

ФАКТОРНЫЙ АНАЛИЗ ДАННЫХ В ЗАДАЧЕ ДИАГНОСТИКИ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ

В. И. Дубовская, К. А. Брыжина
Волгоградский государственный университет. г. Волгоград
dubovskajav@volsu.ru

Аннотация

Данная работа посвящена вопросу использования факторного анализа данных в рамках задачи диагностики рака молочной железы. В статье анализируются возможности применения данного многомерного статистического метода с целью установления влияния диагностических признаков на результат работы классификатора. Проведен вычислительный эксперимент, в ходе которого использовались два классификатора: на основе логистической регрессии и на основе случайного леса. Эксперимент показал, что для одной и той же обучающей выборки структура факторной нагрузки для этих методов совпадает только для тех факторов, которые способствуют правильному результату.

Введение

На данный момент актуальной задачей медицины является ранняя дифференциальная диагностика онкологических заболеваний, в частности рака молочной железы. Одним из способов решения проблемы является использование новых эффективных методов обследования пациенток, среди которых можно отметить метод микроволновой радиотермометрии. Последний базируется на измерении собственного электромагнитного излучения тканей в инфокрасном диапазоне длин волн и микроволновом диапазоне, благодаря чему становится возможным неинвазивное выявление температурных аномалий на глубине нескольких сантиметров [1].

Данному методу посвящен ряд работ, в которых специалистами был описан подход использования методов математического моделирования и искусственного интеллекта по отношению к термометрическим данным с целью выявления подобных аномалий внутри молочных желез [2-4]. К полученным на данный момент результатам, приведенным в работах вышеуказанных авторов, можно отнести: построение эмпирических гипотез поведения температурных полей молочных желез, описание данных гипотез в математической форме, проектирование признакового пространства, используемого в ходе проведения классификации пациенток методами машинного обучения.

В настоящей работе предлагается подход, который позволит исследовать исходный набор признаков с целью установить, на сколько каждый из них влияет на ошибку классификатора. Использование факторного анализа позволит:

- во-первых, интерпретировать результаты классификатора;
- во-вторых, внести изменения в математическую модель классификатора таким образом, чтобы его работа учитывала влияние признака на правильность или ошибочность поставленного диагноза.

Материалы и методы исследования

В данной работе использовалась база термометрических данных, содержащая в себе информацию о пациентках онкологических центров. Для проведения исследования база была преобразована таким образом, что каждый элемент соответствует отдельной молочной железе пациентки. Степень заболевания оценивается по 6-бальной шкале th0-th5. Каждое значение содержит в себе диагностическое описание температурных аномалий внутри органа [5]. Будем считать, что в случаях th0-th2 – молочная железа здорова, в случаях th3-th5 – молочная железа в группе риска.

Вся выборка данных H содержит в себе 9182 элемента. Deskриптивная математическая модель диагностического состояния пациентки описывается 62 функциями. Эти функции вычислялись на основе термометрических данных, полученных при измерении пациентки комплексом РТМ-01-РЭС.

Выборку температурных данных можно представить в виде:

$$S = \begin{pmatrix} t_0^1 & \dots & t_{43}^1 \\ \vdots & \ddots & \vdots \\ t_0^k & \dots & t_{43}^k \\ t_0^{k+1} & \dots & t_{43}^{k+1} \\ \vdots & \ddots & \vdots \\ t_0^n & \dots & t_{43}^n \end{pmatrix},$$

где t_i^j , $i = 0, \dots, 9$ – температуры i -ой точки правой молочной железы j -ой пациентки (РТМ-диапазон);

t_i^j , $i = 10, \dots, 19$ – температура $[i - 10]$ точки правой молочной железы j -ой пациентки (ИК-диапазон);

t_i^j , $i = 20, \dots, 29$ – температура $[i - 20]$ точки левой молочной железы j -ой пациентки (РТМ-диапазон);

t_i^j , $i = 30, \dots, 39$ – температура $[i - 30]$ точки левой молочной железы j -ой пациентки (ИК-диапазон); $t_{40}^j = T_1$,

$t_{41}^j = T_2$ – температуры опорных точек j -ой пациентки (РТМ-диапазон);

$t_{42}^j = T_1$, $t_{43}^j = T_2$ – температуры опорных точек j -ой пациентки (ИК-диапазон).

Для каждого качественного признака онкологического заболевания, к примеру, асимметрии температурных полей, соотношению кожной и глубинной температур, А.Г. Лосевым и В.В. Левшинским был построен набор моделирующих функций от температур, значение которых количественно характеризует признак. Подробнее с процессом построения функций можно ознакомиться в работе [6].

Для проведения классификации были использованы такие методы как логистическая регрессия и случайный лес, реализованные И.Е. Поповым [7]. После проведения классификации были получены две выборки:

выборка R , содержащая в себе выборку данных H о пациентах и полученный методом случайного леса результат классификации (чувствительность алгоритма составила – 87.5%, специфичность – 91.7%).

выборка L , содержащая в себе выборку данных H о пациентах и полученный методом логистической регрессии результат классификации (чувствительность алгоритма составила – 89.6%, специфичность – 91.2 %).

Здесь специфичность – доля вернодиагностированных здоровых пациентов, чувствительность – доля вернодиагностированных пациентов группы риска.

Структура выборок R и L :

В выборке R присутствует 8396 элементов, соответствующие верной классификации (522 – метка «Группа риска», 7874 – метка «Здоров»). 786 элементов – неверной классификации (74 – метка «Группа риска», 712 – метка «Здоров»).

В выборке L присутствует 8367 элементов соответствующие верной классификации (534 – метка «Группа риска», 7833 – метка «Здоров») и 815 – неверной классификации (62 – метка «Группа риска», 753 – метка «Здоров»).

Обозначим с индексом 0 выборку, соответствующую элементам с верной классификацией, а с индексом 1 – неверной. Тогда:

$$R = R_0 \cup R_1. \quad (1)$$

$$L = L_0 \cup L_1. \quad (2)$$

Полученные выборки можно представить в виде:

$$R = \begin{pmatrix} f_1^1 & \dots & f_{62}^1 & d^1 & r^1 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ f_1^k & \dots & f_{62}^k & d^k & r^k \\ f_1^{k+1} & \dots & f_{62}^{k+1} & d^{k+1} & r^{k+1} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ f_1^n & \dots & f_{62}^n & d^n & r^n \end{pmatrix} \quad (3)$$

$$L = \begin{pmatrix} f_1^1 & \dots & f_{62}^1 & d^1 & l^1 \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ f_1^k & \dots & f_{62}^k & d^k & l^k \\ f_1^{k+1} & \dots & f_{62}^{k+1} & d^{k+1} & l^{k+1} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ f_1^n & \dots & f_{62}^n & d^n & l^n \end{pmatrix} \quad (4)$$

где f_i^j – значение рассчитанных признаков;

i – номер признака;

j – номер пациента;

n – количество элементов в выборке;

d^j – метка класса, соответствующая поставленному врачом диагнозу;

r^j – результат классификатора случайный лес для соответствующего пациента;

l^j – результат классификатора логистическая регрессия для соответствующего пациента.

Очевидно, что на полученный диагноз влияют используемые в процессе классификации признаки. Для определения уровня влияния признаков на результат классификации в данной работе был использован факторный анализ, который основан на выявлении посредством анализа вариаций переменных и их корреляций непосредственно не наблюдаемых и не измеряемых характеристик – общих факторов [8].

Для выделения факторов достаточно, чтобы корреляционная матрица была положительно определена, а также, чтобы детерминант этой матрицы не был равен нулю.

Однако на данном наборе признаков данные требования не соблюдаются, что подтверждено проверкой в среде IBM SPSS Statistics. Вызвано это тем, что некоторые признаки имеют между собой сильную корреляцию. Поэтому было решено убрать такие признаки. (В дальнейшем за рамками представляемой работы планируется восстанавливать значения факторных нагрузок по этим признакам).

При отборе признаков, которые будут участвовать в построении факторных моделей.

использовался подход, примененный в работе [9]:

– Рассчитывались корреляции для всех пар признаков.

– Находились корреляции со значениями больше либо равно 0.7 (данное значение было выбрано из таблицы шкал Чеддока, согласно которой корреляция со значениями от 0.7 включительно является сильной).

– Убирали первый признак из пары с сильной корреляцией, второй оставляли.

Ниже представлена табл. 1 с количеством отсеянных признаков и количеством оставшихся признаков для проведения факторного анализа.

Таблица 1 – Количество отобранных признаков

Статус признака	Выборка признаков по выборкам			
	R_0	R_1	L_0	L_1
Отсеян	42	41	42	44
Оставлен	20	21	20	18

Для исследования был выбран метод главных факторов и ортогональный метод вращения Варимакс [8].

Результаты вычислительного эксперимента и его обсуждение

Выделение факторов и поворот факторных нагрузок проведен при помощи готового программного обеспечения IBM SPSS Statistics. Анализ был проведен поочередно для каждой выборки R_0 , R_1 , L_0 и L_1 .

Для отбора наиболее значимых факторов, был использован подход, основанный на совместном использовании критерий Кайзера и Меттелла [10]. Ниже приведена табл. 2 с номерами отобранных факторов и их вкладом в общую (суммарную) дисперсию для выборки R_0 .

Поскольку для данного набора данных было оставлено 20 признаков, то 100% дисперсии последних объясняет 20 факторов. Однако, начиная с 10 фактора, вклад в суммарную дисперсию становится настолько малым, что практически не влияет на погрешность модели.

Аналогичное количество факторов было выбрано и при помощи критерия «каменной осыпи» (рис.1). Согласно данному критерию, строится график собственных дисперсий переменных и выбирается такое количество факторов, где значения перестают сильно меняться [8].

Аналогичным способом определялось количество факторов для остальных выборок R_1 , L_0 и L_1 .

Таблица 2 – Фрагмент данных с объясненной совокупной дисперсией

	Извлечение дисперсий после вращения		
	Всего	% дисперсии	Суммарный %
1	2.159	10.795	10.795
2	1.615	8.075	18.870
3	1.598	7.991	26.861
4	1.509	7.544	34.405
5	1.480	7.402	41.806
6	1.370	6.848	48.654
7	1.369	6.845	55.499
8	0.922	4.609	60.108
9	0.808	4.040	64.149

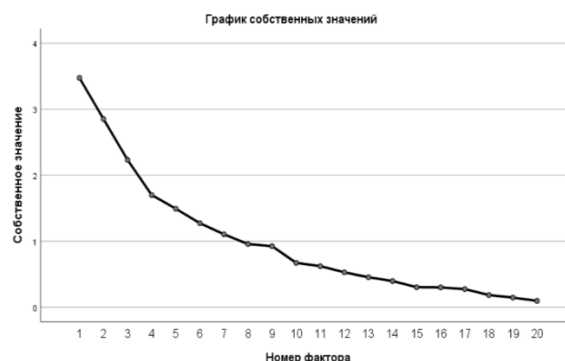


Рисунок 1 - График каменной осыпи для выборки R_0

Для выборок R_0 и L_0 было выделено по 9 факторов, для выборок R_1 , L_1 – по 8 факторов.

В табл. 3 приведены значения суммарной дисперсии для каждой из выборок.

Таблица 3 – Значения суммарной дисперсии

Выборка	Суммарная дисперсия
R_0	64.149
R_1	64.777
L_0	64.673
L_1	67.394

После выделения факторной структуры был применен ортогональный метод вращения Варимакс. Цель использования данного метода – получение более простой факторной структуры, при которой минимизируется число переменных с высокими факторными нагрузками.

Таким образом, упрощается процесс интерпретации результатов – нахождения признаков, которые вносят значимый вклад в фактор. Следовательно, и в результат работы алгоритмов классификации.

В табл. 4 представлен фрагмент матрицы факторных нагрузок после вращения для выборки R_0 .

Таблица 4 – Фрагмент матрицы факторных нагрузок после вращения

Матрица факторных нагрузок после вращения									
	Факторный анализ								
	1	2	3	4	5	6	7	8	9
MG005	-0.031	-0.083	0.668	-0.348	0.000	0.066	0.061	-0.110	0.147
MG013	0.117	0.005	0.089	0.011	-0.054	0.828	0.009	-0.014	0.148
MG014	0.157	-0.057	0.976	-0.045	0.007	0.056	-0.048	0.078	0.038
MG015	0.011	-0.182	-0.172	0.819	0.080	0.044	0.094	-0.124	-0.100
MG017	-0.034	0.036	-0.110	0.715	-0.378	-0.036	-0.003	-0.239	0.042
MG018	-0.076	-0.739	0.075	0.092	0.023	-0.011	-0.153	-0.050	-0.021
MG022	0.266	-0.008	-0.005	0.005	-0.026	0.702	-0.012	0.011	-0.036
MG023	0.442	0.266	0.326	0.121	0.052	-0.037	-0.045	-0.023	-0.137
MG024	0.028	0.312	-0.062	-0.241	-0.133	-0.048	-0.101	0.508	0.139
MG026	0.008	-0.062	0.026	-0.167	0.390	0.036	0.058	0.684	0.020
MG027	0.086	0.860	-0.007	-0.021	-0.011	-0.006	0.180	0.033	0.084
MG035	0.009	0.073	0.022	0.015	0.075	-0.013	-0.108	0.076	0.355
MG050	-0.044	-0.007	0.021	0.184	-0.779	-0.004	0.196	0.101	-0.150
MG051	-0.109	-0.141	0.044	-0.150	0.163	0.041	-0.706	-0.126	0.360
MG053	0.350	-0.024	0.146	-0.179	0.026	0.295	-0.033	-0.053	0.615
MG058	0.746	0.110	0.091	-0.002	-0.044	0.202	0.300	-0.038	0.016
MG059	-0.067	-0.038	0.021	0.061	0.713	-0.091	-0.007	0.221	0.031
MG060	0.151	0.268	0.021	-0.016	-0.060	0.016	0.800	-0.108	-0.029
MG061	0.695	0.016	0.006	-0.036	-0.017	0.132	0.044	0.005	0.024
MG062	0.797	0.027	-0.002	-0.012	-0.001	0.142	0.014	0.066	0.212

После получения итоговых матриц было выполнено отсеивание факторных нагрузок ниже 0.5, как незначимых.

После этого было проведено сравнение значений последних для обоих алгоритмов классификации.

Данные сравнения выборок приведены в табл. 5 и в табл. 6.

При интерпретации вычисленных факторных нагрузок было выявлено:

– при сравнении двух выборок R_0 и L_0 все значимые признаки совпали, значения факторных нагрузок практически идентичны; признаки, которые значимо влияют только на результат дерева решений или логистической регрессии отсутствуют;

– при сравнении двух выборок R_1 и L_1 можно заметить, что, наоборот, совпадающих признаков по факторам практически нет; при сравнении факторных нагрузок большинство значений различаются минимум на 0.1 - 0.3.

Таблица 5 - Сравнение выборок R_0 и L_0

Признак	Номер фактора	R_0	L_0
		Факторная нагрузка	Факторная нагрузка
MG005	3	0.668	0.668
MG013	6	0.828	0.818
MG014	3	0.976	0.974
MG015	4	0.819	0.842
MG017	4	0.715	0.683
MG018	2	-0.739	-0.747
MG022	6	0.702	0.718
MG023	1	0.442	0.427
MG024	8	0.508	0.462
MG026	8	0.684	0.726
MG027	2	0.860	0.856
MG035	9	0.355	0.350
MG050	5	-0.779	0.792
MG051	7	-0.706	-0.661
MG053	9	0.615	0.657
MG058	1	0.746	0.741
MG059	5	0.713	0.688
MG060	7	0.800	0.817
MG061	1	0.695	0.688
MG062	1	0.797	0.792

Таблица 6 - Сравнение выборок R_1 и L_1

Признак	Номер фактора	R_1	L_1
		Факторная нагрузка	Факторная нагрузка
MG004	2	0.706	0.079
	3	-0.008	0.818
MG013	2	0.906	0.022
	3	-0.003	0.784
	8	-0.157	0.592
MG014	6	-0.025	0.818
	7	0.630	-0.028
MG015	5	-0.562	-0.564
MG017	2	-0.101	-0.647
	4	-0.748	-0.003
MG022	1	0.588	0.399
	8	-0.250	0.699
MG023	6	0.010	0.536
	7	0.710	0.167
MG024	5	0.861	0.831
MG025	1	0.811	-
MG026	2	-0.033	0.600
	4	0.668	0.004
MG027	7	-0.056	0.532
MG033	4	0.040	0.837
	6	0.709	0.020
MG035	4	0.095	0.833
	6	0.910	-0.075
MG050	2	-0.068	-0.700
	4	-0.543	0.003
	8	-0.516	0.076
MG051	3	-0.718	-
MG053	1	0.449	0.561
	3	-0.200	0.529
MG057	1	0.788	-
MG058	1	0.778	0.627
MG059	2	-0.022	0.710
	4	0.563	0.143
MG060	3	0.916	0.009
	7	0.097	0.834
MG062	1	0.825	0.901

Выводы

Представляет интерес полученный результат по факторам, влияющим на правильность классификации. Оказалось, что факторные нагрузки по отобраным факторам практически идентичны для обоих методов классификации.

В качестве причины такого явления можно указать гипотезу о том, что факторы, влияющие на правильность, обусловлены не моделью классификатора, а информационной моделью, согласно которой порождены исходные данные.

Для выборок с ошибочными результатами почти все факторные нагрузки различны.

В совокупности это говорит о том, что классификаторы, в целом, опираются на верную информационную модель, а ошибки алгоритмов классификации обусловлены ошибками математической модели классификатора.

Исходя из этого можно сказать о возможности улучшения классификаторов путем изменения математической модели в том направлении, которое обеспечит более значимую роль тех признаков, которые послужат проводниками особенностей информационной модели исходных данных в математическую модель классификатора.

Литература

1. Vesnin, S. Modern Microwave Thermometry for Breast Cancer / S. Vesnin, A. K. Turnbull, J. M. Dixon, I. Goryanin // Journal of Molecular Imaging & Dynamics, 2017. - Vol. 7. Is. 2. DOI: 10.4172/2155- 9937.1000136.
2. Polyakov, M.V. Application of computer simulation results and machine learning in the analysis of microwave radiothermometry data / M. V. Polyakov, I. E. Popov, A. G. Losev, A. V. Khoperskov // Математическая физика и компьютерное моделирование, 2021. - Т. 24. - № 2. - С. 27-37.
3. Замечник, Т. В. Управляемый классификатор в диагностике рака молочной железы по данным микроволновой радиотермометрии / Т. В. Замечник, А. Г. Лосев, А. Ю. Петренко // Математическая физика и компьютерное моделирование, 2019. - Т. 22. - № 3. - С. 53-67.
4. Зенович, А. В. Алгоритмы формирования двумерных признаков диагностики заболеваний молочных желез методами комбинированной радиотермометрии / А. В. Зенович, Н. А. Батурин, Д. А. Медведев, А. Ю. Петренко // Математическая физика и компьютерное моделирование, 2018. - Т. 21. - № 4. - С. 44-56.
5. Lashkari, A., Pak, F., Firouzmand, M. Full Intelligent Cancer Classification of Thermal Breast Images to Assist Physician in Clinical Diagnostic Applications // J Med Signals Sens, 2016. - Vol. 1. - No. 6. - PP. 12-24.
6. Лосев, А. Г. Интеллектуальный анализ данных микроволновой радиотермометрии в диагностике рака молочной железы / А. Г. Лосев, В. В. Левшинский // Математическая физика и компьютерное моделирование, 2017. - Т. 20. - № 5. - С. 49-62.
7. Germashev, I. Fuzzy Inference of the effectiveness factors of the computational model for the diagnosis of the breast cancer / I. Germashev, V. Dubovskaya, A. Losev, I. Popov // 2021 3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), 2021. - PP. 528-533.
8. Окунь, Я. Факторный анализ. Перевод с польского Г. З. Давидовича. - М., 1974.

9. Гермашев, И. В. Факторный анализ влияния признаков на точность диагностики рака молочной железы по данным микроволновой радиотермометрии / И. В. Гермашев, В. И. Дубовская, А. Г. Лосев, И. Е. Попов // Прикаспийский журнал: управление и высокие технологии, 2022. - №1. - С. 139-148.

10. Гермашев, И. В. Модель иерархии признаков в диагностике рака молочной железы по данным микроволновой радиотермометрии / И. В. Гермашев, В. И. Дубовская, А. Г. Лосев // Математические методы в технологиях и технике, 2021. - № 6. - С. 78 - 83.

Дубовская В. И., Брыжина К. А. Факторный анализ данных в задаче диагностики рака молочной железы. Данная работа посвящена вопросу использования факторного анализа данных в рамках задачи диагностики рака молочной железы. В статье анализируются возможности применения данного многомерного статистического метода с целью установления влияния диагностических признаков на результат работы классификатора. Проведен вычислительный эксперимент. в ходе которого использовались два классификатора: на основе логистической регрессии и на основе случайного леса. Эксперимент показал, что для одной и той же обучающей выборки структура факторной нагрузки для этих методов совпадает только для тех факторов, которые способствуют правильному результату.

Ключевые слова: факторный анализ, диагностика рака молочной железы, логистическая регрессия, случайный лес, структура факторной нагрузки

Dubovskaya V. I., Bryzhina K. A. Factor analysis of data in the problem of diagnosing breast cancer. This work is devoted to the issue of using factorial data analysis in the framework of the task of diagnosing breast cancer. The article analyzes the possibilities of using this multivariate statistical method in order to establish the influence of diagnostic features on the result of the classifier. A computational experiment was carried out. during which two classifiers were used: based on logistic regression and based on a random forest. The experiment showed that for the same training sample, the structure of the factor load for these methods is the same only for those factors that contribute to the correct result.

Key words: factor analysis, breast cancer diagnostics, logistic regression, random forest, factor loading structure.

Статья поступила в редакцию 25.05.2022
Рекомендуется к публикации профессором Федяевым О. И.