Research on Algorithm of Dependability Oriented Anomaly Detection of Virtual Machines under Cloud

Hongli Li

Department of Electronic Information Engineering, Tianjin Vocational Institute, Tianjin, 300410, China

email: lee69123@163.com

Keywords: Cloud Platforms; Anomaly Detection of VMs; Kernel Metod; Principal component analysis (PCA); Feature Extraction

Abstract. In this paper, a large-scale cloud platform Virtual machine anomaly detection key technologies. For cloud environments systematic study of the feature extraction technique is proposed based on principal component analysis (PCA) for feature extraction algorithm. The algorithm selects the most efficient or concentrated extract from the original performance, data analysis most useful "features", the first analysis of the anomaly detection problem to be solved.

Introduction

Cloud computing resources on demand, flexible scalable, service-oriented, high cost and other advantages, has become the mainstream of computing and service models. However, with the rapid development of cloud computing, the size and complexity of the cloud platform is growing, along with its frequent outbreaks of accidents, serious impact on the reliability and availability of cloud platforms and reduces its credibility.

In order to enhance the credibility of the cloud platform, real-time acquisition and virtual machine health related performance indicators through data preprocessing, feature extraction, and a series of abnormality detection processing, real-time detection of the virtual machine is abnormal, and by the subsequent abnormal localization and troubleshooting to locate and troubleshoot. Accordingly, the abnormality detecting virtual machine is an important foundation for cloud security platform credibility.

From the performance data effectively detect and locate abnormal its roots, it is not as simple as we expected. As the size and complexity of the continued growth of cloud platforms, automatic identification of abnormal demand continued to grow [1].

In this paper, unsupervised feature extraction algorithm. Detailed analysis and the principles derived based on principal component analysis (PCA) for feature extraction algorithm, and points out the principles and inadequate algorithms, and propose a new feature extraction algorithm. Linear feature extraction algorithm assumes data having a generally linear configuration, so there are some limitations. However, due to the complexity of nonlinear methods are often relatively high, this article is not an in-depth study.

PCA for Feature Extraction Algorithm

PCA is a classic unsupervised feature extraction algorithm, which uses orthogonal transform a group of related variables observed sample is converted into a set of linear independent component (called PCA). PCA was first proposed by Karl Pearson proposed in 1901 [2], and thereafter on PCA as a method widely used in classical statistical data analysis, data dimensionality reduction [3] [4] [5] [6] [7] [8], data compression, and other fields. Document [9] describes the use of user-friendly way of thinking PCA method. This section analyzes the principles of the PCA method, and pointed out its shortcomings.

PCA method Import

Generally, in order to better understand the state of the system, we hope to collect as many performance indicators. However, some performance indicators may be relevant (redundant), while other performance indicators may be present in the system noise.

Suppose the sample matrix $X_{n\times l}$ has zero mean, you can use the following formula to calculate the performance indexes n covariance matrix between C_X ($X_{n\times l}$ if not zero mean, you need to calculate the mean value $\overline{X_i}$ for each performance index X_i; then sample matrix $X_{n\times l}$ each sample value X_{ij}, minus the corresponding mean $\overline{X_i}$):

$$C_X = \frac{1}{l-1} XX^T \tag{1}$$

Find another formula C_X covariance matrix of [10] is (within-class scatter matrix covariance C_X is all samples as a class):

$$C_X = \frac{1}{l-1}\sum_{i=1}^{l}(x_i - \overline{x})(x_i - \overline{x})^T \tag{2}$$

Wherein $\overline{x}$ is the random vector X mean of each sample x_i ($\overline{X_i}$ different from previously described), which is calculated as:

$$\overline{x} = \frac{1}{l}\sum_{i=1}^{l} x_i \tag{3}$$

C_X specific form as:

$$C_X = \begin{bmatrix} \sigma^2_{x_1x_1} & \sigma^2_{x_1x_2} & \cdots & \sigma^2_{x_1x_n} \\ \sigma^2_{x_2x_1} & \sigma^2_{x_2x_2} & \cdots & \sigma^2_{x_2x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_{x_nx_1} & \sigma^2_{x_nx_2} & \cdots & \sigma^2_{x_nx_n} \end{bmatrix} \tag{4}$$

Covariance matrix C_X has the following properties:

①C_X is a symmetric matrix of order $n\times n$;

②$\sigma^2_{x_ix_i}$ of C_X diagonal elements is the variance of the performance indicators X_i;

③$\sigma^2_{x_ix_j}$ of C_X non-diagonal elements is the covariance between the performance of X_i and X_j.

Covariance matrix contains C_X covariance between all performance indicators that measure the correlation between the two performance indicators. These covariance reflects the degree of redundancy and noise of the observational data [9]:

①C_X on the diagonal, the greater its value the corresponding performance indicators more important, it means that its value is smaller secondary performance indicators or may be the presence of noise.

②C_X non-diagonal elements, its value indicates the degree of redundancy (linear correlation) between the size of the corresponding performance indicators right.

Set in the original sample is $X_{n\times l}$ simple orthonormal (orthogonal these groups constitute a matrix of order n) under represented. PCA question to be answered is: Is there another orthonormal basis, which is a linear combination of simple standard orthogonal basis, and can best represent the sample set?

$Y_{s\times l}$ new set of samples obtained after transformation. The so-called "best represent" meaning redundant after transform characteristic between minimizing the covariance matrix corresponding to

the C_Y that diagonal elements as small as possible; at the same time to maximize the signal that corresponds to such covariance on the diagonal matrix C_Y is as large as possible, according to descending order.

Assumptions Made by the PCA Method

Four above ideas actually contained the PCA method assumptions made [9]:

①Linear Hypothesis: PCA made a very tough but very efficient linear assumptions, the group is looking for simple linear combination of orthonormal basis, this limitation makes the problem is greatly simplified.

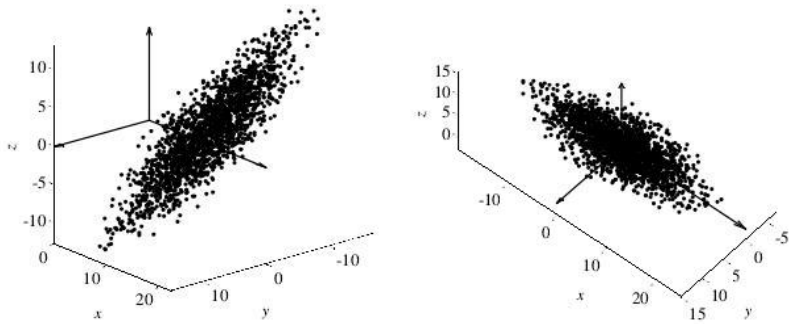
②The mean and variance are sufficient statistics: mathematical form sufficient statistic captures the following description of the concept, a complete description of the mean and variance of the probability distribution. The only class of probability entirely by the first two moments (mean and variance) distribution is described by exponential family distribution (Gaussian, exponential distribution, etc.). To address this hypothesis, performance indicators X_i, must obey the exponential family distribution. If you do not obey, then this assumption is not valid. On the other hand, this assumption also formally guarantee the covariance matrix of the signal to noise ratio (SNR) completely characterizes the redundancy and noise.

③The direction of the largest variance contains our most interesting dynamic: This assumption also implies greater variance data with a higher signal to noise ratio. Therefore, having a large variance PCA representatives meaningful dynamic, and has a smaller variance principal component may be secondary to performance or noise.

④ Primary is orthogonal: PCA methods between assumptions are orthogonal principal component vectors. Real usefulness of this assumption is valid so that the problem of analytical solution. In addition, P is defined by a set of orthonormal basis composition, then P is an orthogonal matrix. P conversion is actually the original n-dimensional coordinate system is a rigid rotation (holding each coordinate relative position and relative orientation between the same axis).

Geometric Interpretation PCA Method

The following illustrates the principles of the PCA method to three dimensions. A set of data points in three-dimensional Gaussian distribution in three-dimensional space is approximately an ellipsoid body, as shown in Figure 1 (a) below. Assuming ellipsoid with three axes a, b, c are the length 20, 8, 3. When using PCA feature extraction method, first find a group based on the assumption 3 vector whose direction is parallel to the long axis of a direction (length of the shaft in a way to characterize the data in the direction of the dynamic, the longer axis data variance in the direction of the greater), this is the first principal component direction. According to the assumption 4, PCA defines the next to find a vector perpendicular to the base before all the basis vectors have been found, so the data dynamic PCA to find the second largest in the direction of the axis in a plane perpendicular to the direction of major axis b. Finally, find the direction of the long axis c. These three directions basis vectors constitute the new coordinate system. PCA raw data rigid rotation (PCA method based on assumptions 1 and 4) to the new coordinate system, as shown in Figure 1 (b) below.



(a) raw data (b) PCA data after conversion

Fig.1. Geometrical explanation of PCA

How to Solve Orthogonal Transform Matrix P

The question now is how to solve the orthogonal transform matrix P, and answer why P is a matrix consisting of eigenvectors C_X thereof.

Since the data matrix $X_{n\times l}$ is zero mean, the new data matrix $Y_{s\times l}$ obtained after orthogonal transformation is a zero mean (because the orthogonal transformation is a rigid transformation, does not change the relative position between the data), it is possible to press $Y_{s\times l}$ calculated covariance matrix C_Y:

$$C_Y = \frac{1}{l-1}YY^T = \frac{1}{l-1}(PX)(PX)^T = \frac{1}{l-1}PXX^TP^T = P\left(\frac{i}{l-1}XX^T\right)P^T \tag{5}$$

$$C_Y = PC_XP^T \tag{6}$$

Since P is an orthogonal matrix, so there is: $P^{-1} = P^T$. Suppose p_i is the column vector P^{-1} and P^T (p_i^T is the row vector of P). There are:

$$P^T = P^{-1} = [p_1 \quad p_2 \quad \cdots \quad p_s], \qquad P = \begin{bmatrix} p_1^T \\ p_2^T \\ \vdots \\ p_s^T \end{bmatrix} \tag{7}$$

We expect C_Y has the following form:

$$C_Y = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_s \end{bmatrix}, \quad \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_s \tag{8}$$

The (7), (8) into (6), there are:

$$P^{-1}C_Y = C_XP^T \tag{9}$$

$$[p_1 \quad p_2 \quad \cdots \quad p_s]\begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_s \end{bmatrix} = C_X[p_1 \quad p_2 \quad \cdots \quad p_s] \tag{10}$$

There are:

$$\lambda_i p_i = C_X p_i, \quad \mathrm{i} = 1,\ 2,\ \cdots,\ \mathrm{s} \tag{11}$$

Therefore, p_i is the eigenvector of C_X (p_i is the column vector of the row vector

corresponding p_i^T), i.e. $p_i = v_i$, λ_i is the corresponding eigenvalues. Thus, solving the eigenvector p_i of C_X , and p_i^T organized into orthogonal transform matrix P (in descending order according to the corresponding feature value arrangement). After new data matrix $Y_{s\times l}$ transformation $Y = PX$ obtained $Y_{s\times l}$, its covariance matrix for the elements on diagonal, diagonal C_X eigenvalues are arranged in descending order.

PCA-based Feature Extraction Algorithm

Algorithm PCA-based feature extraction algorithm

Input: l n-dimensional samples $x_1, x_2, \cdots, x_l$, ie n is the number of performance indicators, l for the total number of samples; Ae contribution rate is h.

Output: $P_{s\times n}$ is orthogonal transform matrix, $Y_{s\times l}$ is feature extraction data matrix after, s is the number of feature-extracted.

Step 1: The sample data is organized into a matrix $X_{n\times l}$ on $n \times l$ order.

Step 2: The sample value for each performance indicator, minus the average of the performance index, the data matrix obtained after zero mean, denoted $X_{n\times l}$.

Step 3: Calculate $X_{n\times l}$ under (1) or (2) covariance matrix C_X .

Step 4: Calculate C_X eigenvalues and corresponding eigenvectors, eigenvalues remember $\lambda_1, \lambda_2, \cdots, \lambda_n$ (in descending order), the corresponding feature vector $v_1, v_2, \cdots, v_n$.

Step 5: The row vector $v_1^T, v_2^T, \cdots, v_n^T$ orthogonal transform matrix composed $P_{n\times n}$, $Y = PX$ transform matrix $Y_{n\times l}$ and new data sets; or column vectors $v_1, v_2, \cdots, v_n$ orthogonal transform matrix composed $P_{n\times n}$, $Y = P^T X$ transformed new data matrix obtained $Y_{n\times l}$.

If the feature extraction Shihai hoping to reduce the dimension of the original data matrix, it should be replaced by the following Step 5 Step 5'.

Step 5': Suppose the first k principal component of the contribution rate and meet thresholds h (general admission 0.85,0.9 or 0.95), namely $\sum_{i=1}^{s} \lambda_i \Big/ \sum_{i=1}^{n} \lambda_i \geq h$, then take the first s eigenvectors constitute transform matrix $P_{s\times n}$, for $X_{n\times l}$ transformed, on obtain data matrix $Y_{s\times l}$ dimensionality reduction.

Conclusion

PCA is the main starting point is to remove the correlation data set (second order dependency); PCA method is simple, can be obtained analytical solution, assuming it four from these advantages made; but also because PCA methods were too many too strong assumptions, such PCA method has many limitations, such as PCA assumes that the original performance indicators exponential family distribution (Gaussian), if you do not obey, the PCA method will lose more information. The introduction of nuclear methods (KPCA [11]) can remove datasets higher order dependencies in the PCA. Another direction of extension of the PCA method is to focus on a more general definition of statistical dependencies in the data, such as requiring data collection in all directions after the dimensionality reduction is statistically independent [9], this extension leads ICA (Independent Component Analysis law).

References

[1] Lan Z L, Zheng Z M, and Li Y W. Toward automated anomaly identification in large-scale systems [J]. IEEE Transactions on Parallel and Distributed Systems, 2010, 21(2): 174-187.

[2] Pearson K. On Lines and Planes of Closest Fit to Systems of Points in Space [J]. Philosophical Magazine, 1901, 2(11): 559-572.

[3] Smith D, Guan Q, and Fu S. An anomaly detection framework for autonomic management of compute cloud systems [C]. Proceedings of 34th Annual IEEE International Computer Software and Applications Conference Workshops, 2010: 376-381.

[4] Guan Q and Fu S. Adaptive anomaly identification by exploring metric subspace in cloud computing infrastructures [C]. Proceedings of the 32nd IEEE International Symposium on Reliable Distributed Systems (SRDS), 2013: 205-214.

[5] Fu S. Performance Metric Selection for Autonomic Anomaly Detection on Cloud Computing Systems [C]. Proceedings of 54th Annual IEEE Global Telecommunications Conference (GLOBECOM), 2011.

[6] Dong S D. Linux kernel-based virtual machine resource-oriented service Exception Monitoring System [D] Chongqing: Chongqing University, 2011.

[7] Ren T. For IaaS virtual machine Anomaly Detection System [D] Chongqing: Chongqing University, 2014.

[8] Pechenizkiy M, Puuronen S, and Tsymbal A. The Impact of Sample Reduction on PCA-based Feature Extraction for Supervised Learning [C]. Proceedings of the 21st Annual ACM Symposium on Applied Computing (SAC), 2006: 553-558.

[9] Shlens J. A Tutorial on Principal Component Analysis [R], April 22, 2009; Version 3.01. http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf.

[10] Duda R O, Hart P E, and Stork D G. Li H D, Yao T X translation. Pattern classification (original book the 2nd edition) [M]. Beijing: Mechanical Industry Press, 2003.

[11] Schölkopf B, Smola A, and Muller K R. Nonlinear component analysis as a kernel Eigenvalue problem [J]. Neural Computation, 1998, 10(5): 1299-1319.