Performance Comparison For Multi Class Classification Intrusion Detection In SCADA Systems Using Apache Spark

Raogo KABORE [1], Yvon KERMARREC [1], Philippe LENCA [1]

1. IMT Atlantique – Lab-STICC UMR CNRS 6285 – UBL
Technopole Brest Iroise
F-29238 Brest cedex
{raogo.kabore, yvon.kermarrec, philippe.lenca}@imt-atlantique.fr

SCADA (Supervisory Control And Data Acquisition).are industrial control systems, that allow the monitoring and control of large indutrial systems [1]. Those systems are more and more subject to cyber attacks due to their interconnexion with corporate networks and the Internet [2], [3], [4], [5], [6], [7]. But the SCADA networks differ from the traditional IT networks because of characteristics like their high availability and real-time requirements [8], [9]. Common IT solutions like Anti viruses, firewalls, or software patches are not always effective for SCADA systems. We are comparing in this work the performances of a SCADA-specific Intrusion Detection system built with apache Spark, using Decision Tree [10] , Random Forest [11] , Naïve Bayes [10] and Multilayer Perceptron [12] approaches. Our comparison criteria are the recall, specificity, precision, training time and detection time. The dataset used is obtained from a testbed of the Mississippi State University SCADA Security Laboratory and Power and Energy Research laboratory which is representing a water storage tank system [13]. The records of the dataset were captured from the control system of the water storage [14]. This dataset contains 28 attacks grouped in 7 categories and attack free records. The general framework of our intrusion detection system (Figure 1) is using Apache Spark ML API with the PySpark (Python for Spark) language. Hadoop HDFS is used to store the raw dataset and Apache Hive is used to enable the conversion of raw dataset into dataframe within Spark. Accuracy, recall (or sensitivity), precision, specificity, training time and prediction time [15], [16] are the measures we have selected to evaluate the performances of the algorithms. The experimentation results (Table 1) show that the Decision Tree classifier has a very good detection rate (recall of 100 %) for all tuples categories except the Denial-of-Service (recall of 0). But the precision of 47.50% of class 5 gives us a hint that the class 6 tuples are rather misclassified. Decision Tree has also a fairly good training and detection time (7.84 s and 0.23 s respectively). The Random Forest also has a good detection rate for all classes apart DoS (from 95% up to 100%). But it has a 60% detection rate for the DoS class and longer training and detection time (24.73 s and 0.34 s respectively). Naïve Bayes and Multilayer Perceptron give overall poor classification results, but Naïve Bayes is very fast at training and detecting (2.96 s and 0.14 s respectively) . Multilayer Perceptron on the other hand, while taking time to train (155.51 s) is very fast in the prediction phase (0.16 s).

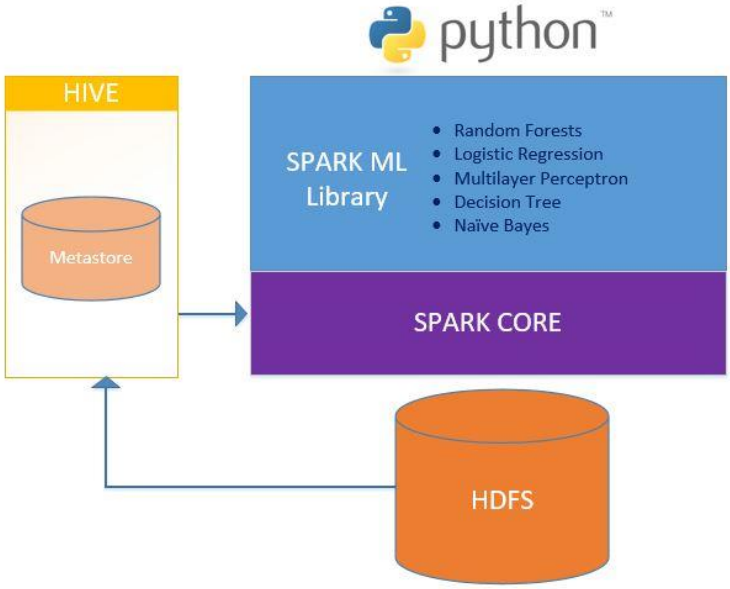

Figure 1 : SCADA IDS Framework

	Decision Tree			**Random Forest**			**Naïve Bayes**			**Multilayer Perceptron**		
Class	*Recall*	*Prec*	*Spec*	*Recall*	*Prec*	*Spec*	*Recall*	*Prec*	*Spec*	*Recall*	*Prec*	*Spec*
0	100	100	100	100	100	100	0	0	100	99.59	92.08	77.10
1	100	100	100	98.79	99.93	99.99	84.70	98.76	99.96	96.61	91.74	99.64
2	100	100	100	99.95	94.91	99.70	99.28	6.64	22.09	0	0	99.83
3	100	100	100	95.68	98.46	99.99	96.42	80.63	98.82	0	0	99.99
4	98.25	100.	100	98.77	99.12	99.99	78.53	98.90	99.99	99.04	67.97	99.23
5	100	47.50	99.45	100	100	100	100	100	100	0	0	100
6	0	0	100	60.05	100	100	0	0	99.72	0	0	100
7	100	100	100	100	100	100	100	100	100	100	99.94	99.99
Training time	7.84 s			24.73 s			2.96 s			155.51 s		
Prediction time	0.23 s			0.34 s			0.14 s			0.16 s		

*Prec : Precison *Spec : Specification

Table 1 : Decision Tree, Random Forest, Naïve Bayes and Multilayer Perceptron comparison.

[1] Keith Stouffer, Joe Falco, and Karen Scarfone, Guide to industrial control systems (ICS) security, in: NIST special publication 800.82 (2011), pp. 1616.

[2] Paul Oman, Edmund Schweitzer, and Deborah Frincke, Concerns about intrusions into remotely accessible substation controllers and SCADA systems, in: Proceedings of the Twenty-Seventh Annual Western Protective Relay Conference, vol. 160, 2000.

[3] Alfonso Valdes and Steven Cheung, Intrusion monitoring in process control systems, in: System Sciences, 2009. HICSS09. 42nd Hawaii International Conference on, IEEE, 2009, pp. 17.

[4] Bonnie Zhu, Anthony Joseph, and Shankar Sastry, A taxonomy of cyber attacks on SCADA systems, in: Internet of things (iThings/CPSCom), 2011 international conference on and 4th international conference on cyber, physical and social computing, IEEE, 2011, pp. 380388.

[5] Jill Slay and Michael Miller, Lessons learned from the maroochy water breach, in: International Conference on Critical Infrastructure Protection, Springer, 2007, pp. 7382.

[6] Bill Miller and Dale Rowe, A survey SCADA of and critical infrastructure incidents, in: Proceedings of the 1st Annual conference on Research in information technology, ACM, 2012, pp. 5156.

[7] Nicolas Falliere, Liam O Murchu, and Eric Chien, W32. stuxnet dossier, in: White paper, Symantec Corp., Security Response 5 (2011), p. 6.

[8] Alvaro A Cardenas et al., Attacks against process control systems: risk assessment, detection, and response, in: Proceedings of the 6th ACM symposium on information, computer and communications security, ACM, 2011, pp. 355366.

[9] Bonnie Zhu and Shankar Sastry, SCADA-specific intrusion detection/ prevention systems: a survey and taxonomy, in: Proceedings of the 1st Workshop on Secure Control Systems (SCS), vol. 11, 2010.

[10] Nahla Ben Amor, Salem Benferhat and Zied Elouedi, Naive bayes vs decision trees in intrusion detection systems, in: Proceedings of the 2004 ACM symposium on Applied computing, ACM, 2004, pp. 420-424.

[11] Leo Breiman, Random forests, in: Machine learning 45.1 (2001), pp.532.

[12] Miroslav Kubat," An Introduction to Machine Learning" in; Springer, 2015.

[13] Wei Gao et al., On SCADA control system command and response injection and intrusion detection, in: eCrime Researchers Summit (eCrime), 2010, IEEE, 2010, pp. 19.

[14] Thomas Morris and Wei Gao, Industrial control system traffic data sets for intrusion detection research, in: International Conference on Critical Infrastructure Protection, Springer, 2014, pp. 6578

[15] Jiawei Han, Micheline Kamber, and Jian Pei, "Data mining: Concepts and techniques" in: Elsevier, 2012.

[16] Chun-Wei Tsai et al., Big data analytics: a survey, in: Journal of Big Data 2.1 (2015), p. 21.