ZESZYTY NAUKOWE POLITECHNIKI BIAŁOSTOCKIEJ 2007
Informatyka – Zeszyt 2

Andrzej Chmielewski[1], Sławomir T. Wierzchoń[2]

ON THE DISTANCE NORMS FOR DETECTING ANOMALIES IN MULTIDIMENSIONAL DATASETS

Abstract: One of the key parameters of algorithms for anomaly detection is the metric (norm) applied to calculate the distance between every two samples which reflect its proximity. It is especially important when we operate on real-valued high dimensional datasets, i.e. when we deal with the problem of intruders detection in computer networks. As observed, the most popular Euclidean norm becomes meaningless in higher than 15-dimensional space. This means that other norms should be investigated to improve the effectiveness of real-valued negative selection algorithms. In this paper we present results for the following norms: Minkowski, fractional distance and cosine.

Keywords: negative selection, anomaly detection, Minkowski norm, fractional distance metric

1. Introduction

In recent years, the problem of calculating distance between two vectors (samples) in highly dimensional space was studied in great detail, for example in [14]. Nearest neighbor search, clustering and indexing are the examples of applications (issues) where behavior of some norms in high dimensions is the major obstacle to implement effective algorithms, and choice of the distance metric is not obvious. As it was shown in [1], some metrics, like i.e. Euclidean norm, loose its meaningfulness of proximity with increasing dimensionality. Thus, they can not be applied to highly dimensional datasets.

This problem was also observed, e.g. in [3] and [11], during the process of anomaly detection in datasets containing descriptions of network connections (e.g. the well-known KDD Cup 1999 dataset with 41 attributes [8]). Even after reduction of dimension for selected subsets to about 20, the average efficiency

[1] Faculty of Computer Science, Technical University of Bialystok, Poland
[2] Institute of Computer Science, Polish Academy of Science, Warsaw, Poland

of applied negative selection V-Detector algorithm never exceed 70%. We showed in [3]-[5] that after some modifications, this algorithm can produce quite good results, comparable with results generated by Support Vector Machine - a very strong classification tool. However, we are convincing that the choice of appropriate metric, especially for high dimensional datasets, is crucial. Therefore, in this paper, we present some experiments with selected norms which, should make possible to gain better results, in comparison to Euclidean distance.

2. Selected distance norms and its behavior in high dimensional spaces

In this section we present metrics (norms) used in the experiments described in Section 5. Below, there are good-known expressions to calculate the distance between two points: $\mathrm{x}=[x_1, x_2, \ldots, x_n]$ and $\mathrm{y}=[y_1, y_2, \ldots, y_n]$ in the space $\Re^n$. We also discuss their behavior in high dimensional space.

2.1. Minkowski norm and fractional distance metric

Minkowski norm of order m (L_m-norm distance) in space $\Re^n$ is defined as:

$$L_m(\mathrm{x},\mathrm{y})=\left(\sum_{i=1}^{n}|x_i-y_i|^m\right)^{\frac{1}{m}} \tag{1}$$

This is the generalized metric distance for $m \geq 1$. When $m=1$, it becomes Manhattan distance and when $m=2$, it becomes Euclidean distance.

Based on Minkowski norm, Aggarwal et al. [1] introduced *fractional distance metric* with $m<1$ which is more appropriate in high dimensional spaces. They proved that for the uniform distribution of k points

$$\left(\frac{C}{(m+1)^{1/m}}\right)\cdot\sqrt{\frac{1}{2\cdot m+1}} \leq \lim_{n\to\infty} E\left[\frac{Dmax_n^m - Dmim_n^m}{n^{1/m-1/2}}\right] \leq \left(\frac{C\cdot(k-1)}{(m+1)^{1/m}}\right)\cdot\sqrt{\frac{1}{2\cdot m+1}} \tag{2}$$

where $E[X]$ is expected value of the random variable X, $Dmax_n^m$ and $Dmin_n^m$ are farthest and nearest distance to the origin (measured by the distance metric L_m), and C is some constant. Eqn. (2) means that in a high-dimensional space

the absolute difference between $Dmax_n^m$ and $Dmin_n^m$ increases at the rate of $n^{1/m-1/2}$ when m decrease, independently of data distribution. Thus, *fractional distance metric* should provide better contrast between farthest and nearest neighbor than Manhattan and Euclidean norms. Moreover, the Euclidean norm should not be used for $n > 5$ and it completely loose its meaningfulness of proximity distance for $n > 15$ (distance is always very close to 0) – see e.g. [2] for a deeper discussion.

The similar proof, but only for Euclidean distance, was presented by Stibor in [10] to explain the very poor results obtained for the V-Detector algorithm (see Section 4) for the already mentioned KDD Cup 1999 dataset.

2.2. Cosine norm

Cosine norm for non-zero vectors x and y is defined as

$$D_{COS}(\mathrm{x},\mathrm{y}) = 1 - \frac{\langle \mathrm{x},\mathrm{y} \rangle}{\|\mathrm{x}\|\|\mathrm{y}\|} \tag{3}$$

where $\langle \mathrm{x},\mathrm{y} \rangle$ is the inner product of vectors x and y.

This norm is a popular distance measure for comparing (classifying) documents in the information retrieval applications. Thus, it seems to be good metric even for high dimensional spaces – see e.g. [2].

3. Negative selection

One of the major algorithms developed within emerging field of artificial immune systems (AIS) is Negative Selection Algorithm, proposed by Forrest et al. in [6]. It is based on the principles of self/nonself discrimination in the immune system. More formally, let U stands for the problem space, e.g. a set of all possible bit strings of fixed length, and S stands for the set of strings representing typical behavior. Then the set of strings characterizing anomalous behavior, N can be viewed as the set-theoretical complement of S:

$$N = U \setminus S \tag{4}$$

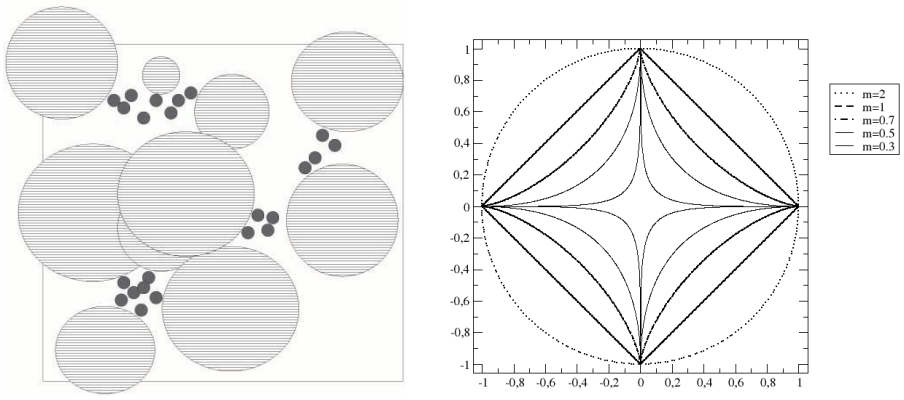
The elements of S are called self, and those of N are termed as non-self.

To apply the negative selection algorithm it is necessary to generate a set $D \subset N$ of detectors, such that each $d \in D$ recognizes at least one element $n \in N$,

and does not recognize any self element. Thus, we must designate a rule, $match(d,u)$, specifying when d recognizes an element u, consult [12] for details. This approach, although intuitive and simple, admits at least two serious drawbacks. First, it is hard to specify the full set S; typically we observe only a subset $S^{'} \subset S$. Second, majority of detection rules induce so-called holes, i.e. regions of N which are not covered by any detector.

Instead of the binary representation of the space U we can use real-valued representation, originally proposed in [7]. This paper is focused only on real-valued detectors, described in Section 4.

4. V-Detector algorithm



(a) (b)

Fig. 1. (a) Examples of performance V-Detector algorithm in 2D. Self samples and detectors are represented as circles. Grey circles denotes self samples, dashed circles denotes v-detectors. (b) Unit spheres for selected L_m norms in 2D.

The V-Detector algorithm was formally proposed by Ji and Dasgupta [9]. It operates on (normalized) vectors of real-values attributes being points in the *n*-dimensional unit hypercube, $U = [0,1]^n$. Each self sample, $s_i \in S$, is represented as a hypersphere with the center at $c_i \in U$ and constant radius r_s, i.e. $s_i = (c_i, r_s)$, $i = 1,\ldots,|S|$, where $|S|$ is the number of self samples. Every point $u \in U$ belonging to any hypersphere is considered as a self element. Also, detectors d_j are represented as hyperspheres: $d_j = (c_j, r_j)$, $i = 1,\ldots,|D|$ where $|D|$ is a number

of detectors. In contrast to self elements, the radius r_j is not fixed but is computed as the Euclidean distance from a randomly chosen center c_j to the nearest self element (this distance must be greater than r_s, otherwise detector is nor created). Formally, we define r_j as

$$r_j = \max\left\{0, \min_{1 \le i \le l} dist(c_j, c_i) - r_s\right\} \tag{5}$$

```
Input:    S = set of self samples
          r_s = self radius
          T_MAX = max. number of V-detectors
          co = estimated coverage
Output:   D = set of generated V-detectors

begin
   D ← ∅
   repeat
      find ← false
      t ← 0
          repeat
             x ← random point from [0,1]^n
             foreach d_i∈D do
                    //calculate the distance between c_di (center of detector d_i) and x

                    //if distance is lesser than r_di (radius of detector d_i)
                    if dist(c_di, x) ≤ r_di then
                       // point x is covered by detector
                       t ← t+1

                       //check, if the estimated coverage (co) was achieved
                       if t ≥ 1/(1-co) then
                              return D;
                    else
                       //point x is not covered by detector
                       find ← true;
                       break;
                    endif
             endfor
          until find = true

          //now, x is the candidate for detector

          //calculate the distance to the nearest self sample (r_d)
          r_d ← ∞
          foreach s_i∈S do
             l ← dist(c_si, x)
             if l - r_s < r_d then
                    r_d = l - r_s
             endif
          endfor

          //radius of detector (r_d) should be equal or grater then r_s
          if r_d ≥ r_s then
             //add new detector d to set D
             D ← D ∪ {d=(x,r_d)}
          endif

   until |D|=T_MAX
end
```

Fig.2. Pseudocode of V-Detector algorithm.

The algorithm terminates if predefined number T_{max} of detectors is generated or the space $U \setminus S$ is sufficiently well covered by these detectors (parameter co). The pseudocode of V-Detector algorithm is presented in Fig. 2.

As mentioned above, in original version, V-Detector utilizes Euclidean distance to calculate similarity between samples, thus both: v-detectors and self samples are represented as hyperspheres (circles in 2D, see Fig. 1(a)). However, this shape will change, when we choose another metric. Fig. 1 (b) presents unit spheres for selected values of m for L_m norm in 2D.

5. Experiments

To evaluate the performance of the original V-Detector algorithm two indices were used: *Detection Rate* (*DR*) and *False Alarm Rate* (*FAR*), computed as follow:

$$DR = \frac{TP}{TP + FN}$$
$$FAR = \frac{FP}{FP + TN} \tag{6}$$

where *TP* (*true positive*) is the number of correctly classified anomalous (nonself) samples, *FN* (*false negative*) is the number of self samples recognized as nonself, *FP* (*false positive*) is the number of nonself samples recognized as self and *TN* (*true negative*) is the number of correctly classified self samples.

It is worth to notice that for V-Detector algorithm *FAR* is always equal 0 when the same dataset is used at learning and classification process.

Experiments were performed for the following metrics: $L_{0.3}, L_{0.4}, L_{0.5}, L_{0.7}$, L_1 (Manhattan), L_2 (Euclidean) and cosine with following values of radius of self samples (r_s): 0.01, 0.001, 0001. As a test dataset we take some parts of KDD Cup 1999, namely, K_{ICMP} and K_{UDP} containing description of ICMP and UDP protocols, respectively. Moreover, these sets were divided into several subsets containing connections specific for particular services (consult [3] for details and reasons of such a decomposition):

$$K_{ICMP} = K_{ICMP\,eco\_i} \cup K_{ICMP\,ecr\_i} \cup K_{ICMP\,red\_i} \cup K_{ICMP\,tim\_i} \cup$$
$$\cup K_{ICMP\,urh\_i} \cup K_{ICMP\,urp\_i}$$
$$K_{UDP} = K_{UDP\,domain\_u} \cup K_{UDP\,ntp\_u} \cup K_{UDP\,other} \cup K_{UDP\,private} \cup K_{UDP\,tftp\_u}.$$

All experiments were repeated 20 times.

6. Results

Results presented in Fig. 3 and Fig. 4 shows that the highest values of *DR* were obtained for $L_{m=0.5}$ norm (with some exceptions). It is worth to notice than for $m < 0.5$, the V-Detetcor algorithm generated much worse results than for $m = 0.5$, in despite of tendered theoretical proofs mentioned in Section 2. And for $m \leq 0.2$ none of nonself samples was detected (*DR*=0) in despite of maximal number of detectors was generated ($T_{\max} = 100000$) – this suggest the existence of optimal value of *m* in the interval [0, 1].

One of the disadvantages of the *fractional distance metric* is that the duration of learning and classification stages rapidly increase with lower values of m (see Fig. 5). For example, duration of learning and classification for $m = 1$ is 2-3 times shorter than for $m = 0.5$. It is related with higher number of generated detectors (see Fig. 6) and complexity of calculation of the distances.

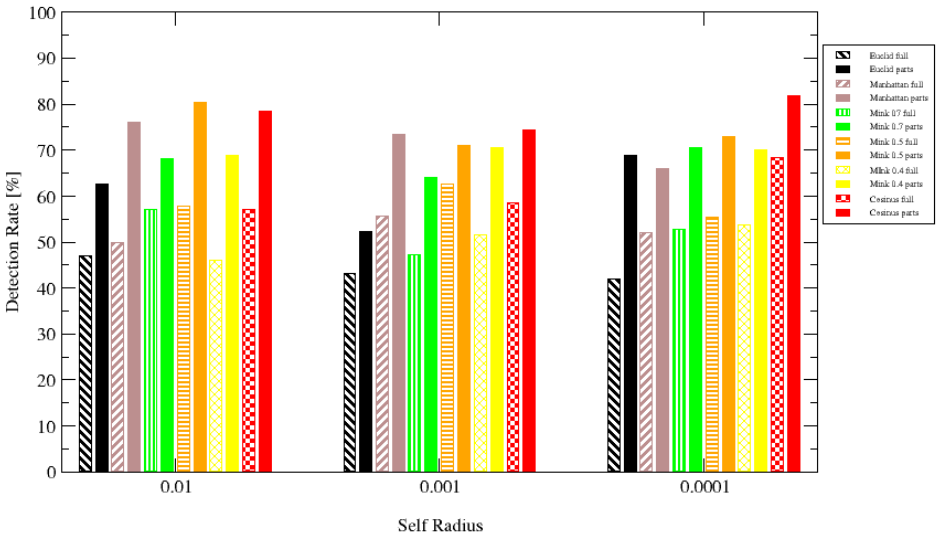


Fig. 3. *Detection Rate* values for ICMP protocol for different norms. "Parts" denotes that overall rate for protocol was calculated from rates obtained for its all subsets.

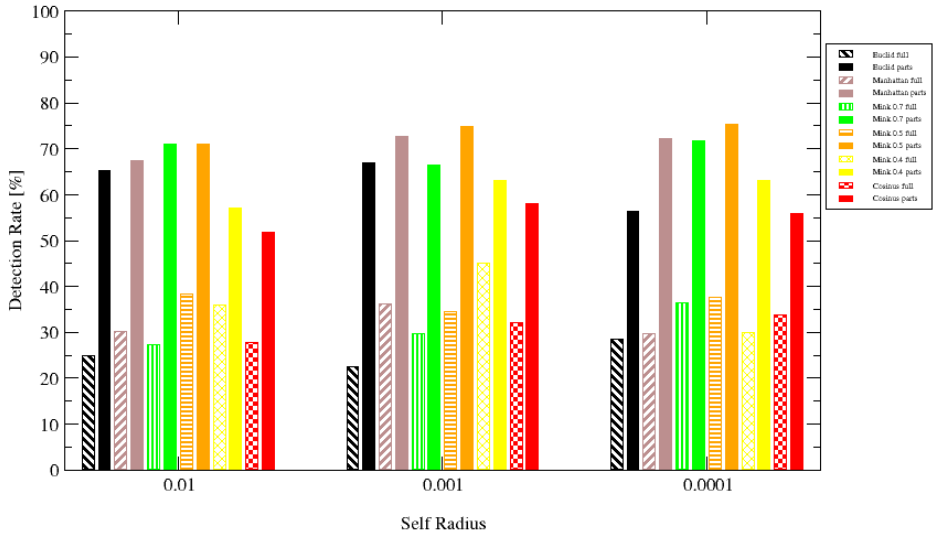


Fig. 4. *Detection Rate* values for UDP protocol for different norms. "Parts" denotes that overall rate for protocol was calculated from rates obtained for its all subsets.

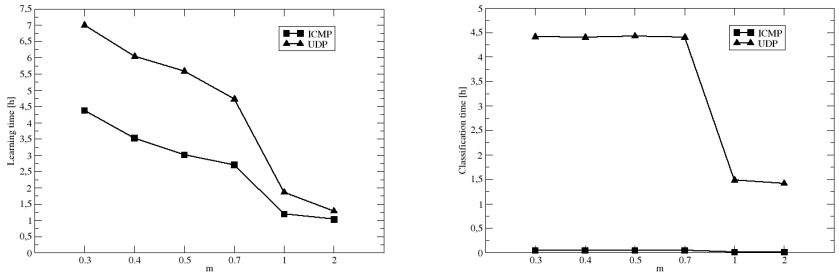


Fig. 5. Duration of learning and classification processes for different values of *m* for L_m-norm.

Cosine metric results in two extreme different values: good results for K_{ICMP} dataset (*DR* over 80%), and poor for K_{UDP}. It seems that this metric can be used only for special type of datasets, after experiment verifying its usability. As cosine metric is usually used to compare documents, it could be very effective in finding anomalies among their.

It worth to notice that, independently on metric, in all cases the *DR* was much grater when testing dataset was divided on subsets (about 2 times).

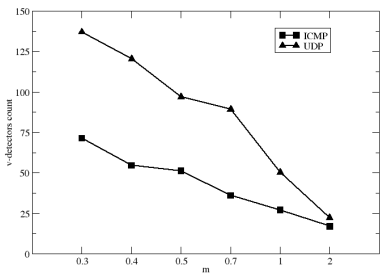


Fig. 6. Number of v-detectors generated for different values of m for of *m* for L_m -norm.

7. Conclusions

The results of experiments affirm that non-Euclidean metrics are more appropriate to calculate distance (proximity) between samples in highly dimensional spaces. Decreasing the value m for L_m-norms causes improvement of the *DR* ratio. However, for low values of m ($m < 0.5$) the efficiency of the algorithm decreases, what implies that the optimal value of m locates somewhere in the interval [0.5, 1.0]; hence, for all the datasets, this value should be properly tuned.

Performed experiments showed also a trade-off between efficiency and time complexity for $L_{m>0.5}$-norms. This is very important information in the case of intrusion detection in computer networks, where the efficiency understood as the time of reaction (to a potential intruder) is of primary interest.

Acknowledgement

Experiments were performed on the computer cluster at Faculty of Computer Science, Bialystok Technical University.

This work was partly supported by Bialystok Technical University grant S/WI/5/03.

References:

[1] Aggarwal C., Hinneburg A., Keim D.A., *On the surprising behavior of distance metrics in high dimensional space*, In Proc. of the 8th International Conference on Database Theory, ICDT 2001, London, pp. 420-434.

[2] Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U. *When is „nearest neighbor" meaningful*, In Proc. of the 7th ICDT, LNCS 1540, Springer 1999, pp. 217-235.

[3] Chmielewski A., Wierzchoń S. T., *Badanie przydatności algorytmu generującego V-detektory do klasyfikacji wybranych zbiorów*, Inżynieria wiedzy i systemy ekspertowe. T.1, Wrocław (Poland), 2006, pp. 13-22.

[4] Chmielewski A., Wierzchoń S. T., *Comparing real-valued negative selection algorithms for intrusion detection applications*, In Proc. of the 13th International Multi-Conference, Advanced Computer Systems (ACS'2006), Vol. 1, Międzyzdroje (Poland), 2006, pp. 387-395.

[5] Chmielewski A., Wierzchoń S.T., *V-Detector algorithm with tree-based structures*, In Proc. of the International Multiconference on Computer Science and Information Technology, Wisła (Poland), 2006, pp. 9-14.

[6] Forrest S., Perelson A., Allen L., Cherukuri R., *Self-nonself discrimination in a computer*. In Proceedings IEEE Symposium on Research in Security and Privacy, Los Alamitos, CA, 1994. IEEE Computer Soc. Press, pp. 202-212.

[7] Gonzalez F., Dasgupta D., Nino L.F., *A randomized real-valued negative selection algorithm*. In: J. Timmis, P.J. Bentley, E. Hart, eds., Proceedings of the 2nd International Conference on Artificial Immune Systems (ICARIS-2003), LNCS 2787, Springer-Verlag, 2003, pp. 261-272.

[8] Hettich S., Bay S. D., KDD Cup 1999 Data (1999), http://kdd.ics.uci.edu.

[9] Ji Z., Dasgupta D., *Real-valued negative selection algorithm with variable-sized detectors*, In Genetic and Evolutionary Computation GECCO-2004, Seattle (USA), Part I. LNCS 3102, Springer-Verlag, 2004, pp. 287-298.

[10] Stibor, T., J. Timmis und C. Eckert: *On the use of hyperspheres in artificial immune systems as antibody recognition regions*, In Proc. of the 5th International Conference on Artificial Immune Systems (ICARIS-2006), LNCS, Springer-Verlag, Oeiras (Portugal), 2006, pp. 215-228.

[11] Stibor, Thomas, Jonathan Timmis und Claudia Eckert: *A comparative study of real-valued negative selection to statistical anomaly detection techniques*, In Proc. of the 4th International Conference on Artificial Immune Systems(ICARIS-2005), Banff (Canada), LNCS, Springer, Berlin/Heidelberg, 2005, pp. 262-275.

[12] Wierzchoń S.T., *Deriving concise description of non-self patterns in an artificial immune system*, In L. C. Jain, J. Kacprzyk, eds., New Learning Paradigms in Soft Computing, Physica-Verlag 2001, pp. 438-458.
[13] Wierzchoń S.T. *Sztuczne systemy immunologiczne. Teoria i zastosowania.* Akademicka Oficyna Wydawnicza ELIT, Warszawa 2001.
[14] Weber R., Schek H.-J., Blott S,, *A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces*, In Proc. of VLDB Conference, 1998.

O METRYKACH ODLEGŁOŚCI DLA WIELOWYMIAROWYCH ZBIORÓW DANYCH WYKORZYSTYWANYCH W ALGORYTMIE SELEKCJI NEGATYWNEJ O WARTOŚCIACH RZECZYWISTYCH

Streszczenie: Jednym z kluczowych parametrów algorytmów wykrywania anomalii jest metryka (norma) służąca do obliczania odległości pomiędzy dwiema próbkami, która odzwierciedla ich podobieństwo. Jest ona szczególnie istotna w przypadkach operowania na zbiorach o wielu wymiarach takich, z jakimi mamy do czynienia w przypadku wykrywania intruzów w sieciach komputerowych. Zaobserwowano, że najczęściej stosowana norma euklidesowa staje się bezużyteczna w przestrzeniach o wymiarach większych niż 15. Oznacza to konieczność stosowania innych norm, które pozwoliłyby na zwiększenie skuteczności algorytmu selekcji negatywnej o wartościach rzeczywistych. W artykule prezentujemy wyniki uzyskane dla normy Minkowskiego, L_m, przy zmianach parametru m w zakresie (0, 2] oraz dla odległości kosinusowej.

Słowa kluczowe: selekcja negatywna, wykrywanie anomalii, norma Minkowskiego z ułamkowym wykładnikiem, odległość kosinusowa