Формирование обучающей выборки при использовании искусственных нейронных сетей в задачах поиска ошибок баз данных

В.В. Галушка, В.А. Фатхи

Неотъемлемой частью современных информационных систем являются базы данных (БД), предназначенные для хранения и извлечения информации. Как правило, БД, находясь на нижнем уровне в структуре многоуровневой информационной системы, является источником данных для средств обработки информации и принятия решений. Соответственно среди требований к данным, хранимым в БД, на первый план выходят такие как полнота, актуальность и достоверность. Существующие подходы к обеспечению достоверности включают ограничения целостности, триггеры и использование хранимых процедур [1, 2]. Они предполагают проверку хранимых в ячейках значений на соответствие известным, заранее заданным пределам, однако даже значения в допустимых пределах могут не отражать реальные характеристики описываемого объекта или явления, приводя, таким образом, к ошибкам, выражающимся в недостоверности данных.

Более точную оценку достоверности данных, связанных с ошибками в таблицах БД можно проводить, используя методы интеллектуального анализа данных, основанных на применении искусственных нейронных сетей (ИНС) [3], ключевой особенностью которых является способность к обучению и обобщению. Особо важным этапом метода поиска ошибок в БД является этап формирования обучающей выборки (или эталонного фрагмента БД). При этом на первый план выходит необходимость обеспечения репрезентативности обучающей выборки.

Репрезентативность — соответствие характеристик выборки характеристикам популяции или генеральной совокупности в целом [4, 5]. Репрезентативность определяет, насколько возможно обобщать результаты

исследования с привлечением определённой выборки на всю генеральную совокупность, из которой она была собрана.

В контексте аналитических технологий под репрезентативностью исходных данных следует понимать наличие достаточного количества разнообразных обучающих примеров, отражающих правила и закономерности, которые должны быть обнаружены моделью в процессе обучения [6]. Она имеет три аспекта [4]:

— достаточность — число обучающих примеров должно быть достаточным для обучения. Для нейронной сети необходимо, чтобы число обучающих примеров было в несколько раз больше, чем число весов межнейронных связей, в противном случае модель может не приобрести способности к обобщению. Кроме этого, размер выборки должен быть достаточным для формирования обучающего и тестового множеств;

— разнообразие — большое число разнообразных комбинаций вход-выход в обучающих примерах. Способность ИНС к обобщению не будет достигнута, если число примеров достаточное, но все они одинаковые, т.е. представляющие лишь часть классов, характерных для исходного множества;

— равномерность представления классов — примеры различных классов должны быть представлены в обучающей выборке примерно в одинаковых пропорциях. Если один из классов будет преобладать, то это может привести к «перекосу» в процессе обучения модели, и данный класс будет определен моделью как наиболее вероятный для любых новых наблюдений [8].

Рассмотрим каждое из требований более подробно.

Исходя из свойств нейронных сетей, число нейронов входного слоя, при использовании таблицы (отношения) БД для обучения ИНС равно числу столбцов в таблице, выбранных для проведения проверки [6, 8]. Оно хранится в метаданных БД и может быть извлечено из служебных таблиц [7,

8]. Для многослойной сети число нейронов в скрытом слое должно превышать число нейронов во входном слое в 1,5 — 2 раза [9], таким образом, общее число нейронов во входном и скрытом слое составляет от 2,5n до 3n. Так как сеть полносвязная, т.е. каждый нейрон предыдущего слоя соединён со всеми нейронами следующего слоя, то число связей между 1-ым и 2-м слоями равно $n \times 2 \times n = 2 \times n^2$.

Общее же число связей для двухслойной сети:

$$K_M = 2 \times n^2 + 2 \times n \times m = 2 \times n \times (n + m),$$

где n — число входных нейронов,

m — число выходных нейронов.

Для сети Кохонена:

$$K_K = n \times m.$$

Однако число нейронов в выходном слое m является неизвестным, так как зависит от количества классов объектов [9, 10], информация о которых хранится в данной таблице. Таким образом, число связей между нейронами скрытого и выходного слоя равное $2n \times m$ невозможно определить до окончания этапа кластеризации.

При $n > 1$ и $m > 1$

$$n \times m < 2 \times n^2 + 2 \times n \times m.$$

Это означает, что для обучения сети Кохонена требуется выборка значительно меньшая, чем для обучения многослойной сети, следовательно, на первом этапе анализа данных можно использовать обучающую выборку, содержащую количество элементов, значительно превышающее минимально необходимое.

Воспользуемся эмпирическим предположением, основанным на предыдущих практических результатах, и примем число классов m равным 10, тогда:

$$K_M = 2 \times n^2 + 20 \times n.$$

Выберем K_M строк из таблицы в качестве обучающей выборки для обучения сети Кохонена, т.е. $K_K = K_M$, тогда

$$m \times n = 2 \times n^2 + 20 \times n,$$

$$m = 2 \times n + 20. \quad (1)$$

То есть, сеть Кохонена обученная на K_M примерах должна иметь возможность выделить до $2n$+20 кластеров, что является достаточным для большинства задач и предметных областей.

Следующие два требования — разнообразие и равномерность возможно обеспечить случайным выбором строк из таблицы для обучающей выборки [11].

Однако метод случайного выбора не может гарантировать стопроцентное выполнение указанных условий. Присутствующий элемент случайности, особенно при достаточно большом количестве классов может вносить значительные погрешности, в том числе, приводящие к нерепрезентативности обучающей выборки. В некоторой степени устранить данный недостаток можно путём формирования 2-х обучающих выборок. В случае если в результате кластеризации двух выборок получено одинаковое число кластеров, можно говорить об их достаточной репрезентативности.

Рассмотрим результаты, полученные при использовании разработанного метода для поиска ошибок в тестовой БД, которая содержит 485 строк с информацией о деятельности некоторой транспортной компании, в 5 из которых намеренно внесены ошибки. Известно, что в таблице представлена информация о n = 3 кластерах различных объектов, тогда, в соответствии с формулой (1) m = 26 строк. Так как объем обучающей выборки должен превышать полученное значение в несколько (т.е. минимум в 2) раз и, так как в соответствии с разработанным методом [1], используются 2 выборки, то обще число строк равно 26 × 2 × 2 = 104.

Номера строк с ошибками заранее известны, необходимо переместить данные строки как можно ближе к началу таблицы, что позволит проверить их в первую очередь.

На графике (рис. 1) видно, что наилучший результат достигается при объёме обучающей выборки, составляющем 20% от всей таблицы, то есть 485 × 0.2 = 97 строк, что соответствует рассчитанному ранее объёму выборки.

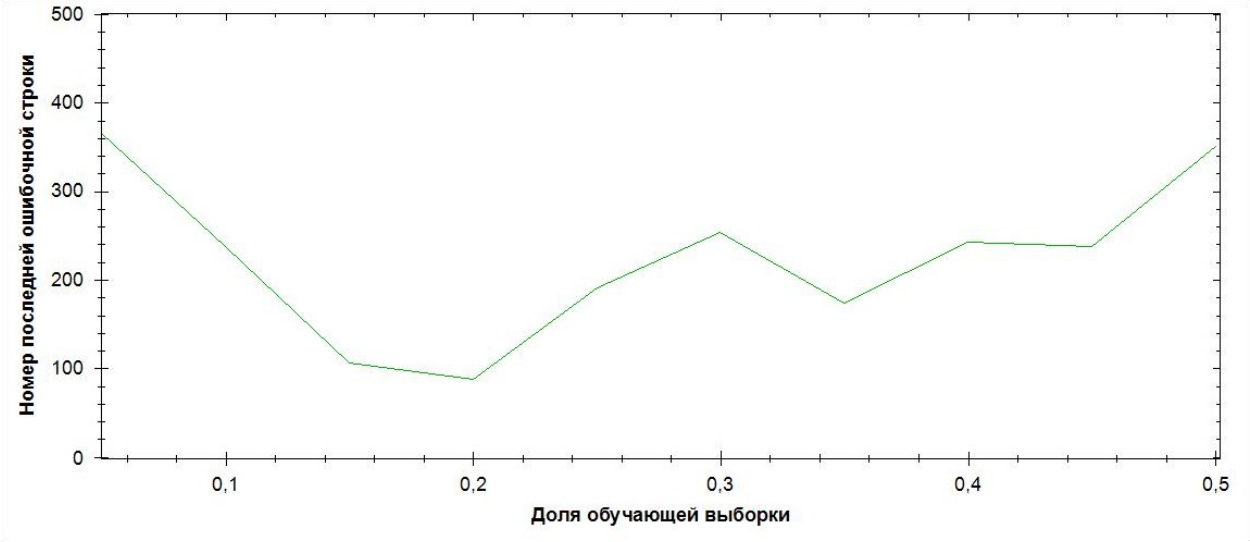

Рис. 1 — Зависимость положения ошибочной строки от объема обучающей выборки

Для указанного объема обучающей выборки были проведены эксперименты по обучению ИНС и классификации строк тестовой таблицы. В результате каждой строке было поставлено в соответствие некоторое число — уверенность ИНС в принадлежности строки к одному из классов данных, выделенных в обучающей выборке. Далее таблица была отсортирована по данному в порядке возрастания данного критерия. График распределения ошибочных строк в таблице на рис. 2. Из него видно, что все строки, содержащие ошибки находятся ближе к началу таблицы. Это означает, что проверка достоверности небольшой части отсортированных строк характеризует достоверность всей таблицы.

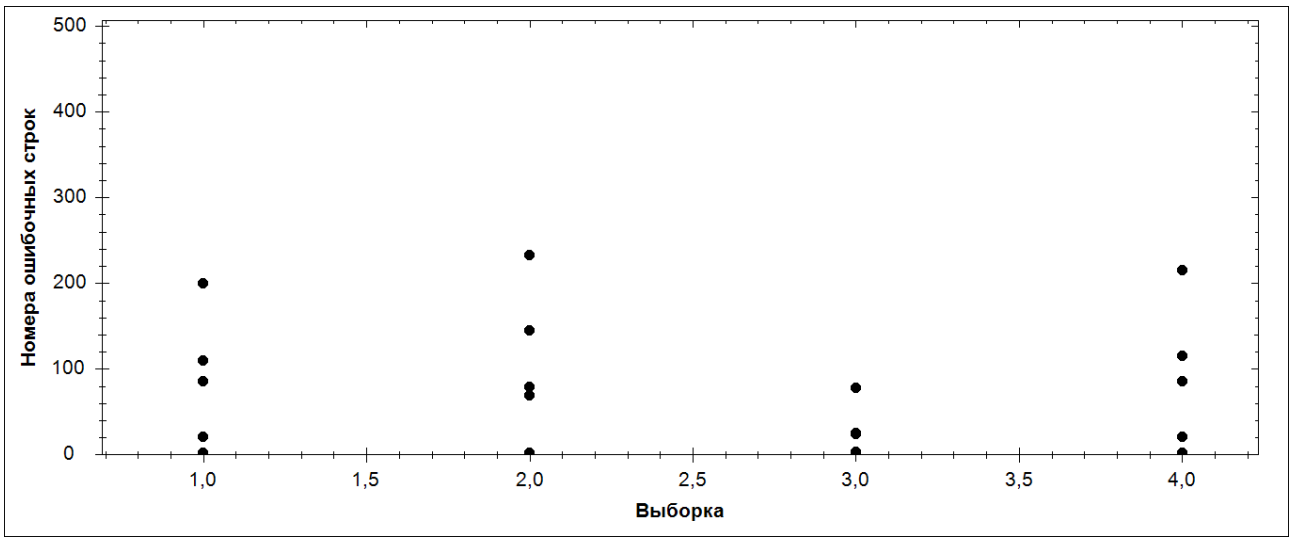

Рис. 2 — Распределение ошибочных строк

Таким образом, результаты экспериментов подтверждают теоретические расчёты достаточного объёма обучающей выборки, а также эффективность случайного выбора строк из таблицы БД, и позволяют сделать вывод о том, что в случае наличия в обучающей выборке достоверных данных, значение принадлежности строки к какому-либо классу является адекватным критерием её достоверности.

Литература

1. Карпова Т.С. Базы данных: модели, разработка, реализация [Текст]. — СПб.: Питер, 2001. — 304 с.; ил.
2. Nicolas J.-M. Logic for improving integrity checking in relational databases, Acta Informica, 18:3 (1982), p. 227 — 253.
3. Галушка В.В., Молчанов А.А., Фатхи А.А. Применение многослойных радиально-базисных нейронных сетей для верификации реляционных баз данных [Электронный ресурс] // «Инженерный вестник Дона», 2012, №1. – Режим доступа: http://ivdon.ru/magazine/archive/n1y2012/686 (доступ свободный) – Загл. с экрана. – Яз. рус.
4. Репрезентативность данных (Representativeness of data) // BaseGroup Labs — Глоссарий [Электронный ресурс] URL:

http://www.basegroup.ru/glossary/definitions/representativeness/ (дата обращения 26.07.2011)

5. Кобзарь Л.И. Прикладная математическая статистика. Для инженеров и научных работников [Текст]. — М.: Физматлит. — 2006 г. — 814 с.
6. Барсегян А.А., Куприянов М.С., Кузнецов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP [Текст] — 2-е изд. перераб. и доп. — СПб.: БХВ-Петербург, 2007. — 384 с.: ил.
7. Сахил Малик Microsoft ADO.NET 2.0 для профессионалов = Pro ADO.NET 2.0. [Текст] — М.: «Вильямс», 2006. — с. 560.
8. Ian H. Witten, Eibe Frank and Mark A. Hall Data Mining: Practical Machine Learning Tools and Techniques. — 3rd Edition. — Morgan Kaufmann, 2011. — p. 664.
9. Саймон Хайкин Нейронные сети: полный курс = Neural Networks: A Comprehensive Foundation. [Текст] — 2-е. — М.: «Вильямс», 2006. — с. 1104.
10. С.П. Алёшин, Е.А. Бородина Нейросетевое распознавание классов в режиме реального времени [Электронный ресурс] // «Инженерный вестник Дона», 2013, №1. — Режим доступа: http://www.ivdon.ru/magazine/archive/n1y2013/1494 (доступ свободный) – Загл. с экрана. — Яз. рус.
11. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. [Текст] — Новосибирск: ИМ СО РАН. — 1999. — 270 с.