УДК 004.492.3

к.т.н., доц. **Графов Р.П.** (ХНУ, г. Хмельницкий)
д.т.н., проф. **Боровик О.В.** (НАГПСУ, г. Хмельницкий)
д.т.н., проф. **Чумаков В.И.** (АВМС, г. Севастополь)

НЕЧЕТКИЕ МОДЕЛИ В ЗАДАЧАХ ПОИСКА ВРЕДОНОСНЫХ АГЕНТОВ В ИНФОРМАЦИОННЫХ СИСТЕМАХ

У статті розглядається методологія вирішення проблеми пошуку шкідливих програм в розподілених інформаційних системах, характеризуються нечіткої інформацією про протікають процеси. Використаний механізм нечіткого логічного висновку, заснованого на продукційних правилах прийняття рішення. Отримані результати можуть бути використані для антивірусного діагностування програмного забезпечення різних комп'ютерних систем.
Ключові слова: інформаційні системи, програмне забезпечення.

В статье рассматривается методология решения проблемы поиска вредоносных программ в распределенных информационных системах, характеризующихся нечеткой информацией о протекающих процессах. Использован механизм нечеткого логического вывода, основанного на продукционных правилах принятия решения. Полученные результаты могут быть использованы для антивирусного диагностирования программного обеспечения различных компьютерных систем.
Ключові слова: Інформаційні системи, програмне забезпечення.

In the article the methodology of the decision of a problem of search of harmful programs in the distributed information systems characterised by the indistinct information on proceeding processes is considered. The mechanism of the indistinct logic conclusion based on condition-action rules of decision-making is used. The received results can be used for anti-virus diagnosing of the software of various computer systems.
Keywords: information systems, the software.

Введение. Развитие и внедрение компьютерных технологий и построение на их основе различных информационных систем (управляющих, обучающих, справочных и др.) и методов доступа к информационным ресурсам отвечает современным требованиям научно-технического прогресса и получили в настоящее время достаточно широкое применение.

Основной задачей при этом является, прежде всего, обеспечение защиты пользовательских программ и операционных систем от влияния различных вредоносных программ (ВП), создающих потенциальную угрозу особенно со стороны удаленных станций.

Особое место среди ВП занимают различные вирусные, а также троянские программы, которые, в отличие от вирусных программ, способны проникать в систему с целью хищения информации, что представляет собой особую опасность.

Имеющиеся многочисленные факты хищения информации в конкретных системах можно объяснить тем, что существующие антивирусные программы во многом ориентированы на выявление известных ВП, которые достаточно исследованы и классифицированы и не достаточно адаптированы на распознавание новых подозрительных объектов, проникающих в систему.

В технологиях поиска ВП, как известно, различаются два независимых аспекта [1]: технический и аналитический. Технический компонент – это система сбора информации (о файловой системе, файлах и их цифровом содержимом и др.), на основании которой аналитическая часть в соответствии с принятой политикой безопасности оповещает пользователя о потенциальной угрозе и запрашивает у него дальнейшие указания относительно ВП. К основным методам сбора информации относятся:

- считывание и работа с файлом как с массивом байтов;
- эмуляция кода программы и запуск команд в виртуальной копии компьютера;

- запуск программы в «песочнице» (sandbox), а также использование технологий виртуализации;
- мониторинг и регистрация всех событий, происходящих в операционной системе;
- поиск системных аномалий относительно некого эталонного состояния.

Рассмотренные средства обработки содержат в себе элементы искусственного интеллекта, требующие достаточно сложных алгоритмов построения аналитической системы, включая экспертные системы или нейронные сети.

Цель статьи. С учетом приведенного анализа актуальность приобретает формирование технологии поиска и идентификации неизвестных ВП, которая сочетает элементы технологии поиска аномалий и экспертных систем. Последнее определяет цель данной работы.

Результаты исследования. Для достижения цели в работе использован математический аппарат нечеткой логики (НЛ), который в последнее время эффективно используется для решения ряда задач со слабоструктурированными данными.

Механизмы нечеткой логики позволяют строить модели предметной области, адекватные реальным, на основе семантического описания объекта исследования и знаний экспертов, что на много проще разработки сложных математических моделей. При этом, с помощью специализированных методов обработки нечисловой информации обеспечиваются достаточно точные решения [2].

Предварительно рассмотрим некоторые понятия теории нечетких множеств, необходимые для дальнейшего изложения. Нечеткое множество A определяется как множество упорядоченных пар (кортежей) вида: $(\mu_A(u),u)$, где $\mu_A(u)$ – степень принадлежности элемента множества $u \in U$ нечеткому множеству A, u – является элементом универсального множества U. Следовательно,

$$A = \{\mu_A(u_i),u_i\};\ i=1,\ldots,k. \quad (1)$$

Степень принадлежности определяется некоторым действительным числом из интервала [0, 1]. Функция принадлежности позволяет для произвольного элемента универсального множества вычислить степень принадлежности его нечеткому множеству. Нечеткое множество связано с лингвистической переменной и их нечеткими переменными, которые позволяют определить ее значение с помощью функции принадлежности.

Например, лингвистическая переменная: “Категория программы“. Её значениями могут быть нечеткие переменные, образуемые с помощью модификаторов (или, не, очень и др.) в виде: не опасная, немного опасная, очень опасная и т.д. Функция принадлежности, если она задана, позволяет оценить далее лингвистическую переменную. Применительно к высказываниям в нечеткой логике применимы функции алгебры логики, которые имеют не численные, а лингвистические значения истинности, которые адекватны степени принадлежности. Основой семантического описания исследуемой модели является выбор и анализ входных и выходных лингвистических переменных, а также их значений, которые получаются с помощью мониторинга и экспертных оценок. Лингвистическая переменная считается заданной, если для нее определены: наименование N, универсальное множество A (область рассуждений), базовое терм–множество T (нечеткие переменные – значения лингвистической переменной), функция принадлежности μ и семантические процедуры преобразования (модификации) переменных термов лингвистической переменной.

Для исследования нечетких моделей используется механизм логического вывода, позволяющий получить четкий результат функционирования объекта, описываемого нечетким множеством переменных. Данный механизм содержит следующие этапы: фаззификация, нечеткий логический вывод, дефаззификация (приведение к четкости) [3,4].

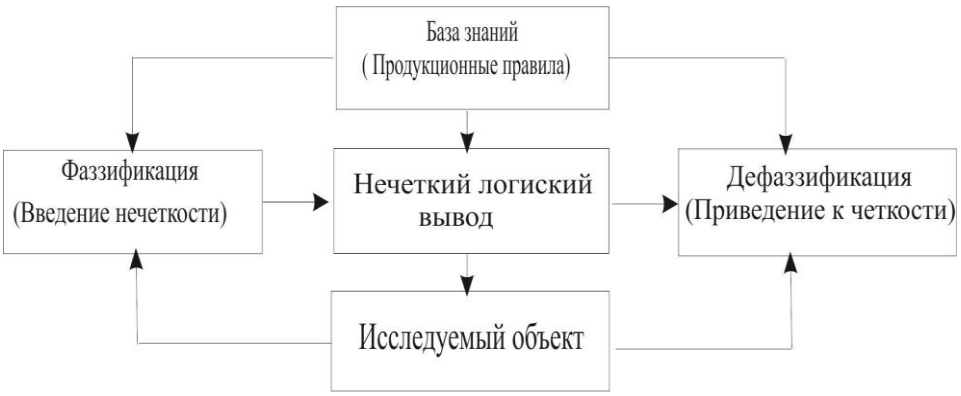

Рис. 1. Обобщенная структура логического вывода

Фаззификация – процесс перевода входных четких величин, полученных на основе экспертного анализа исследуемого объекта, в нечеткие переменные. На этапе нечеткого логического вывода используются нечеткие продукционные (if – then) правила базы знаний, для преобразования нечетких входных данных после фаззификации в нечеткие выходные. Дефаззификация – процесс композиции нечетких множеств логического вывода и приведение к четкости (получению результата) для принятия решения.

Действия ВП рассматриваются на этапах проникновения в систему, активизации и деструктивных действий, которые образуют один трехуровневый жизненный цикл (ЖЦ). Логический вывод о наличии ВП реализуется с помощью нечеткой модели (НМ), которая основана на семантическом описании и анализе множества возможных состояний ВП в течении жизненного цикла:

$$НМ \rightarrow \{(X,Y,V,Z)\}, \quad (2)$$

где:

- X – универсальное множество входных состояний $x \in X$, принимающих ВП в системе пользователя (функции и действия, используемые ВП для проникновения в систему);
- Y – универсальное множество выходных состояний $y \in Y$ (компоненты, которые потенциально могу быть подвержены действию ВП);
- V – множество нечетких отношений (x_i, y_j), характеризующее связи между элементами множеств X и Y;
- Z – характеристические параметры отношений.

На множествах X, Y определены нечеткие множества ($X^п$, $X^а$,$X^д$) и ($Y^п$, $Y^а$,$Y^д$) входных и выходных состояний ВП, где индексы п,а,д соответствуют этапам ЖЦ (проникновение в систему, активизация и деструктивные действия). В соответствии с (1) множества $X^п$, $X^а$, $X^д$ определяются, как множества упорядоченных пар: $X^п = \{\mu_X^п(x),x\}$; $X^а = \{\mu_X^а(x),x\}$; $X^д = \{\mu_X^д(x),x\}$, где $\mu_X^п(x)$, $\mu_X^а(x)$ и $\mu_X^д(x)$ – функции принадлежности, указывающие на степень принадлежности элемента x множества X множествам $X^п$, $X^а$,$X^д$ соответственно. Множества ($Y^п$, $Y^а$, $Y^д$) представляются в виде множеств отношений (x_i, y_j):

$Y^п = \{(x_i, y_j, \mu_Y^п(x_i, y_j)\}$; $Y^а = \{(x_i, y_j, \mu_Y^а(x_i, y_j)\}$; $Y^д = \{(x_i, y_j, \mu_Y^д(x_i, y_j)\}$,

где выражения $\mu_Y^п(x_i,y_j)$, $\mu_Y^а(x_i,y_j)$ и $\mu_Y^д(x_i,y_j)$ являются функциями принадлежности нечетких отношений (x_i,y_j) множествам $Y^п$, $Y^а$,$Y^д$ соответственно, которые определяют степень уверенности в существовании причинно–следственной связи (x_i, y_j). Нечеткие отношения множеств $Y^п$, $Y^а$,$Y^д$ далее и представляются соответствующими матрицами отношений:

$V_п = |x_i, y_j|$, $V_а = |x_i, y_j|$, $V_д = |x_i, y_j|$,

в которых на пересечении x_i, y_j находятся значения функции принадлежности $\mu_Y^п(x_i,y_j)$, $\mu_Y^а(x_i,y_j)$ и $\mu_Y^д(x_i,y_j)$ соответственно.

В качестве входной лингвистической переменной на множестве X принята: “Степень подозрительности”. Ее характеристикой является терм–множество T, которое содержит три

нечетких переменных (значения лингвистической переменной): “Малая подозрительность M(x)”, “Средняя подозрительность C(x)” и “Высокая подозрительность B(x)”, с областью определения [0, 1]. На практике число переменных терма не превышает семи. Для физической реализации лингвистической переменной необходимо определить значения переменных терма T при заданных входных переменных x. Для этого должны быть найдены функции принадлежности: $\mu_X^{п}(M)$, $\mu_X^{п}(C)$, $\mu_X^{п}(B)$; $\mu_X^{а}(M)$, $\mu_X^{а}(C)$, $\mu_X^{а}(B)$, а также $\mu_X^{д}(M)$, $\mu_X^{д}(C)$, $\mu_X^{д}(B)$. На практике с этой целью часто используется ряд типовых форм кривых: треугольная, трапецеидальная и гауссова функция принадлежности [4]. Так, на рис. 3 а) для нечетких переменных M(x), C(x), B(x) на множествах $X^{п}$, $X^{а}$, $X^{д}$ их функции принадлежности определены при входных значениях $x_1 = 2$, $x_2 = 5$, $x_3 = 9$: (0.62, 0.78, 0.37), (0.9, 0.5, 0.25) и (0.42, 0.28, 0.5), соответственно. Входные значения определяются и приводятся к определенному формату экспертом по результату мониторинга. Выходную лингвистическую переменную зададим на множестве Y. В качестве такой переменной принята: “Опасность поражения“, значение которой будет определяться на этапах ЖЦ нечеткой переменной: “Степень поражения“ с функциями принадлежности: $\mu_Y^{п}(x_i, y_j)$, $\mu_Y^{а}(x_i, y_j)$ и $\mu_Y^{д}(x_i, y_j)$. В данной постановке задачи, для оценки лингвистической переменной также учитывается вид поражения: зависание системы, потеря и искажение файлов, нарушение реестра и т.д., что учитывается вектором $Z = \{z_k\}$, в котором компонентам z_k, $k = 1,..,r$, (виды поражения) присваиваются нормированные приоритетные веса $P = \{p_k\}$, ($\sum p_k = 1$), учитывающие уровень их опасности для системы пользователя.

Рассмотрим содержательную часть матриц отношений, построенных с учетом [5]. Так, в матрице отношений $V_п = |x_i, y_j|$ элементы x_i – функции (механизмы) и способы проникновения ВП в систему пользователя, $i = 1,..,k$; элементы y_j – возможные входы проникновения в систему: порты сетевых протоколов прикладного уровня, $j = 1,..,h$. (индексы не совпадают с истинными номерами портов). Например: x_1 – набор функций работы с электронной почтой; x_2 – набор функций работы с системами передачи файлов; x_3 – набор функций отдаленного доступа (через консоль и командные интерпретаторы); x_4 – набор функций работы с web–браузером и т.д. Порт p_{20} – FTP (FileTransferProtocol) – сетевой протокол, предназначенный для передачи файлов в компьютерных сетях; p_{22} – SSH (Secure Shell) – сетевой протокол, который позволяет осуществлять отдаленное управление компьютером и передачу файлов и т.д.

На этапе деструктивных действий используется матрица отношений исследуемого объекта и структурных компонент операционной системы (ОС) : $V_д = |x_i, y_j|$, в которой x_i – действия ВП, направленные на структурные компоненты ОС пользователя; $i = 1,..,\sigma$; y_j – структурные компоненты ОС пользователя, $j = 1,..,\tau$. Например, x_1 – прием и отправление файлов; x_2 – создание и уничтожение файлов; x_3 – создание, запуск и уничтожение процессов и т.д. К структурным компонентам ОС относятся: y_1 – файловая система; y_2 – планировщик процессов (как составляющая ядра операционной системы); y_3 – системные службы (ключи реестра – Windows XP) и т.д.

Из приведенных примеров видно, что поведение ВП на этапах ЖЦ многовариантно, имеет нечеткий характер и, в принципе, не может быть просто прогнозировано. Кроме того, их функции принадлежности не определены, что делает невозможным оценки выходной лингвистической переменной. В работе данные функции определяются, в отличие от типовых подходов, косвенным путем на основе анализа и оптимизации матриц нечетких отношений $V_п$, $V_а$, $V_д$ с привлечением оценок экспертов.

Покажем на примере принцип формирования функции принадлежности $\mu_Y^{п}(x_i, y_j)$ на этапе проникновения. Рассматриваемая задача представляется как нахождение для каждого x_i матрицы $V_п = |x_i, y_j|$ наиболее вероятного порта проникновения y_j при заданных признаках опасности z_k. Данная задача сводится к задаче ранжирования, в которой в отличие от [4], попарное сравнение экспертов производится с учетом оценочных признаков, позволяющих учесть особенности сравниваемых объектов.

Исходным для решения задачи является построение матрицы превосходства $S=|s_{ij}|$, элементы которой s_{ij} выражаются любым положительным числом ($s_{ij} = s_i/s_j$; $0 < s_{ij} < \infty$; $s_{ji} = 1/s_{ij}$; $s_{ii} = 1$; i, j = 1,…,m, где m – число возможных исходов). Элементы s_{ij} матрицы S определяются вычислением значений парных предпочтений по каждому признаку отдельно с учетом их весов $P=\{p_k\}$; k=1,..,r, с использованием выражения:

$$s_{ij} = \frac{\sum_{k=1}^{r} s_{ij}^k \cdot p_k}{\sum_{k=1}^{r} s_{jk}^k \cdot p_k}; \; s_{ji} = \frac{1}{s_{ij}}; \; s_{ii} = 1; \; i, j = \overline{1,m}.$$

С помощью матрицы S определяется собственный вектор $\Pi = (\pi_1,...,\pi_m)$, который соответствует максимальному положительному корню λ характеристического полинома $|S - \lambda \cdot E| = 0$; $S \cdot \Pi = \lambda \cdot \Pi$, где E – единичная матрица. Компоненты вектора Π ($\sum\pi_i = 1$) отождествляются с оценкой $\mu_Y^{\text{п}}(x_i, y_j)$, учитывающей принятые признаки опасности. Подобная процедура производится для всей матрицы $V_{\text{п}} = |x_i, y_j|$, в результате чего получаем оптимизированную матрицу отношений $V_{\text{п}} = |x_i, y_j|$, в которой используются лишь отношения x_i, y_j с наиболее выраженным уровнем опасности, определяемым значением, равным π_{max}, ($0 \le \pi_{max} \le 1$). На основе полученной матрицы строится нормированная кривая функция принадлежности $\mu_Y^{\text{п}}(x_i, y_j)$ выходной переменной y, а также таблица, в которой каждому значению $\mu_Y^{\text{п}}(x_i, y_j)$ ставится в соответствие пара (x_i, y_j), позволяющая идентифицировать, в том числе, ресурсы, которые были использованы вредоносной программой на рассматриваемом этапе (рис. 2). Для компактности отношения (x_i, y_j) заменены символом R.

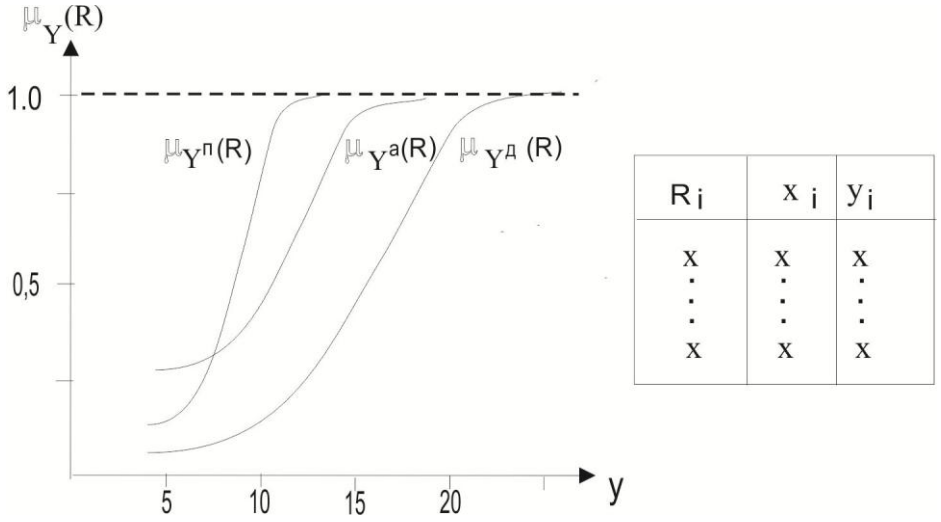

Рис. 2. Примерный вид функций принадлежности выходной переменной

Основой для нечеткого вывода служит база нечетких продукционных правил (НПП), определяющих стратегию решения задачи. Типичное продукционное правило базы правил состоит из посылки (антецедента): нечеткие высказывания в форме «если…» и заключения (консеквента) – в форме «то…». Антецедент может содержать несколько посылок, которые объединяются в зависимости от стратегии посредством логических связок «и» или «или».

Каждое из правил нечетких продукций может иметь некоторый вес $F_i = [0, 1]$, который определяет значимость правила или уверенность в степени истинности заключения, получаемого по отдельному нечеткому правилу. В общем виде база, содержащая m правил, имеет вид:

Π_1: если x_1 это A_{11} … и (или) … x_n это A_{1n}, то y это R_1;

Π_i: если x_1 это A_{i1} … и (или) … x_n это A_{in}, то y это R_i;

Π_m: если x_1 это A_{i1} … и (или) … x_n это A_{mn}, то y это R_m,

где: x_k – входные переменные, k=1,…,n; y – выходная переменная;

A_{ik} – заданные нечеткие множества с функциями принадлежности (i=1,…,m; k=1,…,n).

На основе нечетких высказываний, истинность которых установлена в результате фаззификации, оценивается степень истинности нечетких высказываний, являющихся заключением соответствующих НПП. Далее выполняется процедура (агрегирование) определения степени истинности левых частей (уровней отсечения – a_i) по каждому из правил системы нечеткого вывода. Так, дизъюнкцией нечетких высказываний является логическая операция, результатом которой является нечеткое высказывание, определяемое как:

$$a_i = \max_i (A_{ik}(X_k)).$$

Приведем для примера фрагмент базы правил (три переменные и три правила), в которой для каждого правила использованы связки «или» и значения F_i равны единице:

П1: если x_1 = M, или x_2 = C, или x_3 = B, то y = $R^{п}$;

П2: если x_1 = M, или x_2 = C ,или x_3 = B, то y = $R^{а}$;

П3: если x_1 = M, или x_2 = C, или x_3 = B, то y = $R^{д}$.

Возможны и другие стратегии получения базы правил.

В работе используется стратегия на основе правила max–min композиции и нечеткой операции max – дизъюнкции для оценки одинаковых заключений. В рассматриваемом фрагменте при заданных x_1, x_2, x_3 правила позволяют получить нечеткое заключение о степени истинности подозрительности на исследуемых этапах:

$\mu_1 = \max\{\mu_x^{п}(M); \mu_x^{п}(C); \mu_x^{п}(B)\} = \max\{0.62; 0.78; 0.37\} = 0.78.$

$\mu_2 = \max\{\mu_x^{а}(M); \mu_x^{а}(C); \mu_x^{а}(B)\} = \max\{0.9; 0.5; 0.25\} = 0.90.$

$\mu_3 = \max\{\mu_x^{д}(M); \mu_x^{д}(C); \mu_x^{д}(B)\} = \max\{0.42; 0.28; 0.5\} = 0.5.$

Следующий этап – активизация: процесс нахождения степени истинности каждого из заключений нечетких продукционных правил, который осуществляется усечением функций принадлежности выходной переменной на уровнях μ_1, μ_2, μ_3, как показано на рис. 3 б):

$$\mu_Y^{п}(R)= 0{,}78;\ \mu_Y^{а}(R) = 0{,}90;\ \mu_Y^{д}(R) = 0{,}5.$$

Для композиции (объединения) полученных усеченных функций используется максимальная композиция нечетких множеств

$$M_y(R) = \max_i (\mu_y(R)),$$

где: $M_y(R)$ – функция принадлежности итогового нечеткого множества.

На рис. 3 показана графическая интерпретация нечеткого вывода для трех входных переменных и трех нечетких правил.

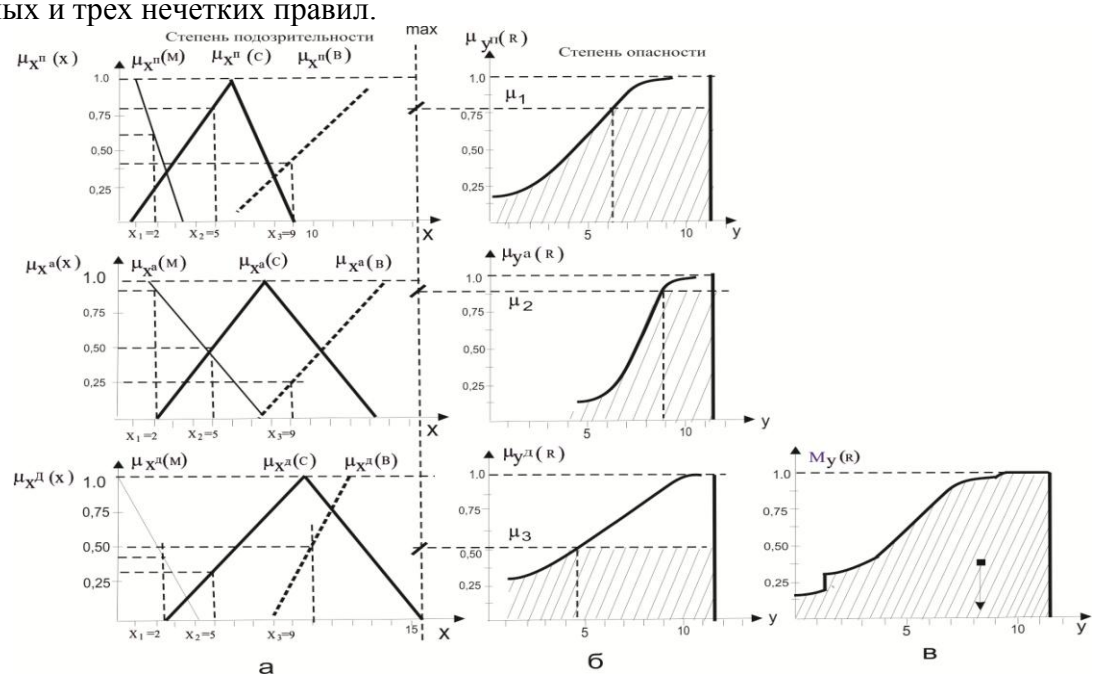

Рис. 3. Графическая интерпретация логического вывода

На рис. 3 этапы интерпретации логического вывода а), б), в) выполняют следующие функции: а) - представление функции принадлежности для нечетких переменных M(x), C(x), B(x) на множествах $X^{п}$, $X^{а}$, $X^{д}$ при входных значениях $x_1 = 2$, $x_2 = 5$, $x_3 = 9$;

б) - нахождение степени истинности заключений нечетких продукционных правил на уровнях отсечения μ_1, μ_2, μ_3;

в) - приведение к четкости логического вывода (получение численного значения).

Заключительным является этап дефаззификации – приведение к четкости (получение численного значения для принятия решениия) (см. рис. 3 в)). Для дефаззификации в работе использован метод центра тяжести [4], при котором значение выходной переменной равно абсциссе центра тяжести площади, ограниченной графиком функции принадлежности итогового нечеткого множества $M_y(R)$.

$$y = \frac{\sum_{i=1}^{n} y_i M_y(R)i}{\sum_{i=1}^{n} M_y(R)i} = \frac{11{,}3 \cdot 0{,}78+9 \cdot 0{,}9+5 \cdot 2{,}5}{0{,}78+0{,}9+0{,}5} = 8,9 .$$

Полученный результат интерпретируется как коэффициент опасности поражения системы со стороны ВП, который сравнивается с некоторым нормированным коэффициентом, определяемым принятой стратегией безопасности. Обобщенная схема реализации предложенной методологии приведена на рис. 4, которая содержит подсистемы мониторинга и принятия решения.

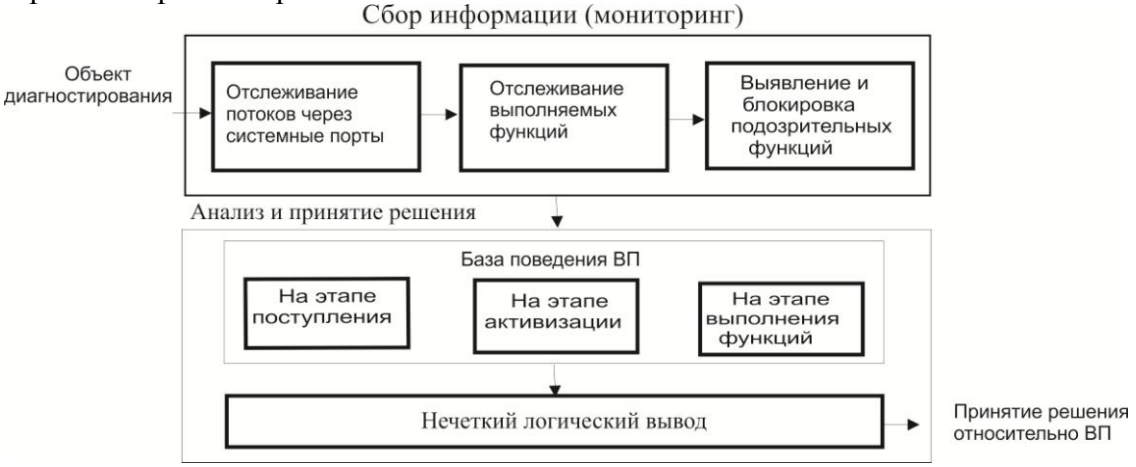

Рис. 4. Обобщенная схема поиска вредоносных программ

Выводы. В работе предложена новая методология и ее использование для поиска и идентификации ВП на основе методов нечеткой логики, позволяющей более просто получать достаточно точные решения на основе семантического описания задачи и нечетких продукционных правил.

Существующим методам присущи следующие недостатки: невысокое быстродействие, низкая вероятность идентификации новых ВП, высокие требования отдельных методов к аппаратному обеспечению, сложность модели и ее реализации.

Установлено, что применение нечеткой логики открывает большие возможности для адаптации к любым моделям ВП, а также принятия оптимального решения путем проектирования соответствующей базы продукционных правил и стратегии их обработки. При этом обеспечивается повышение быстродействия за счет исключения из алгоритмов поиска достаточно сложных компонент и процедур (баз сигнатур, эвристических анализаторов, использование контрольных сумм и т.д.), свойственных существующим методам, а также достаточно высокая реакция и незначительная вероятность пропуска неизвестных ВП.

Результаты работы могут быть использованы разработчиками антивирусного диагностирования. Рабочей средой для разработки и исследования предложенной методологии является программная оболочка FUZZY EXPERT.

ЛИТЕРАТУРА:

1. Технологии обнаружения вредоносного кода. Эволюция. Алиса Шевченко. www.securelist.com

2. Заде Л. А. Понятие лингвистической переменной и его применение к принятию приближенных решений. М: Мир, 1976. – 65с.

3. Леоненков А.В. Нечеткое моделирование в среде MATLAB в fuzzyTECH. – СПб.: БХВ – Петербург, 2005. – 736 с.

4. Штовба С. Д. Проектирование нечетких систем средствами МАТЛАБ. М:, Горячая линия – телеком – 2007. – 290 с.

5. Х–релиз: исходники трояна // Информационный портал Хакер.Ру. – Режим доступа: http://www.xakep.ru/post/17223/

Рецензент: д.т.н., проф. Лєнков С.В.