УДК 004.8

# ИНТЕЛЛЕКТУАЛЬНЫЙ МЕТОД ВЕРИФИКАЦИИ БАЗ ДАННЫХ

**Галушка В.В., аспирант, ассистент; Молчанов А.А., к.т.н., ст. преподаватель**
*(Донской государственный технический университет, г. Ростов-на-Дону, Россия)*

В современных средствах автоматизации и телекоммуникации для хранения данных обычно применяют базы данных (БД). Из практики известно, что с увеличением объема данных в БД повышается вероятность «вкрадывания» в БД недостоверных данных, не соответствующих действительности, которые при использовании могут привести к негативным последствиям (сбой системы, нарушение технологического процесса, выход из строя оборудования, принятие неоптимальных управленческих решений и т.д.), поэтому задача верификации – проверки содержащихся данных в БД на соответствие реальности – является актуальной.

Для обеспечения достоверности хранимых данных в БД главным образом уделяют внимание механизмам работы ограничений целостности и способам их тестирования. В [1] приведен обзор методов и программных средств верификации реляционных БД, а также описывается программа, реализующая подход к обеспечению автоматизированной верификации ограничений целостности БД на основе проверки соответствия формальной спецификации БД и существующих в исходных кодах ограничений целостности и триггеров. Однако, целостность БД принципиально не в состоянии контролировать достоверность БД.

Основными способами обеспечения достоверности информации БД является входной контроль вводимых данных со стороны системы управления БД или прикладного приложения с помощью триггеров или хранимых процедур либо со стороны человека, да и то в ограниченных масштабах, поскольку в ряде случаев люди не обладают полнотой знаний о реальном мире.

В данной работе для определения достоверности вводимого кортежа (строки) в таблицу БД предлагается интеллектуальный метод, основанный на теории искусственных нейронных сетей [2].

Приведем постановку задачи в терминах реляционной алгебры [3]. Имеется БД, представляющая совокупность отношений. Рассмотрим $n$-арное отношение $R$, которое является подмножеством полного декартова произведения $D_1 \times D_2 \times ... \times D_n$ множеств доменов $D_1$, $D_2$, …, $D_n$ ($n \geq 1$), не обязательно различных. Каждый элемент отношения $R = \{r_1, r_2, ..., r_j, ..., r_m\}$ ($m \geq 0$) является кортежем, включающим в себя элементы множеств $D_1$, $D_2$, …, $D_n$: $j$-ый элемент отношения $R$ равен $\langle a_{j,1}, a_{j,2}, ..., a_{j,n} \rangle$, $a_{j,1} \in D_1$, $a_{j,2} \in D_2$, ..., $a_{j,n} \in D_n$. Отношение может быть представлено в виде таблицы, в которой столбцы (поля, атрибуты) соответствуют вхождениям

доменов в отношение, а строки (записи) – наборам из $n$ значений, взятых из исходных доменов. Каждому элементу $r_j$ поставим в соответствие параметр $d(r_j)$ – достоверность данного элемента такое, что $0 \le d(r_j) \le 1$. Необходимо для каждой новой строки $r_{m+1}$, заносимой в таблицу, оценить ее достоверность $d(r_{m+1})$.

Для простоты описания метода определения достоверности вводимого кортежа ограничимся проверкой отношений, содержащих только числовые данные. В случае необходимости поля строкового типа могут быть исключены или заменены числовыми кодами.

Первым этапом верификации является выявление кластеров – классов объектов похожих между собой в пределах группы и максимально отличающихся от объектов, принадлежащих другой группе. Для определения принадлежности элемента к кластеру будем использовать сеть Кохонена [2]. При этом, количество кластеров $T$ соответствует количеству нейронов выходного слоя и определяется в соответствии с методикой, описанной в [4]. В качестве обучающей выборки используются строки БД, достоверность которых равна единице. Результатом кластеризации сетью Кохонена будут для каждого элемента $r_j$ пары $(r_j, C_k)$, где $C_k$ – кластер, к которому принадлежит элемент $r_j$ ($k = 1, 2, ..., T$). Полученное множество пар используется на следующем этапе верификации при обучении радиально-базисной нейронной сети.

Для использования радиально-базисной нейронной сети определим количество нейронов во входном и выходном слое, количество скрытых слоев и нейронов в каждом из них, функцию активации нейронов.

Количество нейронов во входном слое определяется количеством $n$ столбцов анализируемой таблицы. Количество нейронов скрытого слоя определяется для каждого случая отдельно, однако проведенные ранее исследования показывают, что достаточным является количество в 1,5–2 раза превышающее количество нейронов в выходном слое [5]. Количество нейронов в последнем (выходном) слое соответствует количеству кластеров $T$, определённых на предыдущем этапе с помощью сети Кохонена.

Текущее состояние нейрона определяется как взвешенная сумма его входов. Выход нейрона будет определяться как логистическая функция вида $f(s) = (1 + e^{-\alpha s})^{-1}$. В результате обучения радиально-базисной сети методом обратного распространения ошибки с использованием полученного множества пар $(r_j, C_k)$ будет получена сеть, способная классифицировать объекты, информация о которых хранится в анализируемой таблице. Благодаря способности нейронных сетей к обобщению, т.е. распространению выделенных знаний на неизвестные ранее образцы, обученную сеть можно применять для классификации объектов, не входивших в достоверную БД. При этом степень уверенности в принадлежности образца, поданного на вход, к определённому классу определяется максимальным значением выхода нейронов выходного слоя. «Низкое» значение $d$ означает, что поданный на вход образец не относится ни к одному из классов с достаточной степенью уверенности и, следовательно, может содержать ошибки или неточности. Данный параметр можно использовать в качестве значения оценки достоверности.

Помимо рассмотренного случая, возможен вариант, когда имеется таблица, содержащая большое количество строк, достоверность которых неизвестна и её требуется оценить. Данный случай можно свести к описанному ранее, если в качестве подготовительного этапа верификации выбрать из таблицы некоторое число строк и про-

верить их вручную, получив, таким образом, достоверные строки, которые можно использовать в качестве обучающей выборки.

Предложенная методика может применяться для решения задач верификации различных БД любого типа информационных систем.

Перечень ссылок

1. Глухарёв М.Л., Косаренко А.П., Хоменко А.Д. Программа для автоматической верификации ограничений целостности баз данных. Программные продукты и системы. – № 1 – 2011.
2. Каллан Р. Основные концепции нейронных сетей: Пер. с англ. – М. : Издательский дом «Вильямс», 2001. – 287 с. ISBN 5-8459-0210-X
3. Райордан Р. Основы реляционных баз данных /Пер. с англ., М.: Издательско-торговый дом «Русская редакция», 2001, 384 с.
4. Галушка В.В., Фатхи Д.В. Методика определения оптимального числа нейронов выходного слоя сети Кохонена при решении задач кластеризации. Информационная безопасность регионов. – № 2 – 2011, с. 41-44
5. Галушка В.В., Фатхи В.А. Программная модель для исследования возможностей применения искусственных нейронных сетей в агропромышленном комплексе, Состояние и перспективы развития сельскохозяйственного машиностроения: материалы междунар. науч.-практ. конф. г. ДГТУ, Ростов н/Д, 2011