## Постановка задачи

Как видно из обзора подходов к мультибиометрии, в настоящее время это один из основных способов повышения качества распознавания системы в целом. Модули распознавания в мультибиометрической системе могут быть организованы в последовательные или в иерархические структуры с различным числом уровней, причём в ряде случаев последовательное прохождение отпечатком через мэтчеры является следствием технической оптимизации правила объединения [3 статья].

При описании алгоритмов в литературе в качестве оценки качества даётся РХПУ принимающего устройства, которая показывает связь FAR и FRR для единственного сравнения. Аналогично демонстрируются результаты экспериментов мультибиометрических алгоритмов. Между тем, эксплуатационные характеристики реальных систем даются в терминах FAR(n) и FRR(n), которые определяются эмпирически в ходе работы системы. Очевидно, что в случае одноалгоритмической системы обе эти характеристики являются производными от значения порога системы. Способы их нахождения приведены выше. Однако в мультибиометрической системе эти параметры зависят как от множества системных порогов отдельных алгоритмов, так и от используемого способа их объединения [Гайд ту мультибиометрикс]. РХПУ мультибиометрической системы нельзя получить только анализом распределения степеней схожести отдельных алгоритмов, особенно для большого числа мэтчеров.

С практической точки зрения при оптимизации порогов мультибиометрической системы в условиях неизвестной зависимости между ними и качественными характеристиками системы существует риск попасть в некоторый локальный максимум. Наличие модели, которая бы могла предсказать хотя бы приблизительный вид такой зависимости, позволило бы проводить шаг предварительной оптимизации порогов, что позволило бы, в свою очередь, существенно уменьшить этап оптимизации после внедрения системы.

Дополнительное затруднение представляет и утверждение о наличии положительной корреляции между модулями распознавания, работающими по разным алгоритмам на основе одного и того же признака [Что-нибудь]. Как следствие, если один модуль распознавания даёт положительный ответ, то вероятность того, что другой модуль также даст положительный ответ, возрастает.

Таким образом, необходимо построить модель качественных характеристик мультибиометрической системы или определить способ построения таких моделей, которые бы отвечали следующим требованиям:

* В модели учитываются как сами алгоритмы, так и способ их объединения
* На основании модели можно предсказывать поведение системы и делать предварительную оптимизацию до её введения в эксплуатацию

## Теоретико-вероятностные характеристики биометрических систем

При построении нашей модели качества биометрической системы необходимо, прежде всего, изучить теоретико-вероятностные характеристики процесса идентификации.

Идентификация является по определению поиском 1:N. При этом результатом работы модуля сопоставления является численная метрика схожести предоставленного шаблона и шаблона из базы. Итоговое решение об идентификации получается на основе анализа всего множества отдельных решений. В [гайд ту биометрикс] показано, что с точки зрения теории вероятностей результат каждого из сопоставлений не зависит от других результатов, что позволяет рассматривать вероятность некоторого результата работы системы как произведение вероятностей результатов отдельных сопоставлений.

При переходе от одноалгоритмических систем к мультиалгоритмическим системам возникает проблема независимости результатов сопоставления одних и тех же пар шаблонов разными алгоритмами. В [ссылка] показано, что в этом случае между ними существует положительная корреляция, которую необходимо учитывать в модели.

Однако данная проблема была решена в [индепенденс], где экспериментально установлено, что мэтчеры, для которых уровень равной ошибки (уровень FAR или FRR для такой рабочей точки системы, в которых эти показатели равны, Equal Error Rate, EER) не превышает 5% могут быть рассмотрены как статистически независимые без существенной потери точности. Такому уровню равной ошибки соответствует большинство современных алгоритмов, достигших реального применения.

На этом основании мы будем использовать тот же теоретико-вероятностный подход, основанный на независимости отдельных сопоставлений, к мультиалгоритмическим системам, что и к одноалгоритмическим.

## Анализ стратегий объединения на уровне принятия решения

Рассматривая качественные характеристики комбинированных, мультибиометрических подходов к построению систем мы будем рассматривать случай двух модулей распознавания. Такое ограничение связано, во-первых, с практикой внедрения биометрических систем реального времени.

При рассмотрении стратегий мы будем качественно анализировать выигрыш или проигрыш их характеристик распознавания путём нахождения соответствующих отношений конкретных характеристик к таковым для рассмотренного случая одноалгоритмической идентификационной системы.

### Стратегия И

Стратегия И основана на применении конъюнкции к результатам работы двух мэтчеров: принять общее решение о сопоставлении, если есть один и только один шаблон, который был сопоставлен с предоставленным обоими мэтчерами, иначе принять решение о несопоставлении.

Оценим характеристики качества распознавания на основе указанных выше положений.

FAR(m) можно оценить по аналогии с одноалгоритмической системой, если мы будем рассматривать каждый алгоритм как независимую биометрическую систему. Допуск злоумышленника произойдёт в случае ошибки обеих систем, что в случае независимых случайных событий соответствует операции умножения:

Здесь и далее мы будем полагать, что комбинированная система получается из одноалгоритмической добавлением дополнительного алгоритма. Тогда изменение в FAR(m) системы составляет

Следовательно, FAR(m) всегда уменьшается при использовании стратегии И.

Аналогично найдём GAR(m):

Таким образом, стратегия И приводит к уменьшению вероятности корректного доступа. Однако поскольку первый множитель стараются держать близким к единице, это уменьшение, как правило невелико.

Коэффициент ложной идентификации определяется по аналогии:

Вероятность ложной идентификации в мультибиометрической системе со стратегией И существенно меньше таковой в одноалгоритмической системе благодаря множителю , каждый из сомножителей которого близок к нулю.

Поскольку

то

Аналогично

Поэтому

Отсюда следует, что , то есть стратегия И ухудшает кожффициент ложного отказа доступа к системе. Учитывая результат, полученный для GAR(m)и необходимо сделать вывод, что стратегия И безусловно повышает безопасность системы при понижении удобства её использования.

Однако другое достоинство этой стратегии состоит в возможности оптимизации процедуры сопоставления. Если данный шаблон был отвергнут одним из алгоритмов как несопоставленный, нет необходимости вычислять степень схожести вторым алгоритмом. Фактически такая реализация соответствует концепции непрерывной классификации, выдвинутой Люминиа в [Luminia]. Она заключается в отказе от дискретной классификации, то есть разбиения всего множества шаблонов на непересекающиеся классы (например, классификация Гальтона-Генри [Henry 1900]) и отнесении предоставленного шаблона к некоторому классу на основе близости некоторой соответствующей ему метрики метрике класса, в пользу вычисления некоторой метрики близости для всех шаблонов базы. Все шаблоны со значением метрики выше пороговой образуют класс, с которым сопоставлен предоставленный шаблон. Значение порога намеренно делается достаточно низким, чтобы в класс входило несколько шаблонов.

Число шаблонов, которое при такой организации системы попадает на вход второго алгоритма, является случайной величиной, поэтому мы можем оценить только среднее время сопоставления для такой стратегии составляет

Где – среднее число сопоставленных алгоритмом А шаблонов. Это число отличается для случаев, когда к системе обращается злоумышленник и зарегистрированный пользователь:

Но поскольку FRR достаточно близок к нулю, можно считать, что при больших m . Так как по затратам времени – это худший из двух случаев, следует оценивать по нему. На практике множитель растёт медленнее, чем убывает , что позволяет отбросить большую часть слагаемых без значительной потери точности. Например, для базы в 100 шаблонов с FAR=10-3 уже 4-й член суммы будет равен приблизительно .

### Стратегия ИЛИ

Стратегия ИЛИ является дополняющей к стратегии И: принять общее решение о сопоставлении, если есть один и только один шаблон, который был сопоставлен с предоставленным хотя бы одним мэтчером, иначе решение о несопоставлении. При рассмотрении этой стратегии оказывается полезно использовать формулу включений и исключений (inclusion-exclusion). Для краткости для двух вероятностей A и B будем обозначать

Нахождение FAR(m), GAR(m), FIR(m) и FRR(m) осуществляется по тем же принципам, что и раньше.

Из этого соотношения видно, что множитель , всегда не меньший (а на практике строго больший) единицы, приводит к тому, что стратегия ИЛИ всегда хуже с точки зрения FAR для одиночного сравнения. Однако добавления множителя позволяет теоретически достигать лучшего качества распознавания на уровне системы в целом, если

Обозначим , тогда

Так как , то , откуда

Если Q>0,5, то улучшение системного FAR обеспечивается стратегией ИЛИ даже при добавлении к исходному алгоритму менее качественного с точки зрения индивидуального FAR.

Для улучшения качества распознавания зарегистрированных пользователей требуется выполнение неравенства

В данном случае показатель степени играет более значимую роль, чем значение коэффициентов алгоритмов, поскольку правая часть неравенства с уменьшением коэффициентов ложного доступа обоих алгоритмов стремится к . В логарифмической записи

Например, для алгоритмов с m должно быть меньше 11.

Коэффициент ложной идентификации равен

Он становится лучше, если

### Стратегия Top k + И

Стратегия Top k + И является развитием методики непрерывной классификации.

На базе этой идеи предлагается стратегия Top k + И. Вместо использования порога для определения класса шаблонов предлагается относить к итоговому классу N шаблонов с наибольшим значением метрики. В [Гайд ту биометрикс] описан математический аппарат, который используется для описания таких систем. В частности, основной характеристикой является кривая суммарного сходства (Cumulative Match Curve, CMC) – накопленная сумма массы ранговой вероятности, то есть сумма вероятностей, что для входящего запроса правильная личность будет стоять на заданной позиции. Проще говоря, CMC(k) – это вероятность, что в отсортированном по убыванию метрики массиве шаблонов в числе первых k шаблонов окажется соответствующий. В [там же] показано, что для каждой фиксированной базы CMC(k) является постоянной величиной. k выбирается таким образом, чтобы с одной стороны максимизировать CMC(k), а с другой стороны минимизировать время последующей обработки. Предельным случаем является CMC(m)=1, но при этом модуль не фильтрует записи. Для k<m всегда существует вероятность промаха классификатора, равная 1-CMC(k).

Необходимо отметить, что в такой стратегии ранг шаблонов, показывающий относительную степень их схожести с предоставленным, не используется при дальнейших сопоставлениях шаблонов. В отличие от мультибиометрических систем с объединением по рангу при наличии нескольких непрерывных классификаторов в системе на вход последующих подаётся результат работы предыдущих, а не вся первоначальная база.

Когда к системе получает доступ злоумышленник, классификатор отработает «вхолостую»: записи злоумышленника в базе нет. Следовательно, в результате его работы база сокращается до k вариантов, что приводит к оценке вероятности ложного доступа:

Найдём отношение FAR(k) к FAR(m):

Таким образом, вероятность ложного доступа всегда снижается при k<<m.

В системе с классификатором на вероятность корректного доступа будет также влиять и вероятность верной классификации:

GAR и FRR влияют на удобство пользования системой: чем выше GAR и ниже FRR (при том же или меньшем FIR), тем удобнее система для пользователя. Аналогично с предыдущим случаем, с точки зрения GAR система будет удобнее при условии:

Ложная идентификация в системе с классификатором может произойти в двух случаях:

Ошибка классификатора привела к тому, что в отфильтрованной базе нет нужной записи, и произошло ложное распознавание одного из отпечатков;

Классификатор сработал корректно, но вместо верного отпечатка был сопоставлен другой.

Эти случаи описываются следующим соотношением:

Ложная идентификация в системе с классификатором будет лучше, если:

Наконец, коэффициент ложного отказа доступа в системе с классификатором будет равен:

Система становится удобнее при , отсюда:

Так как неравенства совместны, для улучшения всех характеристик вероятность корректной классификации при заданном k должна быть выше максимального из полученных значений.

Временные характеристики данной стратегии определяются сходно со стратегией И, однако здесь время работы второго модуля определяется значением k.

Ранжирование – это сортировка множества пар (ключ шаблона, метрика). На больших объёмах данных время этой сортировки перестаёт быть пренебрежимо малым, поэтому при реализации необходимо выбрать алгоритм сортировки, наиболее полно соответствующий архитектуре вычислительного устройства.