отвечает блок «Запрос количества идентичных людей в базе данных». Этот блок работает по алгоритму прямого частичного сравнения разных наборов реквизитов, например, имени, отчества и даты рождения, формируя тем самым рабочий набор данных для рассматриваемого алгоритма нечеткого поиска. Затем в работу вступает «Блок сравнения реквизитов», ключевые функции которого отводятся логически выделенным процедурам COMPARISON_STRING и COMPARISON_NUMBER, созданным на основе модифицированного метода вычисления метрики Левенштейна, которые позволяют проводить интеллектуальное сравнение двух похожих строк или чисел с учетом возможных неточностей или ошибок ввода. С помощью указанных процедур программа формирует набор совпадений, и по результатам обработки предлагаемой и искомой записей выносит решение об идентичности строк. Например, у человека совпадают имя, отчество, дата рождения и номер паспорта, а в фамилии допущена ошибка в одну букву. В данном случае программа однозначно идентифицирует реквизиты. Данные процедуры могут применяться не только для нечеткого поиска реквизитов, но также везде, где требуется полнотекстовый поиск с нечетко заданными входными данными.

Алгоритм нечеткого поиска аккумулирует так называемый «опыт прошлых идентификаций» и записывает его в специально отведенное место в базе данных для использования в последующих идентификациях. Это позволяет сохранить не только результаты автоматической работы программы, но и решения операторов после отработки ими оставшихся не найденных реквизитов.

Заключение

Рассмотренный метод нечеткого поиска персональных данных на основе модифицированной метрики Левенштейна, позволяет быстро определять людей, используя данные ранее проведенного поиска. Встроенная система приоритета реквизитов позволяет идентифицировать человека в таких случаях, как смена фамилии, имени, переезд, ошибки при ручном вводе данных, а также при частично отсутствующих реквизитах.

Рассмотренную в работе процедуру нечеткого поиска можно рассматривать как часть системы поддержки принятия решений. Процедура не требует вмешательства оператора, накапливает опыт в процессе работы, позволяя тем самым полностью освободить специалистов от низкопрофильной, неэффективной ручной работы напрямую с наборами реквизитов физических лиц, хранящимися в базах данных.

В перспективе данный алгоритм может быть успешно внедрен в системы глобального объединения хранилищ государственных или коммерческих организаций для ведения единой базы данных населения любой страны мира. Логическая структура разработанного алгоритма позволяет реализовать его на любом популярном языке программирования. Масштабируемость алгоритма позволяет применять программные процедуры на его основе, как в малых организациях, так и в крупных корпорациях, везде, где ведется и актуализируется реестр данных физических лиц. Возможные примеры использования: портал госуслуг, медицинские электронные системы, кадровые и бухгалтерские системы учета служащих, банковские системы хранения данных о клиентах и т.п.

Алгоритм реализован на языке PL-SQL системы управления базами данных Oracle 11g. Разработанное программное обеспечение, реализующее метод автоматизированного поиска персональных данных на основе нечеткого сравнения, внедрено и успешно функционирует с 2007 года в муниципальном учреждении «Городской информационный центр» г. Тольятти Самарской области.

Литература

1. Международный фонд автоматической идентификации. Технологии автоматической идентификации [Электронный ресурс]. – Режим доступа: http://www.fond-ai.ru/art1/art223.html, свободный. Яз. рус. (дата обращения 16.06.2012).

2. Хемминг Р.В. Теория кодирования и теория информации: Пер. с англ. / Под ред. Б.С. Цыбакова. – М.: Радио и связь, 1985. – 176 с.

3. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР. – 1965. – Т. 163. – № 4. – С. 845–848.

4. Бойцов Л.М. Анализ строк [Электронный ресурс]. – Режим доступа: http://itman.narod.ru/articles/infoscope/string_search.1-3.html, свободный. Яз. рус. (дата обращения 16.06.2012).

А.А. Литвинов, М.В. Акинин

ОПЕРАТИВНОЕ КАРТОГРАФИРОВАНИЕ МЕСТНОСТИ С ПРИМЕНЕНИЕМ МАШИН ОПОРНЫХ ВЕКТОРОВ

(Рязанский государственный радиотехнический университет)

Одной из наиболее популярных методологий машинного обучения по прецедентам является построение машины опорных векторов, известной в англоязычной литературе под названием SVM (Support Vector Machine). Этот тип методов статистического оценивания функций (или новый вид обучаемых машин) был предложен в середине 1990-х гг. Основы подхода заложены в работах В. Н. Ванника по статистической теории обучения. SVM-алгоритмы приобрели популярность благодаря многим привлекательным свойствам и перспективным практическим приложениям (в биоинформатике — при обработке информации огромных объемов; в области экономики и бизнеса — для прогнозирования временных рядов, оценивания кредитоспособности, а также в сфере финансовой безопасности). SVM-подход к задаче восстановления регрессии преодолевает "проклятие размерности". Источником повышения эффективности SVM-алгоритмов являются оптимальные схемы распределения памяти и эффективные вычислительные процедуры.

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 пространство 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