TNC Transactions on Networks and Communications
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7420

Feature selection using closeness to centers for network intrusion detection

[1]S. Sethuramalingam,[2]Dr. E.R. Naganathan
[1] Department of Computer Science, Aditanar College, Tiruchendur , India
[2]Department of Computer Science, Hindustan University, Chennai , India
[1]seesay@rediffmail.com ; [2]ern_jo@yahoo.com

ABSTRACT

Classification in intrusion detection data set becomes complex due to its high dimensionality. To reduce the complexity, significant attributes for classification called as features in the data set needs to be identified. Numbers of methods are available in the literature for feature selection. In this paper, a new algorithm based on closeness of points to its center is proposed. It is tested with NSL-KDD data set. The algorithm shows better result.

1. INTRODUCTION

Internet provides a lot of services to human being at the same times unauthorized persons can try to access it. It questions the security of the internet. Although many static defense mechanisms are available such as firewalls even the better dynamic security system is needed[1]. The intrusion detection is played a major role. Intrusion detection is based on the principle that intruder features are different from the normal features [2]. It can be divided into two general types known as anomaly detection and misuse detection [3]. Anomaly detection detects threats by judging whether the activity deviate significantly from the known normal behavior. Misuse detection detects threats based on whether the signature of the behavior matching a known threat

pattern or not[4]. In the network based intrusion detection system, a huge amount of data is collected. The data are summarized as connection records. It has number of attributes to describe the record. Therefore the data set is high dimension and volumes of records also large.If the data set directly is used by any machine learning algorithm, it becomes difficult to get expected results since it contains many redundant records.

DOI: 10.14738/tnc.23.286
Publication Date: 12th June 2014
URL: http://dx.doi.org/10.14738/tnc. 23.286

2. RELATED WORK

In the network data set there are features can directly collected from the packets. Such features are referred as basic features example number of bytes, flag , service types etc., . some of the features are created by combing basic features. They are called derived features. The goal of using Derived Features is to find similarities that exist between different TCP connections in the

Network [5]. In order to compute those features, two types of sliding window intervals are used. Time Based Features: are all the derived features computed with respect to the past x seconds, where x is the size of the time window interval [6,7,8,9]. Connection Based Features: are all the derived features computed with respect to the past k TCP connections that were encountered in the network. In [10] authors proposed a feature selection algorithm based on gain ration and correlation. The C4.5 tree uses gain ratio to determine the break and to select the important features. Genetic algorithm is applied as search method with correlation as fitting function.

Li et al. [11] utilized Kmeans clustering to assign the data of each class to k clusters, and then used the new dataset consisting of only the centers of clusters to train SVM, in which k is the upper bound of the number of support vectors in each class. In [12] authors proposed a hybrid based algorithm for future selection which is based on information gain and genetic algorithm. In the proposed algorithm features are selected based on how many values or points closer to its centers. After computing this values, features are selected.

3. PROPOSED ALGORITHM

In this algorithm center of each feature for anomaly class and normal class are computed. Number of values of anomaly closer to anomaly center and similarly number values of normal closer to normal center are computed. Now for each feature number of points closer to its center is known. These values are used to form feature set. The following algorithm compute number of points closer to it center for each feature.

Algorithm closer_points(x,ano)

```
Let x be the given data set and ano is the number of anomaly records
Let s1 and s2 be the number of records and number of columns of x
fp←0;fn←0;tp←0tn←0;
for j=1:s2
   aavg(j)←mean(x(1:ano,j));
   navg(j)←mean(x(ano+1:s1,j));
end
for j=1:s2
   adavg←0;
   ndavg←0;
   for i=1:s1
       adavg←abs(aavg(j)-x(i,j));
       ndavg←abs(navg(j)-x(i,j));
    if i<=ano
      if (adavg<ndavg)
          tp←tp+1;
```

```
    else
       fn←fn+1;
    end
   elseif (i>ano)
    if(ndavg<adavg)
       tn←tn+1;
    else
       fp←fp+1;
    end
   end
  end
 end
```

4. EXPERIMENT RESULTS

In order to remove the influence of dimensions, the data set is standardized [13]. The data set is standardized by subtracting a measure of central location i.e. mean and divided by some measure of spread such as standard deviation. The algorithm closer point is executed with 1000 training data set from NSL-KDD[14] . The results are given below in the table 1.

Table 1. Number of features values closer to its center

Feature no.	Number of normal values closer to anomaly center (Fp)	Number of anomaly values closer to normal center (fn)	Number of anomaly values closer to anomaly center (tp)	Number of normal values closer to normal center (tn)
f3	126	131	219	524
f4	40	63	287	610
f5	11	344	6	639
f6	470	7	343	180
f23	63	116	234	587
f24	524	26	324	126
f25	19	140	210	631
f26	15	140	210	635
f27	28	270	80	622
f28	26	271	79	624
f29	38	83	267	612
f30	43	247	103	607
f31	491	23	327	159
f32	303	50	300	347
f33	161	22	328	489
f34	131	48	302	519
f35	57	299	51	593
f36	125	274	76	525
f37	132	316	34	518
f38	10	140	210	640
f39	4	141	209	646
f40	36	265	85	614
f41	34	271	79	616

From the table features f3,f4,f5,f6 and f38 i.e. service, flags, src_bytes, dst_bytes and dst_host_serror_rate are selected for classification.

The classification algorithm uses 8990 record as training data set and 999 records as testing data set. The testing data set has 470 anomaly records and 529 normal records. The classification algorithm described below is used for classification. The proposed feature selection algorithm is compared with [12] developed by the authors. The results of three different algorithms are tabulated in table 2.

Algorithm fuzzy_compos(trn_amean, trn_astd, trn_nmean, trn_nstd, tstdataset)

Tstdataset: testing data set has m records and n attributes

Trn_amean: mean of anomaly class records in the training data set

Trn_astd: standard deviation of anomaly class records for the training data set

Trn_nmean: mean of normal class records in the testing data set

Trn_nstd: standard deviation of normal class records in the testing data set

```
for each connection record in the testing data set
 py1←1;  py2←2
   for each attribute in the connection record
       y(i,j)←gausmf(x(i,j),[trn_amean,trn_astd])
       y1(i,j)←gausmf(x(i,j),[trn_nmean,trn_nstd])
       py1←py1*y(i,j)
       py2←py2*y1(I,j)
  end
       by1(i)←py1;
       by2(i)←py2;
 end
  for each connection record in the testing data set
  f1(i) ←(by1(i)*trn_astd)+trn_amean;
     f2(i) ←(by2(i)*trn_nstd)+trn_nmean;
    if  (f1(i)>f2(i))
        if   i <= anolimit
          tp←tp+1;
       else
         fn←fn+1;
     end
          if  i> anolimit
            tn←tn+1;
          else
           fp←fp+1;
          end
      end
   end
```

Table 2: comparison of different featuring algorithm

S.No.	Feature selection Algorithm	Features selected	False positive (FP)	False negative (FN)	True positive (TP)	True Negative (TN)
1	Information gain	f1,f2,f3,f4,f5,f6,f23,f24,f25,f26,f27, f28,f29,f30,f31,f32,f33,f34,f35,f36, f37,f38,f39,f40,f41	112	62	408	417
2	Information gain and genetic algorithm	f3,f5,f6,f23,f24,f27,f28,f29,f30,f32, f33,f34,f35,f38,f37,f40	106	49	421	423
3	Number of closer point algorithm	f3,f4,f5,f6,f38	60	85*	385	469

From the table value for false positive (normal as anomaly) and false negative (anomaly as normal) are decreasing. Detection of anomaly (true positive) and Detection of Normal (true negative) are increasing. Therefore third one is performs better than the other two.

5. CONCLUSION

In this paper, a new feature selection algorithm is proposed and testing with testing data set. The results are better than the two algorithms. In the first one the information gain there are 25 features out of 41 features are used to get the result. In the case of information gain and genetic algorithm are used to select features 16 features out of 41 features are selected. In the proposed algorithm uses only five features out of 41 features to get the result. In future fuzzy distances based on closeness of the center can be proposed to improve the results.

REFERENCES

[1]. Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets H. Güneş Kayacık, A. Nur Zincir-Heywood, Malcolm I. Heywood

[2]. Dalhousie University, Faculty of Computer Science, 6050 University Avenue, Halifax, Nova Scotia. B3H 1W5

[3]. Christine Dartigue, Hyun Ik Jang, and and Wenjun Zeng, "A New Data-Mining Based Approach for Network Intrusion Detection," Seventh Annual Conununication Networks and Services Research Conference, May 2009.

[4]. Wenke Lee, Salvatore J. Stolfo, and Kui W. Mok, "A Data Mining Framework for Adaptive Intrusion Detection," Proceedings of the EEE Symposium on Security and Privacy, pp.I20-132, 1999.

[5]. Feature selection and design olintrusion detection system based on k-means and triangle area support vector machine Pingj ie Tang ,Rang-an Jiang, Mingwei Zhao Dept. Computer Science and Engineering Dalian University of Technology Dalian City, China 2010 Second International Conference on Future Networks Iosif-Viorel Onut and Ali A. Ghorbani “A Feature Classification Scheme for

Network Intrusion Detection" Faculty of Computer Science, University of New Brunswickm540 Windsor Street, Fredericton, New Brunswick, PoBox 4400, Postal Code E3B 5A3, CanadaInternational Journal of Network Security, Vol.5, No.1, PP.1–15, July 2007

[6]. P. Dokas, L. Ertoz, V. Kumar, A. Lazarevic, J. Srivastava, and P. Tan, "Data mining for network intrusion detection", in Proceedings of NSF Workshop on Next Generation Data Mining (Baltimore, MD), pp. 21-30, Nov. 2002.

[7]. L. Ertoz, E. Eilertson, A. Lazarevic, P. N. Tan, P.Dokas, V. Kumar, and J. Srivastava, "Detection of novel network attacks using data mining", in ICDM Workshop on Data Mining for Computer Security (DMSEC) (Melbourne, FL), pp. 30-39, Nov. 2003.

[8]. KDD, Kdd-cup-99 task description,The Fifth International Conference on Knowledge Discovery and Data Mining, http://kdd.ics.uci.edu/databases/kddcup99/task.html, last access Oct, 2005.

[9]. S. J. Stolfo, W. Lee and K. W. Mok, "Mining in a data-flow environment: Experience in network intrusion detection", in Proceedings of the 5 International Conference on Knowledge Discovery and Data Mining, pp. 114-124, 1999.

[10]. Asha Gowda Karegowda, A. S. Manjunath & M.A.Jayaram **"C**omparative study of attribute selection using gain ratio and correlation based feature selection International Journal of Information Technology and Knowledge Management July-December 2010, Volume 2, No. 2, pp. 271-277

[11]. Der-Chiang Li, Yao-Hwei Fang, "An algorithm to cluster data for efficient classification of support vector machines," Expert Systems with Applications, Elsevier, vol. 34, issue 3, pp.2013-2018, Apr 200S.

[12]. S.Sethuramalingam and E.R. Naganathan, "Hybrid feature Selection for Network Intrusion", International Journal of Computer Science and Engineering, vol. 3 No. 5 May 2011 pp 1773-1780

[13]. Hai Jin Jianhua Sun, Han Chen, Zongfen Han Cluster and Grid Computing Lab. Huazhong University of Science and Technolory, Wuhan 430074 China. "A Fuzzy Data Mining Based Intrusion Detection Model. Proc. Of the 10th IEEE International Workshop on Feature Trends of Distributed Computing Systems" (FTDCS'04)@2004 IEEE

[14]. http://nsl.cs.unb.ca/NSL-KDD/

