



СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ



А.Е. Сулавко

Машинное обучение в приложениях биометрии

**Статистический подход к анализу и
классификации биометрических
образов**

Конспект лекции

СПбГЭТУ «ЛЭТИ», 2022 г.





СТАТИСТИЧЕСКИЙ ПОДХОД К АНАЛИЗУ И КЛАССИФИКАЦИИ БИОМЕТРИЧЕСКИХ ОБРАЗОВ

В предыдущем разделе был отмечен основной подход к классификации биометрических данных, а именно их деление на две больших группы: статические и динамические. Однако говоря о подобной классификации, более корректным является подход, при котором на две упомянутые группы делятся так называемые **биометрические образы**. Биометрическим образом (примером образа) принято называть биометрические данные человека, подвергающиеся в дальнейшем масштабированию, удалению шумов и другой обработке в целях вычисления вектора биометрических параметров (признаков)[9]. Именно биометрический образ непосредственно участвует в процессе идентификации/верификации пользователя, и именно признаки, извлекаемые из образа, напрямую задействованы в работе алгоритма классификации на два потенциально возможных класса: «Свой»/«Чужой».

Одним из наиболее важных параметров биометрического образа можно считать **показатель уникальности биометрического образа** (характеристика i -го биометрического параметра, отражающая отличие контролируемого параметра от среднестатистического значения этого параметра, характерного для всех пользователей). С дополнительными характеристиками биометрических образов и их определениями можно ознакомиться в серии стандартов ГОСТ Р 52633.

Итак, говоря о биометрических образах и работе с ними, важно отметить следующий (после выбора самого образа) этап - а именно этап извлечения признаков. Согласно определению, представленном в ГОСТ ISO/IEC 19794-1-2015, **извлечением биометрических признаков** называется процесс, применяемый к биометрическому образцу с целью изолирования и вывода повторяющейся и отличительной информации, по которой может проводиться сравнение с другими биометрическими образцами. Тогда признаком a_j назовем конкретную величину, характеризующую образы (классы) и имеющую определенный физический смысл. В свою очередь набор признаков, описывающих биометрический образ, принято называть пространством признаков.





Как правило, большинство биометрических признаков имеют законы распределения, близкие к трем следующим:

нормальное распределение $p(a_j, \mu_j, \sigma_j) = \frac{1}{\sigma_j \sqrt{2\pi}} e^{-\frac{(a_j - \mu_j)^2}{2\sigma_j^2}}$,

где μ_j и σ_j - математическое ожидание и среднеквадратичное отклонение j -го признака;

логнормальное распределение $p(a_j, \mu_j, \sigma_j) = \frac{1}{a_j \sigma_j \sqrt{2\pi}} e^{-\frac{(\ln(a_j) - \mu_j)^2}{2\sigma_j^2}}$,

где μ_j и σ_j - параметр масштаба и формы;

закон распределения Лапласа $p(a_j, \mu_j, \sigma_j) = \frac{\sigma_j}{2} e^{-\sigma_j |a_j - \mu_j|}$,

где μ_j и σ_j - коэффициенты сдвига и масштаба.

Существуют два наиболее значимых параметра, отличающих биометрические признаки друг от друга независимо от области применения и задачи распознавания: информативность и взаимная корреляционная зависимость. Под информативностью признака подразумевается количество информации о распознаваемых образах, которое он несет в себе lbit (то есть данный параметр характеризует то, насколько хорошо биометрический признак подчеркивает различия между распознаваемыми объектами). Для каждого отдельного образа информативность признака может отличаться. Суммарной информативностью признаков называют взвешенную сумму частных оценок информативности признаков с учетом их взаимной корреляции

Среднюю информативность N признаков применительно к задаче верификации образов можно оценить по формуле

$$I(N) = \frac{\sum_{j=1}^N (I_j = \frac{\sum_{i=1}^n Scч_i(j)}{n})}{N}$$





Для задачи идентификации можно воспользоваться формулой

$$I(N) = \frac{\sum_{j=1}^N (I_j = 2 \frac{\sum_{i=2}^n \sum_{l=1}^{i-1} S_{cil}(j)}{n^2 - n})}{N}$$

где $S_{ci}(j)$ – площадь пересечения двух функций плотности вероятности (ФПВ) значений j -го признака, характеризующих, соответственно, образ «Свой» и образ «Чужие» для класса образов под номером i ;

$S_{cil}(j)$ – площадь пересечения двух ФПВ значений j -го признака, характеризующих образ «Свой» для классов образов под номерами i и l ;

n – число классов образов, доступных для исследования (например, количество испытуемых, для которых проводится оценка информативности определенных биометрических признаков). Чем ниже значение I , тем признак (и) информативнее. При необходимости эту оценку можно обратить ($\bar{I} = 1 - I$) или перевести в биты информации: $I_{bit} = -\log_2(I)$.

Давать качественную оценку силе корреляционной зависимости между признаками (или любыми другими случайными величинами) удобно с использованием шкалы Чеддока:

- $|r_{ji}| < 0,1$ – отсутствие связи между признаками a_j и a_i ;
- $0,1 < |r_{ji}| < 0,3$ – слабая связь между признаками a_j и a_i ;
- $0,3 < |r_{ji}| < 0,5$ – умеренная связь между признаками a_j и a_i ;
- $0,5 < |r_{ji}| < 0,7$ – заметная связь между признаками a_j и a_i ;
- $0,7 < |r_{ji}| < 0,9$ – высокая связь между признаками a_j и a_i ;
- $0,9 < |r_{ji}| < 0,99$ – весьма высокая связь между признаками a_j и a_i ;
- при $r_{ji} = 1$ – связь между признаками a_j и a_i функциональная, т. е. $a_j = f(a_i)$ или $a_i = f(a_j)$.





Введем аналогичную градацию информативности биометрических признаков:

$I \leq 0,001$	$9.966 \leq I_{bit}$	Сверхинформативные	$0,999 \leq \bar{I}$
$0,001 < I \leq 0,1$	$3.322 < I_{bit} \leq 9.966$	Высокоинформативные	$0,9 \leq \bar{I} < 0,999$
$0,1 < I \leq 0,3$	$1.737 < I_{bit} \leq 3.322$	Весьма информативные	$0,7 \leq \bar{I} < 0,9$
$0,3 < I \leq 0,5$	$1 < I_{bit} \leq 1.737$	Средней информативности	$0,5 \leq \bar{I} < 0,7$
$0,5 < I \leq 0,7$	$0.515 < I_{bit} \leq 1$	Малоинформативные	$0,3 \leq \bar{I} < 0,5$
$0,7 < I \leq 0,9$	$0.152 < I_{bit} \leq 0.515$	Крайне малоинформативные	$0,1 \leq \bar{I} < 0,3$
$0,9 < I \leq 0,999$	$0.001 < I_{bit} \leq 0.152$	Почти не информативные	$0,001 \leq \bar{I} < 0,1$
$0,999 < I$	$I_{bit} < 0.001$	Неинформативные	$\bar{I} < 0,001$

Рис. 1. Шкала информативности признаков

Информативность биометрических признаков является важным параметром еще и с точки зрения возможности подготовки данных как собственно к процессу извлечения признаков, так и, например, для подачи на вход нейронной сети (в случае нейросетевой классификации). Динамические биометрические образы по большей части (72,5%) состоят из малоинформативных признаков, также они включают существенную часть признаков средней информативности и незначительное количество весьма информативных, а также крайне малоинформативных признаков. Для сравнения можно видеть, что образ лица человека (один из самых малоинформативных статических биометрических образов) содержит гораздо больше информации, так как имеет высокоинформативные признаки и существенную часть весьма информативных параметров. Наиболее информативные статические образы (отпечаток пальца, радужка, трехмерные модели лица и черепа) состоят преимущественно из высокоинформативных признаков [10].

Информативность отдельно взятого признака можно считать аналогом его уникальности. Но на общее количество информации образа влияют корреляционные связи признаков[11]. Часть информации «переходит» в матрицу коэффициентов корреляции между признаками. В зависимости от функционала, лежащего в основе классификатора биометрической системы, наличие корреляционных связей между признаками может как способствовать улучшению процесса классификации, так и ухудшать его.





Поэтому при построении классификаторов важно учитывать корреляционные связи между признаками с целью повышения качества их работы.

Во многом, теория об учете корреляционных связей между признаками построена на байесовском подходе, а сам процесс идентификации/верификации биометрических образов, в таком случае, осуществляется по средствам байесовского классификатора. Под байесовским классификатором понимают широкий класс алгоритмов классификации, основанный на принципе максимума апостериорной вероятности. Для классифицируемого объекта вычисляются функции правдоподобия каждого из классов, по ним вычисляются апостериорные вероятности классов. Объект относится к тому классу, для которого апостериорная вероятность максимальна[12].

Частным случаем байесовского классификатора, применяемого в задачах верификации/идентификации биометрических образов, является так называемый «наивный» байесовский классификатор. Такой классификатор основывается на дополнительном предположении о том, биометрические образы описываются статистически независимыми признаками (что на практике чаще всего оказывается не так).

Начальным этапом построения, например, системы идентификации биометрических образов, основанной на «наивном» байесовском классификаторе, является формирование плотностей апостериорных вероятностей гипотез $P(A_j | H_i)$, которые соответствуют каждому признаку из A_j и определяются по сформированным в процессе обучения распределениям признаков. Последовательным применением формулы гипотез Байеса вычисляются интегральные (финальные) вероятности гипотез:

$$P(H_i | A_j) = \frac{P(H_i | A_{j-1})P(A_j | H_i)}{\sum_{i=1}^n P(H_i | A_{j-1})P(A_j | H_i)}$$

На каждом шаге в качестве априорной вероятности используется апостериорная вероятность гипотезы, вычисленная на предыдущем шаге. На первом шаге все гипотезы считаются равновероятными, т. е. $P(H_i | A_0) = n^{-1}$, где n — количество гипотез. При $j = N$ (N — количество признаков) по





максимальному значению $P(H_i | A_N)$ принимается решение о принадлежности анализируемого образа соответствующему субъекту.

Если выше приведенную формулу на каждой итерации домножать на вес j -го признака W_j (при этом $W_j < 1$), то в таком случае можно говорить о так называемом «осторожном» байесовском классификаторе, так как при его реализации нет «слепого» доверия к входным данным. Тогда каждый шаг расчета апостериорной вероятности гипотез с учетом значения очередного признака будет вычисляться по следующей формуле:

$$P(H_i | A_j) = P(H_i | A_{j-1}) + \left(\frac{P(H_i | A_{j-1}) P(A_j | H_i)}{\sum_{i=1}^n P(H_i | A_{j-1}) P(A_j | H_i)} - P(H_i | A_{j-1}) \right) \times (W_j)$$

Вес должен принадлежать интервалу $[0; 1]$ и может присваиваться шагам, исходя из информативности или коррелированности соответствующих признаков либо иным способом. Вес помогает балансировать пороговое значение апостериорных вероятностей и защищает процесс принятия решений от так называемых сбоев. Например, сбой происходит при $P(A_j | H_i) = 0$, когда значение j -го признака попадет в интервал гистограммы i -й гипотезы (эталона), соответствующий нулевой плотности (относительной частоте). Тогда апостериорная вероятность i -й гипотезы на следующих шагах уже не поднимется выше нуля. При $W_j = 1$ получим классическую формулу Байеса и, соответственно, «наивный» байесовский классификатор.

При построении классификаторов на основе байесовских нейронов для НПБК, также ставится задача разделения нейроном данных по уровню коррелированности. В простейшем случае выделяют три уровня: коррелированные, отрицательно коррелированные и независимые данные. Соответственно такой нейрон будет иметь три варианта значений на выходе функции активации.

Также требуется обеспечить, чтобы для всех «Чужих» пары признаков были сбалансированы по взаимной корреляционной зависимости. Это означает, что должны быть не только признаки со слабой корреляционной зависимостью, но и признаки с заметной и высокой (по шкале Чеддока) взаимной корреляцией. Это требуется, чтобы можно было создать равное





количество нейронов, ориентированных на обработку независимых, положительно и отрицательно зависимых признаков.

Учет всех выше перечисленных условий при проектировании классификатора на основе байесовских нейронов позволяет классификатору работать в пространстве сильно зависимых признаков и, используя подобные зависимости, существенно повышать уровень надежности распознавания образов. Более того такие нейроны способны генерировать ключи потенциально большой длины.

Оценка работы подобных классификаторов производится по окончании сессии обучения, по средствам расчета относительных частот возникновения ошибок 1-ого и 2-ого рода. Как отмечалось во введении курса, в биометрических системах эти показатели принято называть вероятностями ошибок «ложного отказа» (FRR) и «ложного доступа» (FAR). Сравнение биометрических систем выполняется по коэффициенту равной вероятности ошибок ($EER=FRR=FAR$).

Извлечение признаков из временных рядов. Одним из возможных подходов к извлечению признаков из биометрических образов является метод, заключающийся в экстракции значимых признаков из временных рядов, представляющих собой собранный в разные моменты времени статистический материал о значении каких-либо параметров (в простейшем случае одного) исследуемого процесса. Наиболее часто работа с временными рядами осуществляется при использовании голосовой биометрии, что продиктовано удобством обработки такого биометрического материала. Непосредственное извлечение признаков осуществляется при помощи математико-статистических методов анализа, предназначенных для выявления структуры временных рядов и их прогнозирования. К методам, применимым в задачах извлечения признаков из биометрических образов, можно отнести следующие:

Спектральный анализ (разложение ряда на сумму функций синусов и косинусов с различными частотами, для определения тех, появление которых особенно существенно и значимо).





Корреляционный анализ (позволяет выделять различные формы периодичности и взаимовлияния временных процессов, а также осуществлять прогнозирование будущего поведения временного ряда)

Кепстральный анализ (базируется на функции обратного преобразования Фурье от логарифма спектра мощности сигнала)

Вейвлет анализ (сигнал представляется совокупностью волновых пакетов - вейвлетов, образованных на основе некоторой исходной функции).

Например, для голосовых паролей вычисляются амплитудно-частотные характеристики речевого сигнала, которые являются значениями признаков a_j , для подписи — коэффициенты Фурье или вейвлет-преобразования функций координат и давления пера на планшет [10]. В последнем случае (для получения коэффициентов Фурье) такие параметры подписи, как функции положения пера на планшете $x(t)$, $y(t)$ и давления пера на планшет $p(t)$ (где t — время в дискретной форме) на этапе нормирования подвергаются одномерному преобразованию Фурье, а затем обратному преобразованию Фурье.

Еще одними важными параметрами, используемыми при работе с биометрическими признаками, описываемыми временными рядами, являются такие понятия, как шкала-мел и мелкепстральные коэффициенты.

Для представления интенсивности звука в определенной области частот используются единицы [мел]. Мел - единица высоты звука, основанная на восприятии этого звука нашими органами слуха. Мел удобно применять в целях анализа речи человека, так как его использование «приближает» алгоритмы обработки данных к человеческим параметрам восприятия, что благотворно сказывается на качестве распознавания.

Шкала мел отражает эмпирически подобранные значения, отражающие особенности слуха человека. На ее основе рассчитываются мел-частотные кепстральные коэффициенты (MFCC), которые в последнее время получили весьма широкое применение. Последовательность их расчета, следующая:

1. С использованием преобразования Фурье рассчитывается спектр речевого сигнала.
2. Полученный спектр пересчитывается для мел-шкалы.
3. Формируются треугольные перекрывающиеся окна, равномерно расположенные на мел-шкале.





4. Значения векторов спектра сигнала умножаются на оконную функцию после чего находится энергия сигнала, которая попадает в каждое из окон анализа.

5. Полученные коэффициенты (значения энергии) возводятся в квадрат и логарифмируются.

6. Используется еще раз преобразование Фурье, но, как правило, в связи с переходом к расчету значений квадрата комплексного спектра, применяется дискретное косинусное преобразование (ДКП). Особенностью ДКП является тот факт, что оно позволяет эффективно работать с четными функциями, при этом спектр остается вещественным, что очень удобно для его дальнейшего использования[17].

