УДК 004.891

УЧЕТ ОСОБЕННОСТЕЙ ПРАКТИЧЕСКИХ ЗАДАЧ ПУТЕМ НАСТРОЙКИ ПАРАМЕТРОВ МЕТОДА ЛОГИЧЕСКОГО АНАЛИЗА ДАННЫХ

Кузьмич Р.И.

ФГАОУ ВО «Сибирский федеральный университет», Красноярск, e-mail: romazmich@gmail.com

Ключевым требованием к современным методам анализа данных является возможность учета особенностей практических задач. При реализации такого преимущества происходит настройка параметров метода на конкретную задачу. Чем больше параметров метода нуждается в настройке, тем больше возможное число его реализаций. Однако в таком случае становится сложнее настроить такой метод под конкретную задачу. Также следует отметить, что корректная настройка параметров метода позволяет найти компромисс между критериями, которые устанавливает заказчик к результатам работы метода. В работе рассматривается метод логического анализа данных, который позволяет производить настройку под конкретную задачу на этапах построения опорного множества признаков, формирования правил, построения классификатора. В качестве примера осуществляется настройка параметров метода при решении задачи управления приземлением космического корабля. Специфическими особенностями данной задачи являются пропущенные значения признаков и малое число наблюдений в выборке, что, бесспорно, осложняет процесс принятия решения о принадлежности к определенному классу. Также следует указать, что в качестве способа тестирования применяется кросс-проверка, а не процентное разделение. Выбор способа тестирования продиктован малым числом наблюдений в выборке.

Ключевые слова: правило, классификатор, настройка, параметр, кросс-проверка

ACCOUNTING THE FEATURES OF PRACTICAL TASKS BY SETTING PARAMETERS OF THE METHOD OF LOGICAL ANALYSIS OF DATA

Kuzmich R.I.

Siberian Federal University, Krasnoyarsk, e-mail: romazmich@gmail.com

A key requirement for modern data analysis methods is the ability to take into account the specifics of practical tasks. When implementing this advantage, the method parameters are set for a specific task. The more method parameters need setting, the greater the possible number of its implementations. However, in this case, it becomes more difficult to set such a method for a specific task. It should be noted that the correct setting of the method parameters allows finding a compromise between the criteria set by the customer to the results of the method. The paper discusses the method of logical analysis of data, which allows setting for a specific task at the stages of building a support set of features, formation patterns, and building a classifier. As an example, the method parameters are set when solving the task of shuttle landing controlling. Specific features of this task are the missing values of features and a small number of observations in the sample, which undoubtedly complicates the process of making decisions about belonging to particular class. It should be shown that cross-validation is applied as a test method, rather than a percentage split. The choice of testing method is dictated by a small number of observations in the sample.

Keywords: pattern, classifier, setting, parameter, cross-validation

В работе речь пойдет о методе логического анализа данных, принадлежащего к логическим алгоритмам классификации, основная идея работы которых состоит в том, что они выделяют правила из исходных данных. Ранее метод удачно применялся для решения ряда практических задач в различных сферах [1–3]. Суть метода заключается в последовательном выполнении двух операций на области пространства исходных признаков, в которой находятся положительные и отрицательные наблюдения. Первая операция (формирование правил) заключается в формировании семейства малых подмножеств, имеющих типичные положительные и отрицательные особенности. Вторая операция (построение классификатора) заключается в объединении подмножеств, полученных на предыдущем этапе [4].

Отметим, что рассматриваемый метод является достаточно гибким инструментом, который позволяет учитывать предложения заказчика и особенности решаемой практической задачи. Для реализации таких преимуществ на этапах построения опорного множества (множества признаков, позволяющего отделить положительные наблюдения от отрицательных с высокой точностью), формирования правил, построения классификатора имеются параметры метода, которые путем целенаправленной настройки позволяют соблюдать баланс между точностью и трудоемкостью построения правил, распознающей и обобщающей способностями классификатора, интерпретируемостью классификатора и точностью классификации.

Цель исследования: обоснование возможности учета особенностей практических задач методом логического анализа данных путем целенаправленной настройки его параметров.

Ниже приведено описание самих параметров и особенности их настройки на каждом из перечисленных этапов метода [5].

Материалы и методы исследования

Одним из параметров на этапе построения опорного множества является минимальное число признаков, по которым различаются наблюдения разных классов. Изменяя значение параметра, получаем разные пулы признаков для формирования правил. При настройке необходимо соблюдать баланс между точностью классификации и трудоемкостью построения правил. При большом значении данного параметра получаем небольшой пул признаков, сокращая трудоемкость построения правил, поскольку пространство поиска невелико. Однако такая ситуация может привести к неудаче при построении классификатора, который мог бы корректно классифицировать наблюдения тестовой выборки.

Также отметим, что на этапе отбора признаков применяется несколько критериев для оценки каждого из признаков в наборе с целью создания пула выбранных признаков. По каждому рассматриваемому критерию выбираем с целью дальнейшего использования k первых по рангу признаков. Пул состоит из набора этих признаков, которые ранжируются в числе первых k признаков, согласно примененным критериям. Ряд критериев, применяемых для определения значимости признака при разделении положительных и отрицательных наблюдений, описан в [6].

Основанное на наборе правил ранжирование признаков по их значимости служит в качестве основного принципа для итеративной последовательности шагов предлагаемого алгоритма формирования пула признаков. Начиная с исходного пула определяем каждый шаг процедуры формирования нового пула, состоящего приблизительно из половины признаков, имеющих наибольший ранг, в соответствии с ранжированием, базирующимся на наборе правил.

Точность построенного на новом пуле классификатора оценивается на тестовом множестве:

1. Если находится допустимое качество, т.е. если точность приблизительно равна родительской точности, новый пул заменяет предыдущий пул и процесс продолжается.

2. Если точность снижается относительно точности родительского пула, новый пул не является непригодным автоматически. Необходимо проверить его на модифицированном наборе правил. Модификация набора правил с целью его наилучшего представления может быть достигнута варьированием (обычно увеличением) параметров покрытия и степени правила.

3. Если набор правил, классификационная мощность которого имеет допустимую точность, найден, тогда новый пул принят и процесс продолжается.

4. Если не определено наилучшего представления набора правил, тогда новый пул улучшается включением 75% признаков из родительского пула (87,5 или 93,75%).

5. Если этот процесс не создает пул, который определяет набор правил с допустимым качеством, тогда процесс останавливается на родительском пуле как окончательном.

В некоторых случаях требуется, чтобы обеспечивалась робастность, т.е. количество признаков в конечном пуле не должно быть ниже установленного порога. В этих случаях процесс останавливается прежде, чем количество признаков получится слишком маленьким.

В случае наличия выбросов и пропусков значений признаков в выборке эффективно использовать частичные правила, т.е. правила, которые могут покрывать небольшое число наблюдений другого класса. В качестве параметра метода, позволяющего реализовать формирование частичных правил, выступает количество наблюдений другого класса, которое может захватить формируемое правило. При настройке необходимо соблюдать баланс между распознающей и обобщающей способностями классификатора. Отметим, что при низком значении параметра происходит эффект переобучения, т.е. распознающая способность классификатора выше чем обобщающая. Регулируя параметр в сторону его увеличения, необходимо их уравновесить.

В работе [7] исследована и апробирована оптимизационная модель для формирования правил, выделяющих существенно различные подмножества наблюдений выборки. В качестве настраиваемого параметра в данной оптимизационной модели выступает максимальное количество правил, покрывающих каждое наблюдение обучающей выборки в классификаторе. Данный параметр позволяет регулировать количество правил в классификаторе, соблюдая баланс между интерпретируемостью классификатора и точностью классификации. Новый классификатор работает аналогично построенному на базе оптимизационной модели с максимальным покрытием, если значение параметра равно максимальному количеству правил для данного класса. При стремлении значения параметра к 1 в новом классификаторе становится недостаточно правил, чтобы корректно классифицировать вновь поступающие наблюдения, т.е. происходит снижение его обобщающей способности. Эмпирическим путем установлено, что значение параметра надо выбирать в диапазоне от 5 и до значения среднего покрытия правил, построенных с использованием оптимизационной модели с максимальным покрытием, причем чем ниже значение параметра, тем меньше количество правил в классификаторе, что, бесспорно, приводит к росту его интерпретируемости.

В работе [5] приведено описание алгоритмической процедуры выбора базовых наблюдений для формирования правил. Настраиваемым параметром в этой процедуре является количество центроидов для каждого класса, получаемых с помощью применения метода «k-средних» к множеству наблюдений обучающей выборки. При настройке параметра необходимо соблюдать баланс между точностью классификатора и трудоемкостью его построения. Количество правил в классификаторе равно числу полученных центроидов. Таким образом, чем меньше центроидов, тем меньше трудоемкость построения классификатора. С другой стороны, при недостаточном количестве правил в классификаторе точность классификации снижается из-за увеличения числа отказов от классификации. Поэтому, варьируя значение параметра, необходимо следить за изменением числа неклассифицированных наблюдений тестовой выборки.

В работе [8] описана алгоритмическая процедура построения классификатора из набора информативных правил. В качестве параметра в данной процедуре выступает порог информативности. Он позволяет регулировать количество правил в классификаторе, соблюдая баланс между интерпретируемостью классификатора и точностью классификации.

При постепенном увеличении порога информативности интерпретируемость классификатора возрастает, поскольку уменьшается количество правил в нем, но, начиная с определенного значения порога информативности, происходит рост отказов от классификации, следовательно, снижение точности классификации в целом. Причиной увеличения отказов является исключение всех правил, которые ранее покрывали определенные наблюдения тестовой выборки, т.е. появление непокрытых наблюдений при тесте.

Результаты исследования и их обсуждение

Осуществим настройку параметров метода при решении задачи управления приземлением космического корабля [9, 10]. Отметим, что объем выборки для этой задачи равен 15. В табл. 1 приведена выборка для данной задачи, содержащая 6 наблюдений класса с ручным управлением кораблем (0 класс) и 9 наблюдений класса с автоматической посадкой корабля (1 класс). Каждое наблюдение в выборке характеризуется семью признаками: stability, error, sign, wind, magnitude, visibility, class. Как видно, в выборке присутствуют пропущенные значения признаков, которые в табл. 1 обозначены «*».

Пропущенные значения влияют на различные стадии алгоритма. Для начала, если значение исходного признака пропущено, все значения всех бинарных переменных рассматриваются как пропущенные. На стадии построения опорного множества недоступность значений переменной предотвращает использование данной переменной для отличия соответствующего наблюдения от других. Следовательно, в задаче о множественном покрытии ограничение, соответствующее паре, состоящей из наблюдений положительного и отрицательного класса, должно приводить только к тем переменным, чьи значения известны для обоих наблюдений в паре (т.е. коэффициенты для наблюдений с пропущенными значениями должны быть равны 0).

Задача заключается в следующем: необходимо на базе исходной выборки данных извлечь правила для классификации новых наблюдений.

Особенностью настройки метода для задачи управления приземлением космического корабля является выбор способа тестирования. Как правило, для задач классификации применяется процентное разделение – способ тестирования, при котором вся выборка делится на обучающую и тестовую выборки. Но поскольку выборка наблюдений состоит всего из 15 наблюдений, то в качестве способа тестирования в данном случае применяется кросс-проверка.

K-областной метод статистики является одним из методов кросс-проверки. Суть метода заключается в случайном разбиении выборки на k примерно равных подмножеств. При этом классификатор строится на k-1 подмножествах, а потом тестируется на k-м. Так происходит k раз, при этом всегда выбирается новое тестовое подмножество. Мерой качества описанного метода является средняя точность, полученная как среднеарифметическое всех испытаний.

Если число k равно количеству наблюдений в выборке, при этом тестовое множество содержит только одно наблюдение, тогда k-областной метод статистики называется методом поочередного пропуска [11].

Таблица 1

Исходная выборка для задачи управления приземлением космического корабля

stability	error	sign	wind	magnitude	visibility	class
*	*	*	*	*	1	1
1	*	*	*	*	0	0
0	2	*	*	*	0	0
0	1	*	*	*	0	0
0	3	1	1	*	0	0
*	*	*	*	4	0	0
0	4	*	*	1	0	1
0	4	*	*	2	0	1
0	4	*	*	3	0	1
0	3	0	0	1	0	1
0	3	0	0	2	0	1
0	3	0	1	1	0	1
0	3	0	1	2	0	1
0	3	0	0	3	0	0
0	3	0	1	3	0	1

Таблица 2

Примеры правил для задачи управления приземлением космического корабля

stability	error	sign	wind	magnitude	visibility	class
1						0
	<3					0
		1				0
					1	1
				<4		1
	≥3					1

Для формирования правил использовалась модифицированная оптимизационная модель, суть работы которой состоит в том, что правила покрывают небольшое число наблюдений другого класса. Пропущенными значениями признаков в выборке обусловлено применение такой оптимизационной модели. Следует указать успешное использование поисковых алгоритмов оптимизации для решения задач оптимизации, связанных с построением опорного множества и формированием правил. Отличительной чертой таких алгоритмов является применение вычисления функций в точках [12].

Примеры правил, из которых состоит классификатор для метода, приведены в табл. 2. Правила получены с использованием программного приложения, реализованного автором [13].

Согласно полученным результатам, точность классификации составила 80%, т.е. 12 из 15 наблюдений классифицированы правильно. Отметим, что каждое построенное правило состоит из одной переменной, т.е. является легко интерпретируемым. Полученные правила позволяют наглядно обосновать принадлежность данного наблюдения тому или иному классу.

Для сравнения результатов предлагаемого метода по точности данная задача решена в системе анализа данных WEKA с помощью алгоритмов C4.5 [14], RIPPER [14], Adaboost [15]. Количество правильно классифицированных наблюдений для указанных алгоритмов: C4.5 – 9, RIPPER – 9, Adaboost – 11. Таким образом, предлагаемый автором метод в целом показал наилучший результат по точности классификации, кроме того, он характеризуется возможностью учета особенностей практических задач.

Заключение

В работе детально описаны возможности учета особенностей практических задач предложенным методом путем целенаправленной настройки его параметров. На примере задачи управления приземлением космического корабля, специфическим особенностями которой являются пропущенные значения признаков и малое число наблюдений в выборке, приведено эмпирическое подтверждение возможности корректной настройки параметров метода.

Следует подчеркнуть, что особенностями рассматриваемого метода являются обоснование и наглядность принимаемых им решений, возможность выявления новых классов наблюдений, возможность определения важности признака при построении классификатора. Также метод осуществляет всестороннее исследование полного набора признаков, сфокусированного на классификационной мощности комбинации признаков (не ограничивая внимания только индивидуальными признаками), и имеет возможность извлечения новой информации о роли индивидуальных признаков и комбинации признаков через анализ их всесторонних перечислений.

Список литературы

1. Boros E., Hammer P.L., Kogan A., Crama Y., Ibaraki T., Makino K. Logical analysis of data: classification with justification. RUTCOR Technical Report. 2009. no. 5. P. 1–33.

2. Shaban Y., Yacout S., Balazinski M., Jemielniak K. Cutting tool wear detection using multiclass logical analysis of data. Machining Science and Technology. 2017. no. 21 (4). P. 526–541.

3. Herrera J.F.A., Subasi M.M. Logical analysis of multiclass data. RUTCOR Technical Report. 2013. no. 5. P. 1–23.

4. Brauner M.W., Brauner D., Hammer P.L., Lozina I., Valeyre D. Logical analysis of computer tomography data to differentiate entities of idiopathic interstitial pneumonias. RUTCOR Research Report. 2004. no. 30. P. 1–15.

5. Кузьмич Р.И. Модифицированный метод логического анализа данных для задач классификации: дис. … канд. техн. наук. Красноярск, 2016. 136 с.

6. Alexe G., Alexe S., Hummer P.L., Vizvari B. Pattern-based feature selection in genomics and proteomics. RUTCOR Research Report. 2003. no. 7. P. 1–12.

7. Кузьмич Р.И., Масич И.С. Модификация целевой функции при построении паттернов для увеличения различности правил в модели классификации // Системы управления и информационные технологии. 2014. № 2 (56). С. 14–18.

8. Кузьмич Р.И., Масич И.С. Построение модели классификации как композиции информативных паттернов // Системы управления и информационные технологии. 2012. № 2 (48). С. 18–22.

9. UCI Machine Learning Repository [Electronic resource]. URL: http://archive.ics.uci.edu/ml/index.html (01.03.2019).

10. Кузьмич Р.И. Поиск закономерностей при решении задачи управления приземлением космического корабля // Актуальные проблемы авиации и космонавтики: материалы VIII Всероссийской научной-практической конференции творческой молодежи, посвященной 55-летию запуска первого искусственного спутника Земли: в 2 т. Т 1. Технические науки. Информационные технологии. Сообщения школьников. 2012. С. 306–307.

11. Refaeilzadeh P., Tang L., Liu H. On comparison of feature selection algorithms. AAAI Workshop – Technical Report. 2007. no. 05. P. 34–39.

12. Antamoshkin A.N., Masich I.S. Combinatorial optimization and rule search in logical algorithm of machine learning. Engineering & automation problems. 2010. no. 5. P. 52–57.

13. Антамошкин А.Н., Кузьмич Р.И., Масич И.С. Модифицированный метод логического анализа данных. М: Роспатент, 2016. № гос. рег. 2016619162.

14. Vijayarani S., Divya M. An Efficient Algorithm for Generating Classification Rules. International Journal of Computer Science and Technology. 2011. no. 2 (4). P. 512–515.

15. Sun B., Chen S., Wang J., Chen H. A robust multi-class AdaBoost algorithm for mislabeled noisy data. Knowledge-Based Systems. 2016. no. 102. P. 87–102.