Конспект лекции

А.Е. Сулавко

Защищенное исполнение искусственного интеллекта

**Защищенное исполнение нейросетевых алгоритмов на базе корреляционных нейронов**

**ЗАЩИЩЕННОЕ ИСПОЛНЕНИЕ НЕЙРОСЕТЕВЫХ АЛГОРИТМОВ НА БАЗЕ КОРРЕЛЯЦИОННЫХ НЕЙРОНОВ**

Мера Минковского позволяет в некоторых случаях «уравнять» вероятности ошибок распознавания в пространстве коррелированных и независимых признаков. Однако далее будет показано, что корреляция не только искривляет *пространство признаков* относительно каждого класса образов, но и переносит часть информации об образах в *«скрытые» измерения*. Речь идет об информации, касающейся уровней искривления пространства признаков в направлении каждого измерения. Чтобы извлечь данную информацию введем два варианта меры Байеса-Минковского (1), которая оперирует разностями между признаками и таким образом осуществляет анализ данных, которые заключены в *мета-пространстве* с размерностью , т.е. между измерениями исходного пространства признаков.

(1)

Под мета-пространством в данном контексте подразумевается разность вида:

которая фактически является грубой *(точечной) оценкой корреляции* между двумя исходными признаками под номерами j и t. Под *точечной оценкой* понимается оценка, сделанная всего по одному примеру тестовой выборки, но при наличии некоторых априорных знаний, сделанных по обучающей выборке небольшого объема.

«Наивная» схема классификации Байеса является полностью корректной, когда признаки независимы. В этом случае пространство признаков абсолютно не имеет кривизны. Мера Минковского наоборот измеряет расстояние в искривленном пространстве. Новую метрику решено назвать мерой Байеса-Минковского, так как она «преобразует» пространство коррелированных в разной степени признаков в мета-пространство независимых признаков.

Динамика изменения EER(*p*) для меры Байеса-Минковского имеет обратную тенденцию по сравнению с показателями EER(*p*) для меры Минковского. Если признаки коррелированы, мера Байеса-Минковского дает более высокий результат, чем, если признаки независимы. Причем, вероятности ошибок распознавания в пространстве коррелированных признаков для меры Байеса-Минковского ниже, чем уровень ошибок для меры Минковского в случае независимости признаков. Для меры Байеса-Минковского также характерно следующее: чем выше уровень корреляции, тем больше снижается уровень ошибок по сравнению со случаем независимости признаков.

На первый взгляд это противоречит классической теории математической статистики, которая утверждает: корреляция между признаками указывает на то, что часть информации в признаках повторяется. Но учитывая полученные результаты, данное утверждение стоит уточнить: *пространство признаков искривляется из-за корреляционных связей, причем относительно каждого класса образов и каждого измерения характер искривления различен, что влечет за собой уменьшение количества информации об отличии классов, наблюдаемой в пространстве признаков, и появление новой информации в мета-пространстве Байеса-Минковского, но в большем объеме*.

Мера Байеса-Минковского является мерой-антагонистом по отношению к мере Минковского, так как обладает почти противоположными свойствами. Метрика Байеса-Минковского подходит для распознавания образов в пространстве сильно зависимых признаков, настраивая *степенной коэффициент p* можно существенно повысить надежность распознавания образов. При различных значениях p мера Байеса-Минковского работает по-разному (фактически при разных p мы имеем разные меры близости, решения которых тем менее коррелированы, чем сильней отличается этот параметр).

Чтобы обучить нейрон на базе меры Минковского или Байеса-Минковского достаточно вычислить параметры *mj* и *σj* по данным обучающей выборки, что элементарно. Таким образом, сети из таких нейронов будут способны к автоматическому обучению. Очевидным недостатком таких сетей с точки зрения безопасности является необходимость хранить параметры *mj* и *σj*, компрометирующие данные обучающей выборки для класса «Свой» (за исключением информации о корреляции между признаками).

**Мета-пространства признаков Байеса-Минковского, не компрометирующие знания искусственного интеллекта**

Введем меры близости:

(2)

обладающие схожими свойствами с мерой (1) в плане обработки сильно коррелированных признаков. Преимущество метрик (2) заключается в том, что для ее корректной работы требуется хранить только параметры . Однако среднеквадратичные отклонения признаков для класса образов «Свой» допустимо изменить на среднеквадратичное отклонение признаков для класса «Чужие» (назовем его нормирующим коэффициентом ), с учетом этого скорректируем формулу (2):

(3)

*Нормирующий коэффициент*  не компрометирует данные какого-либо класса, так как это интегральная оценка разброса значений признака для всех классов (для класса «Чужие», характеризующего все возможные классы), т.е. таким образом обеспечивается *дифференциальная конфиденциальность* (можно добавить к шум, чтобы немного сместить его значение). Смысл коэффициентов заключается в приведении всех признаков примерно к единой области значений. Под *мета-признаком* в случае использования меры (3) подразумевается разность вида:

(4)

что фактически делает метрику (3) линейным классификатором, работающим без настройки весовых коэффициентов, но в мета-пространстве признаков.

Аналогично можно ввести меры близости:

(5)

где – это *второй нормирующий коэффициент*, который может быть вычислен как математическое ожидание значений j-го признака для класса образов «Чужие» с учетом добавления шума. Под мета-признаком в случае использования меры (5) подразумевается разность вида:

(5)

Меры близости (5) имеют близкие свойства с мерами (1) и (3), но результаты классификации, полученные этими метриками, не будут иметь стопроцентной корреляции.

В мета-пространствах признаков Байеса-Минковского можно строить любые интеллектуальные системы анализа данных (для задач классификации, регрессии и др.). *Для этого нужно применить одно из возможных отображений исходного признакового пространства в* ***спрямляющее мета-пространство*** *признаков Байеса-Минковского , которое имеет* ***большую*** *размерность, а именно:*

,

где *n* – исходная размерность пространства признаков.

Вычислительный эксперимент по идентификации сгенерированных образов на закрытом множестве классов ([2], стр. 66)

**Принципы построения моделей корреляционных нейронов Байеса-Минковского для защищенного исполнения искусственного интеллекта.**

Корреляционный (автокорреляционный) нейрон должен разделять данные по уровню коррелированности. В простейшем случае можно ограничиться разделением данных на положительно коррелированные, отрицательно коррелированные и независимые. Такой нейрон будет иметь три варианта значения на выходе функции активации (функции квантования). Простейший корреляционный нейрон работает в режиме верификации образов и разделяет классы «Свой» и «Чужие». Рассмотрим данный вариант построения нейрона.

Пусть нейрон работает с мета-признаками, которые получены путем отображения (4). Корреляционный нейрон будет иметь два функционала для анализа входных данных и пороговую функцию активации для квантования данных на выходе.

Первый функционал будет разделять положительно коррелированные данные и независимые, его можно определить так:

(6)

Второй функционал будет определять отрицательно и положительно коррелированные данные:

(7)

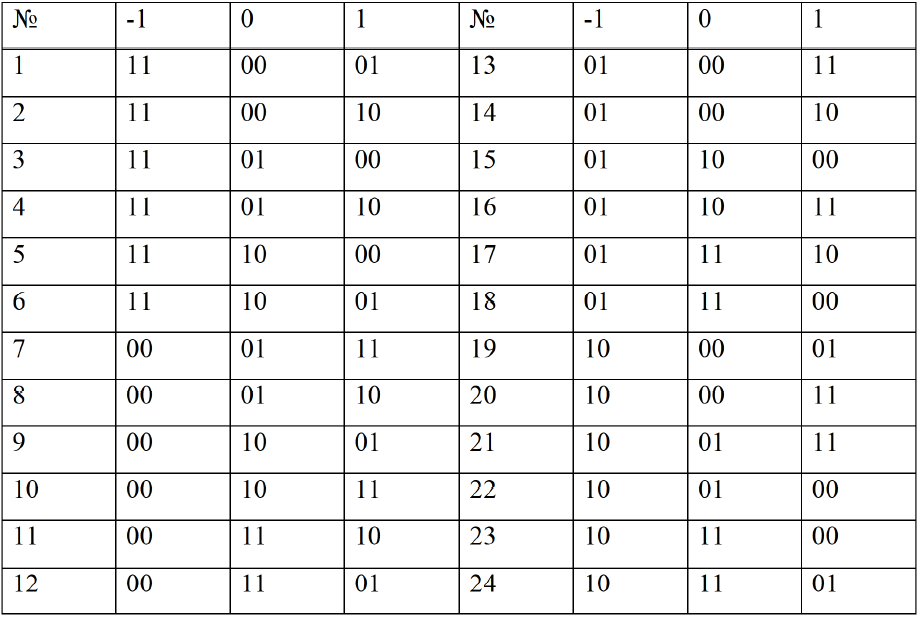
Состояния на выходе нейрона задаются активационной функцией с двумя квантователями:

(8)

Состояния {«-1», «0», «1»} указывают на положительную корреляцию, отсутствие заметной корреляции или отрицательную корреляцию, соответственно. Далее следует применить таблицу перевода этих состояний в двухбитный код. Если мы применим одну таблицу перевода в двоичные состояния {«01», «00», «10»}, то предъявление образов «Чужих» даст последовательность длинных кодов, состоящих в большей степени из нулей. Другими словами вероятность появления состояний «0» (Р(0)) в разрядах выхода ПБК намного выше, чем вероятность появления состояний «1» (Р(1)), что нежелательно. Желательным является наличие равновероятных событий Р(«0») = Р(«1») в разрядах выходов ПБК.

Для достижения желаемого эффекта Р(«0») = Р(«1») необходимо использовать множество таблиц перевода троичных состояний в двоичные коды, имеется всего 24 варианта этих таблиц для рассматриваемого случая.

Таблица 1. – 24 варианта хэширующих преобразований отклика нейрона в двоичный код



При равномерном применении по отношению к нейронам всех 24 таблиц хеширующих преобразований выполняется условие Р(«0») ≈ Р(«1»). Таблица преобразований отклика нейрона выбирается с учетом того, на какие биты настраивается нейрон. Допустим, при верном решении нейрон должен продуцировать состояния «10», а верное решение связано с реакцией нейрона на положительно коррелированные данные. Тогда для него случайным образом задается одна из 6 возможных таблиц преобразования. Конечно сам криптографический ключ или пароль, на связывание с которым настраивается ПБК, также должен состоять примерно из равного количества нулей и единиц, что подразумевается при его генерации в соответствии с принятыми нормами и стандартами.

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 <https://statpsy.ru/correlation/velicina/>) взаимной корреляцией. Это требуется, чтобы можно было создать равное количество нейронов, ориентированных на обработку независимых, положительно и отрицательно зависимых признаков.

Балансировка по корреляции между признаками не обязательно должна быть идеальной. Главное, чтобы можно было создать достаточное и равное число нейронов для обработки слабо коррелированных пар признаков, положительно и отрицательно сильно (или хотя бы заметно) коррелированных пар признаков. Также важно, чтобы количество возможных пар признаков в этих категориях коррелированности также было почти равным (в среднем по всем классам образов, т.е. для класса «Чужие»).

*Обучение корреляционных нейронов сводится к определению не только таблиц хеширующих преобразований, но и таблиц связей нейронов и порогов срабатывания*. Также должны быть вычислены нормирующие коэффициенты по каждому признаку (например, только по ), с помощью которых осуществляется нормализация данных при переходе в мета-пространство признаков Байеса-Минковского.

Вычисление таблиц связей опирается на матрицу коэффициентов корреляции между признаками для образа «Свой». Корреляционная матрица рассчитывается на основании обучающей выборки. Из матрицы выбираются соответствующие пары признаков с определенным уровнем взаимной корреляции (например, слабым, высоким/заметным положительным, высоким/заметным отрицательным). После обучения эта матрица должна удаляться.

Вычисление порогов срабатывания должно опираться на распределение коэффициентов корреляции между признаками, которое строится исходя из аналогичной корреляционной матрицы, но уже для класса «Чужие» (после обучения эту матрицу также целесообразно удалить). Чтобы сеть из корреляционных нейронов Байеса-Минковского обладала хорошими хэширующими свойствами, необходимо, чтобы состояния каждого нейрона были равновероятны (по аналогии с соответствующим требованием для двухуровневого линейного нейрона в ГОСТ Р 52633.5-2011). Для трехуровневых нейронов вероятность каждого состояния на выходе должна быть примерно равной 0,333 (пороги должны быть заданы так, чтобы вероятность возникновения состояний {«-1», «0», «1»} на выходе нейрона была одинакова при поступлении на его входы образа «Чужого»).

*Для корректной работы корреляционных нейронов Байеса-Минковского признаки должны иметь распределение, близкое к нормальному, по крайне мере, в представленном исполнении.* *Таким образом, к блоку извлечения признаков необходимо предъявлять требование о нормальности распределения признаков, что можно реализовать множеством способов (например, используя для анализа изображений и сигналов вариационный автокодировщик).*

Эксперименты показывают, что переход в мета-пространство признаков не ведет к проявлению проблемы «проклятья размерности», если признаки коррелированы. Но когда признаки сильно коррелированы, то при использовании аналогичной обучающей выборки (например, 10 примеров) удается достичь более высоких результатов, если перейти в более высокоразмерное пространство мета-признаков Байеса-Минковского. При этом количество вычислений растет линейно по отношению к увеличению размерности пространства признаков, по крайней мере, при использовании байесовского классификатора. Количество же признаков при переходе в мета-пространство Байеса-Минковского при этом растет не по экспоненте, а по степенному закону *0,5n2-0,5n*.

Каждый ПБК, основанный на представленной модели корреляционных нейронов Байеса-Минковского (возможных вариаций моделей существует множество), строится для определенного класса образов и работает в режиме верификации. Режим идентификации может быть реализован через построение множества ПБК на базе корреляционных нейронов. При этом блок извлечения признаков может быть общим для всех ПБК (например, глубокая нейронная сеть).

*Нейроны обучаются быстро и автоматически.* Данные обученных нейронов не компрометируют обучающую выборку и криптографические ключи. Атака Маршалко для таких нейронов видится неосуществимой, так как нет никаких весовых коэффициентов. Есть только случайные таблицы хеширующих преобразований, таблицы связей и порогов. Сами пороги, возможно, будут стандартные, при условии, что к признакам будут предъявляться относительно жесткие требования по балансировке корреляционных связей. В любом случае, эти данные не дают информации об обучающей выборке и ключе. Нельзя сопоставить, какую зависимость имеют те или иные признаки, так как каждый нейрон продуцирует три двухбитных состояния на выходе. Вполне возможно, что такие нейроны не нуждаются в криптографической защите вообще. Либо нуждаются в ней, но для этих целей может быть адаптирована спецификация или разработана новая, с применением схожего подхода. В любом случае, атаки по ускорению перебора образов «Чужих» с целью извлечения знаний из ИИ уже не будут эффективны (по крайней мере настолько, как для нейросетевых ПБК, обучаемых по ГОСТР 52633). Наличие трех возможных состояний нейрона вместо двух сильно усложняют такой перебор.

Наконец, *преимуществом предлагаемой схемы является высокая потенциальная длина ключа*. Возвращаясь к тому же примеру задачи верификации подписи, мы можем получить 86320 мета-признаков из 416 исходных признаков. Можно ожидать, что для корреляционного нейрона число входов можно снизить примерно в 2 раза по сравнению с классическими нейронами, обучаемыми по ГОСТ Р 52633. Таким образом, при количестве входов нейрона, равном 8, потенциальная максимальная длина ключа (при идеальной балансировке признаков по корреляции) составляет 21580 бит (с учетом того, что каждый нейрон продуцирует два бита ключа на выходе).

**Список источников**