

УДК 51-74

ОЦЕНКА МАТЕМАТИЧЕСКИХ МЕТОДОВ, ИСПОЛЬЗУЕМЫХ ДЛЯ ПОДГОТОВКИ ИСХОДНЫХ ДАННЫХ

Бочкарев С.В., Обухов Е.С.*Пермский национальный исследовательский политехнический университет,
Пермь, e-mail: obuhov2014@bk.ru*

Настоящая статья посвящена исследованию эффективности методов подготовки исходных данных для задач классификации. В результате моделирования классификаторов, на основе подготовленных данных, были рассчитаны показатели точности, позволяющие оценить эффективность методов подготовки данных. Для подготовки исходных данных были использованы корреляционный, факторный анализы и анализ чувствительности. Построение моделей классификаторов и проведения анализов было осуществлено в Deductor 5.3 Trial, StatSoft Statistica 6 Trail. Корреляционный анализ рассматривался на примере данных по параметрам полетного мониторинга авиадвигателя. В факторном анализе применялся метод ортогонального вращения Varimax. В ходе анализа чувствительности вначале использовалось обучение на тестовых значениях с определением ошибки. Обучено 10 нейронных сетей типа многослойный персептрон, из которых выбрано 4 с наименьшей ошибкой. Точность классификации модели, обученной на данных после анализа чувствительности, составила 77,04%. Показано, что модели не являются достаточно точными для выполнения поставленной задачи по классификации технического состояния на примере авиационного двигателя.

Ключевые слова: корреляционный анализ, факторный анализ, анализ чувствительности, классификация, подготовка данных

THE ESTIMATION OF MATHEMATICAL METHODS USED TO PREPARATION OF DATA

Bochkarev S.V., Obukhov E.S.*Perm National Research Polytechnic University, Perm, e-mail: obuhov2014@bk.ru*

This article is dedicated to the study of preparation data methods efficiency. Several classification models were built using prepared data to estimate the efficiency of the preparation data methods efficiency. The estimation of the methods efficiency was based on calculated, before and after applying the methods of preparation data, accuracy rates and recall of classes. To prepare the data was used correlation, factor, and sensitivity analysis. To build the models of classification and making the analysis was used Deductor 5.3 Trial, StatsoftStatistica 6 Trail. Correlation analysis was discussed at the example data on parameters of flight monitoring aircraft engines. Factor analysis method applied in orthogonal rotation Varimax. During the first sensitivity analysis was used to test training values with the definition of the error. 10 trained neural networks for multilayer Perceptron type, of which 4 are selected with the smallest mistake. The accuracy of classification models, trained to data after sensitivity analysis, totaled 77,04%. It is shown that the models are not accurate enough for the task according to the classification of the technical condition of the aircraft engine example.

Keywords: factor analysis, correlation analysis, sensitivity analysis, classification, estimation of preparation data methods

При построении сетей предполагается, что входной и выходные слои заданы, то есть уже известно, что поступает на вход сети и что является ее выходом. То, какие переменные будут выходными, известно всегда (по крайней мере, в случае управляемого обучения). Однако сбалансировать входной набор данных весьма непросто. Чаще всего неизвестно заранее, какие из входных переменных действительно полезны, и выбор хорошего множества входов бывает затруднен целым рядом обстоятельств.

Каждый дополнительный входной элемент сети – это новая размерность в пространстве данных. С этой точки зрения, очевидно, что для того, чтобы иметь достаточную плотность точек и, как следствие, эффективно оценивать структуру данных, нужно иметь довольно много измерений (векторов данных).

Для осуществления контроля технического состояния авиационных двигателей, используется информация, поступающая от датчиков контролирующих параметров. Данная информация именуется как параметры полетного мониторинга, в силу того, что авиационный двигатель является конструктивно сложным элементом и производится контроль множества параметров, данные имеют большую размерность, порядка 40 атрибутов, что влечет за собой высокие требования к объему выборки данных, необходимой для обучения модели оценки технического состояния двигателя. Для того чтобы сократить размерность данных и исключить малозначимые и малоэффективные переменные, применяются следующие методы:

1. Корреляционный анализ [1] применяется для оценки зависимости выходных полей данных от входных факторов и устра-

нения незначущих. Принцип подобного анализа состоит в поиске таких значений, которые в наименьшей степени коррелированы (взаимосвязаны) с выходным результатом. Такие факторы могут быть исключены из результирующего набора данных практически без потери полезной информации. Критерием принятия решения об исключении является порог значимости. Если корреляция (степень взаимозависимости) между входным и выходным факторами меньше порога значимости, то соответствующий фактор отбрасывается как незначущий.

Рассмотрим применение метода на данных по параметрам полетного мониторинга авиадвигателя. Для выполнения данного анализа было использовано специальное программное обеспечение (СПО) Deductor 5.3. На начальном этапе имеется набор данных размерностью 40 атрибутов, один из которых является целевым атрибутом (техническое состояние двигателя). Таким образом, можно составить почти 40 пар входной атрибут и целевой атрибут. Корреляция это число между -1 и $+1$, которое измеряет степень взаимосвязи между двумя атрибутами, применительно к текущей задаче можно оценить корреляцию между целевым атрибутом и одним каждый раз разным входным атрибутом. В табл. 1 показана интерпретация значений корреляции.

Таблица 1
Интерпретация корреляционных значений

Отрицательное значение	Положительное значение	Интерпретация
$-0,2 < V < 0,2$		Очень низкое значение
$-0,5 < V < -0,2$	$0,2 < V < 0,5$	Слабая корреляция
$-0,7 < V < -0,5$	$0,5 < V < 0,7$	Средняя корреляция
$-0,9 < V < -0,7$	$0,7 < V < 0,9$	Высокое значение
$V < -0,9$	$V > 0,9$	Очень высокое значение

На последующем этапе был выбран метод расчета корреляции – коэффициент корреляции Пирсона и рассчитаны значения функции корреляции (см. рис. 1) между каждым входным и выходным (техническое состояние двигателя) столбцами. На пересечении строки с именем входного поля и столбца с именем выходного поля находится значение рассчитанной между ними корреляции.

Исключение незначущих факторов производится на основании вычисленной

корреляции путем задания необходимого уровня значимости. Столбцы, у которых максимальное из рассчитанных значений корреляции меньше порога, будут исключены из выходного набора.

Таким образом, согласно табл. 1, было установлено значение порога 0,5, тем самым были исключены входные столбцы (атрибуты), обладающие уровнем корреляции с целевым атрибутом (выходным атрибутом) ниже среднего. Следовательно, в ходе выполнения корреляционного анализа удалось исключить 15 малозначущих переменных.

2. Факторный анализ. В отличие от корреляционного анализа факторный анализ [4] учитывает взаимосвязь не только между двумя наборами данных (атрибутами). Цель факторного анализа заключается в понижении размерности пространства факторов. Это необходимо в случаях, когда входные факторы коррелированы друг с другом, т.е. взаимозависимы. В факторном анализе речь идет о выделении из множества измеряемых характеристик объекта новых факторов более адекватно отражающих свойства объекта.

Аналогично оценим эффективность метода на наборе данных полетного мониторинга в СПО Deductor 5.3. После выбора входных данных на анализ был определен метод факторного решения – метод ортогонального вращения Varimax (этот метод является наиболее часто применяемым, поскольку он облегчает интерпретацию факторов) и число выбираемых факторов – 25.

Для заданных переменных 25 рассчитаны главные компоненты. Таблица с результатами представлена на рис. 2.

Для того чтобы определить, какие главные компоненты необходимо включить в выходной набор, следует установить порог значимости, рекомендуемые значения которых выделены синим цветом. Регулируя данные значения, количество переменных, заданных на предыдущем шаге, может варьироваться. Последней по списку попадающей в выходной набор главной компонентой станет первая, у которой суммарный вклад в результат превысит порог значимости.

В ходе проведения факторного анализа из входного набора данных удалось выделить 25 главных компонент (факторов). Эффективность данного подхода определим на этапе моделирования классификатора.

Аналогичная задача по понижению размерности может быть выполнена с помощью нелинейных преобразований на базе автоассоциативных сетей (воспроизводят на выходе свои же данные). Число вну-

тренных элементов данного рода сетей небольшое, что и заставляет сжимать данные. Типичная структура сети 5 слоев – средний слой сжимает, крайние выполняют нелинейные преобразования.

3. Анализ чувствительности. При моделировании нейронных сетей в СПО Statistical Neural Networks можно использовать различные комбинации входных переменных. Поэтому имеется возможность игнорировать некоторые переменные, так что полученная сеть не будет использовать их в качестве входов. Так же можно провести анализ чувствительности сети к входным переменным. Такая процедура позволяет сделать вывод об относительной важности входных перемен-

ных для конкретной нейронной сети и при необходимости удалить входы с низкими показателями чувствительности [2].

Используя набор данных полетного мониторинга, оценим эффективность анализа чувствительности. В ходе анализа поочередно будут исключаться входные переменные. Для того чтобы определить чувствительность каждой из переменных, сеть сначала обучается на тестовых значениях (т.е. на всем наборе данных, со всеми доступными переменными) и определяется ошибка $N1$ сети. Затем сеть обучается на тех же значениях, но в этот раз происходит «исключение» наблюдаемых значений с определением ошибки $N2$ сети.

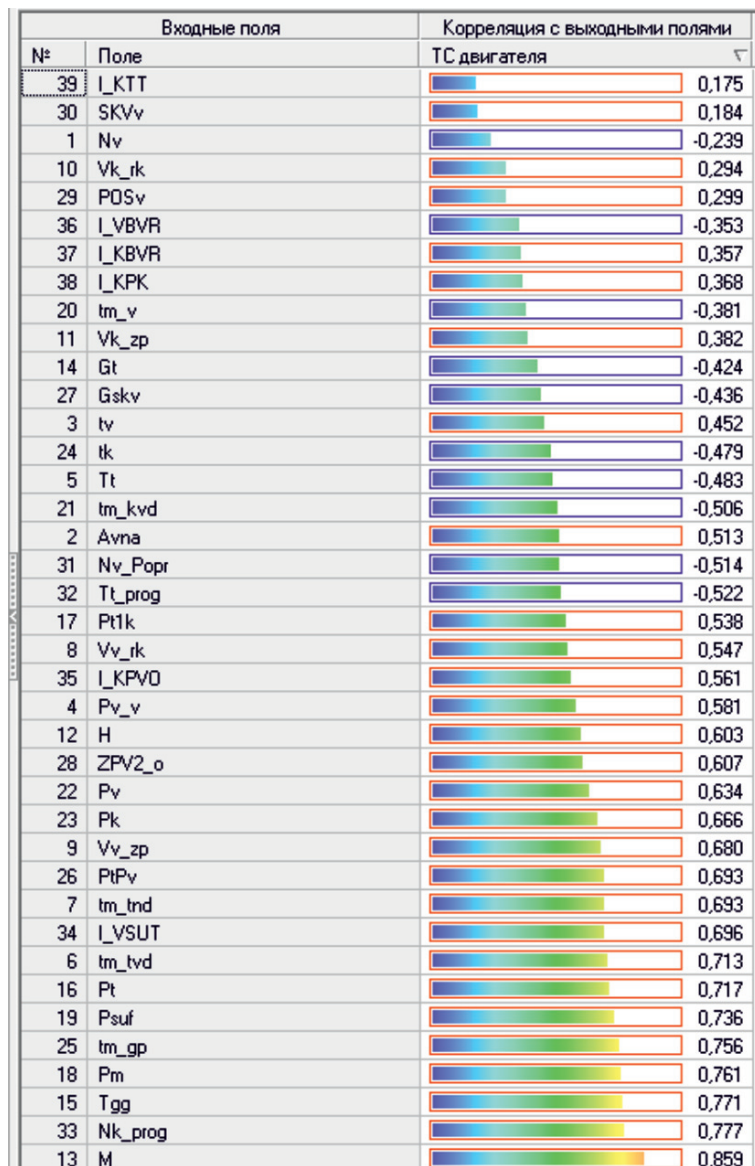


Рис. 1. Матрица корреляции

Главные компоненты	Собственное значение	Вклад в результат	Суммарный вклад
☒ Значение 11	0,600	01,54 %	95,37 %
☒ Значение 12	0,551	01,41 %	96,78 %
☒ Значение 13	0,339	00,87 %	97,65 %
☒ Значение 14	0,306	00,78 %	98,43 %
☒ Значение 15	0,141	00,36 %	98,79 %
☒ Значение 16	0,125	00,32 %	99,12 %
☒ Значение 17	0,078	00,20 %	99,32 %
☒ Значение 18	0,073	00,19 %	99,50 %
☒ Значение 19	0,048	00,12 %	99,62 %
☒ Значение 20	0,036	00,09 %	99,72 %
☒ Значение 21	0,023	00,06 %	99,78 %
☒ Значение 22	0,019	00,05 %	99,83 %
☒ Значение 23	0,015	00,04 %	99,86 %
☒ Значение 24	0,012	00,03 %	99,89 %
☒ Значение 25	0,010	00,02 %	99,92 %
☐ Значение 26	0,008	00,02 %	
☐ Значение 27	0,005	00,01 %	
☐ Значение 28	0,005	00,01 %	
☐ Значение 29	0,004	00,01 %	
☐ Значение 30	0,003	00,01 %	
☐ Значение 31	0,002	00,01 %	

Порог значимости (%) 99,91

Рис. 2. Таблица со списком рассчитанных главных компонент

Таблица 2

Результаты анализа чувствительности

	Ошибка	Vv_zp	Tv	Pk	Pt	Tt	Tm_v	...
Отношение 1	0,396	0,6785	0,6787	0,742	0,6792		1,004	...
Ранг 1		19	16	13	15		3	...
Отношение 2	0,273	0,989	0,961	1,0009	0,983	1,045	1,049	...
Ранг 2		25	33	22	27	11	10	...
Отношение 3	0,363	0,924	1,002	1,0074	1,0004		0,9994	...
Ранг 3		25	13	11	14		17	...
Отношение 4	0,144	0,944	1,013	1,0451	0,984	1,038	0,991	...
Ранг 4		37	20	14	33	17	27	...

Так как мы удалили некоторую информацию, которую использует сеть (т.е. одну из входных переменных), то резонно ожидать ухудшение ошибки. Значение чувствительности – это отношение ошибки $N2$ к исходной ошибке $N1$. Чем чувствительнее сеть к данному входу, тем большее ухудшение можно ожидать и, таким образом, тем больше отношение. Если отношение меньше либо равно 1, тогда отключение переменной не влияет на производительность сети. Полученные результаты показаны в табл. 2.

В ходе анализа было обучено 10 нейронных сетей типа многослойный перцептрон, из них выбрано 4 с наименьшей ошибкой. Для каждой переменной рассчи-

тано отношение ошибок и проставлен ранг. Как видно из таблицы, параметр Tt не участвовал в обучении 1 и 3 сетей, что негативно повлияло на ошибку, следовательно, можно сделать вывод о высокой значимости данной переменной. Так же анализируя таблицу, можно сделать предположение о том, что параметр Vv_zp не является значимым и его можно исключить. Таким образом, был составлен набор данных из 27 переменных (атрибутов).

Для сравнительного анализа эффективности подходов было составлено три набора данных: набор после корреляционного анализа, набор после факторного анализа, набор после анализа чувствительности.

Таблица 3

Результаты моделирования при обучении на исходных данных

	Наст. Неисправен	Наст. Исправен
Прогн. Неисправен	19	4
Прогн. Исправен	27	368
Отзыв класса	41,30 %	98,02 %
Исходные данные. Точность классификации – 92,58 %		

Таблица 4

Результаты моделирования при обучении на данных после факторного и корреляционного анализа

Данные после факторного анализа. Точность классификации 91,39 %		
	Наст. Неисправен	Наст. Исправен
Прогн. Неисправен	25	15
Прогн. Исправен	21	357
Отзыв класса	54,35 %	95,97 %
Данные после корреляционного анализа. Точность классификации 91,39 %		
Прогн. Неисправен	24	8
Прогн. Исправен	22	364
Отзыв класса	52,17 %	97,85 %

Таблица 5

Результаты моделирования при обучении на данных после анализа чувствительности

	Наст. Неисправен	Наст. Исправен
Прогн. Неисправен	10	51
Прогн. Исправен	5	178
Отзыв класса	66,66 %	77,72 %

Для того чтобы оценить эффективность и выявить наиболее результативный из подходов, необходимо осуществить проверку при моделировании классификатора. Сравнивая точности моделей, рассчитанных на основе данных матрицы ошибок, можно сделать вывод о пригодности и целесообразности использования определенного метода подготовки данных.

Итак, в качестве отправной точки нужно определить точность модели [5] на исходных данных, т.е. на данных без каких-либо манипуляций с атрибутами.

Из табл. 3 видно, что неисправные состояния двигателя распознаются с очень низкой точностью, это, несомненно, является проблемой.

Оценивая результаты после факторного и корреляционного анализа, видно, что повысилась точность распознавания неисправного состояния, но данный показатель до сих пор не является достаточно высоким для эффективного функционирования модели.

Точность классификации модели, обученной на данных после анализа чувствительности, составляет 77,04 %. Исключение

некоторых атрибутов положительно повлияло на распознавание неисправного состояния двигателя, но в то же время значительно упала точность классификации модели в целом, в связи с тем, что точность определения класса «исправен» упала на 20 %.

Таким образом, применяя эти подходы по подготовке данных, была улучшена точность распознавания класса «неисправен» на 20 %, однако точность распознавания класса «исправен» упала на 20 %.

В данной статье был проведен сравнительный анализ методов подготовки данных к моделированию классификатора для оценки технического состояния авиадвигателя. Используемые методы целесообразно применять, особенно учитывая тот факт, что приоритет определения действительно существующей неисправности двигателя выше, чем ложное определение его неисправности. Следовательно, повышение точности распознавания неисправного класса оправдывается даже с учетом падения точности определения исправного класса. Оценивая ситуацию в целом, можно отметить, что модели не являются достаточно точны-

ми для выполнения поставленной задачи по классификации технического состояния авиационного двигателя. Причиной этого может быть значительный дисбаланс значений между классами, а именно недостаточное количество данных по классу «неисправен». Анализ данной проблемы описан в статье [3].

Список литературы

1. Гайдешев И. Анализ и обработка данных: специальный справочник. – СПб.: Питер, 2001. – 752 с.
2. Справочные руководства СПО «Statistica». [Электронный ресурс]. – Режим доступа: <http://www.statsoft.ru/resources/support/info.php> (дата обращения: 20.05.16).
3. Egor Obukhov, Handling the problem of unbalanced data sets in the classification of technical equipment states (Material of conference of 4th International Conference on Applied Innovations in IT). Germany. – 2016.

4. Richard L. Factor Analysis, second edition, Hillsdale: Lawrence Erlbaum Associates. 1983. – 454 p.

5. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publ., 2011. – 629 p.

References

1. Gajdyshev I. Analiz i obrabotka dannyh: specialnyj spravochnik. SPb.: Piter, 2001. 752 p.
2. Spravochnye rukovodstva SPO «Statistica». [Elektronnyj resurs]. Rezhim dostupa: <http://www.statsoft.ru/resources/support/info.php> (data obrashheniya: 20.05.16).
3. Egor Obukhov, Handling the problem of unbalanced data sets in the classification of technical equipment states (Material of conference of 4th International Conference on Applied Innovations in IT). Germany. 2016.
4. Richard L. Factor Analysis, second edition, Hillsdale: Lawrence Erlbaum Associates. 1983. 454 p.
5. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publ., 2011. 629 p.