Data Mining Techniques for Threat Detection & Learning Classes

Venkateswara Rao Ch
Research Scholar, Dept. of CSE
V.B.S. Purvanchal University
Jaunpur, Uttar Pradesh, India.

Dr. Sushil Tripathi Ph.D, D.Litt.,
Professor& HoD, Dept. of CSE
V.B.S. Purvanchal University
Jaunpur, Uttar Pradesh, India.

Abstract: **It demonstrates that the ensemble- based approach is significantly more effective than traditional single-model methods; supervised learning outperforms unsupervised learning, and increasing the cost of false negatives correlates to higher accuracy. It shows effectiveness over non sequence data. For sequence data, this dissertation proposes and tests an unsupervised, ensemble based learning algorithm that maintains a compressed dictionary of repetitive sequences found. Throughout dynamic data streams of unbounded length to identify anomalies. In unsupervised learning, compression-based techniques are used to model common behavior sequences. This results in a classifier exhibiting a substantial increase in classification accuracy for data streams containing insider threat anomalies. This ensemble of classifiers allows the unsupervised approach to outperform traditional static learning approaches and boosts the effectiveness over supervised learning approaches. One of the bottlenecks to construct compress dictionary is scalability. For this, an efficient solution is proposed and implemented using Hadoop and MapReduce framework. We could extend the work in the following directions. First, we will build a full fledge system to capture user input as stream using apache flume and store it on the Hadoop distributed file system (HDFS) and then apply our approaches. Next, we will apply MapReduce to calculate edit distance between patterns for a particular user's command sequence data.**

Keywords: **classes, learning, models, GBAD, Threat, supervised, LIBSVM**

1. INTRODUCTION

As new models area unit created and previous ones updated to be additional precise, the smallest amount correct models area unit discarded to invariably maintain associate ensemble of specifically K current models. An alternative approach to supervised learning is unsupervised learning, which might be electively applied to strictly untagged data—i.e., information during which no points area unit expressly identified as abnormal or non-anomalous. Graph-based anomaly detection (GBAD) is one important type of unsupervised learning (Cook and Holder, 2007; Eberle and Holder, 2007; Cook and Holder, 2000), however has historically been restricted to static, finite-length datasets. This limits its application to streams associated with business executive threats that tend to possess limitless length and threat patterns that evolve over time. Applying GBAD to the business executive threat problem thus needs that the models used be reconciling and efficient. Adding these qualities enable effective models to be engineered from huge amounts of evolving information. In this treatise we tend to solid business executive threat detection as a stream mining drawback and professional pose 2 strategies (supervised and unsupervised learning) for efficiently sleuthing anomalies in stream information (Parveen, McDaniel et al., 2013). To deal with concept-evolution, our supervised approach maintains associate evolving ensemble of multiple OCSVM models (Parveen, Wegeret al., 2011). Our unsupervised approach combines multiple GBAD models in associate ensemble of classifiers (Parveen, Evans et al., 2011). The ensemble change method is intended in both cases to stay the ensemble current because the stream evolves. This organic process capability improves the classifier's survival of concept-drift because the behavior of each legitimate and illegitimate agents varies over time. In experiments, we tend to use check information that records supervisor call instruction data for an oversized, Unix-based, multiuser system.

2. DETAILS OF LEARNING CLASSES

This chapter can describe the different categories of learning techniques for non sequence data (Parveen, Evans et al., 2011; Parveen, McDaniel et al., 2013; Parveen, Weger et al., 2011). It serves the aim of providing a lot of detail on specifically however every technique arrives at detection business executive threats and the way ensemble models area unit designed, modified and discarded. The first segment goes over supervised learning thoroughly and therefore the second segment goes over unsupervised learning. each contain the formulas necessary to know the inner workings of every category of learning.

supervised Learning

In a chunk, a model is made mistreatment one category support vector machine (OCSVM) (Manevitz and Yousef, 2002). The OCSVM approach first maps coaching information into a high dimensional feature area (via a kernel). Next, the algorithmic program iteratively finds the peak margin hyperplane that best separates the coaching information from the origin. The OCSVM could also be considered as an everyday two-class SVM. Here the first category entails all the coaching information, and the second category is that the origin. Thus, the hyperplane (or linear call boundary) corresponds to the

Classification rule: f(x) = hw,xi+ b ---------- (1)

Where w is that the traditional vector and b could be a bias term. The OCSVM solves AN improvement problem to find the rule with peak geometric margin. This classification rule are used to assign a label to a check example x. If f(x) < zero, we tend to label x as AN anomaly, otherwise it is labelled traditional. Actually there's a trade off between maximising the gap of the hyperplane from the origin and therefore the variety of coaching information points contained within the region separated from the origin by the hyperplane.

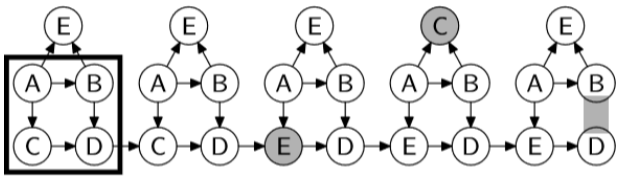

Figure.1. A graph with a normative substructure (boxed) and anomalies (shaded)

unsupervised Learning

Algorithm one uses 3 forms of graph primarily based anomaly detection(GBAD) (Cook and Holder, 2007; Eberle and Holder, 2007; Cook and Holder, 2000; Yan and Han dynasty, 2002) to infer potential anomalies mistreatment every model. GBAD could be a graph-based approach to finding anomalies in information by finding out 3 factors: modifications, insertions, and deletions of vertices and edges. Each distinctive issue runs its own algorithmic program that finds a normative substructure and makes an attempt to find the substructures that area unit similar however not fully a dead ringer for the discovered normative substructure. A normative substructure could be a revenant subgraph of vertices and edges that, once amalgamated into one vertex, most compresses the general graph. The rectangle in Figure 1 identifies AN example of normative substructure for the represented graph. Our implementation uses SUBDUE (Ketkar et al., 2005) to find normative substructures. The best normative substructure may be characterised because the one with

Borderline description length (MDL): L(S,G) = DL(G | S) + DL(S) (2)

wherever G is that the entire graph, S is that the substructure being analyzed, DL(G | S) is that the description length of G once being compressed by S, and DL(S) is that the description length of the substructure being analyzed. Description length DL(G) is that the minimum variety of bits necessary to explain graph G (Eberle et al., 2011). Insider threats seem as little proportion differences from the normative substructures. This is as a result of business executive threats decide to closely mimic legitimate system operations except for small variations embodied by illegitimate behavior. we tend to apply 3 different approaches for characteristic such anomalies, mentioned below.

GBAD-MDL

Upon finding the simplest press normative substructure, GBAD-MDL searches for deviations from that normative substructure in resultant substructures. By analyzing substructures of a similar size because the normative one, deference's within the edges and vertices' labels and in the direction or endpoints of edges area unit identified. The foremost abnormal of those area unit those substructures that the fewest modifications area unit needed to provide a substructure iso morphic to the normative one. In Figure four.1, the shaded vertex labelled E is AN anomaly discovered by GBAD-MDL.

GBAD-P

In distinction, GBAD-P searches for insertions that, if deleted, yield the normative substructure. Insertions created to a graph area unit viewed as extensions of the normative substructure. GBAD-P calculates the chance of every extension supported edge and vertex labels, and therefore exploits label data to get anomalies. The chance is given by

P(A=v) = P(A=v | A)P(A) (3)

Where A represents a foothold or vertex attribute and v represents its price. Chance P(A=v | A) may be generated by a Gaussian distribution:

GBAD-MPS

Finally, GBAD-MPS considers deletions that, if re-inserted, yield the normative substructure. To get these, GBAD-MPS examines the parent structure. Changes in size and orientation within the parent signify deletions amongst the subgraphs. The foremost abnormal substructures area unit those with the littlest transformation price needed to create the parent substructures identical. In Figure four.1, the last substructure of A-B-C-D vertices is identified as abnormal by GBAD-MPS as a result of the missing edge between B and D marked by the shaded parallelogram.

3. EXPERIMENTAL SETUP

Supervised Learning

We used LIBSVM (Chang and designer, 2011) to make our models and to get predictions for our check cases in our supervised approach. First, we'll provide an summary of our use of SVM software, that is standing operating procedure and is well documented in LIBSVMs facilitate files. We chose to use the RBF (radial-based function) kernel for the SVM. it absolutely was chosen as a result of it gives sensible results for our information set. Parameters for the kernel (in the case of two-class SVM, C and γ, and within the case of one-class SVM, ν and γ) were chosen so the F1 live was maximized. We tend to selected to use the F1 live during this case (over alternative measures of accuracy) because, for the classifier to try and do well in step with this metric, it should minimize false positives while conjointly minimizing false negatives. Before coaching a model with our feature set, we used LIBSVM to scale the input file to the vary [0,1]. This was

done to make sure that dimensions which takes on high values (like time) don't outweigh dimensions that strive against low values (such as dimensions that represent categorical variables). The parameters that were used to scale the coaching information for the model area unit a similar parameters that were wont to scale that model's check information. Therefore, the model's check information are within the neighborhood of the vary [0,1]. We conducted 2 experiments with the SVM. The first, as seen in Table 2, was designed to check one-class SVM with two-class SVM for the needs of business executive threat

Table 1. Dataset statistics after filtering and attribute extraction

. Dataset statistics after filtering and attribute e

Statistic	Value
No of vertices	500,000
No of tokens	62,000
No of normative substructures	5
No of users	all
Duration	9 weeks

Table 1. Dataset statistics once filtering and attribute extraction detection, and therefore the second, as seen in Table.3, was designed to check a stream classification approach with a a lot of ancient approach to classification. We'll begin by describing our comparison of one-class and two-class SVM. For this experiment, we tend to took the seven weeks of data, and every which way divided it into halves. We tend to deemed the first 0.5 coaching information and the other 0.5 testing information. We tend to made an easy one-class and two-class model from the training information and recorded the accuracy of the model in predicting the check information. For the business executive threat detection approach we tend to use AN ensemble-based approach that's scored in real time. The ensemble maintains K models that use one-class SVM, each con structed from one day and weighted in step with the accuracy of the models previous decisions. for every check token, the ensemble reports the bulk vote of its models. The stream approach printed on top of is a lot of sensible for detection business executive threats because business executive threats area unit stream in nature and occur in real time. A state of affairs like that within the first experiment on top of isn't one which will occur within the world. Within the world, insider threats should be detected as they occur, not once months of information has heaped-up in. Therefore, it s reasonable to check our change stream ensemble with an easy one-class SVM model constructed once and tested (but not updated) as a stream of recent information becomes offered, see Table.3.

Unsupervised Learning

For our unsupervised approach (based on graph primarily based anomaly detection), we wanted to accurately depict the effects of 2 variables. Those variables area unit K, the amount of ensembles

Table 2. Exp. A: One Class vs. Two Class SVM

	One Class SVM	Two Class SVM
False Positives	3706	0
True Negatives	25701	29407
False Negatives	1	5
True Positives	4	0
Accuracy	0.87	0.99
False Positive Rate	0.13	0.0
False Negative Rate	0.2	1.0

Table 3. Exp. B: Updating vs. Non Updating Stream Approach

	Updating Stream	Non Updating Stream
False Positives	13774	24426
True Negatives	44362	33710
False Negatives	1	1
True Positives	9	9
Accuracy	0.76	0.58
False Positive Rate	0.24	0.42
False Negative Rate	0.1	0.1

Table 4. Summary of data subset A (Selected/Partial)

Summary of data subset A (Selecte

Statistic	Dataset A
User	Donaldh
No of vertices	269
No of edges	556
Week	2–8
Weekday	Friday

maintained, and q, the amount of normative substructures maintained for every model within the ensemble. We tend to used a set of information throughout this wide range of experiments, as represented in Table 4, so as to finish them in a very manageable time. The choice to use the tiny subset of information was acquired because of the exponential growth in price for checking subgraph isomorphism. Each ensemble iteration was run with letter of the alphabet values between one and eight. Iterations were created with ensemble sizes of K values between 1 and 6.

4. RESULTS

Supervised Learning

Performance and accuracy was measured in terms of total false positives (FP) and false negatives (FN) throughout seven weeks of check information as mentioned in Table 4(week 2-week 8). The Lincoln Laboratory dataset was chosen for each its massive size and since its set of anomalies is acknowledge, facilitating AN correct performance assessment via misunderstanding counts. Table 2 shows the results for the first experiment mistreatment our supervised technique. One class SVM outperforms two-class SVM within the first experiment. Simply, two-class SVM is unable to observe any of the positive cases properly. Though the two-class SVM will achieve the next accuracy, it's at the value of getting a 100 percent false negative rate. By varying the parameters for the two-class SVM, we tend to found it potential to extend the

false positive rate (the SVM created a trial to discriminate between anomaly and traditional data), but in no case may the two-class SVM predict even one in every of the actually abnormal cases properly. One-class SVM, on the opposite hand, achieves a moderately low false negative rate (20%), while maintaining a high accuracy (87.40%). This demonstrates the prevalence of one-class SVM over two-class SVM for business executive threat detection. The superiority of one-class SVM over two-class SVM for business executive threat detection more justifies our call to use one-class SVM for our check of stream information. Table 3 gives a summary of our results for the second experiment mistreatment our supervised technique. The updating stream achieves a lot of higher accuracy than the non-updating stream, whereas maintaining an equivalent, and borderline, false negative rate (10%). The accuracy of the change stream is 76%, whereas the accuracy of the non-updating stream is fifty eight. The superiority of change stream over non change stream for business executive threat detection further justifies our call to use change stream for our check of stream information. By using labeled information, we tend to establish a ground truth for our supervised learning algorithmic program. This ground truth permits USA to position higher weights on false negatives or false positives. By advisement one more than the opposite, we tend to penalise a model a lot of for manufacturing that that we've exaggerated the weight for. Once detection business executive threats it's a lot of vital that we tend to don't miss a threat (false negative) than determine a false threat (false positive). Therefore, we tend to weigh false negative more heavily–i.e. we tend to add a FN price. Figures show the results of coefficient the false negatives a lot of heavily than false positives with this established ground truth. This is to say, that at a FN price of fifty, a false negative that's made can count against a model 50 times quite a false positive can. Increasing the FN price conjointly will increase the accuracy

of our OCSVM, change stream approach. we will see that that increasing the FN price up to thirty solely will increase the whole price while not affecting the accuracy, however once this, the accuracy climbs and therefore the total price comes down. Total cost, as calculated by equation, represents the total variety of false positives and false negative once they need been modified by the increase FN price. We tend to see this trend peak at a FN price of eighty wherever accuracy reaches nearly 56% and therefore the total price is at a coffee of 25229.

TotalCost = TotalFalsePositives + (TotalFalseNegatives ∗FNCost) (1)

The false negatives area unit weighted by price a lot of heavily than false positives as a result of it's more vital to catch all business executive threats. False positives area unit acceptable in some cases, but AN business executive threat detection system is useless if it doesn't catch all positive instances of business executive threat activity. this can be why models WHO fail to catch positive cases and manufacture these false negatives area unit corrected, in our greatest case result, eighty times a lot of heavily than those who manufacture false positives.

Table 5 reinforces our call to incorporate FN price throughout model elimination that heavily punishes models WHO manufacture false negatives over those who manufacture false positives. Including FN price will increase the accuracy of the ensemble and provides a far better F2 live.

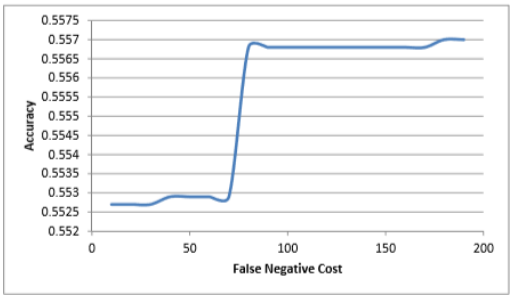

Figure 2. Accuracy by FN price

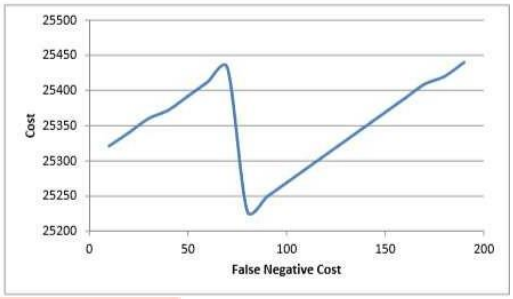

Figure 3. Total price by FN price

	Accuracy	F_2 Measure
w/ FN Cost	0.55682	0.00159
w/o FN Cost	0.45195	0.00141

Table 4. Impact of FN price

Unsupervised Learning

We next investigate the impact of parameters K (the ensemble size) and letter of the alphabet (the variety of normative substructures per model) on the classification accuracy and running times for our unsupervised approach. To a lot of simply perform the larger variety of experiments necessary to chart these relationships, we tend to use the smaller datasets summarized in Table 4 for these experiments. Dataset A consists of activity related to user donaldh throughout weeks 2–8. This user displays malicious business executive activity throughout the individual fundamental quantity. This dataset evince similar trends for all relationships mentioned henceforth; thus we tend to report solely the details for dataset A throughout the rest of the section. Figure 6 shows the link

between the cutoff letter of the alphabet for the amount of normative substructures and therefore the period in dataset A. Times increase or so linearly until letter of the alphabet = five as a result of there area unit solely four normative structures in dataset A. The rummage around for a 5th structure thus fails (but contributes running time), and better values of letter of the alphabet don't have any further effect.

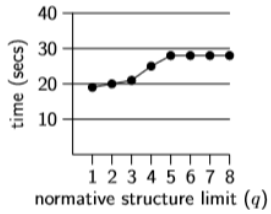

Figure 5. The effect of letter of the alphabet on runtimes for fixed K = half dozen on dataset A

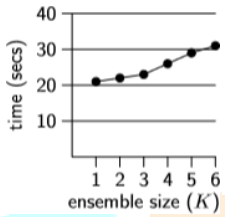

Figure 6. The effect of K on runtimes for fixed letter of the alphabet = four on dataset A

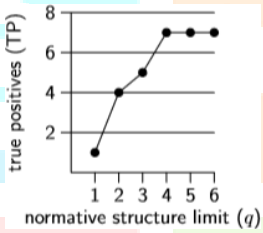

Figure 7. The effect of letter of the alphabet on TP rates for fixed K = half dozen on dataset A

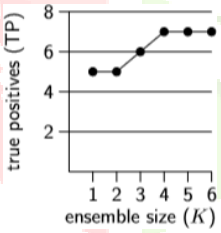

Figure 8. The effect of K on TP rates for fixed letter of the alphabet = four on dataset A

Figure 6 shows the impact of ensemble size K and runtimes for dataset A. of course, runtimes increase or so linearly with the amount of models (2 seconds per model on average during this dataset). Increasing letter of the alphabet and K conjointly tends to help within the discovery of true positives (TP). Figures illustrate by showing the positive relationships of letter of the alphabet and K, severally, to TP. Once q = four normative substructures area unit thought-about per model and K = four models area unit consulted per ensemble, the classifier dependably detects all seven true positives in dataset A. These values of letter of the alphabet and K thus strike the simplest balance between coverage of all business executive threats and therefore the efficient runtimes necessary for prime responsiveness. Increasing letter of the alphabet to four will return at the value of raising a lot of false alarms, however. Figure shows that the false positive rate will increase at the side of actuality positive rate till letter of the alphabet = four. Dataset A has solely four normative structures, thus increasing letter of the alphabet on the far side this time has no effect. This is supported with letter of the alphabet = 4, 5, 6 showing no increase in TP.

REFERENCES

[1]. Abouzeid, A., K. Bajda-Pawlikowski, D. J. Abadi, A. Rasin, and A. Silberschatz (2009). Hadoopdb: An architectural hybrid of mapreduce and dbms technologies for analytical workloads. PVLDB 2(1), 922–933.

[2]. Akiva, N. and M. Koppel (2012). Identifying distinct components of a multi-author document. In EISIC, pp. 205–209.

[3]. Al-Khateeb, T., M. M. Masud, L. Khan, C. C. Aggarwal, J. Han, and B. M. Thuraisingham (2012). Stream classification with recurring and novel class detection using class-based ensemble. In ICDM, pp. 31–40.

[4]. Al-Khateeb, T., M. M. Masud, L. Khan, and B. M. Thuraisingham (2012). Cloud guided stream classification using class-based ensemble. In IEEE CLOUD, pp. 694–701.

[5]. Alipanah, N., P. Parveen, L. Khan, and B. M. Thuraisingham (2011). Ontology-driven query expansion using map/reduce framework to facilitate federated queries. In ICWS, pp. 712–713.

[6]. Alipanah, N., P. Parveen, S. Menezes, L. Khan, S. Seida, and B. M. Thuraisingham (2010). Ontology-driven query expansion methods to facilitate federated queries. In SOCA, pp. 1–8.

[7]. Alipanah, N., P. Srivastava, P. Parveen, and B. M. Thuraisingham (2010). Ranking ontologies using verified entities to facilitate federated queries. In Web Intelligence, pp. 332–337.

[8]. Baron, M. and A. Tartakovsky (2006). Asymptotic optimality of change-point detection schemes in general continuous-time models. Sequential Analysis 25(3), 257–296.

[9]. Kumar Keshamoni and Hemanth. S, January 2017. Smart Gas Level Monitoring, Booking & Gas Leakage Detector over IoT. In Advance Computing Conference (IACC), 2017 IEEE 7th International (pp. 330-332). IEEE.

[10]. Borges, E. N., M. G. de Carvalho, R. Galante, M. A. Gon¸calves, and A. H. F. Laender (2011, sep). An unsupervised heuristic-based approach for bibliographic metadata deduplication. Inf. Process. Manage. 47(5), 706–718.

[11]. Brackney, R. C. and R. H. Anderson (Eds.) (2004, March). Understanding the Insider Threat. RAND Corporation.

[12]. Bu, Y., B. Howe, M. Balazinska, and M. Ernst (2010). Haloop: Efficient iterative data processing on large clusters. PVLDB 3(1), 285–296.