Симптомно-синдромальный подход к решению многофакторных задач с приложением в медицине

Леонович Роман Александрович, 622-я группа

Санкт-Петербургский Государственный Университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель — к.ф.-м.н. доцент **Н.П. Алексеева**
Рецензент — Управляющий директор ПАО банк ВТБ.
Биостатистик, ФГБУ «НМИЦ ПН им. В.М.Бехтерева **Е.П. Скурат**

Санкт-Петербург
2023г.

Цели и задачи

Целью данной работы является исследование методов симптомно-синдромального анализа как способа улучшения качества классификации.

- Расширение пространства признаков с помощью суперсиндромов для улучшения точности классификации данных.
- Исследование расслоения по тривиальным симптомам как метода улучшения классификации.
- Разработка оптимального алгоритма частичной классификации неполных данных.
- Изучение возможностей применения симптомного анализа для выявления факторов риска подпопуляций.

Используемые классификаторы

Пусть $\mathbb{X} \in \mathbb{R}^{n \times p}$ — набор независимых переменных для n индивидов, а $\mathbf{W} \in \mathbb{R}^{p}$ — вектор весов для p признаков.

- Нейронные сети
  1. Архитектура нейронной сети.
     - Количество скрытых слоев в нейронной сети.
     - Количество нейронов на каждом слое.
     - Функции активации применяемые к слоям. В работе используется сигмоидная: $\sigma(x) = \frac{1}{1-e^{-x}}$

     Результатом работы каждого слоя будет значение функции активации от линейной комбинации значений переменных и их весов: $\mathbf{Y} = \sigma(\mathbb{X}\mathbf{W})$, где $\mathbf{Y} \in \mathbb{R}^{n}$ — результат работы классификатора.
  2. Метод оптимизации.
  3. Функция потерь.
- Логистическая регрессия
- Дискриминантный анализ

Основные положения симптомного анализа [Алексеева — 2021]

Пусть $X = (X_0, \ldots, X_k)^T$, такой, что $X_i \in \{0, 1\}$.

$s(X_0, \ldots, X_k) = \alpha_0 X_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k \, (mod 2)$ — симптом.

Совокупность всевозможных симптомов порядка k представленная в виде вектора

$$S\left(\mathbb{X}_{k+1}\right) = S\left(\mathbb{X}_k, X_k, S\left(\mathbb{X}_k\right) + X_k(mod 2)\right) \tag{1}$$

— аддитивный синдром $S\left(\mathbb{X}_{k+1}\right)$ k

Таким же образом, заменяя операцию сложения умножением в [1], можно получить мультипликативный синдром

$$V\left(\mathbb{X}_{k+1}\right) = V\left(\mathbb{X}_k, X_k, V\left(\mathbb{X}_k\right) \cdot X_k(mod 2)\right) \tag{2}$$

Алгоритм построения суперсиндрома

Результат построения аддитивного синдрома $S(\mathbb{X}_{k+1})$ по элементам мультипликативного синдрома $V(\mathbb{X}_{k+1})$ будем называть суперсиндромом $SV(\mathbb{X}_{k+1})$.

Алгоритм построения суперсиндрома по трем признакам:

1. Рассматриваем тройку признаков X_0, X_1, X_2.
2. Мультипликативный синдром:

$$V\left(X_0, X_1, X_2\right) = \left(X_0, X_1, X_{01}, X_2, X_{02}, X_{12}, X_{012}\right)$$

3. Аддитивный синдром (суперсиндром):

$$SV\left(X_0, X_1, X_{01}, X_2, X_{02}, X_{12}, X_{012}\right) = \left(X_0, X_1, X_0 + X_1, X_0X_1, \ldots\right)$$

Частичная классификация

Пусть имеется матрица независимых переменных $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ по которым с помощью классификатора F строится прогноз для вектора зависимых переменных $\mathbf{Y}$.

Определение

Возьмем выборки $\mathbb{X}_\tau = (\mathbf{X}_{t_1} \ldots \mathbf{X}_{t_k})$ из $\mathbb{X}$, где $\tau = (t_1, \ldots t_k) \subseteq (1, 2, \ldots m)$. Результат работы классификатора $F_\tau(\mathbb{X}_\tau)$ будем называть частичным предсказанием, а F_τ— частичным классификатором

Прогнозы в виде апостериорных вероятностей будем усреднять для каждого индвивида $y_i \in \mathbf{Y}$.

Пример улучшения классификации по полным данным

Данные по прогнозированию сердеченого заболевания[1] включают в себя 297 индивидов и 13 признаков из которых 8 являются категориальными. Прогнозируемая переменная отвечает за наличие болезни сердца у пациента (0 – болезни нет, 1 – болезнь есть).

- (a) **Количество сосудов выявленных при коронарографии больше двух.**
- (b) **Бессимптомная боль.**
- (c) **Имеются холодные пятна при талиевом сканировании сердца.**
- (d) Наблюдается аномалия на ЭКГ.
- (e) Наклон сегмента ST не положительный.
- (f) Пол.
- (g) Уровень сахара в крови больше чем 120 мг/дл.
- (h) Стенокардия, вызванная физическими упражнениями.

С помощью элементов суперсинрома построенного из признаков a, b, c было расширено множество признаков исходных данных.

[1] Данные взяты из UCI Machine Learning Repository

Точность классификации при расширении множества признаков

Таблица: Сравнение точности классификации моделей

	Train	Test
Лог. Регрессия	0.894	0.865
Лог. Регрессия при добавлении синдромов S_3	0.860	0.885
Лог. Регрессия при добавлении части синдромов S_3	0.882	0.923
Линейный лискриминантный анализ	0.857	0.867
ЛДА при добавлении синдромов S_3	0.852	0.867
ЛДА при добавлении части синдромов S_3	0.865	0.867
Нейронная сеть	0.793	0.767
Нейронная сеть при добавлении синдромов S_3	0.869	0.883
Нейронная сеть при добавлении части синдромов S_3	0.873	0.883

- Лучше всего на расширение множества признаков реагирует нейронная сеть. Такой эффект достигается за счет проектирования данных при многослойности.

.

Результат улучшения классификации при расслоении по симптому

		Predicted	
		0	1
Actual	0	146	25
	1	14	112
Точность		**0.865**	

Таблица: Матрица ошибок, при $s(a,b,c) = a(b+c) = 0$ (благополучные по основным характеристикам коронарографии и сканирования)

		Predicted	
		0	1
Actual	0	119	13
	1	2	6
Точность		**0.893**	

Таблица: Матрица ошибок, при $s(a,b,c) = a(b+c) = 1$ (неблагополучные по основным параметрам коронарографии и сканирования)

		Predicted	
		0	1
Actual	0	12	2
	1	5	86
Точность		**0.933**	

- При расслоении данных по симптому точность классификаии в обеих подгруппах возрастает.

Примеры наоболее значимых для улучшения классификации симптомов

Таблица: Точность классификации в подпуляциях[2]

i	**Симптомы**	$s_i = 0$	$s_i = 1$
1	$a(b+c)$	0.8928	0.9333
2	$b+c$	0.8898	0.9285
3	$a+c$	0.8774	1.0000
4	ab	0.8805	0.9010
5	$c(a+b)$	0.8880	0.9000
6	$abc+bc+a$	0.8618	0.9489

- Из 127 построенных симптомов, примерно 50% показали прирост в точности классификации одной из подгрупп после раслоения. Симптомы, при расслоении по которым точность классификации увеличилась в обеих подгруппах приведены в таблице.

[2] Здесь, a — Количество сосудов выявленных при коронарографии, b — Бессимптомная боль, c — холодные пятна при талиевом сканировании сердца

Классификация по неполным данным

Данные по восстановления после туберкулеза содержат в себе 109 индивидов и 25 признаков. Прогнозируемая переменная — динамика восстановления после четырех месяцев (1 — хорошая, 2 — плохая).

- Метод работы с неполными данными — частичная классификация.
- Используемый классификатор — нейронная сеть.
- Слои в нейронных сетях: входной, скрытые, выходной. Количество нейронов на входном и скрытом слоях зависит от парамметров частичной классификации.
- Функция активации — сигмоидная.
- Функция потерь — кросс-энтропия.
- Метод оптимизации — Adam.

Таблица: Матрица ошибок классификации

		Predicted	
		1	2
Actual	1	24	21
	2	6	58
Точность	**0.752**		

Основные трудности програмной реализации

- Большое количество пропусков, вследствиие чего — большое количество частичных классификаторов (> 12000).
- Затрачивается много времени на обучение моделей (в среднем 35 классификаторов в минуту).

Методы решения возникших проблем:

- Подбор оптимального количества скрытых слоев и числа нейронов в них.
- Подбор метода оптимизации модели.
- Отбор частичных классификаторов, по количеству индивидов в подвыборках и по эффективности классификации.
- Использование графического ускорителя для уменьшения времени затраченного на обучение модели

Используя вышеперечисленные методы, было отобрано 3000 классификаторов, а время их обучения сократилось до 5 классификаторов в минуту.

Расслоение по тривиальным симптомам в группе с низкой резитентностью

Чувствительность — доля положительных результатов, которые правильно идентифицированы как таковые.

Специфичность — доля отрицательных результатов, которые правильно идентифицированы как таковые.

Таблица: Показатели классификации при расслоении данных в группе с высокой чувствительностью к препаратам

Расслоение	Точность	Чувствительность	Специфичность
При низкой распр.	1.0	1.0	1.0
При высокой распр.	1.0	1.0	1.0
Обе подгруппы	0.96	1.0	0.9

- При расслоении по разной степени распространенности и высокой чувствительности к препаратам точность классификации достигает максимума.

Расслоение по тривиальным симптомам в группе с высокой резистентностью

Таблица: Показатели классификации при расслоении данных в группе с низкой чувствительностью к препаратам

Расслоение	Точность	Чувствительность	Специфичность
При низкой распр.	0.89	1.0	0.75
При высокой распр.	0.84	0	1.0
Обе подгруппы	0.72	0.81	1.0

- При расслоении данных в группе с низкой чувствительностью к медикаментам, точность классификации упала, по сравнению с группой с высокой чувствительностью.
- Показатели чувствительности и специфичности говорят о том, что такое расслоение не дает желаемого результата.

Улучшение классификации при расслоении по нетривиальным симптомам

Таблица: Точность расслоения по симптомам [3]

i	**Симптомы** s_i	$s_i = 0$	$s_i = 1$
1	$lc + rl + cr + lcr + c$	0.944	0.852
2	$cr + l + r$	0.900	0.904
3	$l + c$	0.864	0.895
4	$lc + rl + cr + lcr + c$	0.925	0.857
5	lc	0.864	0.895
6	$lc + cr + lcr + l + c + r$	0.867	0.865

- В ходе исследования были построены 127 симптомов $(s(l, c, r))$, по которым проводилось расслоение данных. Расслоение считалось успешным, если точность классификации **в обеих подпопуляциях** была больше 0.85. Такому условию соответствует $\approx 5\%$ построенных симптомов.

[3] Здесь, l — локализация, c — полость, r — чувствительность к препаратам

Сравнение факторов риска при расслоении по симптому lc

Таблица: Коэффициенты дискриминантной функции при расслоении по симптому $lc = 0$ (группа средней тяжести)

mmp1.L	0.249
mmp9.L	-0.984
timp.L	1.273
elas.L	-3.856
n.pal.L	1.991
n.seg	-0.026
soe.L	0.282
ob	-0.028
mg.L	2.952
pi.L	2.921

Высокий показатель металлопротеиназы 1 с низкой эластазой означает начало репаративных изменений, связанных с фиброзированием.

Таблица: Коэффициенты дискриминантной функции при расслоении по симптому $lc = 1$ (тяжелая группа)

mmp1.L	-0.201
mmp9.L	-0.909
timp.L	0.312
elas.L	-2.984
n.pal.L	2.984
n.seg	0.014
soe.L	-0.517
ob	-0.011
mg.L	-1.003
pi.L	8.376

Компенсирование эластолитической активности. Данный процесс у тяжелых больных дает хороший прогноз на восстановление.

Заключение

- Разработаны алгоритмы позволяющие улучшить точность классификации с помощью симптомо-синдромального анализа.
- В результате проведения симптомно-синдромального анализа выявлена неоднородность данных и определены факторы риска.
- Исследован метод частичной классификации для неполных данных. и разработаны соотвествующие алгоритмы.
- В результате симптомного анализа данных с применением нейронной сети, прирост в точности достиг 20% при условии большого количества пропусков в исходных данных.
- Разработаны комплексы программ на языках программирования Python и R, позволяющие облегчить работу с медицинскими данными, провести их анализ и интерпретировать результаты.