# Feature Selection

Although there are many features which can be extracted from PE files, not all of these will contribute positively to the accuracy of our solution and the inclusion of some features for clustering may, in fact, result in reduced accuracy. To increase the accuracy of our solution we will be carrying out dimensionality reduction [1] via data mining. This will result in a lower dimensional [1] dataset which will contribute to performance improvements when processing the data, as well as the aforementioned accuracy increase.

The experiments carried out by Raman present a useful starting point for feature selection, achieving a 98.57% true positive rate with a 5.68% false positive rate. These results were achieved using only 7 independent features of the PE files which were data mined[2].

The results achieved by Shafiq et al. are marginally better with an average of 99% true positive rate and a 0.5% false positive rate. This increased accuracy was achieved by mining 189 features from the initial dataset [3].

Of the two sets of results cited above, I believe Raman’s process of feature selection was more effective in reducing the dimensionality of the data while still preserving the most important aspects of the initial dataset.

According to the results presented by Raman the following 7 features are the most accurate, independent, indicators of whether a file is malicious or benign [2]:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Description** | **Location** | **Offset (PE/PE32+)** | **Size (bytes)** |
| DebugSize | The size of the debug directory table. Malicious files may have a size of 0. | Optional Header Data Directories | 144/160 | 8 |
| ImageVersion | The file version number. Consisting of both Major and Minor image versions. Benign files tend to have higher version numbers, malicious files may have a value of 0. | Optional Header Windows-Specific Fields | Major: 44/44  Minor: 46/46 |  |
| IatRVA | The relative virtual address of the import address table. Benign files tend to have a value of 4096, malicious files tend to have a value of 0 or a very large value. | Optional Header Data Directories | 192/208 | 8 |
| ExportSize | The size of the export table. Benign files may have a non-zero value. | Optional Header Data Directories | 96/112 | 8 |
| ResourceSize | The size of the resource section. Some malicious files will have a value of zero. Benign files may have a large amount of resources. | Optional Header Data Directories | 112/128 | 8 |
| VirtualSize2 | The size of the second section.  Many malicious files only have one section. | Section Table (Section Headers) | 8 | 4 |
| NumberOfSections | The number of section headers in the section table. | Section Table (Section Headers) | N/A | N/A |

**Fig 1.** Proposed features. Consolidation of information available in [2] & [4].

# References

[1] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

[2] K. Raman, “Selecting Features to Classify Malware,” *InfoSec Southwest 2012*, pp. 1–5, 2012.

[3] M. Zubair Shafiq, S. Momina Tabish, F. Mirza, and M. Farooq, “‘PE-Miner: Mining Structural Information to Detect Malicious Executables in Realtime’ in Recent Advances in Intrusion Detection,” Springer Science + Business Media, 2009, pp. 121–141.

[4] “Microsoft Portable Executable and Common Object File Format Specification,” 2015. [Online]. Available: http://download.microsoft.com/download/9/c/5/9c5b2167-8017-4bae-9fde-d599bac8184a/pecoff\_v83.docx.