IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

Classification of High Dimensional Class Imbalanced Data Using Data mining Techniques

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

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

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

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.

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

I. INTRODUCTION

There are many algorithms available for data classification. In Classification of data there were many major changes happened and had many problem related to this. As increase in size of data features the classification of data become difficult. Due to imbalanced data classification of data is not possible easily. The instance or class is not balance is called imbalanced data. Fault detection problem involves learning a binary classifier that provide two class labels i.e. normal and fault in data mining. If classes are not equally managed then class is imbalanced. Most of algorithm focuses on normal sample and ignore fault sample. 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. Models assume an underlying functional form of the classifier and have some fit parameters are parametric. Models have no explicit assumption about the form of the classifier are non- parametric. Instance of semiconductor information is considered and proposed a fault detection using incremental clustering. Incremental clustering algorithm finds the fault class distribution and process a reduced feature with accuracy and efficiency. The classification clusters normal instances to degrade the accuracy and prerequisites of evaluation. To distinguish potential defective wafers factual outlines are kept up for every group. Mahalanobis separation which is a factual separation measure that considers the connections and contrasts among the information directs utilized toward foresee the class mark of new wafer in multidimensional element. Proposed algorithm is very beneficial when performing flaw recognition in stream information situations with imbalanced information and even under procedure floats. When there is very high dimensional data present, computation cost and storage requirement increases. For this purpose we implement an algorithm for which the redundant and irrelevant features are removed from dataset. For this, based on a minimum spanning tree (MST), Fast Clustering based Feature Selection algorithm is used. For better performance and efficiency feature selection algorithm is used.

II. LITERATURE SUREVY

The majority class represents “normal” Classes, while the minority class represents “Fault” Classes. In multiple classes classification this problem is exists. It 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. Highly imbalanced two-class classification problems occurred i.e. small fraction of records of minority class than the majority class. Conventional methods tend to strongly favour the majority class, and largely ignore the minority class when dealing with an imbalanced data set. This result in systems: First step will have high detection ability with relatively high false alarm rate to produce small data sets with higher concentration lower or no detection of the minority class when directly applied to an imbalanced data set. Second step makes the verification more affordable as pre-screening step narrows down the data samples.
Model assume an underlying functional form of the classifier and have some fit parameters are parametric. The Support Vector Network based on parametric approach for two class classification problem which implements the idea that input vectors are mapped to high dimension feature set. The SVM finds the separating hyper-plane in the feature space that can create maximum distance between the plane and the nearest data of different classes. Consider the case of monitoring semiconductor manufacturing process. Increase in the output and improved product quality is of importance in manufacturing. Quickly detecting faultier

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

and diagnosing the problem is main motive of multivariate statistical process control. Principal component analysis PCA method is popular to notice the problem. The method has some disadvantages. 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.

The problem occurs in real world applications including time series analysis or some industrial process. One Class Support Vector Machines proves efficient in non-stationary classification problem. 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. 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.

In semiconductor manufacturing it is necessary to quickly detect faulty behaviors and consistently improve equipment productivity. In various industries, fault detection is a crucial problem. In classification 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 this paper a non-parametric control charts such as k-nearest neighbour 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. The cumulative distance of this observation to its k nearest neighbours 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 neighbour distance (k-NDD) rule based on local covariance structure of the monitored observations.

A dataset is imbalance if classes are not equally represented. In data mining, fault detection issues involves learning a binary classifier that provide two class labels i.e. normal and fault. 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. 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.

III. THE EXISTING SYSTEM

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. 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 performs following four phases.

A. Initialization:
A new single member cluster, accepts a new sample and it begins the fault detection task for the single member cluster.

B. Classification:
Class label of the new sample is assigned. Decision of labeling the label is made by calculating the distance between the new sample and center i.e. mean of the nearest normal cluster. If number of instances is less than number of dimensions then cluster is considered as immature otherwise it is mature cluster. In classification phase, distance of the instance to be classified from each cluster is calculated. If cluster is immature then Euclidian distance is used otherwise mahalanobis distance [8] is used. If distance is less than thresholds then instance are classified as normal cluster otherwise it is classified as faulty instance.

C. Cluster Update and Generation:
If classification is right then membership of the instance to the nearest cluster is checked. If distance between nearest cluster and instance is less than specific threshold then, the instance is considered as member of the cluster. If instance is member of the cluster then cluster and cluster prototype is updated.
If instance is not the member of the cluster and its actual label is normal then new single member cluster is generated. If classification is wrong and actual label is normal then new single member cluster is generated. If classification is wrong and actual label is fault then that instance is dropped and next instance is processed.

D. Cluster Merge:
As cluster increases, 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.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

IV. PROPOSED SYSTEM

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. 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.

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.MST will be constructed using prim's algorithm [9].

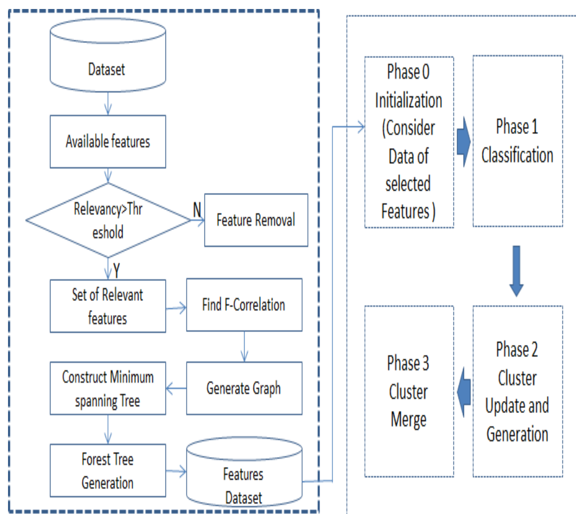


Fig. 1. System architecture

Fig. 2.

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.

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

V. 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 prim's 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 upd0ate(O4)
$n_p^{new} = n_p^{old} +1$
$m_p^{new} = m_p^{old} + 1/ n_p^{new} (x- m_p^{old})$

Copyright to IJARCCE

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

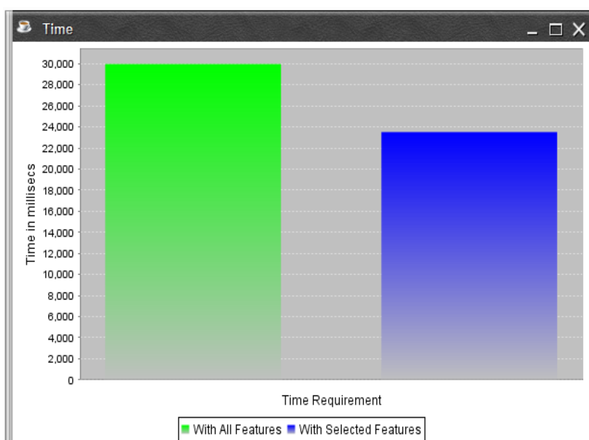
Where, n is number of instances in O9 and p is prototype of O9
P10- Merge clusters.
Cluster p' and cluster pm combined into one cluster p''
Np'' = Np' + Npm
Where,Np'' is number of members in new cluster p''
O = {o1, o2, o3, o4, o5}
O1 – Vector of Symmetric uncertainty from P1
O2 – Remaining features 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

VI. ACTUAL RESULTS

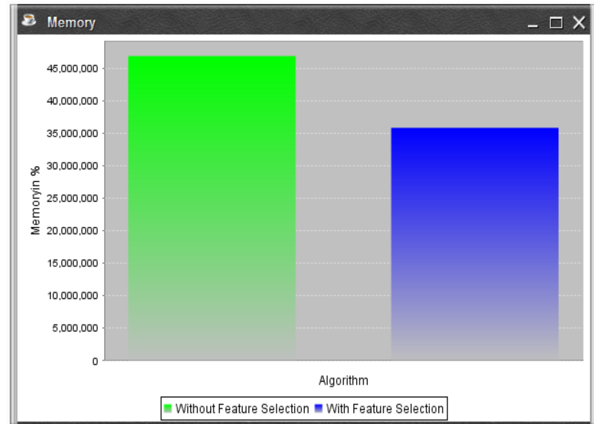
The aim of the performing experiments is to check the effect of the application of feature selection technique before applying the Incremental Clustering-Based Fault Detection (IC-FDM) technique and also to check the memory and time requirements for IC-FDM [6] and for IC-FDM with Feature Selection technique. The proposed method improved the accuracy in case of high dimensional data as redundant and irrelevant features will be removed from it.Kdd99-r2l and kdd-u2r datasets [10] will be used for experiments. First one contains the instances with r2l attack and second one contains instances with u2r attacks. Both dataset contains 41dimensions. First contains 1.45 % outliers and second one contains 0.077 % outliers therefore these datasets are class imbalanced.

Parameters	IC-FDM with Feature Selection	IC-FDM without feature selection
Accuracy	99.99	97.95
Memory (units)	31.07(Kb)	36.25(Kb)
Time (units)	23.52(sec)	29.66(sec)

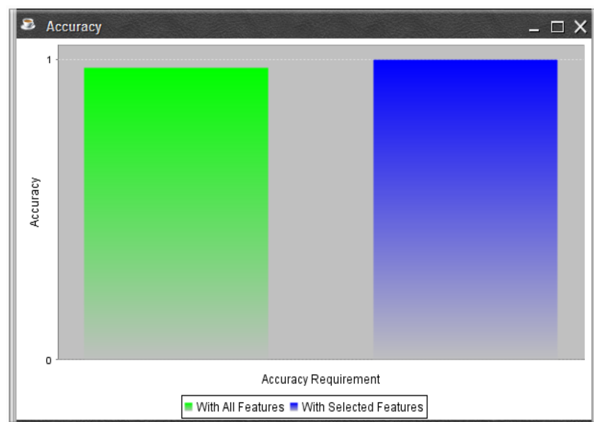
Table1.Comparison between IC-FDM with Feature Selection and IC-FDM without feature selection



Graph 1. Time Comparison between IC-FDM with Feature Selection and IC-FDM without feature selectionconclusion



Graph 2. Memory Comparison between IC-FDM with Feature Selection and IC-FDM without feature selectionconclusion



Graph 3. Accuracy Comparison between IC-FDM with Feature Selection and IC-FDM without feature selection

VII. CONCLUSION

Classification issues with imbalanced data in high dimensional space are addressed using feature selection technique and incremental clustering fault detection method. We have used irrelevant feature removal technique with the incremental clustering based algorithm for fault detection as it 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.

ACKNOWLEDGMENT

It is been rightly said that we are built on the shoulder of others. For everything I achieved, the credit goes to all those who had really helped me to complete this work successfully. I am extremely thankful to my Project Guide Prof. Jyoti N. Nandimath for guidance and review of this paper work. I would also like to thank the all faculty members of SKNCOE, Pune and my friends who helped me for this work.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 7, July 2016

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] C. Cortes and V. Vapnik, "Support vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297.
[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] Jueun Kwak, 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 Guangtao WangFast, "Clustering based Feature Subset Selection algorithm for High-Dimensional data," IEEE Trans. Know. Data Engg., vol 25,no. 1, Jan 2013.
[8] 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.
[9] Prim, R. C. (November 1957), "Shortest connection networks And some generalizations",Bell System Technical Journal 36 (6): 13891401.
[10] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set" 2009.

BIOGRAPHIES

Vidya Omase pursing ME degree in Computer Engineering from Smt. Kashibai Navale College of Engineering, Pune-41 India and BE degree in Computer Engineering from Vidya Pratishthans College of Engineering, Pune University, India, in 2013. Her current research area is data mining.

Prof. Jyoti N. Nandimath received ME degree in Computer Science Engineering from Walchand college of Engg, Sangli India in 2011 and BE degree in Computer Science Engg from Karnataka University, India. Her current research area is data mining. Currently she is working as Assistant Professor in Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, Pune-41 India. She has 20 years of Industry and teaching experience.