Survey Paper on Fault Detection in Sensor Data

Vidya D. Omase[1], Jyoti N. Nandimath[2]

[1]Dept. of Computer Engineering, Smt. Kashibai Navale College of Engineering, Vadgaon Bk, Pune, India

[2]Assistant Professor, Dept. of Computer Engineering, Smt. Kashibai Navale College of Engineering, Vadgaon Bk, Pune, India

Abstract: ***Process of data classification of data suffers with the increasing dimensionality of data. Fault Detection becomes important and critical in several industries. For organizations, it is essential to continuously improve the productivity. In semiconductor manufacturing it is crucial to detect faults at initial stages. So, quick identification of abnormal results is primary objective. Data classification possesses some issues because of unbounded size of data and imbalance nature of the data. Data imbalance means the number of instances in one class greatly outnumbers the number of instances in the other class. In classification, the available standard algorithms tend to favor the majority class and produce low detection of minority class as a result when the class sizes are highly imbalanced. This results in inaccurate classifier generation and wrong prediction of data. In literature many fault detection algorithms are available to address these issues. An online fault detection algorithm based on incremental clustering performs well and efficiently process the data.***

Keywords: Data mining, Classification, Clustering, Class imbalance data

1. Introduction

In recent years Data Mining is of utmost interest. Mining huge available data using different techniques such as classification of data and discovering knowledge from the data has great significance. There are many algorithms available for data classification. In last few years there are major changes done on classification of data to address different issues. With the increasing size of data, classification of data becomes difficult because of unbounded size and imbalance nature of data. Data imbalance means the number of instances in one class greatly outnumbers the number of instances in the other class. A dataset is said to be highly skewed if sample from one class is in higher number than the other.

From the perspective of data mining, fault detection problem involves learning a binary classifier that provide two class labels i.e. normal and fault. A dataset is said to be imbalanced if classes are not equally represented. The most of algorithm are more focusing on classification of normal sample while ignoring or misclassifying fault sample which prevents providing generalized knowledge over the entire fault data space. Machine learning using such data sets is an issue that should be investigated and addressed. The classifications of algorithms are either parametric or non-parametric. Parametric models assume an underlying functional form of the classifier and have some fit parameters. Non-parametric models have no explicit assumption about the form of the classifier.

In paper [6] case of semiconductor data is considered and proposed an online fault detection algorithm based on incremental clustering. The algorithm finds wafer faults in class distribution skews and process sensor data in terms of reductions in the required stages with accuracy and efficiency. The algorithm clusters normal data to reduce the storage and requirements of computation. To detect potential faulty wafers statistical summaries are maintained for each cluster. The Mahalanobis distance which is statistical distance measure that considers correlations and differences among the data points used to predict the class label of new wafer in multidimensional feature space. Algorithm proposed in [6] is highly advantageous when performing fault detection in stream data environments with imbalanced data and even under process drifts. However, when there is very high dimensional data present, computation cost and storage requirement rises.

The paper is organized in the following manner. In the remaining of this paper we have studied literature survey. In section 3 concluded the survey paper after that the references used for the paper are describes.

2. Literature Survey

2.1 Data mining

Daily transactions, operations, process generates huge data at many organization. This large data contains knowledge which can be used for improving the business. Data mining is used for extracting knowledge from large data. There are many concepts in data mining such as Classification, Clustering, Prediction, Regression etc.

2.2 Classification

Classification is widely used data mining technique in which unlabeled data is labeled. For example patient is diabetes positive or negative can be labeled through the classification technique. There are two phases in classification, one is training or learning and another is testing process. In training knowledge is extracted from labeled data and in testing this extracted knowledge is used for labeling unlabeled data. Classifications techniques are differing by way of extraction of knowledge and representation of the knowledge.

2.3 Clustering

In Clustering, similar records or data are grouped in same cluster and records in one cluster should differ from records in another cluster**.** Clustering unsupervised learning in which there is no need of labeled data. In Clustering distance of each record with every other record is calculated using any distance measure. Records with less distance are put into same cluster.

2.4 Class imbalance data

In learning or training phase of the classification technique, if data of one class is very less than another class data then that data is called as Class imbalance data. For example, new disease in found then data of those actual patients which are suffering from that disease is very less than normal patients. In this example data of positive class is less than negative class. Class imbalance data can degrade the performance of the learning classification algorithm.

Due to imbalanced data, classification of data is troublesome. The majority class represents "normal" cases, while the minority class represents "abnormal" cases. This problem exists in many imbalanced two-class classifications. This prevents developing effective classification methods because many traditional algorithms based upon the presumption that training set have sufficient representatives of the class to be predicted. In this paper highly imbalanced two-class classification problems addressed i.e. small fraction of records of minority class than the majority class. Conventional methods tend to strongly favor the majority class, and largely ignore the minority class when dealing with an imbalanced data set. This result in lower or no detection of the minority class when directly applied to an imbalanced data set.

Paper [1] suggests two step systems: First step will have high detection ability with relatively high false alarm rate to produce small data sets with higher concentration of minority class. Second step makes the verification more affordable as pre-screening step narrows down the data samples. The Paper proposed an ensemble-based approach i.e. an Ensemble Classifier for Highly Imbalanced class sizes (ECHI), for highly imbalanced class distributions.

Traditional fault detection methods are based on statistical process monitoring methods and pattern classification based methods. Developed methods are not suitable for semiconductor processes which have unique characteristics like non-linearity and multimodal batch trajectories. This paper proposes [2] diffusion maps based knn fault detection which required less data storage requirement and improves the accuracy of fault detection. Proposed system uses information preservation and feature reduction properties of diffusion map.

This paper [3] considers the case of monitoring semiconductor manufacturing process. Increase in the output and improved product quality is of importance in manufacturing. Quickly detecting abnormalities and diagnosing the problem is main motive of multivariate statistical process control. In such scenario, Principal component analysis (PCA) method is popular to address the issue. But the method has some drawbacks. Paper proposed new substatistical PCA-based method with the application of Support Vector Data Distribution. SVDD is one class classification method for fault detection and the goal is to define boundary around the samples with volume as small as possible which helps to improve performance. Also Correlations between multiway, multi-model, and adaptive submodel methods are discussed in paper.

In data modeling abrupt change is defined as, possibility of variation in the distribution that generate the data, produced in short time. The problem exists in real world applications including time series analysis or some industrial process. One Class Support Vector Machines proves efficient in non-stationary classification problem. One class classifier model describes a single class of object and distinguishes it from all other possible object, also one class SVM assumes that origin in the feature space belong to faulty class hence it aims to maximize the distance between origin and clusters of normal sample in future space. Paper [4] introduced an extension of Time-Adaptive Support Vector Machines (TA-SVM) to one class problems (OC-SVM) which is able to detect abrupt process changes with normal class training data.

In various industries, fault detection is a crucial issue. In semiconductor manufacturing it is necessary to quickly detect abnormal behaviors and consistently improve equipment productivity. For fault detection some statistical methods such as control charts are the most widely used approaches. Due to the number of variables and the possible correlations between them, these control charts need to be multivariate. This paper suggested a non-parametric control charts such as k-nearest neighbor control rule by He and Wang. The method is advantageous because of its ease of understanding and implementation in industrial environment than black box methods such as SVM and neural networks.

The approach [5] is used to evaluate distance of on observation compared to the normal operating region. The cumulative distance of this observation to its k nearest neighbors in the learning sample is calculated. A fault is declared if this distance is too large. The paper proposed new distance, an adaptive Mahalanobis distance for K-nearest neighbor distance (k-NDD) rule based on local covariance structure of the monitored observations. This method achieves fine performance with lesser complexity.

In data mining, fault detection problem involves learning a binary classifier that provide two class labels i.e. normal and fault. A dataset is said to be imbalance if classes are not equally represented. Most of standard algorithms such as Support Vector Machines (SVM) are more focusing on classification of normal sample while ignoring or misclassifying fault sample which prevents providing generalized knowledge over the entire fault data space. Machine learning using such data sets is an issue that should be investigated and addressed.

The Paper [6] proposed an Incremental Clustering Fault Detection Method (IC-FDM) i.e. an online fault detection algorithm based on incremental clustering using Mahalanobis distance which is a statistical distance measure that considers the correlations and differences among the data points. The algorithm provides high accuracy for fault detection even in severe class distribution skews and able to process massive data in terms of reductions in the required storage. Also it is highly advantageous when performing fault detection in stream data environments.

3. Conclusion and Future scope

Many standard fault detection algorithms are available to address the problem of data imbalance in classification of data. In literature, the incremental clustering based algorithm for fault detection is discussed which provides better results. However, existing system focuses on reducing the number of data records; there is a scope to remove number of irrelevant features. Removing irrelevant features [7] i.e. less important variables before applying fault detection incremental clustering based fault detection algorithm improves speed of the process and gives more accurate prediction of data.

References

[1] E. Byon, A. K. Shrivastava, and Y. Ding, "A classification procedure for highly imbalanced class sizes," IIE Trans., vol. 42, no. 4, pp. 288–308,2010.

[2] Y. Li and X. Zhang, "Diffusion maps based k-nearest-neighbor rule technique for semiconductor manufacturing process fault detection," Chemometr. Intell. Lab. Syst., vol. 136, pp. 47–57, Aug. 2014.

[3] Z. Ge and Z. Song, "Semiconductor manufacturing process monitoring based on adaptive substatistical PCA," IEEE Trans. Semicond. Manuf., vol. 23, no. 1, pp. 99–108, Feb. 2010.

[4] G. L. Grinblat, L. C. Uzal, and P. M. Granitto, "Abrupt change detection with one-class time-adaptive support vector machines," Expert Syst. Appl., vol. 40, no. 18, pp. 7242–7249, 2013.

[5] G. Verdier and A. Ferreria, "Adaptive Mahalanobis distance and k-nearest neighbor rule for fault detection in semiconductor manufacturing," IEEE Trans. Semicond. Manuf., vol. 24, no. 1, pp. 59–68,Feb. 2011

[6] JueunKwak, Taehyung Lee, and Chang Ouk Kim, "An Incremental Clustering-Based Fault Detection," IEEE Trans. Semicond. Manuf., vol. 28, no. 3, Aug 2015.

[7] Qinbao Song, Jingjie Ni, and GuangtaoWangFast, "Clustering based Feature Subset Selection algorithm for High-Dimensional data," IEEE Trans. Know. Data Engg.,vol 25,no. 1, Jan 2013.