Множественное восстановление пропущенных данных с помощью глубинных нейробайесовских моделей

Царькова Н.И., кандидат педагогических наук, доцент кафедры «Прикладная информатика», ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Суворов С.В., кандидат экономических наук, доцент кафедры «Прикладная информатика», ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Жиляева И.А., кандидат экономических наук, доцент кафедры «Прикладная информатика», ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Шебанова К.В., магистрант кафедры «Прикладная информатика», ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Аннотация. В данном исследовании рассматриваются фундаментальные концепции, лежащие в основе байесовской статистики, описывается применение данного метода для восстановления пропущенных данных с помощью глубинных нейробайесовских моделей. В качестве модели используется восстановление пропущенных данных при передаче по защищенным каналам связи. Для выбранных наборов результаты исследования показывают удовлетворительный уровень восстановления пропущенных данных.

Ключевые слова: нейробайесовская модель, технология BigData, код Хемминга, нейронная сеть, генетический алгоритм, защищенный канал связи, пропущенные данные.

Multiple Missed Data Recovery Using Deep Neuro Bayesian Models

Tsarkova N.I., candidate of pdagogical sciences, Associate Professor of the

«Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia
Suvorov S.V., candidate of economic sciences, Associate Professor of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia
Zhilyaeva I.A., candidate of economic sciences, Associate Professor of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia
Shebanova K.V., magistrate of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia

Annotation. This study discusses the fundamental concepts underlying Bayesian statistics, describes the application of this method to restore missing data using deep neuro-Bayesian models. As a model, recovery of missing data is used during transmission over secure communication channels. For the selected sets, the results of the study show a satisfactory level of recovery of missing data.

Keywords: Neuro-Bayesian model, Big Data technology, Hamming code, neural network, genetic algorithm, secure communication channel, missing data.

Статистические исследования обеспечивают теоретическую основу для согласованного и тщательного анализа данных, предоставляя математические основы для унификации моделей того, как информация (в том числе «большие данные») используются для описания различных систем. Статистика также необходима для обработки результатов экспериментов, в которых присутствуют неопределенности различного характера, связанные с изучаемыми процессами. Хотя единого универсального статистического подхода не существует, одним из наиболее распространенных в последние 30 лет, является байесовский статистический вывод. Байесовская статистика имеет определенные возможности, которые позволяют использовать её при изучении различных сложных статистических приложений, особенно в машинном обучении, где другие статистические методы могут столкнуться с трудностями.

Как следствие, байесовские подходы в настоящее время широко используют в различных научных и технологических приложениях, включая IT-сферу.

В практической деятельности используют несколько подходов к работе с большими массивами данных, которые содержат пропущенные значения. Первый подход, который является наиболее простым в реализации, состоит в проведении процесса удаления всех записей, содержащих значения с пропусками, из всего рассматриваемого массива и продолжение работы только с полными данными. Использование такого подхода оказывается целесообразным, если пропущенные данных оказываются единичными. Но в таком случае возникает серьезная опасность «исчезновения» важных закономерности при проведении удаления данных. В тех же случаях, когда число пропусков достаточно велико, удаление соответствующего количества записей может привести к недостатку данных или невозможности дальнейшего вычисления требуемых статистических параметров. Второй подход состоит в использовании специальных методов модификации данных, которые допускают наличие пропусков в большом массиве. Третий подход, наиболее распространенный, использует методы оценки значения пропущенного элемента. Данные методики дают возможность заполнения пропусков в массиве, основываясь на определенных предположениях о величине значения отсутствующей информации.

В данном исследовании будет использоваться модификация третьего подхода для восстановление пропущенных данных при передаче по защищенным каналам связи с помощью глубинной нейробайесовской модели.

Построение модели защищенного канала

Для передачи данных по моделируемому защищенному каналу используется модифицированный код Хэмминга. Параметры кодов Хемминга: $n = 2^m - 1,\ \ m > 1; k = n - m; d = 3;$ проверочная матрица $H = (1, \alpha, \alpha^2, ..., \alpha^{n-1}),$ где α – примитивный элемент поля Галуа $GF(2^m).$

Столбцы проверочной матрицы являются элементов поля $GF\left(2^m\right)$, то есть векторами из P_n в базисе $1,\alpha,\alpha^2,\text{К},\alpha^n$ для примитивного элемента α поля $GF\left(2^m\right)$. Если в качестве α взять корень неприводимого полинома x^3+x+1, то матрица:

$$\tilde{H}=\left[1,\alpha,\alpha^2,...,\alpha^{2^m-2}\right]=\begin{bmatrix}0&0&1&0&1&1&1\\0&1&0&1&1&1&0\\1&0&0&1&0&1&1\end{bmatrix}$$

Эту матрицу называют матрицей линейного кода.

Расстояние Хэмминга между двумя векторами $\bar{x},\bar{y}\in P_n$ — это число $dist(\bar{x},\bar{y})$ несовпадающих координат данных векторов. Весом $wt(\bar{x})$ векторов из пространства $\bar{x}\in P_n$ называют число ненулевых координат этих векторов.

Расстояние Хэмминга имеет все свойства обычного линейного расстояния:

1) $dist(\bar{y},\bar{x})=dist(\bar{x},\bar{y})-$ симметричность;

2) $dist(\bar{x},\bar{y})=0$ выполняется тогда и только тогда, когда имеет место $\bar{x}=\bar{y}$;

3) $dist(\bar{x},\bar{z})+dist(\bar{z},\bar{y})\geq dist(\bar{x},\bar{y})-$ треугольниканеравенство.

Минимальное или кодовое расстоянием кода C — это наименьшее из всех расстояний между попарно разными векторами, принадлежащими коду C. Величину значения кодового расстояния определяют помехоустойчивой функцией кодирования.

В том случае, когда минимальное расстояние кода C составляет $d=2t+1$ или $d=2t+2$, то использование кода C дает возможность обнаружения $d-1$ ошибок и исправления t ошибок для каждого принятого вектора-слова, имеющего длину n.

Пусть H является проверочной матрицей двоичного кода C. Минимальное расстояние для такого кодабудет составлять d тогда и только в

том случае, когда любые из $d-1$ столбцов в матрице H являются линейно независимыми, ибудут присутствовать d линейно зависимых столбцов.

Кодом Хемминга называют линейный код C_χ, который имеет проверочную матрицу вида $H_\chi = \left(1, \alpha, \text{K}, \alpha^{2m-2}\right)$. В этом случае α^i – является двоичным вектор-столбцом над полем $GF(2)$ при использовании базиса $1, \alpha, \text{K}, \alpha^{m-1}$ для примитивных элементов α поля $GF\left(2^m\right)$.

Из определения вытекает, что в качестве столбцов матрицы H_χ могут быть всевозможные ненулевые векторы из двоичного пространства P_n. Поэтому произвольные коды Хэмминга характеризуются параметрами $n = 2^m - 1$, $k = n - m$:

$(7, 4); (15, 11); (31, 26); (63, 57); (127, 120); (255, 247); (511, 502); (1023, 1013)$ и так далее.

Код Хэмминга обладает минимальным расстоянием $d = 3.$ Код Хэмминга способен исправлять одиночные ошибки.

Линейные коды C. имеющие длину n и проверочную матрицу вида:

$$\text{H}=\left[\begin{array}{cccc} 1 & \beta^{b} & \beta^{2b} & \beta^{(n-1)b} \\ 1 & \beta^{b+1} & \beta^{2(b+1)} & \beta^{(n-1)(b+1)} \\ \hline 1 & \beta^{b+\delta-2} & \beta^{2(b+\delta-2)} & \beta^{(n-1)(b+\delta-2)} \end{array}\right] = [\beta^{bi}, \beta^{(b+1)i}, ..., \beta^{(b+\delta-2)i}]^T$$

над полем $GF(q)$ называют кодамиБоуза-Чоудхури-Хоквингема (БЧХ-кодом), имеющиеконструктивное расстояние δ. При n = $q^{\text{m}} - 1$ БЧХ-кодыявляются примитивными, и не примитивными, в случае n< $q^{\text{m}} - 1.$

Определении подразумевает, что в данной матрице H все элементы $\beta^i = \alpha^{ci}$ заменяются на соответствующиевектор-столбцы $(b_{m-1}, b_{m-2}, ..., b_0)^T$ поэтому код будет определен над полем $GF(q)$, а матрица H будет обладать конструктивными размерами $m(\delta - 1) \times n.$ Неравенство $m(\delta - 1) < n$ должно гарантировать, что ядро этой матрицы не будет тривиальным и, следовательно,

код C существует, представляя собой линейное пространство размерности, не меньшей, чем $n - m(\delta - 1)$.

Для каждого целого числа n, которое не делится на q, над полем $GF(q)$ существует БЧХ-код с длиной n. Для всякого нечетного $n \geq 3$ существует двоичный БЧХ-код длиной n.

Ранг проверочной матрицы H БЧХ-кода C чаще всего совпадает с числом ее строк $m(\delta - 1)$.

Пусть для некоторого целого t, не делящегося на $q^m - 1$, проверочная матрица H БЧХ-кода C содержит, с точностью до перестановки строк, подматрицу $[\beta^{it}, \beta^{itq}]$. Тогда $rank[\beta^{it}, \beta^{itq}] = rank[\beta^{it}]$.

Это остается справедливым, если в подматрице $[\beta^{it}, \beta^{itq}]$ степень itq заменить на $f(s) = itq^s$ для целых s, $1 \leq s \leq m - 1$.

Минимальным или кодовым расстоянием кода C называется наименьшее из расстояний между попарно различными векторами кода C.

Значение кодового расстояния определяет следующая – фундаментальная в помехоустойчивом кодировании

Если минимальное расстояние кода C равно $d = 2t + 1$ или $d = 2t + 2$, то код C может обнаружить до $d - 1$ ошибок и исправить до t ошибок в каждом принятом векторе-слове длиной n.

Классический примитивный БЧХ-код Сс проверочной матрицей $H = (\alpha^i, \alpha^{3i})^T$, $0 \leq i \leq n - 1$, имеет кодовое расстояние равное 5. Следовательно, этот код корректирует одиночные и двойные ошибки.

Алгоритм восстановления пропущенных данных с помощью нейробайесовской модели

Предлагаемый алгоритмкод обладает следующими свойствами:

1) Все одиночные битовые ошибки могут быть исправлены;

2) Все двойные битовые ошибки могут быть обнаружены;

3) Все соседние битовые двойные ошибки могут быть исправлены;

4) Вероятность неверного исправления для несмежных двойных ошибокснижена.

Характеристики линейного блочного кода полностью определяется его H-матрицей. Для того, чтобы обнаружить все одиночные битовые ошибки, соответствующие синдромы ошибок должен быть уникальными. Следует учитывать, что синдром для однобитовой ошибки в бите с позицией p совпадает с p-й столбцом H-матрицы. Для того, чтобы однозначно идентифицировать все одиночные битовые ошибки, все столбцы H-матрицы должны быть уникальными.

Для того, чтобы обнаружить все двойные ошибки в битах, соответствующий синдромы должны отличаться от всех синдромоводнобитовых ошибок.

Синдром для двойной битовой ошибки определяется операцией исключающее ИЛИ (XOR) соответствующих столбцов H-матрицы. Так что не может быть 3-цикла в H-матрицах. k-цикл относится к набору k линейно зависимых столбцов матрицы проверки на четность, то есть, когда проведены все операции XOR, в результате имеется полностью нулевой столбец. Для исправления всех смежных двойных битовых ошибок, синдромы близлежащих двойных битовых ошибок должны быть отличны друг от друга, а также отличается от синдромов всех однобитовых ошибок.

Определим условия, которым должны удовлетворять H-матрицы для предлагаемого кода:

1) Невсе столбцы нулевые;

2) Все столбцы являются различными;

3) Отсутствует линейная зависимость, включающая 3 или меньше столбцов т.е. отсутствуют 2-циклы, 3 циклы допускаются.

4) Отсутствует линейная зависимость столбцов с участием C_i, C_j, C_k, C_m ,где $m > k > j > i$, такие, что $j = i + 1$ и $m = k + 1$.

5) Кроме того, код пытается минимизировать количество 4-циклов с участием C_i, C_j, C_k, C_m, ,где $m > k > j > i$, такие, что $j = i + 1$ и $m = k + 1$.

Условие 1 гарантирует, что отсутствуют однобитовые ошибки в тривиальном случае.

Условие 2 гарантирует, что все синдромы единичных ошибок являются уникальными. Каждый синдром единичной ошибки соответствует одному из столбцов H-матрицы. Поскольку все столбцы H-матрицы различны, одиночные битовые ошибки однозначноидентифицируются и, следовательно, являются исправимыми. Кроме того, этоусловие гарантирует, что отсутствует пара двойных ошибоквида (i, j) и (j, k) такая, что соответствующие синдромы одинаковы. Предположим, что такие двойные ошибкисуществуют, то $(C_i \oplus C_j) \oplus (C_j \oplus C_k) = 0$, т.е. $(C_i \oplus C_k) = 0$, но это противоречит тому факту, что все столбцы H-матрицы различны. Это гарантирует, что синдромы смежных ошибок вида (i, i + 1) и (i + 1, i+ 2) различны.

Условие 3 гарантирует, что синдромы для всех двойных битовых ошибок отличаются от одиночных битовых ошибок. Синдром для двойной ошибки передачи битов определяется исключающим-ИЛИ из столбцов, соответствующих ошибочному биту позиции. Если H-матрица свободна от 3-циклов, то XOR любых двух столбцов H-матрицы не совпадает с каким-либо столбцом H-матрицы. Это гарантирует, что синдромы всех двойных битовых ошибок отличаются от отдельных синдромы одиночных ошибок, а условие 2 обеспечивает отличие от нуля синдромов двойных битовых ошибок. Следовательно, все двойные битовые ошибки могут быть обнаружены.

Условие 4 вместе с условием 2, гарантирует, что синдром для смежной двойной ошибки отличается от всех других смежных синдромов битовых двойных ошибок. Если предполагается, что имеются ошибки только битовые одиночные или смежные двойные ошибки с H-матрицей удовлетворяющей условиям 1 по 4, то можно однозначно выявить синдромы для всех одиночных битовых ошибок исмежных ошибок, следовательно, можно исправить все одиночные битовые ошибки.

Носиндромы для соседних битовых ошибок совпадают с некоторыми синдромами несмежных двойных ошибок. Это происходит потому, что

некоторые 4-циклы допускаются, чтобы уменьшать затраты ресурсов на битовую проверку. Так что имеется вероятность того, что несмежная двойная бородкой будет принята в качестве смежной двойной битовой ошибки и, следовательно, будет неправильно исправлена (хотя вероятность несмежных двойных ошибок значительно меньше, чем усмежных двойных ошибок). Условие 5 пытается минимизировать вероятность таких событий.

Создание H-матрицы является, по существу, систематическим процессом поиска, чтобы удовлетворить всем условиям, указанным ранее. Для матрицы (r × n), имеется $2^{(r \times n)}$ возможных вариантов, поэтому исчерпывающий метод поиска состоит в нахождении достаточно больших значенийr и n. На рисунке 24 показана H-матрица для такого кода. Но для этого кода, исчерпывающий поиск не является особо практичным, даже если область считать ограниченной. Вес колонки H-матрицы определяется как число единиц вколонка. Если ограничить H-матрицу только 3-весои и1-весом столбцов, тогда имеется $C_3^6 = 20$ вариантов из которых 16 столбцов могут быть выбраны $C_{16}^{20} = 4845\,4845$ способами.

Для каждого выбора есть ($16! > 2 \times 10^{13}$) перестановок столбцов, которые следует производить для поиска лучшего кода.

Таким образом, при исчерпывающем поиске лучшего кода будет использоваться ($4845 \times 16! > 2 \times 10^{13}$) матрицы. Пространство поиска возрастает еще больше, если допускаются произвольно взвешенные столбцы.

Обсуждение результатов

Первый используемый набор данных - передача 120 МБ информации при нормальных условиях эксплуатации сети.

Выборка данных проводилась каждые 0,1 секунды, и было зарегистрировано 200 пропусков. Извлечение данных без пропущенных значений показано в таблице 1.

Таблица 1

Набор данных без пропущенных значений

1.	**2.**	**3.**	**4.**	**5.**
0.11846	0.089431	0.11387	0.6261	0.076995
0.10859	0.082462	0.11284	0.6261	0.015023
0.099704	0.19919	0.14079	0.62232	0.061972
0.092794	0.19164	0.12733	0.6261	0.059155
0.0888845	0.30023	0.13768	0.6261	0.028169
0.0875858	0.63182	0.074834	0.63052	0.079812

Данные были разделены на наборы данных обучения и тестирования. Из-за ограниченности имеющихся данных, одна седьмая данных была сохранена в качестве набора тестов, а остальные представлены для обучения.

Для эксперимента с данными была реализована нейронная сеть - генетический алгоритм (NN-GA) с использованием сети кодеров, обученной 4 скрытыми узлами в течение 200 тренировочных этапов. Генетический алгоритм был реализован с использованием представления в течение 30 поколений, с 20 хромосомами на поколение. Скорость мутации была установлена на значение 0,1. Параметры алгоритма были определены опытным путем. Коэффициент корреляции и точность в пределах 10% от фактических значений приведены в таблице 2.

Таблица 2

Результаты тестирования восстановления пропущенных данных

	Корреляция		10%	
Переменная	**Corr EM**	**Corr NN-GA**	**EM**	**NN-GA**
1.	-	0.9790	-	21.43
2.	0.7116	0.8061	14.29	14.29
3.	0.7218	0.6920	7.14	28.57
4.	-0.4861	0.5093	3.57	10.71
5.	0.6384	0.8776	10.71	7.14

Результаты тестирования показывают, что алгоритм не позволил сделать прогноз для столбца 1 в этом наборе данных. Причина состоит в том, что для данного метода, чтобы сделать прогноз, матрица прогнозирования должна быть положительной. Основной причиной этого является то, что одна переменная линейно зависит от другой переменной. Эта линейная зависимость может иногда существовать не между самими переменными, а между элементами моментов, такие как среднее, дисперсии, ковариации и корреляции. Другие причины этой причины включают ошибки при чтении данных, начальные значения и многое другое. Эта проблема может быть решена путем удаления переменных, которые линейно зависят друг от друга, или путем использования главных компонентов для замены набора коллинеарных переменных на ортогональные компоненты. Для остальных наборов в других столбцах результаты показывают удовлетворительный уровень восстановления пропущенных данных

Нейробайесовские модели являются концептуально естественным подходом для применения в IT-сфере. Современные вероятностные языковые среды программирования для байесовских вычислений еще больше упростили его применение, предоставив интерфейсы для определения потенциально очень сложных моделей даже для неспециалистов. В этом исследовании описаны базовые теоретические основы, необходимые для реализации байесовского

моделирования с упором на приложения в информационной безопасности. Тем не менее, необходимы дальнейшие исследования улучшенных и более быстрых методов байесовских вычислений для больших данных. Байесовское моделирование требует значительного количества предположений о составлении порождающих моделей и уточнении исходных предположений.

Библиографический список

1. Gupta A., Lam M.S. Estimation Missing Values using Neural Networks // J. of Operational Research Society. – 1996. – Vol. 47. – № 2. – C. 229–239.

2. Nelwamondo F.V., Mohamed S., Marwala T. Missing Data: A comparison of neural network and expectation maximization techniques // Current Science. – 2007. – Vol. 93. – № 11. – C.1467–1473.

3. Карлов И.А. Методы восстановления пропущенных значений с использованием инструментария DataMining // Вестник Сибирского гос. аэрокосмического ун-та им. Академика М.Ф. Решетнева. – 2011. – № 7(40) – С.29–33.

4. Карлов И.А., Кошур В.Д. Подходы к построению гибридной модели для оценки значений пропущенных элементов в массивах данных // Нейроинформатика, ее приложения и анализ данных: Матер. XX Всеросс. семинара. – 2012. – С. 174–179.

5. Halkidi M., Batistakis Y., Vazirgiannis M. On Clustering Validation Techniques // J. of Intelligent Information Systems. – 2003. – № 17:2/3. – C.107–145.

6. An Introductions to computing with neural net, Richard P. Lipman, IEE ASP Magazines, April 2017, pages 2-22.

7. A Neural Networks Approaches Toward Intrusion Detections, Kevins L. Foxer, Rondal R. Hennings, Jonatan H. Reeds, Richard P. Sitnonians, Harris Corporations, Government Informations System Divisions, P.O. Box 98000, Melbourne, FL 32902, July 2000.

8. Univariate Economic Time Series Forecasting by Connexionist Methods, A. Varfis and C. Versino, Proceedingsof the International Neural Networks Conference, Paris, 1990, pages 342-345.

References

1. Gupta A., Lam M.S. Estimation of missing values using neural networks // J. of Operational Research Society. – 1996. – Vol. 47. – № 2. – S. 229-239.

2. Nelvamondo F.V., Mohamed S., Marvala T. Missing data: comparison of neural network methods and maximization of expectations // Modern Science. – 2007. – Issue. 93. – № 11. – P.1467-1473.

3. Karlov I.A. Methods for recovering missing values using DataMining tools // Bulletin of the Siberian State. Aerospace University named after Academician M.F. Reshetneva. – 2011. – № 7 (40). – P. 29–33.

4. Karlov I.A., Koshur V.D. Hybrid models for evaluating the values of missing elements in a data array // Neuroinformatics, its applications and data analysis: Mater. XX All-Russian seminar. – 2012. – P. 174-179.

5. Halkidi M., Batistakis Yu., Vazirgiannis M. On cluster validation methods // J. Intelligent information systems. – 2003. – № 17: 2/3. – P. 107-145.

6. Introduction to Neural Network Computing, Richard P. Lipman, IEE ASP Magazines, April 2017, 2–22.

7. Neural Network Approaches to Intrusion Detection, Kevins L. Foxer, Rondal R. Hennings, Jonathan H. Reeds, Richard P. Sitnonians, Harris Corporation, Government Information System Units, P.O. Box 98000, Melbourne, FL 32902, July 2000.

8. One-factor forecasting of economic time series using the methods of connectionism, A. Warfis and K. Versino, Materials of the International Conference of Neural Networks, Paris, 1990, pp. 342-345.