ИДЕНТИФИКАЦИЯ МЕХАНИЗМОВ КОМПЛЕКСНОГО ОЦЕНИВАНИЯ КАК ПОДХОД АНАЛИЗУ ДИСКРЕТНЫХ ДАННЫХ

Сергеев В. А.[1], Коргин Н. А.[2]
(*ФГБУН Институт проблем управления*
им. В.А. Трапезникова РАН, Москва)

Представлен подход к анализу дискретных наборов данных на основе недавно разработанных подходов к идентификации механизмов комплексного оценивания. На учебном примере «Задача о Зените» демонстрируется методика идентификации механизмов комплексного оценивания, рассматриваются пятнадцать допустимых структур, в качестве дополнительного инструмента вводится метод анализа групп эквивалентности. В основе описываемого подхода к идентификации лежит составление и решение оптимизационного функционала, составленного с помощью дерева свертки. Рассмотрение групп эквивалентности может существенно сокращать вычислительные затраты для исследования набора данных. На основе наборов данных оценки жизнеспособности компаний и оценки дизайнерских проектов для ста пяти допустимых структур демонстрируются результаты работы методики идентификации механизмов комплексного оценивания по неполным данным. Освещаются возможности предложенного подхода.

Ключевые слова: идентификация и редукция модели; планирование и контроль производства; моделирование и принятие решений в сложных системах; комплексное оценивание; унитарное кодирование; унитарные функции; неполные данные; анализ дискретных данных.

1. Вступление

На примере рассмотрения трех случаев идентификации механизмов комплексного оценивания (МКО) на основе трех неполных наборов данных демонстрируются возможности анализа данных, предоставляемые предложенным в работе [4] подходом к идентификации МКО. МКО посвящено большое количество работ, одна из последних – [5]. В основном предполагается, что для определения элементов МКО – структуры и матриц – используется экспертный подход. Метод, рассматриваемый в данной

[1] *Владимир Александрович Сергеев, м.н.с. аспирант* (*sergeev.bureau@gmail.com*).
[2] *Николай Андреевич Коргин, д.т.н., доцент, г.н.с.* (*nkorgin@ipu.ru*).

работе, напротив, основан на получении МКО на основе работы с набором входных данных из исследуемой задачи. В этой статье мы сделаем шаг вперед и используем МКО не как механизм принятия решений, а как модель, описывающую набор дискретных данных, под именем МКО. Основным преимуществом по сравнению с нашей предыдущей работой является расширение метода для работы с неполными данными. Мы предлагаем подход, выходящий за рамки стандартных методов работы с неполными данными, используемых в статистике, см. [6]. В отличие от метода деревьев решений (см., например, [9, 8]), мы движемся снизу вверх, чтобы получить комплексную оценку на основе листьев.

На примере идентификации первой модели – набора данных «О Зените» в двоичной шкале – мы продемонстрируем базовый подход к работе с неполными наборами данных. На примерах второй и третьей задач покажем, какие еще подходы к анализу данных открываются при использовании предложенного метода. Данные для второго задания – это оценка финансовой устойчивости ряда российских риэлторских компаний (см., [1]), а для третьего – оценка дизайн-проектов студентов Уральского государственного архитектурно-художественного университета. (УрГАХУ) для условий крайнего севера из Арктической школы дизайна (см., например, [2]). Данные по второму и третьему заданию представлены по тройной шкале.

2. Основные понятия и определения

Использование унитарного кодирования (см., например, [7]) позволяет реализовать подход, в рамках которого любой МКО можно рассматривать как последовательность матричных операций с унитарно закодированными значениями листьев в соответствии со «словом», описывающим соответствующий МКО. Результатом этой последовательности операций является унитарно закодированный результат МКО. Принимая во внимание тот факт, что любое полное двоичное дерево с l листьями должно иметь $l - 1$ внутренних узлов, включая корень, любой МКО можно закодировать в виде «слова» с $2l - 1$ буквами

в упрощенном квадратичном представлении; Буквы $l-1$ должны обозначать матрицы свертки (внутренние узлы), а l – листья. Например, последовательную структуру можно записать как $M1\,l1\,M2\,l2\,M3\,l3\,l4$, а симметричную – как $M1\,M2\,l1\,l2\,M3\,l3\,l4$, см. рис. 1.

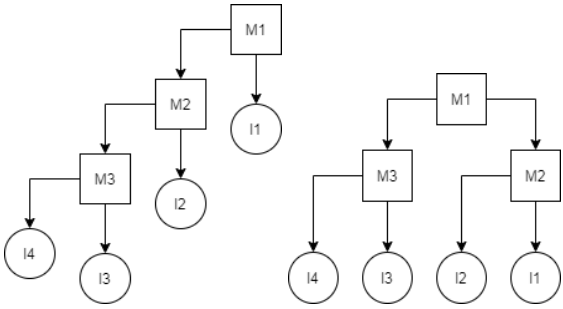

Рис. 1. Типы полных деревьев на четырех именованных листьях

Мы предполагаем, что данные подготовлены заранее и не содержат повторяющихся и противоречивых примеров. На основании входных данных названия столбцов обозначаются как листья. Затем для каждой из всего спектра структур на этом количестве листьев необходимо сформировать и максимизировать полином, см. [4].

2.1. ЗАДАЧА «О ЗЕНИТЕ»

В рассматриваемой задаче из [3] исследуется вопрос прогнозирования результата игры футбольной команды в зависимости от четырех параметров: расположение в турнирной таблице относительно соперника, проводится ли матч дома, пропускает ли кто-либо из лидеров команды матч, идет ли дождь. Всего определено восемь из шестнадцати возможных примеров. Далее для формирования математической модели исходные данные были закодированы в двоичной шкале.

В работе [3] рассматривается хорошо известный подход с использованием деревьев решений. Мы решим эту проблему, определив математическую модель на дереве свертки и матрицах свертки. Далее по тексту используются определения

и обозначения из статьи [4]. Этот пример рассматривается как обучающий набор в унитарной шкале, непротиворечивый и, как упоминалось выше, неполный. Таким образом, у нас есть только восемь обучающих примеров вместо шестнадцати для полного набора, в шкале два на четырех переменных.

В соответствии с изложенным в [4], этот набор данных может быть реализован через определенный МКО, если будут реализованы все примеры, описывающие этот набор. В результате расчета задачи идентификации оказалось, что эту задачу реализуют всего три структуры, из которых две последовательные и одна симметричная.

3. Набор данных по оценке жизнеспособности компаний

Следующий пример – из области оценки жизнеспособности компаний. Основываясь на работе [1], используются данные оценки компаний по недвижимости в зависимости от пяти параметров: дебиторская задолженность, кредиторская задолженность, нераспределенная прибыль, запасы, основные средства. Первоначально рассматривались две сотни российских строительных компаний, из которых сто ликвидированы или находятся в процессе ликвидации в связи с банкротством и сто экономически успешных компаний, по которым дела о банкротстве не возбуждены. Для реализации процесса оценки финансового положения в работе использовались балансы предприятий. В результате обработки данных осталось сорок восемь примеров, дублирующие и противоречивые примеры ранее были удалены, притом что полный набор в троичной шкале на пяти листьях состоит из двухсот сорока трех примеров.

Проверка всех возможных ста пяти структур полных деревьев с именованными листьями показала, что наилучший результат аппроксимации для этой задачи реализует сорок пять из сорока восьми примеров, и нет структуры, которая реализует все сорок восемь примеров. Наилучшее приближение данных достигается по структуре $M1\ l3\ M2\ l4\ M3\ l5\ M4\ l1\ l2$.

4. Набор данных по арктическому дизайну

В третьем случае рассматривается вопрос определения математической модели на основе данных об оценках дипломных проектов о развитии крайнего севера студентов-дизайнеров из УрГАХУ. В предоставленном наборе данных представлен тридцать один пример из двухсот сорока трех возможных для полного набора на пяти листах в троичной шкале. Каждый пример имеет пять параметров: актуальность, экономика, этика/экология, эстетика/имидж, технические характеристики. часть. Повторяющиеся и противоречивые примеры были ранее удалены.

Были исследованы все сто пять возможных структур и найден ряд структур с максимальным значением реализованных примеров. Рассмотрение групп листьев показывает, что сочетание критериев эстетики и технической части, а также эстетики и этики/экологии дает наилучшие результаты для ансамбля.

5. Заключение

В рассмотренных случаях для расчета использовался решатель Гуроби. Для второй и третьей задач в процессе работы с данными было выявлено, что допустимо без потери качества решения установить лимит времени работы решателя для одной структуры в один час. Первая задача решается быстро и без такого ограничения. Природа решаемых задач оптимизации такова, что решение для структур, имеющих реализацию на представленных данных, находится за секунды, а поиск аппроксимаций может занять много времени.

Анализ групп эквивалентности позволяет выделить заведомо нереализуемые структуры и значительно ускорить вычисления.

На данных примерах мы показали, что предлагаемый подход к анализу дискретных наборов данных в качественных или категориальных оценках предоставляет:

• возможность декомпозиционного (древовидного) представления модели, описывающей эти данные;

• структурный анализ вклада отдельных переменных в оценку;

• анализ влияния отдельных переменных и групп переменных на конечное значение агрегированного показателя;

• анализ монотонности обучающей выборки дискретных данных;

• выделение целевых примеров для дальнейшей идентификации модели.

Данный функционал вполне применим к областям, где традиционно используется ЭДВ, таким как маркетинг, здравоохранение, оценка окружающей среды и другим.

Следующий шагом планируется развивать механизм групп эквивалентности для повышения скорости работы через отсев заведомо нереализуемых структур, а также синтез реализуемых структур.

6. Благодарности

Авторы выражают благодарность за частичное финансирование гранта РНФ 17-78-20047.

Литература

1. АЛЕКСЕЕВ А.О., НОСКОВА А.Р. *Исследование достоверности прогнозирования банкротства при введении новой категории финансового состояния предприятий* // Прикладная математика и вопросы управления. – 2020. – №3. – С. 105–122. – DOI: 10.15593/2499-9873/2020.3.06.
2. КРАВЧУК С.Г., ГАРИН Н.П., КУКАНОВ Д.А., ГОСТЯЕВА М.А., КОНЬКОВА Ю.С. *Арктический дизайн основные понятия и практика реализации* // Дизайн и технологии. – 2017. – №62(104). – С. 17–28.
3. НИКОЛЕНКО С., ТУЛУПЬЕВ А. *Самообучающиеся системы*. – М.: МЦНМО, 2009.
4. BURKOV V., SERGEEV V., KORGIN N. *Identification of Integrated Rating Mechanisms as Optimization Problem* // 13th IEEE Int. Conference "Management of Large-Scale System Development" (MLSD-2020). – 2020. – P. 1–5.
5. BURKOV V., BURKOVA I., POLOVINKINA A., SHEVCHENKO L. *Integrated Assessment System Based on Dichotomous Tree* //

Advances in Intelligent Systems and Computing. – 2021. – Vol. 1258. – P. 578–587.
6. KWANG K., SHAO J. *Statistical methods for handling incomplete data*. – Boca Raton: CRC Press. – 2014.
7. OKADA S., OHZEKI M., TAGUCHI S. *Efficient partition of integer optimization problems with one-hot encoding* // arXiv preprint. – arXiv: 1906.07385. – 2019.
8. QUINLAN J. *Induction of Decision Trees* // Machine Learning 1. – 1986. – P. 81–106.
9. ROKACH L., MAIMON O. *Data Mining and Knowledge Discovery Handbook*. – New York: Springer, 2005. – Chapter 9.

IDENTIFICATION OF INTEGRATED RATING MECHANISMS AS AN APPROACH TO DISCRETE DATA ANALYSIS

Vladimir Sergeev, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, research assistant, postgraduate (sergeev.bureau@gmail.com).
Nikolay Korgin, V.A. Trapeznikov Institute of Control Sciences of RAS, Moscow, D.E.Sc., associate professor, principal researcher. (Tel: +7495-335-60-37; e-mail: nkorgin@ipu.ru).

Abstract: An approach to the analysis of discrete data sets based on recently developed method of the identification of integrated assessment mechanisms is presented. On the simple example - the case " Results of football team Zenith ", the method of identifying the mechanisms of complex assessment is demonstrated, fifteen valid structures are considered, as an additional tool, the method of analysis of equivalence groups is introduced. The described approach to identification is based on the compilation and solution of an optimization functional compiled using a convolution tree. Consideration of equivalence groups can significantly reduce the computational costs for exploring a dataset. On the basis of data sets for assessing the viability of companies and evaluating design projects, based on one hundred and five admissible structures, the results of a methodology for identifying integrated assessment mechanisms based on incomplete data are demonstrated. Possibilities of the proposed approach are highlighted.

Keywords: Identification and model reduction; Production planning and control; Modelling and decision making in complex systems; Integrated assessment; One-hot encoding; Incomplete data; Discrete data analysis.

УДК 519
ББК 32.81