Nov 7th, 9:00 AM - 5:00 PM

Randomizing Ensemble-based approaches for Outlier

Lediona Nishani
University of New York Tirana, ledionanishani@gmail.com

Marenglen Biba
University of New York Tirana, marenglenbiba@unyt.edu.al

Randomizing Ensemble-based approaches for Outlier

Lediona Nishani[1] , Marenglen Biba[2]

[1,2]Department of Computer Science University of New York Tirana, Albania
ledionanishani@gmail.com[1], marenglenbiba@unyt.edu.al[2]

Abstract. The data size is increasing dramatically every day, therefore, it has emerged the need of detecting abnormal behaviors, which can harm seriously our systems. Outlier detection refers to the process of identifying outlying activities, which diverge from the remaining group of data.
This process, an integral part of data mining field, has experienced recently a substantial interest from the data mining community. An outlying activity or an outlier refers to a data point, which significantly deviates and appears to be inconsistent compared to other data members. Ensemble-based outlier detection is a line of research employed in order to reduce the model dependence from datasets or data locality by raising the robustness of the data mining procedures. The key principle of an ensemble approach is using the combination of individual detection results, which do not contain the same list of outliers in order to come up with a consensus finding. In this paper, we propose a novel strategy of constructing randomized ensemble outlier detection. This approach is an extension of the heuristic greedy ensemble construction previously built by the research community. We will focus on the core components of constructing an ensemble –based algorithm for outlier detection. The randomization will be performed by intervening into the pseudo code of greedy ensemble and implementing randomization in the respective java code through the ELKI data-mining platform. The key purpose of our approach is to improve the greedy ensemble and to overcome its local maxima problem. In order to induce diversity, it is performed randomization by initializing the search with a random outlier detector from the pool of detectors. Finally, the paper provides strong insights regarding the ongoing work of our randomized ensemble-based approach for outlier detection. Empirical results indicate that due to inducing diversity by employing various outlier detection algorithms, the randomized ensemble approach performs better than using only one outlier detector.

Keywords: outlier detection, ensemble outlier detection, greedy ensemble, randomized ensemble, ELKI

1. Introduction

The exponential growth of large databases has led to the need for monitoring, examining and predicting economical, weather forecast or other various procedures in the whole world. In these processes not so often occur rare events, distinguished from the daily basis processes, which can harm or can deteriorate the respective process. These rare behaviors are called outlier or anomalies, which can happen very infrequently. The most popular definition of outlier is "an observation, which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" [1]. Outlier detection has found a substantial attention from many areas of research respectively in disclosing malicious transactions in the banking system, intrusions detection in the networking system and privacy data in the healthcare databases. Data mining and machine learning algorithms have emerged in order to solve the outlier detection tasks. Data mining communities have categorized outlier detection methods by four different groups [2]: statistical reasoning, distance-based, density-based and modeled- based approaches. In regards to statistical reasoning [3], the data points are represented as a stochastic distribution where outliers are identified based on the relationship they have with the stochastic distribution of the data points. Statistical reasoning approaches undergo through various limitations in higher dimensionality data because they can face significant difficulty in computing the multidimensional distribution of data in these conditions. On the other hand, distance-based approaches proposed in [4], [5], [6], [7], [8], [9], are intended to enhance and mitigate the

shortcomings that statistical approaches pose to outlier detection problems. Their key idea is based on estimating distances between data points and assigning score to data. The event, which has the larger score, is pronounced as outlier. Regarding density- based approaches, various pieces of work are presented, too: [10], [11], [12], [13]. They rely on computing the densities of local neighborhoods. Referring to model- based techniques, they determine the normal behavior by making use of predictive models just like neural networks or unsupervised support vector machines. Making use of just one algorithm such as a density-based or a distance-based algorithm does not summarize the entire kinds of outliers because there are methods, which are good at detecting some kind of outlier and others, which perform better in other domains. Therefore, in order to provide the whole truth, it is appropriate to integrate different outlier detection outcomes by means of an ensemble-based approach coming up with a consensus finding. Ensemble analysis is a kind of method, which aims to reduce the model dependence from the specific dataset or data locality. Usually ensemble approaches are referred as the combination of the data outcomes executed in independent ways. Ensemble-based approaches for outlier detection are more difficult to be employed compared to classification or clustering domain due to the combination of small sample space and the unsupervised nature. This is why the state of-the-art of ensemble analysis for outlier detection has been neglected from the research community and has not been profoundly analyzed in depth. Ensemble for outlier detection has inherited from the ensemble-based classification two major properties of constructing an ensemble named: accuracy (to be better than random) and diversity (to perform different errors in different algorithms) in order to enhance the detection process. Diversity is referred to overcome making the same errors due to the correlated output results. Accuracy consists of having the high performance detection rate and low false detection rates. In the remainder of the paper, related background will provide motivations of the ensemble-based analysis and the most popular and significant line of research from the first attempt to the latest works. By randomizing this ensemble algorithm we intend to enhance the detection rate and to mitigate the false positive rates. After disclosing the motivation that underlies this line of research, the next section deals with the proposed strategy that we will motivate our research by stating the reason why we have chosen to follow this idea. Finally, in concluding of the paper we reasonably indicate some future open issues that we intend to deal with in the future work and fill the gap of the previous researches.

2. Related Work

In this section, we are going to describe some works, which have been extensively cited and have opened new line of researches. Ensemble analysis has raised attention for the first time in the classification problems; in the supervised learning techniques. Extensive research was undertaken in building ensembles from single classifier algorithms in order to enhance effectiveness. This line of research has a sound theoretical background [15], [16], [17], [18], [19]. Besides classification, ensemble analysis has found a considerable attention in constructing ensemble-clustering methods [2]. Referring to the field of outlier detection, there are some strong attempts in the implementation of some outlier detection algorithms. Therefore, this domain has been called an “emerging area” [21]. Usually ensemble methods are considered as meta methods [22] because they process the output of their own methods. One procedure that can induce diversity is dubbed as bagging [14]. This is an ordinary process employed in the classification and clustering techniques. However, this kind of technique has not been so successful and has not found sound theoretical framework for implementation. Instead, various techniques have been presented that tackle the score and the ranking problem. In [14] was proposed a breadth-first traversal through the outlier rankings in order to combine algorithms. Then it is performed the comparability of the retrieved scores. Calibration approaches aimed to fit outlier scores that have been detectors outcome have been converted into probability estimates [23]. In [24] was proposed an algorithm, which has generated some scores centered on their mean and been scaled by their standard deviations.. On the other hand, statistical reasoning has been carried out to make sense of different outlier scores into converting through outlier probabilities [25]. In this paper has been disclosed the likelihood of improvement across combining various approaches, but it lacks on applying of measure of current diversity of correlation among algorithms members. Ensemble

analysis has been deployed successfully in outlier detection for high dimensional data where multiple subspaces are examined in order to identify outliers [26], [12], [27], etc.
Outlier detection methods are categorized in two major groups based on the kind of outlier they intend to discover: global methods and local methods. The distanced- based definition of outlier was proposed for the first time from [28]. It comprises the first data based-oriented approach in the domain of outlier detection. On the other hand, local methods are been extensively explored from significant strategies. Variants of local approaches have been proposed in [13] and [29]. Another work related to detecting the principle of local outlier is LDOF [9]. Rather than in terms of speed and computation time, ensemble based approaches exhibit the capability of enhancing the performance of their algorithms in terms of quality of detection. Shubert et al. [2] referred to similarity measure in order to accurately interpret different outlier rankings and to evaluate the diversity that different outlier detection algorithms generate. Not all this line of research proposed in the body of knowledge has pursued the inducement of diversity from novel methods. Subsequently, they have not found out new tools of inducing diversity for the selected pool of outlier detector. That is why is advocated that the theoretical fundamentals of ensemble based approaches for outlier detection is not advanced and immature even though they have borrowed some sound principles from the rich tradition of supervised leaning techniques for ensemble analysis.
In [30] is introduced subsampling technique as a meaningful mean of inducing diversity between detector members. It is given evidence that executing an ensemble of various algorithms detectors is a subsample of the data is more effective than the other methods of inducing diversity. The paper major contributions is in demonstrating empirically that it can be constructed an ensemble for outlier detection, which outperforms the individual outlier algorithms if they are executed individually.

3. Proposed Strategy

Our strategy derives from the greedy ensemble of Schubert et.al [33]. We have modified this ensemble algorithm by increasing its diversity throughout randomization. The key principle of the greedy ensemble relies on the diversity maximazation and at the same time keeping the size of the ensemble small. It is demonstrated [21] that using the diversity factor can enhance and migliorate the performance of the ensemble substantially. The available outlier detectors utilized as ensemble members are variants of kNN-based outlier detection: Local Outlier Factor (LOF) [10], LDOF [9], LoOP [13], KNN [7], k-NN [4], [32]. Authors have carried out experiments using the ELKI data mining platform. After exploring and examining in depth the greedy approach, we have inspected a crucial gap in this line of research that let us to do further improvement. The key idea is that while greedy approach is iterating in order to find the best outlier detector and thus discarding with no future chance to evaluate the algorithms back, we propose to perform a randomization. In this moment of time, for instance, the ensemble must not choose the best of algorithm closer to the target vector or the less closer, but one random detector. We had this idea due to the logic that the best outlier sometimes does not lead to the best outcome. It happens that while searching for the best result in local climbing search, it may find out just the local maximum data point, but not the global maximum of the whole dataset. Therefore, in order to escape the local maxima, we need to employ randomization techniques to the greedy ensemble. Moreover, by randomizing we make sure that the search will continue in a random data point. We have conceptualized that this kind of methodology can lead to substantial increase of accuracy and diversity. Randomization in practice will be induced by modifying the pseudo code and at the same time, implementing this change in the java code of the greedy ensemble construction, which is provided from the data-mining group. We have used the Netbeans java interface in order to modify the code. In java, the class of random represents randomization. This randomization problem is analogues with the rolling dies problem. We will set a probability α, which will be given externally, and according to this probability are going to be select or to choose randomly the remaining outlier detectors. After selecting a random algorithm, we will test and run the greedy ensemble with the new added detector. Empirical results show that the outputs will be diversified and new kind of outliers not discovered before will be captured from the randomized greedy ensemble approach. Therefore, we can foster that deploying randomization techniques in the greedy ensemble will contribute in identifying and capturing new malicious data points.

Conclusion and Future work

This paper is a short overview of the major steps that the construction of the randomized ensemble-based approach will follow and under what circumstances our research direction is founded. The key idea of our work is that while greedy approach is iterating in order to find the best outlier detector and thus discarding with no future chance to evaluate the algorithms back, we propose to perform a randomization. In this moment of time, the ensemble must not choose the best of algorithm closer to the target vector, but one random detector. Randomization in practice will be induced by modifying the pseudo code and at the same time, implementing this change in the java code of the greedy ensemble construction, which is provided from the data-mining group. We have used the Netbeans java interface in order to modify the code. In future, we plan to construct an ensemble based on the combining and selected different outlier algorithms with different parameters. Numerous experiments are predicted to be carried out. Through various experiments, we are going to build our approach by providing strong experimental results.

References

1. Hawkins, D.: Identification of Outliers. s.l. Chapman and Hall, (1980)
2. Schubert, E., Wojdanowski, R., Zimek, A., Kriegel, H.: On Evaluation of Outlier Rankings and Outlier Scores.. Anaheim, CA : SIAM, Proceedings of the 2012 SIAM International Conference on Data Mining. (2012) pp. 1047-1058
3. Hadi, A. S., Rahmatullah Imon, A., H.M. Werner, M. :Detection of outliers. WIREs Comp.(2009) pp. 57–70
4. Angiulli , F. and Pizzuti, C.: Fast outlier detection in high dimensional spaces. Proc. PKDD (2002) pp. 15–26
5. Knorr, E. M., Ng, R. T., Tucanov, V.: Distance based outliers: Algorithms and applications. VLDB (2000) pp. 237–253
6.Orair, G. H., Teixeira, C., Wang, Y., Meira Jr, W.; Parthasarathy, S.: Distance-based outlier detection:Consolidation and renewed bearing. PVLDB (2010) pp. 1469–1480
7. Ramaswamy, S., Rastogi, R. and Shim, K.: Efficient algorithms for mining outliers from large data sets. Proc. SIGMOD (2000) pp. 427–438
8. Vu , N. H. and Gopalkrishnan, V.: Efficient pruning schemes for distance-based outlier detection. Proc.ECML PKDD (2009) pp. 160–175
9. Zhang, K., Hutter, M. and Jin, H.: A new local distance based outlier detection approach for scattered real world. Proc. PAKDD (2009) pp. 813–822
10.Breunig, M. M., Kriegel, H. P., Ng, R., Sander, J.: LOF: Identifying density-based local outliers. In Proc.SIGMOD (2000) pp. 93-104
11. De Vries, T., Chawla, S. and Houle, M. E.: Finding local anomalies in very high dimensional space. Proc.ICDM (2010) pp. 128–137
12. Keller, F.; Müller, E. and Böhm, K.: HiCS: high contrast subspaces for density-based outlier ranking. Proc.ICDE (2012)
13. Kriegel, H. -P., Kroger, P., Schubert, E., Zimek, A.: LoOP: local outlier probabilities. Proc. CIKM (2009) pp. 1649-1652
14. Lazarevic, A. and Kumar, V.: Feature Bagging for Outlier Detection. Proc. KDD. ACM. Chicago, Illinois (2005) pp. 157-166
15. Breiman, L.: Bagging Predictors. Machine Learning (1996) pp. 123-140
16. Chawla, N., et al.: SMOTEBoost: Improving the Prediction of Minority Class in Boosting. In Proceedings of the Principles of Knowledge Discovery in Databases PKDD (2003)
17. Freund, Y. and Schapire, R.: Experiments with a New Boosting Algorithm. In Proceedings of the 13th International Conference on Machine Learning.Bari : s.n. (1996) pp. 325-332
18. Joshi, M.; Agarwal , R.; Kumar, V.: Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong? Proceedings of the Eight ACM Conference ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
Edmonton, Canada: (2002)
19. Kong, E., Dietterich, T.: Error-Correcting Output Coding Corrects Bias and Variance. In

Proceedings of the 12th International Conference on Machine Learning. San-Francisco CA : (1995) pp. 313-321
20. Ghosh , J., Acharya, A.: Cluster ensembles . Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. (2011) pp. 305–315
21. Schubert, E. Generalized and Effiicient Outlier Detection for Spatial, Temporal, and High-Dimensional Data Mining. Munchen, Germany : s.n. (2013)
22. Aggarwal, Ch. C.: Outlier Ensembles. ACM SIGKDD Explorations Newsletter I2012) pp. 49-58
23. Gao, J. and Tan, P. -N.: Converting Output Scores from Outlier Detection Algorithms into Probability Estimates Hong-Kong : IEEE Data Mining ICDM '06. Sixth International Conference (2006) pp. 212 - 221
24. Nguyen, H. V., Ang, H. H. and Gopalkrishnan, V.: Mining outliers with ensemble of heterogeneous detectors on random subspaces. Proc. DASFAA (2010) pp. 368-383
25. H.-P. Kriegel, P. Kroger, E. Schubert, and A. Zimek.: Interpreting and unifying outlier scores. Proc.SDM (2011) pp. 13-24
26. He, Z., Deng , S., Xu, X.: A Unified Subspace Outlier Ensemble Framework for Outlier Detection. Advances in Web Age Information Management (2005)
27. Muller, E.; Schiffer, M.; Seidl, T.: Statistical Selection of Relevant Subspace Projections for Outlier Ranking. ICDE Conference. (2011) pp. 434–445
28. Knorr , E. M. and Ng, R. T.: A unified notion of outliers: Properties and computation. Proc. KDD (1997) pp. 219-222
29. Papadimitriou, S.; Kitagawa, H.; Gibbons, P. B.: LOCI: Fast Outlier Detection Using the Local Correlation Integral. IEEE 19th International Conference on Data Engineering (ICDE'03) (2003)
30. Zimek, A.; Gaudet, M.; Campello, R. J. G. B.; Sander, J.: Subsampling for Efficient and Effective Unsupervised Outlier Detection Ensembles. Proceeding KDD '13 Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM. New York, USA (2013) pp. 428-436
31. Zimek, A., Campello, R. J. G. B. and Sander, J.: Ensembles for Unsupervised Outlier Detection. ACM SIGKDD Explorations Newsletter (2013) pp. 11-22
32. Kriegel, H. -P.; Schubert, M. and Zimek, A.: Angle-based outlier detection in high-dimensional data. KDD (2008) pp. 444–452
33. Shubert, E., Wojdanowski, R., Zimek, A., Kriegel, H.-P.: On evaluation of outlier rankings and outlier scores. 11th SIAM International Conference on Data Mining (SDM) s.n., (2011)
34. Zhang, K.; Hutter, M.; Jin, H.: A new local distance-based outlier detection approach for scattered. Proc. PAKDD (2009) pp. 813-822.
35. Papadimitriou, S., et al: Fast outlier detection using the local correlation integral. Proc. ICDE (2003) pp. 315-326
36. Ghosh , J., and Acharya, A.: Cluster ensembles. WIREs DMKD (2011) pp. 305–315
37. Nguyen, H. V.; Ang, H. H.and Gopalkrishnan, V.: Mining outliers with ensemble of heterogeneous detectors on random subspaces. In Proc. DASFAA (2010) pp. 368–383.