УДК 004.85

МАШИННОЕ ОБУЧЕНИЕ В ИНЖЕНЕРНЫХ ИССЛЕДОВАНИЯХ

Бутенко Анатолий Иванович
доктор сельскохозяйственных наук, профессор
but_tolik@mail.ru,
Мичуринский государственный аграрный университет
г. Мичуринск, Россия

Аннотация. В статье описываются два метода машинного обучения – «Случайный лес» и «Искусственные нейронные сети» для решения задач классификации и регрессии.

Ключевые слова: машинное обучение, случайный лес, нейронные сети.

В настоящее время всё чаще сообщается, что какие-то задачи решают с помощью искусственного интеллекта. Когда говорят об искусственном интеллекте, часто имеют в виду искусственную нейроннаю сеть, но понятия искусственный интеллект и нейронные сети не тождественны. Дать строгое определение искусственного интеллекта, понятное неспециалисту трудно, хотя всем ясно, что речь идет о создании машин, которые бы выполняли функции, требующие интеллектуальности при их выполнении людьми.

В 1950 году английский математик Алан Тьюринг в работе [1] предложил поведенческий тест для проверки интелектуальности машины. Эксперт в течение 5 минут общается с некоторым собеседником (может быть переписка) и должен решить, машина это была или человек. Тест Тьюринга считается пройденым искусственным интелектом, если 30% экспертов решат, что общались с человеком. С. Рассел и П. Норвиг [2] приводят следующие шесть пунктов, которым должна удовлетворять компьютерная программа, чтобы пройти тест Тьюринга:

1. Средства обработки текстов на естественных языках, позволяющие успешно общаться с компьютером.
2. Средства представления знаний, с помощью которых компьютер может записать в память то, что он узнает или прочитает.
3. Средства автоматического формирования логических выводов, обеспечивающие возможность использовать хранимую информацию для поиска ответов на вопросы и вывода новых заключений.
4. Средства машинного обучения, которые позволяют приспосабливаться к новым обстоятельствам, а также обнаруживать и экстраполировать признаки стандартных ситуаций.
5. Машинное зрение для восприятия объектов.
6. Средства робототехники для манипулирования объектами и перемещения в пространстве.

Далее они отмечают, что перечисленные пункты составляют основную часть направлений исследований по искусственному интеллекту.

Мы хотим подробнее остановиться на пункте 4 – машинном обучении (machine learning). Это направление в настоящее время стало прикладным и широко используется для решения сложных задач в различных областях.

Различают три вида машинного обучения: обучение с учителем, обучение без учителя и глубокое обучение [3-5]. Глубокое обучение предполагает анализ Big Data – настолько большого массива информации, что одного компьютера будет недостаточно. Нейронные сети позволяют разделить одну большую задачу на несколько маленьких и распределить по нескольким компьютерам. Такие задачи, конечно, решает не отдельный человек, а коллектив исследователей. К обучению без учителя относят такие статистические методы, как, например, кластерный анализ или метод k-ближайших соседей. Эти методы использовались раньше, чем обозначилось направление - машинное обучение, и они не связаны со спецификой указанного направления [4, 6]. Подробнее рассмотрим обучение с учителем. Обычно этим методом решают задачу классификации или регрессии. Для обучения классификации нужно иметь достаточно большую по объему выборку значений некоторых признаков x_i у объектов и номера классов y_j, которым принадлежат данные объекты. У каждого объекта может быть несколько признаков x_i (вектор признаков) и только одно значение y_j. Данные разбиваются на две части - обучающую и проверочную выборки. По обучающей выборке программа устанавливает такие параметры модели, чтобы ошибка к отнесению класса на этой выборке было минимальной, а когда такая модель будет построена, она проверяется на проверочной выборке. Если там классификация будет хорошей, значит модель работающая, в противном случае подбирают другие признаки или другие данные. При решении задачи регрессии y_j будут содержать значения некоторой функции от x_i вместо номеров классов. Для обучения управлению механизмами значение y_j может быть некоторыми кодами для управляющих воздействий механизма, например беспилотника, а x_i - это признаки, выработанные программой по видеопоследовательности. В

зависимости от каких-то ситуаций будут вырабатываться те или иные действия механизмов [4, 7, 8].

Рассмотрим подробнее два метода машинного обучения - «случайный лес» и «исскуственные нейронные сети».

Метод «Случайный лес» (Random Forest) был предложен в 2001 году Лео Брейманом и Адель Катлер [3, 9] для решения задач классификации и регрессии.

Для классификации в этом методе используются «деревья решений», каждое из которых рекурсивно формируется по данным обучающей выборки с помощью «жадной» процедуры, на каждом шаге максимально уменьшая значение математической функции, описывающей неоднородность данных, содержащихся в узлах анализируемых «деревьев решений». Есть несколько функций, описывающих неоднородность, например, индекс Джини $f=\sum_{k=1}^{m} p_k(1-p_k)$, где p_k-эмпирическая вероятность встретить метку k–го класса в указанном множестве [2, 10]. На каждом шаге находится некоторый признак x_i и константа c_i такие, что объекты обучающей выборки, для которых выполняется неравенство $x_i<c_i$, имеют метки одного класса, например, первого. В области, задаваемой этим неравенством, функция однородности будет минимальна и все объекты, попадающие в данную область, будут относиться к первому классу. Остальная область, где не выполняется указанное неравенство, имеет высокую неоднородность и компьютер находит какой-то другой признак x_j и соответствующую константу c_j , что в области выполнения неравенства $x_j<c_j$ будут объекты какого-то другого класса, например, второго. Такая повторяющаяся процедура выглядит как дерево с узлами в виде неравенств (рис.1).

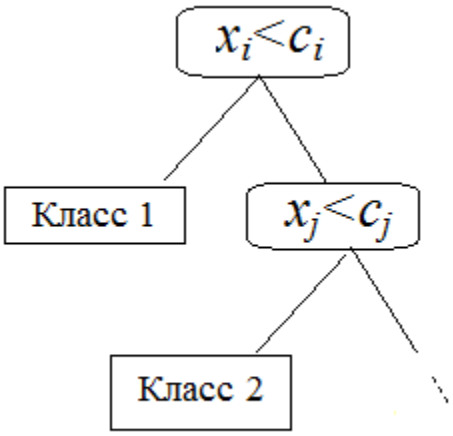

Рисунок 1 – Дерево решений

С помощью дерева мы наглядно можем представить, какие логические процедуры применяет компьютер при разделении области, описываемой обучающей выборкой, на заданные классы. Количество узлов, то есть, цепочка неравенств определяет высоту конкретного дерева решения. После достижения заданной высоты программа прекращает дальнейшее формирование дерева.

Даже на основе одного дерева решений можно проводить классификацию, но алгоритм «Случайный лес» уникален тем, что позволяет значительно улучшить результаты, используя бэггинг (bagging) – специально организованный «ансамбль деревьев решений». В статистике известен метод «размножения» данных - бутстрэп (bootstrap), когда из выборки объема N можно построить большое количество новых выборок того же объема, выбирая случайным образом элементы из исходной выборки. При таком генерировании, в каждой выборке может оказаться много одинаковых объектов, но … состав выборок будет различен.

Поскольку принадлежности объектов к классам в обучающей выборке заведомо известны, то на компьютере может независимо строиться множество разных «деревьев решений». Они дают немного различающиеся между собой результаты классификации [1, 10, 11]. В итоге принимается решение путем «голосования», то есть для каждого объекта выбирается тот класс, на который укажет большинство деревьев, а параметры подбираются так, чтобы была

минимальная ошибка отнесения объектов к заданным классам, на примерах представленных обучающей выборкой.

В конце обучения программа выдает ошибки на обучающей и тестовой выборках. Каждая из них есть отношение количества объектов соответствующей выборки, ошибочно классифицированных, к объему данной выборки. Кроме ошибок распознавания программа выдает относительные важности признаков, которые в сумме дают единицу. Их можно использовать при пересмотре признаков. Если у какого-то признака относительная важность близка к нулю, то его можно отбросить.

С помощью нейронных сетей можно также решать задачу обучения с учителем. Под нейронными сетями подразумеваются адаптируемые и обучаемые распараллеленные вычислительные структуры, состоящие из большого числа элементарных преобразователей - «искусственных нейронов» [12]. В состав нейрона входят умножители (синапсы), сумматор и нелинейный преобразователь. Синапсы осуществляют связь между нейронами и умножают входные сигналы на веса, характеризующие силу связи. Сумматор выполняют сложение сигналов, поступающих по синаптическим связям от других нейронов, и внешних входных сигналов. Нелинейный преобразователь реализует нелинейную функцию одного аргумента — выхода сумматора. Эта функция называется функцией активации или передаточной функцией нейрона. Математическая модель нейрона описывается соотношением

$$y = f(\sum_{i=1}^{n} w_i x_i + b),$$

где w_i– вес синапса; b — значение смещения; x_i - компонент входного вектора (входной сигнал); y - выходной сигнал нейрона, n - число входов нейрона; f - нелинейное преобразование (функция активации или передаточная функция).

Синаптические связи с положительными весами называют воэбуждающими, с отрицательными весами — тормозящими. Передаточная функция выбирается из заданного стандартного набора функций, форма графика у которых немного различается. Среди них могут быть ступенчатые, с углами и гладкие кривые, но общим для них характерно резкое возрастание на

небольшом интервале от постоянного низкого значения (обычно равного (-1) или 0) до постоянного высокого значения (обычно равного 1). Низкое значение соответствует невозбужденному нейрону, а высокое - возбужденному. Как только суммарный сигнал и смещение достигнут пороговой величины, произойдет триггерный эффект возбуждения нейрона.

Искусственная нейронная сеть представляет собой набор соединенных между собой нейронов. Передаточные функции всех нейронов в сети обычно фиксированы, а веса являются параметрами сети и могут изменяться.

Существует много видов нейронных сетей, но, к счастью, необязательно придумывать нейросеть «с нуля» существует несколько десятков различных нейросетевых архитектур, причем эффективность многих из них доказана математически. Чаще всего рассматривается многослойная сеть, у которой первый слой принимает вектор входных сигналов, последний слой формирует выход, а внутренние слои формируют преобразования входных сигналов в выходные.

Обучение нейронной сети происходит также как и методом «случайной лес» - на вход подаются некоторые сигналы, которым соответствуют изображения некоторых объектов и нужно правильно классифицировать эти объекты. Например, мы хотим научиться различать рукописные буквы русского алфавита. Если на вход подаётся изображение некоторой буквы, например, «А», то на выходе должен сформироваться максимальный сигнал у метки, соответствующей этой букве, а другие сигналы должны быть значительно ниже [1, 3, 9]. Вычисляя разность между желаемым и реальным ответами, получаем «вектор ошибки». Алгоритм обучения состоит в том, чтобы эту ошибку уменьшить. Оказывается, что после многократного предъявления примеров веса сети стабилизируется, причём сеть дает правильные ответы почти на все примеры из базы данных обучающей выборки.

В многослойных сетях оптимальные выходные значения нейронов всех слоев, кроме последнего, как правило, неизвестны, и такую сеть невозможно обучить, руководствуясь только величинами ошибок на выходах сети.

Наиболее приемлемым вариантом обучения в таких условиях оказался градиентный метод поиска минимума функции ошибки с рассмотрением сигналов ошибки от выходов НС к ее входам, то есть в направлении, обратном прямому распространению сигналов в обычном режиме работы. Этот алгоритм обучения НС получил название процедуры обратного распространения.

Алгоритм действует циклически (итеративно), и его циклы принято называть эпохами. На каждой эпохе на вход сети поочередно подаются все обучающие наблюдения, выходные значения сети сравниваются с целевыми значениями и вычисляется ошибка [12]. Значения ошибки, а также градиента поверхности ошибок используются для корректировки весов, после чего все действия повторяются. Начальная конфигурация сети выбирается случайным образом, и процесс обучения прекращается либо когда пройдено определенное количесгво эпох, либо когда ошибка достигнет некоторого определенного уровня малости, либо когда ошибка перестанет уменьшаться (пользователь может сам выбрать нужное условие остановки) [6, 11, 12].

Сравнивая рассмотренные два метода машинного обучения, следует отметить, что «случайный лес» использовать проще, так как в нем меньше настраиваемых параметров. В нем также не нужно нормировать признаки. Поэтому начинать машинное обучение для решения конкретной задачи лучше с метода «случайный лес», а если не удасться достигнуть нужной точности, то перейти к нейронным сетям.

Оба метода в виде команд содержаться в свободно распространяемой библиотеке компьютерного зрения OpenCV. В этой библиотеке содержится обширный набор команд для цифровой обработки изображений. Все команды приводятся в двух вариантах: для C++ и для python, поэтому управляющую программу нужно писать на одном из этих языков программирования.

Среди математических пакетов расширения MATLAB есть один специально по сетям - Neural Networks Toolbox.

Список литературы:

1. Turing A. Computing Machinery and Intelligence// Mind, 1950, p. 433–460. (Русский перевод в книге А. ТЬЮРИНГ. МОЖЕТ ЛИ МАШИНА МЫСЛИТЬ? – М.: ФИЗМАТЛИТ, 1960.- 67 с.).
2. Рассел, Стюарт, Норвиг, Питер. Искусственный интеллект: современный подход, 2-е изд..: Пер. с англ. — М. : Издательский дом “Вильямс”, 2006. — 1408 с.
3. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных/ П. Флах. – М.:ДМК Пресс, 2015. – 400 с.
4. Breiman L. Random forests/ L. Breiman // Machine Learning. — 2001. — Vol. 45, no. 1. — Pp. 5–32.
5. Комплекс машин для маточников вегетативно размножаемых подвоев и интенсивного сада / А.И. Завражнов, К.А. Манаенков, В.Ю. Ланцев, В.В. Хатунцев и др. //Достижения науки и техники АПК. - 2009. - № 1. - С. 49-52.
6. Бросалин В.Г. Механизация отделения отводков клоновых подвоев яблони / В.Г. Бросалин, К.А. Манаенков // Вестник Мичуринского государственного аграрного университета. - 2012. - № 3. - С. 198-205.
7. Использование возможностей языка r для реализации алгоритмов машинного обучения в среде MS SQL SERVER 2019 / А.А. Крумкаченко, Д.В. Косенков, В.В. Гавриков, Р.Н. Абалуев // Наука и Образование. – 2020. – Т. 3. – № 2. – С. 2.
8. Средства создания web-сайтов / В.И. Назаров, Н.В. Картечина, Н.В. Пчелинцева, Р.Н. Абалуев // Наука и Образование. – 2020. – Т. 3. – № 2. – С. 27
9. Проектирование и реализация интерактивной специализированной информационно-справочной системы / С.В. Федоров, И.В. Уколов, А.А. Лукин, И.А. Лунев [и др.] // Наука и Образование. – 2020. – Т. 3. – № 2. – С. 3
10. Абалуев, Р.Н. Обзор современных подходов к обеспечению информационной безопасности при создании инфраструктуры интернета вещей

в агропромышленном комплексе / Р.Н. Абалуев, А.А. Крумкаченко // Наука и Образование. – 2019. – Т. 2. – № 2. – С. 289.

11. Абалуев, Р.Н. Перспективы использования аддитивных технологий в агропромышленном комплексе / Р.Н. Абалуев, С.О. Чиркин // Наука и Образование. – 2019. – Т. 2. – № 2. – С. 311.

12. Shcherbakov S.Yu. Drying hawthorn berries in drum dryer using blade agitator / S.Yu. Shcherbakov, P.S. Lazin, I.P. Krivolapov // Amazonia Investiga. - 2019. - Т. 8. - № 21. - С. 588-595.

UDC 004.85

MACHINE LEARNING IN ENGINEERING RESEARCH

Butenko Anatoly Ivanovich
Doctor of Agricultural Sciences, Professor
but_tolik@mail.ru
Michurinsk State Agrarian University
Michurinsk, Russia

Annotation. The article describes two methods of machine learning - "Random forest" and "Artificial neural networks" for solving problems of classification and regression.

Key words: machine learning, random forest, neural networks.