УДК 343.9

С. Н. Нефедов
кандидат технических наук, доцент
НПЦ Государственного комитета судебных экспертиз Республики Беларусь
г. Минск, Беларусь
E-mail: *nefedov@sudexpertiza.by*

В. А. Пархименко
кандидат экономических наук, доцент
E-mail: *parkhimenko@bsuir.by*
М. М. Татур
доктор технических наук, профессор
E-mail: *tatur@bsuir.by*

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Беларусь

Применение методов интеллектуального анализа данных в криминалистике и судебной экспертизе

В статье рассматриваются возможности и существующие подходы к применению методов интеллектуального анализа данных в криминалистике и судебной экспертизе. Кратко изложены основные задачи, решаемые средствами Data Mining & Knowledge Discovery: кластеризация, классификация, регрессия, поиск ассоциативных правил, ранжирование. Предлагаются организационные меры по активизации внедрения данных методов анализа в правоохранительную деятельность.

Ключевые слова: криминалистика, интеллектуальный анализ данных, выявление скрытых зависимостей, кластеризация, классификация, регрессия, поиск ассоциативных правил, ранжирование.

Проникновение информационных технологий (ИТ) во все сферы общественной жизни выразилось, помимо прочего, в фиксации огромного количества разнообразных фактов человеческой деятельности (финансовых транзакций, телефонных звонков и SMS-сообщений, данных фото- и видеорегистрации, происшествий различного рода и т. п.), которые хранятся в форме различных баз данных. Эти базы данных содержат огромное количество информации, которая может быть очень полезной (иногда эта информация может быть исчерпывающей) для решения многих прикладных задач. Однако необходимая (в конкретной ситуации) информация находится в огромном массиве «ненужных» данных, в ряде случаев она может содержаться в неявном виде (например, связи между некоторыми показателями, временные и пространственные зависимости и т. д.). Найти нужную информацию (извлечь из базы данных) часто оказывается невозможным без применения современных методов интеллектуального анализа данных и высокопроизводительных компьютеров. Совокупность этих методов в настоящее время называют Data Mining & Knowledge Discovery (DM&KD)[1], они все более широко применяются в различных отраслях знаний и прикладных сферах.

[1] В англоязычной литературе вместо термина «интеллектуальный анализ данных» обычно используется термин Data Mining – «Добыча данных» (дословный перевод), который часто понимают как «добычу полезных ископаемых», а поиск закономерностей в огромном наборе фактических данных действительно сродни этому процессу. А также близкий термин Knowledge Discovery in Databases – «Обнаружение знаний в больших базах данных». Либо оба термина используют совместно – DM&KD. Такой вариант, в последнее время, часто используют в публикациях на русском языке.

По данной тематике написано ряд обстоятельных книг (учебников и монографий), некоторые из которых уже стали хрестоматийными, отметим лишь одну – [1]. К сожалению, публикаций русскоязычных авторов, особенно учебной литературы, явно недостаточно. Как исключение, отметим книгу В. А. Дюка и А. П. Самойленко [2], но она была издана в 2001 г. В последнее время появились переводы интересных книг зарубежных авторов [3; 4], однако проблема эффективного практического использования аппарата DM&KD продолжает оставаться актуальной.

Data Mining & Knowledge Discovery как методология выявления скрытых зависимостей

DM&KD – это междисциплинарная методология (совокупность методов, технологий и алгоритмов) интеллектуального анализа данных с целью обнаружения скрытой и нетривиальной информации, полезной для принятия решений.

Главная ценность DM&KD – это практическая направленность данной технологии, путь от сырых данных к конкретному знанию, от постановки задачи к готовому приложению, при поддержке которого принимают решения.

Долгое время основным инструментом анализа данных была математическая статистика, а также средства оперативной аналитической обработки данных (online analytical processing – OLAP), которые далеко не всегда позволяют успешно решать такие задачи. Методология DM&KD является развитием этих методов, которая сейчас рассматривается как самостоятельное мультидисциплинарное научное направление. Она базируется, помимо математической статистики и OLAP, на подходах и методах машинного обучения, искусственного интеллекта, проектирования и управления базами данных, а также других смежных областей ИТ (рисунок 1).

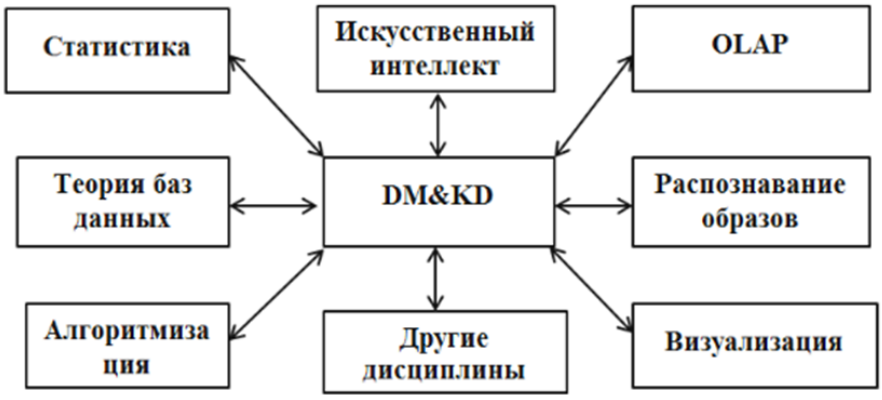

Рисунок 1 – Мультидисциплинарность DM&KD

Цель DM&KD в общем виде формулируют как «обнаружение скрытых закономерностей». В англоязычной литературе в этом контексте речь ведут об обнаружении: скрытых образов или структур (hidden patterns); скрытой информационной ценности (hidden value); скрытых тенденций или трендов (hidden trends); скрытых связей или зависимостей (hidden relationships); скрытых закономерностей (hidden regularities). Представляется целесообразным использовать в качестве основного русскоязычного эквивалента термин *«скрытые закономерности»*.

В данном контексте «скрытый» понимается не как специально спрятанный или зашифрованный (скрываемый с умыслом), а как неочевидный, т. е. не обнаруживаемый явно, прежде всего, по причине большого количества первичных данных.

Закономерность, в соответствии с классическим определением, – это необходимая, существенная, постоянно повторяющаяся взаимосвязь явлений реального мира. Таким образом, еще не известные для исследователя взаимосвязи явлений и объектов реального мира следует называть скрытыми закономерностями.

По всей видимости, случаи существования подобных скрытых закономерностей можно свести к двум типовым, обращая внимание на временной и ресурсный аспекты проблемы:

1. Скрытые до настоящего времени, т. е. пока еще не обратившие на себя внимание[2]. В этом случае вычислительных мощностей, аналитических инструментов, интеллектуальных ресурсов, как правило, достаточно для обнаружения закономерности, но она не обнаруживается, так как просто-напросто такая задача до настоящего момента не ставилась. Например, не рассматривались соответствующие версии.

2. Скрытые, потому что выходят за границы интеллектуальных способностей человека-аналитика («невооруженного» человеческого мозга) и/или наличных в настоящий момент инструментов анализа, вычислительных ресурсов и программных средств.

В DM&KD в первую очередь, по мнению авторов, речь идет о втором случае, т. е. выявление тех закономерностей, которые являются скрытыми потому, что массив анализируемых данных огромен по своему размеру и постоянно растет, содержит множество информационных «шумов», пространство признаков рассматриваемых объектов многомерно и неоднородно. Здесь идет речь не столько о естественной сложности для человека анализировать такую «объемную», «зашумленную» и «многомерную» информацию, сколько о том, что в настоящее время существует большой (быть может – гигантский) разрыв между потоком данных, поступающих от различных источников, и технологиями их обработки и интерпретации, которые применяются для интеллектуальной поддержки принятия решений конечными потребителями (органами государственного управления, силовыми структурами, хозяйствующими субъектами).

Основные типовые задачи Data Mining & Knowledge Discovery

В научной литературе существуют разные определения типов задач, решаемых с помощью методов и алгоритмов DM&KD, однако к числу основных следует отнести: классификацию, кластеризацию, поиск ассоциативных правил, ранжирование, прогнозирование и регрессию. Данные задачи могут считаться типовыми задачами и формулируются в абстрактном виде, подлежат формальному решению методами DM&KD независимо от конкретной предметной области и конкретной прикладной проблемы.

Рассмотрим кратко каждую из задач и приведем некоторые возможные прикладные проблемы, потенциально сводимые к той или иной задаче.

Задачи классификации и кластеризации очень близки, с их использованием множество объектов разбивается на подмножества в зависимости от общности признаков.

Классификация – наиболее простая и распространенная задача DM&KD, которая заключается в системном распределении изучаемых предметов, явлений, процессов по классам (родам, видам, типам), на основе каких-либо существенных признаков.

В ходе решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных – классы; по этим признакам новый объект можно отнести к тому или иному классу. Таким образом, процесс классификации состоит из двух этапов:

1. Конструирование модели: описание множества предопределенных классов и выделение классификационных признаков.

2. Использование модели: классификация новых неизвестных объектов.

В качестве примера рассмотрим разделение объектов (физических лиц) на классы (рисунок 2) в зависимости от уровня доходов (горизонтальная ось) и расходов (вертикальная ось).

На рисунке 2а объекты разделены на два класса только по одному признаку – уровню доходов. Лица с высокими доходами обозначены прямоугольниками, а с низкими – кружками.

На рисунке 2б объекты разделены на три класса с учетом обоих параметров. Между двумя штриховыми линиями – физические лица, у которых уровень расходов примерно со-

[2] В философской терминологии это еще «неопредмеченные» закономерности, т. е. они объективно существуют, но не как предмет человеческой деятельности и мышления.

ответствует уровню доходов. Выше верхней штриховой линии – лица с «подозрительными расходами» (расходы выше доходов), они могут заинтересовать правоохранительные органы. Причем граница между классами может быть нечеткой (в данном примере три объекта находятся на границе разделения классов). Ниже нижней штриховой линии – физические лица, потенциально имеющие накопления (расходы существенно ниже доходов), они могут быть объектами преступных посягательств.

Приведенный пример весьма прост, он иллюстрирует сущность задачи классификации. В реальных ситуациях количество параметров и анализируемых объектов настолько велико, что их невозможно решить без применения методов DM&KD и соответствующих технических средств.

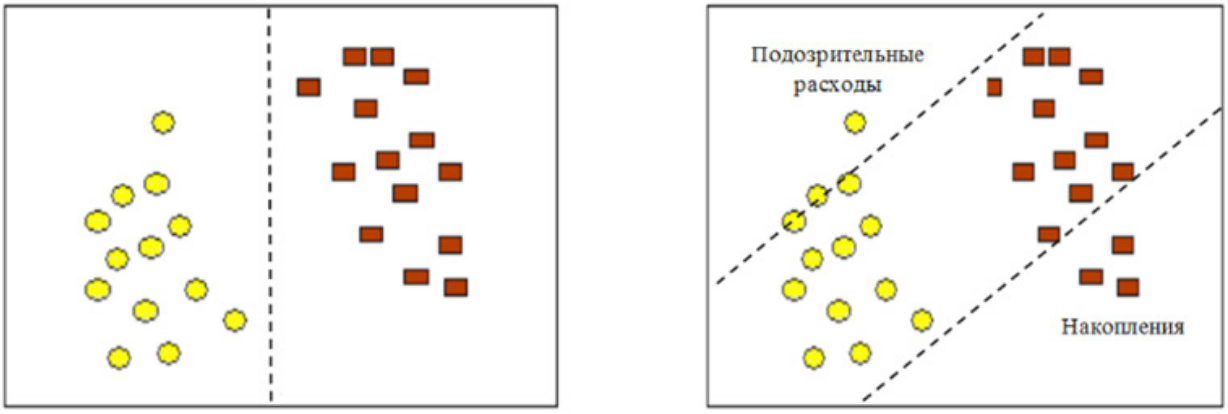

Рисунок 2 – Классификация объетов в зависимости от уровня доходов (горизонтальная ось) и уровня расходов (вертикальная ось)

Кластеризация – это группировка объектов по близости или схожести их характеристик (признаков). Кластеризация является логическим продолжением идеи классификации. Эта задача более сложная, особенность кластеризации заключается в том, что кластеры объектов и их количество изначально не предопределены. Это происходит в ходе решения задачи.

Например, в качестве объектов могут быть рассмотрены конкретные преступления с такими атрибутами, как тип (воровство, кража, грабеж и т. п.), время суток, место, специфика обстоятельств преступления и т. п. Далее с помощью кластеризации могут быть выделены зоны повышенной криминогенной опасности (hotspots) и определен их класс (например, опасность карманных краж, наркомания, мошенничество с подделкой документов и т. д.), с целью дальнейшего их закрепления за конкретными правоохранительными подразделениями, их оснащения и специализации.

Поиск ассоциативных правил – поиск всех значимых зависимостей между признаками объектов, что в данном случае сводится к анализу частоты совместной встречаемости объектов, событий или их атрибутов (признаков и характеристик) с формулировкой итоговых правил по типу «если…, то…». В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных.

Отличие ассоциации от двух предыдущих задач DM&KD заключается в том, что поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно.

Последовательность, или последовательная ассоциация позволяет найти временные закономерности между транзакциями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени (т. е. происходящими с некоторым временным интервалом). Другими словами, последовательность определяется

высокой вероятностью цепочки связанных во времени событий. Фактически, ассоциация является частным случаем последовательности с временным лагом, равным нулю. Эту задачу DM&KD также называют задачей нахождения последовательных шаблонов (sequential pattern). Правило последовательности: после события X через определенное время произойдет событие Y и т. д.

Например, автоматическое выявление финансовых операций по «отмыванию» денег или незаконной деятельности в Интернет может формулироваться как задача последовательной ассоциации: исходя из данных по характеристикам уже выявленных преступлений указанного вида, формируется шаблон, который позволяет автоматически определить для новой финансовой операции или интернет-транзакции, является ли она потенциально незаконной или не является.

Близкой задачей является **ранжирование**, т. е. упорядочение совокупности объектов в соответствии с некоторым критерием или системой критериев.

Ранжирование широко используется в разных областях человеческой деятельности прямо (например, рейтинг музыкальных хитов – хит-парад) или косвенно (при принятии любого решения, требующего выбора среди альтернативных вариантов действий, версий преступлений).

В результате ранжирования исходного множества объектов мы получаем новое множество, путем перестановки объектов исходного множества от «лучшего» до «худшего», от наиболее вероятного события до практически невозможного, и т. д.

Ранжирование, в большинстве случаев, можно рассматривать как способ сведения многокритериального сравнения множества объектов друг с другом к сравнению и упорядочиванию по одному единственному критерию. Этот критерий либо может быть отобран из существующих, либо сконструирован как их функция (например, биржевые индексы). Именно это является объектом внимания в DM&KD.

Задачи **прогнозирования** используются для нахождения оценочных значений пропущенных или же будущих численных показателей на основе зафиксированных данных. Для решения таких задач широко применяются методы математической статистики, один из наиболее распространенных методов – это регрессия.

Регрессия – количественная оценка статистической связи между двумя признаками однородных объектов. В ряде случаев находят зависимость между большим количеством параметров (множественная регрессия).

Например, выявление зависимостей на макроуровне – зависимость между средней заработной платой и числом экономических преступлений.

Помимо указанных задач, также говорят о визуализации, выявлении аномалий и т. п. Подобные задачи либо носят подчиненный (второстепенный) характер, либо напрямую могут быть сведены к одной из типовых задач. В практических приложениях обычно используют несколько методов, поэтому говорят о методологии DM&KD.

Этапы процесса применения Data Mining & Knowledge Discovery

На практике для получения значимых и релевантных для принятия управленческих решений результатов крайне значимым вопросом является не столько то, что методы DM&KD применяются для решения задачи, а то, как они применяются.

В общей процедуре применения DM&KD для решения задачи из любой предметной области можно выделить следующие четыре этапа.

1. **Постановка задачи,** т. е. описание прикладной проблемы из предметной области в терминах типовых задач DM&KD (например, как задачу ранжирования потенциальных подозреваемых на основе неполных свидетельских показаний).

Нередко именно формулировка задачи оказывается самым сложным этапом при реализации процесса анализа для последующего принятия решения, поскольку далеко не все закономерности в данных очевидны с первого взгляда.

2. **Сбор и подготовка данных,** т. е. формирование наборов данных (datasets) в том виде, который требуется для корректной работы методов DM&KD и получения значимых результатов.

Применение методов DM&KD оправданно при наличии достаточно большого количества данных, в идеале – содержащихся в корректно спроектированном хранилище данных (собственно, сами хранилища данных обычно создаются для решения задач анализа и прогнозирования, связанных с поддержкой принятия решений). Структура данных в хранилище должна проектироваться таким образом, чтобы выполнение запросов осуществлялось максимально эффективно.

3. **Процессинг, или DM&KD в узком смысле,** т. е. использование одного из существующих (или модифицированных, или новых) многочисленных алгоритмов для непосредственного интеллектуального анализа данных. Это самый технический и наиболее абстрактный, мало зависящий от предметной области этап. К основным методам (алгоритмам) DM&KD, которые используют для решения рассмотренных выше задач, обычно относят: искусственные нейронные сети; деревья решений; символьные правила; методы ближайшего соседа и k-ближайшего соседа; метод опорных векторов; байесовские сети; линейную регрессию; корреляционно-регрессионный анализ; иерархические и неиерархические методы кластерного анализа, методы поиска ассоциативных правил; метод ограниченного перебора; эволюционное программирование и генетические алгоритмы; разнообразные методы визуализации данных и др.

4. **Оценка, интерпретация и использование результатов,** т. е. проверка валидности результатов и их «перевод» на язык предметной области для принятия решения в рамках той прикладной проблемы, которая послужила исходной причиной для проведения анализа.

Отметим, что с точки зрения авторов, важнейшую роль в перечисленных этапах занимают этапы 1 и 4, на этих этапах основную роль играют специалисты прикладной сферы. Однако они должны иметь определенные знания методологии DM&KD для квалифицированной постановки задачи и корректной интерпретации результатов. Значим и этап 2. А вот этап 3, который, по сути, и представляет собой непосредственную реализацию методов DM&KD, носит все-таки подчиненный характер. На этапах 2 и 3 основную роль играют специалисты ИТ-сферы.

Поясним этот момент подробнее.

Аппарат DM&KD известен уже относительно давно и с точки зрения математики достаточно хорошо описан. Обычно к нему относят широкий перечень методов, начиная с некоторых разделов статистики (регрессия, байессовский классификатор), конечно же, включая методы кластеризации, классификации, ассоциативного поиска, и заканчивая представлением знаний в виде семантических (как вариант, миварных) сетей с процедурами логического и/или нечеткого вывода. Нельзя не отметить и аппарат нейронных сетей (NN), который направлен на решение того же круга задач (часто NN представляют как подмножество методов DM&KD). И, наконец, появление и бурное развитие сверточных NN и глубинного обучения (Deep Learning) обещают решить все проблемы… но, как правило, не решают.

В большинстве случаев провалы связаны с банальной субъективной переоценкой возможностей математики. Другими словами, аналитик изучает один или несколько методов DM&KD и/или NN, осваивает работу со специализированными инструментами (например, R, Intel® Data Analytics Acceleration Library, WEKA 3, SPSS и т. п.), «играется» с открытыми репозиториями (например, UCI Machine Learning Repository) и полагает, что этого достаточно, чтобы решить любую прикладную задачу. Как показывает опыт, с таким багажом можно решить лишь примитивные задачи, т. е. вскрыть зависимости, лежащие на поверхности. Зачастую, подобные зависимости вскрываются умозрительно, а затем аналитик демонстрирует, что и математический аппарат «увидел» это. Теоретическая наука такие результаты принимает, но практике от таких результатов мало толку.

Полагаем, что в рамках решения прикладных задач с помощью DM&KD всегда нужно исходить из некоторых общих принципов.

1. Важно корректно представить решаемую прикладную задачу в виде последовательности (цепочки) процедур DM&KD. При этом, собственно, формальный алгоритм DM&KD (например, *k*-средних, *n*-ближайших соседей и т. п.) может занимать не более 20–40%, а следовательно, и определять результат лишь частично.

2. Важно правильно подготовить исходные данные для формальной обработки. Неполные, «зашумленные» или иным образом неправильно подготовленные данные могут свести к нулю эффективность самых совершенных методик обработки.

3. Большинство сложных для решения прикладных задач содержат, в том числе, неформальные процедуры, поэтому попытки их формального решения обречены на провал.

4. Ряд сложных для решения задач будет представлен итерационными процедурами формальных алгоритмов классификации, кластеризации, ранжирования с последовательным приближением к оптимальному решению.

5. Важно помнить, что анализ данных осуществляется в многомерном пространстве, т. е. существует проблема интерпретации результата. Надо иметь в виду, что поиск неочевидных зависимостей может привести к «открытию» несуществующих связей и закономерностей. Всегда вероятны ошибки первого и второго рода и исследователь должен держать ситуацию под контролем.

Основные направления использования методов и систем DM&KD в борьбе с преступностью

Перенося озвученную логику на предметную область борьбы с преступностью, скажем, что целью DM&KD будет выявление и анализ скрытых закономерностей в данных, связанных с фактами преступлений и противоправных действий, для принятия решений, направленных на предотвращение и/или раскрытие преступлений.

Частными целями DM&KD в области борьбы с преступностью могут быть: выявление общих трендов и закономерностей в сфере преступности; прогнозирование преступлений; объяснение преступных явлений и поведения преступников; автоматизация отдельных этапов аналитической работы криминалистов и замена дорогостоящего труда экспертов; полноценное использование огромного массива накопленной информации о преступлениях; автоматическое выявление незаконной активности в интернете и др.

Анализ академической литературы позволил составить список часто упоминаемых направлений использования методов DM&KD в борьбе с преступностью, в частности методы DM&KD применяются (или могут применяться) для решения широкого круга задач, таких как:

– выявление и прогнозирование территорий/объектов повышенной криминогенности (crime hotspots detection) [5–8];

– оптимизация распределения ограниченных полицейских ресурсов/сил (predictive policing) [5];

– ассоциация преступлений между собой и с конкретным преступником, организацией, транспортным средством (crime linkage) для выявления серийных преступлений и преступных групп, а также ассоциация ранее не раскрытых преступлений с конкретным лицом или группой лиц (crime matching) [5; 6; 9; 10];

– составление профилей преступников (clustering and profiling), в том числе быстрой идентификации преступников, совершающих преступления с одним и тем же «почерком» [11];

– обнаружение зависимостей между характеристиками жертвы преступления, местом, средством и другими обстоятельствами преступления (crime pattern) [6];

– выявление обмана в предоставляемых задержанным данных о себе (criminal identity deceptions detection) за счет сравнения текстовых данных в различных источниках (string comparator techniques) [11];

– ранжирование подозреваемых на основе обработки свидетельских показаний (suspects ranking from multiple witness statements) [10];

– автоматические извлечение структурированной информации из письменных отчетов правоохранительных органов и открытых источников (entity extraction) [11; 12];

– автоматические извлечение характеристик программного обеспечения (extraction of software metrics) с целью выслеживания и установления личностей хакеров [11];

– обнаружение мошенничества (fraud detection) в финансовых транзакциях, страховании, телекоммуникационном секторе и здравоохранении [13];

– обнаружение несанкционированного доступа к компьютерной сети посредством выявления повторяющейся последовательности действий в сетевых транзакциях (sequential pattern mining) [11];

– анализ преступных сетей для выявления связей, ролей, подгрупп в иерархии преступников (criminal network analysis) [11; 14].

Hossein Hassani с коллегами в своей обзорной статье «A review of data mining applications in crime» [14] рассматривают более 100 конкретных примеров[3] использования методов DM&KD.

Помимо большого количества публикаций по применению методов DM&KD в криминалистике и других сферах, в настоящее время разработано достаточно много программных средств, реализующих отдельные алгоритмы. Компания IBM и другие разработчики предлагают различные программные продукты для данной сферы.

Так, IBM I2 COPLINK – модульная программная система от компании IBM, ориентированная на информационную помощь сотрудникам правоохранительных органов разного уровня посредством анализа огромного множества данных, на первый взгляд не связанных между собой. Модуль COPLINK Detect обеспечивает быстрый поиск возможных подозреваемых по всей доступной информации. Модуль COPLINK Activity Correlation идентифицирует подозрительную активность на территории, взятой под наблюдение, исходя из информации, полученной из разных источников. Модуль COPLINK Face Match позволяет идентифицировать подозреваемого по фотографии или фотороботу. Модуль COPLINK Computer Statistics предоставляет различные инструменты для статистической обработки информации и ее визуализации. Модуль COPLINK Incident Analyzer выявляет и визуализирует географическую и временную связь между преступлениями. COPLINK Visualizer реализует визуализацию отношений и ассоциаций между людьми, событиями, местоположениями, организациями и т. п.

Hitachi's Predictive Crime Analytics (PCA) – программное средство, которое на основе анализа в реальном времени различных данных прогнозирует наиболее криминогенные области в городе и визуализирует их на карте.

Financial Crimes Enforcement Network AI System (FAIS) – система, которую использует Сеть по расследованию финансовых преступлений Министерства финансов США.

Насколько известно авторам, в отечественной науке и практике борьбы с преступностью подобного рода методам и системам еще не уделяется полноценного внимания. С одной стороны, эти методы сами по себе относительно новые, с другой стороны, видимо, можно говорить о присущем правоохранительным институтам естественном консерватизме и инерционности. В то же время для активного использования подобных методов далеко не всегда нужно начинать со сложных систем и широкого спектра задач, наоборот, целесообразным представляется поэтапный подход, подразумевающий постепенное, пошаговое внедрение отдельных инструментов для решения некоторых прикладных задач.

В любом случае отправная точка такого внедрения – это понимание фундаментальных основ Data Mining & Knowledge Discovery, о чем шла речь выше.

[3] Отметим ради справедливости, что многие примеры – это лишь академические проекты, потенциально направленные на решение прикладных задач.

Заключение

Полагаем, что озвучивание отдельных, пусть и крайне удачных, примеров из мирового опыта, а также понимание основ DM&KD и готовность использовать эти методы на практике все еще недостаточно для активного внедрения методов DM&KD в отечественную криминалистику. Существует проблема, которая характерна для методологии DM&KD в целом. Суть проблемы состоит в следующем.

Несмотря на то, что в настоящее время существует огромное множество научных статей, монографий, учебников, образовательных курсов, специализированных компаний, профессиональных ассоциаций, а также конференций и других мероприятий, до сих пор является острой проблемой взаимодействия специалистов по DM&KD и специалистов из предметной области: первые, как правило, имеют техническое образование и ориентированы на этап алгоритмического процессинга данных, вторые, как правило, мыслят только в терминах предметной области, не понимая и не имея желания вникать в детали «математики и информатики». Зачастую проблема состоит в том, что ученые не могут объяснить доступным языком специалистам-практикам реальные выгоды от применения интеллектуального анализа данных, а те в свою очередь испытывают недоверие к современным средствам моделирования.

В качестве заключения следует отметить следующее.

Изучение мирового опыта и научных публикаций по интеллектуальному анализу данных дает веские основания полагать, что:

1. Методы DM&KD не только применимы в сфере борьбы с преступностью, но и дают полезные результаты в длинном ряде случаев.

2. В современных условиях всеобщей информатизации значимость DM&KD только возрастает.

3. Следует объединять усилия экспертов (ученых) по DM&KD и специалистов-предметников (криминалистов, криминологов).

4. Не следует рассматривать DM&KD как «волшебную палочку». Без тщательной работы специалистов предметной области значимых и полезных результатов ожидать не приходится.

Авторы полагают, что решение указанной проблемы в такой предметной области, как борьба с преступностью, должно базироваться на широкой популяризации возможностей и достижений DM&KD в среде специалистов-предметников. Авторы планируют подготовить серию статей, в которых постараются в доступной форме рассмотреть основные методы DM&KD и вопросы их практической реализации.

Следующим шагом видится формирование междисциплинарных команд, включающих специалистов ИТ-сферы, криминалистов и экспертов, для выработки направлений работы по решению конкретных прикладных задач предметной области и обсуждения проблемных вопросов, апробации методологии DM&KD в рамках пилотных исследовательских проектов и т. д.

Список использованных источников

1. Data Mining. A knowledge discovery approach / K. Clos [et al.]. – Springer, 2007. – 606 p.

2. Дюк, В. А. Data Mining : учеб. курс / В. А. Дюк, А. П. Самойленко. – СПб. : Питер, 2001. – 368 с.

3. Силен, Д. Основы Data Science и Big Data. Python и наука о данных / Д. Силен, А. Мейсман, А. Мохамед. – Спб. : Питер, 2017. – 336 с.

4. Мюллер, А. Введение в машинное обучение с помощью Python / А. Мюллер, С. Гвидо. – М., 2016–2017. – 393 с.

5. Ellahi, Abida, and Irfan Manarvi. CRIME DATA MINING: AN ANALYSIS OF REAL TIME DATA IN PAKISTAN // Caribbean Journal of Criminology and Public Safety. January&July, 2010. 15(1&2). 195–214.

6. Gupta, Manish, B. Chandra, and M. P. Gupta. Crime data mining for Indian police information system // Proceeding of the 2008 Computer Society of India (2008).

7. Kiani, Rasoul, Siamak Mahdavi, and Amin Keshavarzi. Analysis and prediction of crimes by clustering and classification // International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No. 8, 2015.

8. Nath, Shyam Varan. Crime pattern detection using data mining // Web intelligence and intelligent agent technology workshops, 2006. wi-iat 2006 workshops. 2006 ieee/wic/acm international conference on. IEEE, 2006.

9. Keyvanpour, Mohammad Reza, Mostafa Javideh, and Mohammad Reza Ebrahimi. Detecting and investigating crime by means of data mining: a general crime matching framework // Procedia Computer Science 3 (2011) : 872–880.

10. Porter, Michael D. A statistical approach to crime linkage // The American Statistician 70.2 (2016) : 152–165.

11. Crime data mining: a general framework and some examples / Chen Hsinchun [et al.] // Computer. 37.4 (2004) : 50–56.

12. An experimental study of classification algorithms for crime prediction / Iqbal Rizwan [et al.] // Indian Journal of Science and Technology 6.3 (2013) : 4219–4225.

13. Tasdoven, Hidayet, and Bahadir Sahin. Crime data mining as a decision making tool // International Journal of Public Policy 6.3-4 (2010) : 278–287.

14. A review of data mining applications in crime / Hassani Hossein [et al.] // Statistical Analysis and Data Mining : The ASA Data Science Journal 9.3 (2016) : 139–154.

Дата поступления: 27.11.2017

S. N. Nefedov
Candidate of Technical Sciences, Associate Professor
SPC of the State Forensic Examination Committee of the Republic of Belarus
Minsk, Belarus

U. A. Parkhimenka
Candidate of Economic Sciences, Associate Professor
M. M. Tatur
Doctor of Technical Sciences, Full Professor
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus

Application of data mining techniques in criminalistics and forensics

The article considers the possibilities and existing approaches to the application of methods of data mining in criminalistics and forensics. A brief summary of the main tasks that solved by means of Data Mining & Knowledge Discovery is given: clustering, classification, regression, search for associative rules, ranking. Organizational measures to enhance implementation of data analysis methods in law enforcement are proposed.

Keywords: forensics, data mining, revealing the hidden dependencies, clustering, classification, regression, search for association rules, ranking.