ИСПОЛЬЗОВАНИЕ МЕТОДОВ DATA MINING ПРИ ИССЛЕДОВАНИИ МОДЕЛИ ПРЕСТУПЛЕНИЯ

THE USE OF DATA MINING METHODS IN THE STUDY OF THE CRIME MODEL

S. Sukhov
S. Krygin
S. Kuvychko

Summary. The paper presents the methodology and results of the study of the possibility of using Data Mining methods in the analysis of criminal activity on the example of illegal acts in the field of information and telecommunication technologies. The article describes the peculiarity of using decision trees in the processing of predictor variables of different types, the analysis of the structure of empirical data, the establishment of links between the trace picture of a particular situation and the method of committing a wrongful act in order to further form standard versions of the investigation.

Keywords: Data Mining, decision trees, standard versions, law enforcement.

Сухов Сергей Николаевич
К.ю.н., Нижегородская академия МВД России
amlawdd@yandex.ru
Крыгин Сергей Владимирович
К.ю.н., Нижегородская академия МВД России
kryginsv@mail.ru
Кувычков Сергей Иванович
К.ю.н., доцент, Приволжский филиал Российского государственного университета правосудия, г. Нижний Новгород
redsxrjd@mail.ru

Аннотация. В работе представлена методика и результаты исследования возможностей применения методов DataMining при анализе криминальной активности на примере противоправных деяний в сфере информационно-телекоммуникационных технологий. Описывается особенность применения деревьев решений при обработке предикторных переменных различных типов, анализа структуры эмпирических данных, установление связей между следовой картиной конкретной ситуации и способом совершения противоправного деяния в целях последующего формирования типовых версий расследования.

Ключевые слова: Data Mining, деревья решений, типовые версии, правоохранительная деятельность.

Развитие информационных технологий в современных условиях характеризуется возрастанием как объема информации, так и усложнением структурного построения массивов информации практически во всех сферах человеческой деятельности. Сложная социально-экономическая ситуация в стране, высокий уровень преступности, требует повышения эффективности деятельности правоохранительных органов, разработки систем поддержки решений в правоохранительной сфере, так как использование традиционных способов обработки информационных массивов становится малоэффективным. Одним из актуальных вопросов остается проблема разработки и исследования моделей общественных отношений, возникающих в связи с противодействием преступности. Существуют различные методы исследования моделей — применение деревьев решений, использование искусственных нейронных сетей, эволюционное программирование, применение нечеткой логики и ассоциативной памяти и т.д.

Для повышения эффективности решения поисково-познавательных задач в правоохранительной сфере целесообразно использовать моделирование. В юридической практике крайне востребована и эффективно используется криминалистическая модель минувшего события, которую можно охарактеризовать как искусственно разработанную систему, которая воспроизводит объект, который она заменяет, с определенной степенью сходства. Анализ и исследование данной модели позволяет выделить новые знания об оригинальных объектах и используется в последующем, для более эффективного решения управленческих, поисково-идентификационных и иных задач как в практической деятельности правоохранительных органов, так и в научных исследованиях. Необходимо отметить, что использование методов моделирования при исследовании социальных или экономических объектов сталкивается со значительными трудностями, так как практически невозможно учесть в модели всю совокупность факторов, которые влияют на состояние и особенности функционирования объекта, в отличии, например, от моделирования технических объектов [1].

Любое противоправное деяние можно описать комплексом факторов, которые могут в значительной степени оказывать влияние на состояние или функционирование объекта, который мы исследуем. Накопленный эмпирический материал относительно исследуемого объекта (элементы состава преступления) позволяет построить вектор развития описываемой ситуации, что

приводит к возможности значительно повысить эффективность деятельности правоохранительных органов и, например, выделить типовые версии. Построение типовых версий является одним из основных методов расследования преступлений.

Определение того факта, какая типовая версия будет являть лучшей, определяется на основе анализа качественных свойств и количественных характеристик противоправного деяния. В данной ситуации необходимо исследование n-мерного пространства с целью определения наиболее близко расположенных признаков, с последующим исследованием связи, которая возникает между *зависимой переменной (отклик)* и экзогенными переменными.

Эмпирических данных накапливается все больше, и эксперты уже не справляются с их обработкой, в связи с чем, активно развиваются относительно новые направления по интеллектуализации методов анализа данных, которые в дополнение к традиционным методам анализа и OLAP-системам, способны выявлять в системе скрытые закономерности, находить причинно-следственные связи, определение локации событий, их прогнозирование и наглядное представление, т.е. формируют знания на основе данных. Одним из несомненных плюсов использования методов Data Mining является возможность наглядной демонстрации и доведения полученных результатов вычислений понятным языком лицам, не имеющим математического или технического образования.

Технология интеллектуального анализа данных (Data Mining) относится к междисциплинарной области знаний, широко применяются в различных сферах жизнедеятельности [2].

Исследования в области статистики длительное время с осторожностью относились к данным методам исследования, считалось что подобные методы не дают полноценной картины исследуемого события, связей и закономерностей, считали Data Mining набором методов исследования с применением искусственного интеллекта, анализа массивов данных и статистики. Однако, исследования последнего десятилетия показали высокую практическую значимость описываемых методов в различных направлениях социально-экономического, технического и информационного развития общества [3]. Проблематика исследования анализа данных, обнаружения закономерностей в массивах информации, построение деревьев решений встречается во многих предметных отраслях народного хозяйства [4, 5].

Эффективно использовать методы Data Mining для криминалистических целей. Вопросы разработки типовых версий криминальной активности всегда были предметом поиска ученых криминалистов, описываемые нами методы позволяют значительно расширить традиционные подходы к формированию структуры и содержанию типовой версии.

Для практического использования в деятельности правоохранительных органов недостаточно простого выяснения природы явления, необходимо обеспечить наглядную демонстрацию полученных результатов. Основная направленность рассматриваемых методов — поиск решения, на основе которого возможно перспективное моделирование ситуации, оценка вероятного развития событий.

В зависимости от поставленных задач методы Data Mining можно подразделить на две группы. Первая группа методов объединяет в себе задачи по кластеризации и сегментации, вторая решает задачи прогнозирования. Применение методов из первой группы позволит получить описательные результаты, например, выявить шаблоны данных, которые в последующем можно подвергнуть классическим процедурам анализа и интерпретировать полученные результаты (к таким методам можно отнести: алгоритм k-средних, k-***медианы, иерархичные методы кластерного анализа, самоорганизующиеся карты Кохонена и другие)***. Использование методов второй группы позволяют, основываясь на значениях известных переменных, прогнозировать неизвестные или будущие значения других переменных. При благоприятных условиях возможно построение робастной модели, которую достаточно трудно проверить в практике деятельности правоохранительных органов, в связи с тем, что результаты носят характер эвристик.

Учитывая, что для учета и сравнения большого числа переменных, выявления закономерностей возможно использование самых разнообразных методов, определимся, что более детально в статье остановимся на использовании метода «добычи данных», основанного на применении деревьев решений.

Криминальная активность как объект исследования обладает огромным количеством свойств, в связи с чем, возникают значительные трудности по определению того необходимого минимума характеристик, которые дают возможность исследовать данную активность [6].

Каждому абстрактному образу свойства можно определить переменную, что позволит, в последующем, сформировать определенную систему и связывать характеристики отдельных свойств с многообразием их проявлений в системе.

Сопоставляя множество наблюдений конкретного свойства с заранее определенными базовыми значениями, мы сможем определять

Таблица 1. Результат анализа зависимости способа совершения противоправного деяния от специальности (рода деятельности) лица, которое его совершило

Порядковый номер	Направленность ветвления		Распределение результатов по группам					Класс прогноза
	Вариант ветвления А	Вариант ветвления В	Специальность№ 1	Специальность № 2	Специальность№ 3	Специальность № 4	Специальность № 5	
1	2	3	68	33	27	25	12	1
2	4	5	62	31	4	11	4	1
3			6	2	23	14	8	3
4			55	5	2	7	2	1
5			11	30	0	8	0	2

Многомерность переменных приводит к сложности табулирования данных, что вызывает затруднения при объединении различных переменных в единую систему.

Сложность формирования типовых версий заключается в необходимости выполнения правила, согласно которому необходимо проявление характеристик следов противоправной деятельности не по отдельности, а в совокупности

При формировании типовых версий необходимым условием является одновременное проявление следов противоправной деятельности. Данные особенности и потребности послужили основанием для формирования таких методов моделирования как применение деревьев решений и использование искусственных нейронных сетей, об особенностях работы с которыми мы остановимся более подробно в другой статье.

Для формирования прогноза развития криминальной активности возможно использование деревьев решений, которые являются одним из основных методов Data Mining. Однако, необходимо учитывать, что данные методы необходимо использовать совместно с традиционными методами, такими как нелинейное оценивание, непараметрическая статистика, кластерный анализ и т.д., что в совокупности дает практико-ориентированный инструмент исследования информационных структур.

Достоинством использования деревьев решений выступает их возможность в графическом виде визуализировать и интерпретировать полученные выводы, что облегчает их восприятие по сравнению с классическим числовым форматом, поэтому многие предметные области используют данные методики.

Другим достоинством применения деревьев решений является гибкость данного метода по сравнению с классическими методами анализа. Использование одномерного ветвления при применении деревьев решений позволяет оценивать вклад отдельных переменных и дает возможность обрабатывать предикторные переменные различных типов, такие как интервальные, категориальные, ранговые (порядковые) и другие [7], в отличие, например, от традиционного метода анализа дискриминантных функций, при котором осуществляется поиск линейной комбинации признаков для разделения двух или более классов (событий). В последующем, данную комбинацию возможно использовать как линейный классификатор, например, для снижения размерности перед классификацией.

Материалы и методы исследования

Построение деревьев решений, вычисление критерия χ^2 К. Пирсона и другие вычисления описываемые в данной статье были выполнены в программном комплексе Statistika. Деревья решений строились используя алгоритм CART [8], который позволяет решать задачи классификации и регрессии при построении дерева решений.

Построение бинарного дерева решений осуществлялось по методу «Classification And Regression Trees», при котором происходит полный перебор возможных комбинаций.

Не рассматривая в деталях особенности обработки данных приведем пример исследования зависимости способа совершения противоправного деяния от специальности (рода деятельности) лица, которое его совершило по видам преступлений в сфере информационных технологий.

Найденные правила, которые связывают предикторные переменные с объясняемой переменной выражаются в терминальных вершинах.

В числовом виде структуру дерева решений можно представить в следующей таблице (табл. 1).

Таблица 2. Абсолютные и относительные веса распределения данных в третьей терминальной вершине

Род деятельности субъекта		Распределение данных (абсол.)	Распределение данных (отн. вес)
Номер класса	Специальность		
1	№ 1 Профессиональный программист	6	0,112
2	№ 2 Технический специалист	2	0,044
3	№ 3 Пользователь	23	0,431
4	№ 4 Взаимодействующий персонал	14	0,263
5	№ 5 Руководитель	8	0,152

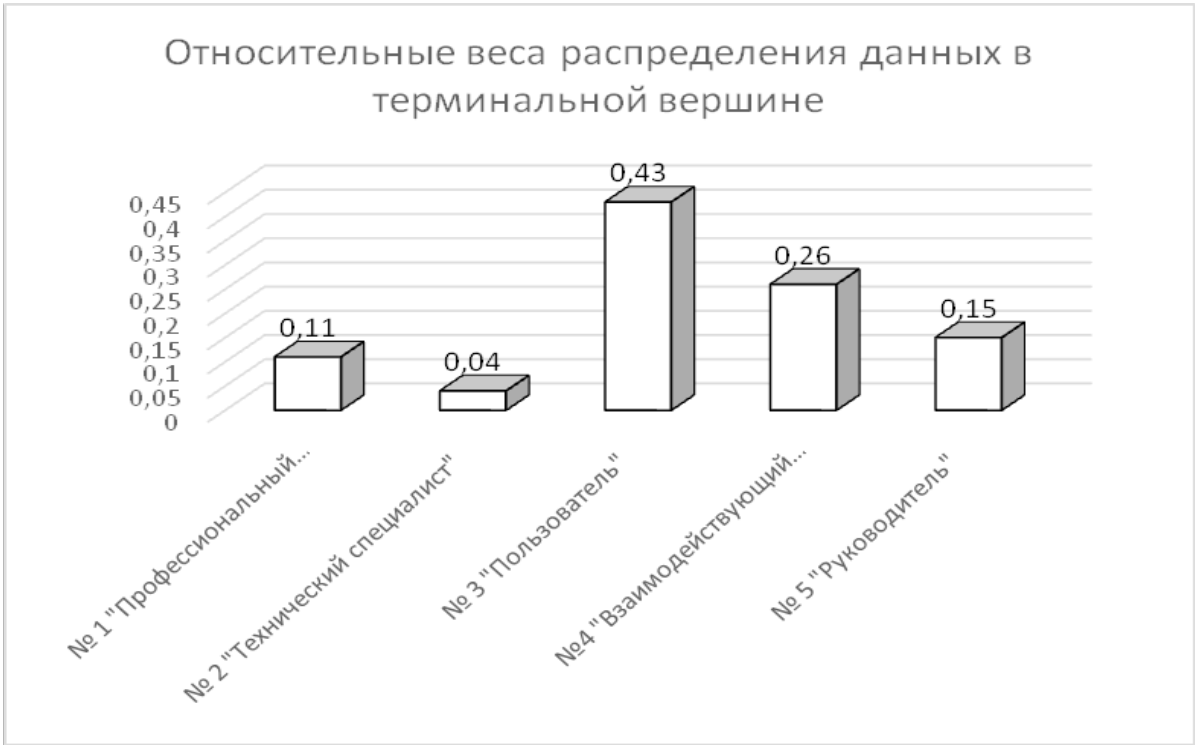

Рис. 1. Относительные веса распределения данных в терминальной вершине

Проведем анализ полученного правила, т.е. выясним, можно ли при таком распределении выделить какое-либо проявление (категорию) зависимой переменной (в нашем примере специальность, рода деятельности лица).

Возможно использование χ^2 К. Пирсона и λ Колмогорова-Смирнова, которые позволяют определить критерии определения расхождения или согласия распределений. В нашем случае наиболее подходящим является использование критерия χ^2 К. Пирсона — непараметрического метода, который позволяет провести оценку статистической значимости различий двух или более переменных.

$$\chi_n^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Критерий Пирсона состоит в расчете суммы отношений квадратов отклонений наблюдаемых O и ожидаемых частот E, к ожидаемой частоте E.

Если частоты соответствуют ожидаемым, то значение критерия будет относительно не большим в виду равенства нулю таких слагаемых. Но если значение критерия оказывается значительным, то это свидетельствует в пользу существенных различий между частотами. Критерий принимает значительную величину, когда появление такого или еще большего значения становится маловероятным. И чтобы рассчитать такую вероятность, необходимо знать распределение критерия при многократном повторении эксперимента, тогда гипотеза о согласии частот верна. Величина χ^2 также зависит от количества слагаемых. Чем их больше, тем большее значение

	Oi	Ei	Oi-Ei	(Oi-Ei)^2	/Ei	Хи-квадрат
	6	10,6	-4,6	21,16	1,996226	25,208
	2	10,6	-8,6	73,96	6,977358	
	23	10,6	12,4	153,76	14,50566	
	14	10,6	3,4	11,56	1,090566	
	8	10,6	-2,6	6,76	0,637736	
сумма	53	53				

Рис. 2. Пример вычисления критерия χ^2с помощью MS Excel

Критические значения критерия Пирсона (χ^2-критерия) для различного уровня значимости q (%) и числа степеней свободы v

Число степеней свободы, v	Уровень значимости, q, %							
	20	10	5	2	1	0,5	0,2	0,1
1	1,642	2,706	3,841	5,412	6,635	7,879	9,550	10,83
2	3,219	4,605	5,991	7,824	9,210	10,60	12,43	13,82
3	4,642	6,251	7,815	9,837	11,34	12,84	14,80	16,27
4	5,989	7,779	9,488	11,67	13,28	14,86	16,92	18,47
5	7,289	9,236	11,07	13,39	15,09	16,75	18,91	20,52

Рис. 3. Критические значения критерия для различного уровня значимости и числа степеней свободы

должно быть у критерия, так как слагаемое внесет свой вклад в суммарный результат.

Для наглядности вынесем в отдельную таблицу (табл. 2) абсолютные и относительные веса распределения данных в третьей терминальной вершине табл. 1.

Относительные веса распределения данных в терминальной вершине три (табл. 1, третья строка) отобразим в отдельной диаграмме (рис. 1).

Результаты исследования

Основываясь на критических значениях критерия Пирсона χ^2 = 18,47 (рис. 3) для степени свободы v = 4 и уровне значимости при распределении частот q = 0,1% вычисленный χ^2 = 25,208 (рис. 2).

Сопоставив значение вычисленного χ^2 с критическим значением для рассматриваемого уровня значимости и числа свободы (рис. 3), можно сделать вывод о том, что в рассматриваемой вершине распределение не равномерное, следовательно, лицо со специальностью № 3 (пользователь) с высокой долей вероятности может совершить противоправное деяние.

В дальнейшем, исключив из распределения частоту, соответствующую специальности № 3, мы сможем вычислить вероятность совершения противоправного деяния лицом со специальностью № 4 (взаимодействующий персонал). Для этого необходимо вычислить вновь χ^2.

При новом уровне значимости q = 5% и степени свободы v = 3 определяем по таблице критическое значение критерия Пирсена (рис. 3), который соответствует χ^2 = 7,8. По описанной ранее методике вычисляем χ^2, который соответствует χ^2 = 10,0. Как и в первом вычислении, сопоставив значение вычисленного хи-квадрата с критическим значением для рассматриваемого уровня значимости и числа свободы, можно сделать вывод о том, что в рассматриваемой вершине распределение также не равномерное, и лицо со специальность № 4 так-

же может совершить противоправное деяние. Распределение вычисленных частот по остальным специальностям носит случайный характер.

Выводы

Рассмотренный пример построения дерева решений для построения типовой версии о специальности субъекта преступления по имеющимся у правоохранительных органов следам, демонстрирует возможность его использования в практической деятельности. При выявлении вида способа совершения противоправного деяния можно предполагать о специальности лица, его совершившего.

Иные особенности способа совершения противоправного деяния, такие как предмет преступления, характер последствий какой-либо новой информации не несут.

Заключение

С целью построения типовых версий возможно использование деревьев решений, при помощи которых осуществляется анализ структуры эмпирических данных, выявление устойчивых связей между способом совершения противоправного деяния и следовой картиной конкретной ситуации, что в последующем поможет разработать информационную систему поддержки принятия решений для нужд правоохранительных органов.

ЛИТЕРАТУРА

1. Умнов А. Е. Методы математического моделирования: Учебное пособие.— М.: МФТИ, 2012.— 295 с.
2. Гудков А. А. Автоматизированная система мониторинга социально-экономической сферы региона на основе технологий обнаружения знаний в базах данных: автореф. дис. канд. тех. наук.— Пенза.— 2008.— 21 с.
3. Применение байесовых сетей в задачах анализа внутренних угроз информационной безопасности the use of bayes / Карпычев В. Ю., Сычев В. М. // Вестник Воронежского института МВД России. 2015. № 1.— С. 125–128.
4. Булычев А. В. Системный подход к анализу скрытых закономерностей в больших массивах слабоструктурированных данных: автореф. дис. канд. тех. наук.— Москва.— 2010.— 23 с.
5. Вахитов А. Р. Математическое и программное обеспечение системы оперативной обработки и интеллектуального анализа данных, использующей нечеткую логику: автореф. дис. канд. тех. наук.— Томск.— 2010.— 26 с.
6. Клир Дж. Системология. Автоматизация решения системных задач. Пер. с англ.— М.: Радио и связь, 1990.— 234 с.
7. Миркин Б. Г. Введение в анализ данных. Учебник и практикум для бакалавриата и магистратуры.— М. ЮРАЙТ, 2014.— 174 с.
8. Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. Classification and regression trees. Monterey, CA: Wadsworth Inc, 1984.— 366 p.