SVM Detailed Analysis

One of the objectives of this study is to perform a more detailed analysis of this corpus using the Support Vector Machines classifier. Therefore, classification using different kernels was performed. The kernels used included radial basis function (RBF), linear kernel, polynomial kernel, and the sigmoidal kernel.

Table 7 – Kernel Method Comparison

Kernel	F-measure Normal	F-measure Anomaly	F-Measure W. Avg.
Linear	0.786	0.756	0.769
Polynomial	N/A	N/A	N/A
RBF	0.777	0.764	0.77
Sigmoidal	0.707	0.685	0.694

Of all these kernels, RBF was the fastest with regards to processing time. The slowest kernel to be processed was the polynomial kernel which was stopped before completion. The results of the classification analysis using these different kernels on the Train+ and Test+ datasets from the NSL-KDD corpus as can be seen on Table 7. Finally, considering performance requirements, the analysis was performed using a subset of the 19 top features as ranked by information gain. This analysis can be seen in the next section.

Reduced Feature Set

A test was conducted with a reduced dataset (Train+ and Test+ datasets from the NSL-KDD corpus). Computational speed is essential in IDS systems that run on routers and network appliances with limited memory and processing power. A test was conducted using a reduced feature set of 19 features. The features were selected based on the information gain feature ranking. After conducting the analysis, the results of the classifier were only 2% lower than with the full set. This result is important because it shows which features are the most important and that not all are needed to maintain relatively good classification accuracies.

Conclusions

The results of the analysis show that Support Vector machines can obtain good classification results with the newly expanded NSL-KDD IDS corpus. Additionally, feature ranking was performed and the best features were identified. The results show that classification with the top half of the features obtained results which are almost as good as when using the full set of features. After conducting the analysis, the results of the classifier were only 2% lower than with the full set. Future work combining intrusion detection systems and machine learning will include the use of sequential methods for classification analysis such as with Hidden Markov Models (HMMs). HMMs can prove to be very useful for this type of analysis because they help to capture knowledge about prior states and how this information can help to predict future outcomes. Additionally, the study of new specific kernels which can be derived automatically will also be explored.

References

Chang, C.-C., Lin, C. 2001. LIBSVM: a library for support vector machines. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Cieslak, D.A.; Chawla, N.V.; Striegel, A. 2006. Combating imbalance in network intrusion datasets. IEEE International Conference on Granular Computing, 10-12, pp.732 - 737

Cortes, C., Vapnik, V. 1995. Support-Vector Networks. Machine Learning, vol. 20, pp. 273-297.

Kayacik, H. G., Zincir-Heywood, A. N., & Heywood, M. I. 2005. Selecting features for intrusion detection: A feature relevance analysis on kdd 99 intrusion detection datasets. In Proceedings of the Third Annual Conference on Privacy, Security and Trust (PST-2005).

Kayacık, G.; Zincir-Heywood, N. 2005. Analysis of three intrusion detection system benchmark datasets using machine learning algorithms. In Proceedings of the 2005 IEEE international conference on Intelligence and Security Informatics (ISI'05), Paul Kantor, Gheorghe Muresan, Fred Roberts, Daniel D. Zeng, and Fei-Yue Wang (Eds.). Springer-Verlag, Berlin, Heidelberg, 362-367. DOI=10.1007/11427995_29 http://dx.doi.org/10.1007/11427995_29

Kendall, K. 1999. A database of computer attacks for the evaluation of intrusion detection systems. Proceedings DARPA Information Survivability Conference and Exposition (DISCEX), MIT Press, pp: 12-26.

Khan, L,; Awad. M.; Thuraisingham, B. 2007. A new intrusion detection system using support vector machines and hierarchical clustering. The VLDB Journal. DOI 10.1007/s00778-006-0002-5

McHugh J. 2000. Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory. ACM Transactions on Information and System Security, Vol. 3 No.4

Perdisci, R. 2006. Statistical Pattern Recognition Techniques for Intrusion Detection in Computer Networks Challenges and Solutions. Ph.D. Dissertation. University of Cagliari, Italy, and Georgia Tech Information Security Center, College of Computing, Georgia Institute of Technology, Atlanta, GA, USA.

Roesch, M. 1999. Snort - Lightweight Intrusion Detection for Networks. In Proceedings of the 13th USENIX conference on System administration (LISA '99). USENIX Association, Berkeley, CA, USA, 229-238.

Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A. 2009. A Detailed Analysis of the KDD CUP 99 Data Set. In proceedings of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA 2009).

Yang, Y., Pederson, J. 1997. A Comparative Study on Feature Selection in Text Categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, pp. 412-420.