A QR Decomposition Approach for Improved Anomaly Detection over Non-linear Data

Lince Rachel Varghese
M.Phil (CS), Research Scholar
School of IT and Science
Dr.G.R.D College of Science, Coimbatore

N. Sudha Bhuvaneswari
Associate Professor, MCA, M.Phil (CS), PhD
School of IT and Science
Dr.G.R.D College of Science, Coimbatore

ABSTRACT

Anomalies are patterns that lack normal behavior. Anomaly detection process can be used to predict changes in input patterns and can prevent different malicious activities. Kernel Principal Component Analysis is an effective, anomaly detection technique. This is an effective technique for non-linear data set. In this, kernel Eigenspace splitting and merging approach is used for predicting the anomalies effectively. Kernel splitting process, can extract smaller KES from larger KES. Here a QR decomposition technique can be used for, KES splitting. This is done to remove the data patterns that no longer fit to the current data distribution.KES splitting is combined with KES update. An adaptive update takes an appropriate sliding window size into consideration and reduces the number of updates required, when a change in the data distribution occurs. Thus an adaptive split-merge KES with QR decomposition, shows a superior performance in predicting anomalies that have a non-linear behavior.

General Terms

Anomaly detection, KPCA, Non-linear, QR decomposition,

Keywords

Anomaly detection, KPCA, QR decomposition, KES

1. INTRODUCTION

Anomalies are set of objects which are considerably different from the remainder of the data [1]. Anomalies can also be called outliers, observations, exceptions, contaminants, peculiarities, discordant, etc in different domains. The main aim of anomaly detection is to identify data that do not conform to the patterns exhibited by the data set [2]. The input data instances are described by binary, categorical or continuous attributes. The output of an outlier detection algorithm can be a score ,which is the level of "outlierness " of data points, or it can be a binary label which shows whether the data point is an outlier or not. Anomalies can be of different types, point, contextual or collective. Most of the researches focus on point anomalies. Based on the availability of user labels, outlier detection methods can be classified as supervised, unsupervised and semi-supervised. Based on the different assumptions made about normal data and outliers, there are different outlier detection methods such as statistical, proximity–based and cluster based [3]. This problem comes under unsupervised [6] classification. QR decomposition based adaptive incremental anomaly detection over non-linear data sets, use split-merge KES algorithm [4] and shows to have a superior performance.

2. RELATED WORK

The first approach to overcome the limitations of global view of outlierness has been the density based local outlierness (LOF).LOF [7]compares the density of each object O of a data set D with the density of k-nearest neighbors' of O .By providing an incremental update to LOF,can add and remove data instances from a model. But all this has less effect on high dimensional data.

Angle Based Outlier detection (ABOD) meets this difficulty in high dimensional data. ABOD [8] access, the variance in angles between an outlier, and all other pairs of points. But the problem of irrelevant attributes, in case of high dimensional data is not addressed by ABOD.

Hyper Graph Based outlier detection [9] is an outlier mining method based on hyper graph model for categorical data. Hyper graphs are able to capture the distribution characteristics in data subspaces.

PCA [10] relays on spectral decomposition. An orthogonal transformation is used to convert, a set of correlated variables into linearly uncorrelated variables which is called principal components. An incremental update and an on-line anomaly detection technique are available for PCA. But the drawback of these models is that, data cannot be removed from the model.

When data set is very large, addition or removal of single outlier instances will never significantly affect the resulting principal direction. So an oversampling PCA (osPCA) [11] can be used for problems, that deals with large scale anomaly detection. Here the target instance is duplicated many times. The idea behind this is, to amplify the effect of outlier than the effect of normal data. The online osPCA method is able to determine the anomaly of the target instances, without satisfying memory and computation efficiency

Kernel Principal Component Analysis KPCA [12] and reconstruction error is found as an effective anomaly detection technique, for non-linear data sets. In the case of a non-stationary environment to represent the current data distribution, the recomputation of kernel eigenspace is required. Recomputation is a complex operation and the challenge is to reduce the computational complexity. Using KPCA, the estimation of non-linear kernel principal components (KPCs) suitable in describing highly complex data distributions are possible. KPCA almost always outperforms PCA in most comparisons. But KPCA require large training data sets .So it is difficult to store and manipulate all data at once. The resulting KPCs are defined implicitly by using linear expansion of the training data. So the data should be saved after training. This cause high cost for storage resources and high computational load. In the case of on-line data processing, the main drawback of KPCA is that, it is computable only in batch manner.

In case of incremental KPCA [13], it requires the update of KES. The main drawback is that the mean of the data in reproducing kernel Hilbert space (RKHS) does not alter. Liu et al. presents an algorithm which is able to update a KES with a single data instance which accounts for a change of means. Khediri et al. [15] presents a technique that allows block updates. Sharma et al. [16] allows the merging of eigenspaces with adaptation to a changing sample mean.

3. METHODOLOGY

3.1 Adaptive KPCA for Anomaly Detection

When working in a non stationary environment, in order to adapt to the changing concepts, the model should be updated. Updating a model consist of three states. In the first stage detect a change in the data distribution. In the second stage, data that no longer represent the current concept has to be removed. In the last stage, add data to the model that represents the current data distribution. In the case of batch KPCA the previous model should be discarded whenever a new model is constructed from the updated training set. The drawback of batch update is the computational cost. To overcome the issue of non stationary environment we can use sliding window [14], [15] of data. Here when an additional M data instance arrives they are incorporated into the model and the oldest M data instances are removed. Now the model is able to adapt to the data distributions. To avoid the unnecessary updates we use the KES to detect when a change has occurred. Whenever a change is detected an update is made by KES split and merge algorithms, to remove and add data to KES. This has a lower computational complexity than batch KPCA [17].

3.2 Splitting and Merging Kernel Eigenspaces

QR decomposition is used for splitting. Splitting removes a pattern that no longer fit to the current pattern concept. QR decomposition extracts a smaller KES from the larger KES to form the new KES. If cholesky decomposition is used for splitting it is unable to provide solution at the time of rounding process and the changes reflects a lot in system. QR decomposition overcomes this problem. In linear algebra QR decomposition is also called QR factorization. Here a matrix A is decomposed into a product A = QR. Q is an orthogonal matrix and R is an upper triangular matrix. QR decomposition is mainly used to solve, linear least square problems and this forms the basis for particular eigenvalue algorithm, the QR algorithm.

If A is $m \times n$ and left-invertible, then it can be factored as A = QR , R is $n \times n$ and upper triangular with $r_{ii} > 0$,Q is $m \times n$ and orthogonal ($Q^TQ = I$) rewrite normal equations $A^TAx = A^Tb$ using QR factorization A = QR

$$A^TAx = A^T b$$

$$R^TQ^TQRx = R^TQ^T b$$

$$R^TRx = R^TQ^T b \ (Q^TQ = I)$$

$$Rx = Q^T b \ (R \text{ nonsingular})$$

3.3 Adaptive Scheme

In the case of a non stationary environment by using a sliding window approach the updates will be performed unnecessarily. So an adaptive scheme is used. Here a minimum size of sliding window is determined first and an adaptive updates scheme is used to perform updates to the model, whenever necessary [17].

3.3.1 *Selection Criteria*

According to [17] mean reconstruction error can be used as a measure, to determine, when an update should occur to the model. Anomalies and data distribution changes cause a large reconstruction error of data instances. Since the data set mainly contain normal data, the reconstruction error of anomalies will only have a small impact on the mean reconstruction error. If the data distribution changes, both the normal and anomaly data in the training set will have a large reconstruction error. Now the mean of the reconstruction error increases, and it allows distinguishing between a data block containing anomalies, and a data block with anomalies and with a changed data distribution.

$$\overline{\varepsilon}\ _{training\ set} = \frac{\sum_{i=1}^{N} \varepsilon(i)}{N}$$

$$\overline{\varepsilon}\ _{update} = \frac{\sum_{i=1}^{M} \varepsilon(i)}{M}$$

$$\overline{\varepsilon}\ _{ratio} = \frac{\overline{\varepsilon}_{update}}{\overline{\varepsilon}_{trainingset}}$$

Fig 1: Calculating the mean of reconstruction error and error ratio

If the reconstruction error ratio is large, it represents that the data does not match the current model. The threshold value ν is predefined and is used to determine when ever an update is required. When a block of data arrives the data can be merged into the model if $\varepsilon_{ratio} > \nu$. In an adaptive anomaly detection technique an adaptive split merge KES is proposed. The adaptive scheme consists of two phases.

1. Initialization
2. Adaptive updates

In the initialization stage, determine the number of training data instances that is required to model the data distribution. Batch KPCA is used on an initial block of data. For this training set, the mean reconstruction error is calculated. Then the mean reconstruction error for the next block of data is calculated. If $\varepsilon_{ratio} > \nu$ it represent that the current block of data does not represent the current model, and this is added to training set by KES merge operation. If the ratio is lower it is not added to the training set.

The operation of the adaptive split merge KES algorithm [17] is as follows

Initialization

Window of N data instances

Batch KPCA on initial M data instances

Calculate $\varepsilon^{-}_{\text{training set}}$

For every set of M data instances

1. Project data onto KPCs
2. Calculate ε^{-}_{update} and ε_{ratio}
3. If $\varepsilon_{ratio} > \nu$

```
            KES Merge
            Calculate $\bar{\varepsilon}_{\text{training set}}$
    Else
            Discard
    Online operation
    For every set of M newly arrived data
1.  Calculate $\bar{\varepsilon}_{\text{update}}$ and $\varepsilon_{\text{ratio}}$
2.  If $\varepsilon_{\text{ratio}} > \nu$
            KES split Remove M oldest Data vectors
            KES merge Add M new Data Vectors
            Calculate $\bar{\varepsilon}_{\text{training set}}$
```

4. EXPERIMENTAL RESULT

Experimental test is conducted on the proposed methodology to find the improvements in terms of relative eigenvalue errors. The performance evaluation was conducted with the help of abalone data set.

4.1 Abalone Data Set

The Abalone data set contains physical measurements of abalone, which are large edible sea snails. The data set comes from 1994 study “The population of biology of abalones”. There are 4177 rows and 9 columns. The columns include 1 categorical predictor (sex), 7 continuous predictors (Length, Diameter, Height, Whole weight, Shucked weight, Viscera weight, Shell weight), and an integer response variable (number of rings).

Although the physical measurements can be used to predict the number of rings (and thus its age) with some accuracy, it is noted that information not present in the dataset (weather patterns and location, hence food availability) could be used to improve the accuracy of predictions.

Number of Attributes: 8

The abalone data set will be divided into 100, 200 and so on in order to detect the anomalies and check the effectiveness of the proposed procedure**.**

The data set $X = (x_1, x_2,\ldots,x_m)$ is firstly divided into M subsets X_i (i =1,…, M), each of which consists of about k = m/M. Without loss of generality, it is denoted:

$X_1 = (x_1, \ldots,x_k)$

$X_2 = (_{xk+1},..x_{2k})$,

. .

$X_M = (x_{(M-1)k+1},\ldots x_m)$

The data set $X = (X_1, X_2, \ldots\ldots, X_M)$ is then transformed into $\Sigma = (\Sigma_1, \Sigma_2,\ldots, \Sigma_M)$, where Σ_i is n x n auto correlation matrix. Then the eigenvalues and eigenvectors of the matrix are computed.

In linear algebra, the Cholesky decomposition or Cholesky factorization is a decomposition of a Hermitian, positive-definite matrix into the product of a lower triangular matrix and its conjugate transpose, useful for efficient numerical solutions. However this methodology is unable to provide solution at the time of rounding process where the changes reflect a lot, in system. Here Cholesky decomposition is replaced by QR factorization technique for the kernel splitting process.

Table 1. Attribute information of abalone data set

Name	**Data Type**	**Measurement**	**Description**
Sex	nominal		M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Rings	Integer		+1.5 gives the age in years

4.2 Calculation of relative eigenvalue error

The relative error is the absolute error divided by the magnitude of the exact value. In figure 2 the X axis represents the number of data points and the Y axis represents the relative eigenvalue error. The relative eigenvalue error is calculated for KPCA with cholesky decomposition and for KPCA with QR decomposition. The QR decomposition shows an improved in performance.

5. CONCLUSION

Real-world data sets are used for evaluating the performance. Data stream consist of normal data and anomalies. In non-stationary environment, a Split-Merge, KES is used for anomaly detection. QR factorization technique is introduced for eigen splitting process. In the splitting process, compared with choleskey decomposition our approach is able to achieve satisfactory results. Thus an adaptive split-merge KES with QR decomposition, shows a superior performance in predicting anomalies that have a non-linear behaviour. Different techniques are available for anomaly detection. The main issue is that, we have to determine, the technique that is most appropriate for our problem in hand. As a future work, KPCA can be combined with an improved Support Vector Machine (SVM), or any other classification methods, and the combination can be used for anomaly detection.

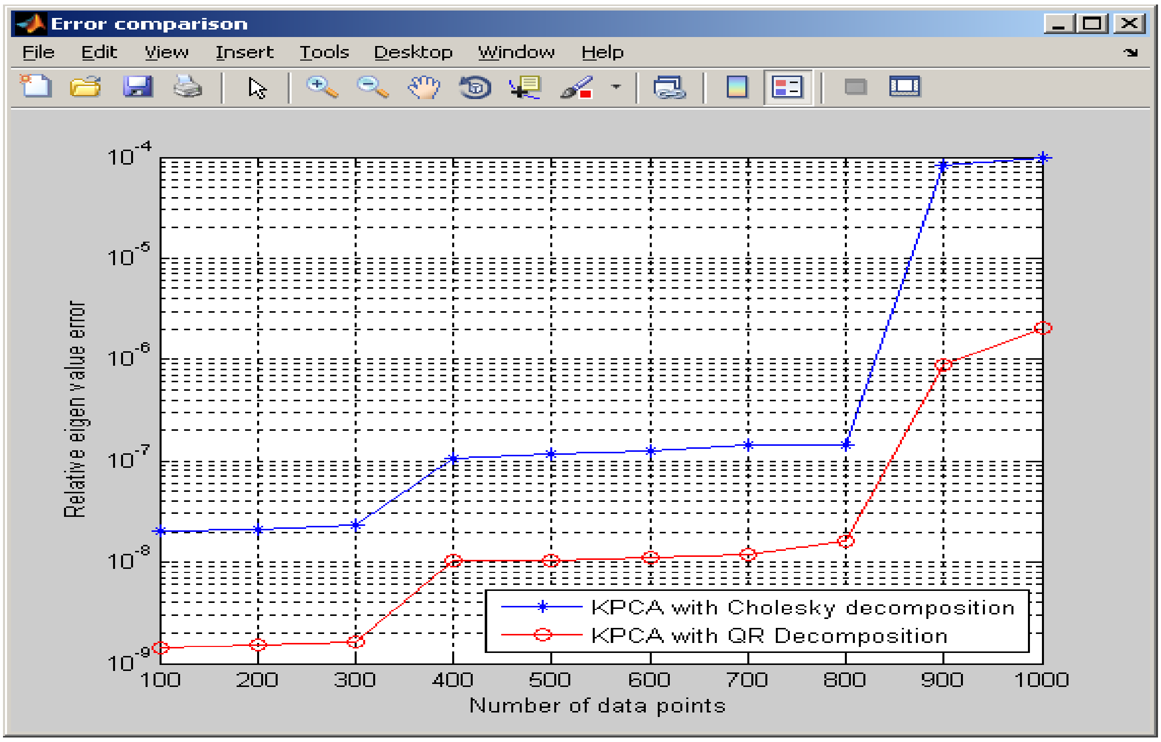

Fig 2: Eigen error rate comparison

6. REFERENCES

[1] V.Chandola,A.Banerjee,andV.Kumar,“Anomalydetection:A survey,” ACM Comput. Surv., vol. 41, no. 3, pp. 15:1–15:58, 2009.

[2] V.Barnett and T.Lewis,Outliers in Statistical Data .New York,NY,USA:Wiley,1994,vol.3.

[3] Jiawei Han,Micheline Kamber and Jian Pei,”Data Mining :Concepts and Techniques”,2nd ed.Morgan Kaufmann,2006.

[4] P. Hall, D. Marshall, and R. Martin, “Adding and subtracting eigenspaces with eigenvalue decomposition and singular value decomposition,” Image Vis. Comput., vol. 20, no. 13, pp. 1009–1016,2002.

[5] Kou, Y., Lu, C., & Sinvongwattana, S. Survey of Fraud Detection Techniques Yo-Ping Huang, 749–754. 2004.

[6] E.Eskin et.al ,A geometric framework for unsupervised anomaly detection ,pp 77-101,2002.

[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, Aug. 2008, pp. 444–452.

[9] Wei, Li, et al. "Hot: Hypergraph-based outlier test for categorical data."*Advances in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2003. 399-410.

[10] I. Jolliffe, Principal Component Analysis. Hoboken, NJ, USA: Wiley, 2005.

[11] 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.

[12] H.Hoffmann,”Kernel PCA for novelty detection, “Pattern Recognit., vol 40,no. 3,pp.863-874,2007.

[13] T.-J. Chin and D. Suter, “Incremental kernel principal component analysis,” IEEE Trans. Image Process. vol. 16, no. 6, pp. 1662–1674,Jun. 2007.

[14] B. Liu, Y. Xiao, P. Yu, Z. Hao, and L. Cao, “An efficient approach for outlier detection with imperfect data labels,” IEEE Trans.Knowl. Data Eng., vol. 26, no. 7, pp. 1602–1616, 2014.

[15] I. B. Khediri, M. Limam, and C. Weihs, “Variable window adaptive kernel principal component analysis for nonlinear nonstationary process monitoring,” Comput. Ind. Eng., vol. 61, no. 3, pp. 437–446, 2011.

[16] G. Sharma, S. Chaudhury, and J. Srivastava, “Bag-of-features kernel eigen spaces for classification,” in Proc. 19th Int. Conf. on Pattern Recognit., Dec. 2008, pp. 1–4.

[17] Colin O’Reilly, Member, IEEE , Alexander Gluhak, and Muhammad Ali Imran, Senior Member, IEEE Adaptive Anomaly Detection with Kernel Eigenspace Splitting and Merging IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 1, 2015.