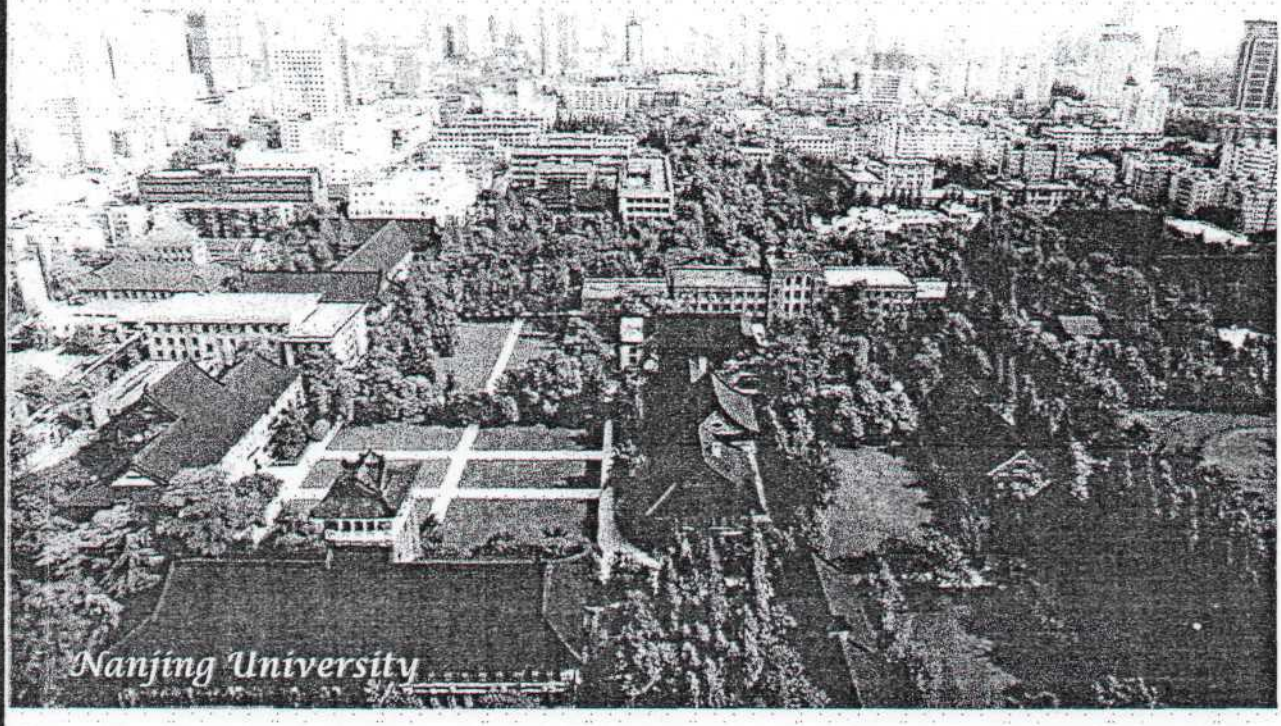
The 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'07)

May, 22-25, 2007
Nanjing, China



Working Notes of Workshop on

Data Mining for Business (DMBiz'07)

Carlos Soares Yonghong Peng Jun Meng (Eds.)

Program

The program includes contributions in exciting application areas, such as management, economics, banking, e-business, health and industry. The problems and techniques addressed include web and text mining, decision support, credit risk management, intrusion and fault detection, personalization and marketing.

Invited Talk

We are very honored to have on our workshop an invited talk by Weiliang Le from Business Objects and Xian Jiaotong University. The title of the talk is "Business Intelligence and Data Mining: Today and Future".

Review Process

A total of 52 papers were submitted to the workshop, including 7 in Chinese. Due to the large number of papers, we have carried out a first analysis, which excluded 10 papers that were out of the scope of the workshop. Each remaining paper was reviewed by at least two reviewers. Based on the reviews, 13 papers were selected for oral presentation and 13 for poster presentation.

Acknowledgements

We wish to thank the organizers of PAKDD for their support, in particular to the Workshops Chair, Takashi Washio, and the Program Committee Chair, Zhi-Hua Zhou.

We thank the members of the Program Committee for the timely and thorough reviews, despite receiving more papers than promised, and for their comments which we believe will be very useful to the authors.

The financial support from Fundação Oriente and project Triana (POCI/TRA /61001/2004) is gratefully acknowledged. We are also grateful to everybody who helped us to publicize the workshop, in particular to Gregory Piatetsky-Shapiro (www.kdnuggets.com), Guo-Zheng Li (Machine Learning Mailing List in China) and KMining (www.kmining.com).

Last, but not least, we would like to thank the valuable help of a group of students from Zhejiang University: Xiangyin Liu (preparation of the working notes), Zhiyong Li and Jinlong Wang (Chinese version of the webpages), and Huilan Luo and Zhiyong Li (support of the review process).

March 26, 2007

Carlos Soares[1]
Yonghong Peng[2]
Jun Meng[3]

[1] LIACC-NIAAD and Faculty of Economics, University of Porto, Portugal
[2] School of Informatics, University of Bradford, UK
[3] College of Electrical Engineering, Zhejiang University, China

Workshop

Carlos Soare
Yonghong Pe
Jun Meng

Program

Alípio Jorge
André Carva
Arno Knobb
Bhavani Thu
Can Yang

Carlos Soare
Carolina Mo
Chid Apte
Dave Watkin
Eric Auriol
Gerhard Paa
Gregory Piat
Jinlong Wan
Jinyan Li
João Mendes
Jörg-Uwe Ki
Jun Meng
Katharina P
Liu Zehua
Lou Huilan
Lubos Popeli
Mykola Pech
Paul Bradley
Peter van de

Petr Berka

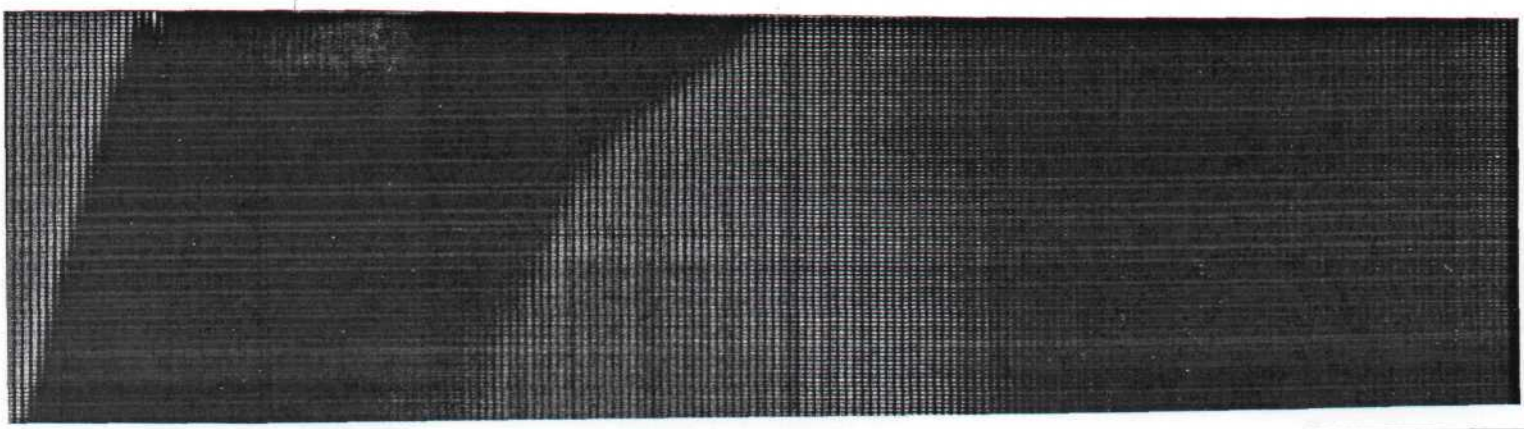
Ping Jiang
Raul Doming
Rayid Ghani
Reza Nakhae
Robert Engel
Rüdiger Wir
Ruy Ramos

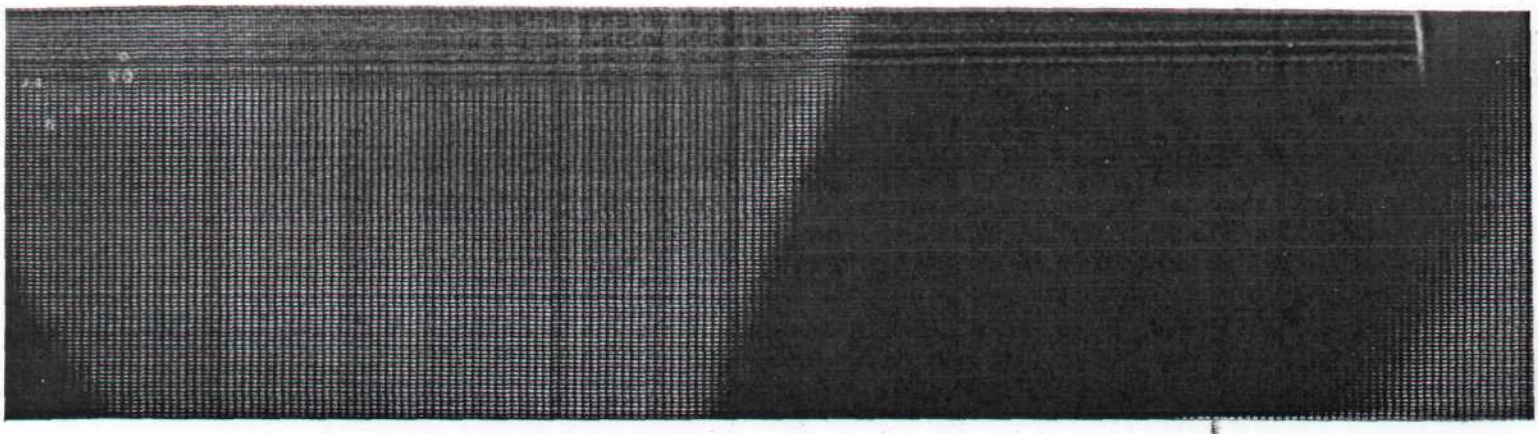
Workshop Organizers

Carlos Soares	University of Porto, Portugal
Yonghong Peng	University of Bradford, UK
Jun Meng	Zhejiang University, China

Program Committee

Alípio Jorge	University of Porto, Portugal
André Carvalho	University of São Paulo, Brazil
Arno Knobbe	Kiminkii/Utrecht University, The Netherlands
Bhavani Thuraisingham	Bhavani Consulting, USA
Can Yang	Hong Kong University of Science and Technology, China
Carlos Soares	University of Porto, Portugal
Carolina Monard	University of S. Paulo, Brazil
Chid Apte	IBM Research, USA
Dave Watkins	SPSS, USA
Eric Auriol	Kaidara, France
Gerhard Paa	Fraunhofer, Germany
Gregory Piatetsky-Shapiro	KDNuggets, USA
Jinlong Wang	Zhejiang University, China
Jinyan Li	Institute for Infocomm Research, Singapore
João Mendes Moreira	University of Porto, Portugal
Jörg-Uwe Kietz	Kdlabs AG, Switzerland
Jun Meng	Zhejiang University, China
Katharina Probst	Accenture Technology Labs, USA
Liu Zehua	Yokogawa Engineering, Singapore
Lou Huilan	Zhejiang University, China
Lubos Popelínský	Masaryk University, Czech Republic
Mykola Pechenizkiy	University of Eindhoven, The Netherlands
Paul Bradley	Apollo Data Technologies, USA
Peter van der Putten	Chordiant Software/Leiden University, The Netherlands
Petr Berka	University of Economics of Prague, Czech Republic
Ping Jiang	University of Bradford, UK
Raul Domingos	SPSS, Belgium
Rayid Ghani	Accenture Technology Labs, USA
Reza Nakhaeizadeh	University of Karlsruhe, Germany
Robert Engels	Cognit, Norway
Rüdiger Wirth	DaimlerChrysler, Germany
Ruy Ramos	University of Porto/Caixa Econômica do Brasil, Portugal





Preface

In the last few years, major Data Mining conferences, such as KDD and PKDD have included in their program workshops that are focused on applications. The success of these workshops has lead us to believe that it is time for another major conference, the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), to host a "Data Mining for Business" Workshop. The events will take place in beautiful and historical Nanjing (China) and the workshop will be held on May 22.

The workshop is organized by Carlos Soares (University of Porto), Yonghong Peng (University of Bradford) and Jun Meng (Zhejiang University). The goal is to gather researchers and practitioners to discuss relevant issues in the application of data mining technology in practice and to identify important challenges to be addressed in future research.

We are sure that this workshop will provide a forum for the fruitful interaction between participants from universities and companies, but we aim to go beyond that! We hope that this workshop will become the starting point for practical projects that involve people from the two communities. The future will tell if we succeeded.

Motivation

Business, government and science organizations are increasingly moving towards decision-making processes that are based on information. In parallel, the amount of data representing the activities of organizations that is stored in databases is also growing. Therefore, the pressure to extract as much useful information as possible from this data is very strong.

Many tools for Data Mining (DM) and Business Intelligence have been developed for that purpose. Additionally, DM methods are increasingly being integrated into other information systems and tools (e.g., customer relationship management, database management systems, network security tools).

Despite the maturity of the field, new problems and applications are continuously challenging both researchers and practitioners. The successful development of solutions for those problems requires that companies and universities work in close contact. Feedback from people with a business-oriented perspective is useful to assess current research results and to provide researchers with new challenges to work on. On the other hand, practitioners as well as decision makers in general need to be in touch with state-of-the-art research. Otherwise, they will not be able to provide the best solutions to their problems or to the problems of their clients. However, contact between these two communities is not as frequent as would be desirable. Although data mining, knowledge discovery and machine learning conferences provide an important contribution, they mostly attract an audience with a more technical and research background.

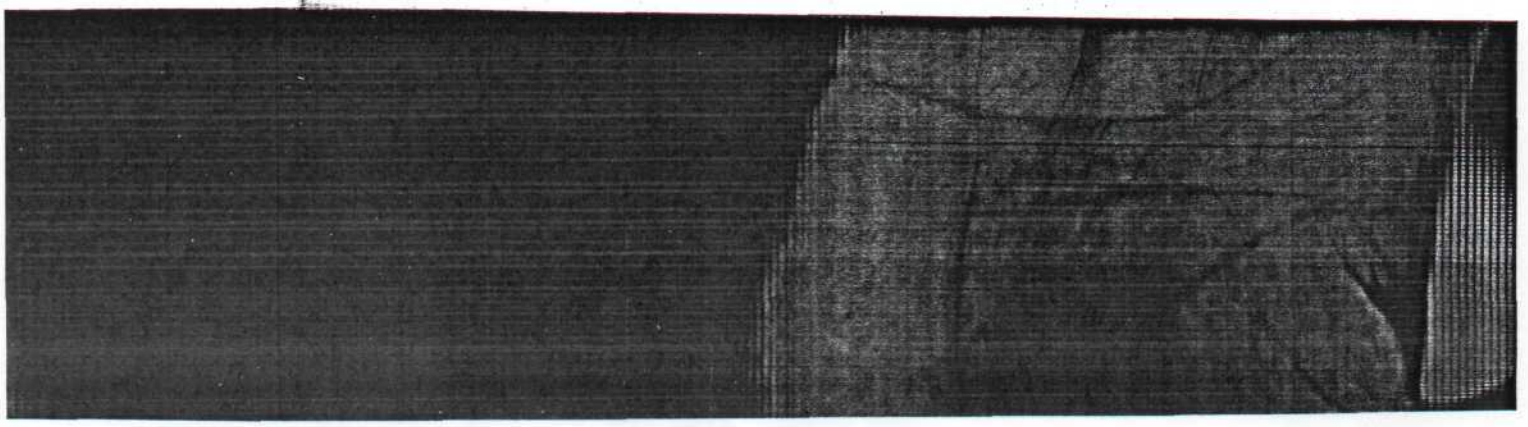


Table of Contents

A KNN-based Evolutionary Algorithm for Intrusion Detection in Networks

Francisco de Toro-Negro, Pedro García-Teodoro, Jesús E. Díaz-Verdejo, and Gabriel Maciá-Fernández

Signal Theory, Telematics and Communications Department,
University of Granada, Spain
{ftoro,pgteodor,jedv,gmacia}@ugr.es

Abstract. This paper addresses the use of an evolutionary algorithm for the optimization of a K-nearest neighbor classifier to be used in the implementation of an intrusion detection system. The inclusion of a diversity maintaining technique embodied in the design of the evolutionary algorithm enables to obtain different subsets of features extracted from network traffic data that lead to high classification accuracies. The methodology has been preliminary applied for Denial of Service attack detection.

1 Introduction

With the increased complexity of security threats, such as malicious Internet worms, denial of service (DoS) attacks, and e-business application attacks, achieving efficient network intrusion security is critical to maintaining a high level of protection. The efficient design of intrusion detection systems (IDS) is essential for safeguarding organizations from costly and debilitating network breaches and for helping to ensure business continuity. An IDS is a program that analyzes what happens or has happened in a computer network environment and try to find indications that the computer has been misused. An IDS will typically monitor network traffic data passing through the network in order to generate an alert when an attack event is taking place. Machine learning algorithms [1] such as binary classifiers can be optimized so they separate two different groups of observations: normal traffic and anomalous traffic (containing some kind of attack) with a certain classification performance.The optimization can involve the setting of certain parameters of the classifier or finding the network traffic features that lead to a good classification performance, the so-called feature selection problem [2].

Evolutionary algorithms (EAs) [3,4] have been showing a great success dealing with optimization problems with several solutions [5, 6] due to its special ability to explore large search spaces and capture multiple solutions in a single run. In this context, the present work tackles the use of an EAs based on deterministic crowding [9] for the optimization of a K-Nearest Neighbour (KNN) binary classifier and its evaluation in an intrusion detection domain.

This work is organized as follows: Section 2, the overall methodology presented in this paper is described. Thus, Section 3 shows some experimental work carried out by

using labeled network traffic data provided by DARPA [7]. Finally, Section 4 is devoted to discuss main results and conclusions.

2 A KNN-based evolutionary algorithm for intrusion detection

In the proposed methodology for intrusion detection, network data (IP packets) are organized into data flows. Then, n features – previously defined- characterizing the data flows are extracted to obtain an n-dimensional feature vector representing the network traffic in a given time window. A KNN binary classifier is optimized to classify each feature vector as belonging to normal traffic or malicious traffic. Feature vectors are normalized to have a value ranging between 0 and 1. A candidate solution to the problem is an n-dimensional weight vector. Each component of the weight vector is a real number ranging between 0 and 1 and will be representing the degree of importance of each feature in the classification process. For the optimization of the classifier, an evolutionary algorithm based on deterministic crowding has been used (see Fig. 1). Due to the fact that the an EA works with a population of k individuals candidate solutions, different choices of weight vectors can be explored in a single iteration of the algorithm. If the necessary diversity maintaining mechanism [8] is incorporated into the EAs, the found solutions will be geometrically different one from each other, providing flexibility to select the features to be considered in the intrusion detection system. This is of great importance due to the fact that the extraction of some features can be less time-consuming than others (i.e. they can be computed by using a smaller window time or they have less complex extraction).

Deterministic_Crowding Procedure
01 Create randomly Population P of Candidate Solutions (individuals) of Size *Popsize*
02 **While** (stop_condition) FALSE
03 $P^* = \emptyset$
04 **While** (Sizeof(P^*) $\neq$ *Popsize*)
05 Select two individuals p_1 and p_2 from P (without replacement)
06 Crossover p_1 and p_2 to obtain h_1 and h_2
07 Mutate h_1 and h_2 to obtain c_1 and c_2 (with mutation probability rate *pmut*)
08 **If**
$[\text{Distance}(p_1,c_1) + \text{Distance}(p_2,c_2)] \leq [\text{Distance}(p_1,c_2) + \text{Distance}(p_2,c_1)]$
09 **If** c_1 is better than p_1 **then** $P^* = P^* \cup \{c_1\}$ **else** $P^* = P^* \cup \{p_1\}$
10 **If** c_2 is better than p_2 **then** $P^* = P^* \cup \{c_2\}$ **else** $P^* = P^* \cup \{p_2\}$
Else
11 **If** c_1 is better than p_2 **then** $P^* = P^* \cup \{c_1\}$ **else** $P^* = P^* \cup \{p_2\}$
12 **If** c_2 is better than p_1 **then** $P^* = P^* \cup \{c_2\}$ **else** $P^* = P^* \cup \{p_1\}$
EndWhile
13 $P = P^*$
14 Evaluate the Performance of each candidate solution in P using the Diagnostic Scheme (K- nearest neighbourn classifier)
EndWhile

Fig. 1. Evolutionary Algorithm based on deterministic crowding used for training the classifier

3 Results

For the purpose of testing the aforementioned classifier methodology, a network database for training purpose provided by DARPA [7] has been used. This database is built with simulated network traffic data containing normal traffic data and 22 different kinds of computer attacks that fall in one of the following groups:

- DoS (Denial of Service): the attacker targets some computing or memory resource and makes it too busy or full to handle legitimate requests, or denies legitimate user access to that resource, for example SYN flood, ping of death, smurf, etc.
- R2U (Remote to User): the attacker exploits some vulnerability to gain unauthorized local access from a remote machine, for example guessing password.
- U2R (User to Root): the attacker has access to a normal user account (obtained legitimately or otherwise) and using this is able to gain root access by exploiting a vulnerability hole in the system, for example buffer overflow attacks.
- PROBE (Probing): attacker scans the network to gather information or find known vulnerabilities. An attacker with a map of machines and services that are available on a network can use this information to look for weak points, for example through port scan.

There are 41 features present in the data set. The first 9 of these are "intrinsic" features which describe the basic features of individual TCP data flows (TCP connections), and can be obtained from raw tcpdump output. The remainder features have been constructed as described in [12, 13]. Thus, features 10 to 22 are content-based features obtained by examining the data portion (payload) of a TCP data flow and suggested by domain knowledge. Features 23 to 41 are "traffic-based" features that are computed using a window. Features 23 to 31 use a two-second time window ("time-based"), and features 32 to 41 are constructed using a window of 100 TCP connections ("host-based"). The reasons for the different windows is that the DoS and PROBE attacks were shown to involve many TCP connections in a short time frame, whereas R2U and U2R attacks are embedded in the data portions of the TCP connections and normally involve a single TCP connection.

In this work, only the detection of attacks falling in the category of DoS attacks is addressed. On other hand, only numerical features have been considered from the original feature set. This way, only 6 out of the 9 "intrinsic" features have been used in this work. As part of the pre-processing of the database, duplicate data were withdrawn. Training and evaluation were performed by a random selection of 500 feature vectors labelled as normal traffic, and 500 feature vectors classified as DoS attacks.

- Classification Accuracy (C): represents the ratio between that correctly classified traffic and the overall traffic.
- Sensitivity (S): represents the ratio between that detected malicious traffic and the total malicious traffic.

The evaluation of the parameters C and S is calculated by applying the leaving one out method [14]. In each cycle, one vector is selected from the database as the test element. This vector is classified with the rest of the individuals in the population serving as classification references.

The crossover operator used in the deterministic crowding procedure is a single-point real-coded operator. Two different types of mutation have been considered: uniform mutation and Gaussian mutation.

The Evolutionary Algorithm ran during 400 iterations with 50 candidate solutions (individuals) with reached classifications accuracies ranging between 95% and 99%. Five of the found solutions (weigh vectors) are shown in Table 1. These results are slightly better than others machine learning algorithms used in IDS applications in the literature, such as k-means clustering [15,16] that has achieved 97.85% of performance accuracy or classifiers based on rules [17] or decision-trees [18] which has achieved 96.9% and 97.5% of performance accuracy respectively. Furthermore, our approach has the important advantage of its easier implementation. Solutions with higher sensitivity are better than others with lower sensitivity. The retrieval of several solutions during the optimization of the classifier enables to give a certain degree of flexibility of choosing the features to be used in the detection (for example those of easier extraction from the network traffic data).

4 Summary

This work addresses attack detection by using a new methodology consisting in a K-Nearest Neighbour binary classifier which produces a decision label (malicious traffic or normal traffic) by processing a feature vector representing the network traffic during a given window time. The features are automatically weighted by using an evolutionary algorithm in order to optimize the performance of the KNN Classifier. The retrieval of more than one solution during the optimization process gives a certain degree of flexibility to choose the features to be used in the detection process. The methodology has been preliminary applied to Denial of Service attack detection contained in DARPA database and has been validated by using leaving one out method. As future work, the authors are intended to explore the performance of this methodology on others network traffic database and set of features.

Acknowledgements. This work has been supported by the Spanish Government through MEC (Project TSI2005-08145-C02-02, FEDER funds 70%) and by the European Community –Research Infrastructure Action under the FP6 "Structuring the European Research Area" Programme– through HPC-EUROPA project (RII3-CT-2003-506079).

Table 1. Classification accuracy (C) and Sensibility (S) of five of the found solutions (weight vectors) to the problem of DARPA Denial of Service attack detection

#Feature	Solution #1	Solution #2	Solution #3	Solution #4	Solution #5
1	0.02	0.80	0.37	0.01	0.01
2	0.73	0.00	0.90	0.75	0.75
3	0.21	0.75	0.14	0.20	0.20
4	0.07	0.06	0.29	0.06	0.06
5	0.83	0.96	1.00	0.82	0.82
6	0.33	0.85	0.87	0.24	0.24
7	0.92	0.46	0.47	0.92	0.92
8	0.99	0.18	0.18	0.99	0.99
9	0.47	0.40	0.75	0.25	0.25
10	0.40	0.10	0.47	0.40	0.40
11	0.87	1.00	0.22	0.87	0.87
12	0.98	0.04	0.02	0.98	0.98
13	0.42	0.63	0.95	0.42	0.42
14	0.75	0.63	0.21	0.96	0.96
15	0.60	0.23	0.95	0.99	0.99
16	0.94	0.91	0.25	0.59	0.59
17	0.46	0.25	0.72	0.91	0.91
18	0.99	0.48	0.62	0.50	0.50
19	0.61	0.75	0.54	0.12	0.12
20	0.90	0.56	0.58	0.86	0.86
21	0.52	0.03	0.91	0.75	0.75
22	0.30	0.60	0.46	0.74	0.74
23	0.75	0.91	0.52	0.50	0.50
24	0.45	0.66	0.03	0.05	0.05
25	0.75	0.29	0.93	0.62	0.62
26	0.15	0.75	0.95	0.16	0.16
27	0.25	0.76	0.80	0.68	0.68
28	0.19	0.09	0.08	0.90	0.90
29	0.64	0.08	0.33	0.58	0.58
30	0.45	0.80	0.49	0.89	0.89
31	0.75	0.25	0.59	0.43	0.43
32	0.75	0.53	0.99	0.95	0.95
33	0.12	0.10	0.58	0.29	0.29
34	0.90	0.04	0.61	0.76	0.76
35	0.42	0.02	0.29	0.78	0.78
36	0.50	0.73	0.51	0.20	0.20
37	0.42	0.62	0.94	0.29	0.29
38	0.98	0.89	0.41	0.93	0.93
C (%)	97.2	97.3	97.2	94.4	94.5
S (%)	98.1	98.4	98.1	99.0	99.2

References

1. M. Sabhnani, G. Serpen: Application of machine learning algorithms to KDD intrusion detection dataset within misuse detection context. In Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications, Las Vegas, 2003, 209-215.
2. H. Liu, H. Motoda: Feature Selection for Knowledge Discovery and Data Mining. Kluwer, 1998.

3. A. E. Eiben, Series, Spring
4. D. E. Goldbe York: Addiso
5. J. P. Li, M. E for multimod 207-234.
6. F. de Toro, J algorithms fc 2004.
7. The UCI KI Irvine, http://kdd.ics
8. B. Sareni, L. on Evolution
9. S. W. Mahl Nature 2. R. 36, 1992.
10. S. W. Mahfc of the Sixth Mateo, CA,
11. A. Petrowsk 1996 IEEE I
12. S. Stolfo, W detection: I Conference
13. W. Lee, S. detection m
14. D. Hand: D
15. M. Sabhnai detection d Conference 2003), Las
16. R. O. Dud: 1973.
17. R. Agarwal mining. Te Minnesota,
18. I. Levin: I Exploratio

l solutions (weight

Solution #5
0.01
0.75
0.20
0.06
0.82
0.24
0.92
0.99
0.25
0.40
0.87
0.98
0.42
0.96
0.99
0.59
0.91
0.50
0.12
0.86
0.75
0.74
0.50
0.05
0.62
0.16
0.68
0.90
0.58
0.89
0.43
0.95
0.29
0.76
0.78
0.20
0.29
0.93
94.5
99.2

o KDD intrusion
the International
ions, Las Vegas,

. Mining. Kluwer,

3. A. E. Eiben, J. E. Smith: Introduction to Evolutionary Computing. Natural Computing Series, Springer, 2003.
4. D. E. Goldberg: Genetic Algorithms in Search, Optimization and Machine Learning. New York: Addison Wesley, 1989.
5. J. P. Li, M. E. Balazs, G. T. Parks, P. J. Clarkson: A species conserving genetic algorithm for multimodal function optimisation. Evolutionary Computation, Vol. 10, No. 3 (2002) 207-234.
6. F. de Toro, J. Ortega, E. Ros, S. Mota, B. Paechter: Parallel processing and evolutionary algorithms for multiobjective optimisation. Parallel Computing, Vol. 30, No.6, 721-739, 2004.
7. The UCI KDD Archive. Information and Computer Science, University of California, Irvine, "KDD cup 1999 data set", http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
8. B. Sareni, L. Krähenbühl: Fitness sharing and niching methods revisited. IEEE Transaction on Evolutionary Computation, Vol. 2, No. 3, 1998.
9. S. W. Mahfoud: Crowding and Preselection Revised, Parallel Problem Solving from Nature 2. R. Manner, B. Manderick (Eds.), Elsevier Science Publishers, Amsterdam, 27-36, 1992.
10. S. W. Mahfoud: A comparison of parallel and sequential niching methods. In Proceedings of the Sixth International Conference on Genetic Algorithms, Morgan Kauffman, San Mateo, CA, 1995.
11. A. Petrowski: A clearing procedure as a niching method for genetic algorithms. In Proc. 1996 IEEE Int. Conf. Evolutionary Computation, Nagoya, Japan, 798-803, 1996.
12. S. Stolfo, W. Lee, A. Prodromidis, P. Chan: Cost-based modeling for fraud and intrusion detection: Results from the JAM project. In Proc. DARPA Information Survavility Conference and Exposition, IEEE Computer Press, 2000, Vol. II, 1130-1144.
13. W. Lee, S. Stolfo and K. W. Mok: A data mining framework for building intrusion detection models. In IEEE Symposium on Security and Privacy, 1999, 120-132.
14. D. Hand: Discrimination and Classification. Wiley & Sons, New York, 1981.
15. M. Sabhnani, G. Serpen: Application of machine learning algorithms to KDD intrusion detection dataset within misuse detection context. In Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications (MLMTA 2003), Las Vegas, NV, June 2003, 209-215.
16. R. O. Duda, P. E. Hart: Pattern Classification and Scene Analysis. New York: Wiley, 1973.
17. R. Agarwal, M. V. Joshi: PNrule: A new framework for learning classifer models in data mining. Technical Report, TR 00-015, Department of Computer Science, University of Minnesota, 2000.
18. I. Levin: KDD-99 Classifier Learning Contest LLSoft's Results Overview. SIGKDD Explorations SIGKDD, January 2000, 1 (2), 67-75.