БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Истомина Т.В., Киреев А.В., Истомина Е.В.* Особенности измерения и интерпретации параметров ПЭС биологических объектов // Методы, средства и технологии получения и обработки измерительной информации: труды Международной научно-технической конференции. – Пенза: Информационно-издательский центр ПензГУ, 2008. – 174 с.
2. *Попечителев Е.П.* Методы медико-биологических исследований. Системные аспекты: Учебн. пособие. – Житомир: ЖИТИ, 1997. – 186 с.
3. *Долгова И.А., Чувыкин Б.В.* Способ экспресс-измерения температуры // Медицинская техника. – 2009. – №1. – С. 12-15.
4. *Фридман А.Э.* Основы метрологии. Современный курс. – СПб.: НПО «Профессионал», 2008 . – 284 с.
5. *Вальд А.* Последовательный анализ. – М.: Физматгиз, 1960.

Истомина Татьяна Викторовна
Пензенская государственная технологическая академия.
E-mail: istom@mail.ru.
440605, г. Пенза, Пр. Байдукова/ул. Гагарина, д.1а/11, тел.: (8412)496155.
Кафедра ИТММБС, зав. кафедрой, профессор, д.т.н.

Istomina Tatiana Viktorovna
Penza state technological academy.
E-mail: istom@mail.ru.
1a/11, Bajdukova /Gagarina, Penza, 440605, Russia, Phone: (8412)496155.
Head of department ITMMBS of PSTA, professor, Doctor of Science.

Ординарцева Наталья Павловна
Пензенская государственная технологическая академия.
E-mail: nat@rclink.ru.
440605, г. Пенза, пр. Байдукова/ул. Гагарина, д.1а/11, тел.: (8412)496155.
Кафедра ИТММБС ПГТА, доцент, к.т.н.

Ordinartseva Natalia Pavlovna
Penza state technological academy.
E-mail: nat@rclink.ru.
1a/11, Bajdukova /Gagarina, Penza, 440605, Russia, Phone: (8412)496155.
Department ITMMBS of PSTA, associate professor, Cand. Eng. Sc.

УДК 621.391.26

А.В. Киреев

НЕКОТОРЫЕ ВОПРОСЫ ПОСТРОЕНИЯ НЕЛИНЕЙНЫХ РЕШАЮЩИХ ФУНКЦИЙ: МЕТОД РАСПРЯМЛЕНИЯ

Предложен метод сокращения размерности пространства описаний, отличающийся гарантированной сходимостью и позволяющий выделить статистически независимые признаки. Дана оценка оптимальной степени сжатия данных.

Пространство описаний; репрезентативность; классификация; регрессия.

A.V. Kireev

SOME QUESTIONS OF CONSTRUCTION OF NONLINEAR DECISION FUNCTIONS: THE METHOD OF STRAIGHTENING

The method of reduction of dimension of space of the descriptions is offered, differing with the guaranteed convergence and allowing to allocate statistically independent attributes. The estimation of an optimum degree of compression of data is given.

Space of descriptions; representative; classification; regress.

Многие задачи анализа биомедицинских данных сводятся к построению многомерных решающих функций. В случае метрического результативного признака эти функции непосредственно используются в качестве регрессионных уравнений. Если же объясняемая переменная не является метрической, то многомерные решающие функции выступают в роли разделяющих поверхностей, являющихся основой для построения классификаторов.

В настоящее время, особенно при большой размерности пространства описаний, наибольшее распространение получили линейные решающие функции. Это связано с двумя причинами. Во первых, для их построения используются весьма эффективные алгоритмы, а во вторых – требуются обучающие выборки минимального объёма, что особенно важно при высокой размерности.

Для построения линейных регрессионных уравнений используется метод наименьших квадратов (классическая линейная регрессия [1]) или метод наименьших квадратов с регуляризацией (гребнёвая регрессия [1]). Для построения разделяющих плоскостей обычно используется линейный дискриминант Фишера [2] либо метод обобщенного портрета (метод максимального зазора) [3].

Основным недостатком линейных решающих функций является то, что они «навязывают» данным линейную структуру, которая в реальных ситуациях практически никогда не наблюдается. В связи с этим, применение линейных моделей всегда связано с потерей дисперсии данных, связанной с их линеаризацией. И это часто становится причиной ограничения числа факторных признаков, включаемых в модель. Если структура данных является существенно нелинейной, то применение линейных решающих функций становится неэффективным.

Нелинейные решающие функции свободны от многих недостатков, свойственных линейным решающим функциям. В большинстве случаев их использование позволяет существенно повысить качество прогнозирования, объяснения или классификации. Однако построение таких функций реализуется значительно сложнее, что и является основной причиной их ограниченного применения. В настоящее время разработке эффективных методов построения нелинейных решающих функций уделяется всё большее внимание со стороны многих авторов [4].

Все известные методы построения нелинейных решающих функций можно условно разделить на нейросетевые и аналитические. Нейросетевой подход основан на процедуре итеративной оптимизации некоторой целевой функции, например на минимизации среднеквадратической ошибки. В настоящее время, благодаря простоте реализации и широким возможностям, он завоевал большую популярность. Основной недостаток нейросетевого подхода заключается в высокой продолжительности процедуры обучения и отсутствии гарантии сходимости обучающих алгоритмов. Кроме того, решающие функции, получаемые в результате обучения нейросети, обычно уступают по основным показателям качества решающим функциям, получаемым аналитическими методами. Например, классификатор на

базе самоорганизующихся сетей Кохонена, как правило, обнаруживает меньшую чувствительность и специфичность по сравнению с линейным классификатором, построенным по методу обобщенного портрета [3].

Независимо от применяемого подхода, для упрощения процедуры построения решающей функции, в её состав часто вводятся алгоритмы снижения размерности пространства описаний. Эти алгоритмы, за счёт предварительной обработки исходных данных, позволяют снять проблемы, связанные с мультиколлинеарностью, а также – улучшить репрезентативность обучающей выборки.

Для бесконечной генеральной совокупности коэффициент доверия выборочных оценок можно определить как

$$t \geq \frac{\Delta}{\sigma} \sqrt[(k/2)]{n}, \qquad (1)$$

где σ, Δ и n – среднеквадратичное отклонение, предельная ошибка и объём выборки; k – размерность данных.

Таким образом, снижение размерности пространства описаний при сохранении общей дисперсии данных повышает достоверность выборочных оценок. Из (1) следует, что достоверность выборочных оценок повышается так же и в случае частичной потери дисперсии данных, но лишь до тех пор, пока степень сжатия не достигнет определённого предела. Оптимальная степень сжатия данных, обеспечивающая максимальную достоверность выборочных оценок, зависит как от самих данных, так и от применяемого алгоритма снижения размерности.

Наиболее распространённым методом снижения размерности данных, позволяющим максимально сохранить их дисперсию, является метод главных компонент [5]. Это линейный метод, основанный на сингулярном разложении ковариационной матрицы данных. При нормальном законе совместного распределения данных метод главных компонент является оптимальным. Он является весьма эффективным и для данных, не удовлетворяющих этому требованию, особенно при их высокой размерности. Однако в этом случае всё же приходится мириться с некоторой потерей дисперсии, связанной с линеаризацией данных.

Нелинейные методы снижения размерности данных, как правило, позволяют более качественное пространство признаков, хотя и реализуются значительно сложнее. Среди них достаточно широкое распространение получил нейросетевой метод «узкого горла». Сеть с «узким горлом» состоит из 4 слоёв. Первый слой образован нейронами сигмовидного типа, число которых превышает число входов сети (этот слой осуществляет процедуру расширения размерности пространства описаний), второй слой – линейный, который собственно и является «узким горлом», число нейронов в нём всегда меньше числа входов сети. На выходе второго слоя формируются сжатые данные. Третий и четвёртый слои используются только в процессе обучения сети и осуществляют процедуру восстановления исходных данных. Собственно процесс обучения заключается в итеративной перестройке коэффициентов сети, с целью минимизации исходных и восстановленных данных.

Недостатком сетей с «узким горлом» является длительность процесса обучения и отсутствие гарантии получения оптимального результата. По этой причине для снижения размерности данных находят применение также нелинейные аналитические методы сжатия, в той или иной мере свободные от перечисленных недостатков. Среди них можно отметить метод главных кривых и многообразий. Следует отметить, что название этих методов отражают цель – определение главных кривых (криволинейных координатных осей), в направлении которых данные обнаруживают максимально возможную дисперсию. Однако способы достижения

этой цели могут быть различными. Так, один из подобных методов использует кусочно–линейную аппроксимацию главной кривой [6]. Он начинается с классического метода главных компонент. На полученной главной прямой располагается несколько узлов, положения которых итеративно уточняется, в результате чего формируется оптимальная кусочно-линейная аппроксимация данных. Строгого математического доказательства сходимости этого алгоритма не существует, однако на практике он сходится обычно довольно быстро.

В настоящей работе рассматривается ещё один нелинейный метод снижения размерности пространства описаний – метод распрямления, отличающийся гарантированной сходимостью и возможностью осуществления контроля характеристик данных непосредственно в ходе его работы. Структура многомерной функции, реализующей снижение размерности пространства описаний, представлена на рис. 1. Как следует из рисунка, структура этой функции аналогична структуре многослойного персептрона, однако способ определения свободных параметров этой функции существенно отличается от нейросетевого.

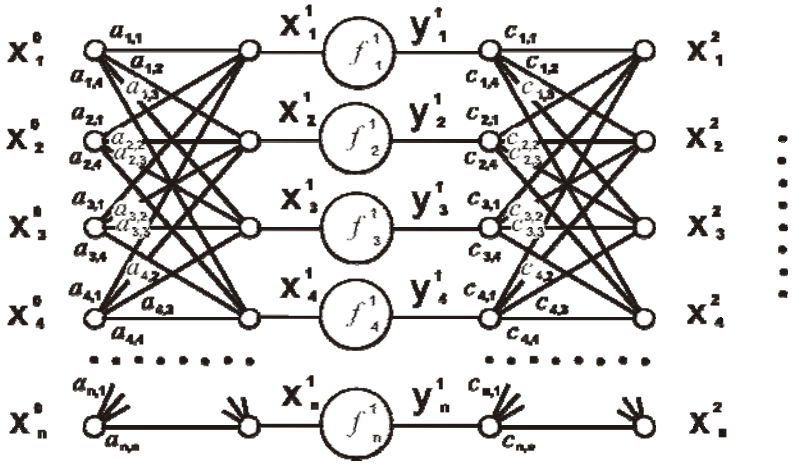

Рис. 1. Структура многомерной функции, реализующей снижение размерности пространства описаний по методу распрямлений

В основе предлагаемого метода лежит всё тот же линейный метод главных компонент, согласно которому на первом этапе производится линейное преобразование исходного пространства признаков $X^0=\{x^0_1, x^0_2, \ldots x^0_n\}$ в новое признаковое пространство $X^1=\{x^1_1, x^1_2, \ldots x^1_n\}$, в котором взаимная корреляция признаков полностью отсутствует (здесь верхние индексы обозначают номер признакового пространства, который не следует путать с показателем степени).

Матрица линейного преобразования, образованная собственными векторами ковариационной матрицы исходного пространства признаков, составляется таким образом, что новые признаки $x^1_1, x^1_2, \ldots x^1_n$ располагаются в порядке убывания величины их дисперсии. При этом признак x^1_1 обладает наибольшей дисперсией среди всех остальных признаков нового пространства описаний.

На втором этапе отдельно для каждого нового признака производится определённое нелинейное преобразование и формируется новое признаковое пространство признаков Y^1. В случае наличия статистической взаимосвязи между признаками пространства X^1, в результате нелинейного преобразования, в пространстве Y^1 возникают корреляции. Таким образом, часть чисто нелинейной взаимосвязи

признаков снова превращается в линейную взаимосвязь, т.е. осуществляется распрямление. Повторное применение метода главных компонент позволяет полностью её устранить, но при этом, в зависимости от формы применяемых нелинейных преобразований, реализуемых функциями $\{f^1_1, f^1_2, \dots f^1_n\}$, происходит перераспределение дисперсии данных между отдельными признаками.

Для того чтобы максимально сконцентрировать дисперсию данных в одном из признаков, признак x^1_1 не подвергается никаким преобразованиям, что эквивалентно $f^1_1=1$. Остальные функции $f^1_2, \dots f^1_n$ выбираются такими, чтобы обеспечить максимальную корреляцию признаков $y^1_2, \dots y^1_n$ с признаком y^1_1. Причём, оптимизация формы этих функций производится отдельно для каждого признака. В результате повторного применения метода главных компонент формируется новое пространство признаков $X^2=\{x^2_1, x^2_2, \dots x^2_n\}$, в котором признак x^2_1 обладает максимальной дисперсией, величина которой не может быть меньше дисперсии признака x^1_1. Таким образом, поочерёдное применение метода главных компонент и процедуры нелинейного преобразования может приводить только к повышению дисперсии главного признака. Одновременно с этим происходит снижение его статистической взаимосвязи с остальными признаками. В пределе главный признак становится статистически независимым, а его дисперсия достигает максимально возможной величины.

После выделения главного независимого признака процедура может быть продолжена. В результате последовательного применения метода главных компонент и нелинейных преобразований к оставшимся признакам выделяется второй по старшинству статистически независимый признак, обладающий максимальной из всех оставшихся признаков дисперсией. Затем аналогично выделяется третий по старшинству признак и т.д. После полного завершения процедуры формируется пространство статистически независимых признаков, часть новых признаков, обладающих пренебрежимо малой дисперсией, отбрасывается, за счёт чего и достигается сокращение размерности. Применение этого метода позволяет осуществить оптимальное сжатие независимо от закона распределения входных данных.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Норман Дреппер, Гарри Смитт.* Прикладной регрессионный анализ. Множественная регрессия = Applied Regression Analysis. – 3-е изд. – М.: Диалектика, 2007.
2. *Клекка У.Р.* Дискриминантный анализ // Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989.
3. *Вапник В., Глазкова Т., Кощеев В., Михальский А., Червоненкис А.* Алгоритмы и программы восстановления зависимостей. – М.: Наука, 1984.
4. Материалы XI Всероссийского семинара «Нейроинформатика и её приложения». – Красноярск, 2003.
5. *Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д.* Прикладная статистика. Классификация и снижение размерности. – М.: Финансы и статистика, 1989.
6. *Caleveland VJS. ind McOill R.* Graphical Perception Theory, Experementation and Application to the Devolopment of Graphical Metods // Journal of the American Statistical Association, September 1984, Vol. 79, № 387, Application Session.

Киреев Андрей Владимирович
Пензенская государственная технологическая академия.
E-mail: kireewska@mail.ru.
440605, г. Пенза, пр. Байдукова/ул. Гагарина, д.1а/11, тел.: (8412)496155.
Инженер, к.т.н.

Kireev Andrey Vladimirovich
Penza state technological academy.
E-mail: kireewska@mail.ru.
1a/11, Bajdukova /Gagarina, Penza, 440605, Russia, Phone: (8412)496155.
The engineer, Cand. Eng. Sci.

УДК 615.47

А.Б. Красковский, В.В. Руденко, О.В. Шаталова

КОМПЛЕКСНАЯ МЕТОДИКА ОЦЕНКИ РИСКА ВОЗНИКНОВЕНИЯ СЕРДЕЧНО-СОСУДИСТЫХ ЗАБОЛЕВАНИЙ

В работе предлагается интеллектуальная система для определения риска сердечно-сосудистых заболеваний, основанная на моделях агрегации частных рисков, полученных в результате тестирования по блокам факторов риска, полученным по психологическим и физиологическим параметрам индивидуума.

Риск сердечно-сосудистых заболеваний; тестовые опросники; регрессионный анализ.

A.B. Kraskovskiy, V.V. Rudenko, O.V. Shatalova

COMPLEX TECHNIQUE OF THE ESTIMATION OF RISK OF OCCURRENCE OF CARDIOVASCULAR DISEASES

In work the intellectual system for definition of risk of the cardiovascular diseases, based on models of aggregation of the private risks received as a result of testing on blocks of risk factors is offered, to the received on psychological and physiological parametres of an individual.

Risk of cardiovascular diseases; test questionnaires; regression the analysis.

В последние десятилетия произошло резкое изменение образа жизни: изменился и продолжает меняться характер быта, труда, питания, уменьшается физическая активность, все чаще встречаются лица с избыточной массой тела; постоянно меняется среда обитания.

Эти факторы не могут положительно влиять на здоровье человека. Они приводят к резкому увеличению психических, психосоматических заболеваний, уменьшению продолжительности жизни, ранней смертности.

Появилась необходимость изучения и своевременного определения факторов риска заболеваний, таких как заболевания сердечно-сосудистой системы.

На данный момент сформировалась концепция факторов риска, основанная на данных эпидемиологических исследований о наличии тесной связи между определенными факторами внутренней и внешней среды и развитием ишемической болезни сердца. Концепция факторов риска является основой для разработки мероприятий по первичной профилактике сердечно-сосудистых заболеваний (ССЗ).

При оценке роли факторов риска в отношении заболеваний надо исходить из данных экспериментальных и клинических исследований и из индивидуального проявления заболевания у каждого отдельного человека. Фактор риска и болезнь не всегда имеют в своей основе причинную связь. Типичным для факторов риска является их групповой характер, а не влияние на отдельного человека.

Все многообразие факторов риска, влияющих на здоровье, можно разделить на две основные группы факторов: внутренние – эндогенные (генетически обусловленные) и факторы внешние – экзогенные (природные и социальные). Это деление на внутренние и внешние факторы является чисто условным.