1 Общие положения

В дипломной работе рассматривается задача распознавания модели автомобиля по фотографии и, хотя на данный момент уже существуют решения в данной области, до сих пор остается не изученной проблема дообучения таких систем. Базовый подход заключается в применении методов машинного обучения (ML — Machine Learning), таких как нейронные сети (NN — Neural Network, или ANN — Artificial Neural Network) и деревья принятия решений (decision tree).

2 Машинное обучение

Машинное обучение можно определить как математическую дисциплину, или набор математических, статистических, численных методов и различных подходов из теории вероятностей для решения задачи автоматического выделения информации (information extraction) и ее интеллектуального анализа (data mining) из неструктурированного набора входных данных. Более подробно — ставится задача сопоставления заданному значению (объекту) некоторого действия (ответа), таким образом можно говорить об обобщении задачи аппроксимации функции.

Машинное обучение активно используется при классификации данных, например, распределение новостных статей по рубрикам. Отдельное внимание уделяется обработке изображений, их автоматическому аннотированию [1] или выделению одного определенного объекта на изображении, задача распознавания рукописного текста (OCR — Optical Character Recognition).

Основные подходы машинного обучения при работе с изображениями — это использование нейронных сетей и деревьев принятия решений. Оба метода занимаются выделением опорных признаков на обрабатываемой картинке, после чего по ним выполняют классификацию. Перед использованием построенной модели ее необходимо обучить на тестовом наборе данных, выработать базу признаков. Алгоритмы обучения решают задачу оптимизации, занимаются подбором неизвестных коэффициентов в математической модели описывающей конкретный метод машинного обучения. Процесс обучения требует большого количества вычислительных операций, что заставляет ис-

кать альтернативные подходы при работе с предметной областью, в которую могут добавляться новые классы объектов. Вместо полного переобучения модели используют алгоритмы частичной корректировки коэффициентов (fine tuning). Такой подход дообучения требует меньше времени, но связан с потерей качества, так как не добавляет в модель признаков объектов новых классов.

3 Деревья принятия решений

Дерево принятия решений является графом, в узлах которого записаны условия определяющие переход в следующую вершину, а в листьях хранятся значения целевой функции (рисунок 1). Таким образом, решение задачи сводится к спуску из корня дерева до его листа.

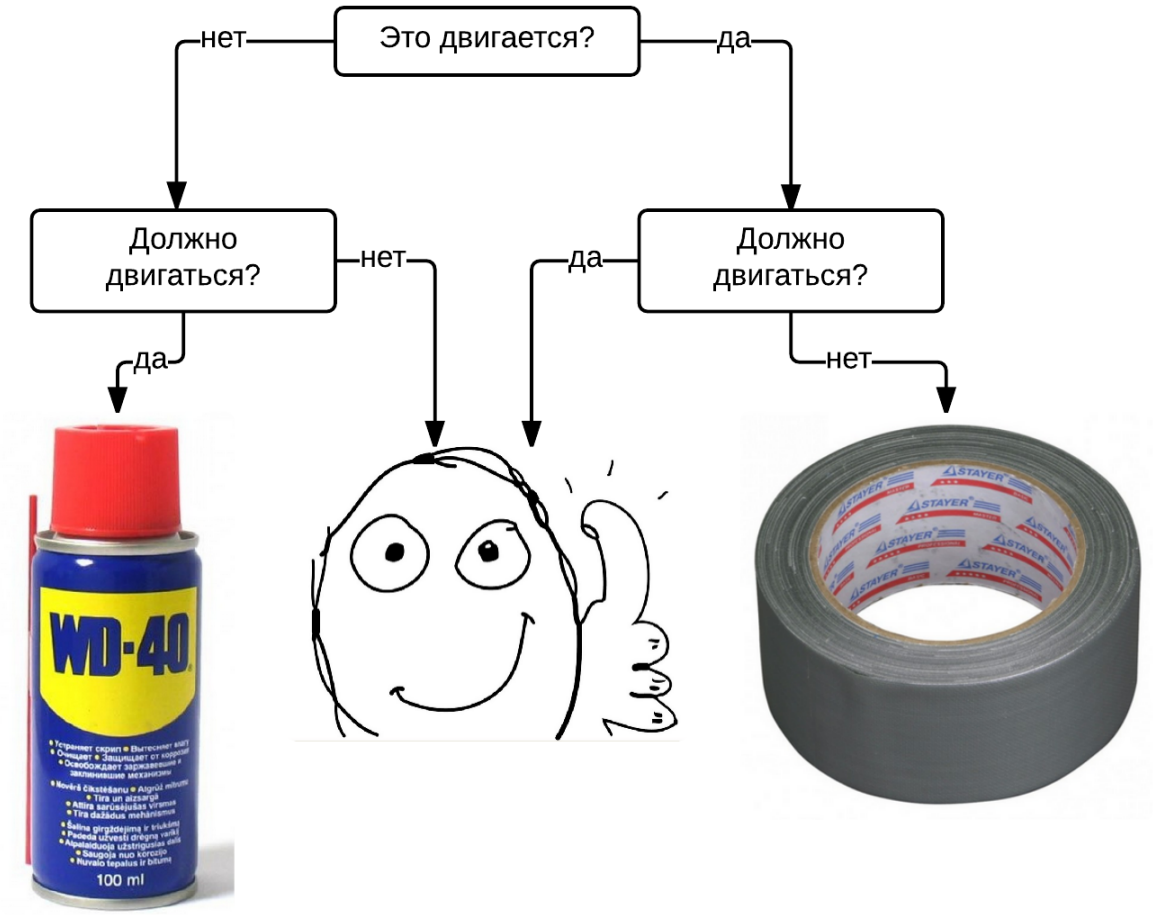

Рисунок 1 – Пример дерева решений (памятка инженерам)

В зависимости от решаемой задачи в листьях дерева могут храниться значение одного из двух основных типов: либо это конкретный класс рассматриваемой предметной области при решении задачи классификации, либо число (регрессионная модель), определяющее вероятность наступления некоторого события (победа команды в предстоящем матче) или прогнозируемую величину исследуемой функции (цена от продажи квартиры). В качестве входных данных дерево решений принимает массив параметров с опорной информацией, используемой в узлах дерева для определения перехода, например, для расчета вероятности выигрыша футбольной команды это будут: информация о положении в турнирной таблице, погоде и предыдущих играх.

3.1 Формирование дерева

Деревья решений имеют очень простую и наглядную структуру, что позволяет формировать их вручную для небольшого набора данных. Тем не менее, существуют разные схемы для автоматического обучения и выращивания деревьев. Каждый из таких методов имеет свои преимущества и недостатки, но один из подходов стал основой большинства других решений, поскольку оказался очень быстродействующим и простым в реализации. Этот алгоритм известен под названием рекурсивного секционирования [2] и действует по принципу инкрементного построения дерева.

Процесс выращивания начинается с пустого дерева и полного набора данных. Первым этапом решается задача поиска приемлемой точки разбиения исходной выборки на подмножества. С точки зрения реализации алгоритма, применяемый способ выбора атрибута разделения не имеет значения. Таким образом создается первый узел в дереве, разбивающий набор данных на несколько подмножеств, после чего данный процесс выполняется повторно, применительно к каждому из подмножеств, для создания поддеревьев. Критерий прекращения выращивания дерева выбирается в зависимости от решаемой задачи.

Для создания узлов принятия решений необходимо установить какими должны быть критерии секционирования. Как правило, при выборе атрибутов разбиения применяется жадный алгоритм, предусматривающий выбор наилучшего из приемлимых атрибутов.

3.2 Процедура обучения

При использовании автоматических методов построения деревьев принятия решений отдельное внимание необходимо уделять обучающей выборке чтобы не допустить переобучения (overfitting) — ситуации, когда во входном наборе данных обнаруживаются некоторые случайные закономерности, отсутствующие в генеральной совокупности. В противном случае, полученная модель является недостаточно общей.

Один из подходов борьбы с проблемой переобучения заключается в заготовке качественных наборов данных, разделенных на три группы:

- Обучающее множество, которое используется в алгоритме рекурсивного секционирования.
- Аттестационное множество, для усовершенствования дерева решений, полученного в результате обучения.
- Проверочное множество, позволяющее выполнить окончательную проверку результатов обучения для подтверждения применимости полученного дерева решений.

3.3 Отсечение веток

Отсечение частей дерева решений (pruning) является алгоритмом улучшения качества и осуществляется на этапе обработки, который следует за обучением. Для его работы необходимо иметь аттестационное множество, отличное от обучающей выборки. Метод основан на предположении, что в дереве существуют ребра, убрав которые произойдет увеличение точности.

Для того чтобы осуществить подобную проверку, необходимо знать, какое промежуточное значение накапливается в каждом узле принятия решений. Далее происходит тестирование дерева на всем аттестационном множестве с подсчетом успешно классифицированных наборов данных в каждой вершине. Если окажется, что некоторый узел имеет лучшую оценку, чем сумма всех его сыновей, то происходит усечение дерева (рисунок 2).

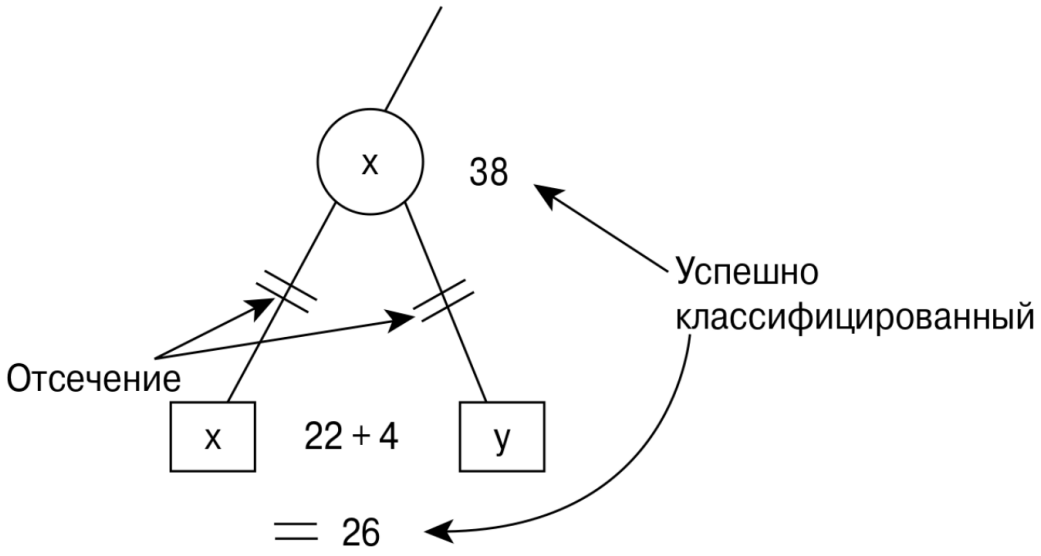

Рисунок 2 – Отсечение ребер дерева после подсчета успешно классифицированных примеров

3.4 Случайный лес

Метод случайного леса (random forest) основан на использовании комитета (ансамбля) деревьев принятия решений. Он был предложен Leo Breiman [3], одним из создателей алгоритма для построения бинарного дерева решений (CART — Classification And Regression Trees). Метод дает высокое качество результата, лучше чем у нейронных сетей [4], эффективно обрабатывает данные с большим числом признаков классов, обладает возможностью оценивать значимости отдельных признаков в модели.

При обучении создается выборка из N примеров, а также задаются размерность пространства признаков M и дополнительный параметр m (для задач классификации обычно выбирается $m \approx sqrt(M)$). Все деревья комитета строятся независимо друг от друга по следующей процедуре:

- Генерируется случайное подмножество с повторениями размером N из множества обучающей выборки.

- Выращивается решающее дерево, классифицирующее примеры данного подмножества, причём в ходе создания очередного узла выбирается признак не из всего пространства признаков размерности M, а лишь из m случайно выбранных.

- Дерево строится до полного исчерпания подмножества примеров и не обрабатывается процедурами улучшения результата (такими как отсечение веток).

Классификация объектов проводится путём голосования: каждое дерево комитета относит классифицируемый объект к одному из классов, и побеждает тот, за который проголосовало наибольшее число деревьев. Размерность леса подбирается таким образом, чтобы минимизировать ошибку классификатора на тестовой выборке.

Случайные леса могут быть естественным образом использованы для оценки важности переменных в задачах регрессии и классификации. Меняя параметры выборки, и определяя точки максимумов можно установить какие из них являются более важными для тренировочного набора.

Использование случайного леса для задачи классификации изображений вместе с алгоритмами выделения регионов интереса (ROI — Region Of Interest) дает хорошие результаты [5].

4 Нейронные сети

Под нейронными сетями понимают математическую модель самообучающейся системы, имитирующей деятельность нервных клеток в биологических нейронных сетях. Понятие возникло при изучении процессов, протекающих в мозге, и при попытке их моделирования. Несмотря на большое разнообразие вариантов нейронных сетей, все они состоят из большого числа связанных между собой однотипных элементов — нейронов, иммитирующих одноименные клетки живых организмов.

Искусственный нейрон, так же, как и живой, состоит из синапсов, связывающих его входы с ядром, которое осуществляет обработку входных сигналов. Связь с нейронами следующего слоя происходит через аксон. Каждый синапс имеет вес, определяющий насколько соответствующий вход нейрона влияет на его состояние, вычисляемое по формуле 1,

$$S = \sum_{i=1}^{n} x_i w_i \tag{1}$$

где n — число входов нейрона, x_i — значение i-го входа нейрона, w_i — вес i-го синапса. Значение аксона вычисляется как $f(S)$. Функция f называется активационной и может быть своей у каждого нейрона сети. На рисунке 3 показан пример нейросети, состоящей из трех слоев: входные нейроны (зеленый цвет), скрытые (синий) и выходной нейрон (желтый).

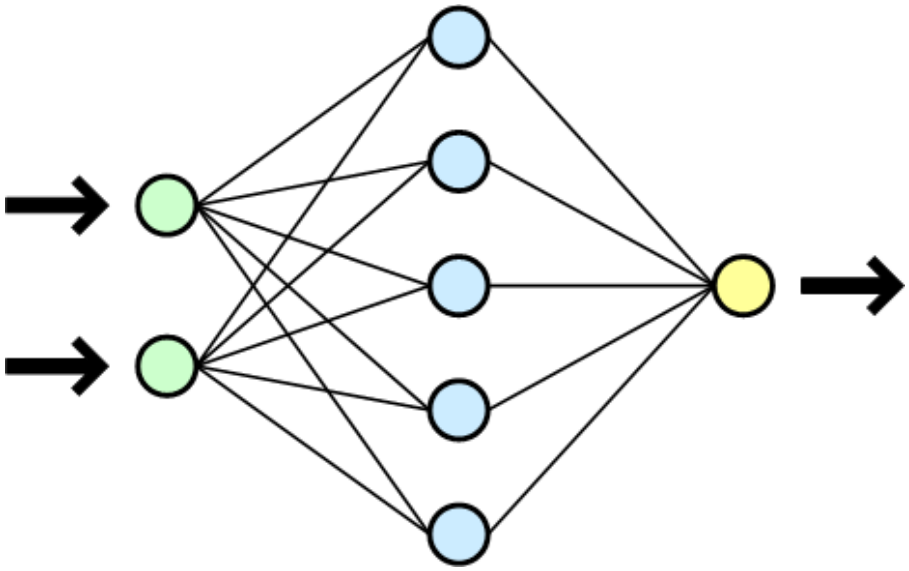

Рисунок 3 – Схема простой нейросети

Нейронные сети активно используются в задачах распознавания образов и предсказывания. Их часто применяют в области машинного обучения, называющейся глубокое обучение (deep learning), занимающейся моделированием высокоуровневых абстракций данных, используя системы со сложной архитектурой из множества нелинейных трансформаций. Под глубиной понимается максимальная длина между входным и выходным узлами графа вычислений модели.

4.1 Нейронные сети обратного распространения

4.2 Сверточные нейронные сети

Используемые обозначения

ANN — Artificial Neural Network, искуственная нейронная сеть (тоже самое что и NN)
CART — Classification And Regression Trees, алгоритм для построения бинарного дерева решений
Data mining — автоматические методы интеллектуального анализа
Decision tree — деревья принятия решений
Deep learning — область машинного обучения, занимающаяся моделированием высокоуровневых абстракций данных, используя системы со сложной архитектурой из множества нелинейных трансформаций
Fine tuning — алгоритмы частичной корректировки коэффициентов модели при ее дообучении
Information extraction — методы автоматического выделения информации
ML — Machine Learning, машинное обучение
NN — Neural Network, нейронная сеть
Pruning — отсечение веток дерева решений, алгоритм улучшения качества
OCR — Optical Character Recognition, задача распознавания рукописного текста
Overfitting — переобучение
Random forest — комитета решающих деревьев для задач классификации и регрессии
ROI — Region Of Interest

Список литературы

[1] Learning hierarchical features for scene labeling / C. Farabet, C. Couprie, L. Najman, Y. LeCun // *Pattern Analysis and Machine Intelligence, IEEE Transactions on.* — 2013. — Vol. 35, no. 8. — Pp. 1915–1929.

[2] *Quinlan, J. R.* Decision trees and decision-making / J. R. Quinlan // *Systems, Man and Cybernetics, IEEE Transactions on.* — 1990. — Vol. 20, no. 2. — Pp. 339–346.

[3] *Breiman, L.* Random forests / L. Breiman // *Machine learning.* — 2001. — Vol. 45, no. 1. — Pp. 5–32.

[4] *Niculescu-Mizil, A.* An empirical comparison of supervised learning algorithms using different performance metrics: Tech. rep. / A. Niculescu-Mizil, R. Caruana: Cornell University, 2005.

[5] *Bosch, A.* Image classification using random forests and ferns / A. Bosch, A. Zisserman, X. Munoz // Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on / IEEE. — 2007. — Pp. 1–8.