Detector coverage under the r-contiguous bits matching rule*

Fernando Esponda
Stephanie Forrest
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131-1386
{fesponda,forrest}@cs.unm.edu

March 27, 2002

Abstract

Some computer anomaly detection systems inspired by the immune system rely on the r-contiguous bits matching rule, whereby two strings of length l are said to match if they have at least r contiguous bits in common. In this paper, we derive a recurrence relation and its closed form solution for computing the number of strings matched by a single such string called a *detector*.

1 Introduction

Anomaly detection is a major concern in computer systems and there is an ever growing body of research in this area. Nevertheless, this problem has always existed in the context of living organisms striving to protect themselves from foreign pathogens. This observation has led some research groups to study the vertebrate immune system as inspiration for building better anomaly detection schemes for computers. Several components make up these systems but at the lowest level they rely on a set of detectors, usually implemented as strings, whose function is to classify novel strings as normal or anomalous by matching them in some fashion. In particular, the r-contiguous bits matching rule —proposed in [2] and used in [**?**, **?**, **?**, **?**]— has a plausible biological foundation. Two strings of length l are said to match under this rule if they exactly match in at least r contiguous bit positions. In this paper, we address the simplest of questions about this rule, namely, how many strings can a single detector match? We set forth a recurrence relation for computing this quantity and, by noting some of its properties, derive a closed form solution given the *binding strength* parameter r.

2 Detector Coverage

In order to proceed we will need to define some notation: [1]

- U the set of all binary strings of length l, $|U| = 2^l$.
- r the number of contiguous bits that must be equal between two strings for them to match, $r > 1$.
- b a string of length r.
- $\hat{b}$ denotes string b stripped of its last bit.
- $v \cdot b$, where $v \epsilon \{0, 1\}$, denotes bit v appended with string b.

*University of New Mexico Technical Report TR-CS-2002-03.
[1] Note that the following procedure is equivalent to that presented in [1] for efficiently generating detectors.

- w is a detector of length l.
- We refer to a substring of w starting at bit position i and ending at position j as $w[i \dots j]$.
- $C_l[b]$= the number of bit strings of length l unmatched by w that end with substring b.
- $T(l)$, the total number of strings of length l unmatched by w.

Given a binary string of length l we will first determine how many strings it doesn't match and then determine the number of matches by simply subtracting this value from $|U|$. The number of strings in U ending in substring b that are unmatched by a given detector w is defined in terms of the unmatched strings of length $l-1$, writing this as a recurrence we have:

$$\begin{aligned} &\text{if } l = r \\ &C_l[b] = \begin{cases} 0, \text{ if } b = w \\ 1, \text{ otherwise} \end{cases} \\ &\text{if } l > r \\ &C_l[b] = \begin{cases} 0, \text{ if } b \text{ is matched by } w[(l-r+1)\dots l] \\ C_{l-1}[0 \cdot \hat{b}] + C_{l-1}[1 \cdot \hat{b}], \text{ otherwise} \end{cases} \end{aligned} \tag{1}$$

Interpreting the binary string index as a decimal number, the amount of strings of length l that are unmatched by w can be expressed as the sum of the unmatched patterns for each of the 2^r possible string endings:

$$T(l) = \sum_{i=0}^{2^r-1} C_l[i] \tag{2}$$

Every detector covers the same amount of space so without loss of generality for the rest of the derivation let $w = 0^l$. Using equations (1) and (2) we can now express $T(l)$ in terms of the number of unmatched strings of length $l-1$ as:

$$T(l) = C_{l-1}[d] + 2\sum_{i=1}^{d-1}(C_{l-1}[i] + C_{l-1}[i+d]) = 2T(l-1) - C_{l-1}[d] \tag{3}$$

where $d = \frac{2^r}{2} = 2^{r-1}$

In order to find an appropriate form for this recurrence we note the following properties:

1. $T(r) = 2^r - 1$
   Proof:
   It is easy to see that if $r = l$ then there is only 1 string out if the 2^r strings of length r that matches w, the remaining $2^r - 1$ strings are unmatched.
2. $T(l) = C_{l+r}[d]$ for $l \geqslant r$
   Proof: We can rewrite $T(l)$ in terms of bins for strings of length $l+1$ in the following way:

   $$T(l) = \sum_{i=0}^{2^r-1} C_l[i] = \sum_{i=0}^{2^{r-1}-1} C_{l+1}[2i+1]$$

   Note from (1) that $C_l[2i] = C_l[2i+1] = C_{l-1}[i] + C_{l-1}[i+d] \quad \forall_i : \quad 0 < i < d$.
   In general, expressing $T(l)$ in terms of bins for strings of length $l+v$ where $0 < v \leqslant r$

   $$T(l) = \sum_{i=0}^{2^{r-v}-1} C_{l+v}[2^v i + 2^{v-1}] \tag{4}$$

   Rewriting the total for strings of length l in terms of bins for strings of length $l+r$ we have:

   $$T(l) = \sum_{i=0}^{2^{r-r}-1} C_{l+r}[2^r i + 2^{r-1}] = C_{l+r}[2^{r-1}] = C_{l+r}[d]$$

3. $C_l[d] = 2^{l-r}$, for $r \leq l < 2r$
   Proof: Using property 2 and equation (4) we write $C_l[d]$ in terms of the unmatched strings of length r:

$$C_l[d] = T(l-r) = \sum_{i=0}^{2^{l-r}-1} C_r[2^{2r-l}i + 2^{2r-l-1}]$$

   note from (1) that $C_r[i] = C_r[j] = 1 \quad \forall_{i,j} : \quad 0 < i, j < 2^r$, hence $C_l[d] = 2^{l-r}$.

The first and third properties, together with equation (3) define a recurrence for $l \leqslant 2r$:

$$T(l) = \begin{cases} 2^r - 1 & if \quad r = l \\ 2T(l-1) - 2^{l-r-1} & if \quad r < l \leqslant 2r \end{cases}$$

Solving this equation yields:

$$T(l) = 2^l - 2^{l-r} - (l-r)2^{l-r-1} \text{ for } r \leqslant l \leqslant 2r \tag{5}$$

Combining this result with the second property and equation (3) we get the final recurrence:

$$T(l) = \begin{cases} 2^l - 2^{l-r} - (l-r)2^{l-r-1} & if \quad r \leqslant l \leqslant 2r \\ 2T(l-1) - T(l-r-1) & if \quad l > 2r \end{cases} \tag{6}$$

The total coverage provided by any one detector of length l for the r-contiguous bits matching rule is $D(l) = 2^l - T(l)$. Solving the recurrence for a given r implies solving $s^{r+1} - 2s^r + 1 = 0$ and substituting the initial conditions. As can be readily seen from this equation the amount of coverage provided by a single detector diminishes exponentially as r increases.

3 Conclusions

Some anomaly detection systems that rely on string matching have considered the r-contiguous bits matching rule for their purposes. In this paper we presented a way to determine, for the set of all binary strings of length l, how many are matched by a single detector under this rule. We have successfully derived a recurrence relation describing this quantity as a function of string length in the appropriate form as to derive its closed form solution given the *binding strength* parameter r.

4 Acknowledgments

The authors gratefully acknowledge the support of the National Science Foundation (grants CDA-9503064, and ANIR-9986555), the Office of Naval Research (grant N00014-99-1-0417), Defense Advanced Projects Agency (grant AGR F30602-00-2-0584), the Intel Corporation, and the Santa Fe Institute. F.E. thanks Consejo Nacional de Ciencia y Tecnología (México) grant No. 116691/131686 for its financial support. We would also like to thank the adaptive systems group and in particular Paul Helman, Justin Balthrop and Matthew Glickman.

References

[1] P. Helman and S. Forrest. An efficient algorithm for generating random antibody strings. Technical Report CS-94-07, University of New Mexico, Albuquerque, NM, 1994.

[2] O. E. Percus J. K. Percus and A. S. Perelson. Predicting the size of the antibodycombining region from consideration of efficient self/nonself discrimination. In *In Procedings of the National Academy of Science 90*, pages 1691–1695.