Автоматическое выделение признаков в задаче классификации сигналов

© В.А. Викулин

Московский государственный университет имени М.В. Ломоносова,
Москва, Россия

va.vikulin@physics.msu.ru

Аннотация. Рассмотрена задача классификации сигналов. Предложен стохастический алгоритм, позволяющий выделять качественное признаковое пространство в этой задаче. Показан принцип разложения признаков сигнала через базисные функции и представлен примерный набор базисных функций, который можно использовать в задачах анализа сигналов. Алгоритм строит каждый признак с помощью максимизации некоторого функционала качества, оптимизируя данное разложение. Предложено несколько вариантов таких функционалов качества. Стохастически проводя такую процедуру, можно синтезировать качественное признаковое пространство. Алгоритм проверен на задаче классификации ЭКГ сигналов, в которой по кардиограмме пациента определялось наличие у него ишемической болезни сердца.

Ключевые слова: анализ сигналов, выделение признаков, задача классификации.

Automatic Feature Extraction for Signals Classification

© V. Vikulin

Lomonosov Moscow State University,
Moscow, Russia

va.vikulin@physics.msu.ru

Abstract. The article is concerned with the signal classification problem. The article suggests an algorithm which allows to create feature space in this task. It shows a strategy performing features with the basic functions, some set of basic functions provided to be used in multiple signal processing problems. The algorithm creates each feature by maximizing some function of feature quality and some sets of possible feature quality metrics for solving signal processing tasks are recommended. If you repeat this procedure randomly you will create feature subspace. The algorithm was tested on ECG classification problem in which algorithm defined the presence of coronary disease in patients.

Keywords: signal processing, feature extraction, classification.

1 Введение

Сигнал – последовательность измерений некоторой величины. Задача классификации сигналов часто встречается во множестве различных прикладных задач – от медицины до приборостроения [5].

Не так давно на рынке стало доступно множество мобильных приборов, которые могут непрерывно записывать кардиограмму человека. Количество подобных приборов растет с каждым днем, ведь такие медицинские измерения не только удобны, но и позволяют своевременно обнаружить болезнь. Необходимо не только быстро обрабатывать все эти показания, но и быстро находить среди них больных людей, а это возможно сделать только методами анализа сигналов. Разработка алгоритмов классификации сигналов становится важной и актуальной задачей. Одним из популярных подходов к решению задачи классификации сигналов является нахождение оптимального признакового пространства, в котором объекты (сигналы) могут наиболее просто быть разделены с помощью классических алгоритмов классификации.

Рассмотрим постановку задачи классификации. Пусть заданы множество объектов X, множество допустимых ответов Y, и существует функция $y: X \to Y$, значения которой известны только на конечном подмножестве объектов $\{x_1 \dots x_l\} \subset X$. Задача заключается в том, чтобы по имеющимся парам объект–ответ восстановить исходную зависимость, то есть построить решающую функцию $a: X \to Y$, которая приближала бы

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

целевую функцию y, причём не только на известных объектах, но и на всём множестве X. Признаком объекта x назовем результат измерения характеристики объекта. Другими словами, признаком называется отображение $f : X \to D_f$, где D_f – множество допустимых значений признака. Нахождению оптимального множества таких отображений $\{f\}$ посвящена настоящая работа.

В данном случае под множеством объектов X будем всегда иметь в виду множество сигналов, то есть конечную последовательность вещественных чисел, а под множеством допустимых ответов Y – двухэлементное множество $\{-1, 1\}$.

Нахождение оптимального признакового пространства является крайне сложной задачей. В большинстве случаев в задачах анализа сигналов для конструирования этого пространства применяются классические приемы, основанные в своем большинстве на преобразовании Фурье и вейвлет-преобразовании [6, 8]. Такой подход имеет ряд существенных недостатков. Он требует глубокого понимания от исследователя природы сигнала, и исследователь должен сам подбирать необходимое спектральное разложение; не существует некоторого универсального преобразования, которое бы позволяло всегда выделять оптимальное признаковое пространство из сигнала. Из-за этого недостатка те качественные признаки, которые были найдены в предыдущей задаче анализа сигнала, в новой задаче могут быть абсолютно неприменимы. В следующей задаче анализа сигналов исследователю необходимо с нуля конструировать признаковое пространство, опираясь исключительно на свою интуицию и опыт, полученный при решении предыдущих задач

Данная работа посвящена методу автоматического построения признакового пространства с помощью максимизации критерия качества признака (здесь и далее признаком сигнала будем называть любую вещественную функцию от сигнала). Нами использовался метод оптимизации, который является обобщением «жадного поиска», но использование конкретно этого метода оптимизации совершенно не обязательно. Поиск оптимального признакового пространства при этом являлся стохастическим, то есть в самом алгоритме заложена рандомизация, что позволяет постоянно генерировать новые признаки, отличающиеся от предыдущих. Критериев качества для оценки признака было несколько, и они тоже выбирались для каждого признака случайно. Это также помогало генерировать непохожие друг на друга признаки, так как не существует универсального метода оценки качества, при этом нельзя максимизировать их все сразу. Благодаря стохастике, данный алгоритм позволял за N итераций почти всегда найти N непохожих друг на друга признаков. Далее синтезированное множество признаков может использоваться любым классическим классификатором.

Данный подход уже применялся несколько раз в анализе сигналов [1, 3, 4, 7]. В этих работах используется генетический алгоритм для нахождения оптимального признакового пространства. Генетический алгоритм является алгоритмом оптимизации, который базируется на механизмах, в какой-то степени аналогичных механизмам эволюции в живой природе. В качестве функции, которая оптимизируется генетическим алгоритмом, выступает какая-либо мера качества признака. Например, качество предсказания алгоритма, построенного на синтезированном признаке на кросс-валидации. Основным недостатком данных работ является четкая привязка как к методу оптимизации, так и к выбору оценки качества признака. Из-за жесткой привязки к оценке качества генетический алгоритм является хорошим выбором, потому что он не старается наивным образом подобрать себе лучшее решение, как это делает, например, жадный алгоритм. Если бы в этих работах использовался «жадный» алгоритм, то признаковое пространство было бы бедным и все время одинаковым, так как этот алгоритм не обладает нужной вариативностью. Одной из важнейших задач данной работы является построение метода, в который легко бы встраивался абсолютно любой метод оптимизации, то есть предлагается обеспечивать нужную вариативность не с помощью метода оптимизации, а с помощью стохастической природы поиска оптимального признакового пространства. В этом случае метод оптимизации может быть любым, он не будет определяющим в конструкции.

2 Стохастический алгоритм синтеза признакового пространства

Рассмотрим предлагаемый алгоритм синтеза признакового пространства.

2.1 Представление признака через базисные функции

Напомним, что признаком сигнала называется функция от сигнала, которая ставит в соответствие сигналу какое-то число. Будем раскладывать каждую такую функцию через набор заранее определенных базисных функций, в рамках которых мы и будем проводить оптимизацию. Таким образом, выбор базисных функций однозначно определит пространство, в котором будет происходить оптимизация. Каждый признак при этом будет представлять собой суперпозицию базисных функций, которые будут применяться поочередно, формируя в итоге значение признака. Пусть мы выбрали множество $\{b\}$ мощности N базисных функций, тогда любой признак сигнала может быть представлен в виде:

$$f(x) = [b_1][b_2] \dots [b_{last}](x), \quad (1)$$

где b_i – очередная базисная функция из множества $\{b\}$, прямоугольные скобки используются для разделения базисных функций и отдельного смысла не несут. Далее будем подразумевать, что функции в выражении (1) применяются слева направо. Эта форма записи не согласуется с привычными правилами записи подобных выражений в математике, но была выбрана из соображений наглядности.

Заметим, что в формуле (1) каждый признак может быть представлен через любое число базисных функций. Конкретно взятая базисная функция может быть использована в представлении признака неограниченное, но обязательно конечное число раз.

Отметим также, что если мы определили, что признаком сигнала является число, то последняя из базисных функций в формуле (1) обязана быть вида $b_{last}: R^m \to R$. Остальные функции должны быть согласованы по областям задания и областям значений: $b_i: R^{m_i} \to R^{m_{i+1}}, b_{i+1}: R^{m_{i+1}} \to R^{m_{i+2}}$. В данной работе из-за специфичности задачи анализа сигналов любой признак описывается не формулой (1), а ее несколько усложненным вариантом, что позволяет как сузить оптимизируемое пространство, так и использовать априорные знания о том, какие базисные функции вообще должны применяться в задаче обработки сигналов, в каком порядке они должны применяться.

- Функции инициализации – множество $\{i\}$. Это функции, с которых должен начинаться каждый признак в представлении (1). В представлении сигнала должна быть ровно одна функция инициализации. В множество функций инициализаций стоит включить те преобразования, которые в предметной области чаще всего используются для предобработки данных, это позволит напрямую использовать знания о предметной области при поиске признакового пространства. В задачах анализа сигналов часто сначала делают предварительную обработку сигналов: сглаживание или применение фильтра низких частот. Это функции $i : R^m \to R^n$.
- Функции трансформации – множество $\{t\}$. Это функции, которые отвечают за преобразования сигнала, который прошел через инициализацию. В каждом сигнале их может быть любое количество, число функций трансформаций может быть ограничено только из соображений вычислительной сложности получаемых признаков. Эти функции представляют собой по большей части нелинейные преобразования. Это функции $t : R^m \to R^n$
- Функции агрегации – множество {a}. Для того чтобы получить из сигнала число, необходимо в конце цепочки базисных функций поставить функцию, которая бы агрегировала всю полученную информацию в одно число, поэтому необходимо ввести функции агрегации. Функциями агрегации могут быть, например, среднее значение последовательности, максимальное значение последовательности и так далее. Это функции $a : R^m \to R^n$.

Таким образом, формула (1) может быть переписана в виде

$$f(x) = [i][t_1][t_2] \dots [t_{last-2}][a](x), \qquad (2)$$

где i – функция инициализации t_j – какая-то из функций трансформации, a – функция агрегации.

Задача генерации признакового пространства заключается в том, чтобы найти X признаков, представимых в форме (2), которые были бы оптимальны с точки зрения оценки качества алгоритма, обученного на этом пространстве признаков. В свою очередь это означает, что для каждого из N признаков необходимо найти функцию инициализации, последовательность функций трансформации и функцию агрегации для данного признака.

Примеры функций инициализации: тождественная; медианное сглаживание; фильтр верхних частот; фильтр нижних частот.

Примеры функций трансформации: логарифмирование; возведение в степень; конечная разность; абсолютное значение; стандартизация сигнала (вычитание среднего и деление на дисперсию).

Примеры функций агрегации: среднее значение; медиана; дисперсия; максимум, минимум. центр масс сигнала (скалярное произведение индексов на значения сигнала).

2.2 Оценка качества признака

Чтобы построить хорошее признаковое пространство, удовлетворяющее условию (2), необходимо четко определить критерий, по которому будет проходить поиск нового признака. Таким образом, нам необходимо ввести критерий качества признака. Такие критерии сильно связаны с методами фильтрации признаков, которых на данных момент известно уже немало. Похожие подходы можно использовать и в оценке качества признака.

Самый простой способ оценить качество признака – проверить, насколько статистически признак связан с целевой переменной. В этой области существует невероятно большое число исследований. Перечислим лишь несколько способов, которые в дальнейшем будем использовать для экспериментов.

- Количество неправильно ранжированных пар целевой переменной при сортировке ее по значениям данного признака. Самая простая оценка качества. Полагаем, что значения признака есть выход некоторого классификатора. Отсортируем целевую переменную по данному признаку и проверим качество этой сортировки.
- Корреляция Пирсона между целевой переменной и признаком, то есть мера линейной зависимости признака от целевой переменной. При этом разумно брать модуль, так как нам не важен знак этой линейной зависимости.
- Взаимная информация между целевой переменной и признаком, то есть величина, описывающая количество информации, содержащегося в целевой переменной относительно признака. В качестве оценки качества разумно брать нормированное на отрезок [0,1] значение.

Статистические методы обладают важным достоинством – они очень быстро считаются. Из-за этого они получили широкое распространение в задачах, где признаковое пространство состоит из огромного числа признаков, но при этом не очень важно выделить оптимальное подмножество признаков из пространства, а гораздо важнее убрать совершенно бесполезные или даже вредные признаки. Основным недостатком этих методов является недостаточная описательная способность, любой статистический критерий не способен исчерпывающе описать степень зависимости одной величины от другой, очень высок риск ошибки в оценке качества.

Существует другой обширный класс методов оценки качества признаков, который проверяет качество алгоритма, обученного на одном этом признаке. В экспериментах использовался один из самых простых алгоритмов – алгоритм *k* ближайших соседей (*k* nearest neighbors, KNN). Описать этот алгоритм довольно просто – объект относится к тому классу, к которому относится большинство из его *k* соседей, то есть *k* ближайших к нему объектов обучающей выборки, в данной работе использовалась стандартное евклидово расстояние. Оценка качества проводилась методом скользящего контроля с исключением объектов по одному (leave-one-out, LOO). Это очень популярный метод оценки качества алгоритма *k* ближайших соседей. В этом методе каждый объект по очереди исключается из обучающей выборки, для него происходит предсказание, вычисляется оценка качества, а затем это качество усредняется.

Последний метод заключается в проверке того, что признак хорошо используется алгоритмом. В экспериментах использовался алгоритм дерева решений. К тестируемому признаку прибавлялся случайный признак, затем на этих двух признаках строилось дерево решений фиксированной глубины, оценивалось, во сколько раз тестируемый признак лучше, чем случайный признак, с помощью оценки уменьшение impurity (impurity – мера качества сплита, которая вычисляется при выборе разбиения в дереве) по разбиениям дерева решений.

2.3 Схема метода

Алгоритм (1) описывает работу метода с помощью псевдокода. Метод работает так:

- Случайно взяли *k*-элементную подвыборку из множества сигналов.
- Применили к этой подвыборке случайную функцию инициализации.
- Выбрали случайный критерий качества признака.
- Установили параметры в функциях трансформации. Например, возведение в степень *p* имеет параметр *p*. В экспериментах параметры трансформации брались случайно из заранее выбранного множества, но можно использовать любой другой подход.
- Нашли новый признак методом оптимизации. Если нужно больше признаков, начали процесс сначала (с новой случайной подвыборки).

Алгоритм 1 Стохастический алгоритм синтеза признакового пространства

```
function find_features(sigs, N, k, init_funcs,
trans_funcs, agg_funcs, criteria)
  i = 0
  features = {}
  while i !=N  do:  // Ищем N признаков
    subs = random_subsample(sigs, k)
    new_init = get_random(init_funcs)
    init_subs = new_init(subs)
    new_crit = get_random(criteria)
    set_parameters(trans_funcs)
    new_feat = optimize(init_subs, new_crit, \
                        trans_funcs, agg_funcs)
    if new_feat not in features then:
      i = i + 1
      features.insert(new_feat)
  return features
```

Построение признака по случайной подвыборке решает одновременно несколько задач. Признаки будут не очень похожи друг на друга, так как они подстраивались под разные множества. Уменьшается риск построить признаковое пространство, которое работает только на определенном наборе объектов. Позволяет избавиться от проблемы несбалансированных классов, можно брать подвыборку с равным количеством объектов каждого класса. Уменьшает вычислительную сложность оценки качества признака, которая во многих методах очень большая.

Алгоритм (2) иллюстрирует работу жадного оптимизатора. Он наращивает функции трансформации жадным образом. Наращивает до тех пор, пока не превысит заранее установленный лимит, или пока качество не перестанет расти. При добавлении новой трансформации просматриваются все возможные функции агрегации.

Алгоритм 2 «Жадный» оптимизатор

```
Function optimize(init_subs, new_crit,
trans_funcs, agg_funcs)
  best_qual = -inf
  found_trans = {}
  features = {}
  while len(found_trans) != MAX_SIZE do
    for new_trans in trans_funcs do
      found_better = False
      for new_agg in agg_funcs do
        feature = create(found_trans + \
                  {new_trans}, new_agg)
        if qual(feature) > best_qual then
          found_better = True
          new_best_trans = new_trans
          best_agg = new_agg
    if not found_better then
      break
    found_trans.insert(new_best_trans)
  return found_trans, best_agg
```

3 Вычислительные эксперименты

Эксперименты проводились на сигналах, которые представляют собой электрокардиограммы пациентов. Для каждой кардиограммы известно, болен ли пациент ишемической болезнью сердца. Это классическая задача бинарной классификации, где класс 1 означает, что пациент с данной кардиограммой болен, класс -1 – здоров. Выборка

состояла из 1798 сигналов, из которых 743 сигнала принадлежало больным, а 1055 сигналов принадлежало здоровым пациентам. Таким образом, нам необходимо ввести критерий качества признака.

Оценка качества синтезированного множества признаков будет происходить с помощью измерения качества алгоритма, обученного на этих признаках. Для этого может использоваться любой классический классификатор. В качестве базового классификатора был выбран случайный лес [2]. Это ансамбль решающих деревьев. Каждое решающее дерево строится по случайным подвыборкам, полученным в результате сэмплирований с возвращениями объектов обучающей выборки.

Для оценки качества будем использовать 20-кратную кросс-валидацию. Важно отметить, что в обучающей выборке многим пациентам принадлежит сразу несколько кардиограмм, поэтому валидация проводилась таким образом, чтобы кардиограммы любого пациента не могли попасть и в обучение, и в контроль одновременно. Это более честная оценка, так как кардиограммы одного и того же пациента очень похожи, и алгоритму проще выдать правильный ответ, так как он уже ранее видел похожую кардиограмму.

Размер случайной подвыборки составлял 100 объектов (50 объектов каждого класса). Функционал качества классификации – точность предсказания по пациентам. Вычисляется он так: для каждой кардиограммы каждого пациента делается предсказания о наличие у пациента болезни, затем для каждого пациента считается процент правильно классифицированных его кардиограмм, затем все эти значения усредняются по пациентам.

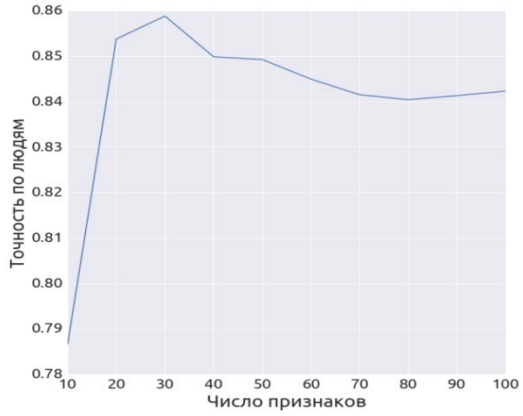

Рисунок 1 Зависимость точности по пациентам от количества признаков

Все эксперименты проводились с алгоритмом «случайный лес», который состоял из 100 деревьев. На Рис. 1 показана зависимость точности от количества признаков с шагом в 10 признаков. Видно, что оптимальное количество признаков находится около 30. Дальнейшее увеличение признакового пространства не приводит к росту качества. Это свидетельствует о том, что многие из сгенерированных признаков являются шумовыми.

Важно отметить, что алгоритм постоянно создает новые признаки, они не совпадают с уже построенными. Максимальное значение точности по пациентам – 0.859.

Все дальнейшие эксперименты проводились для множества, состоящего из 300 синтезированных признаков. На рисунках 2 и 3 показаны самые часто встречаемые функции трансформации и агрегации. Как видно из этих рисунков, среди функций трансформаций нет определенной доминирующей функции, среди функций агрегаций с большим отрывом выигрывает медиана.

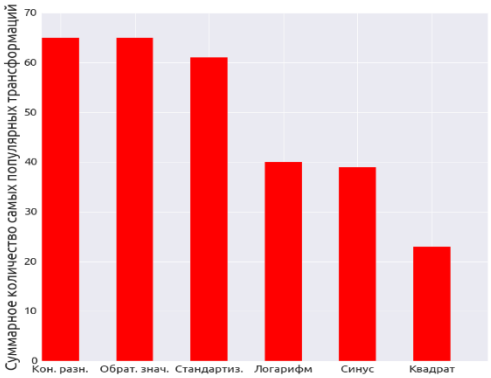

Рисунок 2 Самые популярные функции трансформации

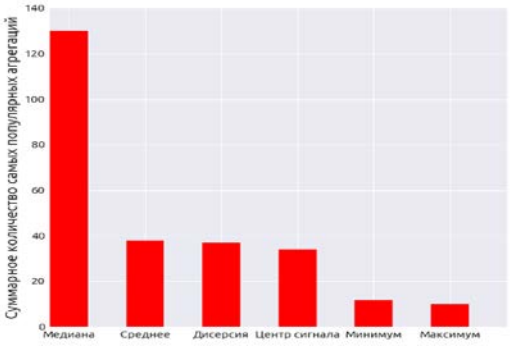

Рисунок 3 Самые популярные функции агрегации

Рисунки 4 и 5 показывают различие в поведении жадного алгоритма при различных методах оценки качества признака. Обозначения: ДНРП – доля неверно ранжированных пар (целевая переменная сортируется по значению признака), качество по ДР – качество признака по оценке дерева решений (см. раздел 2.2 про эти и другие оценки качества). Процент увеличения качества считается по формуле $\frac{FinalQual-InitQual}{InitQual}$, где $FinalQual$ – финальное качество признака, $InitQual$ – начальное качество. Начальное качество определяется качеством лучшей функцией агрегации при отсутствии функций трансформации.

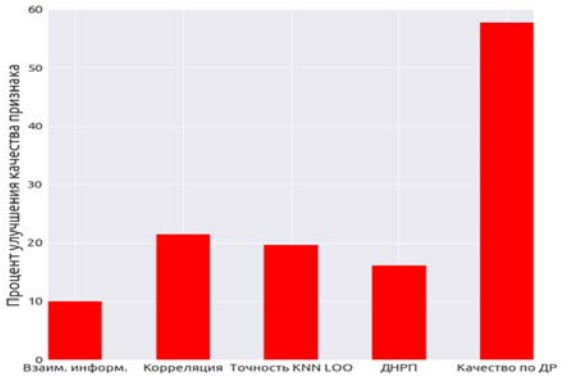

Рисунок 4 Средний процент увеличения качества признака

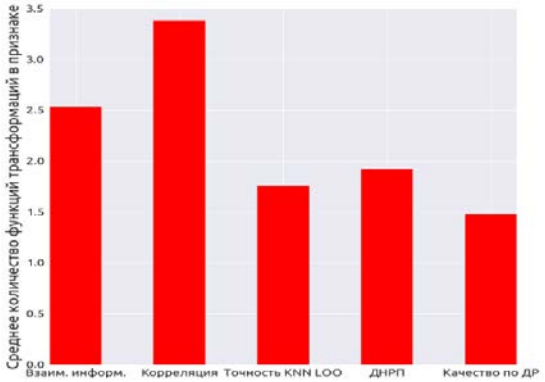

Рисунок 5 Средняя длина трансформаций в признаке

4 Выводы

Предложен алгоритм автоматического построения признакового пространства, проведены вычислительные эксперименты для задачи бинарной классификации кардиограмм. Данный алгоритм в вычислительных экспериментах показал свою способность конструировать признаковое пространство, которое позволило бы решать задачу классификации сигналов с высокой точностью. Перечислим основные достоинства данного подхода.

Алгоритм создает нужные признаки, используя только оценки качества этих признаков. Работа исследователя заключается только в выборе базисных функций, которые специфичны в его задаче. Например, исследователь может использовать фильтр, хорошо работающий конкретно для одного типа данных, но для сигналов этот фильтр не применим.

Алгоритм состоит из нескольких отдельных частей: начальный набор базисных функций, метод оценки качества признака, оптимизатор. Те варианты модулей, которые были приведены в данной работе, являются не более чем тестовыми вариантами, для каждой задачи они могут подбираться индивидуально.

Возможности алгоритма не ограничиваются его применением исключительно в задаче классификации сигналов. При изменении функций инициализации, трансформации и агрегации он может быть применен в любой дугой задаче распознавания неструктурированных данных, например, в задаче классификации текстов или изображений.

Благодаря своей стохастической природе алгоритм с каждой новой итерацией создает признак, который сильно отличается по своему методу построения от предыдущих. Чем больше итераций проведет алгоритм, тем больше вероятность, что среди полученных признаков будет подмножество действительно качественных.

Более подробное описание проблемы выделения признаков в задаче классификации сигналов можно найти в [9]. Настоящая работа содержит наиболее важные результаты вышеупомянутой.

Литература

[1] Al-Sahaf, H., Neshatian, K., Zhang, M.: Automatic Feature Extraction and Image Classification using Genetic Programming. In The 5th Int. Conf. on Automation, Robotics and Applications, pp. 157-162 (2011)

[2] Breiman, L.: Random Forests. Machine Learning (2001)

[3] Dal Seno, B., Matteucci, M., Mainardi, L.: A Genetic Algorithm for Automatic Feature Extraction in p300 Detection. 2008 IEEE Int. Joint Conf. on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 3145-3152 (2008)

[4] Guo, L., Rivero, D., Dorado, J., Munteanu, C., Pazos, A.: Automatic Feature Extraction using Genetic Programming: An Application to Epileptic Eeg Classification. Experts Systems with Applications: An Int. J. (2011)

[5] Kohler, B., Hennig, C., Orglmeister, R.: The Principles of Software qrs Detection. IEEE Engineering in Medicine and Biology Magazine (2002)

[6] Lallo, P. R. U.: Signal Classification by Discrete Fourier Transform. MILCOM 1999. IEEE Military Communications, pp. 197-201 (1999)

[7] Morik, K., Mierswa, K.: Automatic Feature Extraction for Classifying Audio Data. Machine Learning (2005)

[8] Prochazka, A., Kukal, J., Vysata, O.: Wavelet Transform use for Feature Extraction and Eeg Signal Segments Classification. 2008 3rd Int. Symposium on Communications, Control and Signal Processing, pp. 719-722 (2008)

[9] Викулин, В. А.: Автоматическое выделение признаков в задаче классификации сигналов. ВМК, МГУ имени М.В. Ломоносова (2017). http://www.machinelearning.ru/wiki/images/3/37/CourseVikulin.pdf