

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»

Кафедра социальной информатики

**МОДЕЛИРОВАНИЕ НЕЙРОННОЙ СЕТИ ДЛЯ РЕШЕНИЯ
ЗАДАЧ КЛАССИФИКАЦИИ И ВОЗМОЖНОСТИ ЕЕ
ПРИМЕНЕНИЯ В СОЦИОЛОГИИ**

(автореферат бакалаврской работы)

студента 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Газинского Игоря Владимировича

Научный руководитель
кандидат физико-математических наук, доцент _____ Л.Б. Тяпаев
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2022

ВВЕДЕНИЕ

Актуальность проблемы: решения с помощью искусственных нейронных сетей в последнее время востребованы по всему миру, а интерес к ним только возрастает, поэтому считаю выбранную тему особенно актуальной для изучения.

Объект исследования: массивы данных.

Предмет исследования: классификация данных с помощью нейронной сети.

Цель: разработка искусственной нейронной сети, решающей задачу классификации на основе данных, предварительно обработанных в программе статистического анализа SPSS с помощью факторного анализа, а также анализа возможностей применения сетей в социологии.

Задачи работы:

1. Проанализировать предметную область;
2. Изучить теоретические основы классификации с помощью ИНС;
3. Обработать данные для использования их в задаче классификации;
4. Спроектировать архитектуру нейронной сети;
5. Провести тестирование;
6. Раскрыть возможности использования ИНС в социологии.

Теоретическая значимость исследования заключается в возможности использования основных положений и выводов данной работы для дальнейшего изучения выбранной проблематики.

Структура бакалаврской работы. Выпускная квалификационная работа состоит из введения, двух глав по 4 и 3 параграфа соответственно, заключения, списка использованных источников и приложений.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе производится обзор предметной области.

Искусственные нейронные сети по своему строению напоминают биологические нейронные сети, представленные неким количеством,

соединенных между собой, нервных клеток, которые выполняют какие-либо физиологические функции.

Если рассматривать искусственные нейронные сети, то нейроном в них будет нелинейная функция, аргументом которой будет являться линейная зависимость, объединяющая все входные сигналы, такой аргумент называют функцией активации, результат этой функции идет на аксон, одни объединенные такие нейроны с другими и называют искусственной нейронной сетью. Связи в искусственной нейронной сети между нейронами имеют определенные веса, когда сигнал проходит по определенной связи, то он умножается на этот вес. Работа нейронной сети напрямую зависит от функции активации нейронов, она отражает зависимость сигнала на аксоне от суммарных сигналов на входах нейрона.

Выделяют три основные функции активации:

1) Сигмоидальная функция

$$\sigma(x) = 1/(1+e^{-x})$$

Также называется логистической, т.к. является нелинейной, ее могут использовать в многослойных сетях, используя алгоритм обратного распространения ошибки, ограничена нулем и единицей.

2) Кусочно-линейная функция

$$f(x) = \begin{cases} 1, & x \geq +\frac{1}{2}; \\ |x|, & +\frac{1}{2} > x > -\frac{1}{2}; \\ 0, & x \leq 0 - \frac{1}{2}; \end{cases}$$

3) Функция Хевисайда

$$f(x) = \begin{cases} 1, & \text{если } x \geq 0; \\ 0, & \text{если } x < 0; \end{cases}$$

В итоговой модели будет использоваться Сигмоидальная функция и ReLU, она расшифровывается как Rectified Linear Unit, она возвращает ноль, если принимает отрицательный аргумент, а при положительном принимает само значение аргумента.

$$f(x) = \max(0, x)$$

Выделяют 6 типов задач, которые решают искусственные нейронные сети:

1) Распознавание образов и классификация.

В задачах распознавания и классификации за образы принимают любые сущности, они могут быть текстом, изображениями, аудио и т.д. В процессе обучения участвуют заранее подготовленные образцы, соотнесенные к какому-либо классу, сеть постепенно учится относить их сама по типовым признакам. Суммарно все признаки должны определять конкретный класс, к которому необходимо отнести образ. При постановке такой задачи может возникнуть проблема, что при недостаточности типовых признаков сеть может отнести образ к более, чем одному классу одновременно. По завершении обучения можно подавать в сеть ранее нераспознанные объекты, сеть в свою очередь будет их классифицировать.

Структура сети будет отличаться тем, что количество классов, к которым необходимо отнести образы должно совпадать с количеством нейронов в выходном слое. Во время того, как в сеть подает какой-либо объект, на определенном выходе должен появиться признак, который определит класс, к которому этот объект относится, а на других признаках, который указывает на то, что к этим классам объект не принадлежит.

2) Принятие решений и управление.

Задача принятия решений и управления напоминает собой задачу распознавания, в ней классифицируют ситуацию, признаки которой подаются на вход искусственной нейронной сети, на выходе появляется типовой признак, отражающий решение, принятое сетью. Признаки на входе, в свою очередь, отражают описание состояния системы управления.

3) Кластеризация.

Под кластеризацией понимается классификация без учителя, при которой нет выборки, содержащей примеры, соотнесенных объектов к классам для обучения. Пример сетей, решающих данную задачу, являются нейронные сети Кохонена.

4) Аппроксимация функций.

Искусственные нейронные сети можно использовать для аппроксимации функции непрерывного автомата.

5) Прогнозирование.

При прогнозировании в задаче участвуют дискретные отсчеты в определенные моменты времени, по итогу прогнозируется значение в какой-либо момент времени.

б) Сжатие данных и ассоциативная память.

За счет того, что искусственные нейронные сети способны отражать связь между нейронами, можно спроецировать большие массивы данных намного компактнее. Ассоциативной памятью же называют реверсивный процесс, когда с некоторого набора частей данных можно восстановить изначальный объект.

Процедура обучения искусственной нейронной сети затрагивает значения весов связей, изменяя их для того, чтобы минимизировать фактическую ошибку между ожидаемым и наблюдаемым результатом работы ИНС.

В общем случае процедура проходит в несколько этапов:

- 1) Формируются выборки для обучения, состоящая из пар векторов входа и векторов выходов, под ними подразумеваются желаемые результаты.
- 2) Выбирается пара, также, из вектора входа и выхода.
- 3) Подается на вход искусственной нейросети.
- 4) Определяется различие между ожидаемым и наблюдаемым результатом.
- 5) Веса изменяются для минимизации ошибки.
- 6) Неоднократно повторяются шаги со второго по пятый для достижения желаемой величины ошибки, по достижению сеть считается обученной, и тогда в сеть подается вектор, отсутствующий в изначальной выборке.

Во второй главе «Построение ИНС для решения задачи классификации» описывается прикладная часть исследования.

Как уже отмечалось, для корректного обучения нейросети желателен

большой объем выборки для ее обучения, исследование возможностей работы искусственной нейронной сети происходило на примере реальной базы данных, собранной по известному происшествию, связанному с крушением корабля “Титаник”. База данных лежит в открытом доступе и содержит в себе несколько сотен строк с информацией о пассажирах корабля. Столбцы включают в себя пол, возраст, фамилию, имя, отчество, класс поездки, стоимость билета, порт посадки, кол-во детей или родственников на борту и кол-во супруг или братьев, а также информацию о том, выжил пассажир или нет. Цель: обучение нейронной сети корректно классифицировать данные, определяя, кто из пассажиров выжил, а кто нет по некоторому набору входных данных. Для минимизации ошибки строки с недостающими данными были подвергнуты исключению, а также, так как стоимость билета напрямую отражает класс поездки, было принято решение исключить и его из выборки для искусственной нейронной сети.

Также были выделены основные факторы, влияющие на выживаемость с помощью факторного анализа, а именно метода главных компонент, в программе статистического анализа SPSS, в результате факторного анализа основными признаками были выделены: пол, стоимость билета, возраст и класс поездки, так как стоимость билета напрямую отражает класс поездки, было принято решение исключить его из выборки для искусственной нейронной сети.

Для построения ИНС мной был выбран язык Python, а среда разработки – Pycharm v3.7.

Были использованы такие библиотеки как: keras, numpy, pandas, matplotlib и tensorflow.

В качестве входных данных были взяты возраст, пол и класс поездки, в качестве выходных информация о том, выжил ли пассажир. Все данные были закодированы значениями от 0 до 1 для корректного восприятия нейронной сетью. Функция активации на первом слое с входными данными была выбрана RELU функция, на втором слое с выходным - сигмоида. Функция потерь изначально была выбрана mse, это среднеквадратичная ошибка, а в качестве ее оптимизатора sgd, стохастический градиентный спуск, так как данных не так

много, ИНС обучалась и тестировалась на одной выборке, сперва в количестве эпох (циклов), равных 200 и затем 2000, при такой комбинации точность не поднималась выше 0,72, что не самый лучший результат. Было принято решение изменить функцию потерь на бинарную кроссэнтропию, а в качестве оптимизатора взять adam, адаптивную оценку момента. По итогу при эпохах в количестве 2000 итераций, мною была достигнута точность в размере 0.9, что уже является показателем хорошей модели.

Далее, идет описание возможностей применения сетей в социологии и пара примеров исследований, с их использованием. Так, за полтора года до выборов в 2008 году была спрогнозирована победа Дмитрия Медведева пермскими студентами, взяв за основу их программу “Нейросимулятор 5.0”, а также выборку, которую они использовали при обучении сети, модель была применена на базе данных, созданной по недавно прошедшим выборам во Франции, модель успешно справилась с задачей и отдала победу Макрону.

Также, была рассмотрена возможность классифицировать тональность текста с помощью сверточных сетей, Возможность таких исследований открывает новые возможности для социологов, так, можно будет определить мнение по какой-либо теме посредством запуска такой сети для обработки массива сообщений на тематическом ресурсе, люди, зачастую, в социальных сетях, даже без вопроса, готовы сами рассказать намного больше, чем если бы во с ними проводили анкетирование, как очное, так и заочное, к тому же, это будет в разы менее затратно, как в отношении денежного эквивалента, так и затраченного времени.

ЗАКЛЮЧЕНИЕ

Нейронные сети — достаточно старое направление сферы IT, которое начало стремительно развиваться относительно недавно. Несмотря на то, что с момента их проектирования прошло уже много лет, нейронным сетям пришлось долго ждать технических решений для их реализации. Однако в последнее время наблюдается стабильный рост интереса к технологии.

Современные компании все чаще прибегают к нейросетевым технологиям для анализа данных для улучшения связи со своей аудиторией, привлечения новых клиентов, прогнозирования деятельности, поиска решений, стоящих перед ними задач, анализа конкретных областей. На сегодняшний день, нейросетевые технологии используются множеством компаний, владеющих необходимыми навыками их применения, используют их как в профессиональной сфере, так и для развлечений. И хотя применение нейросетевого инструментария на практике достаточно эффективно с точки зрения анализа данных, но не всегда оправдано их применение относительно возможной ручной работы. Поэтому можно сделать вывод, что не каждая задача может быть эффективно решена с помощью нейросети, в основном, этим занимаются крупные успешные компании, которые могут себе позволить опытных специалистов. Хотя в условиях современной рыночной конкуренции применение методов нейросетевого анализа данных может дать значительные преимущества перед компаниями-конкурентами, вскрыв какие-либо новые детали.

Что касается российского рынка, присутствие на нем компаний или же групп специалистов, использующих в своих продуктах или исследованиях нейросетевые технологии, остается весьма небольшим. Хотя перспективы для освоения этих технологий, безусловно есть, инициативные исследователи, например, приведенные в пример во второй главе, которые активно внедряют нейросетевые технологии в свою деятельность.

Стоит отметить, что нейронные сети смогли реализовать свои способности и раскрыть свой потенциал относительно недавно, но уже успели вызвать интерес как с академической, так и коммерческой стороны, поскольку достижения в области технологий анализа огромных массивов данных и их способности к решению различного рода задач сделали их эффективным инструментом в руках профессионалов.

Исходя из проанализированной в этой работе литературы и примеров использования нейросетевых технологий в исследованиях, можно сделать

выводы относительно современных методов обработки данных с их применением. Процессы, проходящие в искусственной нейронной сети спроектированы аналогично процессам, проходящим внутри человеческого мозга, зная механизмы этих процессов, возможно в будущем, при совершенствовании технологического прогресса, специалисты смогут и дальше обновлять и делать лучше технологии нейронных сетей, делая их аналогичными или даже лучше их естественного предка. Но даже сейчас, они успешно справляются и помогают, как сократить время исследования, так и глубже проанализировать какую-либо область.

Раскрывая возможности, конкретно для социологии, можно, с уверенностью, сказать о необходимости внедрения ее в протекающие процессы. Только существующие сети, созданные для определения тональности текста открывают массу новых возможностей, в интернете есть множество ресурсов, анализ которых будет занимать много времени, а с помощью программного решения можно анализировать любое множество сайтов параллельно, количество которых может быть ограничено лишь техническими характеристиками системы, на которой будет запущено программное обеспечение. Хотя в работе не были приведены в пример сети, которые анализируют графические источники, но и они могут активно применяться в социологии, обучив такую сеть на определенных данных, можно будет определять, например, контингент, посетивший какое-либо событие или мероприятие, характеризуемый по возрасту или полу, в настоящее время пришлось бы проводить опрос или просматривать весь хронометраж записанного на камеру видеоматериала, нейросеть же смогла бы определить все в режиме online, гораздо сильнее сократив временные затраты на эту задачу.