Advances in Intelligent Systems Research, volume 132
2nd International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2017)

An Anomaly Detection Method for Stateful Stream Processing System

Guanghui Chang, Lu Zhao*, Jun Liu and Peizhen Li
School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
*Corresponding author

***Abstract*—The SSPS (stateful stream processing system) is a distributed complex event processing system implemented by the framework of stream processing system. However, due to the nature of uncertainty, randomness and burstiness of the data stream and the complexity of complex event processing, the SSPS faces great challenges. In order to address the problem, this paper analyzes the problems of SSPS, on this basis, an anomaly detection method based on FKPCA-SVM is proposed. Experimental results show that the proposed method can efficiently and reliably detect SSPS, and ensure the system can operate normally.**

Keywords- stream processing; complex event processing; kpca; svm

I. INTRODUCTION

In recent years, with the rapid development of emerging information technology and application mode such as Cloud computing, Internet of Things and Mobile Internet, the amount of data sharply increases, thus which promotes the human society into the era of big data. When the data increase abruptly, the technical problems of high throughput, low latency and complex computing cannot be solved by traditional centralized CEP (complex event processing). In this context, distributed CEP technology has been widely studied [1]. A high performance distributed CEP method based on probability is proposed in reference [2], this method is more effective than SASE, but the query scheme is not flexible and difficult to use. A distributed algorithm based on query planning is proposed to form distributed CEP system in reference [3], but this method is effective only when the sliding window is large. Recently, the academia proposes that the distributed CEP system is implemented by using stream computing. In reference [4], a distributed CEP framework is designed and implemented by Storm, which improves greatly the processing ability of CEP system. In this paper, a distributed CEP system based on stream computing technology is called SSPS.

However, due to the nature of uncertainty, randomness and burstiness of the data stream and the complexity of complex event processing, the SSPS arisen easily abnormity. So it is a basic and important work to monitor the SSPS [5]. At present, many techniques are applied to anomaly detection, such as Bayesian detection model, artificial neural network and machine learning model [6]. However, the training set for abnormal monitoring appears the characteristics of nonlinear and high dimension, which makes that traditional classifiers, such as Bias classifier, neural network classifier and linear discriminant classifier, are difficult to achieve desired results. SVM (Support vector machine) is based on statistical learning theory, which has good generalization ability in small sample, high dimension and nonlinear data space [7]. However, SVM has many shortcomings, for example, when there is a large difference between sample attributes, a large number of support vectors will be generated to cost a lot of training time. On the other hand, the SVM classification results are also affected by the high dimension of sample.

Therefore, an anomaly detection method based on FKPCA-SVM is proposed in this paper. In order to reduce the dimension of feature vector and shorten the training time of SVM, the input data of SVM can be preprocessed by KPCA, Furthermore, SVM classifier will be updated by feedback mechanism to improve the classification accuracy.

II. THE ANALYSIS OF SSPS

Definition 1: A distributed CEP system based on stream computing technology is called SSPS.

A distributed CEP system that is implemented by Storm is shown in Figure1. In terms of the composition of the system, SSPS is analyzed form Storm, CEP and performance index.

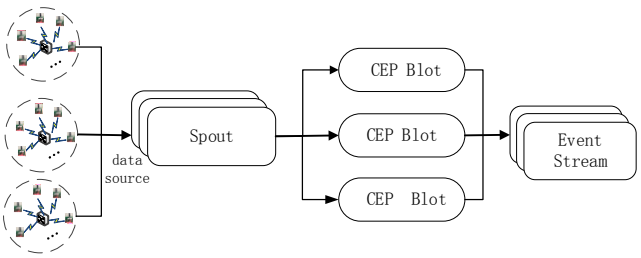

FIGURE I. THE FRAMEWORK OF SSPS

A. The Analysis of Storm

Storm is a distributed real-time computing system. Due to the nature of uncertainty, randomness and burstiness of the data stream, Storm faces unprecedented challenges in load balancing, data throughput and etc.

The load balancing mechanism of Storm is not perfect, but the system load balancing mechanism is a key factor to restrict the system stable operation, high throughput calculation and fast response. Although Storm provides a default scheduler, this scheduling strategy can ensure that the number of Worker on multiple work nodes is the same, but it does not take into

account of the differences between Topology and the difference between Spout or Bolt components in Storm, their computing resources, storage resources, cyber source are not the same.

Storm system has some limitations in throughput. Although the throughput can be improved by increasing the number of worker, the number of worker and clusters are related, so the number of worker are limited. With the increase of data stream quantity, the system faces great challenges in throughput.

B. The Analysis of CEP

Event detection is the core of CEP, which is the process of finding all the basic events that can constitute complex events. At present, the main event detection models of CEP system include NFA detection model, FSM detection model, Petri network detection model, and detection model based on tree or graph [8]. Through investigation and analysis, CEP generally uses NFA detection model, as shown in Table 1.

TABLE I. COMPARATIVE ANALYSIS OF EVENT DETECTION MODEL

System	Event detection model				
	NFA	FSM	Graph	Tree	Petri
SASE	✓				
Cayuga	✓				
ZStream				✓	
Esper	✓				
TPN					✓
RCEDA			✓		

When the event streams that meet the condition of the query arrive, the state of the event streams will be migrated. And only when all the event streams that meet the condition of the query arrive, the event streams can reach the final state to form a more abstract event, as shown in Figure 2.

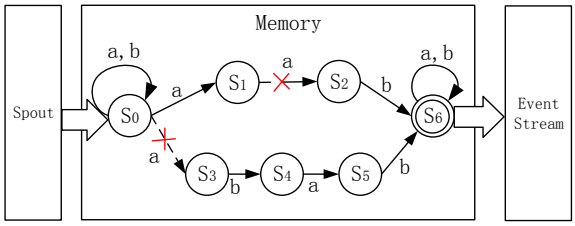

FIGURE II. NFA DETECTION MODEL

Spout is a data source of SSPS, which forms a large number of tuples. These tuples are generally sorted by timestamp. Due to the characteristics of data stream, this time relationship becomes more complex, and this temporal relationship affects the correctness and efficiency of event detection to some extent. Further, in the process of forming complex events, data stream is always processed in memory, there will be a lot of tuples into memory, meanwhile, and the intermediate state of event will be stored in memory. In a larger time window, memory consumption will greatly increase, which increases greatly the probability of system failure.

C. The Performance Index of SSPS

Performance index is the key factor to measure system running status. In order to carry out a comprehensive monitoring for SSPS, a number of monitoring indexes are selected in this paper, as shown in Table 2.

TABLE II. THE PERFORMANCE INDEXS OF SSPS

Monitoring items	Description
Memory/available bytes	Available physical memory bytes
Pages/sec	The sum of the read/written page files per second
% Disk time	Disk usage rate
I/O read bytes/sec	The speed of reading bytes from the I/O
%Processor time	CPU occupancy rate
Processor /packets/sec	The rate of sending and receiving packets
Throughput	The sum of the amount of data transmitted over the network during a single test.
Detection delay	The time required to form a complex event

Throughput is one of the most important indexes to measure the SSPS. Assuming that all atomic events are sent to a single source SSPS, the processing speed of data stream is the most important manifestation of SSPS performance. Therefore, throughput is an index that must be collected in SSPS. Data stream is always processed in memory, there will be a lot of tuples into memory, and meanwhile, in the process of forming complex event, the intermediate state of event also will be stored in memory. If some measures are not taken, memory will grow rapidly, so memory usage can reflect whether the system is abnormal. Detection delay is the total time overhead that a pattern is detected, that is the time overhead from sending the first event to generating the complex event, and this index is the key factor to reflect whether the system has low delay characteristic. In term of SSPS, CPU is the important supporting element of hardware. In the process of data stream processing, a lot of computing is needed, which consumes a lot of CPU.

III. ANOMALY DETECTION METHOD BASED ON FKPCA-SVM

The main idea of this method is to preprocess the input data of SVM and standardize the main features of input data by KPCA, to reduce the dimension of feature vector and shorten the training time of SVM. Furthermore, SVM classifier will be updated by feedback mechanism to improve the classification accuracy. The overall architecture of the method is shown in Figure 3.

Input: The original training set X and test set Y, which comes from the data acquisition module.

Output: Detection accuracy Z and detection results R.

Step1: Data preprocessing. The training set X and test set Y was standardized by Zscore method, to eliminate the influence of different dimension on the prediction results, then the training set X_1 and test set Y_1 will be obtained.

Step2: The training set X_1 and test set Y_1 are processed by using KPCA algorithm. The number of principal components that contribution rate belta is higher than 0.9 are obtained, then the dimension reduction training set X_2 and test set Y_2 are achieved by KPCA.

Step3: The training set X_2 and test set Y_2 are trained by using SVM algorithm, to get an optimal classification plane P that is used to detect the test set Y_2, then the detection results R and detection accuracy Z are obtained.

Step4: The SSPS is detected in real time. If the detection results R is abnormal, then the corresponding abnormal sample N is stored in the update module.

Step5: Classifier update. When the detection accuracy Z of the SVM classifier is reduced, the abnormal sample N in the update module and the training set X are trained again to optimize SVM classifier, then the classification plane P will be replaced by P_1 that is the optimized classification plane.

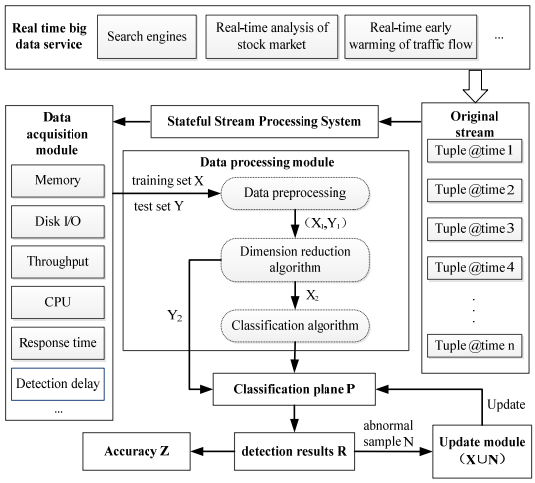

FIGURE III. OVERALL FRAMEWORK FOR ANOMALY DETECTION

IV. EXPERIMENTS AND ANALYSIS

A. Data Acquisition and Preprocessing

In this paper, the experimental platform is set up by using sensor, Storm and CEP, as shown in Figure 4.

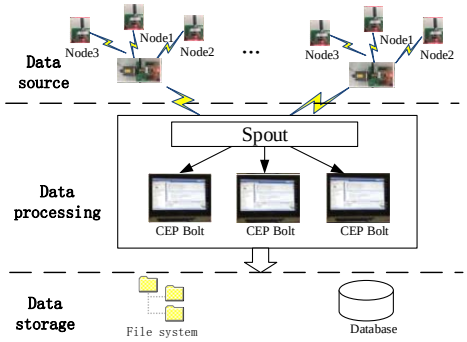

FIGURE IV. EXPERIMENTAL PLATFORM

A sensor node is a data source, these data are forwarded to SSPS by the coordinator. The worker node in SSPS is monitored by Perfmon that is a performance monitoring tool of Windows, such as CPU, memory, throughput, disk and other attributes.

914 samples are selected randomly in this paper. The normal samples accounted for 70%, a variety of abnormal samples accounted for 30% of the total. 600 samples are treated as training samples, the rest are treated as test samples.

B. The Processing and Analysis of KPCA

In this paper, in order to reduce the dimension of feature vector and shorten the training time of SVM, the input data of SVM is preprocessed and standardized by KPCA. The partial dataset of dimension reduction are shown in Table 3.

TABLE III. THE PERFORMANCE INDEXS OF SSPS

Dataset	Principal Component				
	f_1	f_2	f_3	f_4	f_5
P_1	797600	923840	10340	10399	134.5
P_2	-965400	-658660	-7296.7	-6031.9	252.22
P_3	-518180	-256060	-10808	-8651.7	827.21
P_4	621970	767420	894.04	2264.9	-89.766
P_5	586160	735380	-206.55	1298.9	-127.28
P_6	123860	320820	-7985.3	-5708.7	-301.39
P_7	-338440	-94561	-10051	-7855.1	25.524

The cumulative contribution rate of principal component can reach 80%, which shows that the principal components can reflect the original dataset. In this paper, we selected five principal components, the cumulative contribution rate reaches 90%. Compared with the original dataset, the dimensionality of the dataset is reduced by 76.1%, which reduces the complexity of SVM classifier.

C. The Analysis of Kernel Function of SVM

The core of SVM algorithm lies in the kernel function. In this paper, the SVM algorithm is simulated by using the LibSVM toolbox. The time overhead and accuracy of four different kernel functions are shown in Figure 5.

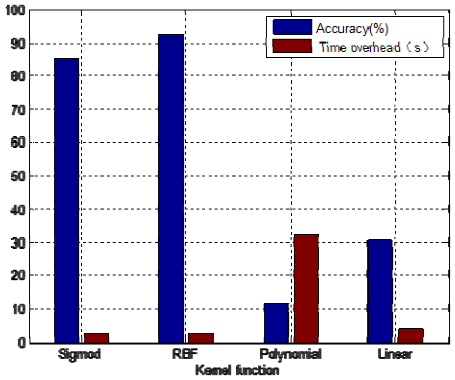

FIGURE V. KERNEL FUNCTION ANALYSIS OF SVM

The classification accuracy of Sigmod kernel function is 87%, the time overhead of Sigmod kernel function is 2.73s; the classification accuracy of RBF kernel function even reaches 98%, meanwhile, the time overhead of RBF kernel function is almost the same when compared with Sigmod; the classification accuracy of polynomial kernel function is only 11.8%, and the time overhead of polynomial kernel function is up to 32.5s; the classification accuracy of linear kernel function is general 30.4%, meanwhile, the linear kernel function does

not exist advantage in time overhead. Experimental results show that RBF kernel function is most suitable for this system.

D. Method Comparative Analysis

In order to prove the superiority of this method, on the one hand, the SVM and PCA-SVM algorithms are compared with KPCA-SVM according to F-Measure, as shown in Figure 6. F-Measure is the weighted harmonic mean of precision P and recall R.

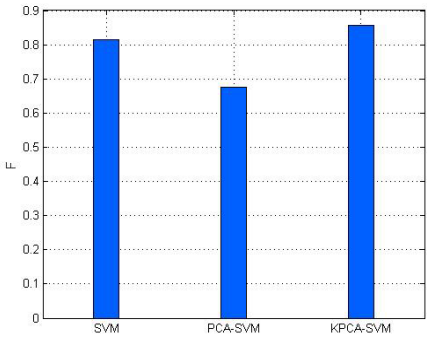

FIGURE VI. THE CONTRAST OF F-MEASURE

The value of F of KPCA-SVM is the highest. Obviously, compared with other, the detection efficiency of KPCA-SVM is better. On the other hand, the SVM and PCA-SVM algorithms are compared with KPCA-SVM in this accuracy and time overhead, as shown in Table 4.

TABLE IV. COMPARISON OF THREE METHODS

Sample category	Detection item	SVM	PCA-SVM	KPCA-SVM
Training sample	Accuracy	93.12%	81.00%	95.62%
	Training time	4.32s	2.41s	2.73s
Test sample	Accuracy	86.65%	72.71%	90.56%
	Testing time	3.16ms	1.32ms	2.36ms

1) The accuracy of SVM is better than PCA-SVM, therefore, in term of the nonlinear samples, the PCA-SVM cannot get good feature extraction and classification results.

2) By comparison of SVM and KPCA-SVM, the performance of KPCA-SVM is better than the former.

3) By comparison of PCA-SVM and KPCA-SVM, the time overhead of KPCA-SVM is more than PCA-SVM, but the accuracy is improved.

Thus, the dataset shows strong nonlinear nature, PCA cannot get good results through linear feature extraction, which will affect the classification results of SVM; but because SVM contains nonlinear mapping ability, SVM can get good classification results; the dataset also shows a strong high dimension feature, aggravated the burden of SVM classifier. KPCA has a very strong nonlinear feature extraction ability, so the performance of anomaly detection will be improved by combining KPCA with SVM.

V. CONCLUSIONS

This paper studies and analyzes SSPS. On this basis, an anomaly monitoring method based on FKPCA-SVM is proposed in this paper. The experimental results show that KPCA can effectively reduce the dimension of dataset, and improve the detection rate of SVM. Furthermore, the accuracy of anomaly detection is improved by the feedback mechanism. In the future, we will improve and optimize further the anomaly monitoring method.

ACKNOWLEDGMENT

The authors acknowledge support from the Program for Innovation Team Building at Institutions of Higher Education in Chongqing (Grant No. CXTDG201602010); the China Postdoctoral Fund (Grant No. 2014M562282); the Project Postdoctoral Supported in Chongqing (Grant No. Xm2014039); the Wenfeng Leading Top Talent Project in CQUPT, the New Research Area Development Programme (Grant No. A2015-44); the Science and Technology Research Project of Chongqing Municipal Education Committee (Grant No. KJ1400422, KJ1500441 and KJ1400431); the Common Key Technology Innovation of Important Industry by Chongqing Science and Technology Commission (Grant No. CSTC2015ZDCY-ZTZX40001); the Collaborative Innovation Center for Information Communication Technology (Grant No. 002); the Social Livelihood Science and Technology Innovation Special Projects of Chongqing (Grant No. cstc2016shmszx40001); the University Outstanding Achievements Transformation Funding Project of Chongqing (Grant No. KJZH17116).

REFERENCES

[1] Cugola G, Margara A. Processing flows of information: From data stream to complex event processing[J]. ACM Computing Surveys (CSUR), 2012, 44(3): 15.

[2] Wang Y H, Cao K, Zhang X M. Complex event processing over distributed probabilistic event streams [J]. Computers & Mathematics with Applications, 2013, 66(10): 1808-1821.

[3] Wang Y, Yang S. Plan Based Distributed Complex Event Processing for RFID Application [J]. International Conference on Computational Intelligence & Software Engineering, 2009:1-4.

[4] Shang Y M. The Research of key technologies on distributed complex event processing [D]. North China Electric Power University, 2015.

[5] Flouris, I., Giatrakos, N., Garofalakis, M., & Deligiannakis, A. (2015, August). Issues in Complex Event Processing Systems. In Trustcom/BigDataSE/ISPA, 2015 IEEE (Vol. 2, pp. 241-246). IEEE.

[6] Liu C, Ghosal S, Jiang Z, et al. An unsupervised spatiotemporal graphical modeling approach to anomaly detection in distributed CPS[C]//2016 ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS). IEEE, 2016: 1-10.

[7] Feshki M G, Sojoodi O, Anvary M D. Managing Intrusion Detection Alerts Using Support Vector Machines[J]. no, 2015, 9: 266-273.

[8] Jastrzab T, Czech Z J, Wieczorek W. Parallel induction of nondeterministic finite automata[C]//International Conference on Parallel Processing and Applied Mathematics. Springer International Publishing, 2015: 248-257.