работки данных, при построении имитационных моделей сложных систем [1]. Решение задачи идентификации является одним из вариантов применения нейронных сетей. В данной работе рассматривается идентификация законов распределения многослойным персептроном, обучаемым методом обратного распространения ошибки [2] и методом QuiqProp [3].

Для решения задачи был выбран многослойный персептрон с одним скрытым слоем. Многослойный персептрон широко используется для поиска закономерностей и классификации образов. Цель обучения сети состоит в подборе таких значений весов, чтобы при заданном законе распределения на выходе получить значения сигналов, которые будут совпадать с ожидаемыми значениями. Входными данными для нейронов сети послужили значения высот столбцов гистограммы[4].

Для обучения многослойного персептрона использовались следующие методы. Метод обратного распространения ошибки – итеративный градиентный алгоритм, который используется с целью минимизации ошибки работы многослойного персептрона и получения желаемого выхода. Его минусом является неопределённо долгий процесс обучения. В данной работе также рассматривается метод быстрого распространения QuiqProp. QuiqProp – один из эвристических методов, являющийся модификацией метода обратного распространения ошибки и ускоряющий процесс обучения.

Была разработана автоматизированная система идентификации законов распределения, проведен ряд исследований работы данной системы, произведено сравнение производительности методов обучения многослойного персептрона. Интерфейс системы представлена на рисунке 1.

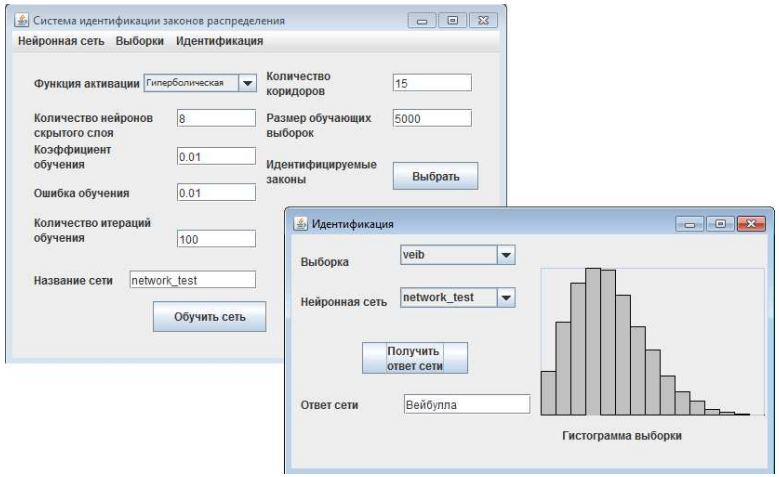

Рисунок 1 – Интерфейс системы

По результатам проведенных исследований можно сделать вывод о хороших идентификационных возможностях многослойного персептрона. Процент распознанных законов при различных параметрах сети находится в диапазоне 75-95%. Для сети с 12 нейронами в скрытом слое, обученной методом обратного распространения ошибки, процент распознанных законов распределения равен 95%. Для сети, обученной методом QuiqProp процент распознанных законов - 93%. Количество итераций обучения меньше у сети, обученной методом QuiqProp - 34 итерации против 75 по сравнению с сетью, обученной методом обратного распространения ошибки.

Литература

1. Проблемы идентификации моделей распределения случайных величин с применением современного программного обеспечения [Электронный ресурс] // http://www.rae.ru/use/?section=content&op=show_article&article_id=7981699

2. Осовский, С. Нейронные сети для обработки информациих [Текст]: учеб. - справоч., пособие / С.Осовский -М.: Издательский дом «Финансы и статистика», 2002. - 51 с.

3. Эвристические алгоритмы обучения многослойного персептрона [Электронный ресурс] // http://ai-news.ru/2015/07/evristicheskie_algoritmy_obucheniya_mnogoslojnogo_perseptrona_343719.html

4. Лёзина, И.В. Автоматизированная система идентификации законов распределения многослойным персептроном [Текст]/И.В. Лёзина, Н.А. Николаева//Наука и образование в жизни современного общества, том 5: сб. научных трудов по материалам международной научно-практической конференции 30 апреля 2015 г - 2015. - С. 77-78.

Н.И. Лиманова, М.Н. Седов

ОБ ОДНОМ МЕТОДЕ НЕЧЕТКОГО ПОИСКА ОБЬЕКТОВ В БАЗАХ ДАННЫХ НА ОСНОВЕ МЕТРИКИ ЛЕВЕНШТЕЙНА

(Поволжский государственный университет
телекоммуникаций и информатики)

Введение

В процессе межведомственного информационного обмена возникает проблема согласования основных реквизитов (ФИО, даты рождения, адреса, паспортных данных и т.п.) физических лиц в базах данных различных ведомств, обменивающихся информацией. Проблема нечеткого поиска персональной информации в базах данных приобретает наибольшую актуальность для физических лиц, у которых частично или полностью не совпадают реквизиты.

Для удобства обработки данных каждому набору реквизитов в базах данных присваивается так называемый персональный идентификационный номер (ПИН). В случае обработки или передачи данных о физическом лице вся привязка осуществляется именно к этому ПИНу. В России, к сожалению, пока нет

единой базы с реквизитами всех жителей, поэтому в разных ведомствах ведется свой отдельный реестр физических лиц и заводятся свои ПИНы. Проблема возникает при осуществлении обмена информацией о жителях между организациями, так как необходимо выполнить привязку входящих реквизитов к уже имеющимся. Для однозначной привязки необходимо выполнить интеллектуальный поиск физического лица в базе-приемнике, который должен учитывать множество факторов: и потенциальные ошибки при ручном вводе, и отсутствующие или устаревшие реквизиты, и т.п. Естественно предположить, что подобный поиск целесообразно реализовать в виде специализированного программного обеспечения [1].

Традиционно данная проблема решается путем анализа тождественности основных реквизитов физического лица. Таких реквизитов несколько: фамилия, имя, отчество, дата рождения, серия, номер паспорта и адрес. Однозначно определив совпадение существующих и новых реквизитов с помощью нечеткого поиска, можно выполнить идентификацию физического лица в базе данных. Данный метод поиска выполняется вручную только в том случае, когда объем передаваемой информации невелик (количество физических лиц не более 30). При больших объемах передаваемых данных используется автоматизированное сравнение тождественности реквизитов. Такой подход позволяет определить в среднем 50–60% от общего числа идентифицируемых физических лиц. Оставшиеся 40–50% представляют собой персональные данные, в которых частично или полностью не совпадают реквизиты. Такую информацию вручную обрабатывать еще сложнее.

Неверные результаты нечеткого поиска могут привести также к большому количеству данных в отчете для ручной отработки, к присвоению ПИНа не тому человеку и к добавлению излишних данных. Последствия таких ошибок в худшем случае могут полностью парализовать работу учреждения на неопределенное время, в лучшем – отнять более 10% рабочего времени специалистов на исправление ошибок. Так как большинство реквизитов физических лиц имеет строковый тип, то естественно предположить, что необходимый метод должен анализировать именно строковые значения.

Математическая модель

Известно несколько видов метрик, отражающих интуитивное понятие схожести строк. Наиболее распространены расстояния Хемминга, метрика Левенштейна и расстояние редактирования [2–4].

Для использования метрики Левенштейна для задач нечеткого поиска потребовалось модифицировать метрику таким образом, чтобы расстояние между строками зависело, в том числе, и от длины сравниваемых строк [3].

Теорема 1: Обозначим при помощи величины $p(s_1,s_2)$ метрику Левенштейна, а величиной $||s_i||$ – длину строки s_i. Тогда функция

$$r(s_1,s_2) = \frac{p(s_1,s_2)}{\max\{||s_1||,||s_2||\}} \qquad (1)$$

является метрикой.

Доказательство: поскольку $p(s_1,s_2)$ – метрика, то имеем:

$$p(s_1,s_2) \geq 0,\ p(s_1,s_2) = p(s_2,s_1),\ p(s_1,s_2) + p(s_2,s_3) \geq p(s_1,s_3)$$

для любых строк s_1,s_2 и s_3. Учитывая эти соотношения и равенство (1), приходим к выводу, что $r(s_1,s_2)$ удовлетворяет первым двум аксиомам, определяющим метрику. Остаётся доказать, что для любых строк s_1,s_2 и s_3 функция $r(s_1,s_2)$ удовлетворяет неравенству треугольника: $r(s_1,s_2) + r(s_2,s_3) \geq r(s_1,s_3)$.

Запишем это неравенство в виде:

$$\frac{p(s_1,s_2)}{\max\{||s_1||,||s_2||\}} + \frac{p(s_2,s_3)}{\max\{||s_2||,||s_3||\}} - \frac{p(s_1,s_3)}{\max\{||s_1||,||s_3||\}} \geq 0.$$

Возможны следующие случаи:

1. $||s_1|| \leq ||s_2|| \leq ||s_3||$ 2. $||s_2|| \leq ||s_3|| \leq ||s_1||$ 3. $||s_3|| \leq ||s_1|| \leq ||s_2||$ 4. $||s_2|| \leq ||s_1|| \leq ||s_3||$
5. $||s_1|| \leq ||s_3|| \leq ||s_2||$ 6. $||s_3|| \leq ||s_2|| \leq ||s_3||$

Рассмотрим первый случай. Имеем:

$$\frac{p(s_1,s_2)}{\max\{||s_1||,||s_2||\}} + \frac{p(s_2,s_3)}{\max\{||s_2||,||s_3||\}} - \frac{p(s_1,s_3)}{\max\{||s_1||,||s_3||\}} = \frac{p(s_1,s_2)}{||s_2||} + \frac{p(s_2,s_3)}{||s_3||} - \frac{p(s_1,s_3)}{||s_3||} \geq$$

$$\geq \frac{1}{||s_3||}(p(s_1,s_2) + p(s_2,s_3) - p(s_1,s_3)) \geq 0.$$

Таким образом, для первого случая неравенство треугольника выполняется. Поскольку второй случай аналогичен первому, на основании подобных выкладок делаем вывод, что для второго случая неравенство треугольника также выполняется.

Перейдём к рассмотрению третьего случая. Итак, в третьем случае имеем:

$$r(s_1,s_2) + r(s_2,s_3) - r(s_1,s_3) = \frac{1}{||s_2||}(r(s_1,s_2) + r(s_2,s_3)) - \frac{1}{||s_1||}r(s_1,s_3)). \qquad (2)$$

Следовательно, в третьем случае для функции $r(s_1,s_3)$ также выполняется неравенство треугольника. Остальные случаи аналогичны рассмотренным выше. Таким образом, функция $r(s_1,s_2)$ является метрикой, заданной на множестве строк. Теорема доказана.

Замечание: функция $r(s_1,s_2)$ принадлежит отрезку [0,1] для любых строк s_1 и s_2.

В предложенном алгоритме данная метрика применяется для работы со строковыми реквизитами физических лиц, к которым относятся ФИО, адрес, документ и т.д. В связи с этим построенная с использованием данной метрики лингвистическая переменная позволяет обрабатывать запросы поиска для человека, похожего на другого человека по реквизитам. Приняв от пользователя такой запрос, мы фактически получаем два значения: значение искомого реквизита и радиус поиска.

Алгоритм нечеткого поиска

В реализации алгоритма на языке PL-SQL СУБД Oracle 11g за предварительную выборку всех записей, отдаленно похожих на искомую,

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.