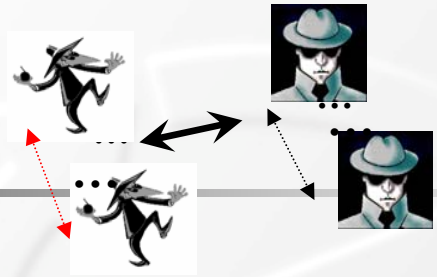
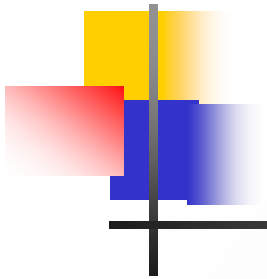


СПИИРАН

КОМБИНИРОВАНИЕ МЕТОДОВ DATA MINING ДЛЯ СТАТИЧЕСКОГО ДЕТЕКТИРОВАНИЯ MALWARE

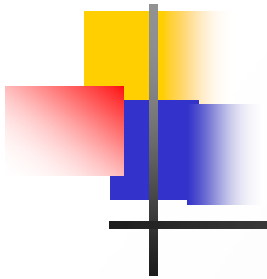
Комашинский Д.В., Котенко И.В.

Санкт-Петербургский институт информатики и автоматизации РАН



Содержание

- Фокус работы
- Общий подход к статическому анализу
- Релевантные работы
- Проведенные эксперименты
- Заключение



Фокус работы

Работа посвящена следующим вопросам:

«Существует ряд существенно различающихся статических подходов к обнаружению malware, основанных на применении методов Data Mining и доказавших свою потенциальную жизнеспособность. Каждый из них воплощает некоторую стратегию принятия решения, эффективность которой определяется как минимум используемым набором статических признаков и используемым классификатором. ***Насколько можно повысить общую точность принятия решения при формировании комбинации классификаторов? Какие из базовых методов комбинирования результатов работы классификаторов наиболее эффективны? Является ли эффективной стратегия комбинирования путем построения иерархической схемы принятия решения?»***

Общий подход к статическому анализу

Существует два основных подхода, нацеленных на выполнение задачи детектирования вредоносного ПО:

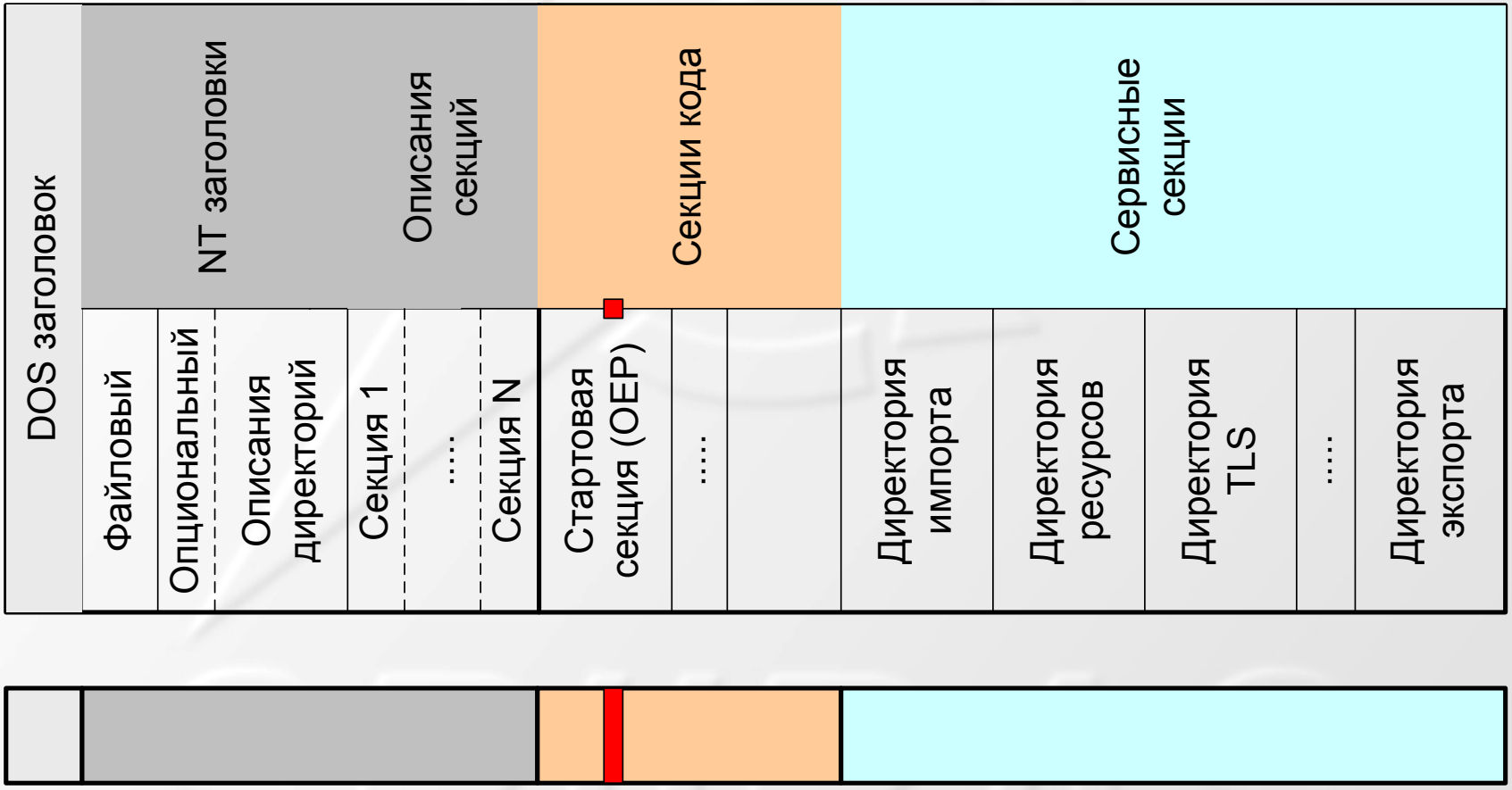
- решение на основе данных, которые могут быть получены без выполнения анализируемой компьютерной программы (так называемая группа методов статического анализа)
- решение на основе данных, полученных при ее выполнении (группа методов динамического анализа или обработки динамической или поведенческой информации).

Общий подход к статическому анализу

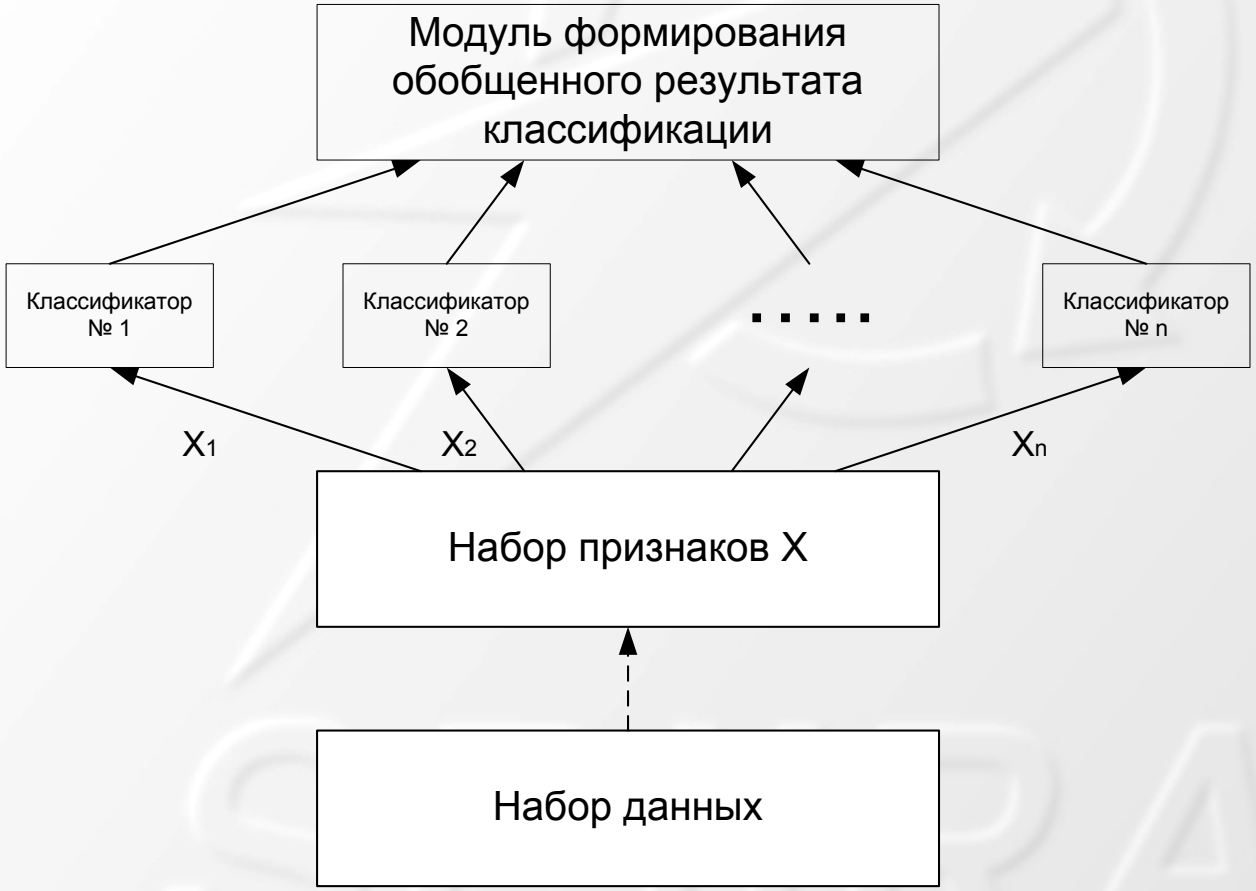
Формат исполняемых файлов PE32 предоставляет большое количество информации, которое может быть использовано при выделении исследователем потенциально значимых признаков:

- Заголовочная информация, описывающая структурные аспекты объекта;
- Программный код, описывающий низкоуровневые функциональные аспекты объекта;
- Вспомогательные таблицы, позволяющие сделать прогноз о высокоуровневых функциональных аспектах объекта;
- Данные, используемые объектом при выполнении и интеграции в исполняющую среду.

Потенциально значимые признаки



Уровни комбинирования при построении мета-классификаторов



1. Комбинационный уровень

2. Классификационный уровень

3. Признаковый уровень

4. Уровень данных

Релевантные работы

- Dai J., Guha R., Lee J. **Efficient Virus Detection Using Dynamic Instruction Sequences** // Journal Of Computers, Vol. 4, NO 5, May 2009.
- Kolter J., Maloof M. **Learning to Detect Malicious Executables in the Wild** // Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.
- Schultz M., Eskin E., Zadok E., Stolfo S. **Data Mining Methods for Detection of New Malicious Executables** // Informatics and Computer Science, Volume 172, Issue 1-2, 2001.
- Wang J.-H., Deng P.S., Fan Y.-S., Jaw L.-J., Liu Y.-C. **Virus Detection using Data Mining Techniques** // Proceedings of IEEE 37th Annual 2003 International Carnahan Conference, 2003.

Релевантные работы

- Komashinskiy D., Kotenko I. **Malware Detection by Data Mining Techniques Based on Positionally Dependent Features** // Proceedings of the 18th Euromicro International Conference on Parallel, Distributed and Network-Based Computing
- Menahem E., Shabtai A., Rokach L., Elovici Y. **Improving Malware Detection by Applying Multi-Inducer Ensemble**. Computational Statistics and Data Analysis, Vol. 53 Iss. 4.

Проведенные эксперименты

Фактически **производится воспроизведение ряда известных работ**, использующих следующие в качестве признаков:

- данные, доступные из базовых заголовков, характеризующих структурные особенности анализируемого файла;
- данные секции импорта, представляющие базовый набор функций операционной системы, используемых приложением при функционировании;
- данные о наличии в исполняемых секциях анализируемого объекта тех или иных байтовых последовательностях (т.н. n-грамм);
- позиционно-зависимые данные, извлекаемых из ограниченного региона вблизи точки входа (Entry Point) в анализируемый файл.

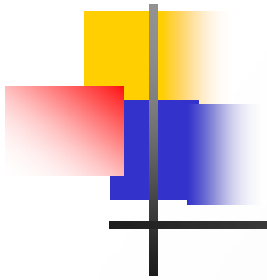
Проведенные эксперименты

Выделение из всего множества доступных признаков из указанных группа наиболее значимых производилось средствами критерия, основанного на вычислении коэффициента информационного усиления каждого признака, их приоритезации и выделения ограниченного набора из верхней части полученного списка.

Проведенные эксперименты

Для обучения отдельных решающих экспертов были применены методы классификации, продемонстрировавшие оптимальные показатели качества предиктивной функции в экспериментах, проведенных в данных работах: Naive Bayes и Decision Tree (C4.5).

Задача построения общего комбинированного классификатора решалась с использованием методов обобщения их конечных результатов: большинства голосов (Majority Vote), взвешенного большинства голосов (Weighted Majority Vote) и Байесовского комбинирования.

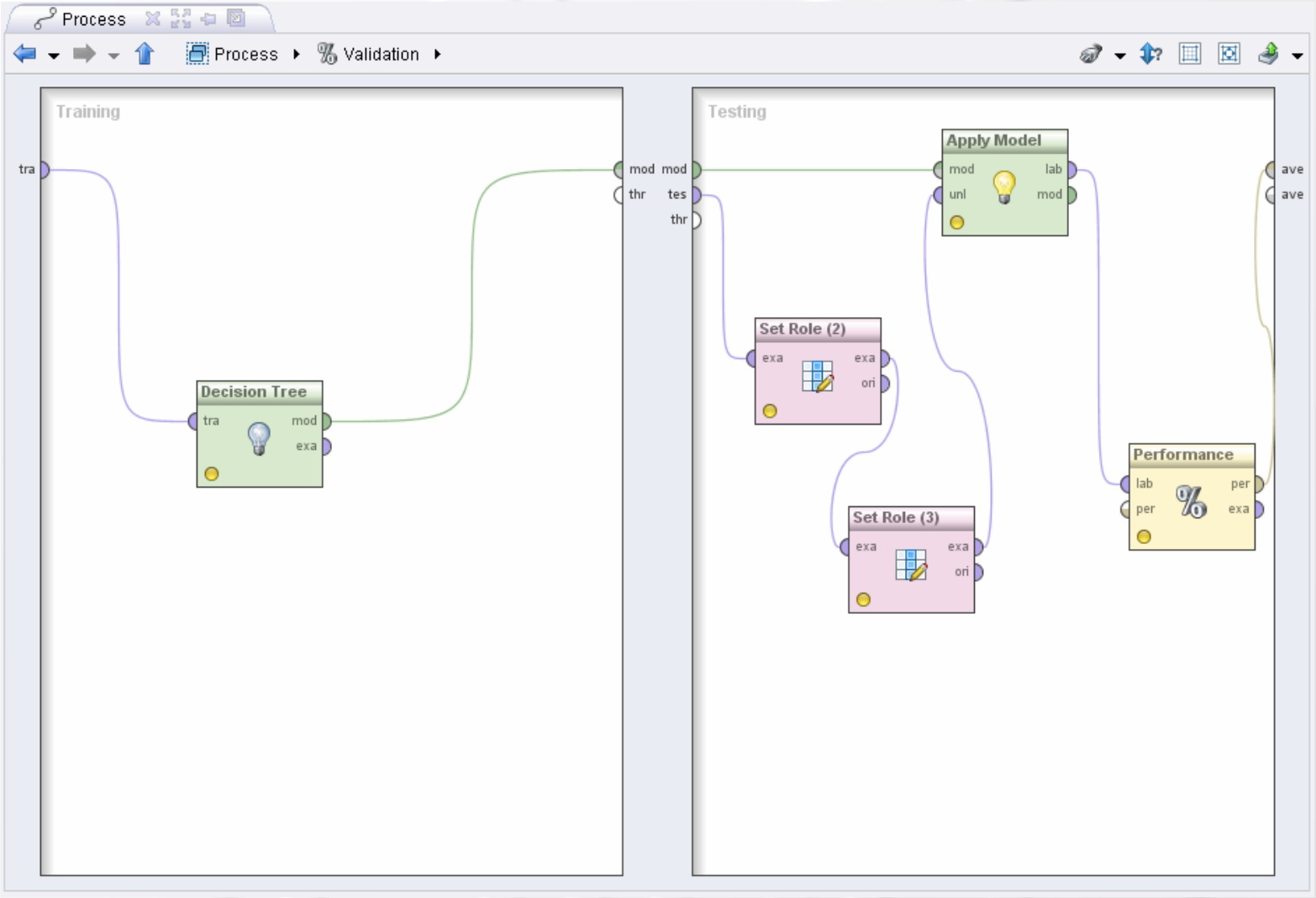


Проведенные эксперименты

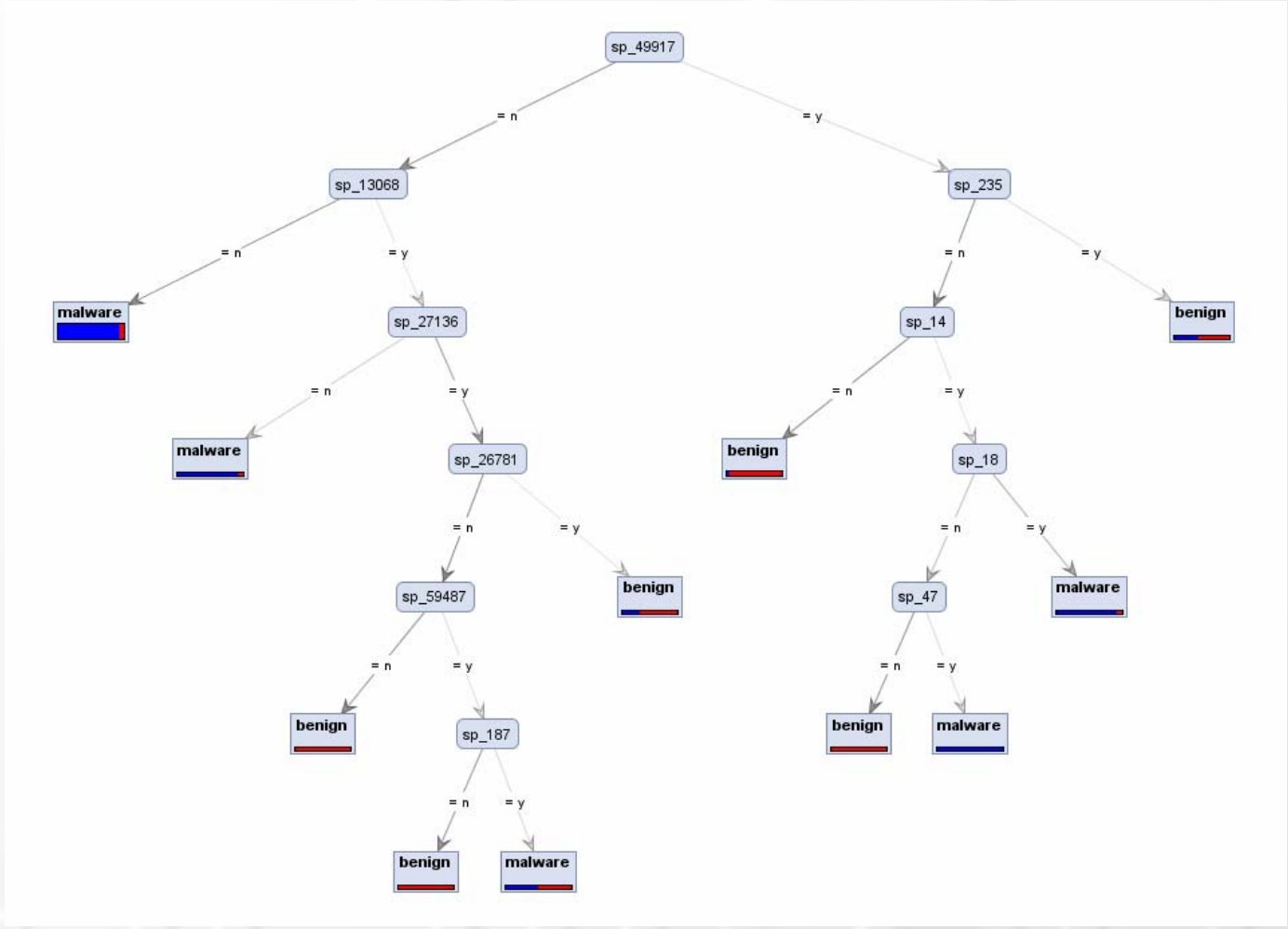
Для обучения классификаторов использовались файлы PE32, полученные с сайта VXHeavens, из системных каталогов операционной системы Windows XP.

Инструментальная поддержка экспериментов осуществлялась средствами программного пакета RapidMiner 5.0, собственными разработанными средствами парсинга файлов формата PE32 и форматирования исходных данных.

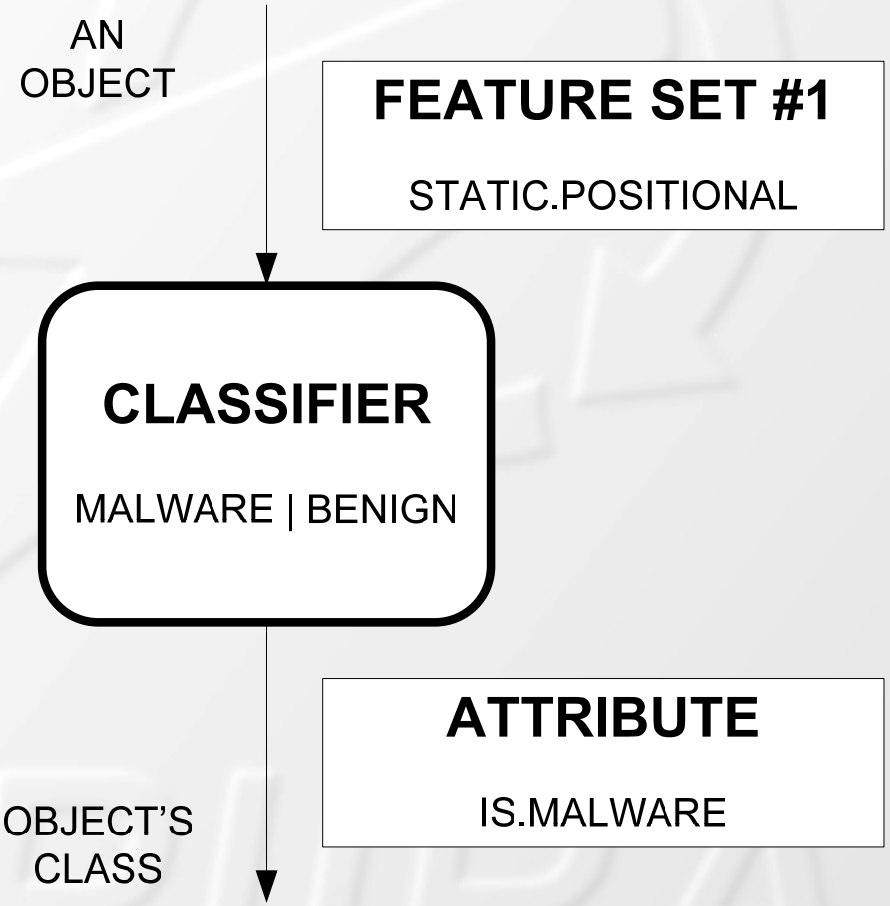
RapidMiner: пример схемы использования классификатора



RapidMiner: пример дерева решений



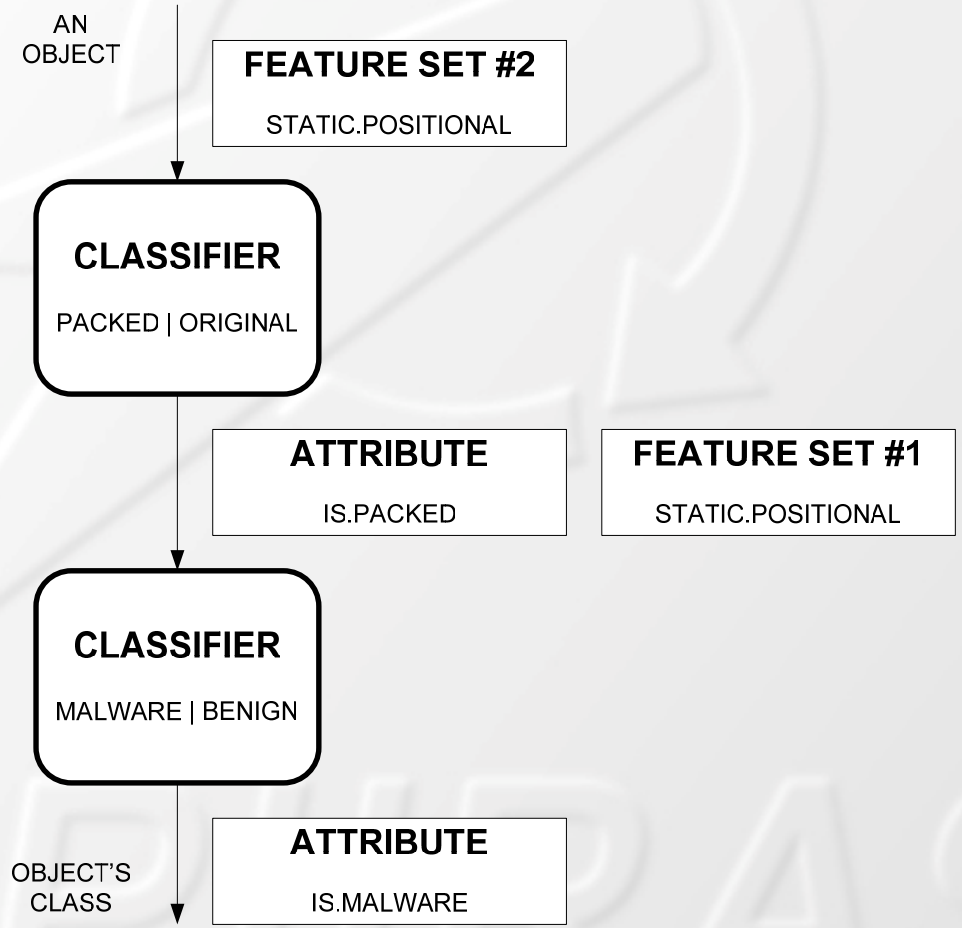
Проведенные эксперименты. Схема 1



Проведенные эксперименты. Схема 1

C4.5 Naive Bayes	True malware	True benign	Class precis.
Predicted malware	10354 10440	896 1176	92.04% 89.88%
Predicted benign	172 86	1277 997	88.13% 92.06%
Class recall	98.37% 99.18%	58.77% 45.88%	

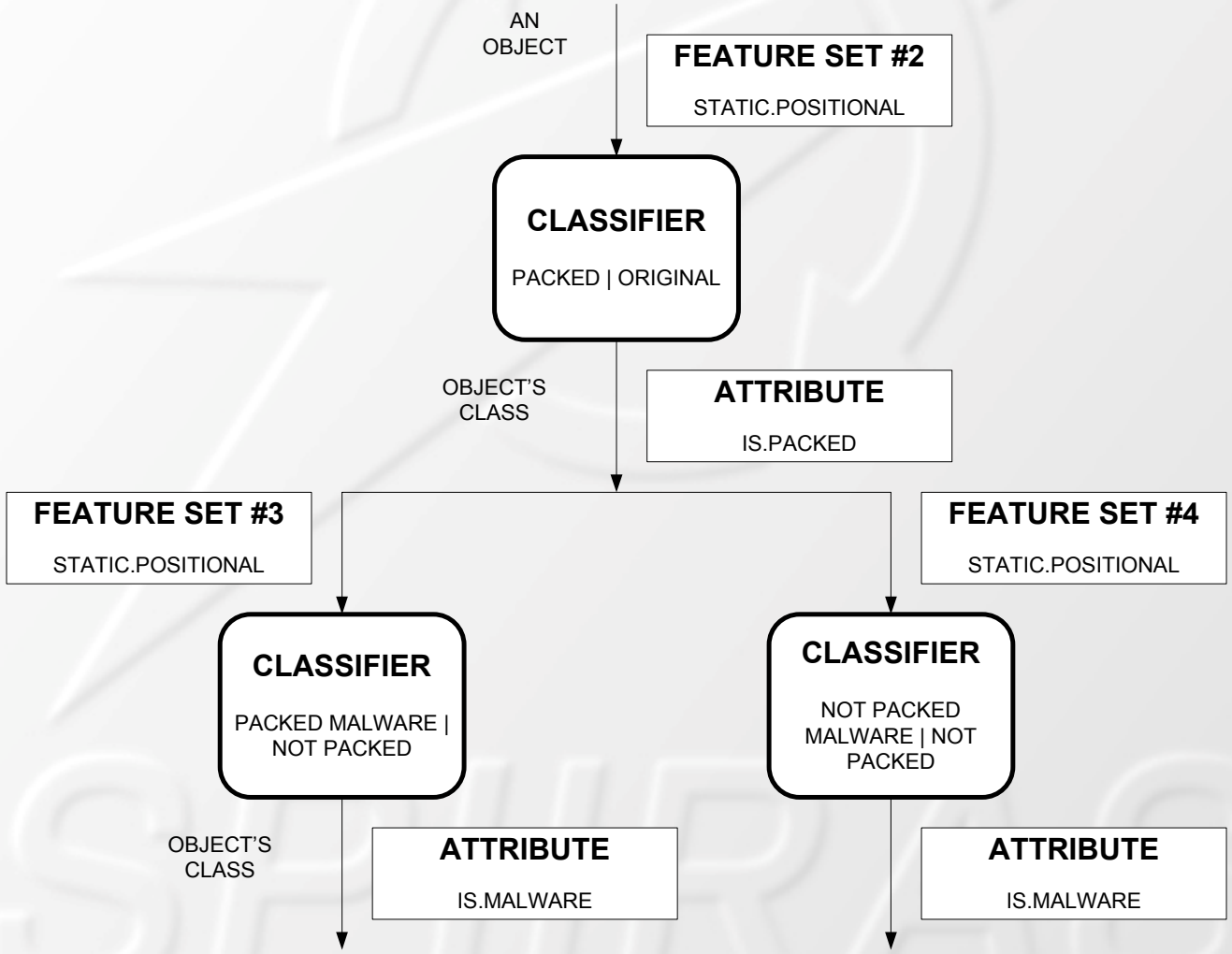
Проведенные эксперименты. Схема 2



Проведенные эксперименты. Схема 2

C4.5 Naive Bayes	True malware	True benign	Class precis.
Predicted malware	9768 9922	454 970	95.55% 91.01%
Predicted benign	758 604	1719 1203	69.39% 66.57%
Class recall	92.96% 94.26%	79.10% 55.36%	

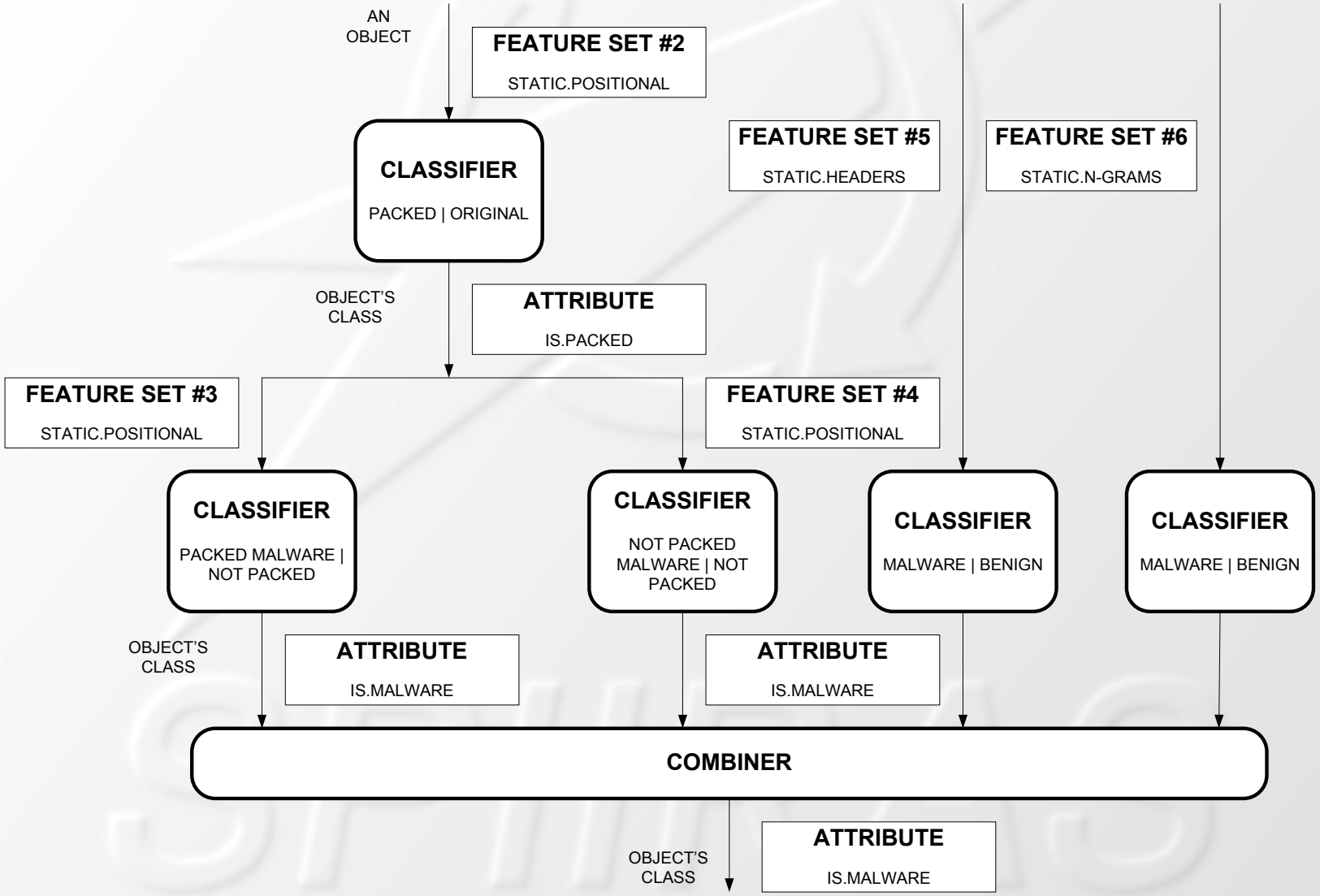
Проведенные эксперименты. Схема 3



Проведенные эксперименты. Схема 3

C4.5 Naive Bayes	True malware	True benign	Class precis.
Predicted malware	10174 9214	295 1239	97.18% 88.14%
Predicted benign	352 1312	1878 934	84.21% 41.58%
Class recall	96.65% 87.23%	86.46% 43.00%	

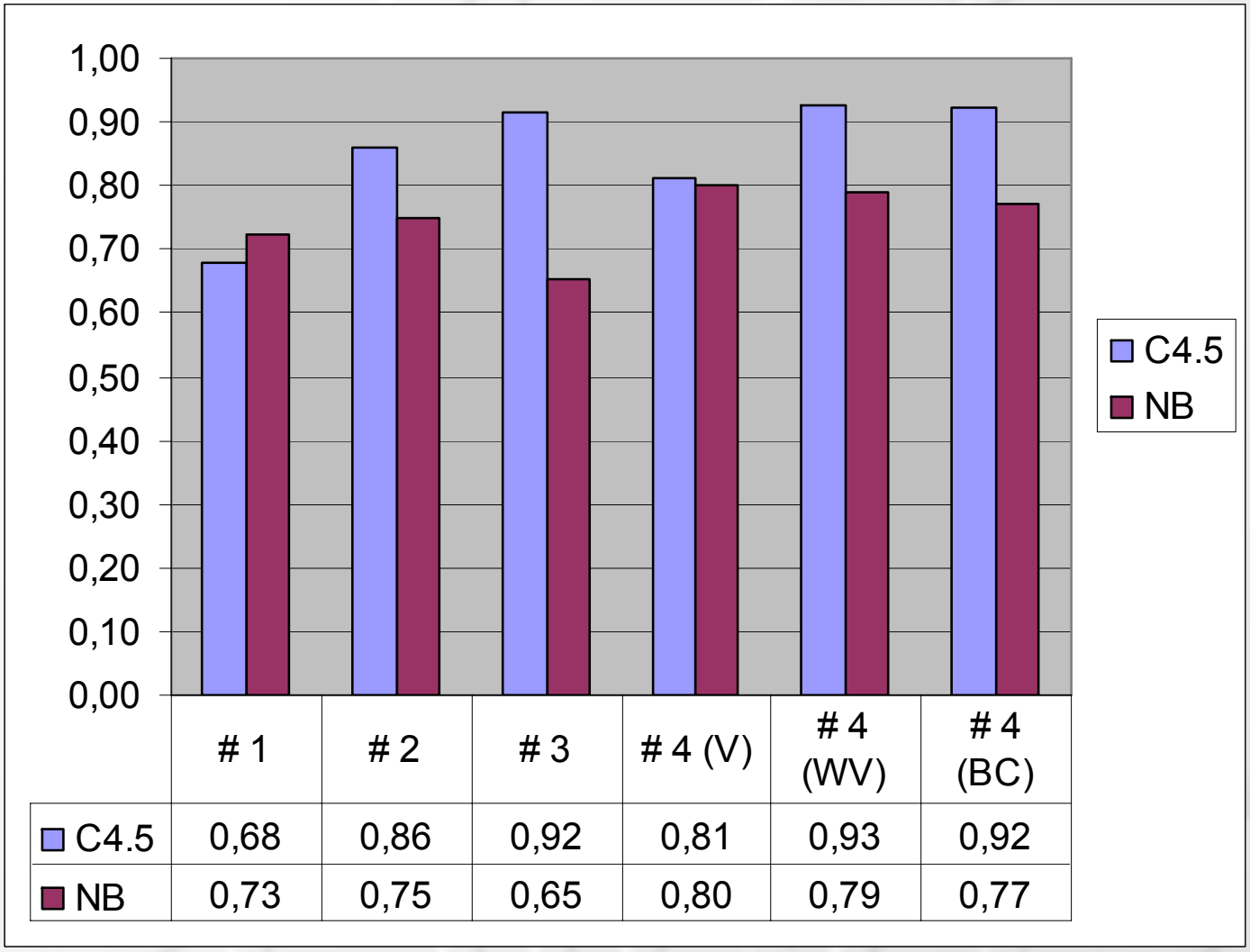
Проведенные эксперименты. Схема 4



Проведенные эксперименты. Схема 4

C4.5 Naive Bayes	True malware	True benign	Class precis.
Predicted malware	10252 9617	262 730	97.50% 92.94%
Predicted benign	274 909	1911 1443	87.45% 61.35%
Class recall	97.39% 91.36%	87.94% 66.40%	

Графическое представление результатов



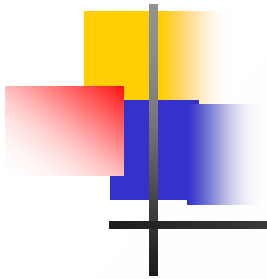
	# 1	# 2	# 3	# 4 (V)	# 4 (WV)	# 4 (BC)
C4.5	0,68	0,86	0,92	0,81	0,93	0,92
NB	0,73	0,75	0,65	0,80	0,79	0,77

Заключение

Обоснована необходимость использования иерархических моделей и методов комбинирования при использовании эвристических методов анализа.

Ближайшими задачами в данном направлении являются:

- Реализация методов и средств автоматической идентификации использованных средств защиты исполняемых файлов;
- Расширение пространства признаков дополнительными элементами, являющимися потенциально важными при осуществлении процессов обучения классификаторов;
- Оценка применимости данных улучшений при построении иерархической модели принятия решения.



Контактная информация

Комашинский Дмитрий Владимирович (СПИИРАН)
komashinskiy@comsec.spb.ru
http://comsec.spb.ru/Komashinskiy/

Котенко Игорь Витальевич (СПИИРАН)
ivkote@comsec.spb.ru
http://comsec.spb.ru/Kotenko/