НАХОЖДЕНИЕ ОПТИМАЛЬНЫХ НАБОРОВ ПРИЗНАКОВ В ЗАДАЧАХ КЛАССИФИКАЦИИ ВОЗДЕЙСТВИЙ НА ВИБРАЦИОННЫХ ДАТЧИКАХ

DETERMINATION OF OPTIMUM FEATURE SETS FOR VIBRATION-BASED SENSOR EVENTS CLASSIFICATION

D. Chickrin
S. Golousov
N. Glavatskiy
D. Ermakov
A. Stepanov
P. Kokunin

Summary. The paper presents comparative analysis of machine learning feature extraction methods in relation to the problem of three-class classification using the experimental sample. Input data sampling represents a uniformly discretized sequence of normalized amplitudes of actions received from a vibration sensor. The k-nearest neighbors algorithm is used for the classification method. As a result of the investigation, the optimum feature set and optimum metrics of distance from the viewpoint of minimization of an error of classification are determined; based on the considered features, the level of their positive or negative impact on the classification process is found.

Keywords: classifcation, machine learning, vibrations, k-nearest neighbors algorithm, wavelet, cepstrum, kurtosis, skewness.

Чикрин Дмитрий Евгеньевич
К.т.н., доцент, ФГАОУ ВО «Казанский (Приволжский) Федеральный Университет»
dmitry.kfu@gmail.com
Голоусов Святослав Владимирович
Аспирант, ФГАОУ ВО «Казанский (Приволжский) Федеральный Университет»
sgolousov@gmail.com
Главацкий Никита Владимирович
ФГАОУ ВО «Казанский (Приволжский) Федеральный Университет»
hanouchh@gmail.com
Ермаков Дмитрий Владимирович
ФГАОУ ВО «Казанский (Приволжский) Федеральный Университет»
aginum0@gmail.com
Степанов Андрей Николаевич
Аспирант, ФГАОУ ВО «Казанский (Приволжский) Федеральный Университет»
pk-kzsol@mail.ru
Кокунин Петр Анатольевич
К.т.н., доцент, ФГАОУ ВО «Казанский (Приволжский) Федеральный Университет»
PAKokunin@kpfu.ru

Аннотация. В работе представлен сравнительный анализ методов извлечения признаков машинного обучения применительно к задаче трехклассовой классификации по экспериментальной выборке. Входная выборка данных представляла собой равномерно дискретизированную последовательность нормированных амплитуд воздействий, получаемых с вибрационного датчика. В качестве метода классификации был использован метод ближайших соседей. В результате исследований был найден оптимальный набор признаков и оптимальная метрика расстояния с точки зрения минимизации ошибки классификации; по рассматриваемым признакам определена степень их позитивного или негативного влияния на процесс классификации.

Ключевые слова: классификация, машинное обучение, вибрации, метод ближайших соседей, вейвлет, спектральное разложение, кепстр, эксцесс, асимметрия, метрика городских кварталов.

Введение

Задача детектирования и классификации в последнее время активно решается исследователями по всему миру применительно к разнообразным системам, как гражданского назначения [1, 2] так и государственного [3]. Множество научных работ посвящено исследованию способов классификации и методов извлечения признаков.

Для успешности реализации алгоритма классификации необходимо выбрать признаки, с помощью которых объекты разных классов будут хорошо различимы. На сегодняшний день существует множество методов извлечения признаков, каждый из которых зарекомендовал себя с положительной стороны. Нас будут интересовать признаки, которые могут быть извлечены из временных рядов, полученных путем дискретизации сигнала вибрационного датчика. В данной работе рас-

сматривается лишь несколько из них, среди которых будет найден набор признаков минимизирующий ошибку классификации.

Самым простым видом признаков является сам временной ряд [3, 4], однако часто по виду временного ряда как такового бывает трудно или невозможно построить надежный алгоритм классификации [5]. В этом случае эффективными могут оказаться статистические признаки, такие как математическое ожидание [6], дисперсия [7] и другие статистические признаки более высокого порядка [8–10].

Другим важным источником признаков является частотное представление сигнала, которое в некоторых случаях достаточно хорошо характеризует объект. Также существуют такие преобразования временного ряда, как вейвлет преобразование и кепстральное разложение, которые являются богатыми источниками для генерации признаков. Так, вейвлет коэффициенты широко используются для классификации целей при использовании сейсмического датчика [11, 12], в то время как кепстральные коэффициенты хорошо себя зарекомендовали в задачах распознавания речи [13, 14].

В качестве метода классификации мы будем использовать метод ближайших соседей и проанализируем влияние его различных параметров на качество классификации.

Во второй части формулируется задача, объясняются признаки и метод классификации. В третьей части описывается сам эксперимент. В четвертой части обсуждаются результаты эксперимента. В последней части подводятся итоги и делаются выводы.

1. Постановка задачи

В данной работе решается задача анализа и классификации воздействий, выраженных в виде временных рядов $x(n)$, поступающих с чувствительного элемента — вибрационного датчика.

Стоит отметить, что обычно при практической реализации данные поступают блоками по несколько отсчётов (N), поэтому мы будет разделять сигнал на независимые блоки для анализа.

Каждый блок данных будем рассматривать как отдельный эксперимент. Множество таких экспериментов и будет выборкой, на основании которой будем решать задачу классификации.

Рассматриваемые признаки для анализа. Методы извлечения признаков играют важную роль для задачи классификации, поскольку состоятельность вектора признаков напрямую влияет на качество классификации.

Мы будем использовать несколько признаков разного рода.

1.1. Временные признаки. В первую очередь рассмотрим признаки, извлекаемые напрямую из временных рядов $x(n)$, где $1 \le n \le N$.

1.1.1. Математическое ожидание. Первым статистическим моментом случайной величины является её математическое ожидание:

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x(n); \tag{1.1}$$

Математическое ожидание представляет собой оценку среднего значения случайной величины [6].

1.1.2. Дисперсия. Вторым статистическим моментом случайной величины является её дисперсия:

$$\sigma = \frac{1}{N}\sum_{n=1}^{N}\left(x(n)-\mu\right)^2; \tag{1.2}$$

Дисперсия показывает насколько отклоняются реализации случайной величины от её математического ожидания [7].

1.1.3. Коэффициент асимметрии. Третьим статистическим моментом случайной величины является её коэффициент асимметрии:

$$\gamma = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{x(n)-\mu}{\sigma}\right)^3; \tag{1.3}$$

Коэффициент асимметрии показывает насколько симметрично распределение случайной величины. Положительная величина γ соответствует большим хвостам справа от μ, отрицательная — слева от μ и равна нулю при симметричном распределении [15].

1.1.4. Коэффициент эксцесса. Четвёртым статистическим моментом случайной величины является её коэффициент эксцесса [16]:

$$\beta = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{x(n)-\mu}{\sigma}\right)^4 - 3; \tag{1.4}$$

Коэффициент эксцесса показывает меру остроты пика случайной величины.

1.1.5. Интегральные признаки. Интегральные признаки сигнала могут быть использованы для клас-

сификации сигналов [17]. Так в задаче детектирования высокая энергия сигнала обычно прямо свидетельствует об обнаружении цели. Энергия сигнала вычисляется по формуле:

$$E_{signal} = \sum_{i=1}^{N} x^2(i) \quad (1.5)$$

1.2. Частотные и частотно-временные признаки. Во-вторых, рассмотрим признаки извлекаемые из спектра сигнала

$$X(\omega) = F(x(t), \omega)$$

с помощью дискретного преобразования Фурье F. Аналогично временным рядам из спектра сигнала могут быть извлечены статистические моменты μ_s, σ_s, γ_s, β_s с использованием частотных рядов.

1.2.1. Shape statistics. Дополнительно к этому будем использовать shape statistics, которые представляют собой статистики более высокого порядка [8]:

$$\mu_{s,shape} = \frac{1}{S}\sum_{i=1}^{N} iX(i); \quad (1.6)$$

$$\sigma_{s,shape} = \sqrt{\frac{1}{S}\sum_{i=1}^{N}\left(i - \mu_{s,shape}\right)^2 X(i)}; \quad (1.7)$$

$$\gamma_{s,shape} = \frac{1}{S}\sum_{i=1}^{N}\left(\frac{i - \mu_{s,shape}}{\sigma_{s,shape}}\right)^3 X(i); \quad (1.8)$$

$$\beta_{s,shape} = \frac{1}{S}\sum_{n=1}^{N}\left(\frac{i - \mu_{s,shape}}{\sigma_{s,shape}}\right)^4 X(i) - 3; \quad (1.9)$$

где

$$S = \sum_{n=1}^{N} X(i).$$

1.2.2. Вейвлет анализ. Вейвлет преобразование широко используется для анализа сигналов [11, 12]. С помощью специальных функций $\Psi(n)$ исходный дискретный сигнал может быть преобразован к представлению:

$$y(a,b) = \frac{1}{\sqrt{a}}\sum_{n=-\infty}^{+\infty}\Psi\left(\frac{n-b}{a}\right)x(n) \quad (1.10)$$

где параметры a, b называются параметром растяжения (масштаба) и параметром положения соответственно.

Функции $\Psi(n)$ называются материнскими вейвлетами. Они представляют собой полосовые фильтры, выбор параметра a позволяет получить нужную полосу пропускания. Обычно используют банк фильтров:

$$\Psi_{jk}(n) = \frac{1}{\sqrt{2^j}}\Psi\left(\frac{n - k2^j}{2^j}\right); \quad j,k \in \mathbb{Z} \quad (1.11)$$

В результате преобразования полученные коэффициенты $y(a, b)$ с одинаковым a ($a = 2^j$) рассматривают вместе и называют детальными коэффициентами j-го порядка $y_j(b)$. В качестве характеристики сигнала можно использовать сами коэффициенты вейвлет преобразования или их статистические моменты, например, $\mu(y_1)$, $\mu(y_2)$, …, $\mu(y_K)$, ограничившись K-м порядком детальных коэффициентов.

1.3. Кепстральные признаки. В этой секции рассмотрим кепстр как источник еще одного класса признаков, поскольку он широко используется в задачах классификации акустических сигналов [14]. Обычно для подсчета кепстра применяется следующая формула:

$$X_C(q) = \Re(F^{-1}(log\,|X(\omega)|^2)) \quad (1.12)$$

где R(z) означает действительную часть комплексного числа z.

Не все кепстральные коэффициенты одинаково значимы для классификации, а лишь несколько первых. Так в задачах автоматического распознавания речи исторически принято использовать первые 12 коэффициентов, поскольку остальные коэффициенты вносят незначительную роль в качество распознавания.

2. Классификация. Задача классификации обычно формулируется следующим образом. Имеется множество объектов O и множество классов C. Каждому объекту $o \in O$ ставится в соответствие определенный класс $c_o \in C$. Для конечного подмножества объектов обычно называемого обучающей выборкой известно соответствие классам. Требуется построить алгоритм, который сможет классифицировать любой объект из исходного множества.

Для сравнения объектов используются их характеристики или признаки, таким образом задача классификации объектов сводится к задаче классификации набора признаков присущих этим объектам. Часто признаки представляют собой числовые величины и могут быть количественно сравнимы. Числовые признаки принято

объединять в векторы признаков и производить классификацию таких векторов.

2.1. Метод ближайших соседей. Данный метод является метрическим классификатором, основанным на сходстве объектов. При классификации объект относят к классу ближайших к нему классифицированных объектов.

Широко используется взвешенный метод k ближайших соседей [1, 2, 4, 8, 13]. Для неопознанного $x \in O$ находят k его ближайших соседей $x_1, \ldots, x_k$ в порядке удаления от x, и расстояния до них $d_1, \ldots, d_k$ в возрастающем порядке. Тогда i-му объекту назначается вес w_i:

$$w_i = \begin{cases} \dfrac{d_k - d_i}{d_k - d_1}, & \text{если } d_k \neq d_1 \\ 1, & \text{иначе.} \end{cases} \quad (1.13)$$

Класс набравший наибольший вес присваивается объекту x.

2.2. Метод опорных векторов. Данный метод относится к линейным классификаторам, основанным на построении линейной разделяющей поверхности [14].

Задача для двух классов ставится следующим образом. Каждый объект представляется вектором в p-мерном пространстве и классом принадлежности. Требуется построить гиперплоскость размерности $(p - 1)$ разделяющие объекты разных классов. Гиперплоскостей построенных таким образом может быть больше одной, в связи с чем естественно рассматривать такую гиперплоскость, минимальное расстояние до которой будет максимально среди всех таких гиперплоскостей.

$$\begin{cases} \|w\|^2 \to min \\ c_i(\vec{w} \cdot \vec{x_i} - b) \geq 1, \quad 1 \leq i \leq n, \end{cases} \quad (1.14)$$

где

$\vec{w}$ — вектор нормали к разделяющей гиперплоскости, b — модуль расстояния от начала координат до разделяющей гиперплоскости,

$c_i \in \{-1, 1\}$ — класс x_i.

Для обобщения задачи на случай многих классов на практике обычно используют технику «один против остальных».

2.3. Метод главных компонент. Одним из основных способов уменьшения размерности пространства признаков с минимальными потерями информации является метод главных компонент [12, 15]. Для случая двух признаков x_1 и x_2 требуется найти единственную прямую, которая эффективно соответствует обоим признакам. Затем мы заменяем x_1 и x_2 на единственный признак x. Случай двух признаков может быть обобщен на случай n признаков.

Цель метода главных компонент минимизировать среднее расстояние по всем признакам до прямой. Алгоритм:

1. Подготовка данных. Осуществляется масштабирование и нормализация данных.
2. Вычисляется матрица ковариаций Σ.
3. Осуществляется сингулярное разложение $\Sigma = USV^*$.
4. Формируется матрица проецирования U взяв первые k столбцов матрицы U
5. Осуществляется проецирование n-мерного пространства признаков X на новое k-мерное пространство $Z = XU$

Обычно количество главных компонент k выбирают таким образом, чтобы ошибка проецирования составляла не более 1% от суммарной вариации исходных данных.

2.4. Определение расстояния. Особняком стоит вопрос выбора метрики для определения расстояния между объектами. Так как объекты сравниваются на основании многомерных векторов особенностей в качестве расстояния разумно использовать обыкновенное евклидово расстояние [13, 14, 15]:

$$d(p,q) = \sqrt{\sum_i (x_p(i) - x_q(i))^2} \quad (1.15)$$

Дополнительно к евклидовой метрике рассмотрим метрику городских кварталов и метрику Чебышева [18]:

$$d(p,q) = \sum_i |x_p(i) - x_q(i)| \quad (1.16)$$

$$d(p,q) = \max_i |x_p(i) - x_q(i)| \quad (1.17)$$

3. Эксперимент

В качестве источника данных будем использовать вибрационный датчик, который в равномерном дискретном времени измеряет одномерную характеристику воздействия на систему. Датчик производит измерения с частотой 2048Гц. Для анализа и классификации воздействий будем использовать кадр данных длинной 1с, что соответствует 2048 отсчетам.

Обучающая выборка состоит из 200 кадров, каждый из которых относится к одному из трёх классов: отсут-

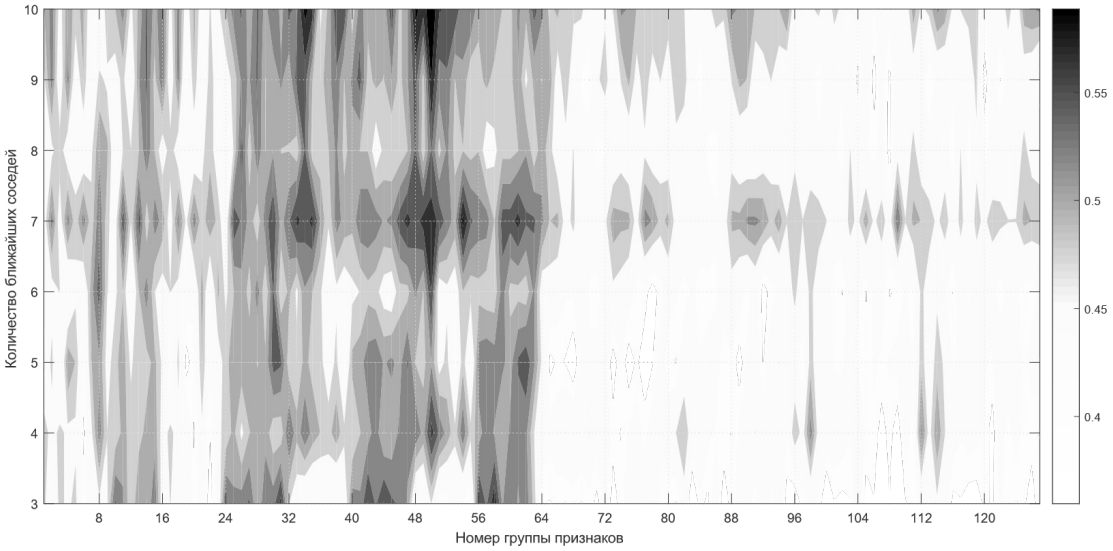

Рис. 1. Влияние выбора признаков и количества ближайших соседей на уровень потерь классификации представленный на рисунке цветовым градиентом.

ствие воздействия (60 кадров), импульсное воздействие (70 кадров) и распределённое воздействие (70 кадров). В качестве критерия оценки качества классификации будем использовать уровень потерь по методике «один против остальных»:

$$CL = \frac{Количество\ неправильно\ классифицированных\ кадров}{Общее\ количество\ кадров} \quad (1.18)$$

3.1. Выявление оптимального набора признаков. Все рассмотренные признаки, извлекаемые из временных рядов, нашли свое применение в разнообразных предметных областях, в которых требуется решать задачу классификации объектов. В данной работе предлагается найти подмножество признаков минимизирующее CL. Для этого рассмотрим все возможные подмножества признаков и для каждого из них вычислим CL.

3.2. Влияние параметров метода классификации. От выбора количества k ближайших соседей и метрики близости также зависит качество классификации. В данной работе предлагается найти оптимальные параметры, которые минимизируют CL. Количество k будем варьировать от 3 до 10. Метрики расстояния были рассмотрены выше: метрика Чебышева, метрика городских кварталов, евклидова метрика.

4. Результаты

Для поиска оптимального набора характеристик сигнала для решения задачи классификации было сформировано 7 подгрупп признаков[2]:

1. статистические моменты временного ряда (μ, σ, γ, β)
2. энергия временного ряда (E_{signal})
3. статистические моменты спектра (μ_s, σ_s, γ_s, β_s)
4. shape statistics (μs, shape, σs, shape, γs, shape, βs, shape)
5. энергия спектра ($E_{s,\ signal}$)
6. вейвлет коэффициенты ($\mu_{w,1}$, $\mu_{w,2}$, $\mu_{w,3}$, $\mu_{w,4}$, $\mu_{w,5}$)
7. кепстральные коэффициенты (C_1, C_2, …, C_{12})

Объединяя эти подгруппы в группы полным перебором, было составлено 127 итоговых групп признаков. Для каждой из групп был запущен алгоритм классификации, основанный на методе ближайших соседей. На Рис. 1 представлены изолинии уровня потерь классификации (CL) в зависимости от номера группы признаков и количества ближайших соседей.

Было установлено, что наличие кепстральных коэффициентов в наборе заметно улучшает классификацию: первая половина групп (1–63) не включает кепстральные коэффициенты, в то время как вторая (64–127) включает, что сопровождается уменьшением CL в среднем на 7%.

Наоборот, подгруппа признаков shape statistics негативно влияет на уровень потерь. Особенно хорошо это заметно в первой половине Рис. 3[3].

Аналогично, присутствие признаков, основанных на вейвлет разложении в среднем ухудшает уровень классификации.

[2] подробно описаны в разделе 1

[3] Группы признаков содержащие подгруппу shape statistics имеют номера: 8–15, 24–31, 40–47, 56–63, 72–79, 88–95, 104–111, 120–127.

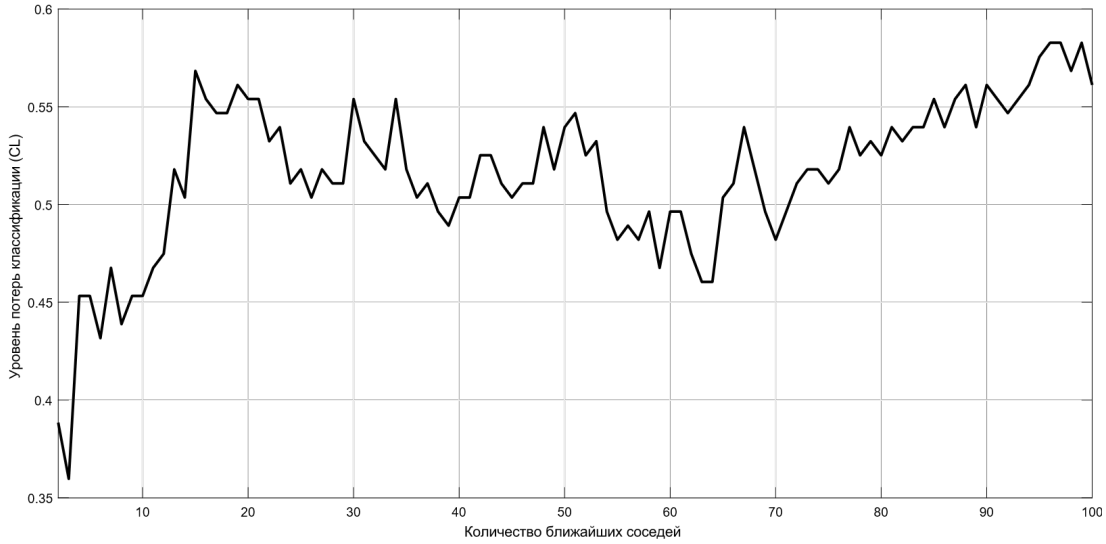

Рис. 2. Зависимость CL от количества ближайших соседей.

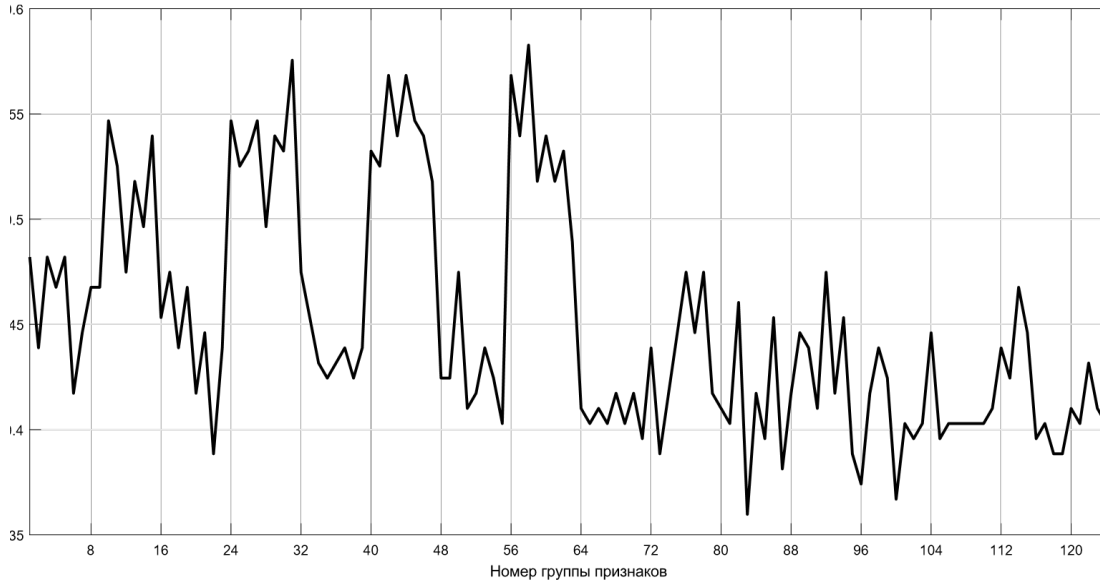

Рис. 3. Влияние выбора признаков на уровень потерь классификации при k = 3.

Минимальный уровень потерь 35,97% достигается на группе признаков № 83, состоящей из статистических моментов временного ряда, его энергии, энергии спектра и кепстральных коэффициентов. Этот набор признаков является оптимальным для данной задачи.

При варьировании количества ближайших соседей в пределах от 2 до 100 качество классификации значительно меняется. На Рис. 2 представлена зависимость CL от k. Оптимальное значение k равно 3, причем при его увеличении заметен большой рост уровня потерь классификации. На Рис. 3 представлен детальный график зависимости уровня потерь классификации от выбора группы признаков для k равного 3.

При использовании метода главных компонент в среднем по всем наборам качество классификации улучшается на 5%, но минимальный уровень потерь остается на прежнем уровне 35,97%. При такой конфигурации при k равном 5 наблюдается сразу несколько оптимальных групп признаков: 76, 79, 93.

Наилучшие результаты достигаются при использовании метрики городских кварталов. При применении

других рассмотренных метрик уровень потерь классификации возрастает в среднем на 4%.

Метод опорных векторов не дает существенного результата. Скорее всего это вызвано линейной неразделимостью выборки.

Заключение

В результате исследования было установлено следующее. При решении задачи классификации методом ближайших соседей выборки, полученной из сигнала вибрационного датчика, наилучшим признаком является признак на основе кепстральных коэффициентов. Применение признаков на основе вейвлет преобразования и shape statistics снижает качество классификации.

Применение метода главных компонент существенным образом не улучшает классификацию.

Среди рассмотренных метрик наилучшие результаты достигаются при использовании метрики городских кварталов.

Группа признаков, при которой достигается минимальный уровень потерь классификации 35,97%, состоит из моментов временного ряда, его энергии, энергии спектра, а также кепстральных коэффициентов.

ЛИТЕРАТУРА

1. Bashir S., Doolan D., Petrovski A.— ClusterNN: A Hybrid Classification Approach to Mobile Activity Recognition // Proceedings of the 13th International Conference on Advances in Mobile Computing and Multimedia. MoMM 2015. New York, NY, USA: ACM, 2015. С. 263–267.
2. Jayasree T., Bobby M., Muttan S.— Sensor Data Classification for Renal Dysfunction Patients Using Support Vector Machine // Journal of Medical and Biological Engineering. 2015. Т. 35, № NOVEMBER. С. 759–764.
3. Bhattacharyya A., Saraswat V. K., Manimaran P. [и др.] — Evidence theoretic classification of ballistic missiles // Applied Soft Computing Journal. 2015. Т. 37. С. 479–489.
4. Seth N., Johnson D., Taylor G. [и др.] — Robotic pilot study for analysing spasticity: clinical data versus healthy controls. // Journal of neuroengineering and rehabilitation. 2015. Т. 12. с. 109.
5. Xi X., Eamonn K., Shelton C. [и др.] — Fast time series classification using numerosity reduction // Proceedings of the 23rd international conference on Machine learning (ICML). 2006. С. 1033–1040.
6. Kumar M. M., Puhan N. B.— Off-line signature verification: upper and lower envelope shape analysis using chord moments. // 2014. № April. С. 347–354.
7. Jing Y., Meng Q., Qi Q. [и др.] — A novel olfactory neural network for classification of Chinese liquors using electronic nose // 2015 IEEE SENSORS — Proceedings. 2015. С. 0–3.
8. Tian Y., Qi H., Wang X.— Target detection and classification using seismic signal processing in unattended ground sensor systems. // Т. 4. 2002. с. IV/4172.
9. Bump W.M.— The Normal Curve Takes Many Forms: A Review of Skewness and Kurtosis. 1991.
10. Liang Z., Wei J., Zhao J. [и др.] — The statistical meaning of kurtosis and its new application to identification of persons based on seismic signals // Sensors. 2008. Т. 8, № 8. С. 5106–5119.
11. Averbuch A., Hulata E., Zheludev V. [и др.] — A wavelet packet algorithm for classification and detection of moving vehicles // Multidimensional Systems and Signal Processing. 2001. Т. 12, № 1. С. 9–31.
12. Liu C.—Classification Fusion in Wireless Sensor Networks. 2006. Т. 32, № 6. С. 1–9.
13. Bano S, Ravi Kumar K.— Decoding Baby Talk: Basic Approach for Normal Classification of Infant Cry Signal // International Journal of Computer Applications. 2015. С. 24–26.
14. Barkana B.D., Zhou J.— A new pitch-range based feature set for a speaker's age and gender classification // Applied Acoustics. 2015. Т. 98. С. 52–61.
15. Saha A., Konar A.— Data-point and Feature Selection of Motor Imagery EEG Signals for Neural Classification of Cognitive Tasks in Car-Driving. // 2015.
16. Айзикович А.А., Корякин А. В.— Методы распознавания типа нарушителя, применяемые в сейсмоакустических периметровых охранных системах // Интеллектуальные системы в производстве. 2013. Т. 1 (21). С. 5–8.
17. Yang B., Lei Y.— Vehicle detection and classification for low-speed congested traffic with anisotropic magnetoresistive sensor // IEEE Sensors Journal. 2015. Т. 15, № 2. С. 1132–1138.
18. Рузибаев О.Б., Эшметов С. Дж. Исследование и анализ алгоритмов на основе нечеткого метода k ближайших соседей с применением различных метрик при диагностике рака молочной железы // Наука и Мир. 2016. Т. 5 (33). С. 102–107.

© Чикрин Дмитрий Евгеньевич (dmitry.kfu@gmail.com), Голоусов Святослав Владимирович (sgolousov@gmail.com),
Главацкий Никита Владимирович (hanouchh@gmail.com), Ермаков Дмитрий Владимирович (aginum0@gmail.com),
Степанов Андрей Николаевич (pk-kzsol@mail.ru), Кокунин Петр Анатольевич (PAKokunin@kpfu.ru).
Журнал «Современная наука: актуальные проблемы теории и практики»