ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«МОСКОВСКИЙ ИНСТИТУТ ЭЛЕКТРОННОЙ ТЕХНИКИ»

ОТЧЕТ ПО ПРОИЗВОДСТВЕННОЙ ПРАКТИКЕ

Направление подготовки — 01.03.04 «Прикладная математика»

Профиль — «Применение математических методов к решению инженерных и экономических задач»

|  |  |  |
| --- | --- | --- |
| Выполнил студент Димаков Владислав Сергеевич | | Группа: МП-40 |
| Оценка руководителя практики от кафедры ВМ-1  Козлитин Иван Алексеевич, доцент, к.ф.-м.н. | \_\_\_\_\_\_\_\_\_\_\_\_\_ | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  | (оценка) | (подпись) |

Москва

2017

Оглавление.

[Введение. 2](#_Toc481350350)

[Задача классификации. 2](#_Toc481350351)

[Обзор существующих методов классификации. 3](#_Toc481350352)

[Деревья решений. 4](#_Toc481350353)

[Алгоритм CART. 5](#_Toc481350354)

[Оптимальное расщепление вершин. 5](#_Toc481350355)

[Алгоритм построения случайного леса. 6](#_Toc481350356)

[Признаки Хаара. 6](#_Toc481350357)

[Создание обучающей выборки. 7](#_Toc481350358)

[Тестирование классификатора. 7](#_Toc481350359)

[Заключение. 8](#_Toc481350360)

[Список использованных источников. 9](#_Toc481350361)

# Введение.

В настоящее время от систем видеонаблюдения требуется не только предоставление возможности воспроизведения и записи видеопотока с камеры, но и возможности решения в автоматическом режиме множество задач без участия человека, начиная от простого детектирования движения в области наблюдения, заканчивая высокоточным подсчётом проехавших машин или прошедших людей. Одной из таких задач является задача классификации объектов, присутствующих на цифровой видеозаписи.

# Задача классификации.

Вероятностная постановка задачи классификации выглядит следующим образом. Пусть множество пар «объект, метка класса» является вероятностным пространством с неизвестной вероятностной мерой . Имеется конечная обучающая выборка наблюдений , сгенерированная согласно вероятностной мере . Требуется построить алгоритм , способный классифицировать произвольный объект

Далее под обучающей выборкой будем понимать независимую выборку из некоторого неизвестного распределения . Здесь – векторы признаков, координаты которых представляют собой значения признаков, измеряемых на некотором объекте; – метки классов, .

# Обзор существующих методов классификации.

Существуют различные методы решения задачи классификации, такие как наивный байесовский классификатор, метод ближайших соседей, искусственные нейронные сети с прямой связью и случайные леса.

Наивный байесовский классификатор **–** вероятностный классификатор, основанный на применении Теоремы Байеса со строгими предположениями о том, что объекты описываются статистически независимыми признаками. Предположение о независимости существенно упрощает задачу, так как оценить одномерных плотностей гораздо легче, чем одну -мерную плотность. К сожалению, данное предположение крайне редко выполняется на практике, что влечет за собой относительно низкое качество классификации в большинстве реальных задач.

Метод ближайших соседей – метод метрической классификации, основанный на оценивании сходства объектов. Классифицируемый объект относится к тому классу, к которому принадлежат ближайших к нему объектов обучающей выборки. Недостатками метода явлюется сложность выбора параметра и высокая вычислительная трудоемкость, которая увеличивается квадратично с ростом числа записей в наборе данных.

Искусственная нейронная сеть с прямой связью являются универсальным средством аппроксимации функций, что позволяет использовать ее в решении задач классификации. Как правило, нейронные сети оказываются наиболее эффективным методом классификации, так как способны генерировать большое число моделей, которые используются в решении задач классификации статистическими методами. Главными недостатками нейронных сетей являются сложность реализации их структур и подбор параметров моделей.

Случайный лес – статистический метод, предназначенных для решения задач классификации и регрессии. Метод основан на построении большого числа деревьев решений, каждое из которых строится по выборке, получаемой из обучающей с помощью бутстрепа (т.е. выборки с возвращением). В отличие от классических алгоритмов построения деревьев решений, в методе случайных лесов при построении каждого дерева на стадии расщепления вершин используется только фиксированное число случайно отбираемых признаков и строится полное дерево, т.е. дерево без усечения.

Данный метод обладает следующими достоинствами:

* Способность эффективно обрабатывать данные с большим числом признаков и классов;
* Нечувствительность к масштабированию значений признаков;
* Одинаково хорошая обработка как непрерывных, так и дискретных признаков;
* Для построения случайного леса по обучающей выборке требуется задание всего двух параметров – количества деревьев в ансамбле и числа случайно отбираемых признаков на стадии расщепления вершин.

Несмотря на то, что данный метод также как и нейронные сети требует построения модели большого размера, он наилучшим образом подходит для решения поставленной задачи классификации объектов, присутствующих на цифровой видеозаписи.

# Деревья решений.

Обозначим – пространство образов, т.е. множество всех возможных значений векторов признаков. Тогда деревом решений будет называться дерево, с каждой вершиной которого связаны:

* Некоторое подмножество . С корневой вершиной связывается все пространство образов .
* Подвыборка обучающей выборки , такая, что

. С корневой вершиной связывается вся выборка .

* Некоторая функция (правило) (здесь – количество потомков вершины ), определяющая разбиение множества на непересекающихся подмножеств. С листьями дерева не связывается никакая функция.

Цель построения дерева решений состоит в классификации векторов из распределения . Процесс принятия решений начинается с корневой вершины и состоит в последовательном применении правил, связанных с вершинами дерева. Результатом этого процесса является определение листа такого, что . В этом случае вектор относится к классу, являющемуся мажорантным (наиболее часто встречающимся) в подвыборке , соответствующей данному листу.

# Алгоритм CART.

Алгоритм CART (Classification and Regression Tree) предназначен для решения задач классификации и регрессии построением бинарного дерева решений. На каждом шаге построения дерева правило , формируемое в узле , делит обучающую выборку на две более однородные подвыборки:

Обычно вместо меры однородности используется противоположная по смыслу мера загрязненности. Пусть – некоторая вершина дерева решений, – подвыборка, связанная с этой вершиной, – загрязненность вершины. Необходимо потребовать, чтобы загрязненность вершины была равна , если содержит прецеденты только одного класса и была максимальной в случае, если содержит одинаковое число прецедентов каждого класса.

Одной из наиболее часто используемых является мера загрязненности вершины, формализованная в индексе Gini:

где – доля примеров класса в подвыборке .

# Оптимальное расщепление вершин.

Правило разбиения множества , связанное с каждой вершиной дерева решений, называется расщеплением. Бинарное расщепление вершины можно рассматривать как функцию , где в случае вектор относится к первому (левому) потомку, а в случае – ко второму (правому).

Расщепление подвыборки естественно осуществлять таким образом, чтобы максимально уменьшить загрязненность. Уменьшение загрязненности вершины для бинарных деревьев определяется как

где и – доли примеров подвыборки , соответствующие левому и правому потомкам ( и ). Наилучшим расщеплением вершины считается разбиение, которое максимизирует величину , т.е. расщепление выполняется таким образом, чтобы .

# Алгоритм построения случайного леса.

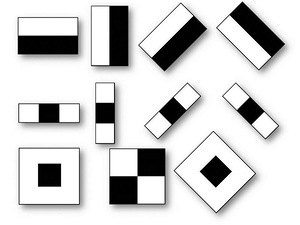
Алгоритм построения случайного леса может быть представлен в следующем виде:

Для (здесь – количество деревьев в ансамбле) выполнить:

1. Сформировать бутстреп выборку размера по исходной обучающей выборке .
2. По бутстреп выборке построить неусеченное дерево решений с минимальным количеством наблюдений в терминальных вершинах равным , рекурсивно следуя следующему подалгоритму:
3. из исходного набора признаков случайно выбрать признаков (в задачах классификации обычно );
4. из признаков выбрать признак, который обеспечивает наилучшее расщепление;
5. расщепить выборку, соответствующую обрабатываемой вершине, на две подвыборки;

# Признаки Хаара.

Признаки Хаара – это признаки цифрового изображения, представляющие собой прямоугольные области, состоящие из смежных прямоугольных подобластей. Своим названием они обязаны сходством с вейвлетами Хаара.



Значение двух-прямоугольного признака вычисляется, как разность между интегральными суммами пикселей в двух смежных прямоугольных подобластях. Для трех-прямоугольного признака значение вычисляется, как интегральная сумма двух внешних подобластей, вычитаемая из суммы в центральной подобласти. Значение четырех-прямоугольного признака вычисляется, как разность между суммами диагональных пар подобластей.

Рассмотрим вектор признаков , координаты которого представляют значения признаков Хаара, измеряемых в некоторой области, полученной в результате сегментации изображения. Значение признака примем равным единице в случае, если разность интегральных сумм подобластей неотрицательна, и нулю – в противоположном случае.

# Создание обучающей выборки.

Цифровая видеозапись представляет собой последовательность двумерных матриц растровых полутоновых изображений , где – номер кадра в последовательности, а каждый элемент матрицы характеризует цвет соответствующего пикселя.

В качестве обучающей видеозаписи была выбрана видеозапись движения цели – радиоуправляемой машинки. Данная видеозапись включала в себя временные промежутки с изменением освещенности области наблюдения и временные промежутки с изменением размеров цели вследствие ее удаления от видеокамеры.

Для создания обучающей выборки на основе видеозаписи был выполнен следующий алгоритм:

1. Для каждого кадра видеозаписи была выполнена сегментация изображения;
2. Для всех полученных сегментов были вычислены значения случайно сгенерированных векторов признаков Хаара ;
3. Значение метки класса принималось равным единице, если – вектор признаков, соответствующий сегменту, охватывающему цель, нулю – в противоположном случае.

# Тестирование классификатора.

Для создания тестовой выборки была использована видеозапись, сходная с обучающей. Данная видеозапись имела отличную от обучающей видеозаписи область наблюдения и включала в себя иные траектории движения цели.

По окончании тестирования был проведен анализ полученных результатов классификации и построен график ошибки.



# Заключение.

Применение классификатора на основе случайного леса для решения задачи обнаружения объектов, присутствующих на цифровой видеозаписи, показало высокие результаты, стойкие к масштабированию объектов, изменению освещенности в области наблюдения, исчезновению объектов из области наблюдения. При этом точность классификации оказалась выше при использовании большого числа деревьев в ансамбле.

Существенным недостатком метода является то, что фазы обучения и тестирования разделены. Так как на практике обучающие данные чаще всего поступают последовательно (например, в приложениях трекинга), то в таких ситуациях алгоритм обучения классификатора должен работать on-line, т.е. «на лету». On-line обучение имеет большое число преимуществ перед off-line методом, например, требует меньший объем выделяемой памяти. Также on-line метод показывает высокие результаты в случае, если распределение, лежащее в основе обучающей выборки, меняется с течением времени.

# Список использованных источников.

1. С.П. Чистяков, «Случайные леса: обзор». Труды Карельского научного центра РАН №1. 2013. С. 117-136.
2. L. Breiman, J.H. Friedman, R.A. Olshen, and C.J Stone. Classification and Regression Trees. Wadsworth, Belmont, Ca, 1983.
3. A. Saffari, C. Leistner, J. Santner, M. Godec, and H. Bischof, «On-line Random Forests», 3rd IEEE ICCV Workshop on On-line Computer Vision, 2009.