УДК 519.854.33

© 2010 г. **И.С. Масич**, канд. физ.-мат. наук
(Сибирский государственный аэрокосмический университет, Красноярск)

МОДЕЛЬ ЛОГИЧЕСКОГО АНАЛИЗА ДЛЯ ПРОГНОЗИРОВАНИЯ ОСЛОЖНЕНИЙ ИНФАРКТА МИОКАРДА

Исследуется метод классификации данных, основанный на поиске и использовании логических правил. Решающее правило классификации базируется на модели, получаемой в результате решения ряда задач комбинаторной оптимизации. Строится модель классификации для прогнозирования осложнений инфаркта миокарда.
Ключевые слова: инфаркт миокарда, классификация, комбинаторная оптимизация, прогнозирование, псевдобулевая функция.

Введение

Большое количество задач распознавания, привлекающих внимание исследователей как в медицине, так и во множестве других областей может быть сформулировано следующим образом. Имеется выборка данных, которая состоит из двух непересекающихся множеств Ω^+ и Ω^- n-мерных векторов. Каждый вектор соответствует некоторому пациенту, векторы множества Ω^+ соответствуют пациентам, находящимся в некотором медицинском состоянии (например, болен или имеет осложнение заболевания), а векторы Ω^- не соответствуют этому состоянию. Компоненты векторов (называемые признаками, переменными, характеристиками, или атрибутами) представляют собой результаты определенных измерений, тестов или просто показывают присутствие или отсутствие определенных симптомов. Эти компоненты могут быть численными, номинальными или бинарными.

Задача состоит в том, чтобы на основании имеющейся выборки данных (классифицированных ранее наблюдений) извлечь информацию о состоянии «нового» пациента, наблюдение которого не содержится в выборке. Главная цель решения таких задач – на основе анализа данных и вычислительных систем диагностики и прогнозирования определить индивидуальную терапию для пациента.

Для решения этой задачи исследуется метод анализа данных, в основе которого лежит принцип вывода логических закономерностей или правил. Каждое правило должно покрывать достаточно много объектов одного класса и практически не покрывать объекты другого класса. Взяв вместе некоторое количество правил, можно получить алгоритм (модель, решающее правило), который будет

решать поставленную задачу классификации.

Построение эффективных правил и модели классификации является сложной комбинаторной задачей. Результаты ее решения определяются видом сформированных критериев и ограничений, а также используемыми алгоритмами оптимизации.

Главной особенностью использования такого подхода является то, что в результате работы метода из базы данных извлекаются правила, с помощью которых можно классифицировать объекты и без помощи компьютера и вычислительной системы.

Осложнения инфаркта миокарда

Инфаркт миокарда (ИМ) – распространенное и грозное заболевание. Бурное распространение этого заболевания за последние полвека сделало его одной из наиболее острых проблем современной медицины. Заболеваемость инфарктом миокарда (ИМ) остается высокой во всех странах. Особенно это касается городского населения высокоразвитых стран, испытывающего стремительный ритм современной жизни и подвергающегося хроническому воздействию стрессовых факторов, нерегулярного и не всегда сбалансированного питания.

Несмотря на то, что внедрение современных лечебно-профилактических мероприятий несколько снизило смертность от инфарктов, она продолжает оставаться довольно высокой. Около 15-20% больных острым ИМ погибают на догоспитальном этапе, еще 15% – в больнице [1], т.е. общая летальность при остром ИМ 30-35%. Настораживает высокая смертность в госпитальный период (т.е. во время нахождения больного в клинике), которая, по данным различных российских авторов, составляет от 10 до 20%.

Течение заболевания у пациентов с ИМ различно. ИМ может протекать без осложнений или с осложнениями, не ухудшающими долгосрочный прогноз. В то же время около половины пациентов в острый и подострый периоды имеют осложнения, приводящие к ухудшению течения заболевания и даже летальному исходу. Предвидеть развитие этих осложнений не всегда может даже опытный специалист. В связи с этим прогнозирование осложнений ИМ с целью своевременного проведения необходимых профилактических мероприятий представляется актуальной задачей.

В исследование включены 1700 больных острым инфарктом миокарда, проходивших лечение в отделении реанимации и интенсивной терапии и в кардиологическом отделении городской клинической больницы [2]. Информация об анамнезе пациентов и течении ИМ получена из историй болезни и сконцентрирована в 117 полях электронной таблицы (базы данных). База данных содержит информацию о возрасте, поле пациента, локализации и глубине ИМ, данных анамнеза, изменениях ЭКГ, количестве калия, натрия, некоторых ферментов крови, особенностях клиники в первые часы заболевания.

За период пребывания больных в стационаре фибрилляция предсердий (ФП) наблюдалась у 170 (10,0%) больных, фибрилляция желудочков (ФЖ) у 71 (4,2%), отек легких (ОЛ) у 159 (9,4%), разрыв сердца (РС) у 54 (3,2%), летальный

исход (ЛИ) у 271 (15,9%). Данные осложнения и ЛИ были выбраны для прогнозирования.

Ранее для прогнозирования коллективом ученых [2] была создана экспертная система на основе искусственных нейронных сетей. В результате настройки алгоритма обучения и параметров сети на данной выборке была достигнута точность прогнозирования 70-90%.

Логический анализ данных

В основе предлагаемого подхода к классификации данных лежит метод, происходящий из теории комбинаторной оптимизации и называемый логическим анализом данных (Logical Analysis of Data – LAD) [3]. Этот метод успешно использовался для решения ряда задач из различных областей [4 – 7]. Основная идея метода заключается в совместном использовании действий по «дифференцированию» и «интегрированию», производимых на области пространства исходных признаков, содержащей заданные *позитивные* и *негативные* наблюдения. На шаге «дифференцирования» определяется семейство малых подмножеств, обладающих характерными позитивными и негативными чертами. На шаге «интегрирования» формируемые определенным образом объединения этих подмножеств рассматриваются как аппроксимации областей пространства признаков, содержащих позитивные и, соответственно, негативные наблюдения.

Ниже приведены последовательные элементы метода [6]:

а) для исключения избыточных переменных в исходной выборке данных в множестве переменных определяется некоторое подмножество S, используя которое можно отличать позитивные наблюдения от негативных. Далее для работы метода используются проекции Ω_S^+ и Ω_S^- множеств Ω^+ и Ω^- на S. Такая процедура используется во многих методах классификации и анализа данных. Особенностью осуществления ее в LAD является то, что происходит выделение не только значимых по отдельности признаков, но и определение комбинаций признаков, которые оказывают коллективное влияние на результат;

б) множество Ω_S^+ покрывается семейством однотипных подмножеств уменьшенного пространства, каждое из которых имеет значительное пересечение с Ω_S^+, но не пересекается с Ω_S^-. Такие подмножества называются «позитивными паттернами». Аналогично множество Ω_S^- покрывается «негативными паттернами»;

в) определяется подмножество позитивных паттернов, объединение которых покрывает все наблюдения Ω_S^+, и подмножество негативных паттернов, объединение которых покрывает все наблюдения Ω_S^-. Совокупность этих двух подмножеств называется «моделью»;

г) позитивный или негативный характер некоторого наблюдения, покрываемого объединением двух подмножеств модели, определяется с помощью решающего правила, основанного на этих подмножествах.

Отличительной особенностью предлагаемого метода является то, что вме-

сто того, чтобы просто ответить на вопрос, к какому из классов принадлежит новое наблюдение, он строит аппроксимацию областей пространства признаков, содержащую наблюдения соответствующих классов. Наиболее важные преимущества такого подхода – это возможность дать объяснение для любого решения, полученного методом, возможность выявления новых классов наблюдений, возможность анализа роли и природы признаков. Более подробно особенности и преимущества LAD описаны в [6].

Бинаризация

Рассматриваемый метод предназначен для работы с выборками данных, в которых признаки принимают бинарные значения. Так как исходная выборка может состоять из метрических переменных, необходимо воспользоваться *методом бинаризации*. Существует несколько способов кодирования, описанных в [8].

Один из простейших способов бинаризации состоит в следующем. Каждой метрической переменной ставится в соответствие несколько бинарных переменных. Бинарная переменная принимает значение 1, если значение соответствующей метрической переменной принимает значение выше определенного порога, и наоборот. Такой способ в [8] называется «единичным». Его недостатком является то, что существует большое число комбинаций бинарных переменных, которым не соответствуют точки в исходном пространстве ($2^n - n - 1$). Этот недостаток затрудняет использование такого способа для кодирования переменных критериальной функции при решении задач оптимизации, так как большое число решений будет недопустимым. Но в данном случае для классификации это не имеет значения, так как бинарные переменные получаются путем кодирования заданных метрических переменных. Основным же преимуществом такого способа является соответствие расстояний в исходном и бинарном пространствах. Это значит, что точки, близкие в исходном пространстве, являются близкими и в бинаризованном. А это, в свою очередь, позволяет еще в процессе бинаризации минимизировать число порогов (и, соответственно число бинарных переменных), ставя в соответствие близким значениям исходной переменной одно и то же значение в бинарном пространстве (при условии, что позитивные и негативные множества наблюдений останутся непересекающимися).

Модель классификации

Поиск паттернов.

В основе рассматриваемого подхода лежит понятие паттерна. *Позитивным паттерном* называется подкуб пространства булевых переменных B_2^t, который пересекается с множеством Ω_S^+ и не имеет общих элементов с множеством Ω_S^-. Негативный паттерн задается аналогично.

Позитивный щ-паттерн для $\omega \in \{0,1\}^t$ – это паттерн, содержащий в себе точку *щ*. Для каждой точки $\omega \in \Omega_S^+$ найдем *максимальный щ-паттерн*, т.е. по-

крывающий наибольшее число точек Ω_S^+.

Соответствующий подкуб зададим с помощью переменных y_j:

$$y_j = \begin{cases} 1, \text{ если } x_i \text{ зафиксирована в подкубе,} \\ 0, \text{ в противном случае.} \end{cases}$$

То есть путем фиксирования l переменных исходного куба размерностью t получаем подкуб размерностью (t-l) и с числом точек 2^{t-l}.

Условие, говорящее о том, что позитивный паттерн не должен содержать ни одной точки Ω_S^-, требует, чтобы для каждого наблюдения $\rho \in \Omega_S^-$ переменная y_j принимала значение 1 по меньшей мере для одного j, для которых $\rho_j \neq \omega_j$:

$$\sum_{\substack{j=1 \\ \rho_j \neq \omega_j}}^{t} y_j \geq 1 \text{ для любого } \rho \in \Omega_S^-.$$

Усиление ограничения для повышения устойчивости к ошибкам производится путем замены числа 1 в правой части неравенства на целое положительное число d.

С другой стороны, позитивное наблюдение $\sigma \in \Omega_S^+$ будет тогда входить в рассматриваемый подкуб, когда переменная y_j принимает значение 0 для всех индексов j, для которых $\sigma_j \neq \omega_j$. Таким образом, число позитивных наблюдений, покрываемых *ц*-паттерном, может быть вычислено как

$$\sum_{\sigma \in \Omega_S^+} \prod_{\substack{j=1 \\ \sigma_j \neq \omega_j}}^{t} (1 - y_j).$$

Таким образом, имеем задачу условной псевдобулевой оптимизации с алгоритмически заданными функциями

$$\sum_{\sigma \in \Omega_S^+} \prod_{\substack{j=1 \\ \sigma_j \neq \omega_j}}^{t} (1 - y_j) \to \max, \qquad (1)$$

$$\sum_{\substack{j=1 \\ \rho_j \neq \omega_j}}^{t} y_j \geq d \text{ для любого } \rho \in \Omega_S^-,\ y \in \{0,1\}^t. \qquad (2)$$

Целевая функция задачи является унимодальной монотонной псевдобулевой функцией [9], т.е. имеет единственный безусловный максимум, находящийся в точке $y^0 = (0,0,...,0)$, и убывает при удалении от точки максимума (при смене любой компоненты с 0 на 1). Причем функция эта задана алгоритмически, так как для ее вычисления необходимо перебрать все наблюдения выборки Ω_S^+. Функция ограничения является также унимодальной и монотонной псевдобулевой функцией, заданной алгоритмически.

Аналогично формулируется задача нахождения максимальных отрицательных паттернов.

Решающее правило.

В итоге получаем семейство максимальных паттернов, число которых ограничено мощностью выборки данных $\left|\Omega^+ \cup \Omega^-\right|$. Обозначим $M_1^+,...,M_p^+$ и $M_1^-,...,M_q^-$ соответственно множества позитивных и негативных паттернов.

Чтобы классифицировать новое наблюдение, воспользуемся следующим решающим правилом:

1. Если наблюдение удовлетворяет условиям одного или нескольких позитивных паттернов и не удовлетворяет условиям ни одного из негативных, то оно классифицируется как позитивное.

2. Если наблюдение удовлетворяет условиям одного или нескольких негативных паттернов и не удовлетворяет условиям ни одного из позитивных, то оно классифицируется как негативное.

3. Если наблюдение удовлетворяет условиям p' из p позитивных паттернов и q' из q негативных, то «знак» наблюдения определяется как $p'/p - q'/q$.

4. В случае, если наблюдение не удовлетворяет условиям ни одного паттерна, позитивного или негативного, то оно остается неклассифицированным.

В случае большого объема выборки данных число различных паттернов может быть велико и решается задача определения *модели*, состоящей из некоторого числа паттернов – так, чтобы она была способна классифицировать те же наблюдения, которые можно классифицировать с помощью полной системы паттернов [10].

Устойчивость к выбросам.

Найденный паттерн характеризуется *покрытием* – числом объектов определенного класса, которые он захватывает, и *степенью* – количеством фиксированных переменных, которые определяют этот паттерн. Согласно приведенной выше оптимизационной модели (1), (2) паттерн не покрывает ни одного объекта другого класса (из обучающей выборки).

Наибольшую ценность представляют паттерны, которые имеют наибольшее покрытие. Чем больше покрытие, тем лучше паттерн отображает образ класса.

Степень паттерна обычно не должна быть слишком большой. Как видно из модели, при уменьшении степени происходит увеличение покрытия, но уменьшать степень можно лишь до тех пор, пока не происходит захвата ни одного объекта другого класса.

Специфика описанной выше задачи прогнозирования состоит в том, что база данных имеет большое число неизмеренных значений (пропущенных данных), а сделанные измерения могут быть неточны либо ошибочны. Шумы и выбросы приводят к тому, что объекты различных классов «накладываются» друг на друга, попадая в «область» противоположного класса. В результате вычисляемые паттерны получаются с большей степенью и с существенно меньшим покрытием, чем если бы выбросов и неточностей не было, а итоговая модель состоит из большого числа маленьких паттернов (с малым покрытием). Это не позволяет построить эффективную модель классификации с «хорошо интерпретируемыми» правилами (в которых участвует небольшое число признаков) и с высокой точно-

стью распознавания.

Для повышения устойчивости метода к выбросам следует ослабить ограничение (2) – сделать возможным, чтобы паттерн захватывал некоторое малое число объектов другого класса. Тогда степень вычисляемых паттернов уменьшится, а покрытие увеличится.

Ограничение оптимизационной модели будет выглядеть следующим образом.

$$\sum_{\rho\in\Omega_S^-} z_\rho \le D, \text{ где } z_\rho = \begin{cases} 0, \text{если} \sum\limits_{j=1}^{t} y_j \ge d, \\ \rho_j \ne \omega_j \\ 1, \text{в противном случае,} \end{cases} \qquad (3)$$

где D – число объектов другого класса, которые допускаются быть покрытыми паттерном (целое неотрицательное число).

Функции (1)-(3) построенной модели оптимизации задаются алгоритмически, т.е. вычисляются через определенную последовательность операций. Для решения задачи оптимизации использовались алгоритмы оптимизации, основанные на поиске граничных точек допустимой области [11 – 15]. Эти алгоритмы были разработаны специально для этого класса задач и основаны на поведении монотонных функций модели оптимизации в пространстве булевых переменных. Алгоритмы поиска граничных точек являются поисковыми, т.е. не требуют задания функций в явном виде, с помощью алгебраических выражений, а используют вычисления функций в точках.

Результаты

Результаты сравнительных испытаний метода рассмотрим на примере прогнозирования осложнения инфаркта миокарда – фибрилляции предсердий.

Для проведения испытаний использовалась выборка данных, состоящая из 164 пациентов с осложнением (позитивные объекты) и 193 объектов без указанного осложнения (негативные объекты). Десятая часть из них (16 и 20 пациентов соответственно) использовалась для контроля и в построении решающей модели не участвовала.

В результате бинаризации из 113 признаков различного типа (бинарных, номинальных, численных) было получено 207 бинарных признаков.

В табл. 1 представлены результаты испытаний с использованием двух оптимизационных моделей: с ограничением (2), исключающим захват паттерном объекта другого класса, и с ограничением (3), позволяющим покрытие паттерном нескольких объектов (D=10) другого класса. В таблице приведены средние значения покрытий и степени для набора паттернов.

Точность классификации определяется двумя значениями: чувствительностью – точностью определения пациентов с осложнением, и специфичностью – точностью определения пациентов без осложнений. На практике к чувствительности предъявляются большие требования, чем к специфичности; полученные результаты классификации в должной мере соответствуют этому.

Таблица 1

Задача оптимизации	Множество паттернов	Покрытие негативных объектов	Покрытие позитивных объектов	Степень паттерна	Точность классификации, %
Целевая функция (1), ограничение (2)	негативные паттерны	12	0	6	**55**
	позитивные паттерны	0	9	6	**69**
Целевая функция (1), ограничение (3)	негативные паттерны	35	10	6	**60**
	позитивные паттерны	10	25	6	**88**

Если необходимо, чувствительность можно повысить за счет специфичности. Для этого нужно внести веса в решающее правило (которое описано выше): $\alpha \cdot p'/p - \beta \cdot q'/q$.

Примеры правил, из которых строится решающая модель, приведены в табл. 2 (примечание: в таблице приведены лишь некоторые признаки – те, которые используются в этих правилах).

Таблица 2

Паттерны	Признаки											
	AGE	SEX	STENOK_AN	DLIT_AG	S_AD_ORIT	ANT_IM	INF_IM	N_R_ECG_P	N_R_ECG_P_4	ROE	TIME_B_S	NOT_NA_2_N
Негатив-				≥4	≥140		≥1					
	<75	1				<1	≥1		0			
				<6				0	0	<13		
	<75		≥2							≥22		
		1					<3	0				
Позитивные			≥3								≥3	<1
				≥4	≥110		<1					
			≥3			≥3	<3					
	≥62				<160		<1				<5	
				≥6			<1					<1

Заключение

Таким образом, модификация условий при поиске правил позволяет находить паттерны с более высоким покрытием, из которых строится более точная модель распознавания. Применение такого подхода необходимо при решении задач с наличием выбросов и шумов и с большим количеством пропусков в выборке данных.

Задача прогнозирования осложнений инфаркта миокарда решена с точностью, сопоставимой с точностью решения посредством искусственных нейронных сетей. При этом логический анализ данных дает ряд преимуществ при практическом использовании. Прежде всего в явном виде известны правила, по кото-

рым принимается решение о принадлежности к какому-либо классу. Кроме того, при применении модели классификации к новому пациенту по тому, каким числом паттернов покрываются его данные, можно судить о вероятности возможной ошибки при распознавании.

ЛИТЕРАТУРА

1. *Кардиология* в таблицах и схемах / под ред. М.Фрида, С.Грайнс, пер. с англ. – М.: Практика, 1996.
2. *Осложнения* инфаркта миокарда: база данных для апробации систем распознавания и прогноза / А.Н. Горбань, В.А. Шульман, Д.А. Россиев и др. – Красноярск, Вычислительный центр СО РАН: Препринт №6, 1997.
3. *Hammer P.L.* The Logic of Cause-effect Relationships // Lecture at the International Conference on Multi-Attribute Decision via Operations Research-based Expert Systems. – Passau, Germany, 1986.
4. *Coronary* Risk Prediction by Logical Analysis of Data // S. Alexe, E. Blackstone, P.L. Hammer and others / Annals of Operations Research. – 2003. – 119. – P.15-42.
5. *An Implementaion* of Logical Analysis of Data // E. Boros, P.L. Hammer, T. Ibaraki and others / IEEE Transactions on Knowledge and Data Engineering. – 2000. – 12(2). – P. 292-306.
6. *Hammer P.L., Bonates T.* Logical Analysis of Data: From Combinatorial Optimization to Medical Applications – RUTCOR Research Report 10-2005, 2005.
7. *Hammer P.L., Kogan A., Lejeune M.* Modeling Country Risk Ratings Using Partial Orders – RUTCOR Research Report 24-2004, 2004.
8. *Растригин Л.А., Фрейманис Э.Э.* Решение задач разношкальной оптимизации методами случайного поиска // Проблемы случайного поиска. – 1988. – Вып. 11. – С. 9-25.
9. *Antamoshkin A.N., Masich I.S.* Identification of pseudo-Boolean function properties // Engineering & automation problems (Проблемы машиностроения и автоматизации). – 2007. – № 2. – P. 66-69.
10. *Масич И.С.* Комбинаторная оптимизация в задаче классификации // Системы управления и информационные технологии. – 2009 – № 1.2(35). – С. 283-288.
11. *Антамошкин А.Н., Масич И.С.* Неулучшаемый алгоритм условной оптимизации монотонных псевдобулевых функций // Электронный журнал "Исследовано в России". – 2004. – № 64. – С. 703-708. http://zhurnal.ape.relarn.ru/articles/2004/064.pdf.
12. *Антамошкин А.Н., Масич И.С.* Гриди алгоритмы и локальный поиск для условной псевдобулевой оптимизации // Электронный журнал "Исследовано в России". – 2003. – № 177. – С. 2143-2149. http://zhurnal.ape.relarn.ru/articles/2003/177.pdf.
13. *Масич И.С.* Приближенные алгоритмы поиска граничных точек для задачи условной псевдобулевой оптимизации // Вестник СибГАУ. – 2006. – 8470, 1(8). – С. 39-43.
14. *Antamoshkin A.N., Masich I.S.* Heuristic search algorithms for monotone pseudo-boolean function conditional optimization // Engineering & automation problems (Проблемы машиностроения и автоматизации). – 2006. – V. 5, N. 1. – P. 55-61.
15. *Antamoshkin A.N., Masich I.S.* Pseudo-Boolean optimization in case of unconnected feasible sets // Models and Algorithms for Global Optimization. Series: Springer Optimization and Its Applications, Vol. 4, edited by A. Tцrn, J. Ћilinskas. – Springer, 2007, XVI. – P. 111-122.

Статья представлена к публикации членом редколлегии Ю.М. Перельманом.

E-mail:
Масич И.С. – i-masich@yandex.ru