Федеральное государственное автономное образовательное учреждение высшего образования «Уральский федеральный университет имени первого президента России Б.Н. Ельцина»

На правах рукописи

Гибилинда Роман Владимирович

РАЗРАБОТКА АВТОМАТИЗИРОВАННЫХ МЕТОДОВ АНАЛИЗА ВОЗДЕЙСТВИЙ НА ФАЙЛЫ В ЗАДАЧЕ РАССЛЕДОВАНИЯ ИНЦИДЕНТОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

2.3.6. Методы и системы защиты информации, информационная безопасность

АВТОРЕФЕРАТ
диссертации на соискание ученой степени кандидата технических наук

Екатеринбург – 2021

Работа выполнена в учебно-научном центре «Информационная безопасность» Института радиоэлектроники и информационных технологий - РтФ ФГАОУ ВО «Уральский федеральный университет имени первого Президента России Б. Н. Ельцина».

Научный руководитель: кандидат технических наук, доцент,
СИНАДСКИЙ Николай Игоревич

Официальные оппоненты: **БАРАНКОВА Инна Ильинична**, доктор технических наук, доцент, ФГБОУ ВО «Магнитогорский государственный технический университет им. Г.И. Носова», г. Магнитогорск, заведующая кафедрой информатики и информационной безопасности;

ТИТОВ Сергей Сергеевич, доктор физико-математических наук, профессор, ФГБОУ ВО «Уральский государственный университет путей сообщения», г. Екатеринбург, главный научный сотрудник кафедры «Информационные технологии и защита информации»;

СОКОЛОВ Александр Николаевич, кандидат технических наук, доцент, ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)», г. Челябинск, заведующий кафедрой «Защита информации» Высшей школы электроники и компьютерных наук

Защита диссертации состоится «16» ноября 2021 г. в 11:00 часов на заседании диссертационного совета УрФУ 2.3.05.13 по адресу: 620002, г. Екатеринбург, ул. Мира, 19, ауд. И-420 (зал Ученого совета).

С диссертацией можно ознакомиться в библиотеке и на сайте ФГАОУ ВО «Уральский федеральный университет имени первого Президента России Б.Н. Ельцина»: https://dissovet2.urfu.ru/mod/data/view.php?d=12&rid=2855

Автореферат разослан «____» _____ 2021 года.

Ученый секретарь
диссертационного совета Сафиуллин Николай Тахирович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. Вступление в силу Федерального закона № 187-ФЗ «О безопасности критической информационной инфраструктуры Российской Федерации» и других сопутствующих нормативных правовых актов обязало владельцев информационных систем (ИС), входящих в состав критической информационной инфраструктуры (КИИ), применять меры по защите информации, в том числе проводить расследование инцидентов информационной безопасности (ИБ).

Несмотря на существование множества решений, предназначенных для обеспечения требуемого (согласно принятой в организации политики безопасности) уровня безопасности информации, злоумышленники не снижают своей активности по реализации различных видов компьютерных атак, направленных на получение несанкционированного доступа к ИС и нарушение конфиденциальности, целостности и доступности обрабатываемых в ИС сведений.

Для определения действий злоумышленника и последствий возникшего инцидента ИБ проводится расследование инцидента. Результаты расследования могут указывать на не выявленные недостатки системы обеспечения ИБ и действующей политики безопасности организации и составляют основу рекомендаций по совершенствованию мер по защите информации, направленных на недопущение возникновения подобных инцидентов ИБ в будущем.

В процессе расследования инцидента ИБ, определяемого как совокупность[1] событий ИБ, возникает потребность в определении воздействий на информацию, обрабатываемую в ИС.

Хранилищем информации в ИС является файл. Таким образом, для определения воздействий на информацию следует обнаружить и классифицировать воздействия на файлы, предварительно их идентифицировав.

Немаловажным фактором, влияющим на результаты расследования инцидента, является время, поэтому обнаружение и классификация воздействий на файлы должны происходить в сжатые строки. Основой для определения воздействий является информация, содержащаяся в разноформатных массивах данных операционных систем (рисунок 1), которые могут содержать десятки-сотни тысяч записей / полей / иных различных информационных структур.

Автоматизация необходима для ускорения процесса расследования инцидента ИБ, в т.ч. представляемого в виде совокупности воздействий на файлы. Расследование инцидентов может быть разделено на две части: проведение экспресс-анализа массивов данных для обнаружения[2] и классификации[3] воздействий на файлы с предварительной их идентификацией[4], а также детализированный анализ в случае необходимости углубленного изучения последствий инцидента.

---

[1] Стандарт Банка России СТО БР ИББС-1.3-2016 «Обеспечение информационной безопасности организаций банковской системы Российской Федерации. Сбор и анализ технических данных при реагировании на инциденты информационной безопасности при осуществлении переводов денежных средств» [Электронный ресурс]. URL: http://garant.ru/products/ipo/prime/doc/71457690 (дата обращения: 16.04.2021).

[2] Под обнаружением воздействия на файл понимается факт нахождения признаков воздействия в массиве данных компьютерной системы, где произошел инцидент.

[3] Классификацией в рамках работы является процесс отнесения обнаруженного воздействия к классу аномальных или нормальных для определения хода инцидента ИБ.

[4] Под идентификацией воздействий на файлы понимается процесс, в результате которого определяется порядок изменения признаков, характеризующих файл.

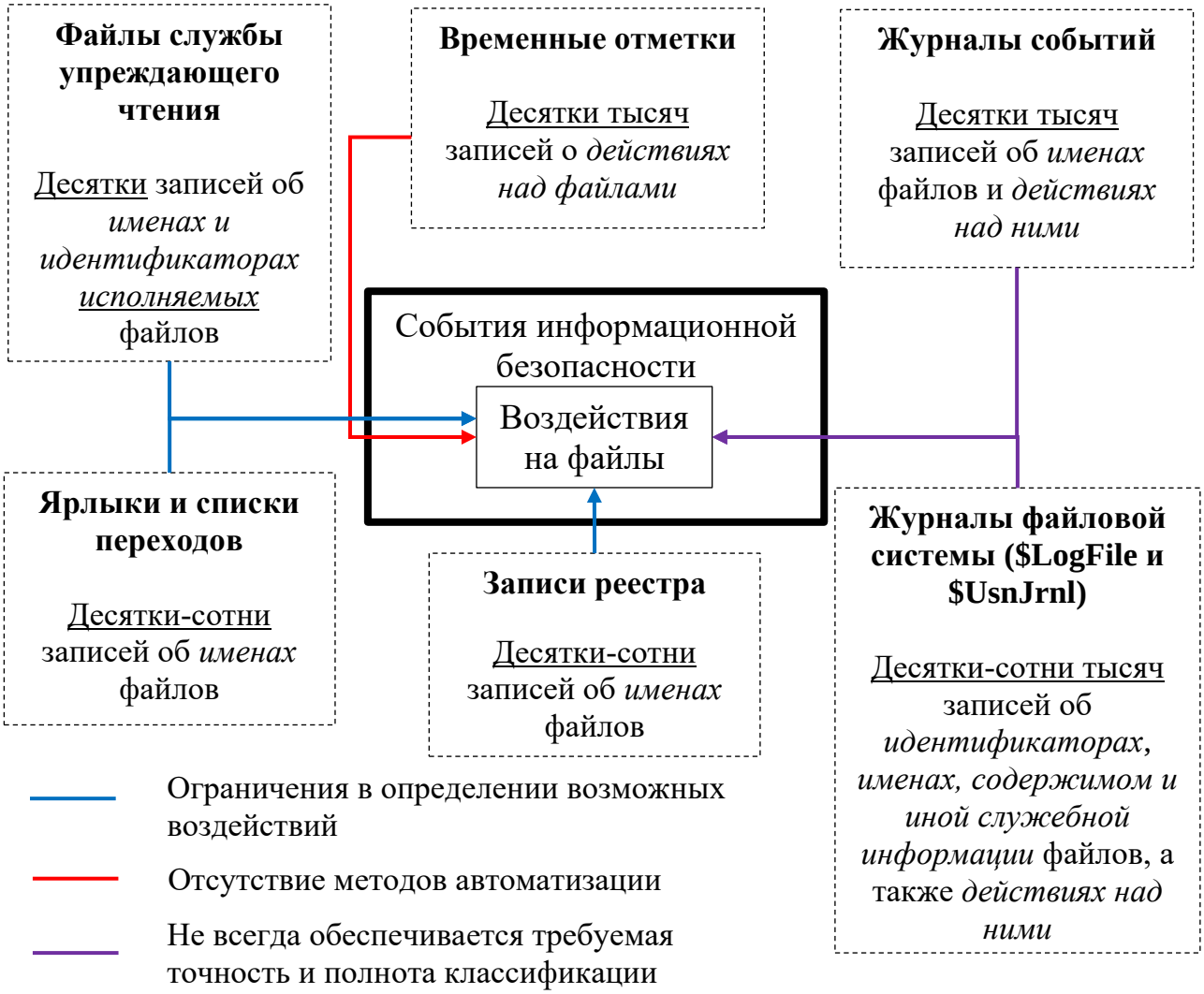

Рисунок 1. Массивы данных и проблемы их анализа

Процесс обнаружения и классификации, проводимый в ходе экспресс-анализа, связан с процедурой сравнения данных с некоторым эталоном. Например, антивирусное программное обеспечение (ПО) обнаруживает вредоносное ПО с использованием баз сигнатур вирусов, система обнаружения атак выявляет компьютерные атаки в сетевом трафике на основании списка правил. Для ускорения процесса обнаружения и классификации заранее подготавливается база данных эталонов воздействий на файлы, чтобы в рамках расследования инцидента произвести сравнение имеющихся данных с эталонными. В качестве эталона выбран т.н. шаблон воздействия на файл – совокупность значений признаков, характеризующих файл и совершаемых файловых операций, являющихся составными частями воздействия. Процесс генерации шаблона основан на предварительной идентификации воздействия и последующей его декомпозиции.

Таким образом, для решения задачи автоматизации процесса обнаружения и классификации воздействий на файлы использование существующих методов анализа данных, применяемых при расследовании инцидентов ИБ, либо не обеспечивает необходимой точности и/или полноты, либо имеет ограничения по набору определяемых воздействий (рисунок 1). Возникла потребность в разработке новых методов анализа данных, учитывающих порядок изменения признаков, характеризующих файлы, в рамках осуществленных воздействий. Для разрешения возникшей потребности был выбран массив данных, обладающий наиболее полной

информацией о воздействиях на файлы – журнал изменений тома $UsnJrnl файловой системы NTFS операционной системы Windows, создан алгоритмический и методологический аппарат автоматизированного анализа воздействий, менее подверженный указанным ранее недостаткам существующих методов, и на его основе разработан специальный комплекс программных средств, позволяющий ускорить процесс расследования инцидентов ИБ [1-8].

Степень разработанности темы исследования. Расследование инцидентов ИБ представляет собой сложный процесс, требующий наличия у специалиста обширных знаний об особенностях функционирования ИС, способах обеспечения ИБ и оценки качества принятых мер защиты, математических методах анализа данных, связанных с инцидентами.

Применению математических методов и алгоритмов, направленных на обеспечение ИБ (в том числе используемых при анализе данных в рамках расследования инцидентов ИБ) посвящено множество научных работ, среди которых необходимо отметить труды С.С. Титова, А.А. Захарова, В.В. Богданова, Ю.Д. Королькова, П.Н. Девянина, И.И. Баранковой, В.А. Баранского, Н.А. Гайдамакина, С.В. Поршнева, В.Н. Зуева, А.Н. Соколова, О.И. Шелухина, Д. Феррейры (D.R. Ferreira), В. Сатиша (V. Sathish), Е. Чуа (E. Chuah), Ф. Юана (F. Yuan), С. Йена (S. Yen), К. Берлина (K. Berlin), Р. Вааранди (R. Vaarandi), Б. Штейна (B. Stein), Х. Штудиавана (H. Studiawan). Авторами разработаны методы анализа информации, в том числе основанные на теории графов, рассмотрено применение метрик в контексте обработки информации, связанной с событиями ИБ, предложены новые и адаптированы существующие методы и алгоритмы машинного обучения для анализа разноформатных массивов данных: сетевого трафика, журналов событий и др. Часть указанных работ содержат количественные оценки результатов работы алгоритмов, что позволяет выбрать наилучший для решения задач анализа данных.

Целью исследования является разработка научно-обоснованных автоматизированных методов расследования инцидентов ИБ, их программных реализаций и методики использования.

Для достижения поставленной цели сформулированы и решены следующие **задачи**:

1. Анализ состояния предметной области и инструментов для автоматизированной идентификации, обнаружения и классификации воздействий на файлы.
2. Разработка математического аппарата для анализа воздействий на файлы, в частности: математической модели процесса идентификации воздействий на файлы; кластеризационного метода идентификации воздействий; метода экспресс-анализа событий ИБ.
3. Создание комплекса программных средств, реализующих алгоритмы разработанных математических методов для анализа воздействий на файлы, направленных, в том числе, на нарушение действующей политики безопасности.

Объект исследования – процесс расследования инцидентов ИБ.

Предмет исследования – автоматизированные методы и алгоритмы расследования инцидентов ИБ.

Научная новизна работы. В рамках проведенного исследования получены следующие новые научные результаты:

1. Разработана модель процесса идентификации воздействий на файлы, основанная на математическом аппарате сетей Петри, позволяющая формализовать

набор признаков, характеризующих файл, для их последующего анализа в рамках расследования инцидентов ИБ (соответствует специальности в части создания событийной модели процесса идентификации воздействий на файлы, в т.ч. направленных на нарушение действующей политики ИБ, применяемой при мониторинге состояния объекта, находящегося под воздействием угроз нарушения его ИБ).

2. Разработан кластеризационный метод идентификации воздействий на файлы, направленных, в том числе, на нарушение действующей политики ИБ (соответствует специальности в части разработки кластеризационного метода идентификации воздействий на файлы с целью создания шаблонов воздействий как нового элемента контроля за влиянием на информацию в рамках процесса управления событиями ИБ).

3. Разработан метод экспресс-анализа событий ИБ, связанных с воздействиями на файлы, позволяющий ускорить процесс обнаружения и классификации воздействий (соответствует специальности в части разработки принципа по созданию нового метода определения нормальных, аномальных и условно аномальных событий ИБ в целях формирования рекомендаций по совершенствованию мер, направленных на обеспечение ИБ).

Теоретическая значимость работы выражается в развитии научно-методического аппарата для анализа воздействий на файлы в рамках расследования инцидентов ИБ.

Практическая значимость работы заключается в разработке комплекса программных средств, обеспечивающего автоматизацию процесса анализа воздействий на файлы, в том числе направленных на нарушение действующей политики ИБ.

Методология и методы исследования. В диссертации представлены результаты исследований, полученные с помощью математического аппарата сетей Петри, теории вероятностей, кластерного анализа и алгоритмов классификации.

Положения, выносимые на защиту:

1. Процесс идентификации воздействий на файлы, направленных, в том числе на нарушение действующей политики ИБ, описывается математическим аппаратом на основе сетей Петри [3].

2. Предложенный кластеризационный метод, в котором используются адаптированные к структуре массивов данных алгоритмы подготовки входной информации и определения оптимального количества кластеров, позволяет автоматизировать процесс идентификации воздействий на файлы [4].

3. Метод экспресс-анализа событий информационной безопасности, связанных с воздействиями на файлы, и реализующий его комплекс программных средств позволяют автоматизировать процесс выявления нормальных[5], аномальных[6] и условно аномальных[7] воздействий на файлы [1].

Границы исследования – операционная система (ОС) Windows с файловой системой (ФС) NTFS и журналом изменений тома $UsnJrnl, являющимся массивом данных о воздействиях на файлы.

---

[5] Нормальным считается воздействие, возникающее в процессе штатного функционирования компьютерной системы.

[6] Аномальным считается воздействие, которое не может возникать в процессе штатного функционирования компьютерной системы и свидетельствует о факте нарушения действующей политики безопасности в отношении файлов.

[7] Условно аномальным считается воздействие, которое может быть отнесено к классу нормальных или аномальных при выполнении дополнительных условий, заданных специалистом, проводящим расследование инцидента ИБ.

Достоверность и обоснованность полученных результатов подтверждается адекватным выбором математического аппарата задачам исследования и результатами экспериментальной апробации предложенных моделей и методов анализа воздействий на файлы.

Апробация исследования. Основные результаты диссертационных исследований докладывались на различных семинарах и совещаниях, а также на международной научной конференции, в том числе:

1. II Всероссийская научная конференция (с приглашением зарубежных ученых) «Фундаментальные проблемы информационной безопасности в условиях цифровой трансформации» (FISP-2020), 30 ноября 2020 г., Россия, г. Ставрополь [7];
2. 16-я Юбилейная международная молодежная научно-техническая конференция «Современные проблемы радиоэлектроники и телекоммуникаций, РТ-2020», 12-16 октября 2020 г., Россия, г. Севастополь [8];
3. 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), 14-15 мая 2020 года, Россия, г. Екатеринбург [2].

Внедрение результатов исследования. Результаты работы используются в ООО «Уральский центр систем безопасности», Екатеринбург, Россия (акт об использовании результатов от 16.03.2021); в Уральском федеральном университете имени первого Президента России Б.Н. Ельцина, Екатеринбург, Россия (акт об использовании результатов от 10.03.2021); в ОКБ «Новатор», Екатеринбург, Россия (акт об использовании результатов от 23.12.2020).

Публикации. По теме диссертации опубликовано 6 научных работ, из них 4 статьи опубликованы в рецензируемых научных журналах и изданиях, определенных ВАК РФ и Аттестационным советом УрФУ, включая 1 статью в издании, индексируемом в международной цитатно-аналитической базе Scopus. Имеются 2 свидетельства о государственной регистрации программы для ЭВМ.

Личный вклад автора. Все результаты и положения, выносимые на защиту, получены лично автором. Все алгоритмы, обсуждаемые в работе, разработаны и экспериментально исследованы автором самостоятельно. Научный руководитель принимал участие в постановке цели и задач исследования, их предварительном анализе, планировании экспериментов. Подготовка к публикации полученных результатов проводилась совместно с соавторами, при этом вклад диссертанта был определяющим.

Структура и объем работы. Диссертация состоит из введения, трех глав, заключения и приложений. Общий объем диссертации составляет 178 страниц, включая 31 рисунок и 23 таблицы. Список использованных источников содержит 151 наименование.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследования, показаны практическая и теоретическая значимость полученных результатов, представлены положения, выносимые на защиту.

Первая глава состоит из шести частей. В **первой части** введены понятия расследования инцидента ИБ, потребности в проведении расследования. Определено,

что в рамках работы под *расследованием инцидента ИБ* понимается реагирование[8] на него, но без осуществления правовой оценки.

Во второй части введены следующие определения: состояние файла, файловая операция и воздействие на файл. В дальнейшем указанные определения использованы по тексту работы.

Инцидент ИБ *A* описывается выражением:

$$A = \left\{ \left\langle S_1, ..., S_p \right\rangle \right\}, \quad (1)$$

где $S_1, ..., S_p$ – события ИБ, p – их количество.

В рамках исследования каждое событие ИБ представляется как совокупность воздействий на файлы и описывается кортежем:

$$S_k = \left\langle \left\langle Y_{fi}^k \right\rangle_{i=1}^{l} \right\rangle_{f=1}^{n}, \quad (2)$$

где Y_{fi}^k – воздействие i на файл f, относящееся к событию ИБ k, l – количество воздействий, n – количество файлов.

В дальнейшем определены какие признаки, характеризующие файл, используются в качестве основы шаблона воздействия на файл. Считается, что произвольный файл f в каждый момент времени t имеет следующие признаки:

- $I_f(t)$ – идентификатор файла – уникальное числовое значение, содержащееся в служебной информации о файле, используемое драйвером ФС для однозначного определения файла;
- $D_f(t)$ – идентификатор родительского каталога файла – уникальное числовое значение, используемое драйвером ФС для установления однозначного соответствия между файлом и каталогом, в котором файл расположен;
- $N_f(t)$ – имя файла – битовая строка, используемая драйвером ФС для представления файла пользователю;
- $C_f(t)$ – содержимое файла – битовая строка, являющаяся информацией, хранимой в файле;
- $X_f(t)$ – иная служебная информация о файле – набор числовых значений, являющихся служебной информацией о файле, зависящий от типа ФС.

По тексту работы указанные величины используются без указания времени для удобства восприятия. Признаки, характеризующие файл f, образуют вектор $V_f(t)$, описываемый выражением:

$$V_f(t) = V_f = \left\{ I_f, D_f, N_f, C_f, X_f \right\}. \quad (3)$$

При рассмотрении некоторого файла f в рамках расследования инцидента ИБ определяются значения компонентов вектора V_f, которые описывают состояние файла на момент события ИБ.

Определение 1. *Состояние файла – значения компонентов вектора V_f признаков, характеризующих файл, присущих файлу f в определенный момент времени t.*

Изменение значений компонентов вектора V_f происходит не спонтанно, а в соответствии с процессом, называемым файловой операцией. По тексту работы файловая операция обозначена символом *O*. Пусть в момент времени t_1 начат процесс, в

[8] ГОСТ Р ИСО/МЭК ТО 18044-2007. «Информационная технология. Методы и средства обеспечения безопасности. Менеджмент инцидентов информационной безопасности». — М. : ФГУП «Стандартинформ», 2007. — 50 с.

результате которого произошло изменение значений компонентов вектора V_f, и к моменту времени окончания процесса t_2 файл, описываемый начальным состоянием $V_{f1} = V_f(t_1)$, перешел в конечное состояние $V_{f2} = V_f(t_2)$, причем $V_{f1} \neq V_{f2}$.

Определение 2. *Файловая операция O – процесс модификации значений компонентов вектора V_f признаков, характеризующих файл, приводящий к переходу файла от одного состояния к другому.*

Примерами файловых операций являются: создание, удаление, переименование, изменение содержимого, изменение служебной информации, перемещение, копирование.

При осуществлении воздействий на файлы возможно проведение как одной, так и нескольких файловых операций, которые приводят к появлению нескольких состояний файла, являющихся последовательным переходом от состояния до начала воздействия к окончательному состоянию.

Определение 3. *Воздействие на файл – совокупность файловых операций, связанных по назначению, разделенных по времени и приводящих к изменению состояний одного или нескольких файлов*.

Определив признаки, характеризующие файл, и их взаимосвязь с воздействиями на файлы, были рассмотрены массивы данных, используемые в качестве источника информации о событиях ИБ при расследовании инцидентов. Обзор массивов приведен в **третьей части** первой главы: временные отметки, журналы событий ОС Windows, ярлыки последних открытых файлов и списки переходов, файлы службы упреждающего чтения Superfetch, записи реестра, журналы ФС NTFS $LogFile и $UsnJrnl. Приведено описание ряда существующих исследований, использующих данные массивы. Основным недостатком указанных массивов данных является отсутствие полного набора признаков, характеризующих файл, что, в свою очередь, не позволяет однозначно идентифицировать воздействие на файл. Анализ показал оптимальность выбора журнала изменений тома $UsnJrnl как массива данных о воздействиях на файлы, обладающего наиболее полной информацией о признаках, характеризующих файл.

В четвертой части рассмотрена возможность применения известных методов анализа применительно к массивам данных: деревья решений, метод главных компонент, нейронные сети, логистическая регрессия, метод опорных векторов, кластерный анализ. В рамках решения задач по определению воздействий на файлы для автоматизации расследования инцидентов ИБ используются как методы уменьшения размерности данных, так и методы классификации. В ходе рассмотрения сделан вывод о том, что среди методов, применяемых для анализа данных и сокращения объема обрабатываемой информации, перспективно использование кластерного анализа. В процессе исследования существующих процедур автоматического подбора оптимальных параметров кластеризации выявлены недостатки и сделан вывод о необходимости разработки нового метода подбора параметров.

Сложности, возникающие при применении алгоритмов классификации, связаны с уменьшением точности получаемых результатов ввиду учета незначимых признаков в анализируемой информации, а также с ограниченным использованием методов для определения лишь наличия аномалий в массиве данных. Любое изменение в наборе признаков влечет за собой процедуру переобучения классификатора, которая может оказаться ресурсоемкой и затратной по времени. Таким образом, в процессе классификации воздействий на файлы следует динамически определять значимые признаки, по которым анализируется информация, в целях повышения точности получаемых результатов. Решением этой задачи являются шаблоны воздействий на файлы, представляющие собой декомпозицию данных о воздействии.

В связи с введением новой структуры данных (шаблонов) разработан метод, который позволяет проводить классификацию воздействий на файлы с учетом имеющихся шаблонов и обеспечивать экспресс-анализ событий ИБ.

В пятой части уделено внимание методам и средствам моделирования. С учетом дискретного характера изменения признаков, характеризующих файл, и наличия взаимосвязи между значениями признаков и файловыми операциями, в качестве основы для построения модели идентификации воздействий на файлы выбран математический аппарат сетей Петри.

В шестой части представлен перечень ПО, используемого при расследовании инцидентов ИБ. Сделан вывод о необходимости разработки комплекса программных средств, позволяющих проводить экспресс-анализ воздействий на файлы.

На основании результатов анализа состояния предметной области **в седьмой части** сформулированы цель и задачи исследования.

Во второй главе рассмотрен порядок изменения значений признаков, характеризующих файл, при совершении файловой операции, как составляющей воздействия на файл. Определение зависимости в изменении признаков, характеризующих файл, от совершенной файловой операции произведено с помощью алгоритма (рисунок 2). Представленный алгоритм не учитывал сложную природу компьютерной системы, в которой выполняются различные по типу и назначению процессы: последовательные и параллельные, синхронные и асинхронные. Среди множества процессов встречаются такие, которые связаны с воздействиями на файлы, результат работы которых необходимо фиксировать. Для моделирования процесса идентификации как простых, так и сложных воздействий на файлы, а не только файловых операций, найден математический аппарат, позволяющий учесть множество возможных комбинаций состояний файлов. Для формализации алгоритма классификации файловой операции (рисунок 2), являющегося базовым для построения событийной модели процесса идентификации воздействий на файлы, применен математический аппарат сетей Петри (рисунок 3).

Сеть Петри, которая позволяет осуществить принятие решений в соответствии с условными переходами, должна обладать т.н. маркировкой – способом определения порядка задействования переходов. Такая сеть задается пятью параметрами – P, T, E, Q, μ, где: P – множество позиций; T – множество переходов; E – входная функция ($T_j \rightarrow P_j$,); Q – выходная функция (Q: $P_j \rightarrow T_j$); μ – маркировка сети фишками[9].

Рассматриваемая событийная модель процесса идентификации воздействий на файлы дополняется двумя параметрами: δ, который является внешним по отношению к сети «накопителем» файловых операций, и временной задержкой $\tau = t_2 - t_1$, где t_1 соответствует времени до начала файловой операции, а t_2 – времени ее окончания. Таким образом, событийная модель, основанная на сети Петри, обозначается как $Net = (P, T, E, Q, \mu, \delta, \tau)$.

В конечное множество позиций $P = P_{fs} \cup P_{al} \cup P_{fin} \cup P_{add}$ включены:

- позиции состояний файла P_{fs} ($P_1 \in \langle I_{f1}, D_{f1}, N_{f1} \rangle$, $P_2 \in \langle I_{f2}, D_{f2}, N_{f2} \rangle$, P_6);
- позиции состояний алгоритма (рисунок 2) P_{al} (P_3 — P_5, P_{19} — P_{28});
- позиции классификации файловых операций P_{fin} (P_{29} — P_{36});
- вспомогательные[10] позиции P_{add} (P_7 — P_{18}, P_{37}, P_{38}).

---

[9] Под фишкой понимается способ определения порядка задействования переходов.

[10] Предназначенные для корректного функционирования сети Петри.

В конечное множество переходов $T = T_{fin} \cup T_{al} \cup T_{add} \cup T_{delay} \cup T_{fo}$ включены:

- переходы «индикации» получения значений состояний файла T_{fin} (T_1, T_2);
- переходы алгоритма (рисунок 2) T_{al} (T_3 — T_8, T_{13} — T_{16});
- вспомогательные переходы T_{add} (T_9 — T_{12});
- переход имитации временной задержки при совершении файловых операций T_{delay} ($T_{25} \in \tau : \tau = t_2 - t_1$);
- переходы классификации файловых операций T_{fo} ($T_{17}, T_{18}, T_{19}, T_{20}, T_{21}, T_{22}, T_{23}, T_{24}$).

В рамках рассматриваемой событийной модели все переходы, кроме T_{25}, являются примитивными, т.е. выполняются мгновенно. Входные и выходные функции связаны как с ветвлениями алгоритма (рисунок 2), так и со вспомогательными позициями и переходами.

Предложенная модель позволяет идентифицировать воздействие на файл на основании данных из вышеуказанных массивов. После извлечения данных специалист проводит оценку равенства компонент I_f, D_f, N_f, определяет признак Z_f и получает результат в виде классифицированной файловой операции, которая помещается в накопитель δ. Модель также обеспечивает верификацию (проверку) массивов данных на предмет наличия искажений данных.

Представленная сеть смоделирована и протестирована в специализированном программном обеспечении Platform Independent Petri Net Editor. Соответствие результата, полученного после выполнения сети, реальным файловым операциям экспериментально подтверждено на специальном стенде с установленными ОС Microsoft Windows 7 Professional SP1, Microsoft Windows 8.1 Professional, Microsoft Windows 10 Professional. Вместе с тем, способ моделирования процесса идентификации воздействий на файлы с использованием математического аппарата сетей Петри является универсальным и может быть применен в других ОС и ФС с учетом имеющихся в них признаков, характеризующих файлы.

Для автоматизации процесса идентификации воздействий на файлы **во второй главе** предложен кластеризационный метод. Идея метода заключается в отображении записей журнала изменений тома $UsnJrnl в виде точек на плоскости для подготовки данных к анализу. Проведению кластеризации способствует нелинейный характер зависимости формирования номера записи U_f журнала $UsnJrnl от времени ее появления T_f. В основе метода использован алгоритм *k*-средних, выбранный исходя из результатов проведенных экспериментов.

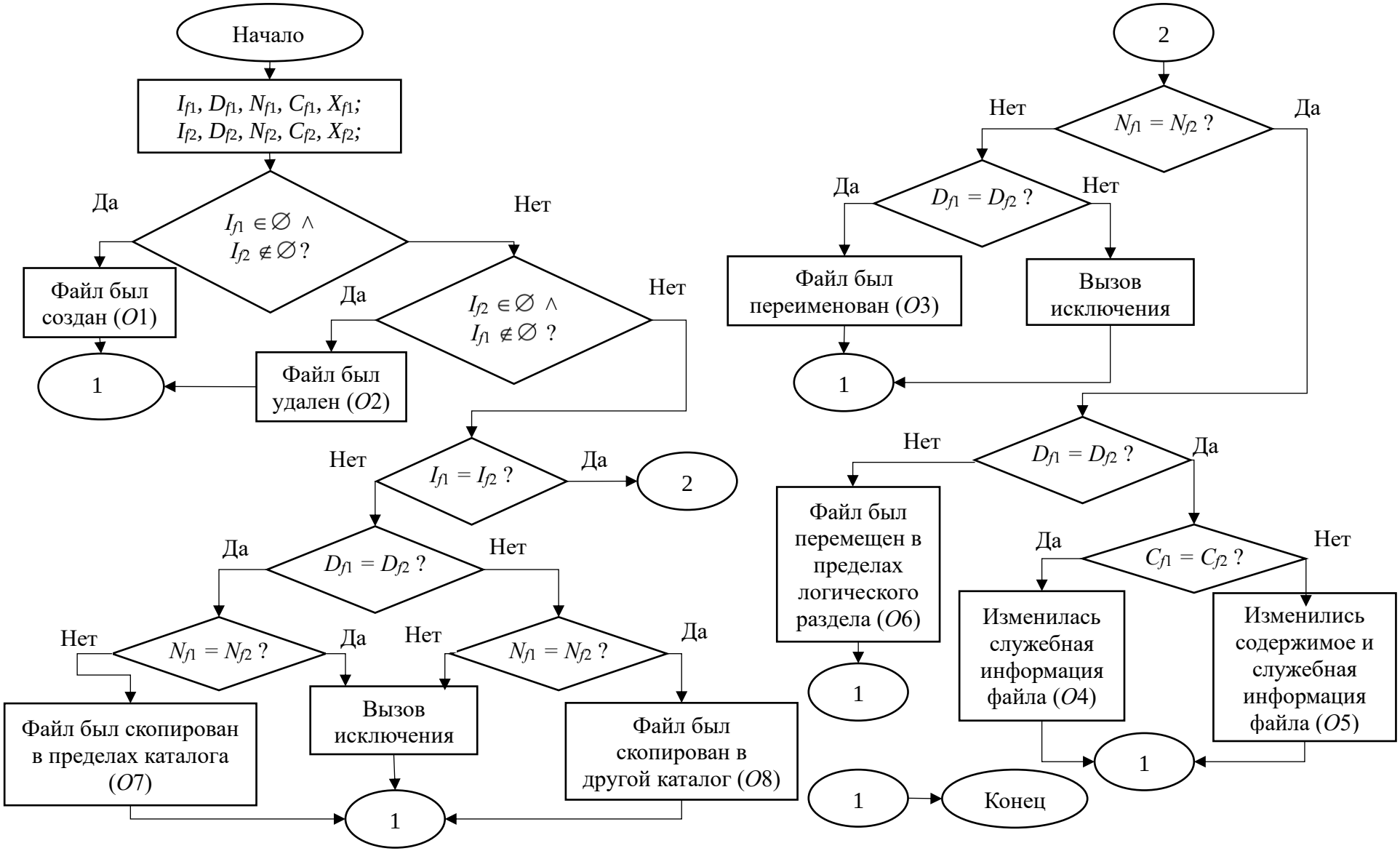

Рисунок 2. Блок-схема алгоритма классификации файловой операции

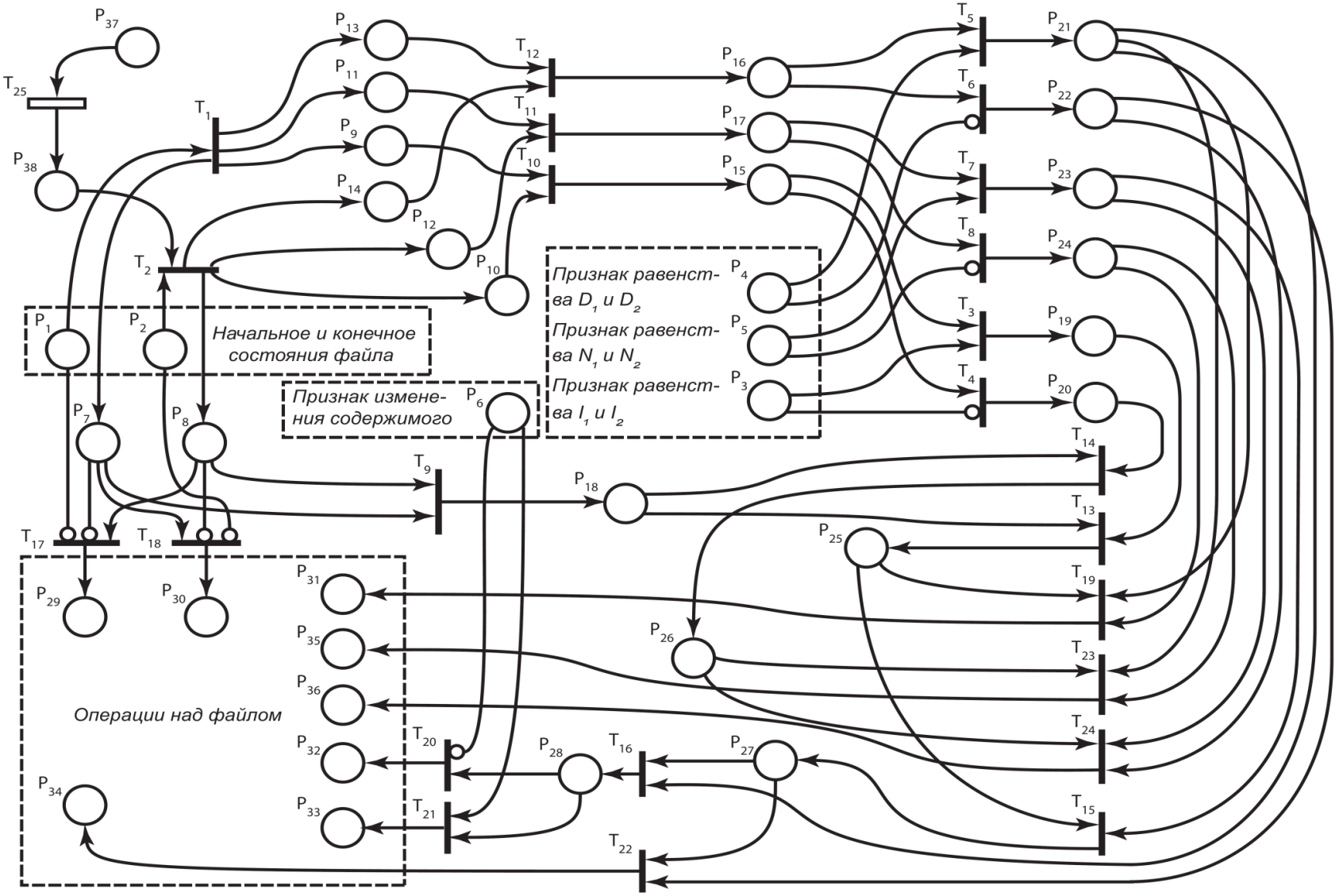

Рисунок 3. Сеть Петри, описывающая событийную модель процесса идентификации воздействий на файлы

В состав кластеризационого метода входят 3 алгоритма. Первый алгоритм, графическое пояснение работы которого представлено на рисунке 4, предназначен для подготовки входных данных – разделения записей журнала $UsnJrnl на блоки в соответствии со значением I_f.

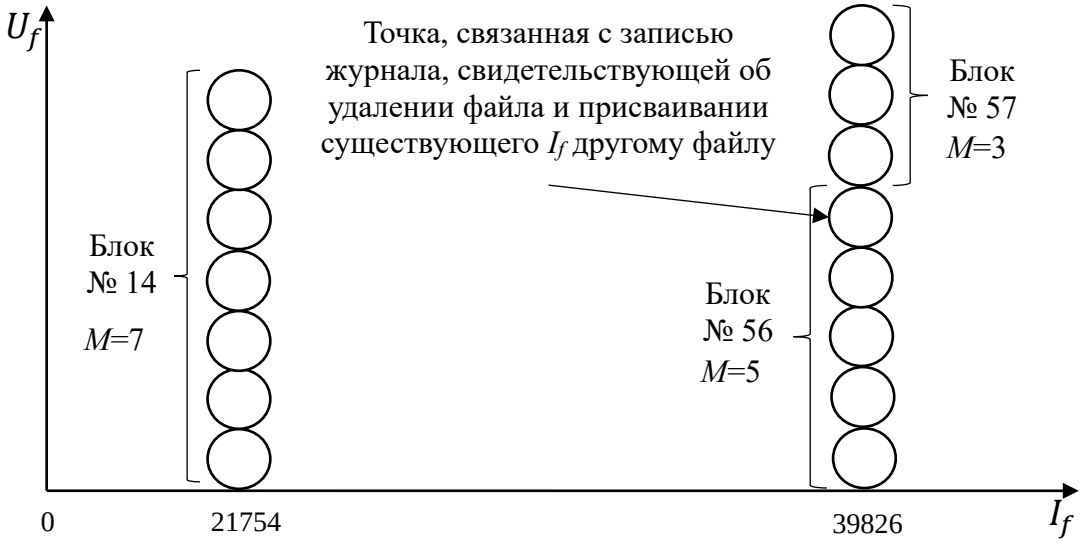

Рисунок 4. Графическое пояснение назначения первого алгоритма

Назначением второго алгоритма является автоматический подбор оптимального значения количества кластеров k – одного из параметров алгоритма кластеризации k-средних. Графическое пояснение алгоритма представлено на рисунке 5.

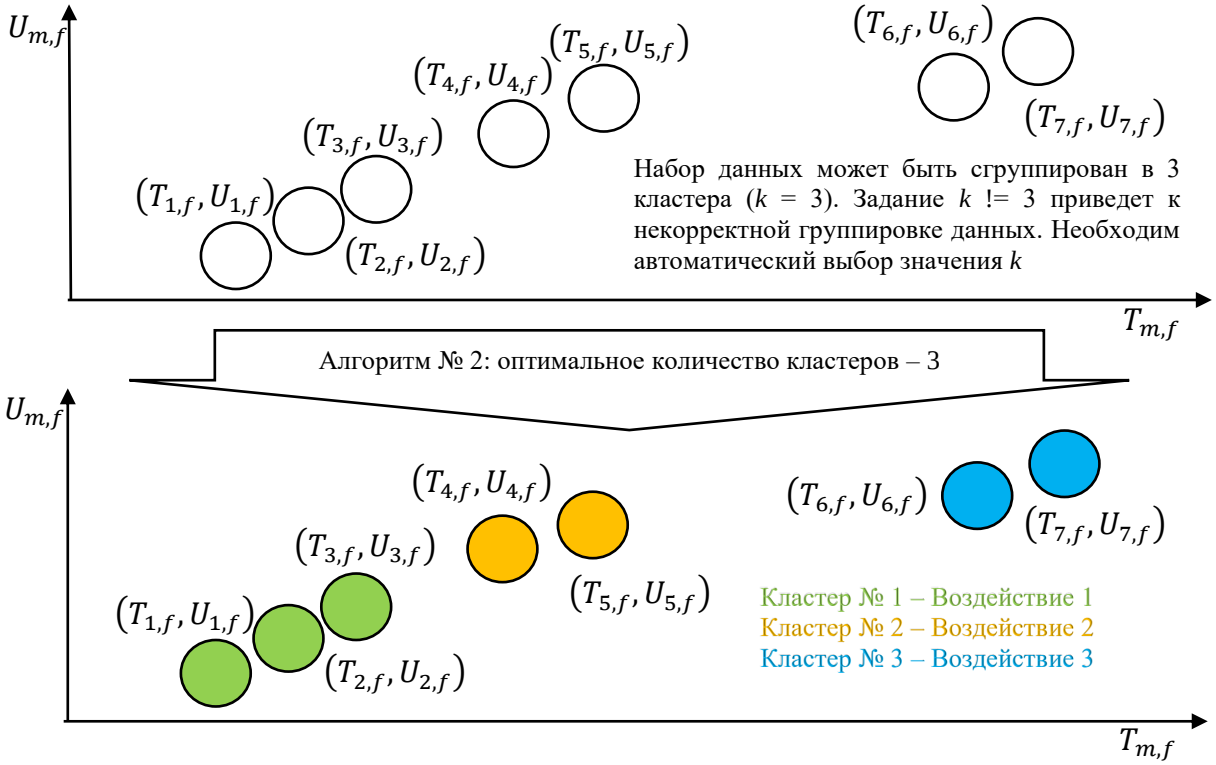

Рисунок 5. Группировка точек на плоскости в k = 3 кластера вторым алгоритмом

Второй алгоритм основан на идее о том, что набор данных можно описать меньшим количеством информации (т.е. кластерами), нежели тем, что используется изначально (т.е. точки на плоскости, соответствующие записям журнала $UsnJrnl).

Указанная идея ранее представлена J. Rissanen в виде т.н. принципа минимальной длины описания (MDL). Его суть заключается в расчете количественной характеристики *L(x)*, называемой длиной описания.

По результатам экспериментов исходная формула для расчета *L(x)* претерпела изменения и выглядит следующим образом:

$$L(x_m, c) = -\sum_{m=1}^{M} \log_2 p(\|x_m - c\|) + k \log_2 M, \quad (4)$$

где x_m – точка с координатами $(T'_{m,f}, U'_{m,f})$, c – центр кластера (определенный с помощью алгоритма k-средних) с рассчитанными координатами (T'_c, U'_c), с которым x_m ассоциирована, $p(\|x_m - c\|)$ – плотность распределения вероятности нормы вектора между x_m и c, k – количество кластеров, M – количество записей $UsnJrnl в блоке, т.е. количество анализируемых точек.

Согласно принципу MDL, оптимальное значение k соответствует минимальному значению $L(x_m, c)$, рассчитанному по формуле (4). По результатам нахождения оптимального значения k к блоку записей журнала $UsnJrnl применяется выбранный ранее алгоритм кластеризации k-средних.

Третий алгоритм кластеризационного метода, представленный на рисунке 6, предназначен для поиска взаимосвязей между полученными кластерами (т.е. простыми[11] воздействиями на файлы), с целью выявления среди них сложных комплексных[12] воздействий.

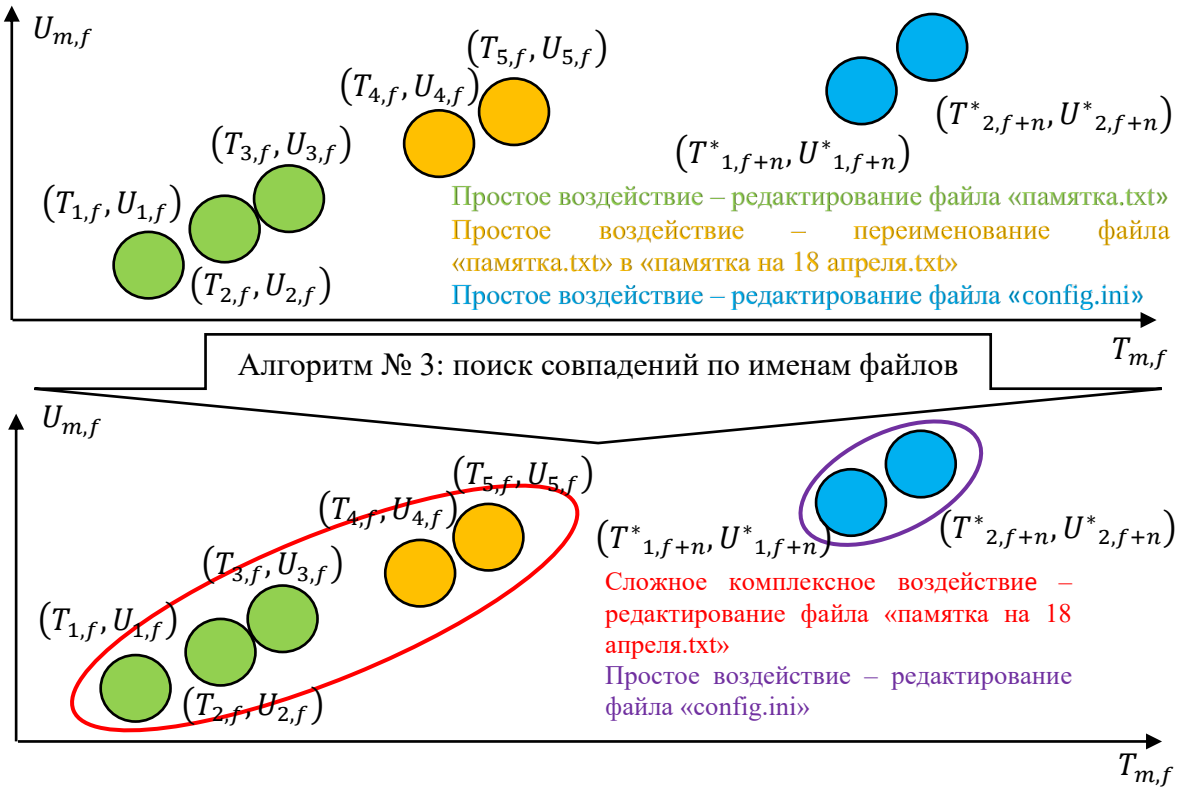

Рисунок 6. Графическое пояснение результата работы третьего алгоритма

[11] Простым является воздействие на файл, состоящее из одной файловой операции.

[12] Сложным комплексным называется такое воздействие, которое состоит из нескольких простых и может иметь отношение к нескольким файлам.

Полученные после работы третьего алгоритма воздействия на файлы «верифицированы» (проверены) с использованием событийной модели процесса идентификации воздействий на файлы, представленной ранее. Предложенный кластеризационный метод идентификации воздействий на файлы является универсальным и может быть применен к другим типам журналов при соблюдении следующих условий: наличие временной отметки появления записи в журнале, номера записи (допускается пронумеровать записи в порядке возрастания) и возможности разделения записей на блоки по дополнительному критерию (аналогично использованию I_f в алгоритме №1 кластеризационного метода).

Результаты работы кластеризационного метода использованы в качестве основы предложенной эталонной структуры, ранее названой шаблоном воздействия на файл, которая представляет собой декомпозицию данных воздействия на значимые признаки.

Значения полей записей журнала составляют основу шаблона G, который представляет собой совокупность фиксированных значений, описываемых вектором:

$$G = \{\langle G_{name}, G_{anomaly}, G_{priority}\rangle, \langle I_G, I_{sign}\rangle, \langle D_G, D_{sign}\rangle, \langle N_G, N_{sign}\rangle, \langle R_G, R_{sign}\rangle\}, \quad (5)$$

где в отношении к рассматриваемому шаблону воздействия:

- G_{name}, $G_{anomaly}$, $G_{priority}$ – название, признак аномальности и приоритет шаблона воздействия;
- I_G, D_G, N_G, R_G – множества значений, описывающих: идентификаторы файловых записей, идентификаторы родительских каталогов, имена файлов, расширения имен файлов и используемые специальные символы[13] (при их наличии), идентификаторы операций, полученных из поля «Идентификатор операции» записей журнала $UsnJrnl;
- I_{sign}, D_{sign}, N_{sign}, R_{sign} – признаки значимости соответствующих множеств значений I_G, D_G, N_G, R_G для воздействия на файл, описываемого шаблоном G.

Признак «аномальность» непосредственно связан с процедурой экспертной оценки и определяется специалистом-аналитиком, подготавливающим шаблон воздействия на файл с целью последующих их (воздействий) классификации при расследовании инцидента ИБ.

Признаки значимости I_{sign}, D_{sign}, N_{sign}, R_{sign} определяют необходимость использования значений из соответствующих множеств I_G, D_G, N_G, R_G при экспресс-анализе воздействий на файлы с использованием шаблонов. Каждый признак значимости принимает значение 0 или 1, определяемое специалистом-аналитиком, подготавливающим шаблон, в зависимости от того, важны ли значения соответствующего множества при идентификации воздействия на файл с применением генерируемого шаблона. Пример шаблона идентификации воздействий на файлы представлен на рисунке 7.

[13] Символы, не являющиеся буквами, цифрами и пробелами, относятся к категории специальных при рассмотрении имен файлов.

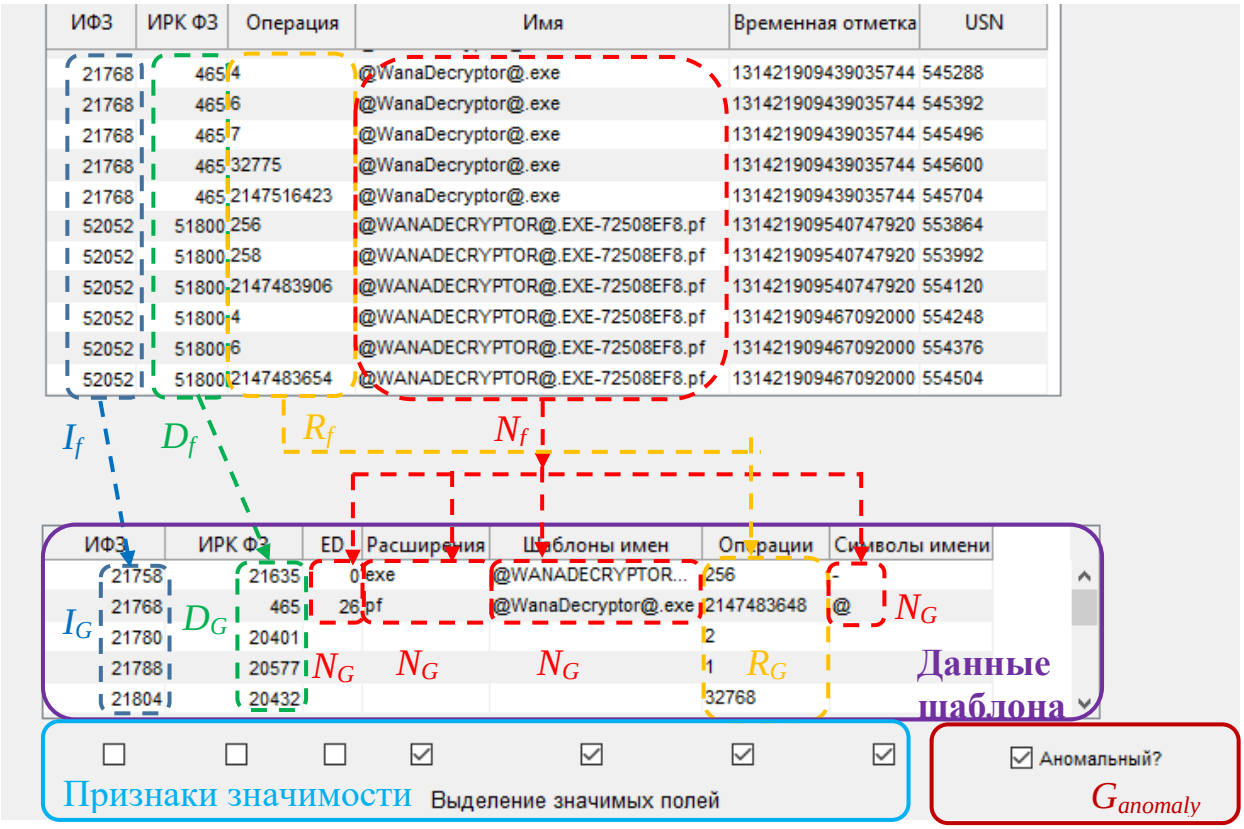

Рисунок 7. Пример генерации шаблона воздействия на файл

Использование шаблонов является основой разработанного метода экспресс-анализа событий ИБ, связанных с воздействиями на файлы, и позволяет сократить временные затраты на обнаружение и классификацию событий, за счет отказа от ресурсоемких процедур переобучения классификатора в пользу простого порогового сравнения записей журнала $UsnJrnl с базой данных шаблонов, как показано на рисунке 8.

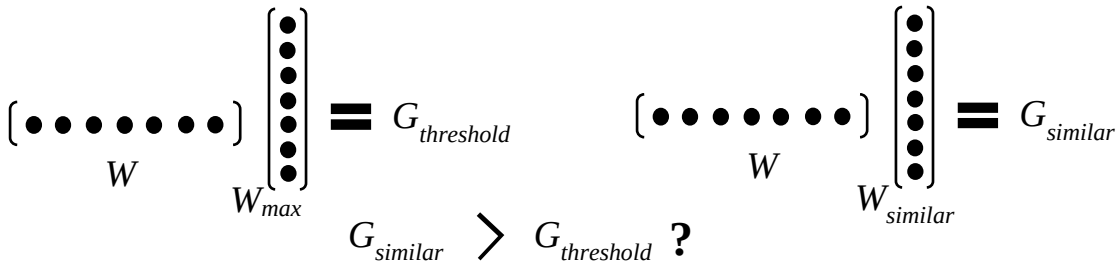

Рисунок 8. Графическое пояснение алгоритма порогового сравнения

На рисунке 8 введены следующие обозначения: W – вектор-строка весовых коэффициентов, соответствующих признакам значимости шаблона G, W_{max} – вектор-столбец весовых коэффициентов всех признаков значимости шаблона G, $W_{similar}$ – вектор-столбец весовых коэффициентов, которые не равны 0, если содержимое записей журнала $UsnJrnl совпало с признаками значимости шаблона G, $G_{threshold}$ и $G_{similar}$ – пороговый коэффициент и коэффициент совпадения записей с признаками шаблона.

Схема применения метода экспресс-анализа представлена на рисунке 9. Разработанный метод экспресс-анализа событий ИБ универсален и может быть использован в других типах ОС и ФС с учетом особенностей изменения признаков, характеризующих файл, которые должны быть отражены в шаблоне воздействий.

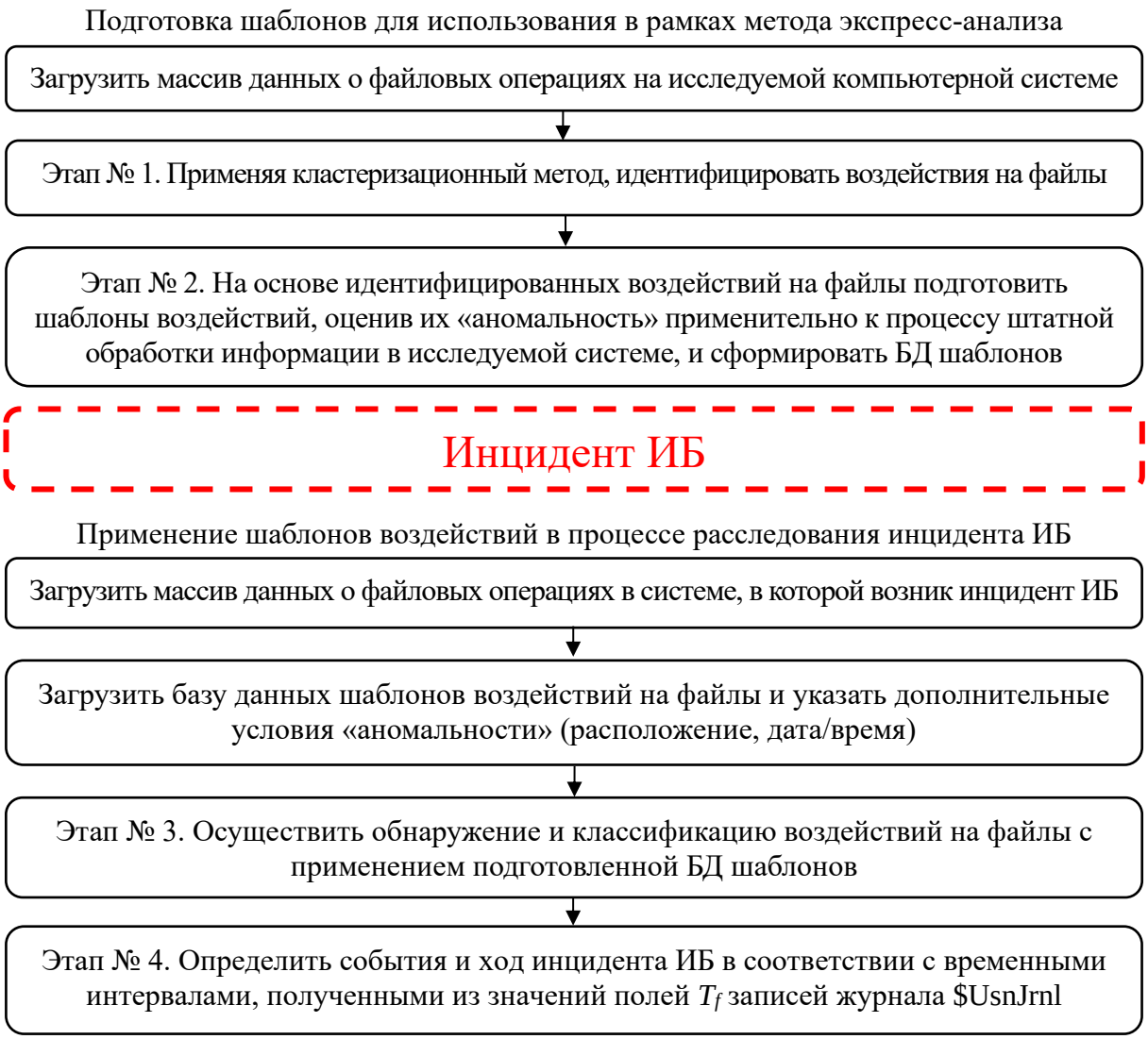

Рисунок 9. Поэтапная схема метода экспресс-анализа событий ИБ

Предложенные событийная модель процесса идентификации воздействий на файлы, кластеризационный метод и метод экспресс-анализа реализованы в виде комплекса программных средств, описанного **в третьей главе**, где в т.ч. рассмотрен порядок совместного использования комплекса с существующим ПО, применяемым при расследовании инцидентов ИБ, и приведены результаты экспериментов по оценке качества кластеризации и классификации предложенных кластеризационного метода и метода экспресс-анализа соответственно.

IDEF0-схема комплекса программных средств представлена на рисунке 10.

В главе приведены рекомендации по использованию комплекса, состоящего из 3 программных средств:

- генерации шаблонов воздействий на файлы, используемого при расследовании инцидентов информационной безопасности в рамках первого и второго этапов метода экспресс-анализа событий ИБ (рисунок 10);
- обнаружения событий информационной безопасности, использующего шаблоны воздействий на файлы, применяемого в рамках третьего и четвертого этапов метода экспресс-анализа событий ИБ (рисунок 10);
- генерации компьютерных атак и осуществления воздействий на файлы, предназначенного для имитации активности пользователя / процесса в отношении файлов компьютерной системы.

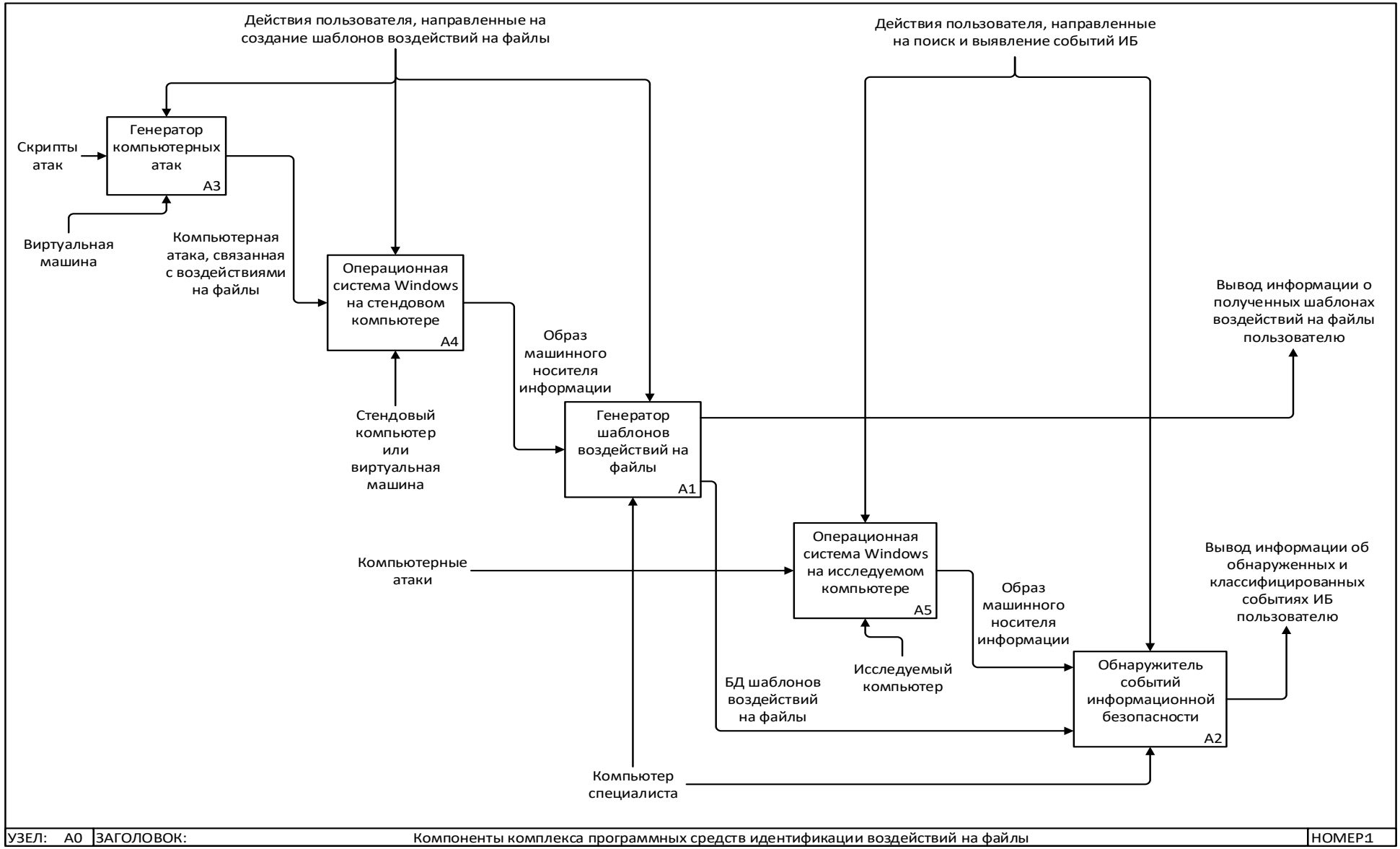

Рисунок 10. IDEF0-схема разработанного комплекса программных средств

Проведенный сравнительный анализ на основании расчета значений метрик Precision, Recall и *F*-score[14] предложенного кластеризационного метода и существующих методов кластеризации представлен в таблице 1.

Таблица 1. Расчет Precision (*P*), Recall (*R*) и *F*-score (*F*) для нескольких выборок экспериментальных данных

Алгоритм кластеризации	**10 точек / 3 кластера**			**3 точки / 2 кластера**			**26 точек / 22 кластера**			**18 точек / 2 кластера**		
	***P* %**	***R* %**	***F* %**	***P* %**	***R* %**	***F* %**	***P* %**	***R* %**	***F* %**	***P* %**	***R* %**	***F* %**
k-средних	88.8	91.6	90.2	75.0	75.0	75.0	90.9	91.6	91.2	92.8	83.3	87.8
k-средних++	88.8	91.6	90.2	75.0	75.0	75.0	90.9	91.6	91.2	92.8	83.3	87.8
k-медоидов	77.7	82.2	79.9	75.0	75.0	75.0	86.3	87.1	86.7	89.2	78.5	83.5
DBSCAN	77.7	82.2	79.9	25.0	25.0	25.0	97.7	97.7	97.7	92.8	83.3	87.8

Из таблицы 1 видно, что представленные алгоритмы дают схожие значения Precision, Recall и *F*-score, но итоговый результат зависит от выбранных параметров алгоритма и входных данных. В каждом примере путем *ручного* изменения параметров алгоритма можно добиться значений Precision, Recall и *F*-score, равных 1. В рассматриваемых примерах значение количества кластеров *k* определялось *автоматически* в соответствии с алгоритмом № 2 кластеризационного метода. Для оценки качества кластеризации количество кластеров *k* при работе алгоритмов *k*-средних++ и *k*-медоидов выбиралось аналогично *k*-средних. Автоматический подбор параметров ε и minPts алгоритма DBSCAN не описан в существующих работах, поэтому были выбраны и зафиксированы такие значения, которые давали приемлемые результаты на большинстве блоков данных.

Проведенный сравнительный анализ на основании расчета значений метрик Precision, Recall и *F*-score предложенного метода экспресс-анализа и некоторых существующих методов классификации представлен в таблице 2.

Таблица 2. Расчет Precision (*P*), Recall (*R*) и *F*-score (*F*) для нескольких выборок экспериментальных данных

Метод классификации	**1000 точек / 6 классов / 3 признака**			**100 точек / 4 класса / 4 признака**			**25 точек / 4 класса / 5 признаков**		
	***P* %**	***R* %**	***F* %**	***P* %**	***R* %**	***F* %**	***P* %**	***R* %**	***F* %**
Ансамбль деревьев решений	92.7	96.6	94.6	92.7	96.6	94.6	93.2	97.3	95.2
Метод опорных векторов	65.5	92.1	77.3	72.0	89.3	79.7	78.6	80.6	79.5
Метод экспресс-анализа	91.2	96.0	93.5	92.3	96.6	94.4	95.7	96.0	95.8

Исходя из результатов, представленных в таблице 2 видно, что предложенный метод экспресс-анализа показывает бóльшую точность классификации, чем метод опорных векторов с линейным ядром, но в некоторых ситуациях уступает ансамблю деревьев решений, что может быть связано с некорректно заданными значениями признаков вектора *G*. Вместе с тем возможность изменения набора выявляемых классов путем добавления/удаления шаблонов воздействий на файлы позволяет не проводить процедуру обучения классификатора повторно, что является преимуществом предлагаемого метода экспресс-анализа в сравнении с существующими методами классификации.

Полученные результаты оценки предложенных методов в сравнении с существующими подтверждают правильность их выбора в качестве основы для

[14] C.J. van Rijsbergen: Information Retrieval. 2nd Edition. Butterworth, 1979.

автоматизации процессов идентификации, обнаружения и классификации воздействий на файлы при расследовании инцидентов ИБ.

В заключении изложены результаты выполненного исследования, рекомендации, перспективы дальнейшей разработки темы.

В ходе исследования получены следующие результаты:

1. Проведен анализ состояния предметной области с указанием недостатков существующих методов, алгоритмов и средств анализа массивов данных, содержащих информацию о воздействиях на файлы.
2. Предложены математические методы и модель, позволяющие в автоматизированном режиме идентифицировать, обнаруживать и классифицировать воздействия на файлы:
    a. впервые обоснована возможность применения математического аппарата сетей Петри для описания процесса идентификации воздействий на файлы, что позволяет формализовать инцидент ИБ как совокупность событий, связанных с воздействиями на файлы;
    b. разработан кластеризационный метод идентификации воздействий на файлы, впервые использующий адаптированные алгоритмы предварительной подготовки входных данных и определения оптимального количества кластеров, позволяющий оптимизировать решение задачи анализа данных о воздействиях на файлы;
    c. разработан метод экспресс-анализа событий ИБ, связанных с воздействиями на файлы, позволяющий выявлять аномальные, условно аномальные и нормальные воздействия на файлы в ходе расследования инцидента ИБ.
3. Разработан комплекс программных средств, позволяющий автоматизировать процесс идентификации, обнаружения и классификации воздействий на файлы в целях определения событий ИБ, лежащих в основе расследуемого инцидента.

Перспективными задачами исследования являются адаптация предложенной событийной модели процесса идентификации воздействий на файлы в других ФС, модификация предложенных кластеризационного метода идентификации воздействий на файлы и метода экспресс-анализа событий ИБ для обработки записей журналов других ФС, доработка комплекса программных средств для функционирования в других типах ОС.

Список публикаций автора по теме диссертации

Статьи, опубликованные в рецензируемых научных журналах и изданиях, определенных ВАК РФ и Аттестационным советом УрФУ:

1. Гайдамакин Н.А. Метод экспресс-анализа событий, связанных с воздействиями на файлы, предназначенный для расследования инцидентов информационной безопасности / Н.А. Гайдамакин, **Р.В. Гибилинда**, Н.И. Синадский // Вестник СибГУТИ. — 2020. — № 4. — С.3-13. (0,68 п.л. / 0,23 п.л.).

2. Gaidamakin N.A. File Operations Information Collecting Software Package Used in the Information Security Incidents Investigation / Nikolay A Gaidamakin, **Roman V Gibilinda** and Nikolay I Sinadskiy // 2020 Ural Symposium on Biomedical

Engineering, Radioelectronics and Information Technology (USBEREIT). – 2020. – № 9117671. – pp. 559-562. (0,25 п.л. / 0,08 п.л.) (Scopus).

3. Гайдамакин Н.А. Событийная модель процесса идентификации воздействий на файлы при расследовании инцидентов информационной безопасности, основанная на математическом аппарате сетей Петри / Н.А. Гайдамакин, **Р.В. Гибилинда**, Н.И. Синадский // Вестник СибГУТИ. — 2020. — № 1. — С.73-88. (1 п.л. / 0,33 п.л.).

4. **Гибилинда Р.В.** Кластеризационный метод идентификации воздействий на файлы с применением алгоритма *k*-средних, используемый при расследовании инцидентов информационной безопасности / **Р.В. Гибилинда** // Вестник УрФО. Безопасность в информационной сфере. — 2020. — Вып. 35 — № 1. — С. 35-47. (0,81 п.л.).

Свидетельства о государственной регистрации программ для ЭВМ:

5. Свидетельство № 2020616110 Российская Федерация. Свидетельство о государственной регистрации программы для ЭВМ «Программное средство обнаружения событий информационной безопасности, использующее шаблоны воздействий на файлы» / **Р.В. Гибилинда**. – Заявка № 2020613614 от 26.03.2020; дата гос. регистрации в Реестре 09.06.2020. – Реестр программ для ЭВМ. – 1 с.

6. Свидетельство № 2020616109 Российская Федерация. Свидетельство о государственной регистрации программы для ЭВМ «Программное средство генерации шаблонов воздействий на файлы, используемое при расследовании инцидентов информационной безопасности» / **Р.В. Гибилинда**. – Заявка № 2020613613 от 26.03.2020; дата гос. регистрации в Реестре 09.06.2020. – Реестр программ для ЭВМ. – 1 с.

Другие публикации:

7. **Гибилинда Р.В.** Автоматизация процесса идентификации воздействий на файлы с применением кластеризационного метода при расследовании инцидентов информационной безопасности / **Гибилинда Р.В.**, Синадский Н.И. // II Всероссийская научная конференция (с приглашением зарубежных ученых) «Фундаментальные проблемы информационной безопасности в условиях цифровой трансформации» (FISP-2020). — 2020. — С. 284-287 (0,31 п.л. / 0,16 п.л.).

8. **Гибилинда Р.В.** Идентификация воздействий на файлы и верификация массивов данных, содержащих информацию о воздействиях, при расследовании инцидентов информационной безопасности / **Гибилинда Р.В.**, Синадский Н.И. // Современные проблемы радиоэлектроники и телекоммуникаций: сб. науч. тр. / под ред. Ю. Б. Гимпилевича. — Москва-Севастополь: Изд-ва: РНТОРЭС им. А.С. Попова, СевГУ. — 2020. — № 3. — С. 219-219 (0,06 п.л. / 0,03 п.л.).

Гибилинда Роман Владимирович

Разработка автоматизированных методов анализа воздействий на файлы в задаче расследования инцидентов информационной безопасности

Автореф. дис. на соискание ученой степени кандидата технических наук

Подписано в печать __.__._____. Заказ № ___

Формат 60х×90/16. Усл. печ. л. 1 . Тираж 70 экз.

Типография ____ _____