

сели от конкретного разбиения, производится N разбиений и полученные значения усредняются. На рис. 2 изображена зависимость $AUC(k)$ для данного способа классификации.

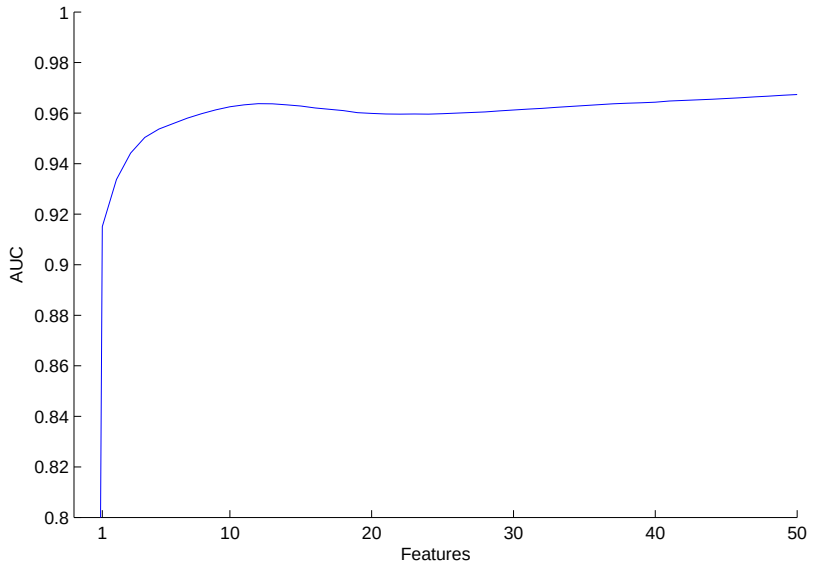
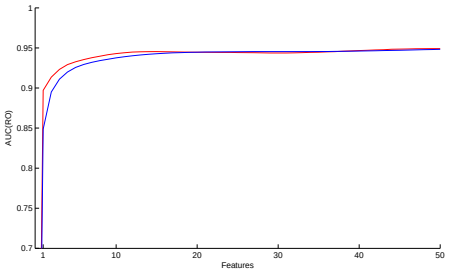


Рис. 3: На графике изображена зависимость значения AUC при классификации с использованием в качестве признаков часот встречаемости p_w триграмм в кодограмме от числа отобранных признаков k для больных ишемической болезнью сердца при логарифмической формуле весов. N=700.

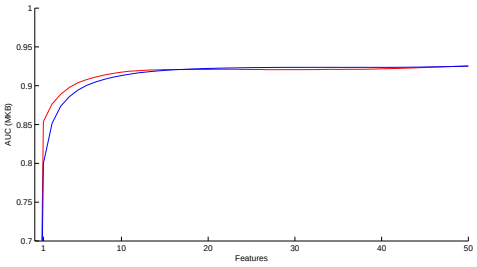
3.4 Оценка качества классификации объектов по долям покрытия

Теперь рассмотрим модель классификатора, содержащую в качестве признаков только доли покрытия $\hat{\theta}_k$ кодограмм k отобранными триграммами диагностического эталона. Рассмотрим зависимость значения AUC от числа триграмм, входящих в диагностический эталон. Сравним зависимость значения AUC от числа триграмм, входящих в диагностический эталон при классификации с помощью признаков — частот n -грамм и с помощью признаков — долей покрытия (Рис.4). Все веса настраиваются по обучающей выборке T, а значение AUC считается на контрольной выборке $\mathfrak{D} \setminus T$. Для того, чтобы результаты не зависели от конкретного разбиения, производится N разбиений и полученные значения усредняются.

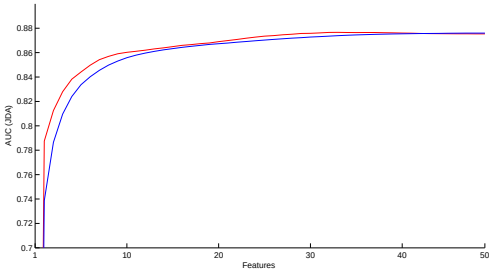
Рис. 4: Зависимость значения AUC при классификации с использованием в качестве признаков долей покрытия $\hat{\theta}_k$ от числа отобранных признаков k при логарифмической формуле весов. N=700.



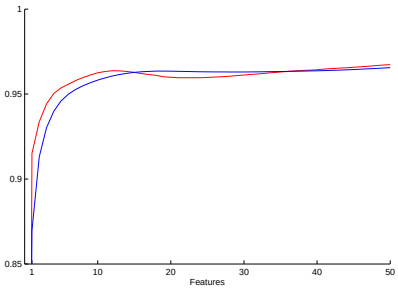
a) РО



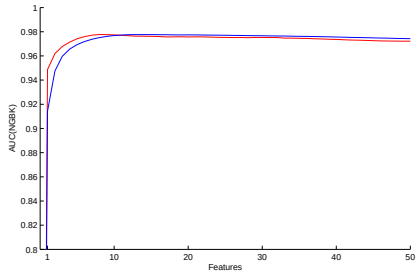
b) МКБ



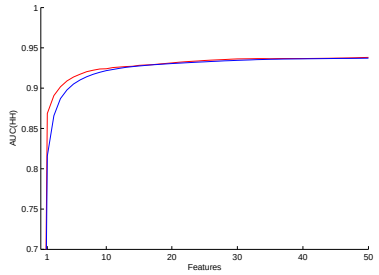
c) Железодефицитная анемия



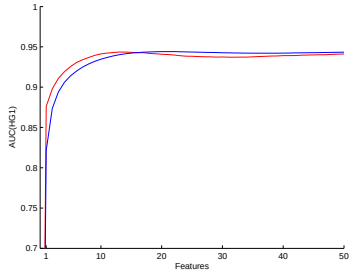
d) Ишемическая болезнь сердца



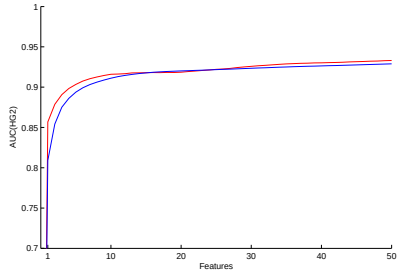
e) НГБК



f) ХХ

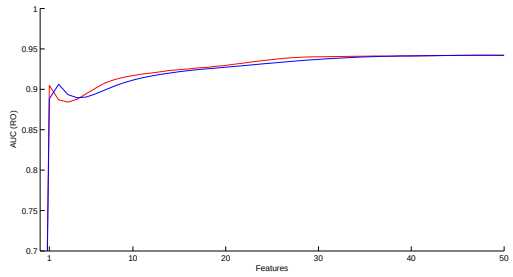


g) ХГ1

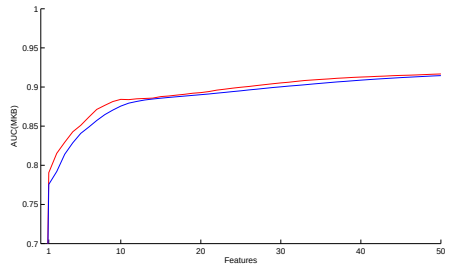


h) ХГ2

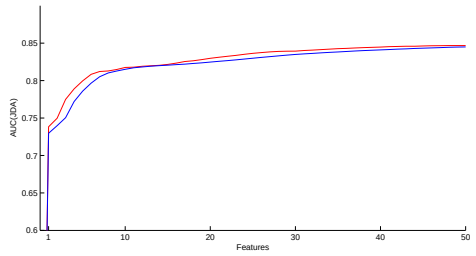
Рис. 5: Зависимость значения AUC при классификации с использованием в качестве признаков долей покрытия $\hat{\theta}_k$ от числа отобранных признаков k при формуле весов $F_w(X_m) - F_w(X_0)$. N=700.



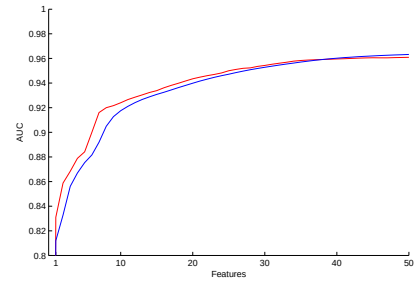
a) РО



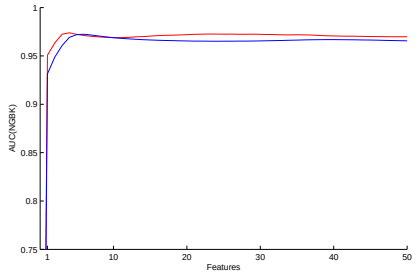
b) МКБ



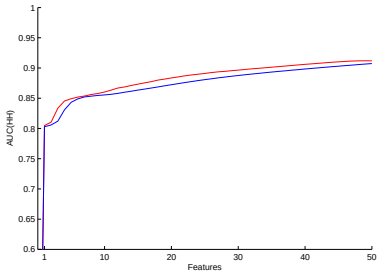
c) Железодефицитная анемия



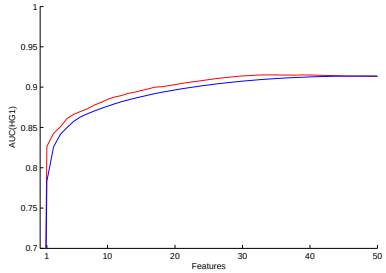
d) Ишемическая болезнь сердца



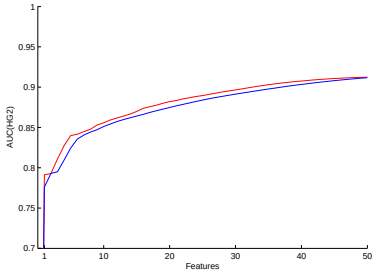
e) НГБК



f) ХХ



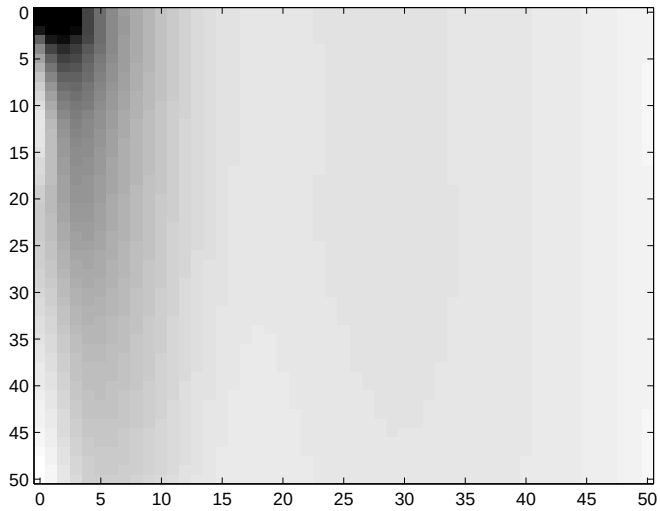
g) ХГ1



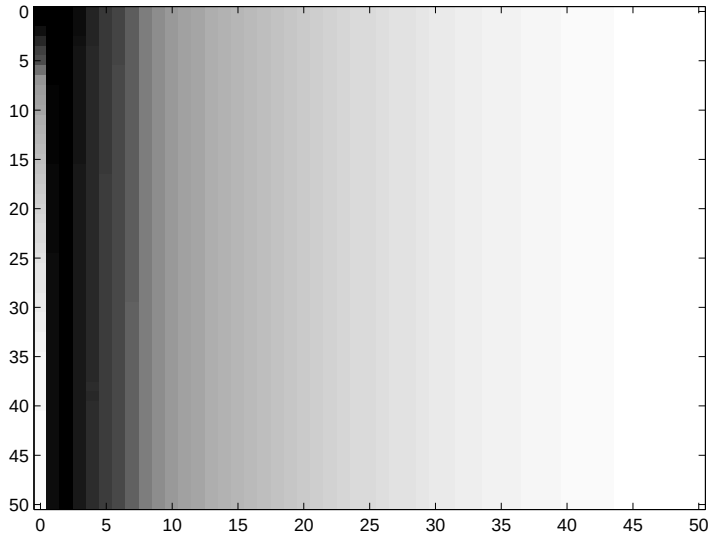
h) ХГ2

3.5 Оценка качества составного метода классификации

В данном эксперименте рассмотрим составную модель классификатора — модель, содержащую k_1 отобранных признаков-триграмм и k_2 признака-доли покрытия $(\hat{\theta}_1, \cdots, \hat{\theta}_{k_2})$. Варьируя k_1 и k_2 получаем различные модели классификатора. На рис. 4 отображена зависимость значений AUC от числа k_1 отобранных признаков - триграмм и числа k_2 отобранных признаков-долей покрытия $\hat{\theta}_{k_1}$.



a)



b)

Рис. 6: Зависимость значения AUC от числа k_1 отобранных признаков -триграмм и числа k_2 отобранных признаков-долей покрытия $\hat{\theta}_{k_1}$ для больных ишемической болезнью сердца при формулах весов а) $\log \frac{F_w(X_m)}{F_w(X_0)}$*, b)* $F_w(X_m) - F_w(X_0)$*. Черный цвет соответствует значению AUC менее 0,95, белый цвет соответствует значению AUC 0,9658. N=700*

В таблицах содержатся результаты экспериментов, которые отображают полученные значения AUC для различных болезней при различных формулах весов при классификации тремя рассмотренными способами.

Болезнь	AUC (триграммы)	AUC (покрытия)	AUC (триграммы, покрытия)
ВСД	0,8803 (50)	0,8804 (50)	0,8803 (8,37)
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)
ЖКБ	0,9037 (50)	0,9031 (50)	0,9042 (41,12)
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)
ХХ	0,9381 (50)	0,9372 (50)	0,9377 (23,25)
ЯБ	0,8800 (50)	0,8793 (50)	0,8811 (37,17)

Таблица 1: В таблице приведены значения AUC для наилучших моделей при трех способах классификации: с помощью признаков-частот триграмм, с помощью признаков-покрытий и составным методом. В скобках указаны параметры модели - количество признаков каждого типа. Рассматривается бинарная классификация для различных болезней. Используется формула весов $\log \frac{F_w(X_m)}{F_w(X_0)}$.

Болезнь	AUC (триграммы)	AUC (покрытия)	AUC (триграммы, покрытия)
ВСД	0,8957 (50)	0,8965 (50)	0,8963 (26, 25) (50)
ГБ	0,9335 (50)	0,9304 (50)	0,9327 (31,18)
ДГПЖ	0,9281 (50)	0,9286 (50)	0,9287 (1,49)
ДЖВП	0,8973 (50)	0,8932 (50)	0,8951 (41,6)
ЖДА	0,8449 (50)	0,8466 (50)	0,8456 (23,27)
ЖКБ	0,9186 (50)	0,9187 (50)	0,9184 (45,3)
ИБС	0,9657 (50)	0,9657 (50)	0,9658 (35,12)
МКБ	0,8891 (50)	0,8887 (50)	0,8892 (4,44)
НГБК	0,9724 (50)	0,9738 (50)	0,9732 (39,6)
РО	0,9422 (50)	0,9417 (50)	0,9420 (10,40)
СД	0,9107 (50)	0,9108 (50)	0,9108 (17,32)
ХГ1	0,9139 (50)	0,9152 (50)	0,9144 (3,43)
ХГ2	0,9123 (50)	0,9117 (50)	0,9117 (6,39)
ХХ	0,9073 (50)	0,9119 (50)	0,9091 (12,45)
ЯБ	0,8764 (50)	0,8766 (50)	0,8771 (7, 45)

Таблица 2: В таблице приведены значения AUC для наилучших моделей при трех способах классификации: с помощью признаков-частот триграмм, с помощью признаков-покрытий и составным методом. В скобках указаны параметры модели - количество признаков каждого типа. Рассматривается бинарная классификация для различных болезней. Используется формула весов $F_w(X_m) - F_w(X_0)$.

Болезнь	AUC (триграммы)	AUC (покрытия)	AUC (триграммы, покрытия)
ВСД	0,8387	0,8384	0,8389 (18,32)
ГБ	0,9231 (50)	0,9214 (50)	0,9229 (31,18)
ДГПЖ	0,9281 (50)	0,9286 (50)	0,9287 (43,6)
ДЖВП	0,8973 (50)	0,8932 (50)	0,8951 (6,41)
ЖДА	0,8559 (50)	0,8569 (50)	0,8556 (9,42)
ЖКБ	0,8342 (50)	0,8338 (50)	0,8345 (48,6)
ИБС	0,9657 (50)	0,9657 (50)	0,9658 (35,12)
МКБ	0,8930 (50)	0,8933 (50)	0,8933 (24,30)
НГБК	0,9724 (50)	0,9738 (50)	0,9732 (39,6)
РО	0,9422 (50)	0,9427 (50)	0,9434 (19,29)
СД	0,8912 (50)	0,8913 (50)	0,8912 (17,30)
ХГ1	0,9139 (50)	0,9152 (50)	0,9144 (13,38)
ХГ2	0,9274 (50)	0,9271 (50)	0,9276 (8,46)
ХХ	0,9098 (50)	0,9099 (50)	0,9097 (11,39)
ЯБ	0,9103 (50)	0,9105 (50)	0,9105 (15,33)

Таблица 3: В таблице приведены значения AUC для наилучших моделей при трех способах классификации: с помощью признаков-частот триграмм, с помощью признаков-покрытий и составным методом. В скобках указаны параметры модели - количество признаков каждого типа. Рассматривается бинарная классификация для различных болезней. Используется формула весов $F_w(X_m)$.

Получаем, что при добавлении новых признаков в модель при одном и том же суммарном количестве признаков качество классификации у смешанной модели для определенных моделей оказывается выше, что говорит о том, что целесообразно добавлять признаки-покрытия к набору информативных триграмм для повышения качества классификации.

4 Заключение.

В качестве признаков модели линейного классификатора для символьных последовательностей можно использовать не только частоты встречаемости последовательностей из n букв — n-грамм, но и доли покрытия символьной последовательности этими n-граммами. Модель, содержащая в качестве признаков только доли покрытия символьной последовательности классифицирует объекты достаточно хорошо. При этом при добавлении новых признаков-долей покрытия в модель, содержащую только признаки n-граммы, при одном и том же суммарном количестве признаков качество классификации у смешанной модели оказывается выше, чем у модели с признаками-n-граммами, что говорит о том, что целесообразно добавлять признаки-покрытия к набору информативных триграмм для повышения качества классификации.

5 Литература

Список литературы

[1] Gorban A.N., Popova T.G., Sadovsky M.G. Classification of symbol sequences over thier frequency dictionaries: towards the connection between structure and natural taxonomy *Open System and Inform. Dyn.* 2000. Vol. 7, N 1. P. 1–17

[2] Успенский В.М. Информационная функция сердца в диагностике заболеваний внутренних органов. *Военно-медицинский журнал*, — Т. 188. — 2010. — № 9. – С. 45- 51.

[3] V. Uspenskiy. Diagnostic System Based on the Information Analysis of Electrocardiogram. *Proceedings of MECO 2012. Advances and Challenges in Embedded Computing. Bar, Montenegro,* June 19-21, 2012, p. 74-76.

[4] Гельфанд М. С. Компьютерный анализ последовательностей ДНК. *Молекулярная биология*, 1998.

[5] Романов А. В. Методика идентификации автора текста на основе аппарата опорных векторов. *Аудит информационной безопасности.*, 2009

[6] Успенский В.М. Информационная функция сердца. *Клиническая медицина*, — 2008. – Т. 86. — №5. – С. 4-13.

[7] Успенский В.М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов методом информационного анализа электрокардиосигналов.- *М.: «Экономика и информация»*, 2008. -116 с.

[8] Успенский В.М. Информационная функция сердца. Теория и практика диагностики заболеваний внутренних органов *Вестник МГАДА. Серия «Философские, социальные и естественные науки». М.*, 2011, № 1(7). — С. 104-112.

[9] V. Uspenskiy. Information Function of the Heart. A Measurement Model *Proceedings of the 8-th International Conference, Slovakia.* 2011, p. 383-386.

[10] Успенский В. М., Воронцов К. В., Целых В. Р. Статистические обоснования информационного анализа электрокардиосигналов для диагностики заболеваний внутренних органов. *Интеллектуальный анализ данных,* 2014.

[11]

6 References