



Метод опорных векторов — набор схожих алгоритмов обучения с учителем, использующихся для задач классификации и регрессионного анализа. Принадлежит к семейству линейных классификаторов, может также рассматриваться как специальный случай регуляризации по Тихонову. Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора, поэтому метод также известен как метод классификатора с максимальным зазором.

Основная идея метода — перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей наши классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Разделяющая гиперплоскость — это гиперплоскость, которая отделяет группу объектов, имеющих различную классовую принадлежность.

Для построения оптимальной гиперплоскости, SVM прибегает к итерационному алгоритму обучения, использующемуся для минимизации функции ошибок.

SVM применяется в таких задачах как задача о понятие оптимальной разделяющей гиперплоскости, линейно разделимая выборка, линейно неразделимая выборка, ядра и спрямляющие пространства, алгоритмы настройки.

Так же существует ряд ядер, которые могут быть использованы в моделях метода опорных векторов. Они включают в себя линейные, полиномиальные, радиальные базисные функции (RBF) и сигмовидные.

SVM, прежде всего, отличный метод, который решает задачи классификации с помощью построения гиперплоскостей в многомерном пространстве. SVM поддерживает как регрессионный анализ, так и задачи классификации, и может работать с несколькими непрерывными и категориальными переменными.

Наиболее развитая и популярная реализация SVM на C++. Существуют адаптированы библиотеки для большинства выборок, включены стандартные ядерные функции, допускается использование предварительно вычисленных матриц ядерных функций, линейная классификация и регрессия.

Целью данного исследования является разработка метода обучения распознаванию образов, максимально близкого к классическому методу опорных векторов (SVM), но использующего только некоторую метрику, заданную на множестве объектов распознавания, которое может быть конечным или бесконечным. Предполагается, что метрика удовлетворяет требованиям, отличающим класс так называемых евклидовых метрик. Всякая такая метрика погружает исходное множество объектов в некоторое, вообще говоря, большее метрическое пространство мощности континуума. Выбор произвольного элемента как нулевого превращает это метрическое пространство в линейное простран-



ство со специфическим скалярным произведением, но с исходной евклидовой метрикой.

SVM могут быть применены в качестве классификатора для выполнения картографирования местности с применением беспилотного летательного аппарата вертолетного типа. В ходе исследований было проведено несколько полётных сессий над Ореховым озером (Рязанская область, г. Рязань), в ходе которых были сняты следующие характерные сцены:

- плоская поверхность;
- почти плоская поверхность с иррегулярными небольшими перепадами высот (кочки, канавы, ямы);
- почти плоская поверхность с произрастающей на ней травой;
- технический колодец;
- насыпь (резкий уклон);
- водный объект;
- линии электропередач.

Результаты данной сессии полётов были использованы для проведения экспериментальных исследований метода оперативного картографирования на базе SVM.

По результатам экспериментальных исследований разработанный метод картографирования показал высокую точность (в среднем — смещение до 2-х метров относительно действительного положения объектов, время обработки данных — не более 1,35 секунд на кадр).

А.И. Лян, А.В. Куприянов

## ПРИМЕНЕНИЕ ТЕХНОЛОГИИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ ДЛЯ СТАТИСТИЧЕСКОГО ТЕКСТУРНОГО АНАЛИЗА ИЗОБРАЖЕНИЙ

(Самарский национальный исследовательский университет  
имени академика С.П. Королёва. г. Самара)

Гистограмма изображения представляет распределение интенсивности отдельно взятых пикселей изображения [1]. С ее помощью можно получить необходимую информацию об изображении. Двумерная гистограмма представляет собой распределение интенсивности пар пикселей (пикселя и его соседа по выбранному направлению - например, горизонтали). Соответственно, трехмерная гистограмма представляет собой распределение интенсивности троек пикселей (сосед выбирается по обоим направлениям).

Трехмерная гистограмма необходима для расчетов различных статистических текстурных признаков, которые в свою очередь необходимы для решения задачи классификации текстур. В общем случае, статистические признаки характеризуют вероятностное распределение уровней яркости изображения [2].

Сложность расчета трехмерной гистограммы заключается в большом объеме решаемой задачи, поскольку необходимо обработать большие объемы



данных - а именно, число таких троек будет равно квадрату общего числа пикселей изображения. Имея даже изображение 10x10, нам потребуется обработать 10000 троек. Сложность задачи растет быстро - как квадрат площади изображения.

В качестве решения данной проблемы можно предложить распределенную обработку изображения на кластере при помощи технологии MapReduce. Кластер это группа компьютеров, объединенных в общую сеть при помощи линий связи. Это могут быть как быстрые Infiniband, так и более медленные Ethernet.

MapReduce это подход к обработке данных, сформированный компанией Google. Ее принцип заключается в том, чтобы разделить все операции с данными на два этапа: map и reduce. На map шаге рабочие узлы получают задание и головного на предварительную обработку данных. На reduce шаге рабочие ноды осуществляют свертку результатов обработки после map шага. После выполнения этого шага, головной узел получает итоговый результат.

В качестве фреймворка для разработки можно выбрать Apache Hadoop [3]. Он создавался с учетом этой модели и позволяет создать надежную систему на большом числе узлов (вплоть до тысяч и сотен тысяч), и единственным ограничением является объем памяти головного узла, которая хранит метаданные. Кроме того, он достаточно прост в разработке и отладке.

В качестве файловой системы используется распределенная HDFS [4], оптимизированная для однократной записи и многократного считывания. Файл хранится в виде набора реплицируемых (для отказоустойчивости) блоков. Hadoop предоставляет API и консольную программу для работы с данной файловой системой.

Hadoop реализован на языке программирования Java и имеет свою реализацию MapReduce. Поэтому для определения своих операций map и reduce необходимо реализовать определенные классы - Mapper и Reducer. Это интерфейсы, которые имеют методы map и reduce соответственно. Для обмена данными используется класс Context.

На рисунке 1 приведен результат расчета гистограммы при помощи указанной технологии. Она является результатом суммирования трехмерной гистограммы по двум координатам.

Достоинством данного подхода является масштабируемость. На кластере можно за короткое время обрабатывать как крупные изображения, так и множество небольших. При недостатке производительности достаточно добавить одну или несколько новых нод. Еще один плюс - относительная дешевизна, поскольку как правило новое оборудование стоит дешевле дополнительных человеко-часов, которые необходимо затратить на улучшение существующего ПО.

У данного метода есть недостатки. Очевидно, что с учетом накладных коммуникационных расходов между отдельными нодами в кластере. Это приводит к тому, что вычисление при помощи других технологий на одной ЭВМ (например, с помощью CUDA) оказывается более эффективным.

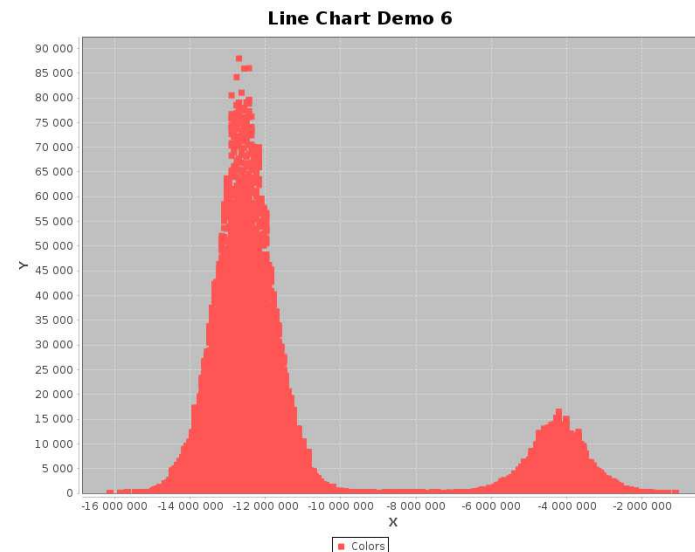


Рисунок 1 – Гистограмма изображения

### Литература

1. Zone System & Histograms [Электронный ресурс]. - Режим доступа: <http://www.illustratedphotography.net/basic-photography/zone-system-histograms>
2. Куприянов, А.В. Анализ текстур и определение типа кристаллической решетки на наномасштабных изображениях [Текст] / А.В. Куприянов // Компьютерная оптика - 2011. - Т. 35.-2.-С.145-152.
3. Apache Hadoop 2.7.2 [Электронный ресурс]. - Режим доступа: <http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html>
4. HDFS Architecture Guide [Электронный ресурс]. - Режим доступа: [https://hadoop.apache.org/docs/r1.2.1/hdfs\\_design.html](https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html)

Д.О. Маркин, А.С. Галкин, П.А. Архипов

### ПОСТРОЕНИЕ АНОНИМНОЙ СЕТИ НА БАЗЕ ТЕХНОЛОГИИ ВЕБ-ПРОКСИ

(Академия Федеральной службы охраны Российской Федерации)

Существующие современные условия удаленного доступа к информационным ресурсам позволяют говорить о том, что простое обращение к информационному сервису оставляет значительное количество "следов" такого обращения в информационных log-файлах провайдеров услуг связи, промежуточных узлов на пути следования данных, а также программном обеспечении и удален-