Заключение

Разработан метод, который применим для заполнения пробелов и ремонта данных с пробелами. Представлены три различные вариации метода, начиная с простейших линейных моделей и заканчивая методом главных кривых для данных с пробелами. Нейросетевая реализация метода позволяет легко строить его параллельные реализации.

Приведенный алгоритм заполнения пробелов не требует их предварительного априорного заполнения – в отличие от многих других алгоритмов, предназначенных для той же цели. Однако он требует предварительной нормировки ("обезразмеривания") данных – перехода в каждом столбце таблицы к "естественной" единице измерения. Важное замечание – в задаче обработки данных с пробелами невозможно перейти к однородной задаче центрированием данных.

Большой интерес вызывает вопрос: сколько слагаемых (главных кривых) следует брать для обработки данных? Существует несколько вариантов ответов, но большинство из них подчиняется эвристической формуле: *число слагаемых должно быть минимальным среди тех, что обеспечивают удовлетворительное (терпимое) тестирование метода на известных данных.* Такой принцип "минимальной достаточности" характерен для многих нейросетевых приложений [32, 33, 47].

Разработанный метод выступает в форме некоторого "анзатца" – предложения, а не в виде серии теорем. Это не случайно – предлагается технология построения *правдоподобных* оценок пропущенных данных, а не их неизвестного истинного значения. Практическая ценность методов такого рода должна оцениваться и взвешиваться пользователями технологии. Разработано соответствующее программное обеспечение. Оно хорошо зарекомендовало себя при решении трудных задач с большим числом пропущенных данных, а в более простых (стандартных) случаях приводит к тем же результатам, что и классические методы статистического анализа.

В итоге основными результатами работы можно считать следующие.

1. Для решения задачи заполнения пропусков и ремонта искаженных данных разработан метод итерационного моделирования неполных данных с помощью многообразий малой размерности. Приведены три вариации метода: начиная с простейших линейных многообразий, продолжая построенными на их основе квазилинейными многообразиями и заканчивая методом главных кривых для данных с пробелами.

2. Для параллельной реализации метода итерационного моделирования данных с пробелами разработан способ построения нейронного конвейера, решающего задачи заполнения пробелов и ремонта данных.

3. Разработаны программные продукты FAMaster и ModelAnalyzer, реализующие предложенные технологии.

4. Численные эксперименты показали высокую эффективность итерационного моделирования неполных данных с помощью многообразий малой размерности. Метод хорошо зарекомендовал себя при решении трудных

задач с большим числом пропущенных данных, а в более простых (стандартных) случаях приводит к тем же результатам, что и классические методы статистического анализа.

Литература

1. Россиев А.А. Моделирование данных при помощи кривых для восстановления пробелов в таблицах // Методы нейроинформатики: сборник научных трудов / Под ред. А.Н. Горбаня. – Красноярск: КГТУ, 1998, – С. 6–22.
2. Горбань А.Н., Макаров С.В., Россиев А.А. Нейронный конвейер для восстановления пробелов в таблицах // Нейронные сети и модели: Труды международной НТК «Нейронные, реляторные и непрерывнологические сети и модели» (19-21 мая 1998 г.), Т.1 / Под ред. Л.И. Волгина. – Ульяновск: УлГТУ, 1998. – с.3.
3. Россиев А.А. Моделирование данных для восстановления пробелов в таблицах // Материалы конференции молодых ученых Института вычислительного моделирования СО РАН, апрель 1998 г. – Красноярск: ИВМ СО РАН, 1998, – с. 46–61.
4. Горбань А.Н., Макаров С.В., Россиев А.А. Нейронный конвейер для восстановления пробелов в таблицах и построения регрессии по малым выборкам с неполными данными // Математика. Компьютер. Образование. Вып. 5. Часть II. Сборник научных трудов / Под ред. Г.Ю. Ризниченко. М.: Изд-во Прогресс-Традиция, 1998. С. 27–32.
5. Горбань А.Н., Макаров С.В., Россиев А.А. Итерационный метод главных компонент для таблиц с пробелами // Третий сибирский конгресс по прикладной и индустриальной математике (ИНПРИМ-98), 22-27 июня 1998. Тезисы докладов. Ч.5. Новосибирск: Изд-во Института математики СО РАН, 1998. – с.74.
6. Горбань А.Н., Макаров С.В., Россиев А.А. Применение линейного и нелинейного факторного анализа, мозаичной регрессии и формул Карлемана для предобработки данных с пробелами // Нейроинформатика и ее приложения: Тезисы докладов VI Всероссийского семинара, 2-5 октября 1998 г. – Красноярск: КГТУ, 1998, – с.197–198.
7. Россиев А.А. FAMaster – программный продукт для моделирования неполных данных и заполнения пробелов в них // Нейроинформатика и ее приложения: Тезисы докладов VI Всероссийского семинара, 2-5 октября 1998 г. – Красноярск: КГТУ, 1998, – с.155.
8. Gorban' A.N., Rossiev A.A. Neural Network Iterative Method of Principal Curves for Data with Gaps. Journal of Computer and System Sciences International, 1999, Vol. 38, No. 5, P. 825–850.
9. Горбань А.Н., Макаров С.В., Россиев А.А. Заполнение пробелов в данных при помощи линейного и нелинейного факторного анализа, мозаичной регрессии и формул Карлемана // Всеросс. научно-техн. конф. Нейроинформатика-99. Сборник научных трудов. В 3 частях. Ч.1. – М.: МИФИ. 1999. 276 с. – С. 25–31.
10. Россиев А.А. Нейросетевая реализация метода главных кривых для данных с пробелами. // "Студент и научно-технический прогресс": Информационные

технологии. Материалы XXXVII международной научной студенческой конференции. –Новосибирск: НГУ, – 1999. С. 90–91.

11. Горбань А.Н., Россиев А.А. Итерационный метод главных кривых для данных с пробелами // Проблемы нейрокибернетики. Материалы XII Международной конференции по нейрокибернетике. – Ростов-на-Дону: Изд-во СКНЦ ВШ. 1999. 323 с. С. 198–201.
12. Россиев А.А. Нейросетевой подход к итерационному методу главных кривых для данных с пробелами // Материалы конференции молодых ученых Института вычислительного моделирования СО РАН, март 1999 г. – Красноярск: ИВМ СО РАН, 1999. – с. 92–94.
13. Горбань А.Н., Россиев А.А. Самоорганизующиеся кривые и нейросетевой итерационный метод главных кривых для данных с пробелами // Нейроинформатика и ее приложения: Тезисы докладов VII Всероссийского семинара, 1999 / Под ред. А.Н. Горбаня. Красноярск. КГТУ. 1999. – 167 с. – С. 32–33.
14. Горбань А.Н., Россиев А.А., D.C. Wunsch II, Самоорганизующиеся кривые и нейросетевое моделирование данных с пробелами // 2-я Всероссийская научно-техническая конференция “Нейроинформатика-2000”. Сборник научных трудов. Ч.1. М.: МИФИ.– 2000. С.40–46.
15. Зиновьев А.Ю., Питенко А.А., Россиев А.А. Проектирование многомерных данных на двумерную сетку. // 2-я Всероссийская научно-техническая конференция “Нейроинформатика-2000”. Сборник научных трудов. Ч.1. М.: МИФИ.– 2000. С.80-88.
16. Дергачев В.А., Макаренко Н.Г., Куандыков Е.Б., Горбань А.Н., Россиев А.А., Восстановление пробелов методами нейроинформатики, International Conference on Problems of Geocosmos, St. Peterburg, 2000, Book of Abstracts.
17. Gorban A.N., Rossiev A.A. Wunch II D.C. Neural Network Modelling of Data with Gaps // Радіоелектроніка. Інформатика. Управління, Запорожье. 2000, № 1, С. 47–55
18. Головенкин С.Е, Матюшин Г.В., Россиев А.А. Выявление факторов, влияющих на течение и прогноз заболевания у больных с сочетанными поражениями проводящей системы сердца // Нейроинформатика и ее приложения. Тезисы докладов VIII Всероссийского семинара. – Красноярск: КГТУ. – 2000. – С.44.
19. Горбань А.Н, Россиев А.А. Нейросетевое итерационное моделирование данных с пробелами самоорганизующимися многообразиями малой размерности // Нейроинформатика и ее приложения. Тезисы докладов VIII Всероссийского семинара. – Красноярск: КГТУ. – 2000. – С.45–48.
20. Артюхов И.П., Виноградов К.А., Россиев А.А., Россиев Д.А. Кластерный анализ регионов Красноярского края по показателям здоровья и здравоохранения // Моделирование неравновесных систем – 2000: Материалы III Всероссийского семинара. – Красноярск: КГТУ. – 2000. – С. 208–210.

21. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы моделирования и первичная обработка данных. М.: Финансы и статистика, 1983. – 471с.
22. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. М.: Финансы и статистика, 1985. – 488с.
23. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989. – 607с.
24. Айзенберг Л.А. Формулы Карлемана в комплексном анализе. Первые приложения. Новосибирск: Наука, 1990. 248 с.
25. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ. – М.: Мир, 1982. – 488с.
26. Вапник В.Н. Восстановление зависимостей по эмпирическом данным. – М.: Наука, 1979. – 448с.
27. Горбань А.Н., Миркес Е.М., Свитин А.П. Метод мультиплетных покрытий и его использование для предсказания свойств атомов и молекул // Журнал физической химии. – 1992. – Т.66, №6. – с.1503-1510.
28. Горбань А.Н., Миркес Е.М., Свитин А.П. Полуэмпирический метод классификации атомов и интерполяции их свойств. Препринт ВЦ СО АН СССР №19, Красноярск, 1989, 29с.
29. Горбань А.Н., Миркес Е.М., Свитин А.П. Полуэмпирический метод классификации атомов и интерполяции их свойств // Математическое моделирование в биологии и химии. Новые подходы. – Новосибирск: Наука. Сиб. отделение, 1992. – с.204-220.
30. Горбань А.Н., Новоходько А.Ю. Нейронные сети в задаче транспонированной регрессии, Второй Сибирский Конгресс по Прикладной и Индустриальной Математике, Тезисы докладов. Новосибирск, 1996. С.160-161.
31. Горбань А.Н., Новоходько А.Ю., Царегородцев В.Г. Нейросетевая реализация транспонированной задачи линейной регрессии, Нейроинформатика и ее приложения: Тезисы докладов IV Всероссийского семинара. Красноярск, 1996, с.37–39.
32. Горбань А.Н., Россиев Д.А. Нейронные сети на персональном компьютере. Новосибирск: Наука, 1996.
33. Ежов А.А., Шумский С.А. Нейрокомпьютинг и его приложения в экономике и бизнесе. М.: МИФИ, 1998. – 224 с.
34. Енюков И.С. Методы, алгоритмы, программы многомерного статистического анализа. – М.: Финансы и статистика, 1986.
35. Жамбю М. Иерархический кластер-анализ и соответствия: Пер. с фр. – М.: Финансы и статистика, 1988. – 342 с., ил.
36. Жанатаусов С.У. Методы прогностических переменных. – Машинные методы обнаружения закономерностей, Новосибирск: 1981, вып. 88, Вычислительные системы. С. 151-155.
37. Загоруйко Н.Г. Методы обнаружения закономерностей

38. Загоруйко Н.Г., Ёлкина В.Н., Лбов Г.С. Алгоритмы обнаружения эмпирических закономерностей. – Новосибирск: Наука, 1985. – 110с.
39. Загоруйко Н.Г., Елкина В.Н., Лбов Г.С., Емельянов С.В. Пакет прикладных программ ОТЭКС. М.: Финансы и статистика, 1986.
40. Загоруйко Н.Г., Ёлкина В.Н., Тимеркаев В.С. Алгоритм заполнения пропусков в эмпирических таблицах (алгоритм “ZET”) // Вычислительные системы. – Новосибирск, 1975. – Вып. 61. Эмпирическое предсказание и распознавание образов. – С. 3-27.
41. Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. – М.: Наука, 1976. – 736 с.
42. Кендалл М., Стьюарт А. Статистические выводы и связи. – М.: Наука, 1973. – 900 с.
43. Лбов Г.С. Методы обработки разнотипных экспериментальных данных. – Новосибирск: Наука, 1981. – 157с.
44. Литл Р.Дж.А., Рубин Д.Б. Статистический анализ данных с пропусками. М.: Финансы и Статистика, 1991.
45. Макаренко Н.Г. Многообразия, погружения и трансверсальность//сб. Проблемы солнечной активности, Ленинград, 1991, 13-28
46. Матюшин Г.В. Сочетанные поражения проводящей системы сердца (распространенность, клинико-электрокардиографические варианты, клиническое течение, прогноз) // Диссертация на соискание ученой степени доктора медицинских наук, КрасГМА, 2000.
47. Нейроинформатика / А.Н. Горбань, В.Л. Дунин-Барковский, Е.М. Миркес и др. Новосибирск: Наука (Сиб. отделение), 1998.
48. Пфанцагль И. Теория измерений. – М.: Мир, 1976. – 246с.
49. Рао С.Р. Линейные статистические методы. – М.: Наука, 1968. – 548 с.
50. Растригин Л.А., Пономарев Ю.П. Экстраполяционные методы проектирования и управления. – М.: Машиностроение, 1986. – 120 с.
51. Самарский А.А. Введение в численные методы. М.: Наука. Главная редакция физико-математической литературы, 1982. – 272 с.
52. Справочник по прикладной статистике. В 2-х т., под ред Э.Ллойда, У.Ледермана, Ю.Н.Тюрина – М.: Финансы и статистика, 1989, 1990.
53. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере / Под ред. В.Э.Фигурнова – М.: ИНФРА-М, 1998. – 528 с., ил.
54. Факторный, дискриминантный и кластерный анализ. – М.: Финансы и статистика, 1989. – 215 с.
55. Afifi A.A., Elashoff R.M. Missing observations in multivariate statistics. – J. Amer. Statist. Assoc., 1966, vol. 61. pp. 595-604.
56. Beale E.M., Little R.J. Missing values in multivariate analysis. – J. Roy. Statist. Soc. B., 1975, vol. 37. pp. 129-145.
57. Buck S.F. A method of estimation of missing values in multivariate data. – J. Roy. Statist. Soc. B., 1960, vol. 22. pp. 202-206.
58. Delicado P., Principal Curves and Principal Oriented Points, Tech. rep. 309, Department d'Economia i Empresa, Universitat Pompeu Fabra, 1998.

59. Dempster A.P., Laird N.M., Rubin D.B. Maximum likelihood from incomplete data via the EM-algorithm. – J. Roy. Statist. Soc. B., 1977, vol. 39. pp. 1-38.
60. Dodge Y. Analysis of experiments with missing data. – New York, Wiley, 1985. 498 p.
61. Engelman L. An efficient algorithm for computing covariance matrices from data with missing values. – Communs Statist. B., 1982, vol. 11. p. 113-121.
62. Frane G.M. Some simple procedures for handling missing values in multivariate analysis. – Psychometrika, 1976, vol. 41. pp. 409-415.
63. Glasser M. Linear regression analysis with missing observations among the independent variables. – J. Amer. Statist. Assoc., 1964, vol. 59. p. 834-844.
64. Gleason T.C., Staelin R. A proposal for handling missing data. – Psychometrika, 1975, vol. 40. pp. 229-252.
65. Gorban A.N., Novokhodko A.Yu. Neural Networks In Transposed Regression Problem, Proc. INNS WCNN ’96.
66. Gorban A.N., Waxman C. Neural Networks for Political Forecast. Proceedings of the WCNN'95 (World Congress on NeuralNetworks'95, Washington DC, July 1995), PP.176- 178.
67. Hartley H.O., Hocking R.R. The analysis of incomplete data. – Biometrics, 1971, vol. 27. pp. 783-808.
68. Hastie T. and Stueltze, Principal Curves, Journal of the American Statistical Association, vol. 84, no. 406, pp. 502-516, 1989.
69. Hastie T. Principal Curves and Surfaces, PhD thesis, Stanford University, 1984.
70. Hocking R.R., Marx D.L. Estimatiom with incomplete data: an improved computational method and the analysis of nested data. – Communs Statist. A., 1979, vol. 8. pp. 1151-1181.
71. Huseby J.R., Schwertman N.C., Allen D.M. Computation of the mean vector and dispersion matrix for incomplete multivariate data. – Communs Statist. B., 1980, vol. 9. pp. 301-309.
72. Kegl B., Krzyzak A., Linder T. and Zeger K. Principal Curves: Learning and convergence, in Proceedings of IEEE International Symposium on Information Theory, p. 387, 1998.
73. Kegl B., Krzyzak A., Linder T., and Zeger K., Learning and Design of Principal Curves, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999.
74. Kohonen T. Self-Organizing Maps. Springer: Berlin – Heidelberg, 1997.
75. Kramer M.A. Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal. 1991. V.37, No. 2. PP. 233-243.
76. LeBlank M., Tibshorany N. Adaptive principal surfaces. Journal of the American Statistical Association. 1994, Mar. V. 89, No. 425. PP. 53-64.
77. Lichtman A.J., Keilis-Borok V.I., Pattern Recognition as Applied to Presidential Elections in U.S.A., 1860-1980; Role of Integral Social, Economic and Political Traits, Contribution N 3760. 1981, Division of Geological and Planetary Sciences, California Institute of Technology.
78. Little R.J., Rubin D.B. Statistical analysis with missing data. – New York, Wiley, 1987. 430 p.

79. Little R.J., Schlushter M.D. Maximum likelihood estimation for mixed continuous and categorical data with missing values. – Biometrika, 1985, vol. 72. pp. 497-512.
80. Little R.J., Smith P.J. Editing and imputation for quantitative survey data. – J. Amer Statist. Assoc., 1987, vol. 82. pp. 58-68.
81. Mardia K.V., Kent J.T. and Bibby J.M. Multivariate Analysis. London: Academic Press, 1979.
82. Srivastava M.S. Multivariate data with missing observations. – Communs Statist. Theory and Method, 1985, vol. 14. pp. 775-792.
83. Tibshirani R., Principal Curves revisited, Statistics and Computation, vol. 2, pp. 183-190, 1992.
84. Titterington D.M., Jiang J.M. Recursive estimation procedures for missing data problems. – Biometrika, 1983, vol. 70. pp. 613-624.
85. Walsh J.E. Computer-feasible method for handling incomplete data in regression analysis. – J. of ACM, 1961, vol. 18. pp. 201-211.
86. Wilks S.S. Moments and distributions of estimates of population from fragmentary samples. – Ann. Math. Statist., 1932, vol.3. pp. 163-195.