

Обработка текстов

А.В. ГАВРИЛОВ, В.М. КАНГЛЕР, С.А. ЗАЙЦЕВ

Новосибирский государственный технический университет

E-mail: Avg@vt.cs.nstu.ru

ПРОГРАММА АНАЛИЗА БАЗ ДАННЫХ С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

Аннотация

В статье приводится описание программного обеспечения позволяющего проводить анализ данных с помощью различных моделей нейронных сетей. В программе реализовано решение, позволяющее использовать ее в качестве графической оболочки для изучения и использования различных моделей нейронных сетей без изменения ее кода. Данное свойство позволяет выбрать наиболее подходящую модель для решения конкретной задачи или подключить собственный модуль нейронной сети.

Постановка задачи

Идея применения искусственных нейронных сетей все чаще возникает при решении различных типов задач. Но реализация этой идеи требует наличия специальной подготовки (нужно правильно выбрать модель нейронной сети и ее параметры, разработать программное обеспечение для работы с этой моделью или освоить достаточно сложное и дорогое ПО, каким являются оболочки для работы с нейронными сетями), что затрудняет использование аппарата нейронных сетей, особенно в нетехнических областях применения. В статье рассмотрена архитектура программного обеспечения, при разработке которого сделана попытка упростить применение искусственных нейронных сетей для анализа баз данных, и которое позволяет использовать различные модели искусственных нейронных сетей при решении различных классов задач анализа данных.

При разработке архитектуры мы исходили из того, что программное обеспечение должно удовлетворять следующим требованиям:

- Программа должна обеспечивать проведение различных видов анализа данных: прогноза, ассоциативного поиска, кластеризации и распознавания.
- Программа должна выполнять роль тренажера, с помощью которого будут изучаться и разрабатываться модели нейронных сетей применительно к анализу данных (решению вышеперечисленных задач).
- В качестве исходных данных для анализа используется содержимое реляционных баз данных, так как это достаточно удобный и универсальный источник данных.
- Программа должна содержать средства для оценки качества проведенного анализа, которые должны позволять определить характер влияния параметров модели на результат работы сети.
- Программа должна состоять из оболочки и динамически подключаемых модулей. Оболочка должна обеспечивать взаимодействие пользователем и базами данных и поддержку этапов анализа данных — создания, обучения и тестирования (использования) сети. Модули должны реализовывать различные модели нейронных сетей и поддерживать интерфейс взаимодействия с оболочкой. Таким образом, после создания оболочки можно будет сосредоточиться на работе с моделями сетей, не отвлекаясь на решение вопросов взаимодействия с пользователем.
- Для взаимодействия оболочки с динамическими модулями необходим унифицированный интерфейс, который должен позволять использовать в программе как можно большее число моделей нейронных сетей.

Архитектура программы

Программа состоит из оболочки, реализующей функциональность, общую для различных задач (доступ к исходным данным, получение команд пользователя, вывод результатов работы и т.д.), и из модулей нейронных сетей. Такое разделение позволяет упростить работу пользователя, так как ему предлагается единый интерфейс для использования различных моделей нейронных сетей, а также освободить разработчика новых модулей нейронных сетей от необходимости реализовывать пользовательский интерфейс.

Чтобы программа могла работать со множеством моделей нейронных сетей было реализовано следующее решение. Каждая модель представляет собой динамически линкуемую библиотеку, которая должна поддерживать интерфейс взаимодействия с оболочкой. Библиотека должна уметь решать

одну или несколько задач анализа (прогноз, кластеризация, распознавание и ассоциативный поиск) и для каждого из них поддерживать свой протокол обмена данными. Чтобы сделать интерфейс максимально гибким по отношению к расширению набора моделей, библиотека с моделью нейронной сети реализована как активное звено. Для этого применяется взаимное связывание оболочки и библиотеки: библиотека экспортирует функции, реализующие выполнение высокоуровневых команд (например, запустить работу сети, обучить сеть, задать параметры сети, сохранить сеть и прочие), число которых не превышает десяти; оболочка, в свою очередь, при создании нейронной сети передает в библиотеку набор указателей на данные и методы, которые позволяют получить исчерпывающие данные о выполняемой пользователем операции, а также управлять процессом чтения исходных данных

Программа-оболочка решает следующие задачи.

1. Чтение исходных данных из базы данных. Данные выбираются с помощью SQL-запроса. Поддерживается автоматизированное создание запроса, не требующее знания SQL.
2. Взаимодействие с пользователем (получение команд и вывод результата) для решения задач прогнозирования временных рядов, распознавания, классификации и ассоциативного поиска.
3. Работа с различными типами данных: целые числа, вещественные числа с плавающей и фиксированной (денежный формат) точкой, строки.
4. Предварительная обработка данных. Для повышения качества анализа исходные данные могут быть предварительно обработаны. Используется два типа предобработки: интервальное кодирование и вычисление тренда. Интервальное кодирование группирует близкие значения, что снижает чувствительность к шуму в исходных данных. Тренд применяется только для решения задачи прогноза и является первой производной. Тренд может быть полезен, когда зависимость проще обнаружить в изменении величины, нежели в ее значении. Применяется два вида тренда: количественный и качественный. Количественный тренд — это изменение величины, а качественный — направление изменения
5. Поддержка словаря — множества значений полей, на которых производится обучение и использование нейронной сети. Словарь необходим для кодирования данных сетью. Например, модель Хопфилда требует двоичного кодирования информации (одним из вариантов может быть использование номера в словаре).

6. Оценка качества проведенного анализа. При решении всех типов задач формируются отчеты, позволяющие пользователю оценить результат. При прогнозировании на тренировочной выборке производится вычисление ошибки прогноза и формирование отчета по ошибкам. Также строится график или диаграмма плотности распределения ошибки прогнозирования (рис. 1.).

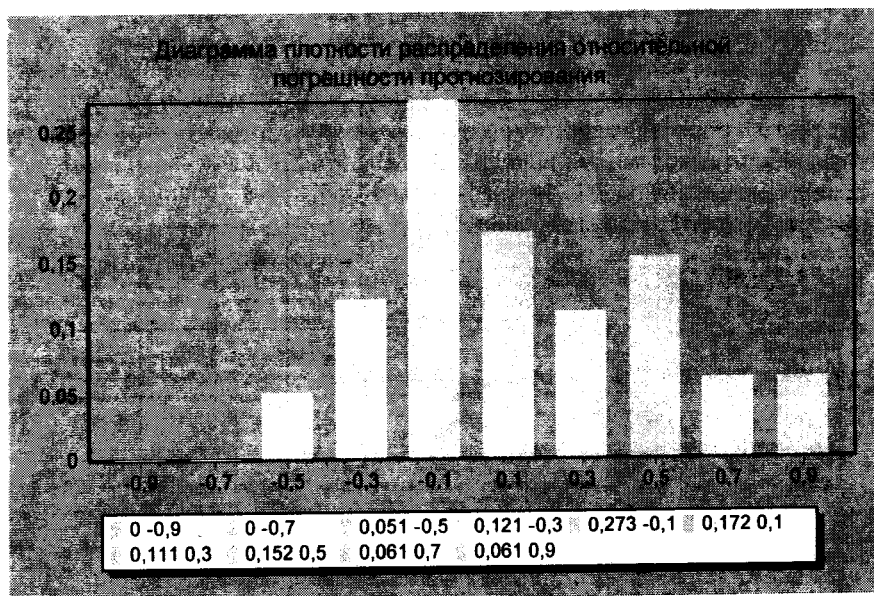


Рис. 1. Диаграмма плотности распределения относительной погрешности прогнозирования

7. Динамическое подключение модулей нейронных сетей.

Использование программы

Работа с программой состоит из двух этапов: подготовительного и основного. Подготовительный этап состоит из создания и обучения нейронной сети (а также из шагов, необходимых для создания сети), второй этап — это этап использования сети и оценки результатов ее работы.

Подготовительный этап состоит из следующих шагов:

1. Создание нового проекта. В основе проекта лежит SQL-запрос, определяющий данные, на которых будут обучаться нейронные сети, и которые будут служить исходными данными для работы сети. Для создания проекта нужно выбрать источник данных — это может быть BDE-алиас или папка, в которой расположены таблицы DBF. После того как источник данных (база данных) выбран, нужно создать SQL-запрос, извлекающий нужные данные.

2. Создание словаря.

3. Создание нейронной сети. Для создания нейронной сети необходимо выбрать нужную модель сети. Затем выбирается тип задачи, которую будет решать модель. После этого выбираются поля, которые связаны с входами и/или выходами нейронной сети. Выделив входное или выходное поле, можно выбрать тип предварительной обработки данных. На рис.2 показана форма создания нейронной сети.

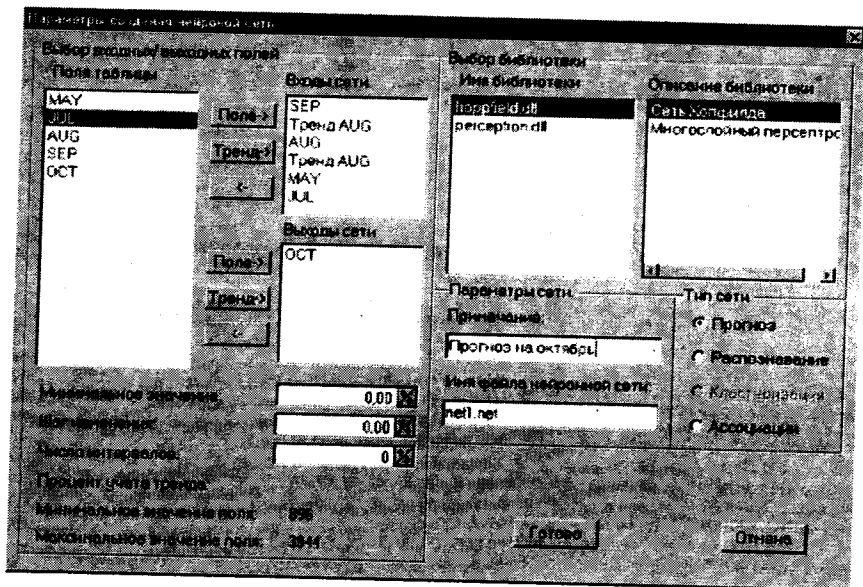


Рис.2. Окно создания нейронной сети

4. Обучение нейронной сети. Для обучения нужно задать подмножество записей, на которых будет производиться обучение. После обучения сеть готова к использованию.

Основной этап заключается в использовании обученной сети. Для анализа данных, используются разные формы в зависимости от решаемой сетью задачи.

Заключение

На данный момент описываемая программа находится в состоянии тестирования и доработки. Но уже сейчас с ее помощью проводятся эксперименты с моделями Хопфилда и многослойного перцептрона.

Оболочка может использоваться для отладки программных моделей нейронных сетей и для оценки влияния различных параметров на качество анализа.

Дальнейшее развитие проекта планируется в направлении оптимизации уже созданных модулей и разработки библиотек, содержащих различные модели нейронных сетей.

Работа частично поддержана грантом Минобразования ГОО-4.1-48.

Список литературы

1. Гаврилов А.В., Канглер В.М., Катомин М.Н., Коротенко А.И. Обнаружение ассоциативных взаимосвязей между полями в базах данных с использованием нейронной сети // Тр. Междунар. науч.-техн. конф. «Научные основы высоких технологий». Новосибирск, 1997. Т.2. С.210.
2. Гаврилов А.В., Канглер В.М. Использование искусственных нейронных сетей для анализа данных // Сб. науч. трудов НГТУ. Новосибирск: Изд-во НГТУ. 1999. № 3(16). С.56-63.
3. GavriloV A.V., Kangler V.M. The use of Artificial Neural Networks for Data Analysis // The Third Russian-Korean International Symposium on Science and Technology. Novosibirsk: NSTU. 1999. Proceedings. Vol.1. P.257-260; Abstracts. Vol.1. P.192.
4. Гаврилов А.В., Губарев В.В. Применение модели Хопфилда для решения задачи прогнозирования на примере анализа притока реки Обь // 2-я Всероссийская научно-техн. конф. «Нейроинформатика-2000». М., 2000. С.33-38.
5. Продукты для интеллектуального анализа данных // Рынок программных средств. 1997. № 14, 15. С.32-33.
6. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. М.: Горячая линия — Телеком. 382 с.