УДК 004.056

БУРЬКОВА ЕЛЕНА ВЛАДИМИРОВНА,
ИЗВЕКОВА ЛЮБОВЬ АНДРЕЕВНА

ПРИМЕНЕНИЕ МЕТОДА КЛАСТЕРИЗАЦИИ ДАННЫХ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ОЦЕНКИ РИСКОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

АННОТАЦИЯ

Построена классификация и проведен анализ методов оценки рисков информационной безопасности. Дана характеристика метода кластеризации данных в контексте оценки рисков. Разработана математическая модель метода кластеризации для оценки рисков. Рассмотрен пример применения данного метода.

Ключевые слова: анализ рисков; классификация методов; кластеризация данных; взаимная информация.

BURKOVA E.V.,
IZVEKOVA L.A.

APPLICATION OF DATA CLUSTERING METHOD FOR SOLVING THE RISK ASSESSMENT PROBLEM INFORMATION SECURITY

ABSTRACT

The classification was built and the analysis of information security risk assessment methods was carried out. The characteristic of the data clustering method in the context of risk assessment is given. A mathematical model of the clustering method for risk assessment has been developed. An example of the application of this method is considered.

Keywords: risk analysis; classification of methods; data clustering; mutual information.

В условиях роста количества и степени опасности угроз утечки информации проблема обеспечения информационной безопасности приобретает все большую значимость для предприятий и организаций. По данным аналитического центра InfoWatch процент утечек в первом полугодии 2018 года по сравнению с 2017 годом вырос на 12% (рис. 1) [1].

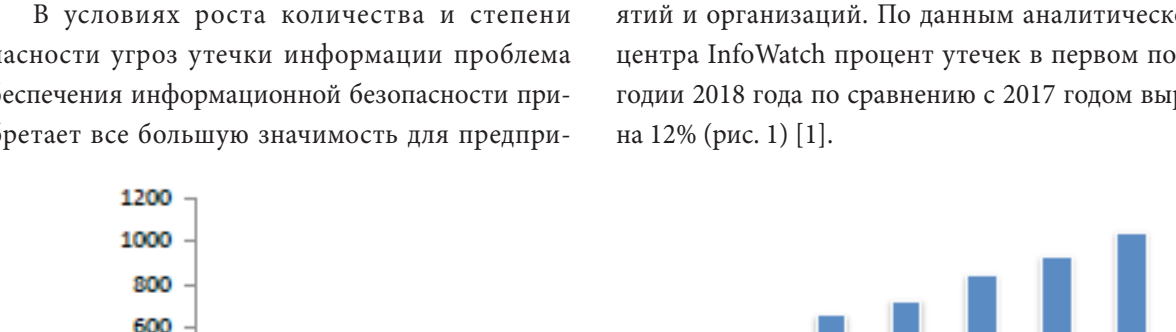

Рисунок 1 – Число утечек информации в первом полугодии 2006 – 2018 гг.

В связи с повышением опасности реализации угроз актуализируется задача проведения аудита безопасности на объектах и проведение модернизации системы защиты информации с целью предотвращения возможного ущерба. Оценка рисков является начальным этапом проектирования системы защиты на объектах и позволяет:

- определить критический уровень безопасности защищаемых активов информационной системы;
- обнаружить уязвимости в системе защиты информации;
- провести оценку эффективности мер по защите информации;
- выработать рекомендации по совершенствованию системы защиты.

На сегодняшний день в данной области отсутствуют единые методы оценки, учитывающие специфику каждого предприятия. Существует проблема сбора достаточного объема статистических данных о вероятности реализации той или иной угрозы. В связи с названными факторами задача оценки рисков информационной безопасности является актуальной.

Согласно ГОСТ ИСО/МЭК 27005-2010, риск информационной безопасности – возможность того, что данная угроза сможет воспользоваться уязвимостью актива или группы активов и тем самым нанесет ущерб организации [2]. Для того чтобы оценить риск необходимо провести идентификацию угроз, уязвимостей и активов. Идентификация угрозы состоит из определения частоты возникновения угрозы, идентификация уязвимости определяется степенью тяжести данной уязвимости информационной системы, актив – определяется его ценностью.

Существует ряд проблем, которые ставят перед специалистами задачу разработки новых методов оценки, такими проблемами являются:

- отсутствие единой методики расчета информационных рисков;
- неадекватность нечетких и лингвистических данных, полученных от экспертов во время диалога;
- проблема построения списка актуальных угроз и уязвимостей в рамках данного предприятия [3].

Проблема оценки рисков информационной безопасности рассматривалась многими учеными и ими были применены различные методы: логико-вероятностный, метод теории игр, нечеткой логики и другие [3, 4, 5, 6]. Нами была построена классификация методов оценки рисков информационной безопасности (рис. 2).

Методы оценки рисков подразделяются на количественные и качественные. Количественной оценкой риска называют процесс присвоения значений вероятности и последствий риска. Качественные оценки риска позволяют выявить риски и их факторы, однако данную оценку приводят к количественным характеристикам, например, ущерб от рисков, таким образом, появляются комбинированные методы, в которых вербальные характеристики степени риска дополняются значениями ущерба.

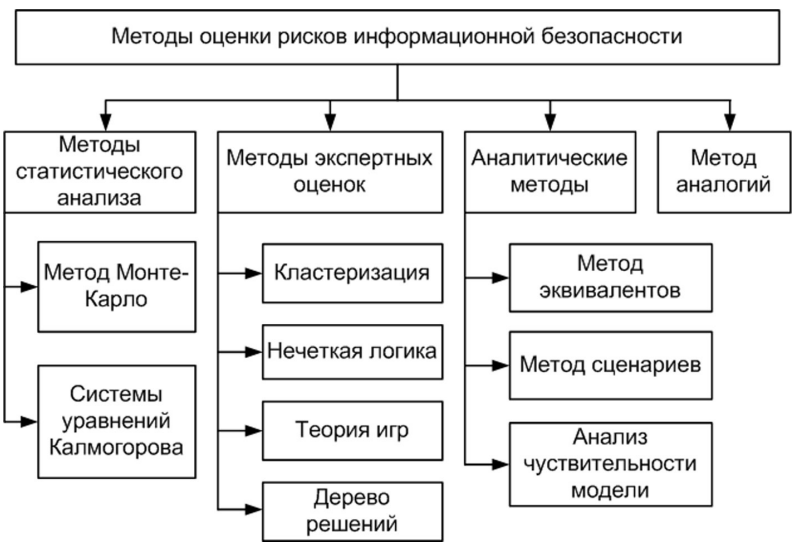

Рисунок 2 – Классификация методов оценки

Экспертные методы – способ сбора исходной информации для построения модели риска выполняется экспертами, имеющими для этого необходимые знания. Статистические методы – накопление статистических данных о реализации угроз и последующих денежных потерях, имевших место на данном предприятии. Аналитические методы – заключается в построении кривой риска, сложный и доступен только профессионалам, проводят оценку рисков на уровне бизнес-процессов. Обычно выбирается показатель чувствительность которого вычисляется в зависимости от различных факторов. Метод аналогов применяют, когда другие методы не дают результатов, проводят построение базы аналогичных объектов, определение общих связей и перенос результатов на исследуемый объект.

В нашей работе для оценки рисков рассматривается метод кластеризации данных. Кластеризация – процесс разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Цель кластеризации: упрощение обработки данных для принятия решений, применяя к каждому кластеру свой метод анализа. Также благодаря кластеризации можно обнаружить нетипичные объекты, которые не удается присоединить ни к одному из кластеров [8].

Рассмотрим алгоритм кластерного анализа в контексте информационной безопасности. Предполагается прохождение нескольких этапов.

1. Определение защищаемых активов. Под активом понимаются база данных поставщиков (клиентов), данные матрицы доступа, политика безопасности организации и другая конфиденциальная информация. Активы и центры кластеров находятся в N-мерном пространстве. N-количество факторов риска, рассматриваемые на предприятии. Центры кластеров можно устанавливать несколькими способами. Установим для начала центры произвольным образом.
2. Ассоциация между активами и уровнями риска на основе определения евклидового расстояния от каждого элемента до текущего центра кластера;
3. Вычисление новых центров кластеров.
4. Повторение данных шагов до того, пока центр не перестанет изменяться или не произойдет количество заданных итераций.

Кластеризация данных при оценке рисков информационной безопасности в примере будет проводится по следующим уровням: высокая безопасность, средняя безопасность, опасно. Иллюстрация кластеризации показана на рисунке 3.

Математическая модель метода представлена в таблице 1.

Рисунок 3 – Демонстрация работы кластеризации

Таблица 1

Математическая модель метода

Этапы оценки	Формальное представление	Характеристика данных	Результат этапа
Этап 1. Идентификация активов, угроз и уязвимостей ИС	A = {a1, a2 … an}	Множество активов, подлежащих защите an – защищаемый актив n – количество активов	Перечень активов
	T = {t1, t2 … tm}	Множество угроз информационной безопасности tm – угроза ИБ m – количество угроз	Перечень угроз безопасности
	V = {v1, v2 … vz}	Множество уязвимостей информационной безопасности vz – уязвимость ИБ z – количество уязвимостей	Перечень уязвимостей информационной безопасности
Этап 2. Определение актуальных угроз и уязвимостей ИС	R = T∩V R = {r1, r2 … rj}	Множество факторов риска rj – фактор риска j – количество факторов риска	Перечень факторов риска ИС
Этап 3. Формирование центров кластеров.	$I(a_k, d) = p(a_k\|d) lb \frac{p(a_k, d)}{p(a_k)p(d)}$	Взаимная информация между значением фактора риска и уровнем риска	Начальные центры кластеров
Этап 4. Определение принадлежности активов к центрам и установка нового центра кластеров	L = {l1, l2, l3, l4}	L1 – высокая безопасность L2 – средняя безопасность L3 – подозрительно L4 – опасно	
	$d(X, C) = \sqrt{\sum_{i=1}^{N} (x_i - c_i)^2}$	X_i – точка данных (факторов риска актива) C_i – центр кластера N – количество факторов риска	Принадлежность активов к центрам
Этап 5. Выявление наименее защищенных активов			Формирование таблицы распределения активов по уровням риска

№ данных	b_1	b_2	b_3	b_4	Уровень безопасности
D1	0, 1	0,3	0,3	0,2	L1
D2	0,5	0,3	0,4	0,3	L2
D3	0,5	0,3	0,4	0,3	L2
D4	0,3	0,7	0,5	0,3	L4
D5	0,3	0,3	0,2	0,3	L2

Совместно с методом кластеризации используется понятие взаимной информации для определения статистической степени корреляции между значениями фактора риска и уровнем риска, определяется по формуле

$$I(a_k, d) = p(a_k|d) lb \frac{p(a_k,d)}{p(a_k)p(d)},$$

где $I(a_k, d)$ – взаимная информация

a_k – конкретное значение фактора риска

d – один из уровней риска (центр кластера)

$p(a_k) = \frac{a_k}{N}$ – отношения числа появления a_k во всех факторах оценки подготовительных данных и общего количества значений факторов оценки.

$p(d) = \frac{d}{N}$ – вероятность подготовительных данных

$p(a_k|d) = \frac{Na_k \cap d}{d_l}$ – условная вероятность (вероятность наступления значения фактора при условии, что фактор принадлежит уровню d); Отношение числа выборочных данных, значение которых a_k, и уровень риска принадлежит уровню d к общему числу выборочных данных, чей уровень принадлежит d.

$p(a_k, d) = \frac{Na_k \cap d}{N}$ – представляет собой вероятность появления a_k, в то время как данные с атрибутом a_k принадлежат уровню риска d [9].

Рассмотрим пример на основе математической

модели, построенной ранее. Для начала сформируем множества активов и факторов риска.

Таблица 2

Входные данные для анализа

Активы	Факторы риска
A1 – информационная система персональных данных	R1 – кража информации
A2 – база данных поставщиков	R2 – внедрение вредоносных программ
A3 – данные матрицы доступа	R3 – отказ в обслуживании
A4 – парольная информация	R4 – угроза ошибочных действий администратора безопасности

Далее эксперту необходимо сформировать матрицу факторов риска (таблица 3), в которой указаны вероятности реализации угроз.

Таблица 3

Матрица факторов риска

	R1	R2	R3	R4	L
A1	0.1	0.3	0.3	0.2	?
A2	0.5	0.3	0.4	0.3	?
A3	0.3	0.6	0.5	0.7	?
A4	0.5	0.3	0.3	0.5	?

Последний столбец указывает на уровень безопасности, то есть указывает эксперту наименее защищенный актив. Для формирования данных уровней воспользуемся мерой взаимной информации, чтобы определить центр каждого из четырех кластеров. Пример для вычисления взаимной информации между значением риска 0.1 и центром кластера L1.

$$p(0.1) = \frac{1}{16}$$

$$p(L1) = \frac{0.2}{4}\, p(0.1|L1) = \frac{1}{9}\; p(0.1, L1) = \frac{1}{16}\; I(0.1, L1) = 0.48$$

Аналогично проводятся вычисления для других значений и центров. Следующий этап – вычисление расстояния от актива до центра, для определения принадлежности к кластерам.

$$d(X, C) = \sqrt{\sum_{i=1}^{N} (x_i - c_i)^2} = \sqrt{0.099324} = 0.3151 (A1)$$

$$d(X, C) = \sqrt{\sum_{i=1}^{N} (x_i - c_i)^2} = \sqrt{0.840924} = 0.91\ (A3)$$

Актив с наибольшими вероятностями атак (A3) находится на самом большом расстоянии от центра L1 и является самым слабозащищенным. Результат показан в таблице 4.

Таблица 4

Результат работы метода

	R1	R2	R3	R4	L
A1	0.1	0.3	0.3	0.2	L1
A2	0.5	0.3	0.4	0.3	L2
A3	0.3	0.6	0.5	0.7	L4
A4	0.5	0.3	0.3	0.5	L3

При добавлении новых данных центры необходимо пересчитать. В результате данный актив наименее защищен, эксперт должен предложить рекомендации по защите данных.

Таким образом, применение данного метода позволяет определить наименее защищенные активы с целью выработки рекомендаций по совершенствованию системы защиты для обеспечения высокого уровня защищенности объекта информатизации.

Список литературы

1. Аналитический центр InfoWatch. Глобальное исследование утечек конфиденциальности информации в I полугодии 2018 года.

2. ГОСТ Р ИСО/МЭК 27005-2010. Информационная технология. Методы и средства обеспечения безопасности. Менеджмент риска информационной безопасности. [Электронный ресурс] – Режим доступа: http://docs.cntd.ru/document/gost-r-iso-mek-27005-2010 (дата обращения 16.03.2019).

3. *Бурькова, Е. В., Гайфуллина Д.А., Хакимова Э.Р.* Прикладная программа оценки физической защищенности объекта на основе логико-вероятностного подхода. Материалы VI Международной научно-практической конференции «Информационные ресурсы и системы в экономике, науке и образовании». Пенза: Общество «Знание» России, Приволжский дом знаний. – 2016. – С. 10-14.

4. *Аткина В.С., Воробьев А.Е.* Подход к оценке рисков нарушения информационной безопасности с использованием иерархического подхода к ранжированию ресурсов предприятия // Информационные системы и технологии. – 2015. – № 1 (87). – С. 125-131.

5. *Черезов Д.С., Тюкачев Н.А.* Обзор основных методов классификации и кластеризации данных // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. –2009. – № 2. – С. 25-29.

6. *Юрьев В.Н.* Игровой подход к оценке риска и формированию бюджета информационной безопасности предприятия // Прикладная информатика. – 2015. – Т.10. – №2(56). – С. 121-126.

7. *Волосенков В.О., Гаврилова Т.Н.* Способы оценки рисков информационной безопасности распределенных вычислительных систем // Проблемы безопасности российского общества. – 2015. – № 1. – С. 69-74.

8. *Прудковский Н.С.* Кластеризация данных методом К-средних. Безопасные информационные технологии // Сборник трудов Восьмой всероссийской научно-технической конференции. НУК «Информатика и системы управления». Под. ред. М. А. Басараба. – 2017. – С. 347-350.

9. *Козлова Е. А.* Оценка рисков информационной безопасности с помощью метода нечеткой кластеризации и вычисления взаимной информации // Молодой ученый. – 2013. – № 5. – С. 154-161.

Статья поступила в редакцию 23 марта 2019 г.
Принята к публикации 28 мая 2019 г.