1 • 2005

РОССИЙСКАЯ
АССОЦИАЦИЯ
ИСКУССТВЕННОГО
ИНТЕЛЛЕКТА

НОВОСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ТЕМА НОМЕРА

МОДЕЛИРОВАНИЕ СИСТЕМ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

URSS

Задача уменьшения размерности пространства исходных данных при прогнозировании характера течения острого панкреатита

Г. Ф. Филаретов, Д. С. Лебедев

Аннотация. В статье рассмотрена возможность сокращения размерности пространства исходных данных методом сжатия исходной информации с некоторыми несущественными потерями с помощью многослойной автоассоциативной нейронной сети с нелинейными функциями активации. Проведено сравнение эффективности сжатия с помощью автоассоциативной нейронной сети и методом главных компонент. Показана высокая эффективность сжатия исходных данных с помощью автоассоциативной нейронной сети.

Введение

Работа посвящена исследованию проблемы сокращения размерности пространства исходных данных при прогнозировании характера течения острого панкреатита (ОП) методом нейронных сетей (НС).

При обработке информации нередко встречаются ситуации, когда объект описывается очень большим числом признаков. Как правило, эти признаки взаимосвязаны (коррелированы) и в информационном смысле избыточны. Размерность данных можно уменьшить двумя способами:

1) удалить переменные, которые несут наименьшее количество информации для разделения объектов прогнозирования на классы,
2) трансформировать пространство переменных с высокой размерностью в пространство переменных с меньшей размерностью с минимальными потерями полезной информации.

С целью оценки информативности количественных переменных для разделения объектов на классы можно использовать *t*-критерий. В этом случае сравниваются средние значения переменных в разных классах с последующим отбором тех из них, которые достоверно различаются. Применение *t*-критерия требует нормальности распределения переменных внутри групп и равенства дисперсий в груп-

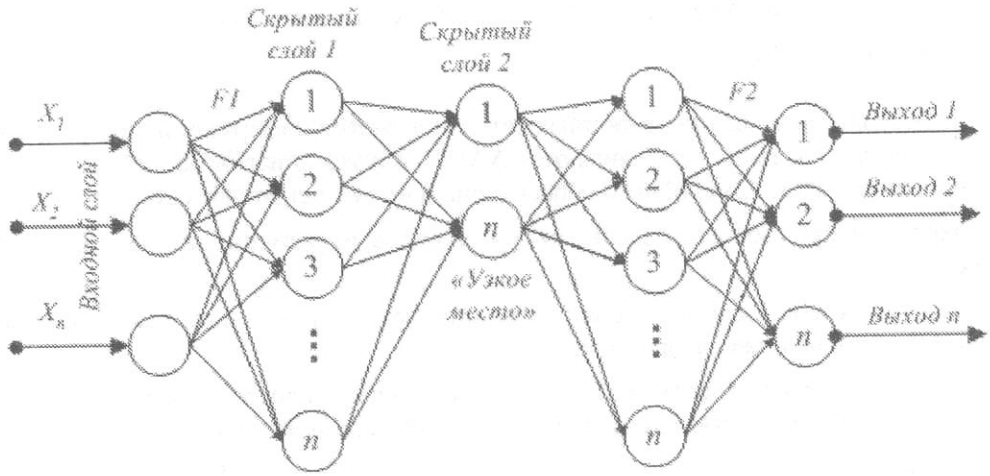

Рис. 2

При использовании во всех слоях нелинейных передаточных функций сеть может осуществлять нелинейное преобразование при сжатии, если исходные данные обладают нелинейной структурой. Если данные имеют линейную структуру, то сеть будет осуществлять линейное преобразование, подобное МГК.

Показатель информативности данных после сжатия методом ААНС рассчитывается по следующей формуле:

$$\tilde{I}_m^{(1)} = \frac{D(OUT_1) + D(OUT_2) + \ldots + D(OUT_n)}{D(X_1) + D(X_2) + \ldots + D(X_n)}, \quad (2)$$

где $D(OUT_k)$ — дисперсия значения на выходе k нейронной сети ($k = 1,\ldots,n$), когда в «узком месте» имеется m нейронов.

С учетом ошибки обучения сети в качестве показателя информативности данных после сжатия рекомендуется использовать формулу:

$$\tilde{I}_m^{(2)} = 1 - \frac{D(e_1) + D(e_2) + \ldots + D(e_n)}{D(X_1) + D(X_2) + \ldots + D(X_n)}, \quad (3)$$

где $D(e_k)$ — дисперсия ошибки воспроизводимости e_k на выходе k нейронной сети: $e_k = (T_k - OUT_k)$; T_k — требуемое выходное значение на выходе k нейронной сети.

Основные положения данной работы иллюстрируются на примере нейросетевой модели прогнозирования характера течения острого панкреатита (ОП). Задача прогнозирования была сформулирована как задача классификации. С целью классификации была разработана НС типа многослойного персептрона. Исходными данными являлись 24 переменные, характеризующие тяжесть патологического процесса (данные клинического, инструментального и лабораторного методов обследования). По характеру течения ОП все больные были разделены на 2 класса.

- Класс 1 — тяжелый характер течения заболевания (инфицированный и неинфицированный панкреонекроз, органная недостаточность);
- Класс 2 — нетяжелый характер течения заболевания (отечный панкреатит).

Все примеры выборки разделены на обучающее и тестовое множество. Число обследованных больных составило 57. Каждый больной обследовался при поступлении, в 1, 2 и 3 сутки госпита-

лизации. Число примеров составило — 228 (57 × 4). Все примеры были рандомизированы. В обучающее множество вошли 156 примеров. Перед построением нейронной сети провели оценку достаточности объема выборки с учетом числа требуемых входов сети, равному числу выбранных переменных. Оценка числа исправлений весовых коэффициентов (числа степеней свободы) сети равна 2^{24}. Для выборки из 228 примеров число входных сигналов сети (признаков заболевания) должно быть примерно 7–8 ($2^8 = 256$). Число выбранных переменных избыточно. Резкое уменьшение числа признаков (с 24 до 8) методом простого исключения не представляется возможным, так как по мере увеличения выборки может измениться влияние переменных. В связи с этим была поставлена и решена задача уменьшения числа переменных методом сжатия с потерями. Сжатию подвергнуто пространство из 12 лабораторных показателей методом главных компонент и ААНС.

Целью данной работы является сопоставление информативности переменных после сжатия методом главных компонент и с помощью ААНС. Сопоставление провели исходя из показателей, рассчитанных по формулам (1), (2) и (3).

Для решения задачи сжатия данных была построена и обучена четырехслойная ААНС, с двенадцатью входными нейронами (n), пятнадцатью нейронами первого слоя (h) и тремя нейронами второго слоя (m) ($n = 12$; $h = 15$; $m = 3$). Обучение сети проводилось с помощью алгоритма обратного распространения ошибки. После обучения провели прогон входной последовательности и фиксировали значения на выходах нейронов второго слоя. В табл. 1 приведены рассчитанные средние значения информативности данных после сжатия методами МГК и ААНС.

Таблица 1

Информативность данных после сжатия методом главных компонент и ААНС

m	I_m	$\tilde{I}_m^{(1)}$	$\tilde{I}_m^{(2)}$
1	0,24	0,32	0,34
2	0,42	0,62	0,50
3	0,56	0,67	0,70

Нетрудно заметить, что информативность данных после выделения одинакового числа главных компонент методом ААНС заметно выше, чем для МГК, ($I_m \prec \tilde{I}_m^{(1)} \prec \tilde{I}_m^{(2)}$).

Расположение точек в двух- и трехмерном пространствах, получаемых на выходах нейронов второго слоя при $m = 3$, показано на рис. 3.

Рассеяние экспериментальных точек на плоскостях главных компонент, полученных методом главных компонент в двух- и трехмерном пространствах при $m = 3$, показано на рис. 4.

На рис. 3 и 4 видно, что результаты, полученные с помощью ААНС, подобны результатам, полученным МКГ, однако информативность данных после сжатия методом ААНС существенно выше, чем информативность данных, полученных МГК. Можно предположить, что ААНС обнаруживает в массиве исходных данных нелинейные эффекты. Первые три главные компоненты, выделенные ААНС, описывают примерно 67 % общего рассеивания исходных данных, тогда как первые три главные компоненты, выделенные МГК — примерно 56 . Степень информативности данных после сжатия ААНС остается достаточно высокой даже при выделении 2 главных компонент.

В результате сжатия методом ААНС число значений лабораторных показателей, подаваемых на вход классифицирующей нейронной сети, было уменьшено с 12 до 3 главных компонент, а общее число значений, подаваемых на вход сети, — с 24 до 15.

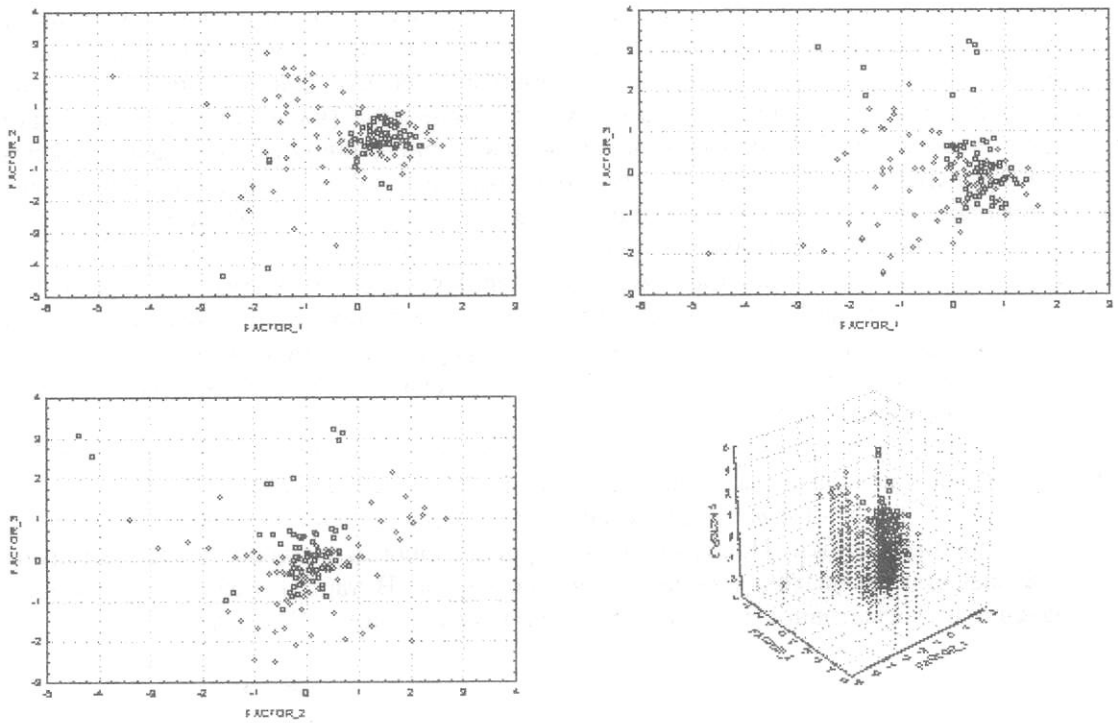

Рис. 3

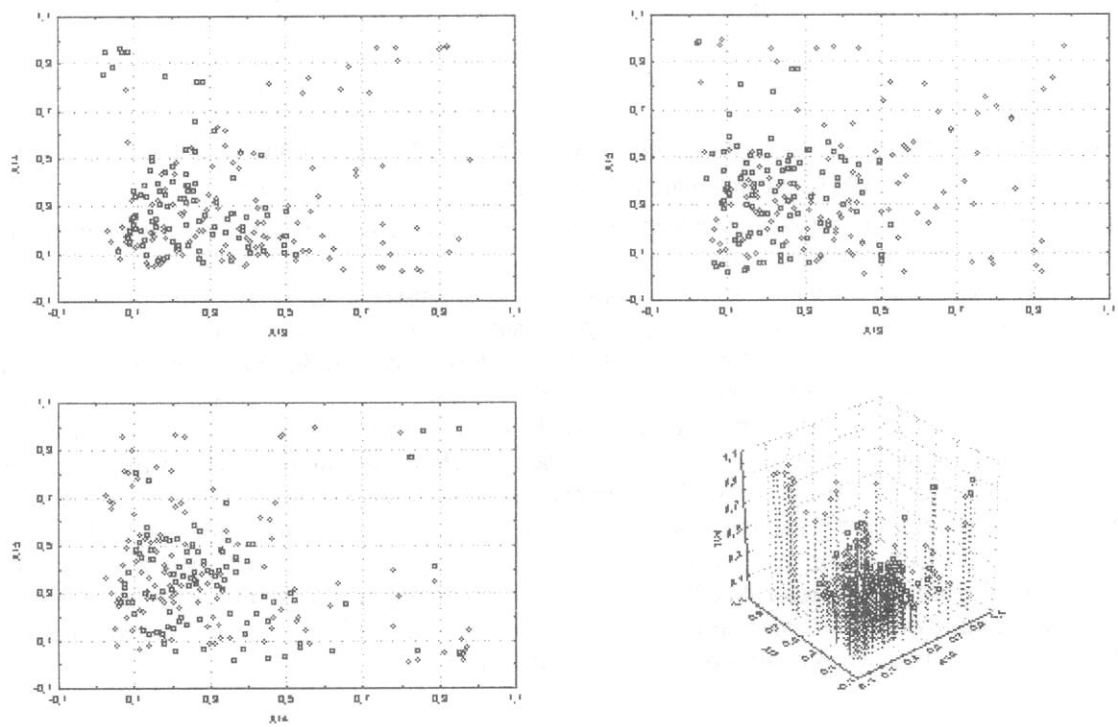

Рис. 4

Заключение

В статье была рассмотрена возможность сокращения пространства исходных данных методом сжатия с потерями с помощью многослойной ААНС с нелинейными функциями активации. Показано, что с помощью многослойной ААНС при выделении равного числа главных компонент может быть получена большая информативность сжатых данных, чем с помощью МГК.

Литература

1. *Айвазян С. А., Мхитарян В. С.* Прикладная статистика и основы эконометрики. Учебник для вузов. М.: ЮНИТИ, 1998. 1022 с.
2. *Сошникова Л. А., Томашевич В. Н., Уебе Г., Шеффер М.* Многомерный статистический анализ в экономике: Учеб. пособие для вузов. М.: ЮНИТИ-ДАНА, 1999. 598 с.
3. *Филаретов Г. Ф., Джордан Б.* Применение автоассоциативных нейронных сетей для сжатия информации. XXX международная конференция «Информационные технологии в науке, образовании, телекоммуникации, бизнесе». Украина, Крым, Ялта — Гурзуф, 19–28 мая 2003 г.
4. *Bourland H., Kamp Y.* Auto-association by Multilayer Perceptrons and Singular Value Decomposition. Biological Cybernetics, 59: 291–294, 1988.
5. *Oja E.* Principal components, minor components, and linear neural networks. Neural networks, 5: 927–935, 1992.

Филаретов Геннадий Федорович. Доктор технических наук, и. о. Генерального директора Государственного научно-исследовательского института системной интеграции.

Лебедев Дмитрий Сергеевич. Врач анестезиолог-реаниматолог высшей категории, заместитель директора Медицинского центра Банка России. E-mail: Lebedev@medcenter.msk.ru.