DOI 10.37882/2223–2966.2021.01.26

ПОСТРОЕНИЕ ОБОСНОВАННЫХ КЛАССИФИКАЦИОННЫХ МОДЕЛЕЙ ПРИ ПРИНЯТИИ РЕШЕНИЙ И ПРОГНОЗИРОВАНИИ

BUILDING JUSTIFIED CLASSIFICATION MODELS FOR DECISION MAKING AND FORECASTING

K. Ponomareva
A. Stupina
A. Fedorova
L. Korpacheva

Summary. Classification is an important task of data mining, where the value of a discrete (dependent) variable is predicted based on the values of some independent variables. Classification models must provide correct predictions for new data instances. This article focuses on the key requirements for such models in any field where the model must be tested before it can be implemented. The main requirements for classification models are clarity and validity, which reflects the model's compliance with existing knowledge of the subject area under consideration. Providing clarity and validity of classification models contributes to their practical application in areas where previously such models were considered too theoretical and incomprehensible. So, a classification model that is accurate, understandable, and efficient is defined as acceptable for implementation.

Keywords: data mining, model productivity, classification, comprehensibility, justifiability, forecasting, medical diagnostics.

Пономарева Катерина Андреевна
Аспирант, Сибирский федеральный университет, Красноярск
ponomareva_katerina@mail.ru
Ступина Алена Александровна
Д.т.н., профессор, Сибирский федеральный университет, Красноярск
h677hm@gmail.com
Федорова Александра Витальевна
К.г.-м.н., доцент, Сибирский федеральный университет, Красноярск
alvitfedorova@gmail.com
Корпачева Лариса Николаевна
К.т.н, доцент, Сибирский федеральный университет, Красноярск
korp_0777@mail.ru

Аннотация. Классификация является важной задачей интеллектуального анализа данных, где значение целевой переменной предсказывается, основываясь на значениях некоторых независимых переменных. Классификационные модели должны обеспечивать правильные прогнозы для новых экземпляров данных. Статья посвящена ключевым требованиям, предъявляемым к подобным моделям в любой области, где модель должна быть проверена, прежде чем она может быть реализована. Основными требованиями к классификационным моделям выступают понятность и обоснованность, что отображает соответствие модели существующим знаниям рассматриваемой предметной области. Обеспечение понятности и обоснованности классификационных моделей способствует их практическому применению в областях, где ранее такие модели считались слишком теоретическими и непонятными. Таким образом, модель классификации, которая является точной, понятной и эффективной, может быть определена как приемлемая для реализации.

Ключевые слова: интеллектуальный анализ данных, производительность модели, классификация, понятность, обоснованность, прогнозирование, медицинская диагностика.

С каждым годом значительно возрастает количество создаваемой и перерабатываемой информации, которой оперирует в процессе деятельности человек. Стремительно растущие объемы данных требуют систематизации для повышения эффективности их использования. Однако одна лишь систематизация не решает задач оперативной и всесторонней обработки информации для ее использования с заданной целью. Актуальными сегодня являются автоматизация проведения анализа имеющихся данных и прогнозирование возможных событий на основе имеющихся наблюдений. Решению подобных задач, наряду с методами экспертных оценок и методами искусственного интеллекта, способствует интеллектуальный анализ данных.

Понятие интеллектуального анализа данных связано с широко распространенным понятием Data Mining, часто интерпретируемым как добыча данных, глубокий анализ данных, нахождение знаний, извлечение знаний в базах данных. Data Mining обеспечивает, в том числе, поддержку принятия решений, основанную на поиске

в больших массивах разнородных данных скрытых закономерностей (шаблонов информации) [4].

Относительно того, какие задачи следует относить к Data Mining, единое мнение на текущий момент не сформировано. В многочисленных литературных источниках рассматриваются такие задачи, как регрессия, классификация, анализ правил ассоциации, прогнозирование и кластеризация [10].

Практический интерес для принятия решений по большому кругу вопросов представляет решение проблемы классификации, под которой подразумевается задача присвоения набора данных предопределенному классу или группе в соответствии с его прогностическими характеристиками. Целью задачи классификации является построение модели, позволяющей в соответствии с установленными правилами, классифицировать будущие точки данных на основе набора конкретных характеристик. Такую модель называют моделью классификации или классификационной моделью.

Решение задачи классификации можно выполнять множеством различных методов. Наиболее часто применяются следующие методы: алгоритм C4.5, алгоритм CART, логистическая регрессия, линейный и квадратичный дискриминантный анализ, алгоритм κ-ближайших соседей (KNN), искусственные нейронные сети и метод опорных векторов (SVM) [6, 7].

Методы классификации нашли практическое применение в оценке кредитоспособности заемщиков, в медицинской диагностике (например, для прогнозирования деменции, классификации доброкачественной или злокачественной опухолей, диагностики осложнений инфаркта миокарда, выборе лучшего эмбриона для оплодотворения). [1]

В последнее время, благодаря имеющимся в распоряжении исследователей и практиков массивов данных, накопленных за значительные временные периоды, интеллектуальный анализ данных все чаще используется для таких направлений, как биоинформатика, маркетинговые исследования, избирательные кампании, для решения задач, направленных на борьбу с терроризмом и др. Это наглядно показывает широкий спектр и возможности применения интеллектуального анализа.

Для использования моделей классификации в решении задач существует ряд требований, связанных с производительностью: обеспечение правильных прогнозов, понятность и обоснованность [5]. Выполнение первого требования обеспечивает способность модели хорошо обобщать и обеспечивать правильные прогнозы для новых «невидимых» экземпляров данных. Правильность прогнозов обычно оценивается процентом правильно классифицированных экземпляров данных. Помимо этого, необходимо учитывать чувствительность и специфичность, которые формируются на основе матрицы путаницы. Для оценки прогностических качеств модели обычно используются ROC-кривая и площадь области под этой кривой AUC.

Исследования показали, что в целом наиболее точные прогнозы дают нелинейные модели, так как они способны фиксировать нелинейности в данных, что наиболее часто характерно для эмпирических данных. Одновременно это выступает и их главным недостатком, поскольку модель считается «черным ящиком», что существенно затрудняет или делает невозможным понимание логики, стоящей за принимаемыми решениями. Это порождает второе требование производительности — понятность.

Понятность является ключевым требованием для понимания логики, заложенной в основу прогностической модели. Понимание логики прогнозирования необходимо в любой области. Прежде всего это требуется для проверки разработанной модели, прежде чем она сможет быть фактически реализована. В некоторых областях, таких как кредитный скоринг и медицинская диагностика, отсутствие понятности является ключевой проблемой и вызывает негативное отношение к использованию классификационной модели или полный отказ от модели.

Таким образом, понятность обеспечивает снижение негативного восприятия при внедрении модели в практику. Основными факторами понятности выступают тип вывода и размер выходного сигнала. Тип ввода связан с понятностью конкретного типа выходных данных, который в значительной степени зависит от предметной области. Наиболее понятными являются классификаторы, разработанные на основе четко сформулированных правил, исходя из практического опыта решения конкретных задач предметной области. При рассмотрении размера выходного сигнала предпочтение отдают моделям меньшего размера.

Понятность можно измерить на основе ранжирования. Различные типы выходных данных можно проранжировать по степени понятности. Исходя из этого можно утверждать, что модели, основанные на правилах, являются более сложными, чем линейные, которые более понятны, чем нелинейные. Однако этот вывод не всегда верен. Линейная модель с одной переменной является более понятной, чем модель, разработанная на основе правил. И чем больше правил заложено в модель, тем она более сложна с позиции понимания (например, модель с более чем 20 правилами). Кроме того, сложность

в понимании модели связана с типом применяемого классификатора и области применения (решаемой задачей). Например, понятность классификатора ближайшего соседа более сложная. Если рассматривать 1NN классификатор, то он может быть логичным и понятным в одной области и довольно бессмысленным в другой, например, для сети с многомерными данными, где даже самый похожий обучающий экземпляр все еще отличается по многим переменным. Однако, как правило, это ранжирование по типам выходных данных будет истинным и наблюдаемым [8].

Разработка понятных классификаторов может осуществляться различными способами. В качестве наиболее приемлемых форматов для классификационной модели рассматриваются правила, деревья и методы, которые индуцируют такие модели, что обусловлено их лингвистической природой и легкостью в понимании для неэкспертов.

Еще один подход, применяемый для получения понятного классификатора — визуализация. Визуализация данных подразумевает отображение информации в графическом или табличном формате. Она направлена на обеспечение наилучшей интерпретации и, как следствие, валидации информации человеком, с целью получения приемлемого классификатора [2, 10]. Визуализация может быть представлена графиками, самоорганизующимися картами, таблицами и диаграммами решений.

Как было уже сказано выше, понятность необходима для проверки соответствия модели накопленным знаниям рассматриваемой предметной области. Наряду с понятностью необходима и обоснованность — интерпретируемость и доказательность модели.

В контексте интеллектуального анализа данных модель обоснована, если она соответствует существующему домену знаний. Поэтому для обеспечения обоснованности модели она должна быть подтверждена экспертом в рассматриваемой предметной области. Экспертное подтверждение, в свою очередь, обеспечивается понятностью модели. Обоснованность модели определяется, прежде всего, обоснованностью элементов ее структуры и основных применяемых в модели правил. Данные аспекты можно считать установленными в случае получения с помощью модели характеристик, которые соответствуют или близки характеристикам реальной системы. В качестве параметров, отражающих степень такого соответствия, выступают адекватность модели, оценить которую можно по средним значениям отклика модели и реальной системы, по дисперсиям отклонений откликов модели от среднего значении откликов системы, по максимальному значению относительных отклонений отклика модели от отклика реальной системы, а также устойчивость модели.

Необходимо понимать, что только способность модели предсказывать состояние реальной системы в некоторый определенный будущий момент не может выступать достаточно убедительным свидетельством полезности модели [3]. Для решения проблемы обоснованности было предложено несколько вариантов адаптации существующих методов классификации и мера, которая позволяет определить степень соответствия разработанной модели задаваемым ограничениям.

Для измерения обоснованности [9] предложена следующая формула:

$$\text{Обоснованность} = 1 - \sum_{i=1}^{n} w_i \sum_{j=1}^{|pr_i|} \frac{1}{|pr_i|} * I(pr_{i,j}). \quad (1)$$

Мера обоснованности задается уравнением

$$\sum_{i=1}^{n} w_i = 1,$$

таким, что мера обоснованности ограничена в пределе между 0 и 1. Набор данных состоит из n переменных V_i, где

n — число переменных;

w_i — определенный пользователем вес, показывающий относительную важность переменной V_i;

$|pr_{i,j}|$ — обще число профилей для переменной V_i;

$I(pr_{i,j})$ — оператор, присваивающий 1, если несоответствие присутствует в j профиле переменной i, и 0 в противном случае.

Данная мера может быть использована и для линейных классификаторов, без использования таблиц решений. В этом случае для каждой переменной существует ровно один профиль. Несоответствие имеет место, если знак переменной V_i не соответствует ожидаемому знаку. Для определения w_i можно использовать аналогичный подход, как и для классификаторов, основанных на правилах. Для статистически предлагаемых w_i можно использовать нормализованные частные коэффициенты корреляции, которые являются корреляцией, когда все другие переменные сохраняются на фиксированных значениях, как статистически обоснованное предположение.

Установка весов (w_i) является сложным, но необходимым этапом при расчетах обоснованности. Установка весов полностью статистически обоснована и основана только на интуиции эксперта в исследуемой предметной области. Используя только статистику, можно точно воспроизвести относительную важность переменных в наборе данных. Поскольку невозможно добиться

Таблица 1. Переменная «Помнит название улицы»

Возраст	Количество лет обучения	Помнит название улицы	Нарушение	Норма
≤ 70	≤ 5	—	×	—
	> 5	—	—	×
> 70 и ≤ 80	≤ 5	—	×	—
	> 5	да	×	—
		нет	—	×
> 80	≤ 5	да	—	×
		нет	×	—
	> 5	—	—	×

Таблица 2. Переменная «Количество лет обучения»

Возраст	Помнит название улицы	Количество лет обучения	Нарушение	Норма
≤ 70	—	≤ 5	×	—
		> 5	—	×
> 70 и ≤ 80	да	—	×	—
	нет	≤ 5	×	—
		> 5	—	×
> 80	да	—	—	×
	нет	≤ 5	×	—
		> 5	—	×

Таблица 3. Переменная «Возраст»

Количество лет обучения	Помнит название улицы	Возраст	Нарушение	Норма
≤ 5	да	≤ 80	×	—
		> 80	—	×
	нет	—	×	—
> 5	да	≤ 70	—	×
		> 70 и ≤ 80, > 80	×	—
			—	×
	нет	—	—	×

идеального качества данных (всегда присутствуют шум, ограниченная доступность данных и т.д.), то ошибочно полностью полагаться только на эти меры. Так, например, отсутствие корреляционной связи между факторной и целевой переменными не означает, что вес (w_i) будет равен 0. Так же в случае, если только одна переменная, из нескольких рассматриваемых в модели, идеально предсказывает целевую переменную, не означает, что вес (w_i) этой переменной равен 1, а всех остальных 0.

Одновременно, при принятии решения об обоснованности модели, не рекомендуется полагаться только на мнение эксперта по предметной области. Это связано с тем, что интуиция и знания эксперта о влиянии различных факторов так же имеют ограничения. Наиболее целесообразным рассматривается объединение этих двух оценок. В этом случае статистически обоснованные веса корректируются в соответствии с мнением эксперта или эксперт осуществляет выбор из имеющегося множества возможных конфигураций весов. Определение влияния изменения веса на прогноз значения целевой переменной, рекомендуется объединить с анализом чувствительности. Это позволит зафиксировать влияние на меру обоснованности даже незначительных изменений в настройках веса.

Установление меры обоснованности можно рассмотреть на примере задачи медицинской диагностики. Так,

$$\text{Обоснованность} = 1 - \left[\begin{array}{c} \frac{1}{3} * \left(\frac{1}{2} I(pr_{\text{помнит},1}) + \frac{1}{2} I(pr_{\text{помнит},2}) \right) + \\ + \frac{1}{3} * \left(\frac{1}{3} I(pr_{\text{образование},1}) + \frac{1}{3} I(pr_{\text{образование},2}) + \frac{1}{3} I(pr_{\text{образование},3}) \right) + \\ + \frac{1}{3} * \left(\frac{1}{2} I(pr_{\text{возраст},1}) + \frac{1}{2} I(pr_{\text{возраст},2}) \right) \end{array} \right] = \quad (2)$$

$$1 - \left[\begin{array}{c} \frac{1}{3} * \left(\frac{1}{2} * 1 + \frac{1}{2} * 0 \right) + \frac{1}{3} * \left(\frac{1}{3} * 0 + \frac{1}{3} * 0 + \frac{1}{3} * 0 \right) + \\ + \frac{1}{3} * \left(\frac{1}{2} * 1 + \frac{1}{2} * 1 \right) \end{array} \right] = 0{,}5 = 50\%$$

$$\text{Обоснованность} = 1 - \left[\frac{1}{3} * \left(\begin{array}{c} \frac{1}{3} I(pr_{\text{помнит}}) + \frac{1}{3} I(pr_{\text{возраст}}) + \\ + \frac{1}{3} I(pr_{\text{образование}}) \end{array} \right) \right] = 1 - \left[0 + \frac{1}{3} + 0 \right] = 0{,}67 = 67\% \quad (3)$$

например, можно рассмотреть классификатор для выявления пациентов с деменцией и здоровых пациентов. В этом случае для медицинского эксперта релевантными могут быть рассмотрены следующие переменные:

- количество лет, в течение которых пациент обучался;
- способность пациента вспомнить название улицы, на которой он живет;
- возраст пациента.

Набор правил для прогнозирования деменции представлены следующим образом:

1: Если «Помнит название улицы» = нет и «Количество лет обучения» > 5,
2: то пациент = нормальный,
3: Иначе если «Помнит название улицы» = да и «Возраст» > 80,
4: то пациент = нормальный,
5: Иначе если «Помнит название улицы» = да и «Количество лет обучения» > 5 и «Возраст» ≤ 70,
6: то пациент = нормальный;
7: Другой пациент = диагностирована деменция.

Данный классификатор на основе правил содержит три правила для диагностики состояния пациента. Ожидания относительно названия улицы очевидны: факт, что пациент помнит название улицы, на которой живет, говорит о его нормальном душевном состоянии. Для возрастной переменной также установлены прямые ожидания: чем старше пациент, тем выше вероятность возникновения деменции. И можно ожидать, что более образованные пациенты проявляют более высокую степень умственной деятельности и поэтому, менее подвержены деменции. Ниже в таблицах 1, 2, 3 показаны профили для трех рассмотренных независимых переменных.

Для переменной «Помнит название улицы» есть два профиля. Первый профиль $pr_{помнит,\ 1}$ указывает на то, что для двух пациентов с одинаковыми характеристиками по возрасту (> 70 и ≤ 80) и количеству лет обучения (> 5), но с разными значениями переменной «Помнит название улицы», тот, кто вспоминает название своей улицы, имеет заболевание, а тот, кто фактически не помнит название, классифицируется как нормальный. Это противоречит здравому смыслу и поэтому $I(pr_{помнит,\ 1}) = 1$. Для второго профиля $pr_{помнит,\ 2}$ выполняется ограничение домена и поэтому $I(pr_{помнит,\ 2}) = 0$.

Аналогично, в таблицах 1, 2, 3 показано, что все профили для переменной «Количество лет обучения» удовлетворяют ожиданиям, в то время как ни один из профилей для переменной «Возраст» не соответствует ожиданиям, что приводит к следующей мере обоснованности (предполагает равные веса) (2).

Поскольку переменная «Помнит название улицы» может быть только 0 (пациент не помнит название улицы) или 1 (пациент помнит название улицы), ожидается положительная совместная эффективность: то, что пациент помнит название улицы, увеличивает вероятность быть здоровым. Для возраста ожидается отрицательный коэффициент, так как более старшие пациенты имеют повышенный шанс наличия нарушений. Ожидание большего количества лет обучения оказывает положительное влияние на психическое состояние пациента. Таким образом, обоснованность линейной классификационной модели можно представить как (3).

Рассмотренная метрика обоснованности позволяет не только сравнивать классификаторы, но и требовать минимального порога обоснованности для модели клас-

сификации, реализованной в системе поддержки принятия решений. Порог может доходить даже до 100%.

Таким образом, понятность и обоснованность являются важными требованиями для моделей классификации в различных областях. К таким моделям предложено несколько подходов — от простых методов индукции правил до передовых инкрементных подходов. Понятные, обоснованные классификационные модели повышают их применимость и практическую значимость в тех областях, где ранее модели классификации считались слишком теоретическими и непонятными. Таким образом, наряду с новыми возможностями для интеллектуального анализа данных, становится неоспоримым практическое применение данного анализа в поддержки принятия решений и прогнозировании.

ЛИТЕРАТУРА

1. Антамошкин А.Н., Масич И. С. Алгоритмы псевдобулевой оптимизации для выявления информативных закономерностей в данных: научная статья / А. Н. Антамошкин, И. С. Масич // Седьмая международная конференция «Системный анализ и информационные технологии (САИТ-2017)».— Светлогорск, 2017.— 8 с.
2. Визуализация границ решения классификатора на основе изображений. URL: https://habr.com/ru/post/483608/ (дата обращения 19.10.2020)
3. Форрестер, Дж. Основы кибернетики предприятия (Индустриальная динамика). URL: https://bzbook.ru/OSNOVY-KIBERNETIKI-PREDPRIYaTIYa.1.html#a1.DZh-FORRESTER-OSNOVY-KIBERNETIKI-PREDPRIYaTIYa-INDUSTRIALJNAYa-DINAMIKA (дата обращения 19.10.2020)
4. Чернышова Г. Ю. Интеллектуальный анализ данных: учебное пособие // Саратовский государственный социально-экономический университет.— Саратов, 2012.— 92 с.
5. Чубукова И. А. Data Mining: курс лекций / И. А. Чубукова // Киевский национальный экономический университет им. Вадима Гетьмана.— Киев.— 326 с.
6. Duda R.O., Hart P. E., Stork D. G. Pattern Classification // John Wiley and Sons, New York, second edition, 2015.
7. Hastie T., Tibshiran R., Friedman J. The Elements of Statistical Learning, Data Mining, Inference, and Prediction.— Springer, New York, 2015.
8. Maimon O.O., Rokach L. Decomposition methodology for knowledge discovery and Data Mining: theory and applications (Machine Perception and Artificial Intelligence) // World Scientific Publishing Company, 2015.
9. Robert Stahlbock, Sven F. Crone, Stefan Lessmann Data Mining. Special Issue in Annals of Information Systems // Springer Science+Business Media, LLC, 2010.
10. Tan P.-N., Steinbach M., Kumar, V. Introduction to Data Mining.— Pearson Education, Boston, MA, 2016.