Введение. Компьютерная аналитика: от классических методов к машинному обучению

Сделанный в конце XX века рывок в компьютерных технологиях анализа данных достиг в наши дни своего пика. Начавшиеся с разработки AT&T Bell Labs объектно-ориентированного языка программирования S+, технологии компьютерного анализа данных стремительно развивались и совершенствовались. Одна из главных целей создателей аналитического программного обеспечения на первом этапе заключалась в разработке максимально полного аналитического и графического функционала, интеграции с базами данных, оптимизации вычислительных процедур и тщательное тестирование методов.

Описанные в книгах и учебниках специализированные аналитические и статистические методы воплотились в компьютерные технологии, что было значительным успехом.

Функционал современного программного обеспечения компьютерной аналитики включает тысячи тщательно тестированных аналитических процедур, постоянно пополняющихся усовершенствованными методами анализа.

Современная версия программы *STATISTICA* содержит более 10 тыс. вычислительных аналитических процедур, используемых во всех областях человеческой деятельности – от промышленных и бизнес-приложений, ритейла до биомедицины и геостатистики: описательные статистики, проверка гипотез, непараметрическая статистика, разнообразные критерии согласия, анализ таблиц сопряженности, различные вариации дисперсионного анализа, множественной регрессии, нелинейного оценивания, факторный анализ, анализ главных компонент, многомерное шкалирование, десятки методов классификации и кластеризации, реализации GLM моделей, нейронные сети, радиальные базисные функции, многослойные персептроны, методы опорных векторов, методы анализа временных рядов, спектральный анализ и многие другие.

Перечисление методов от описательных до углубленных методов анализа и методов построения предиктивных моделей, включая машинное обучение, займет несколько десятков страниц.

Параллельно развивается мир информационных технологий как таковых, цифровая реальность вошла в жизнь, интернет-вещей, облачные решения стали нормой современного производства, оснащенного множеством датчиков, контроллеров, сенсоров. Сбор и хранение данных от различных источников стал надежным, недорогим и эффективным.

Как только мир становится цифровым и наполненным данными, возникает потребность в развитой компьютерной аналитике.

Оказалось, что наряду с классическими методами эффективными являются методы машинного обучения, позволяющие использовать сами данные для настройки и оптимизации алгоритмов. Идея машинного обучения оказалась чрезвычайно плодотворной.

Методы машинного обучения синтезируют методы статистики, теории вероятностей, численных методов, теории оптимизации и других дисциплин. Это актуальная и интересная область, наполненная практически важными задачами и открывающая для исследователей широкое поле деятельности. Многие задачи машинного обучения могут быть поставлены как оптимизационные задачи: мы строим модель, используя *обучающее* множество, и проверяем ее качество на *тестовом* наборе данных.

Отдельная глава книги посвящена методам оптимизации и их приложениям.

Основной тренд компьютерной аналитики

Переход от анализа данных к анализу потоков данных и использованию методов машинного обучения наряду с классическими технологиями составляет суть современной

компьютерной аналитики. Аналитика нужна не сама по себе, а для создания реально работающих моделей и выработки правильных решений в ситуациях, описываемых большим набором данных.

Итак, с одной стороны имеется множество данных, зачастую слабо структурированных или вовсе не структурированных, поступающих от различных источников, датчиков, контроллеров, переключателей, расходомеров, с другой стороны – множество аналитических методов, интегрированных в рамках единой системы или доступных с помощью объектно-ориентированных языков программирования, где методы анализа также являются и объектами.

Попробуем описать применение компьютерных аналитических технологий в конкретных областях.

Промышленность. Текущий и прогностический мониторинг процессов является стандартом для современного производства. Например, современное производство алюминия – это сложная многофункциональная энергоемкая система, требующая тщательного технологического контроля и управления на основе мониторинга параметров и контроля содержание глинозема, регулировки расхода электроэнергии.

Автоматизированные системы мониторинга и управления процессом могут включать цифровые камеры и использовать методы корреляционного и регрессионного анализа для определения степени зашлакованности поверхности расплава по яркости излучения.

Оценка зависимостей составляющих оптического спектра от переизбытка фторида алюминия и расчета криолитового отношения и других параметров важная задача. Возникающие зависимости являются нелинейными и полезными оказываются технологии машинного обучения. Именно в направлении предиктивного моделирования и всестороннего мониторинга промышленного производства интенсивно развиваются ведущие мировые компании.

STATISTICA обладает уникальными графическими возможностями. В *STATISTICA* вы можете импортировать изображения двумя нажатиями кнопки, далее применить к ним разнообразные методы анализа, реализованные в рамках удобного пользовательского интерфейса в виде последовательно открывающихся диалоговых окон (рис. В.1). Доступны форматы изображений: bmp, jpeg, png, gif, tiff.

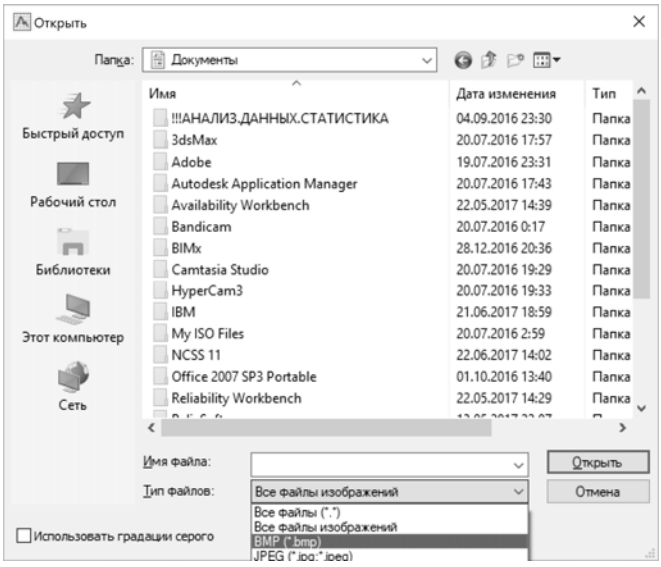

Рис. В.1

Соответствующая опция позволяет рассматривать изображения в градации серого цвета. Единственным параметром такого монохроматического света является *яркость*, которая изменяется в пределах от черного до белого с промежуточными оттенками серого цвета. С помощью удобных графических настроек можно построить уровни интенсивности, определить положение дефектов (рис. В.2).

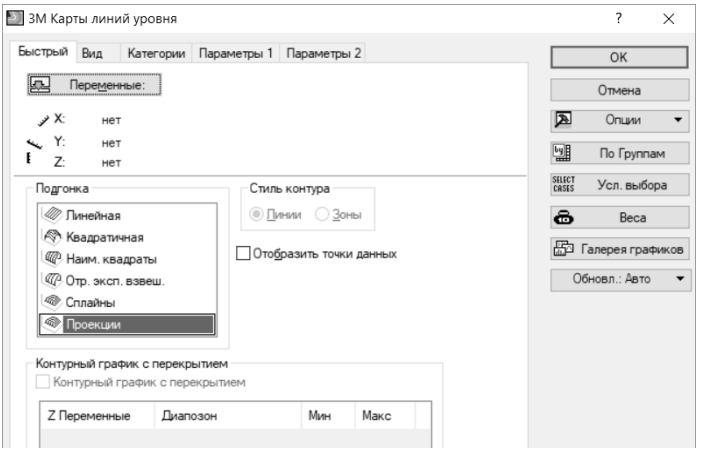

Рис. В.2

В промышленных приложениях эти возможности используются в проектах, связанных с цифровым зрением, что актуально для современных методов контроля качества (рис. В.3–6).

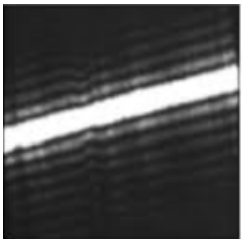

Рис. В.3

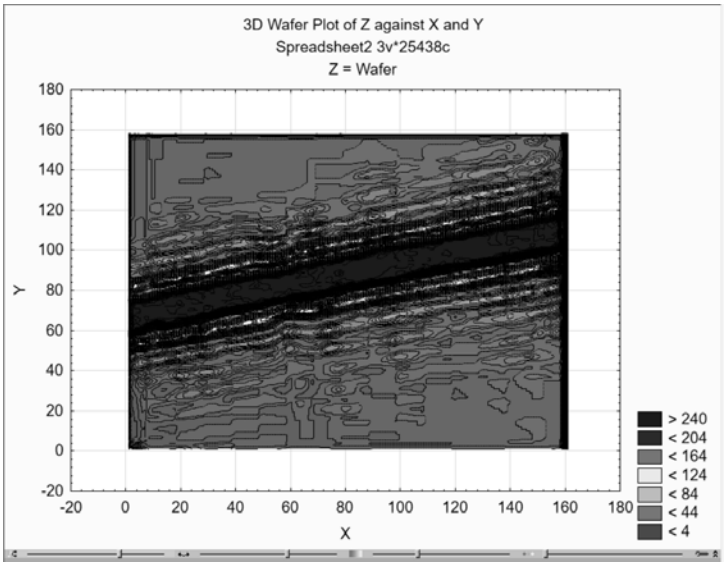

Рис. В.4

В металлургическом производстве оптический контроль с использованием камер высокого разрешения используется для контроля качества поверхности стального листа, позволяя определять наличие дефектов, классифицировать их и находить причины брака.

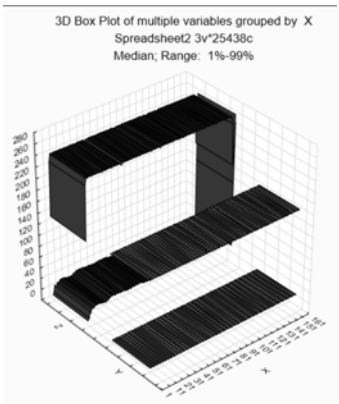

Рис. В.5

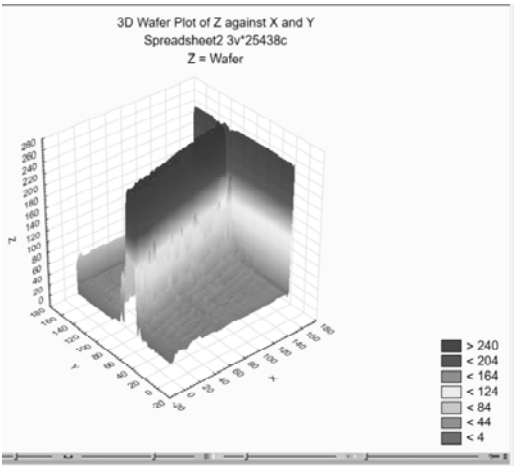

Рис. В.6

Аналогичные задачи возникают при производстве гипсокартона, сварки, компоновке изделий и т. д. Контроль качества гипсокартонных листов проводится с помощью светодиодного света и высокоскоростных камер, фиксирующих дефекты поверхности листа, которые далее анализируются по величине, положению, частоте возникновения, классифицируются в режиме реального времени, используя методы анализа изображений, что позволяет вести производство высокого качества.

В медицинских приложениях это могут быть разнообразные снимки с целью дальнейшей диагностики.

Известная модель RGB (Red – красный, Green – зеленый, Blue – синий) используется для регистрации *цветных* изображений. После проведения предобработки к оцифрованным визуальным данным применяются разнообразные статистические методы: классические методы кластерного анализа, метод главных компонентов, методы машинного обучения, включая нейронные сети.

STATISTICA обладает уникальным инструментарием для визуализации данных. Приведем некоторые примеры графиков, которые каждый пользователь легко построит в программе (рис. В.7).

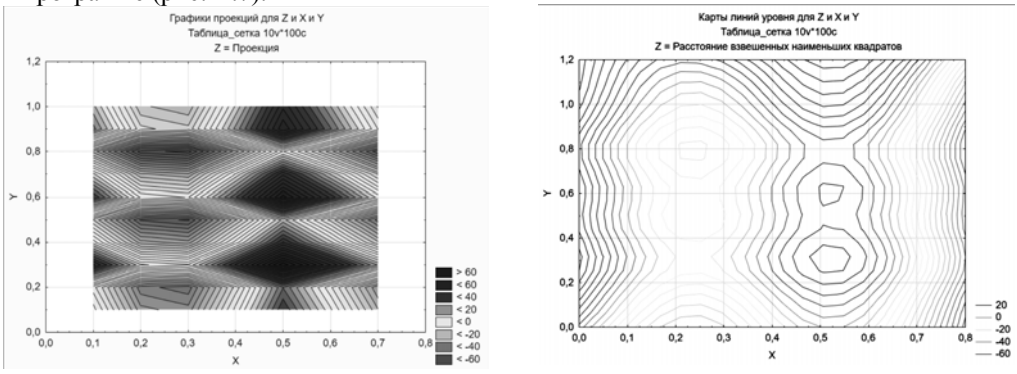

Рис. В.7

Повернутые под разными углами графики позволяют в цвете увидеть особенности объекта исследования (рис. В.8).

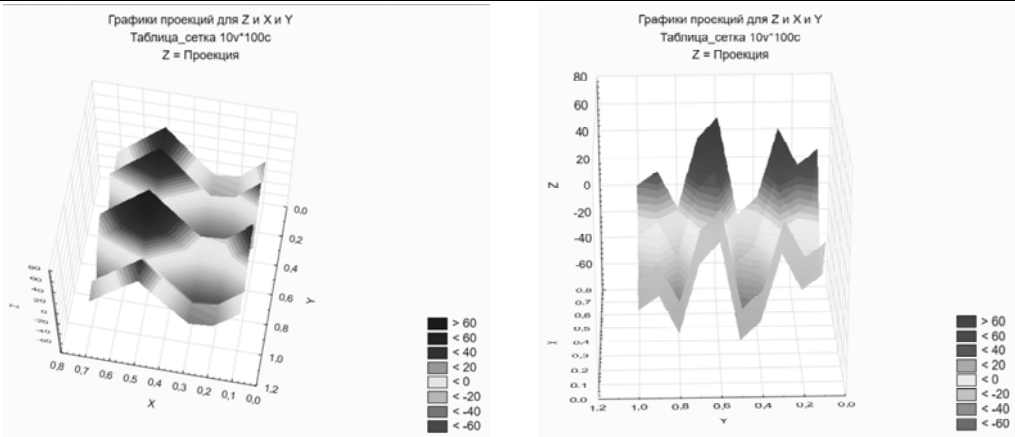

Рис. В.8

Текстильная отрасль. Контроль натяжения нитей, оптимальные способы окраски тканей, описательный и предсказательный мониторинг показателей датчиков, прогнозирование и минимизация обрывов.

Техническая диагностика. Оценка надежности оборудования, прогнозирование ремонтных работ и оптимизация работы сервисных служб на предприятии. Детерминированные модели не работают: различные условия и интенсивность эксплуатации делают необходимым применение прогностических моделей и статистических методов.

Нефтегазовая отрасль. Измерение количества сырого газа, поступающего на вход в установку подготовки газа (датчики, расходомеры), влажности, регулирование потоков газового конденсата в печь, прогноз продукта на выходе, контроль качества на основе измерений, диагностика насосного оборудования.

Энергетика. Контроль за энергосбереженим, оптимальный расчет энергопотребления на предприятии. Предсказательный мониторинг подачи топлива в зависимости от сезона, погодных условий, сырья, оптимизация расхода газа или электроэнергии при выплавке металла.

Маркетинг. Сбор и анализ информации по данной категории товаров, выявление причины падения продаж, причины негативных тенденций, формирование аналитической отчетности и т. д. Аналитика нужна для эффективного управления сетью магазинов, расположенных в одном центре и на больших расстояниях друг от друга.

Продажа запчастей к автомобилям требует склада на несколько десятков тысяч наименований, профицит и дефицит деталей прогнозируется, формирование плана закупок, заказ деталей с учетом того, что требуемые детали должны быть поставлены в заданные сроки (покупателю нужен товар здесь и сейчас).

Сбор и систематизация больших объемов информации, разработка отчетов и прогнозирование продаж, закупок, перевозки товаров и т. д. Потребность в проданной детали или группы деталей может прогнозироваться по сроку службы, все это требует применения предиктивной аналитики.

Аналогичные задачи возникают в современном сельском хозяйстве: полив растений в зависимости от температуры окружающей среды, внесение удобрений, умные теплицы и другие проекты.

Глобальная экономика. Прогнозирование цен на нефть, данные о максимальном числе действующих скважин, зафрахтованных танкерах и т. д. Сланцевая революция