International Journal of Modern Computer Science (IJMCS) ISSN: 2320-7868 (Online)
Volume 7, Issue 3, June, 2019

A Survey on Anomaly Detection Behaviour with Kernel Mapping and Inclusion

Dr R.Umagandhi
Associate Professor & Head
Department of Computer Technology
Kongunadu Arts and Science College
Coimbatore, India
E-mail: umakongunadu@gmail.com

K.Gomathi
M.Phil Research Scholar
Kongunadu Arts and Science College
Coimbatore, India
E-mail:gomathi.karupuswamy@gmail.com

Abstract: Anomaly detection is a significant problem that has been researched within various research areas and application domains. Many anomaly detection methods have been particularly examined for certain application domains, as others are more standard. This survey papers is describes an anomaly detection technique for unsupervised data sets accurately reduce the data from a kernel Eigen space lacking performing a batch re-computation. For each anomaly behavior activities is to identify the key factors, which are used by the methods to differentiate between normal and abnormal actions. This survey paper provides a best and brief understanding of the techniques belonging to each anomaly and kernel mapping category. Further, for each grouping, to identify the improvements and drawbacks of the techniques in that category. It also provides a discussion on the computational complexity of the techniques since it is an important issue in real application domains hope that this survey will provide a good understanding of the many directions in which research has been done on this topic.

Keywords: Adaptive, non-stationary, anomaly detection, outlier detection, kernel principal component analysis, kernel methods.

I. Introduction

Non increasing number of real-world applications, data used in techniques such as machine learning and data mining is provided by a data stream. A data stream is a continuous flow of data measurements from a source where the measurements are provided one instance at a time. A data stream provides a more natural representation of a machine learning problem where the environment is changing as data is continuously generated. The characteristics of data streams mean the entire data set is not available at any one time. Subsets of the data set are used as training sets, with testing sets being drawn from the same distribution. It is usually the case that subsets are formed from contiguous data instances. An important aspect of data streams is the non-stationary nature of the data, meaning that the concepts to be learned evolve in time. This requires an update to the model that is being used to classify data. A batch approach to the problem requires a reconstruction of the model each time an update is required. The training phase is often the most computation-ally costly operation. Incremental learning overcomes this issue by using the previous model as the basis for an update.

One learning problem is anomaly detection, an also known as outlier detection. An anomaly in a data set is defined by Barnett and Lewis as “an observation (or sub-set of observations) which appears to be inconsistent with the remainder of that set of data” [1]. Anomaly detection aims to identify data that do not conform to the patterns exhibited by the data set [2]. The problem is often viewed as an unsupervised one-class classification problem. This equates to a problem with two important characteristics. Labels are not available for data instances in the training set. In addition, there is a class imbalance in the training set where the number of normal data exceeds that of the anomaly data. An adaptive incremental anomaly detection scheme based on Kernel Principal Component Analysis (KPCA) [3] is proposed. An accurate incremental split to a kernel Eigen Space (KES) is proposed that is shown to be more accurate than state-of-the-art methods. This is coupled with a KES merge to form a Split-Merge KES algorithm that allows the addition and removal of data instances to an anomaly detection model. The aim of the anomaly detector is to identify data in the testing set that is drawn from a different data distribution than the normal data in the training set.

A non-stationary environment is considered where the data distribution of the normal data changes with time. An adaptive version determines an appropriate sliding window size and reduces the number of updates that are required by detecting when a change has occurred and therefore only updating when necessary. Split-Merge KES and Adaptive Split-Merge KES are compared with other state-of-the-art batch and incremental anomaly detection techniques and are shown to have superior performance.

A kernelized version of the Eigen space splitting algorithm of Hall et al. [4] is reported. This is shown to be more accurate than state-of-the-art techniques. A detailed evaluation of inaccuracies introduced by incremental KPCA algorithms is provided. A new anomaly detection technique based on incremental KPCA is developed. The technique is shown to be more accurate both in terms of the tracked KES and anomaly detection. An adaptive algorithm is presented which uses there construction error to determine the window size and when an update should occur.

II. LITERATURE SURVEY

Peter Hall et al .[4]discussed the algorithms for adding and subtracting Eigen spaces, thus allowing for incremental updating and down-dating of data models. Importantly, and unlike previous work, we keep an accurate track of the mean of the data, which allows our methods to be used in classification applications. The result of adding Eigen spaces, each made from a set of data, is an approximation to that which would obtain were the sets of data taken together. Subtracting Eigen spaces yields a result approximating that which would obtain were a subset of data used. Using our algorithms, it is possible to perform 'arithmetic' on Eigen spaces without reference to the original data. Eigen spaces can be constructed using either EVD (Eigen Value Decomposition) or SVD (Singular Value Decomposition). We provide addition operators for both methods, but subtraction for EVD only, arguing there is no closed-form solution for SVD. The methods and discussion surrounding SVD provide the principle novelty. We illustrate the use of our algorithms in three generic applications, including the dynamic construction of Gaussian mixture models.q2002 Elsevier Science B.V. All rights reserved.

Rodriguez et al. [5] says that any change in the classification problem in the course of on-line classification is termed changing environments. Examples of changing environments include change in the underlying data distribution, change in the class definition, adding or removing a feature. The two general strategies for handling changing environments are (i) constant update of the classifier and (ii) re-training of the classifier after change detection. The former strategy is useful with gradual changes while the latter is useful with abrupt changes. If the type of changes is not known in advance, a combination of the two strategies may be advantageous. We propose a classifier ensemble using Winnow. For the constant-update strategy we used the nearest neighbour with a fixed size window and two methods with a learning rate: the online perception and an online version of the linear discriminate classifier (LDC). For the detect-and-retrain strategy we used the nearest neighbour classifier and the online LDC. Experiments were carried out on 28 data sets and 3 different scenarios: no change, gradual change and abrupt change. The results indicate that the combination works better than each strategy on its own.

Zhang et al. [6] says that Data collected by WSNs (Wireless Sensor Networks) are inherently unreliable. Therefore, to ensure high data quality, secure monitoring, and reliable detection of interesting and critical events, outlier detection mechanisms are needed to be in place. The constraint nature of resources available in WSNs necessities that unlike traditional outlier detection techniques performed off-line, outliers to be identified in an online manner. This means that outliers in distributed streaming data should be detected in (near) real time with a high accuracy while maintaining the resource consumption of the WSN to a minimum. The authors discussed the outlier detection techniques based on one-class quarter-sphere support vector machine meeting constraints and requirements of WSNs. To reduce the false alarm rate while increasing the detection rate and to enable collaborative outliers detection, we take advantage of spatial and temporal correlations that exist between sensor data. Experiments with both synthetic and real data show that our distributed and online outlier detection techniques achieve better detection accuracy and lower false alarm compared to an earlier distributed, batch outlier detection technique designed for WSNs.

Breunig et al. [7] Address the problem For many KDD applications, such as detecting criminal activities in E-commerce, finding the rare instances or the outliers, can be more interesting than finding the common patterns. Existing work in outlier detection regards being an outlier as a binary property. To contend that for many scenarios, it is more meaningful to assign to each object a degree of being an outlier. This degree is called the LOF (Local Outlier Factor) of an object. It is local in that the degree depends on how isolated the object is with respect to the surrounding neighborhood. We give a detailed formal analysis showing that LOF enjoys many desirable properties. Using real-world datasets, we demonstrate that LOF can be used to find outliers which appear to be meaningful, but can otherwise not be identified with existing approaches. Finally, a careful performance evaluation of our algorithm confirms we show that our approach of finding lo-cal outliers can be practical.

Kriegel et al. [8] address the problem for Detecting outliers in a large set of data objects is a major data mining task aiming at finding different mechanisms responsible for different groups of objects in a data set. All existing approaches, however, are based on an assessment of distances (sometimes in-directly by assuming certain distributions) in the full-dimensional Euclidean data space. In high-dimensional data, these approaches are bound to deteriorate due to the notorious "curse of dimensionality". The authors proposed a novel approach named ABOD (Angle-Based Outlier Detection) and some variants assessing the variance in the angles between the difference vectors of a point to the other points. This way, the effects of the "curse of dimensionality" are alleviated compared to purely distance-based approaches. A main advantage of our new approach is that our method does not rely on any parameter selection influencing the quality of the achieved ranking. In a thorough experimental evaluation, we compare ABOD to the well-established distance-based method LOF for various artificial and a real world data set and show ABOD to per-form especially well on high-dimensional data.

Ding et al. [9] says that Random projection is widely used as a method of dimension reduction. In recent years, its combination on with standard techniques of regression and classification has been explored. Here we examine its use for anomaly detection in high-dimensional settings, in conjunction with PCA (Principal Component Analysis) and corresponding subspace detection methods. We assume also-called spiked covariance model for the underlying data generation process and a Gaussian random projection. We adopt a hypothesis testing perspective of the anomaly detection problem, with the test statistic defined to be the magnitude of the residuals of a PCA analysis. Under the null hypothesis of no anomaly, we characterize the relative accuracy with which the mean and variance of the test statistic from compressed data approximate those of the corresponding test statistic from uncompressed data. Furthermore, under a suitable alternative hypothesis, we provide expressions that allow for a comparison of statistical power for detection. Finally, whereas these results correspond to the ideal setting in which the data covariance is known, we show that it is possible to obtain the same order of accuracy when the covariance of the compressed measurements is estimated using a sample covariance, as long as the number of measurements is of the same order of magnitude as the reduced dimensionality.

Lee et al. [10] says that Anomaly detection has been an important research topic in data mining and machine learning. Many real-world applications such as intrusion or credit card fraud detection require an effective and efficient framework to identify deviated data instances. However, most anomaly detection methods are typically implemented in batch mode, and thus cannot be easily extended to large-scale problems without sacrificing computation and memory requirements. To proposed an online osPCA (Over Sampling Principal Component Analysis) algorithm to address this problem, and we aim at detecting the presence of outliers from a large amount of data via an online updating technique. Unlike prior PCA (Principal Component Analysis) -based approaches, we do not store the entire data matrix or covariance matrix, and thus our approach is especially of interest in online or large-scale problems. By oversampling the target instance and extracting the principal direction of the data, the proposed osPCA allows us to determine the anomaly of the target instance according to the variation of the resulting dominant eigenvector. Since our osPCA need not perform eigen analysis explicitly, the proposed framework is favored for online applications which have computation or memory limitations. Compared with the well-known power method for PCA and other popular anomaly detection algorithms, our experimental results verify the feasibility of our proposed method in terms of both accuracy and efficiency.

Hoegaertsa et al. [11] says that the dominant set of eigenvectors of the symmetric kernel Gram matrix is frequently used in many important kernel methods (like e.g. kernel Principal Component Analysis, feature approximation, de-noising, compression, prediction) in the machine learning domain. Yet in the case of dynamic and/or large scale data, the batch calculation nature and computational demands of the eigenvector decomposition limit these methods in numerous applications. To present an efficient incremental approach for fast calculation of the dominant kernel Eigen basis, which allows totrack the kernel Eigen space dynamically. Experiments show that our updating scheme delivers a numerically stable and accurate approximation at every iteration in comparison to the batch algorithm.

Tax et al. [12] says that two useful extensions of the incremental SVM in the context of online learning. An online support vector data description algorithm enables application of the online paradigm to unsupervised learning. Furthermore, online learning can be used in the large-scale classification problems to limit the memory requirements for storage of the kernel matrix. The pro-posed algorithms are evaluated on the task of online monitoring of EEG data, and on the classification task of learning the USPS dataset with a-priori chosen working set size.

III. Conclusion

Anomaly detection in data mining areas is efficient and effective task to ensure the quality and right decisions. A selection of anomaly detection models was proposed in the literature survey; however, most of them suffer from high dimensional datasets effectiveness or high outliers. This survey shows the challenges that face the design of an efficient and effective anomaly detection model for synthetic and real-world data sets in data mining domain that should be satisfied to design such models.

REFERENCES

1. V. Barnett and T. Lewis, “Outliers in Statistical Data,” New York, USA, vol. 3,1994.
2. V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” vol. 41, no. 3, pp. 15:1–15:58, 2009.
3. B. Scheolkopf, A. Smola, and K.R. Meuller, “Nonlinear component analysis as a kernel Eigen value problem”, Neural Comput., vol. 10, no. 5, pp. 1299–1319, 1998.
4. P. Hall, D. Marshall, and R. Martin, “Adding and subtracting Eigen spaces with Eigen value decomposition and singular value decomposition,” Image Vis, vol. 20, no. 13, pp. 1009–1016, 2002.
5. J. J. Rodrıguez and L. I. Kuncheva, “Combining online classification approaches for changing environments,” in

Proc.Int. Workshop Struct., Syntactic, Statist. Pattern Recognit, 2008, pp. 520–529.

6. Y. Zhang, N. Meratnia, and P. J. Havinga, "Ensuring high sensor data quality through use of online outlier detection techniques," vol. 7, no. 3, pp. 141–151, 2010.
7. M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," ACM SIGMOD Rec., vol. 29, no. 2, pp. 93–104, 2000.
8. H.-P. Kriegel, A. Zimek, and M. Schubert, "Angle-based outlier detection in high-dimensional data," in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 444–452, Aug. 2008.
9. Q. Ding and E. D. Kolaczyk, "A compressed PCA subspace method for anomaly detection in high-dimensional data," IEEETrans. Inf. Theory, vol. 59, no. 11, pp. 7419–7433, Nov. 2013.
10. Y. Lee, Y. Yeh, and Y. Wang, "Anomaly detection via online oversampling principal component analysis," IEEE Trans.Knowl. Data Eng., vol. 25, no. 7, pp. 1460–1470, Jul. 2013.
11. L. Hoegaerts, L. De Lathauwer, I. Goethals, J. A. Suykens, J. Vandewalle, and B. De Moor, "Efficiently updating and tracking the dominant kernel principal components," Neural Netw., vol. 20, no. 2, pp. 220–229, 2007.
12. D. M. Tax and R. P. Duin, "Support vector data description," Mach. Learn., vol. 54, no. 1, pp. 45–66, 2004.