International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 01 | Jan -2017 www.irjet.net p-ISSN: 2395-0072

Fault Detection Of Imbalanced Data Using Incremental Clustering

Bhagwat Tambe[1], Asma Chougule[2], Sikandar Khandare[3] , Prof. Gargi Joshi[4]

[1] *Department of Information Technology, Dr. D. Y. Patil College Of Engineering, Ambi*
[2] *Department of Information Technology, Dr. D. Y. Patil College Of Engineering, Ambi*
[3] *Department of Information Technology, Dr. D. Y. Patil College Of Engineering, Ambi*
[4] *Assistant Professor, Dept. of IT Engineering, Dr. D. Y. Patil College Of Engineering, Ambi, Pune, Maharashtra, India*

Abstract - *As increase in data dimensionality classification of data increased. In industries or organizations fault detection is important task. Due to imbalanced of data classification process has problem. In standard algorithm of classification majority classes have priority for classification and minority classes have less priority for classification therefore it is not suitable for minority classes fault detection from data is applied for only majority classes and less for minority classes. Incremental clustering algorithm solved this problem but it reduced data attribute. To maximize the accuracy, time, and memory for this we proposed a feature selection algorithm for better performance of classification and fault detection.*

Key Words: **Classification, Class imbalanced data, Clustering, Data mining**

1.INTRODUCTION

Data mining is a largely studied subject of research subject. data mining is a mining of competencies from large amount of information. there are lot of issues exists in massive database such as information redundancy, missing data, invalid knowledge and many others., some of the primary obstacle in data circulate study discipline in dealing with high dimensional dataset. outlier detection is a department of data mining, which refers to the quandary of finding objects in a big dataset that range from different information objects. outlier detection has been used to detect and take away undesirable information objects from big dataset. clustering is the method of grouping a suite of data objects into lessons of similar knowledge objects. the clustering techniques are incredibly precious to discover the outliers so known as cluster based outlier detection. the data movement is a brand new arrival of research subject in knowledge mining. the information stream refers to the process of extracting talents from nonstop rapid developing knowledge records.

Data mining, typically, offers with the invention of non-trivial, hidden and fascinating competencies from exclusive forms of data. with the development of understanding applied sciences, the quantity of databases, as well as their dimension and complexity grow rapidly. it's integral what we'd like automated analysis of excellent amount of knowledge. a data flow is an ordered sequence of objects x1,..,xn. the foremost change between a natural database and a data stream management process (dsms) is that rather of members of the family, we have now unbounded data streams. applications, comparable to fraud detection, community float monitoring, telecommunications, knowledge management, and many others., the place the information arrival is continuous and it's either pointless or impractical to store all incoming objects. 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 a statistical distance measure that considers the 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. To avoid this, we can reduce the number of variables by removing the number of irrelevant features and eliminating the redundancy of features. For this, based on a minimum spanning tree (MST), Fast Clustering based Feature Selection algorithm is used. This aims to achieve better efficiency in time and improved result comprehensibility

2. LITERATURE REVIEW

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. The category imbalance situation has obtained enormous concentration in areas similar to desktop studying and sample attention in recent years. A two-category knowledge set is implicit to be imbalanced when probably the most lessons within the minority one is heavily below represented in contrast to the opposite category within the majority one. This obstacle is on the whole major in actual world applications the place it is steeply-priced to misclassify examples from the minority category, similar to detection of fraudulent mobile phone calls, prognosis of infrequent ailments, understanding retrieval, textual content categorization and filtering tasks [1].

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. From the data-mining perspective, the fault detection concern entails learning a binary classifier that outputs two category labels: average and fault. The classification algorithms are both parametric and non-parametric. Parametric units assume an underlying sensible form of the classifier and have some fit parameters. Non-parametric items have no explicit assumption about the type of the classifier. The aid vector desktop (SVM) is among the most well-known and promising parametric algorithm the SVM finds the isolating hyper-airplane within the function area that can create highest distance between the plane and the closest knowledge of distinctive courses[2].

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 sub-statistical 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 multi-way, 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.

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 [5] 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. THE EXISTING SYSTEM

Approach presented in existing system addresses the problem of data imbalance in classification of data. A data imbalance is the unequal representation of classes' i.e. 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 other. Existing system proposed solving approach: online fault detection algorithm based on incremental clustering. For detecting faults in semiconductor data, class labels of new wafer are detected using Mahalanobis distance method. The Mahalanobis distance is a statistical distance measure that considers the correlations and differences among the data points.The Incremental Clustering-Based fault detection method (IC_FDM) performs following four phases.

1. Phase 0 (Initialization): Algorithm creates a new single member cluster, accepts a new sample and it begins the fault detection task for the single member cluster.

2. Phase 1 (Classification): In this phase class label of the new sample is assigned. Decision of assigning the label is made by calculating the distance between the new sample and center i.e. mean of the closest normal cluster. This distance is calculated using Mahalanobis distance. A threshold is decided using probability distribution of the squared distance. If a distance calculated is less than this threshold then the new sample is said to be normal else it is considered as faulty. But at the early stages of model training, creation of un-matured cluster cannot be avoided when the number of members in the cluster does not exceed the number of features.

3. Phase 2 (Cluster Update and Generation): If the distance calculated is less than threshold, statistical summaries i.e. prototype of the closest cluster is updated with consideration of new sample. Because of this cluster grows incrementally. If the cluster is classified as faulty but the actual class is normal, the algorithm creates a new single member cluster whose center point is the sample.

4. Phase 3 (Cluster Merge): As cluster grows incrementally, computational overhead to find nearest cluster is increases. This phase maintains a small number of clusters by repeating the merge of two adjacent clusters until the merge condition is satisfied.

When the available data is very high dimensional there is increase in storage requirement and cost overhead. As number of variables is large in size, there are possibilities of incorporating features which are irrelevant results in inappropriate results. This motivates the introduction of proposed system.

4. PROPOSED SYSTEM

Proposed system aims at fault detection with consideration of imbalanced nature of data and increasing learning accuracy, improving result quality, removing irrelevant data, reducing dimensionality in efficient way by choosing subset of strongly related features and discarding irrelevant features.

Redundant and irrelevant features affect the speed and accuracy of learning. Feature subset selection achieved by identifying and removing irrelevant and redundant features improves prediction accuracy. To achieve this, based on a minimum spanning tree (MST), Fast Clustering based Feature Selection algorithm is used. Algorithms efficiently and effectively deal with irrelevant features removal and eliminate redundant features. It involves:

1) Select available features from the data set.

2) Relevancy of feature is calculated using mathematical rule and compared with relevancy threshold. If this relevancy is greater than the threshold then the feature is added to feature set.

3) Selected features are divided into clusters by using graph-theoretic clustering methods.

4) Construction of the minimum spanning tree (MST) from a weighted complete graph.

5) Partitioning of the MST into a forest with each tree representing a cluster; and

6) the most representative feature that is strongly related to target classes is selected from each cluster to form final subset of features. Features in different clusters are relatively independent; the clustering based strategy of FAST has a high probability of producing a subset of useful and independent features.

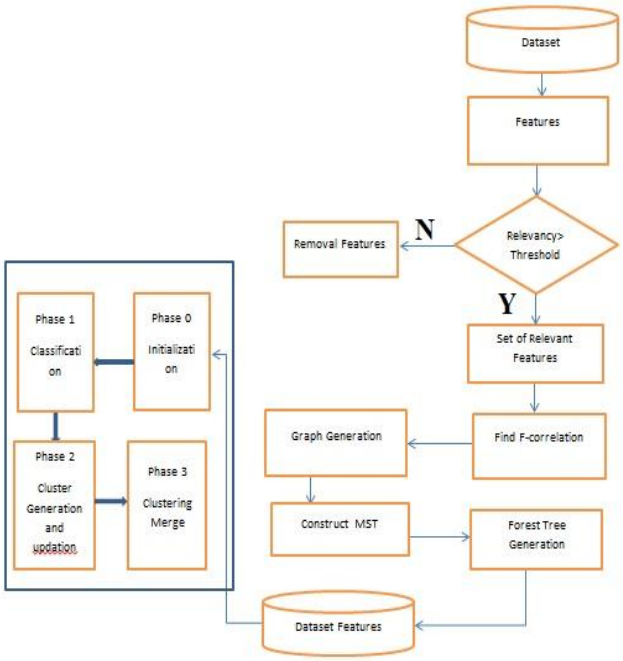


Fig -1: System Architecture

Final set of selected features is considered as the input for the further process as described in above existing system section.

5. MATHEMATICAL MODEL

Let, S is the Fault detection System for high dimensional data having Input, Processes and Output. It can be represented as,

S = {I, P, O}

Where, I is a set of all inputs given to the System, O is a set of all outputs given by the System, P is a set of all processes in the System

I = {I1, I2, I3}

where,

I1 is set of instances with feature set F ={f1,f2,…..,fn} with n features and m tuples.

I2 is distance threshold for classification.

P = {P1,P2,….,P10}

P1 = Symmetric uncertainty of each feature with class variable is calculated using,

SU (X,Y) = 2 * Gain(X|Y) / H(X) +H(Y)

where, H(X) is entropy of discrete random variable X. $H(X)= \sum_{x \epsilon X} p(x) \log p(x)$

Where, p(x) is prior probability for all values of X.

Gain(X|Y) = H(X) – H(X|Y)

P2 = Remove features whose SU is less than threshold SU Output will be the remaining feature set.

P3= SU of each feature with each other feature in O2 is calculated and G(V,E,W) is created.

where, V is set of vertices i.e. set of features and E is set of edges Eij. Eji is edge between Vi and Vj with Wij Symmetric uncertainty.

P4 =Minimum spanning tree calculated for O3 using prims algorithm.
The output will be MST.

P5=For each edge Eij

If SU(Fi,Fj) < SU(Fi,C) ^ SU(Fi,Fj) < SU(Fj,C)

then remove Eij

P6 =Initialization phase

Input to this step will be instance i from I1 and i is considered as Single member cluster.

C0 = {i}

P0 = i,

$\sum_{p0}^{-1}$ = 1/ tij

where C0 is single member cluster, P0 is prototype of C0 and $\sum_{p0}^{-1}$ is estimated covariance matrix.

P7 =Mahalanobis Distance.

When instance i received for classification,

MahalanobisDist(i,p) = $(i - mp)^T \sum_{p0}^{-1} (i- mp)$

Mahalanobis Distance of I is calculated from each cluster

P8= Nearest cluster P using O2 is derived

If MahalanobisDist(i,P) < threshold

i is normal

else i is faulty.

Output will be O7 and O8.

P9 =Membership of instance i will be checked with O4

If member (i, O4) == true

then update(O4)

$n_p^{new} = n_p^{old}+1$

$m_p^{new} = m_p^{old} + 1/ n_p^{new}(x- m_p^{old})$

Where, n is number of instances in O9 and p is prototype of O9

P10-=Merge clusters.

Cluster p' and cluster pm combined into one cluster p'' Np'' = Np' + Npm

Where,Np'' is number of members in new cluster p''

O = {o1, o2, o3, o4, o5}

O1 – Vector of Symmetric uncertainty from P1

O2 – Remaining features from P2

O3 –Undirected Graph G (V, E, W)

O4 – MST of G

O5 – Set of selected features

O6 – Single member cluster

O7 – Vector of Mahalanobis Distance of instance from each cluster.

O8 – Class of the instance.

O9 – Nearest cluster

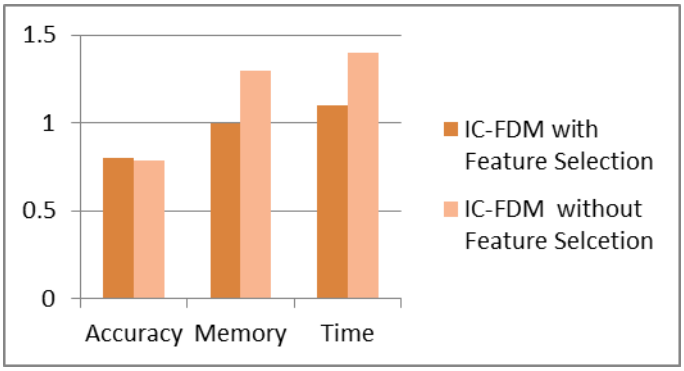
O10 – New cluster from merging in process P5

6. EXPECTED RESULTS

The aim of the performing experiments is to check the effect of the application of feature selection technique before applying the Incremental Clustering-Based Fault Detection (IC-FDM) technique and also to check the memory and time requirements for IC-FDM [6] and for IC-FDM with Feature Selection technique. The proposed method will improve the accuracy in case of high dimensional data as redundant and irrelevant features will be removed from it.Kdd99-r2l and kdd-u2r datasets will be used for experiments. First one contains the instances with r2l attack and second one contains instances with u2r attacks. Both dataset contains 35 dimensions. First contains 1.45 % outliers and second one contains 0.077 % outliers therefore these datasets are class imbalanced. It is expected that, If IC-FDM takes 1.3 units time and 1.4 units memory to complete a task then the proposed method will take 1 unit time and 1.1 units memory respectively.

Parameters	IC-FDM with Feature Selection	IC-FDM without feature selection
Accuracy	0.801	0.786
Memory (units)	1	1.3
Time (units)	1.1	1.4

Table1. Comparison between IC-FDM with Feature Selection and IC-FDM without feature selection



Graph 1. Comparison between IC-FDM with Feature Selection and IC-FDM without feature selection conclusion

7. CONCLUSION

For organizations, Fault Detection becomes important and critical. Many standard fault detection algorithms are available to address the problem of data imbalance in classification of data. However, existing system focused on reducing the number of data records here irrelevant feature removal technique is used with the incremental clustering

based algorithm for fault detection is which provides better results. Removing irrelevant features i.e. less important variables before applying fault detection incremental clustering based fault detection algorithm improves speed of the process and reduces computation and storage requirements and gives more accurate prediction of data.

REFERENCES

[1] V. Garcia, J.S. Sanchez, R.A. Mollineda, R. Alejo, J.M. Sotoca, "The class imbalance problem in pattern classification and learning", Pattern Analysis and Learning Group, Dept.de Llenguatjes iSistemes Informatics, Universitat Jaume I.

[2] Ramesh Nallapati," Discriminative Models for Information Retrieval", nmramesh@cs.umass.edu.

[3] 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.

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

[5] D. Ververidis and C. Kotropoulos, "Information loss of the Mahalanobis distance in high dimensions: Application to feature selection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 12, pp. 2275–2281, Dec. 2009.

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