На правах рукописи

Федоренко Юрий Сергеевич

Модели, алгоритмы и программное обеспечение для выбора персонализированных предложений в сети интернет в режиме реального времени

Специальность 05.13.11
Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей (технические науки)

Автореферат
диссертации на соискание учёной степени
кандидата технических наук

Москва — 2021

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана).

Научный руководитель: **Гапанюк Юрий Евгеньевич**
кандидат технических наук, доцент

Официальные оппоненты: **Конышев Михаил Юрьевич,**
доктор технических наук, доцент,
Федеральное государственное унитарное предприятие «Научно-технический центр «Орион» федеральной службы безопасности Российской Федерации,
научный консультант

Елисеев Владимир Леонидович,
кандидат технических наук,
Акционерное общество «Информационные технологии и коммуникационные системы»,
руководитель Центра научных исследований и перспективных разработок

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский технологический университет «МИСиС»

Защита состоится «27» мая 2021 года в 13 часов 00 минут на заседании диссертационного совета Д 999.216.02 при МАИ и МГТУ им. Н.Э. Баумана по адресу: 105005, г. Москва, 2-я Бауманская ул., д. 5, стр. 1, зал Ученого совета ГУК МГТУ им. Н.Э. Баумана.

С диссертацией можно ознакомиться в библиотеке МГТУ им. Н.Э. Баумана и на сайте http://bmstu.ru.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба направлять по адресу: 105005, г. Москва, 2-я Бауманская ул., д. 5, стр. 1, Ученому секретарю диссертационного совета Д 999.216.02.

Автореферат разослан «__» _____ 2021 года.

Ученый секретарь
диссертационного совета,
д.т.н., доцент А.Н. Алфимцев

Общая характеристика работы

Актуальность работы. Задача построения программных систем для выбора персонализированных предложений в интернете особенно актуальна в связи с проникновением интернета во все сферы жизни, включая деловую сферу, покупки, развлечения и т.д. Одно из подтверждений этому – динамика оборота рынка интернет-рекламы в России, в котором до 2021 года ожидается ежегодный рост оборота на уровне 12%. Также согласно данным системы цитирования Google Scholar за последние 5 лет было сделано более 15 тысяч публикаций по теме CTR prediction (прогнозирование частоты клика по рекламному объявлению). При этом более половины публикаций были сделаны после 2016 года. Продолжающийся академический интерес к данной области свидетельствует о том, что проблема важна и окончательно не проработана.

Выбор персонализированных объявлений для пользователей интернета (т.е. таких объявлений, на которые они будут кликать наиболее часто), является задачей машинного обучения, в которой на основании исторических данных о пользователе (и похожих на него пользователей) требуется подобрать очередную рекламную выдачу. Как известно, успешное решение таких задач во многом определяется используемым признаковым описанием входных данных. Существует множество методов извлечения, конструирования и отбора признаков, однако сложности возникают при применении этих методов к задачам онлайнового (динамического) обучения в высоконагруженных системах, где требуется параллельно обрабатывать множество запросов, укладываясь в жесткие ограничения по времени отклика. В частности, популярные деревья решений с трудом способны обучаться в режиме реального времени. Глубокие нейросети имеют много параметров, восстанавливают сложные функции с большим количеством локальных минимумов и, как следствие, плохо приспособлены к обучению в режиме реального времени. Также они вычислительно сложны, что делает их с трудом применимыми в системах выбора персонализированных предложений, где требуется быстро получать ответ на множество постоянно поступающих запросов. В результате при построении подобных систем, работающих в интернете в режиме реального времени, популярным подходом остается применение линейных моделей (например, логистической регрессии), которые хорошо подходят для задач динамического обучения. Однако при использовании исходных первичных признаков такие модели показывают неудовлетворительные результаты, поскольку их работа подразумевает проведение разделяющей гиперплоскости, в то время как данные редко оказываются линейно разделимыми. По этой причине требуется строить производные признаки, например, хешируя комбинации исходных признаков. Однако недостаток такого подхода заключается в необходимости трудоемкого процесса ручного отбора нужных комбинаций, которые

также может потребоваться модифицировать с течением времени. Данное диссертационное исследование направлено на разработку алгоритмического и программного обеспечения, позволяющего в системах выбора персонализированных предложений автоматически находить нужные комбинации признаков, обновляя их по мере поступления новых данных.

Целью диссертационного исследования является разработка моделей, алгоритмов и программного обеспечения для выбора персонализированных предложений в сети интернет без ручного конструирования признаков, что позволяет сократить объем работы, выполняемый экспертами предметной области. Для достижения поставленной цели решаются следующие задачи:

1. Разработка нейросетевой модели со специализированной архитектурой с поддержкой конструирования признаков при обучении в режиме реального времени. Создание метода прогнозирования частоты кликов пользователя по рекламному объявлению в интернете без ручного конструирования признаков на базе разработанной нейросети.
2. Моделирование системы выбора персонализированных предложений как системы массового обслуживания для выбора наилучшей конфигурации нейросети в рамках предложенной архитектуры.
3. Разработка алгоритмов для обеспечения работы предложенной нейросетевой модели со специализированной архитектурой с учетом заданных временных ограничений. Проектирование и создание программного обеспечения на базе разработанных алгоритмов.
4. Разработка методики статистического тестирования для сравнения качества работы моделей машинного обучения на фиксированной тестовой выборке и её применение для анализа качества работы предложенного метода прогнозирования частоты кликов пользователя.
5. Апробация разработанных моделей и алгоритмов для выбора персонализированных предложений в интернете.

Объектом исследования является программная система выбора персонализированных предложений в сети интернет. **Предметом исследования** являются модели и алгоритмы для выбора персонализированных предложений в сети интернет.

Научная новизна:

1. Разработана нейросетевая модель со специализированной архитектурой, позволяющая осуществлять конструирование комбинаций первичных признаков при обучении в режиме реального времени. Предложен метод прогнозирования частоты кликов пользователя по рекламному объявлению в интернете без ручного конструирования признаков на базе разработанной модели.
2. Предложена аналитическая модель системы выбора персонализированных предложений, позволяющая найти наилучшую конфигурацию

нейросетевой модели со специализированной архитектурой с учетом характеристик и требований к конкретной системе.

3. Разработана методика статистического тестирования для сравнения качества работы моделей на фиксированной тестовой выборке.

Практическая ценность полученных результатов заключается в устранении необходимости ручного подбора комбинаций признаков при построении модели прогнозирования частоты кликов, что может быть применено в системах выбора персонализированных предложений. Разработанные модели и алгоритмы позволяют найти наилучшую конфигурацию в рамках предложенной нейросети с учетом параметров конкретной системы и более эффективно по сравнению с другими подходами оценить качество её работы. Разработанная программная реализация предложенной нейросетевой модели со специализированной архитектурой превосходит в плане быстродействия (время обучения на одном примере и время прогнозирования) аналогичные реализации на базе нейросетевых фреймворков.

Положения, выносимые на защиту:

1. Метод прогнозирования частоты кликов пользователя по рекламному объявлению в интернете в режиме реального времени, который не требует ручного построения производных признаков.

2. Нейросетевая модель со специализированной архитектурой с поддержкой конструирования признаков при обучении в режиме реального времени, особенности её обучения и методика регуляризации.

3. Аналитическая модель системы выбора персонализированных предложений как системы массового обслуживания, позволяющая найти наилучшую конфигурацию нейросетевой модели со специализированной архитектурой с учетом характеристик и требований к конкретной системе.

4. Алгоритмы программной реализации предложенной нейросетевой модели со специализированной архитектурой, позволяющие увеличить в несколько раз производительность по сравнению с реализациями на базе нейросетевых фреймворков.

5. Методика статистического тестирования для сравнения моделей машинного обучения с заданным уровнем значимости на основе значений аддитивных метрик на тестовой выборке, которая менее затратна в вычислительном плане, чем традиционная кросс-валидация по k блокам.

Соответствие **паспорту специальности**. Результаты исследования соответствуют пунктам:

1 - «Модели, методы и алгоритмы проектирования и анализа программ и программных систем, их эквивалентных преобразований, верификации и тестирования»,

3 - «Модели, методы, алгоритмы, языки и программные инструменты для организации взаимодействия программ и программных систем»
паспорта научной специальности 05.13.11 - Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Личный вклад. Основные результаты диссертационной работы получены автором лично, что подтверждено личными публикациями и отражено в совместных статьях. Программное обеспечение разработано автором лично, и на него получено свидетельство о государственной регистрации программы для ЭВМ «Библиотека для работы с разреженной нейронной сетью со специализированной архитектурой (CustomSparseNN)».

Достоверность полученных результатов следует из применяемых строгих математических методов теории вероятностей, математической статистики, теории оптимизаций и массового обслуживания и искусственных нейронных сетей. Теоретические результаты подтверждаются экспериментальными исследованиями. Разработанная программная реализация показала свою эффективность в практических задачах.

Апробация работы. Основные положения и результаты диссертационной работы докладывались и обсуждались на следующих конференциях и семинарах:

1. XXII Международная конференция «Нейроинформатика», онлайн, 2020;
2. XXI Международная конференция «Нейроинформатика», Долгопрудный (МО), МФТИ, 2019;
3. 16th International Symposium on Neural Networks, Москва (Сколково), 2019;
4. IX Международная научно-практическая конференция «Интегрированные модели и мягкие вычисления», Коломна (МО), 2019;
5. XVIII, XIX и XX Международная конференция «Нейроинформатика», Москва (МИФИ), 2016, 2017, 2018;
6. Научный семинар в Научно-исследовательском институте Системных Исследований (НИИСИ) РАН;
7. Междисциплинарный научный семинар «Экобионика» в МГТУ им. Н.Э.Баумана.

Внедрение результатов работы. Разработанные автором модели и предложенные алгоритмы были использованы в научно-производственной деятельности компании Mail Ru Group для решения задачи прогнозирования частоты кликов пользователей по рекламным баннерам в интернете. Предложенная нейросетевая модель внедрена в разработанную в Mail Ru Group систему для анализа значимости новых признаков, что подтверждено актом о внедрении. Теоретические результаты использованы в учебном процессе в МГТУ им. Н.Э. Баумана, что также подтверждено соответствующим актом.

Публикации. Всего по теме диссертации опубликовано 9 научных работ (из них 4 входят в перечень ВАК РФ и 5 индексируются в SCOPUS) общим объемом **3,5 п.л.**

Объем и структура работы. Диссертация состоит из введения, четырех глав, общих выводов и заключения, списка литературы и прило-

жения. Объём диссертации составляет 170 страниц, включая 64 рисунка и 9 таблиц. Список литературы содержит 127 наименований.

Содержание работы

Во **Введении** обосновывается актуальность исследований, проводимых в рамках данной диссертации, формулируется цель и ставятся задачи работы, аргументируются научная новизна и практическая значимость.

В **Главе 1** рассмотрены особенности программной системы выбора персонализированных предложений (рекламной системы), дан критический обзор существующих методов извлечения признаков и проанализирована их применимость в рассматриваемой области. На Рис. 1 изображена схема системы выбора персонализированных предложений. Когда пользователь заходит на сайт, отправляется запрос на показ рекламы, который обрабатывается серверами подбора (их может быть несколько десятков). При обработке запроса определяются баннеры, которые могут быть показаны пользователю. Затем при помощи специализированной модели прогнозируется частота кликов для каждой пары <пользователь, баннер>, после чего выбираются баннеры, для которых полученная величина максимальна. Информация о совершенных показах и кликах записывается в логи. По ним в режиме реального времени производится обучение модели, которая регулярно передается в сервера подбора рекламы.

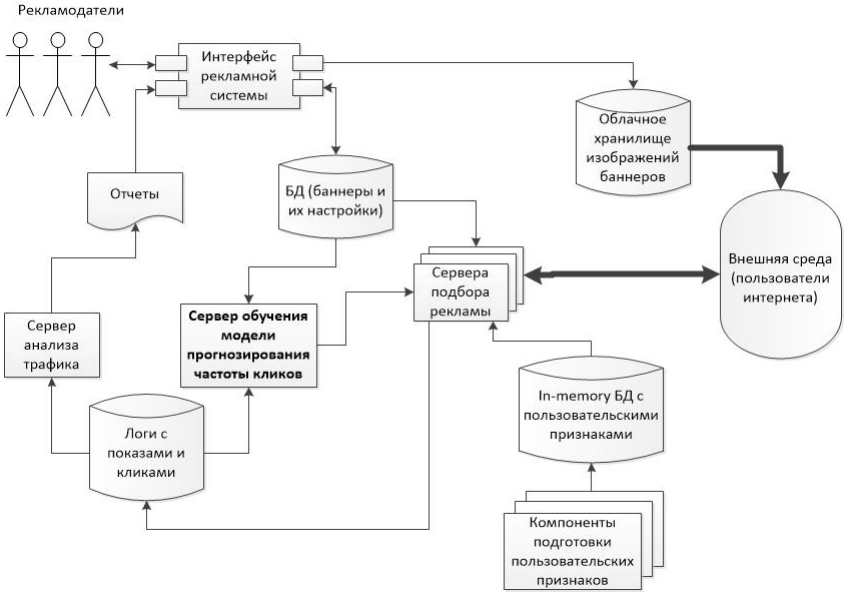


Рис. 1. Схема системы выбора персонализированных предложений

В качестве исходных (первичных) признаков модель прогнозирования частоты кликов использует признаки рекламного объявления, признаки пользователя и признаки места размещения рекламы (сайта). Большинство признаков категориальные. Для их представления используется бинарное кодирование, что приводит к размерности входных данных порядка нескольких миллионов. Другой важной особенностью системы выбора персонализированных предложений является большое количество обрабатываемых запросов. В частности, в работе показано, что в вечернее время нагрузка достигает 40 тыс. запросов в секунду, а время прогноза для одной пары <пользователь, баннер> не должно превышать 1.8 мс.

В работе решается задача регрессии: для произвольного пользователя U, описываемого вектором $(fu_1, fu_2,..., fu_n)$, и баннера B_i с признаковым описанием $(fb_{i1}, fb_{i2},..., fb_{im})$ необходимо спрогнозировать частоту кликов $fr(U, B_i)$ (речь идет о прогнозе, а не об оценке, поскольку в обучающей выборке ни показов, ни кликов для пары $< U, B_i >$ может не быть). При этом $B_i \in B$, где B - множество баннеров, которые могут быть показаны пользователю U (при подборе предложений необходимо спрогнозировать частоту кликов по всем таким баннерам, чтобы из них выбрать наилучшие). Также в данной работе считается, что частота кликов пользователя по каждому баннеру не зависит от других баннеров в рекламной выдаче (обоснование приведено во второй главе диссертации). Несмотря на выбор N баннеров с максимальной частотой кликов, просто отранжировать их недостаточно. Важны и реальные значения полученных частот, поскольку именно на основании них на ряде сайтов производится аукцион между разными системами выбора персонализированных предложений.

В диссертации указано, что с учетом специфики систем выбора персонализированных предложений применяемый в них метод для решения данной задачи предполагает использование линейной модели с хешированием признаков, когда значения хеш-функции от комбинаций признаков используются в качестве бинарных признаков на входе модели (Рис. 2).

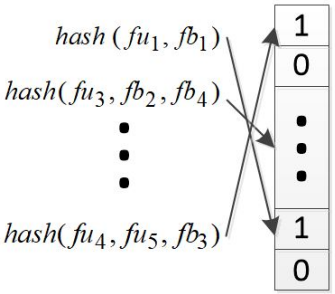


Рис. 2. Пример хеширования комбинаций признаков

Такой метод с переходом в другое признаковое пространство поддерживает обучение в режиме реального времени и позволяет производить сложные разделяющие поверхности за счет составных комбинаций первичных признаков. Серьезный недостаток заключается в том, что подбирать эти комбинации требуется вручную. Предложенные в диссертации модели и алгоритмы позволяют решить данную проблему.

Глава 2 диссертации посвящена разработке основного метода, моделей, а также предложенных методик и алгоритмов. На Рис. 3 представлена модель процесса выбора рекламных баннеров. Обучение модели произво-

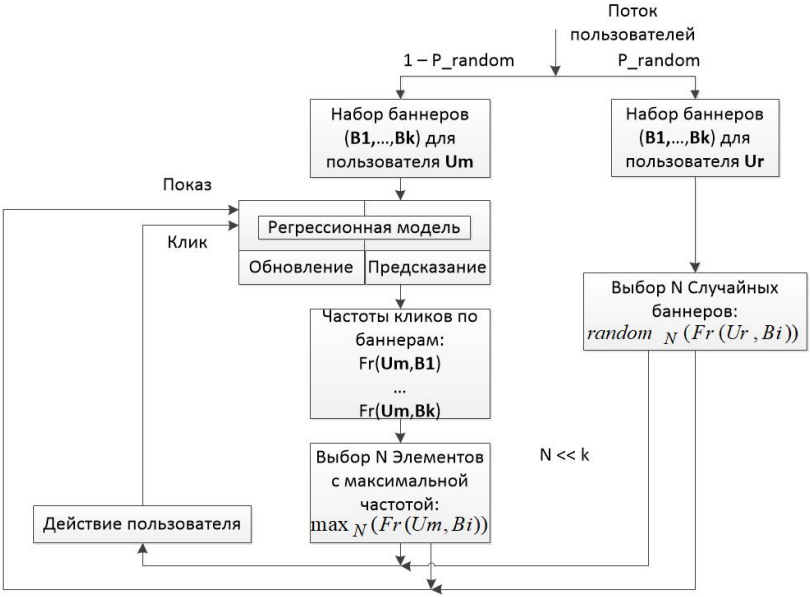


Рис. 3. Модель процесса выбора рекламных баннеров

дится по логам на основе показов и кликов. Небольшая часть запросов не проходит через модель, что приводит к случайному выбору баннеров. Это увеличивает разнообразие данных, т.к. позволяет попасть в обучающую выборку также и тем баннерам, для которых модель выдает низкий прогноз частоты кликов.

Разработанная нейросетевая модель оценки частоты кликов является нейронной сетью с разреженными связями с определенной структурой. Выходной слой нейросети состоит из одного нейрона с функцией активации «сигмоид», который выдает значение частоты клика для заданного набора входных параметров в диапазоне от 0 до 1. На вход нейронная сеть принимает набор категориальных признаков, каждому из которых отводится свой диапазон значений во входном векторе. В базовом варианте архитектуры на скрытом слое сети каждый нейрон связан с одной или двумя зона-

ми, отвечающими за конкретный признак (перебираются все возможные пары). Схема такой нейросети изображена на Рис. 4. Обучаясь на исторических данных в реальном времени, предложенная нейросеть реализует перебор всех пар признаков, назначая каждой паре соответствующий вес, который может изменяться с течением времени.

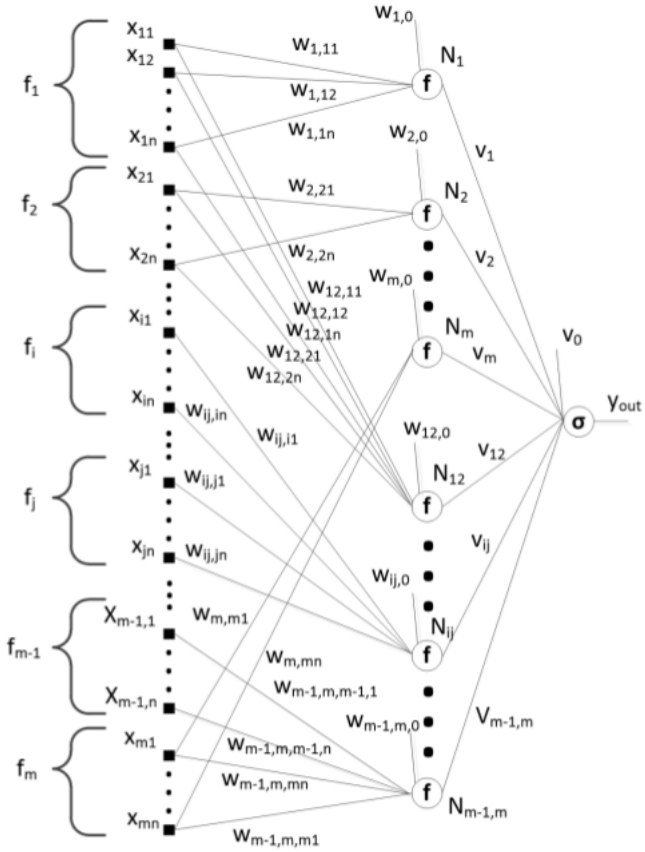


Рис. 4. Схема предложенной нейросетевой модели со специализированной архитектурой

Обучение нейронной сети производится методом стохастического градиентного спуска (SGD) или его модификациями. При обучении используется логистическая функция потерь (логлосс):

$$J(y_{OUT}) = -y_T \cdot \ln(y_{OUT}) - (1 - y_T) \cdot \ln(1 - y_{OUT}),$$

где y_T - истинный тип события (0 для показа и 1 для клика), а y_{OUT} - ответ модели.

В диссертации показано, что минимальное значение данной функции достигается в том случае, когда прогноз частоты клика равен реальной частоте клика $\frac{clicks}{(clicks+shows)}$ на каждой подвыборке с фиксированным признаковым описанием, где shows, clicks - количество показов и кликов на данной подвыборке.

Ниже приведен пример расчета некоторых градиентов. Градиенты весов нейрона выходного слоя:

$$\frac{dJ}{d\nu_i} = \frac{dJ}{dy_{OUT}} \cdot \frac{dy_{OUT}}{dy} \cdot \frac{dy}{d\nu_i} = \frac{y_{OUT} - y_T}{y_{OUT} \cdot (1 - y_{OUT})} \cdot y_{OUT} \cdot (1 - y_{OUT}) \cdot f(N_i)$$

Градиенты весов нейронов скрытого слоя, соединенных со значениями двух признаков:

$$\frac{dJ}{d\omega_{ij_{ik}}} = \frac{dJ}{dy_{OUT}} \cdot \frac{dy_{OUT}}{dy} \cdot \frac{dy}{d\omega_{ij_{ik}}} = (y_{OUT} - y_T) \cdot \nu_{ij} \cdot f'(N_{ij}) \cdot x_{ik} \qquad (1)$$

Обновляются веса только нейронов с ненулевыми входами.

Вышеописанную архитектуру можно расширить до учета троек и бóльшего числа комбинаций признаков. Для выбора наилучшей конфигурации с учетом реальных показателей конкретной системы (требования на время ответа, число запросов, количество и кардинальность признаков и т.д.) в работе предлагается аналитическая модель системы выбора персонализированных предложений как многоканальной системы массового обслуживания с неограниченной очередью. Произведен расчет функции распределения времени пребывания заявки в системе. При этом накладывается требование, чтобы за заданный лимит времени T было обработано 1 - γ запросов. В результате расчетов было получено выражение:

$$e^{-T \cdot \mu} \cdot (1 + c) - c \cdot e^{-T(m \cdot \mu - \lambda)} < \gamma,$$

где λ и μ - интенсивность поступления запросов и обслуживания заявок соответственно; m - количество обслуживающих устройств; $c = \frac{p_w \cdot \mu}{(m-1) \cdot \mu - \lambda}$, p_w - вероятность ожидания в системе. В работе было оценено максимальное время прогнозирования для одного примера и минимальное количество обслуживающих устройств m, необходимых для обеспечения работы системы на требуемом уровне качества, для архитектур с парами и тройками признаков. В результате было показано, что при рассматриваемых параметрах системы выбора персонализированных предложений затраты от использования более медленной архитектуры с учетом троек признаков не покрываются потенциальным приростом от повышения качества работы системы. Также было показано, что создание собственного программного обеспечения оправдано ещё и потому, что реализации предложенной нейросетевой модели со специализированной архитектурой на базе нейросетевых фреймворков не позволяют получить приемлемое время прогнозирования для одного примера.

Также были проведены сравнения различных функций активации в предложенной архитектуре, по результатам которых получилось, что функция активации ReLU дает лучший результат по сравнению с гиперболическим тангенсом (Tanh). Однако, как видно из формулы 1, если аргумент под функцией ReLU становится отрицательным, обновление весов нейрона перестает производиться, поскольку производная от функции ReLU в отрицательной области равняется 0. Как следствие, нейрон уже не может вернуться в работу, из-за чего модель деградирует с течением времени. Для преодоления этого в работе предлагаются две методики: использование функции активации ReLU с утечкой и применение стратегии «выбивания» нейронов. В последнем случае заданной доле неактивных нейронов назначаются веса из равномерного случайного диапазона с определенной периодичностью. Если выведенный из неактивного состояния нейрон так и остался ненужным, он постепенно будет обратно возвращен в неактивное состояние. В экспериментальной части работы показано, что данные методики не ухудшают значения метрик при краткосрочном обучении.

В параграфе 2.7 предлагается методика статистического тестирования для сравнения значений метрик на тестовой выборке. При простом сопоставлении значений метрик остается неясным, является ли наблюдаемое различие реальным, или же имеют место случайные флуктуации. Следует отметить, что здесь речь идет о тестовых выборках из одной генеральной совокупности. Это базовое требование в машинном обучении, равно как и то, что обучающая и тестовая выборка должны быть взяты из одной генеральной совокупности. Поскольку распределения реальных данных многомерны и достаточно сложны по структуре, сравнение моделей и в рамках тестовых выборок из одной генеральной совокупности является важной и не очень простой задачей. По этой причине часто осуществляется кросс-валидация по k блокам, однако данная процедура является трудоемкой, т.к. требуется обучать модель в каждом эксперименте. В диссертации предложена вычислительно более простая методика, однако она применима только для аддитивных метрик (метрика считается аддитивной, если её значение на выборке равняется сумме значений метрики по каждому примеру, входящему в выборку). Логлосс, используемый в работе в качестве основной метрики, является аддитивным.

В предложенной методике тестовая выборка разбивается на N блоков, после чего на каждом блоке вычисляется среднее значение метрики по всем примерам. В результате получаются реализации случайной величины, которая равна сумме большого числа случайных величин (поскольку метрика аддитивна). Согласно центральной предельной теореме сумма большого числа независимых случайных величин имеет нормальное распределение (значения метрик на разных примерах являются независимыми, поскольку каждый пример на тестовой выборке используется только в одном из блоков и не участвует в обучении). Для определения необходимого размера

каждого блока, при котором выполняется условие нормальности, предлагается использовать статистические тесты на нормальность распределения, а также строить график «квантиль-квантиль». Если значения метрики на блоках данных распределены по нормальному закону, можно тестировать гипотезу о равенстве математических ожиданий значений метрик на двух выборках при помощи t теста Стьюдента. Получив значение p_value, меньшее заданного порога, можно отвергнуть нулевую гипотезу с заданным уровнем значимости, тем самым сделав вывод, что наблюдаемое расхождение значений метрик является статистически значимым. Также в работе предложена упрощенная методика сравнения значений метрик, основанная на построении доверительного интервала для математического ожидания значений метрики базовой модели.

В **Главе 3** описаны особенности программной реализации предложенной нейросетевой модели со специализированной архитектурой на языке C++11. Особое внимание уделялось быстродействию. Архитектура нейросети задается статически и хранится в двумерном массиве. Для быстрого вычисления сумм при обработке примеров была создана матрица соединений, в которой каждому номеру признака на входном слое соответствует список связанных номеров нейронов скрытого слоя. Расчет взвешенной суммы для нейрона выходного слоя изображен на Рис. 5. Полученный набор индексов переиспользуется для прогнозирования и обновления. При суммировании сначала рассчитывается вклад во все нейроны скрытого слоя от первого признака, затем - от второго и т.д. Это позволяет получить ускорение за счет локального расположения в памяти слагаемых (весов).

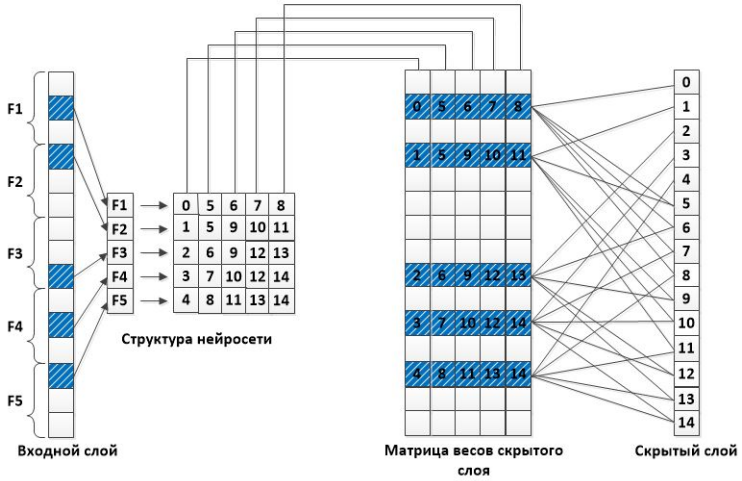


Рис. 5. Организация суммирования для нейронов скрытого слоя

Были проведены замеры времени, в которых сравнивались специализированная реализация логистической регрессии с хешированием комбинаций первичных признаков, вышеописанная реализация предложенной нейросети, а также реализация предложенной нейросети на базе библиотек Lasagne и Pytorch. В экспериментах количество коэффициентов, отводимых на каждый первичный признак при бинарном кодировании, изменялось от 2^8 до 2^{16} (первичные признаки категориальные). Всего использовалось 15 первичных признаков, поэтому общий размер входного вектора изменялся от 4 тыс. до 1 млн. признаков. Измерения проводились для нейросети, перебирающей пары признаков, данная архитектура обозначена как 1x2. Замерялось время получения прогноза и обновления модели для 100 тыс. примеров. Замеры производились на сервере с процессором Intel Xeon CPU E5-2667 3,30GHz и оперативной памятью Micron 36KSF2G72PZ-1 1333 MHz (0,8ns) объемом 256 ГБ. Результаты замеров при работе модели в режиме прогнозирования приведены на Рис. 6, где видно, что разработанная реализация нейросети работает практически так же, как логистическая регрессия и в 2-3 раза быстрее аналогичной реализации на базе Lasagne. Реализация на базе Pytorch при увеличении размера вектора начинает работать значительно медленнее. Таким образом, использование разработанной программной системы имеет смысл в высоконагруженных системах. Для обеспечения параллельной обработки запросов использовались thread_local переменные и thread_specific_ptr указатель из библиотеки boost для хранения полученных значений индексов отдельно для каждого потока. При проектировании программного обеспечения использовался унифицированный язык моделирования UML. Разработанные диаграммы приведены в диссертации.

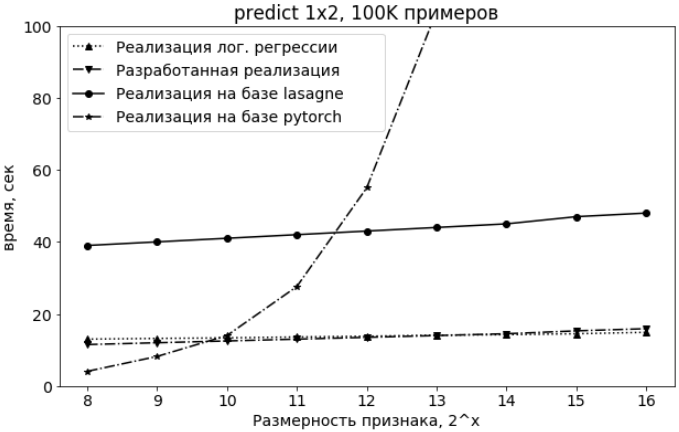


Рис. 6. Сравнение времени прогнозирования разных моделей для 100 тыс. примеров

Глава 4 посвящена экспериментальному оцениванию предложенной нейросетевой модели со специализированной архитектурой в задаче прогнозирования частоты кликов. В проведенных экспериментах качество моделей оценивалось по метрикам логлосс (значение логистической функции потерь) и площадь под ROC кривой (AUC). AUC хорошо подходит для оценки качества решения задач с несбалансированными классами, что имеет место в данном случае. Значение логлосс позволяет определить, насколько хорошо спрогнозированные значения частоты соответствуют реальным значениям.

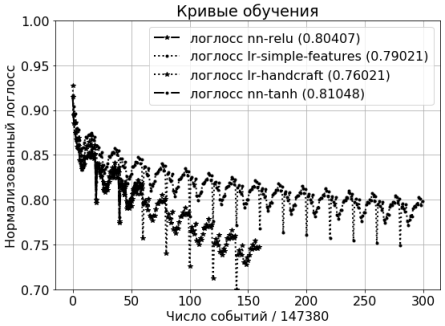
На первом шаге были проведены эксперименты для выбора наилучшего режима работы для каждой модели (разные модификации SGD, однопроходное и многопроходное обучение, использование L2 регуляризации и раннего останова). Затем в наилучшем режиме работы (ранний останов, алгоритм обучения RMSprop) было произведено сравнение предложенной нейросети (nn-relu, nn-tanh) и логистической регрессии с вручную подобранными комбинациями первичных признаков (lr-handcraft) и с набором непосредственно первичных признаков (lr-simple). Также были проведены эксперименты с предложенными стратегиями борьбы с «выключением» нейронов и сравнение различных конфигураций предложенной нейросетевой архитектуры. В качестве примера на Рис. 7 приведены эксперименты с использованием раннего останова (обучение модели прекращалось после того, как ошибка на валидационной выборке начинала возрастать) при обучении алгоритмом RMSprop. Для основных экспериментов было проведено статистическое тестирование согласно предложенной методике. Результаты одного из тестов представлены в Таблице 1.

Интерпретируются полученные результаты следующим образом (на примере сравнения моделей lr-handcraft и lr-simple): вероятность получить наблюдаемую разницу в значениях логлосс моделей lr-handcraft и lr-simple в случае, если они работают одинаково, составляет 0,023. Поскольку логлосс при этом ниже у модели lr-handcraft, можно утверждать, что она работает статистически значимо лучше, чем lr-simple на уровне значимости 0,05. Также видно, что полученный результат сравнения математических ожиданий метрик моделей lr_handcraft, nn_relu не противоречит ну-

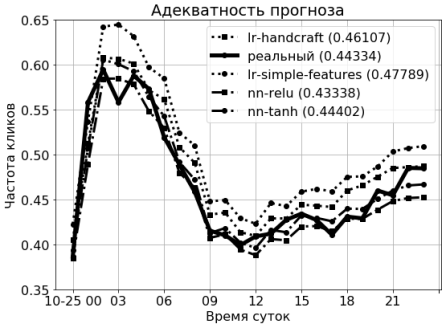
Таблица 1.

Результат сравнения средних значений метрик моделей на тестовой выборке на основе теста Уэлча (обучение с ранним остановом на основе RMSprop)

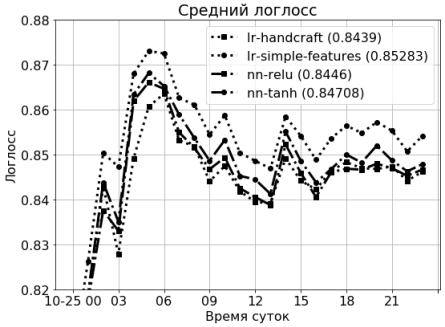
Модель1	Модель2	p_value	Модель1	Модель2	p_value
lr_handcraft	lr_simple	0,023	lr_handcraft	nn_relu	0,67
lr_handcraft	nn_tanh	0,369	lr_simple	nn_relu	0,065
lr_simple	nn_tanh	0,169	nn_relu	nn_tanh	0,639



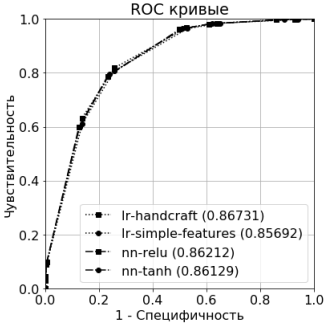
(а) Кривые обучения



(б) Реальные и ожидаемые значения частоты клика



(в) Логистическая функция потерь на тестовой выборке



(г) ROC-кривые и значения площади под ними на тестовой выборке

Рис. 7. Обучение моделей RMSprop с ранним остановом

левой гипотезе. Таким образом, можно принять нулевую гипотезу о том, что логистическая регрессия с вручную подобранными признаками и предложенная нейросетевая модель со специализированной архитектурой с автоматическим конструированием комбинаций признаков для хеширования (в работе показано, что это справедливо и для её модификаций, решающих проблему «выключенных» нейронов) работают примерно с одинаковым качеством. При этом предложенная нейросетевая модель позволяет значительно сэкономить время исследователя, не требуя ручного подбора комбинаций признаков для хеширования.

Общие выводы и заключение

1. Проанализированы методы извлечения признаков и выявлены ограничения их применимости в системе выбора персонализированных предложений. Обоснована необходимость разработки нового метода прогнозирования частоты кликов пользователя по рекламным объявлениям в сети интернет.

2. Разработана нейросетевая модель со специализированной архитектурой, позволяющая решать задачи динамического машинного обучения с конструированием составных признаков в процессе обучения, и описаны особенности её использования.

3. На основе разработанной нейросетевой модели предложен метод прогнозирования частоты кликов пользователей по рекламному объявлению в интернете, не требующий ручного конструирования составных признаков и тем самым позволяющий значительно сократить необходимое время работы экспертов предметной области. Описана методика поиска наилучшей конфигурации в рамках предложенной нейросетевой модели со специализированной архитектурой на основе моделирования системы персонализированных предложений как системы массового обслуживания.

4. Разработаны алгоритмы заполнения матрицы соединений предложенной нейросети; нахождения индексов, участвующих в вычислениях при обработке примера; расчета прогнозируемой частоты клика и обновления модели при обработке одного примера. На основе разработанных алгоритмов спроектировано и реализовано программное обеспечение, позволяющее сократить время прогнозирования и обучения на одном событии в 2-3 раза по сравнению с реализациями на базе нейросетевых библиотек. Получено авторское свидетельство о регистрации программы для ЭВМ.

5. Разработана методика статистического тестирования для сравнения значений аддитивных метрик, полученных различными моделями на фиксированной тестовой выборке. Показано, что применение данной методики в вычислительном плане примерно в k/2 раз проще, чем использование традиционной кросс-валидации по k блокам (где k обычно выбирается порядка 10-20).

6. Проведена апробация предложенной нейросетевой модели и метода прогнозирования частоты кликов пользователя по рекламному объявлению в интернете на данных компании Mail.Ru Group в режиме исследований и на реальной системе. Показано, что при использовании предложенной модели достигаются аналогичные значения метрик по сравнению с моделью, использующей вручную подобранные производные признаки (p_value составило 0,67, что не противоречит нулевой гипотезе о равенстве средних значений метрик обеих моделей). Следовательно, благодаря предложенной модели устранена необходимость ручного построения производных признаков.

Таким образом, поставленные в работе задачи решены, а цель достигнута. Внедрение результатов диссертации позволит избавиться от необходимости формирования производных признаков для модели, что высвободит кадровые ресурсы и повысит эффективность работы подразделения, осуществляющего построение модели прогнозирования частоты кликов пользователей.

Публикации автора по теме диссертации

В рецензируемых изданиях из перечня ВАК РФ

1. Федоренко Ю. С. Методика статистического тестирования для сравнения качества работы моделей машинного обучения // Вестник компьютерных и информационных технологий. 2019. № 12. С. 10—17. (0,5 п.л.)
2. Федоренко Ю. С. Проектирование быстрой программной реализации специализированной нейросетевой архитектуры с разреженными связями // Программные продукты и системы. 2019. № 4. С. 639—649. (0,7 п.л.)
3. Федоренко Ю. С., Гапанюк Ю. Е. Кластеризация данных на основе самоорганизующихся растущих нейронных сетей и марковского алгоритма кластеризации // Нейрокомпьютеры: разработка, применение. 2016. № 4. С. 3—13. (0,6 п.л./0,5 п.л.)
4. Федоренко Ю. С., Гапанюк Ю. Е. Анализ особенностей глубоких нейронных сетей на примере задачи распознавания цифр // Нейрокомпьютеры: разработка, применение. 2017. № 2. С. 24—30. (0,5 п.л./0,4 п.л.)

В изданиях, входящих в международную базу цитирования Scopus

5. Fedorenko Y. S. Using a Sparse Neural Network to Predict Clicks Probabilities in Online Advertising // Advances in Neural Computation, Machine Learning, and Cognitive Research IV. 2020. Vol. 925. (Springer, Cham). P. 276—282. (0,3 п.л.)
6. Fedorenko Y. S., Chernenkiy V. M., Gapanyuk Y. E. The Neural Network for Online Learning Task Without Manual Feature Extraction // Advances in Neural Networks. 2019. Vol. 11554. (Springer, Cham). P. 67—77. (0,5 п.л./0,3 п.л.)
7. Fedorenko Y. S., Gapanyuk Y. E. The Neural Network with Automatic Feature Selection for Solving Problems with Categorical Variables // Advances in Neural Computation, Machine Learning, and Cognitive Research. 2018. Vol. 799. (Springer, Cham). P. 129—135. (0,3 п.л./0,2 п.л.)
8. Fedorenko Y. S., Gapanyuk Y. E., Minakova S. V. The Analysis of Regularization in Deep Neural Networks Using Metagraph Approach // Advances in Neural Computation, Machine Learning, and Cognitive Research. 2017. Vol. 736. (Springer, Cham). P. 3—9. (0,4 п.л./0,2 п.л.)
9. Fedorenko Y. S., Gapanyuk Y. E. Multilevel neural net adaptive models using the metagraph approach // Optical Memory and Neural Networks. 2016. Vol. 25. Issue 4.. С. 228—235. (0,5 п.л./0,3 п.л.)