Cluster Detection in Text Pattern Over Emerging Social Networking Sites

[1]**Dr. A. Suresh,** [2]Reyana A, [3]Vaishnav Namboothiri T A,[4] Dheeraj R
[1]Professor & Head, [2]Assistant Professor, [3,4] UG Scholars
[1,2,3,4]Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, Coimbatore
[1]prisu6esh@yahoo.com, [2] reyareshmy@gmail.com, [3] vaishnavrema@gmail.com,[4]dheerajradhakrishnan@hotmail.com

Abstract: We propose an algorithm for anomalous detection of high dimensional discrete data usingan approach of clustering anomalies from the discrete data sets. Rather than the normal AD algorithm that detects the set of points which collectively exhibit abnormal patterns, providing a systematic way of detecting an anomaly with in an unanimity concern. The proposed algorithm emphasis efficient and powerful detection of anomalies. Unlike the existing techniques of finding each word separately the algorithm here uses a clustering method to detect each word that possess a maximum deviance from the normal pattern collectively in the batch of a given text document. Thereby resulting in more advantageous and effectual way of anomalous discovery over social networking sites preventing abnormal patterns.

Index Terms—Anomaly Detection, Pattern Detection, Clustering method, Discrete data, Statistical analysis, Unethical pattern detection

Introduction

The Clustering Anomalies (CA) is the procedure of detecting a word or a group of words that possess some deviations from the given standard or a pattern. There are many problems and interrupts in the seclusion. The clustering anomalies focus to solve these issues by using an approach of anomalous topic detection. An anomalous cluster is a set of data sample which apparent similar patterns of unconventional words. Each of the sample may not feel so unorthodox by itself but when it is altogether it possess a maximum deviance from the normal pattern or behaviour. The proposed method provides a substructure to detect such breed of anomalies and the pattern which they actually exhibit. While taking some cases no prior knowledge about the normal behaviour will be available and the intention is to find the incongruity in a single data set which consist of a congruous data and similarly with a possibility of atypical patterns without any gloss of which one possess an unethical pattern. The substructure has significant application in a variety of domains. For exemplification consider an important article or another some kind of document have been posted over some sites and some other company may try to post article in this repository to enhance their product and the business. Normally we are having some technique for detecting these kinds of unwanted post like advertisements. However, the one posting these types of articlewill act in such a way that the post will entirely like as a replica in the basis of their context. In this case this can be detected by using a clustering anomalies method. Normally a cluster means the collective of some words or some other which obeys some certain patterns or deviate from that patterns. The proposed method rules with this advantage called clustering. As a glance, we are maintaining a database which consist of a cluster of words possess maximum deviation from the ethical pattern.Clustering anomalies technique typically abnormal patterns exhibited by anomalous groups of clusters. Thereby the algorithm proposes a framework to detect such groups of anomalies and the atypical patterns they exhibit. Moreover, we consider the case where the anomalous pattern may manifest on only a small subset of the features, not on the entire feature space; i.e. samples in the anomalous cluster may be far apart from each other

measured on the full feature space, but on a subset of the feature space (the salient features), they exhibit a similar pattern of abnormality. Some other potentially important applications of our framework are: detecting similar patterns in malware and spyware (that were uploaded to a public software tool repository) to identify sources of attacks; studying patterns of anomalies in consumer behaviour to discover emerging consumer trends; finding shared patterns of tax avoidance to reveal loopholes in the law; and detecting organized malicious activities in social media.

Related Works

The proposed and test domain-independent methods that combine consensus clustering and anomaly detection techniques. Benchmark the efficacy of these methods on simulated insider threat data. Experimental results show that combining anomaly detection and consensus clustering produces more accurate results than sequentially performing the two tasks independently [1]. Due to a rapid advancement in the electronic commerce technology, the use of credit cards has dramatically increased. As Credit card becomes the most popular mode of payment for both online as well as regular purchase cases of fraud associated with it are also raising. This model the sequence of operations in credit card transaction processing using a Hidden Markov Model (HMM) and show how it can be used for the detection of frauds. If an Incoming credit card transaction is not accepted by the trained HMM with sufficiently high probability, it is considered to be fraudulent. At the same time, we try to ensure that genuine transactions are not rejected. This method improves detection accuracy by replacing binary feature thresholds with anomaly scores and by modelling the tail region of the distribution where anomalies occur [5]. Similarly, a parsimonious topic model for text corpora. In related models, such as Latent Dirichlet Allocation (LDA), all words are modeled topic-specifically, even though many words occur with similar frequencies across different topics. Our modeling determines salient words for each topic, which have topic-specific probabilities, with the rest explained by a universal shared model. Further, in LDA all topics are in principle present in every document. By contrast our model gives sparse topic representation, determining the (small) cluster of relevant topics for each document balancing model complexity and goodness of fit. This minimize the topic-specific words, document-specific topics, all model parameter values, and the total number of topics – in a wholly unsupervised fashion. Results on three text corpora and an image dataset show that our model achieves higher test set likelihood and better agreement with ground-truth class labels, compared to LDA and to a model designed to incorporate sparsity [6]. Explore factors that contribute to the success of the ensemble method, such as the number and variety of unsupervised detectors and the use of prior knowledge encoded in scenario-based detectors designed for known activity patterns. We report results over the entire period of the ensemble approach and of ablation experiments that remove the scenario-based detectors [10]. This survey tries to provide a structured and comprehensive overview of the research on anomaly detection. We have grouped existing techniques into different categories based on the underlying approach adopted by each technique. For each category, we have indented key assumptions, which are used by the techniques to differentiate between normal and anomalous behavior. Anomalous topic discovery for document databases represents a challenging domain due to the high feature dimensionality, with many candidate low-dimensional subspaces that may exhibit anomalous patterns. We develop our proposed framework focusing on topic models [11], [12]. Topic models are a class of

statistical models often used for discovering latent patterns (topics) in a collection oftext documents. Each topic specifies a pattern of words; i.e. words that appear more or less frequently than others under that topic [11], which posits document-specific mixing proportions over the topics, with each topic a multinomial distribution over the given vocabulary.

Implementation of Clustering Anomaly Detection

Anomalous Word Detection: Detection anomalous words is carried out in each test documents S (candidate anomalous cluster) that exhibits the pattern with maximum "deviance" from normal topics. Then, we conduct a statistical test to measure the significance of S and the topic exhibited by it, compared to the normal topics hypothesis. If the cluster candidate is determined to be significantly anomalous, we declare it as detected, we remove all documents in S from the test set, and then repeat this process until no statistically significant anomalous topic is found, thereby in the detection phase, it detects all patterns in the test documents that are anomalous (unusual) with respect to the normal topics. A document with anomalous contents, however, will have low likelihood under N_0. Thus,as a quantitative measureto characterize how well documents in S fit it computes: where w€s, and S contains

$$N_0(S) = \sum_{w \in S} \log(w|M_0) = \sum_{w \in S} N_0(w).$$

one new topic which is significantly different from the N normal topics in M_0. The specific structure for the alternative model consistent with the assumption that anomalous documents need not only contain anomalous contents – only a subset of an anomalous document may contain novel topics, with the remaining words well-generated from normal topics. S is unknown and has to be discovered by searching over the test documents. Since the size of S is not known, we begin constructing S by choosing a document in the test set which has the lowest likelihood to S if our test determines that d∗ significantly belongs to S. If the test reveals that contents of d∗ are not significantly related to the anomalous topic, we do not add d∗ to S and we stop adding further documents to the cluster. At each step, after adding a new document to S, we re-initialize all parameters of the alternative model on S. The anomaly score of the cluster score: where M_1, the maximum number of words in the document and N_0 the normal words that does not belong to anomalies set $N_0 \neq \{W\}$

$$(S) = \sum_{w \in S} (M_1(d) - N_0(d))$$

If S is found significantly anomalous, the cluster is reported as detected and we then remove all documents in S from the test set; the algorithm is then repeated on the new test set, until no significant cluster is found. Word Identification:There are different possible methods to determine if a document significantly belongs to S. One naive approach is to consider each document a random draw from a multinomial distribution over all words in the dictionary and then use Pearson's chi-squared test to determine significance of the difference between the observed counts (words in document d) and the expected counts (word probabilities under the null or alternative models). A major problem with this approach is that the length of each document L_d is typically much smaller than the vocabulary size N. While Pearson's chi-squared test relies on L_d! 1, in our problem generally L_d< N.Cluster Creation: Depending upon different discovered anomalous, they been clustered and given in to final output. After growing of a cluster has terminated, we need to determine whether the anomalous topic exhibited by the documents in that cluster is significant. Again, we note that due to small sample size, asymptotic distributions commonly known for the likelihood ratio test, do not hold. Instead, we perform bootstrap testing to compare

significance of a candidate cluster S compared to normal clusters.Thus, we test the null hypothesis that topic M +1 is insignificant in document d∗ versus the alternative hypothesis that it is significant. To test this hypothesis, we conduct a bootstrap algorithm to generate a set of normal documents from the null model, compute the topic contribution in those documents, and compare them with the topic contribution in the candidate document. Here we hold out a portion of the training set for the purpose of generating bootstrap samples – since the training set is used in learning the null model, also using it in our bootstrap algorithm may introduce bias in our significance test. To conduct a fair bootstrap test, we need to ensure that the bootstrap documents have similar topic proportions to those of the candidate document. Moreover, the bootstrap documents and the candidate document should have the same length. We generate |S| bootstrap documents based on the null distribution from a collection of validation documents and compare the likelihood ratio score of this bootstrap cluster.

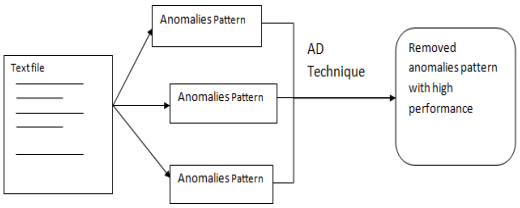

Figure 1. Clustering Anomaly Technique

Experimental Results

In each data set, we choose some classes as anomalous and take all documents from those classes out of the training and validation sets. We then randomly select some documents from normal classes and some documents from anomalous classes to create the test set. Our goal is to detect clusters of documents from the anomalous classes in the test set. The number of true detected anomalies from the majority anomalous class in that cluster divided by the total number of true anomalies and the number of true detected ones divided by the size of the cluster. Since the exact posterior is not available, we approximate this score using the variational distribution of θ_d. The score of a cluster is then the average of the scores of all documents in that cluster.

Conclusion and Future Work

The method detects clusters of anomalous documents which jointly manifest atypical topics on a small subset of (salient) features. The performance isachieved in a greater manner. The database which is maintained is flexible to add a new word or a group of word which possess anomalies. We are looking forward because the existing system provides the detection of anomalous patterns only by single words rather than a cluster pattern. Due to the inability of accessing the existing social networking sites now it cannot be implemented for the stuffs. The main intention of this project is to implement them on the existing and famous social networking sites. Future it will be implemented with effective changes over the social networking sites like Facebook, snapchat etc. Thus, the method accurately detects such anomalies by detecting clusters of anomalies.

References

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," *ACM Computing Surveys (CSUR)*, vol. 41, pp. 1–58, 2009.

[2] K. Wang and S. Stolfo, "Anomalous Payload-Based Network Intrusion Detection," in *Recent Advances in Intrusion Detection*, pp. 203– 222, 2004.

[3] F. Kocak, D. Miller, and G. Kesidis, "Detecting Anomalous Latent Classes in a Batch of Network Traffic flows," *in Information Sciences and Systems (CISS)*, 2014 48th Annual Conference on, pp. 1–6, 2014.

[4] R. Yu, X. He, and Y. Liu, "GLAD: Group Anomaly Detection in Social Media Analysis," *In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 372–381, 2014.

[5] A. Srivastava and A. Kundu, "Credit Card Fraud Detection Using Hidden Markov Model," *IEEE Transactions on Dependable and Secure Computing*, vol. 5, no. 1, pp. 37–48, 2008.

[6] H. Soleimani and D. J. Miller, "Parsimonious Topic Models with Salient Word Discovery," *Knowledge and Data Engineering, IEEE Transaction* on, vol. 27, pp. 824–837, 2015.

[7] X. Dai, Q. Chen, X. Wang, and J. Xu, "Online Topic Detection and Tracking of financial News Based on Hierarchical Clustering," *in Machine Learning and Cybernetics (ICMLC)*, 2010 International Conference on, pp. 3341–3346, 2010.

[8] Q. He, K. Chang, E.-P. Lim, and A. Banerjee, "Keep it Simple with Time: A Re-examination of Probabilistic Topic Detection Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1795–1808, 2010.

[9] M. Zhao and V. Saligrama, "Anomaly Detection with Score Functions Based on Nearest Neighbour Graphs*," in Advances in neural information processing systems*, pp. 2250–2258, 2009.

[10] J. Major and D. Riedinger, "EFD: A Hybrid Knowledge/StatisticalBased System for the Detection of Fraud," *Journal of Risk and Insurance*, vol. 69, no. 3, pp. 309–324, 2002.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] D. Blei, L. Carin, and D. Dunson, "Probabilistic Topic Models," *Communications of the ACM*, vol. 55, pp. 77–84, Nov. 2012.

[13] H. M. J. Almohri, D. D. Yao, and D. G. Kafura"Identifying Native Applications with High Assurance",i*n ACM Conference on Data and Application ecurity and Privacy (CODASPY),* pages 275–282. ACM, 2012.

[14] K. Borders and A. Prakash. Web Tap: Detecting Covert Web Traffic, *In Proceedings of the 11th ACM Conference on Computer and Communication Security*, pages 110–120, 2004.

[15] K. Xu, H. Xiong, C. Wu, D. Stefan, and D. Yao, "Data-Provenance Verification for Secure Hosts", *IEEE Transactions on Dependable and Secure Computing*, 9:173–183, 2012.

[16] K. Xu, D. Yao, Q. Ma, and A. Crowell. Detecting Infection Onset with Behavior-Based Policies. i*n Proceedings of the Fifth International Conference on Network and System Security (NSS)*, September 2011.

[17] H. Zhang, W. Banick, D. Yao, and N. Ramakrishnan. User Intentionbased Traffic Dependence Analysis for Anomaly Detection,*Technical Report, Department of Computer Science, Virginia Tech*, Blacksburg, Virginia, February 2012.

[18] J. Glasser and B. Lindauer. "Bridging the Gap: A Pragmatic Approach to Generating Insider Threat Data," *in Proceedings of the Workshop on Research for Insider Threat, IEEE CS Security and Privacy Workshops*, San Francisco, CA, 23-24 May 2013.

[19] W T. Young, "Use of Domain Knowledge to Detect Insider Threats in Computer Activities," *in Proceedings of the Workshop on Research for Insider Threat, IEEE CS Security and Privacy Workshops*, San Francisco, CA, 23-24 May 2013.

[20] A.Suresh (2013), "An Efficient Conversion of Epigraphical Textual Image to User Readable Text", International Journal of Engineering Research & Technology (IJERT), ISSN 2278-0181,*Vol. 2, No.9, September 2013* pp. 1301-1304.