Credit Card Fraud Detection Using Anomaly Techniques

Dr.V. Ceronmani Sharmila
Head of Department-Department of Information Technology
Hindustan Institute of Technology and Science
Chennai, India
csharmila@hindustanuniv.ac.in

Kiran Kumar R
Department of Information Technology
Hindustan Institute of Technology and Science
Chennai, India
kirankumar.kk1711@gmail.com

Sundaram R
Department of Information Technology
Hindustan Institute of Technology and Science
Chennai, India
sundaramsidharth@gmail.com

Samyuktha D
Department of Information Technology
Hindustan Institute of Technology and Science
Chennai, India
samyuvarma26@gmail.com

Harish R
Department of Information Technology
Hindustan Institute of Technology and Science
Chennai, India
harish.gha@gmail.com

***Abstract*—Credit card fraud transaction detection system is a method used for determining the fraudulent transactions that take place every once in a while. The project uses a test data set of around 27,000 credit card transactions which have been taken from Caltech (Kaggle). The project comprises of primarily 2 major algorithms and uses anomaly detection as a method to classify the fraudulent transactions. The Local Outlier Factor (LoF) defines the various parameters that have to be used in determining the criteria for fraudulent transactions. It then checks upon the different transactions for the various parameters present in the given LoF. This factor then gives each transaction a score based on the various transactions that have or will have taken place. These scores can range from 0 - 1. Each transaction is thus given a score which is based on the various parameters given in the LoF. The second part of the project is isolation forest algorithms which is an algorithm that isolates the transaction which have a high rate of anomaly detected in them. Thus, these transactions are isolated and then checked with various parameters to be labelled as either fraudulent or real transactions. The algorithm also uses charts to check for spikes in the average transaction. We also uses technique like data visualization in order to show the output in more understandable ways which may include histograms, graphs and matrix .Through these two algorithms and with help of data visualization technique we can detect the fraudulent transactions from correct transaction and obtain results in quick time , Since these algorithms are much more time efficient than other machine learning algorithms in this type of tasks .**

I. INTRODUCTION

The machine learning algorithms are used in various sectors of industry specially when it comes to computing field. These algorithms are developed by machine learners who are specialized in machine learning field with proper study of all the tools handled in the respective field. Using machine learning tools, we can perform various instructions to execute the system works. Machine Learning is now playing an important aspects in mobile communication field too. Since several big and popular companies are totally depended on these experts and in future, they need more and more of these peoples in order to match the rivalry with their opponents. These machine learning algorithms now used to find aspect such as emotion detection, fraud detection, path finding etc. such projects can be achieved by using machine learning algorithms .However there is lack of Data scientist in coming future .Generally Data scientist are group of people who study about machine learning, data processing ,data analyst etc. Though machine learning are now further improvised and now we have deep learning concepts which is more accurate and uses tools like TensorFlow etc. and advanced algorithms which makes it more complicated than our general machine learning algorithms and also it takes much more time in writing code but it is more time efficient than machine learning algorithms with much more higher level of accuracy.

These algorithms are very highly used in financial sectors in order to perform predictions on the basis of transactions. These generally leads to detection of fraud or prediction of results and also stimulation of the processes. Since the outputs cannot be in understandable form, so in order to understand the output clearly, we use some technique or methodology in data science, and it is known as data visualization.

Data Visualization is the technique which is mainly used for showing the output in more conventional method or in more understandable form. These are represented in form of histograms, matrix, and many more graphs. In machine Learning or Deep Learning in order to import data or perform execution we have several libraries which contain packages which are predominantly used performing importing or various other functions. These libraries are preinstalled in compiler, but some may need to download and extract by user manually while in case of packages through libraries. These packages contain functions that are used for importing, extracting, processing and performing predictions. Libraries such as pandas, sci-learn, NumPy and matplot are mainly used in our project in order to find fraudulent credit card transactions. Each library has its unique features and functions such as pandas are mainly used for importing datasets from internet or from your system, similarly for matploit that are used for plotting mathematical graphs and numpy for array data entry likewise each library in machine learning has its own functions and features. In our project we use panda in order to import all 27,000 credit card transaction datasets which we get from Kaggle site and downloaded that file in our system. This file is csv type file

which is commonly known as comma separated values so each of the credit card transactions in that file are separated through comma. Then once we imported all the datasets now, we execute these datasets in two-way process, which generally include the two main sets, training set and test set.

Once we done these processing now we need to predict the output in our case it is either fraudulent transaction or correct transactions .These can either be verified through binary values such as 0s and 1s which basically means if the transaction result is 0 then it is correct transaction and if the transaction result is 1 than it is fraudulent transaction.

So, in order to easily understanding we use data visualization techniques in our project which compromises our last and final stage of our project that shows the output in form of graphs, histogram, matrix etc. which makes easy for everyone to understand and quickly come up with conclusion. Since we have similar algorithms to perform these parameters but these two algorithms (local outlier factor and isolation forest algorithms) are more time and space efficient in order to come up with results much quickly as compared to other algorithms and also it is the simplest among other various algorithms such as decision tree ,clustering and techniques like neural network.

The banking sector in the world is the most valuable economically diverse sector. An average of 1 billion transactions can take place every minute. When such a diverse system is operated the security for such an operation required is immense. Each transaction can contain amounts leading from just 1 rupee up to even a few crores. Such systems require huge amounts of security in it. There are various measures that have taken place for the security of such banking firms which can range from.

Hardware and IoT concepts to various machine learning algorithms the efficiency of all of this depends on the monetary value that it presents. An IOT based solution would cost more money and will involve a lot of hardware and feasibility for it. Machine learning algorithms on the other hand will involve a lot of data and processing power which will lead to computational cost. Our project aims to improve the efficiency of both the costs and security involved in our transaction. We do this by taking a sizeable sample data of transactions and apply two anomaly detection algorithms to it. The first one detects the anomalies and separates them from the rest. The second one segregates them into a separate group and can be classified as the abnormal transactions or “misfits” based on them LoF scores. This thus helps to check a trend of transactions, test it out with 30 odd parameters and giving out a score to predict its value as either 0 or 1.

II. RELATED WORKS

The industry is filled with practices that are fraudulent in nature. The core objective is to primarily detect the various credit card frauds and then, check on the different algorithms and make an informed decision. The objective is to show and analyze recently published outcomes in credit card fraud detection. This shows the common terms in credit card fraud and shows key statistics and figures in this field. Based on the versions of fraud incurred by the institutions, many precautions are taken and put into work. The ideas given in this paper bring about a lot of efficiency. The ability of the various methodologies taken a look here in the limitation of credit card. Although there are many issues when a real credit card customers are wrongly framed as anomalies.

The issue with creating enterprise from the web have it in such a way that both the card and the holder need not be present in the premises. It is thus very hard for the business person to find out whether the buyer is the true owner or not. Payment card fraud is now a major issue across the globe. Enterprises and companies shed huge costs yearly because of fake and seamsters continuously finding alternate ways to do unethical activities.

The part to look forward is the fact that these kinds of malicious activities have certain types of formats that they follow and are somewhat easier to find out its root path and further details about it. In this article we plan to check malicious transaction through the algorithms as well as with the genetic algorithm. As we will see that artificial neural network. Artificial intelligence can actually be programmed and trained in such a way that they can imitate a real brain, however it is nearly impossible for AI to reach the level of subtleness and detail as the human brain, it is similar to the unit in our brain that makes core decisions.

A. Genetic algorithm

It is used in the making of important directions on the network topology, number of unseen layers, A lot of nodes which will be apparently used for the design of the network required for the credit card fraud transaction. Various parameters are taken into account in this case and thus various decisions are made on it based out of this. To learn more about artificial neural network and its applications the system implements supervised learning algorithm with forward back algorithm. In the end we will see what the further course of levels can be done in checking for the fraud detection.

Credit card fraud is classified as either unauthorized account activity by a person for which the account was not intended. Additionally, this is a program for which intentions are to be defined and various steps should be taken to halt the misuse in the future and bring about various potential management ways to protect against similar cases in later years. In layman terms, Credit Card Fraud is defined as when one person uses the other person’s credit card for various purposes whilst the actual holder of the card and the card issuer aren’t educated of the fact that the card is being used. The people using the card will have no type of relationship with the cardholder and have very wrong intentions of using the card for personal gains and purposes.

B. Decision tree

Neural networks are those which cause various fraudulent detections. Envisioned a very feasible and something that has much access to the data of fraud happening on the web. Fraud detection environment, environment, which is made out of the core technology of a classifier. However, the original problem/issue is that information has to be put together. Such concepts are: Card watch[3]Back-propagation of error signals[5],FDS(Ghosh & Reilly,SOM(Quah et al.) which helps in improving detection efficiency "mis-detections" (Kim & Kim). Data mining is an efficient way to probe such events ,such as Hadoop which enables the utilisation of various NN algorithms, These have thus been used in many cases of fraud detection. Various networks are also used as a technique to check for malicious transactions , and have been sent to find irregularities in the communications platforms(Ezawa & Norton,1996) as well as in the credit card operations, Although the time is a very important constraint and is considered to be one of the biggest drawbacks of such a method. Primarily when compared with neural networks(Maes et al.,). Furthermore, Trained and efficient systems have also been used in credit card fraud using a rule-based expert system(Leonard)

C. Decision tree

[7] and [12] Given another grouping techniques. The squint category analysis is the system that permit able account which are acting apart from the others at the moment of time while we saw they are acting identical before. The accounts then marked as suspicious. The fake analyst then has to explore the cases. The theory of squint group analysis that if accounts acting identical for a definite amount of time then the account is acting significantly apart, that account need to be addressed. The turning point analysis uses another perspective. The theory denotes that if a certain change of card usage is identified for single basis, then the account needs to be addressed immediately. In another case, that can be based on transactions for an individual single card, the break point analysis can be noticed different behavior. The signal of different acting is of immediate transaction for an very high amount with a very high frequency of usage.

III. LOIF BASED ANOMALY ALGORITHM

The proposed technique we use comprises of two major anomaly detection algorithms such as Local outlier factor and Isolation forest algorithm. It comprises of two parts in it. The first part deals with testing of the outliers and gives it a score. The second part of an isolation technique which isolates the anomalies and separates them from the general pack. The dataset consisting of 28,000 odd transactions are first sent into the first algorithm which determines the outlier factor based out of 28 parameters. Each transaction is then observed into the various parameters and given a score out of 0 and 1 where 0 determines that it is a fake transaction and 1 determines that it is a real transaction.

This transaction is then moved on towards the second algorithm: Forest Isolation algorithm which isolates the bad values and segregates them into separate groups. These techniques are then alerted into isolation and intimated to the user. The software used for this project is Jupyter notebooks which is a sub product of Anaconda. Jupyter notebook is completely online and is hosted locally on the cloud. The initial step is to first import the various packages that are required for this process such as NumPy, Scipy, MatplotLib, Seaborn, Pandas.

Pandas are used for importing data into the Jupiter notebook. Pandas are used for manipulating and putting data into the canvas. Pandas are core and vital to our project as it is how the 27,000 transactions and datasets can be imported into the Jupyter notebooks environment through Pandas. NumPy is another library which is used for doing mathematical calculations in the dataset. Every mathematical transaction done using NumPy.

Scipy is another library which is used for scientific calculations such as sin and cos operations in the same dataset, Matplotlib is a mathematical graph plotter which is used for the purpose of data visualization. Different types of data visualization that can be done using matplotlib are Histograms, Double matrix squares, Anomaly graphs etc. Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. It is used for making scatter plot diagrams and other details as such.

IV. ALGORITHM FOR ANOMALY DETECTION

We have two algorithms and their working is explained separately in the following:

A. Local Outlier Factor (LOF)

It is an unsupervised machine learning algorithm. The expected rating of each and every sample is called Local Outlier Factor, In short LOF. It calculates the local deviation of density of a given sample with respect to its nearest neighbors. It basically compares with local neighborhood Instead of global. It is local so that the expected or anomaly rating totally depends on how isolated the sample is with respect to the surrounding neighborhood. The number of neighbors are basically considered, is typically chosen through Two major factors a) greater than the minimum number of sample a group has to hold, so that other sample can be local outliers comparable to this group, and 2) smaller than the maximum number of nearby sample that can possibly be local outliers.

B. Isolation Forest Algorithm (IFA)

The Isolation Forest separates study by randomly choosing an attribute and then again randomly picking a split value between the maximum and minimum values of the selected

attributes. Since recursive separation can be represented by a tree structure, the number of splitting required to separate sample should be identical to the path length from the root node to the ending node. This path length, averaged over a forest of such random trees, is estimation of normality and our conclusion function. Random separation produces attentively shorter paths for variation. Hence, when a forest of random trees mutually produces shorter path lengths for certain samples, they are highly likely to be anomalies. This algorithm basically follows 3 stages of process

- Forest
- Isolation Tree
- Evaluation (Path Length)

Each of these step carried out with special purpose of its own for example forest step will help it with clustering technique means grouping while isolation tree reduce the isolated transaction and finally evaluation perform the execution of both the algorithms and comes with the output in the form of graph using data visualization technique like histograms, correlation matrix etc. .The evaluation step also known for its path length which defines the procedures of the both (Local Outlier Factor and Isolation Forest Algorithm).So, with the help of these we finally come up with the solution or outcome of our project .

V. SIMULATION AND ANALYSIS

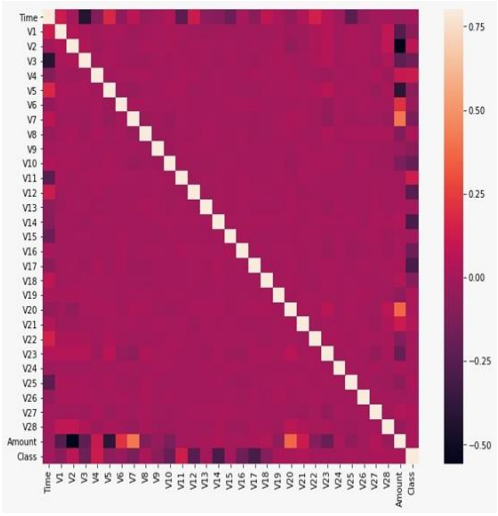

Fig. 1. Heat Map (*Correlation*)

This is called a correlation matrix diagram which is used primarily to check if there is any correlation between various different parameters and various variables in our dataset. This is done so as to keep it in check before building our networks and fitting the model. It helps show the features which are core for this specific project and the overall classification process. It is a pyplot figure which uses Seaborn and a SNS heatmap. This turns our correlation matrix into a visual display. SNS heatmap shows the various pressure points present in our data. It also helps to realise which are important and which are not. It consists of all the V parameters on both sides of X and Y axis with a scale which scales from -0.75 to + 0.50.

Since most of our transactions are real time and are pretty much legitimate the heat map does show a lot of significance near the 0 points. However the heat map does show us that there is a lot of correlation and differences in the class column in all the V points. These heatmaps show the various correlations that are present in our classifications. The one part which is not having a lot of similarity is the class and the amount , both of which vary in different ways as per the requirements. Our sample space for the program has been diluted to 10% of the actual credit card database to around 28,490 from around 280,000 to get a clearer accuracy and to have a The transactions are also using pictorial representation of the various parameters used in our dataset. We use histograms to pictorially represent the various anomalies in each parameter as per the given agenda and requirement.

The various parameters involved in the database include :

- Time
- Class
- Amount
- Avg transaction
- Location
- Transaction Limit
- Merchant vendor
- Country
- Issuing bank etc.

Amount:

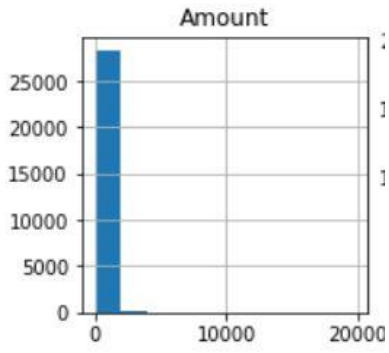

Fig 2. Amount Bar chart

The histogram in the above graph shows the average amount transacted by all the customers and the number of transactions that are fraudulent and the number of transactions that are legitimate. If the transaction is at 0 means it is a legitimate one and if it is at 1 it is a fraudulent transaction. As

per our graph more than 98% of the transactions are legitimate and only less than 2% is considered as fraudulent.

Average Transaction:

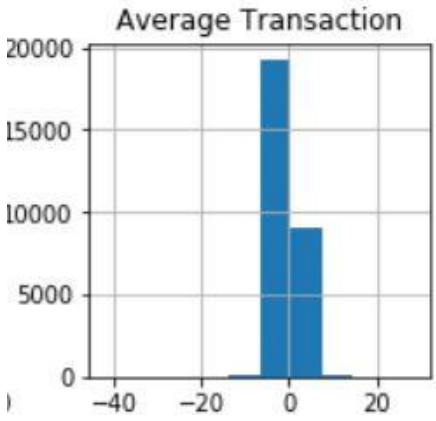

Fig 3 Average Transaction

The above transactions are considered with their average transactions and can see a spike, or a lower spike based on the transactions. If the transactions presented are unusually higher or lower than the actual average transaction it will give out a "0" or a "1". If it is "0" there seems to be no transaction errors and are completely legitimate and on the other hand if the transactions are having "1" or near to "1" it is considered a fraudulent or an anomalous transaction.

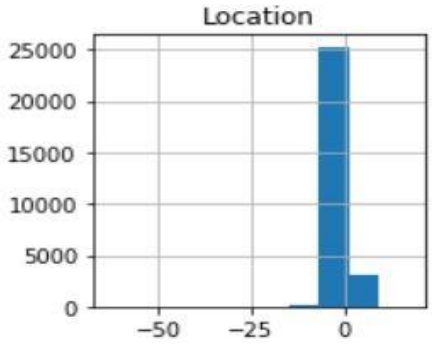

Fig 4 Location

The location histogram shows the bar graph of the whereabouts of the transactions and it will determine the various locations of the transactions based out of it. If the transaction is from an unusually different location which is not the same compared to that of the current location it will flag the location and let the customer know that it could be a fraudulent transaction. If the value given is "0" it means the transaction is well within its location permits, whereas if the location is from an unusually new location it will show "1" which again will flag and give a "1". This process shows that more than 22000 transaction are from the original or usual locations and less than 5% of it are from questionable locations.

Transaction Limit:

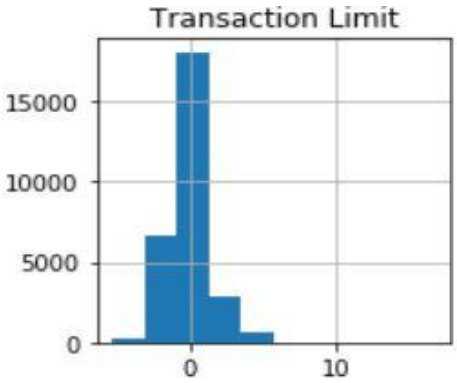

Fig 5 Transaction Limit

The transaction limit of the project shows the various transactions by the customers and also shows the number of customers who have exceeded their average transaction limits. For example: If a customer on average makes transactions of around 15000 per month and all of a sudden there is a slump in transaction or if there is a spike in transaction which can be deemed as a unusual dimension can lead to being flagged and gets 1. If the case is 0 then it is a true transaction and if it is a 1 it can be a fraudulent one and can be flagged. In this dataset it shows that around 80% of the transaction is outrightly well within the limit with less than 10% leaning slightly outside it and about 5% leaning way out of the limit.

Issuing Bank:

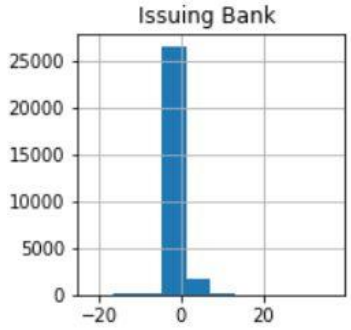

Fig 6 Issuing Bank

Issuing bank shows weather the customer uses the card in the ATM's and outlets of the same designated bank or of any other bank. It should be able to flag details if a customer constantly uses ATMs of banks which are generally not used or are not affiliated with the issuing bank. 0 suggests that the bank is the issuing bank and 1 suggests that the bank is a different one and could be fraudulent. Our data set shows that more than 85% of transactions take place within the legitimate transaction cap and only 15% go to different banks and could* be fraudulent.

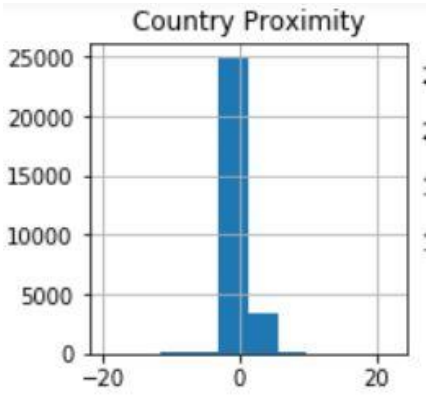

Fig 7 Country Proximity

This is primarily for International citizens who travel around a lot. The credit card company gets access to the frequent locations where the credit card is used for transactions. It also shows whether the transactions is done in unusually different countries. Those who use the transactions in unusual countries will get a "1" and those who use it in the stipulated locations get a "o". In our data set 80% falls under the usual case whereas 20% of them have some or the other sort of credit card fraud.

Card Type:

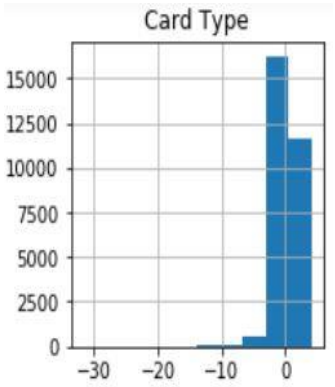

Fig 8 Card Type

The credit card type shows the different types of Card, It could be Visa, Mastercard or even RuPay. It classifies the cards into its different classes and varieties. In this case if the card is changed or is of different type it is given a 0 or else given a 1. In this case all the types are legitimate with a 100% accuracy to it.

In the end, these are just a few of the core parameters used in our project. There are various other parameters that are used but have been hidden to respect the privacy of the customer's transactions. It has been hidden using a PCA dimensionality reduction system. It is also known as Principal component analysis. There are other core parameters we use in this profile. These are then connected together and a correlation towards it is developed using Pyplot as a heat map showing the various correlations present in it.

VI CONCLUSION

Thus, using various dimensionality analysis, we check and find out the different sections of the project. There are more than 30 parameters in our project for which an amalgamation is created as per the requirement. This is then used to a training model such as Local Outlier factor and Isolation forest algorithms and then it is tested out for each algorithms and then it is first accumulated together. This accumulated factions are then applied to isolation forest algorithms and anomalies are detected as per requirement. The isolated anomalies are then referenced as "1" and are classified as fraudulent and the others which are legitimate are classified as "0"

REFERENCES

[1] Signals and system, "Practicing data science and algorithms"
[2] Data science for Dummies "It uses the various data science implementations"
[3] Y. Sahin and E. Duman, "Distinguishing charge card misrepresentation by choice trees and bolster vector machines", Proceedings of the International MultiConference of Engineers and Computer Scientists 2011 Vol I, IMECS 2011, March 2011.
[4] Elkan, C. (2001). Mystical reasoning in information mining: exercises from COIL test 2000. Proc. of SIGKDD01, 426-431.
[5] Mohammed, J. Zaki., and Wagner, Meira Jr. (2014). Information mining and examination: essential ideas and calculations. Cambridge University Press. ISBN 978-0-521-76633-3.
[6] [F. N. Ogwueleka. (2011). Information mining application in charge card misrepresentation location framework. Diary of Engineering Science and Technology, Vol. 6, No. 3 (2011) 311 - 322.
[7] V. Bhusari and S. Patil. (2011). Use of concealed markov show in Visa misrepresentation discovery. Worldwide Journal of Distributed and Parallel Systems (IJDPS) Vol.2, No.6.
[8] S.J. Stolfo, D.W. Fan, W. Lee, A.L. Prodromidis, and P.K. Chan. (1998). Visa extortion recognition utilizing meta-learning: issues and beginning outcomes, Proc. AAAI Workshop AI Methods in Fraud and Risk Management, pp. 83-90.
[9] Sen, Sanjay Kumar., and Dash, Sujatha. (2013). Meta learning calculations for charge card extortion location. Universal Journal of Engineering Research and Development Volume 6, Issue 6, pp. 16-20.
[10] Maes, Sam, Tuyls Karl, Vanschoenwinkel Bram and Manderick, Bernard. (2002). Charge card extortion location utilizing bayesian and neural systems. Proc. of first NAISO Congress on Neuro Fuzzy Technologies. Hawana.
[11] A.C. Bahnsen, Aleksandar, Stojanovic., D. Aouada and Bjorn, Ottersten. (2013). Cost delicate Visa extortion discovery utilizing bayes least hazard. twelfth International Conference on Machine Learning and Applications.
[12] Michael Berthold, David J. Hand," Intelligent Data Analysis", Springer,2007.