|  |
| --- |
| FORM 2 THE PATENTS ACT 1970  (39 of 1970)  &  THE PATENT RULES, 2003 COMPLETE SPECIFICATION (See section 10 and rule 13) |
| 1. **TITLE OF THE INVENTION: -**   **TECHNIQUE FOR OPTIMIZING COMPLEXITY AND REDUNDANCY OF BIGDATA** |
| **2. APPLICANT (S)**   |  |  |  | | --- | --- | --- | | NAME | NATIONALITY | ADDRESS | | **DR. BRIGHT KESWANI** | **INDIAN** | **ASSOCIATE PROFESSOR,**  **DEPARTMENT OF COMPUTER APPLICATIONS,**  **SURESH GYAN VIHAR UNIVERSITY, MAHAL JAGATPURA, JAIPUR, INDIA.** | | **PRITY VIJAY** | **INDIAN** | **RESEARCH SCHOLAR, SURESH GYAN VIHAR UNIVERSITY, MAHAL JAGATPURA, JAIPUR, INDIA.** | |
| **3. PREAMBLE OF THE DESCRIPTION** |
| The following specification particularly describes the invention and the manner in which it is to be performed |

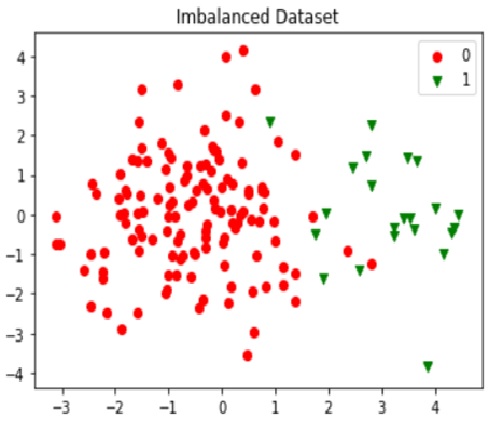
**1.1 Introduction**

1. Dataset state of imbalance occurs when more numbers of examples for one class is present in it. A class having examples more in number are called majority class and other is called minority class [1].
2. This problem is not new during mining the data. Since past years the researchers are continuously looking towards finding a solution for learning from imbalanced data but still, this issue is an important topic for research [2].
3. With the alighting of big data, both data mining and machine learning technology over- emphasizing to grow into one of the extensive tools for eradicating deep insight from imbalance dataset [3]. Dataset imbalance problem, on the other hand, gained lots of importance with the advancement of big data and with it also bought much threat while gathering valuable information from it.
4. Up gradation of data level and algorithm level methods have been noticed with many new hybrid approaches [3]. Many domains do not have balanced dataset. Minority class consists of the most important information and extracting that piece of information from the big dataset is an actual challenge.
5. During the processing of these datasets minority class does not get importance and therefore classification shows inaccurate results. These real-life imbalance dataset challenges galvanize researchers and scholars to focus an effective and real-time solution for such a problem.
6. Messy data is another problem for machine learning algorithm. Data redundancy [10][11], outliers [9], missing records, etc. are main cause of unclean data. Therefore data cleaning is first and foremost task before analysis and time consuming too.
7. Before classification of imbalance complex datasets, resampling methods are headed for balancing such dataset. Old resampling methods like oversampling and under-sampling techniques balance the dataset but have lots of disadvantages. Oversampling when applied can cause problem of over-fitting and under-sampling cause’s information loss.
8. New HPRT, an algorithm which will automatically balance the data set without much information loss and at the same time, the classifier will not cause any over-fitting problem. Our method uses both under-sampling and oversampling, where undersampling occurs at majority class through reduction of redundancy and extreme outliers. This step can reduce majority class, almost up to 40% and then in the next step SMOTE oversampling is applied to a minority class. This two-step approach can balance the dataset while cleaning it and as a result machine learning algorithm performance can enhance.

**1.2 Imbalanced Data Class: Problem and Related Solution**

Most of the real datasets are imbalanced such examples can be found in various domain finance (fraud detection), Medical (identification of disease), network intrusion, telecommunication (equipment failure) natural disaster etc. Lots of challenges occur during the processing of imbalanced learning [10], some of which are listed below:

1. When the dataset is dealing with imbalance problem generalized matrix such as accuracy cannot evaluate the performance of classifier properly because in such cases costs of different errors are diverse. Different matrices such as confusion metrics, F-Score, etc. can be used for better performance.
2. Another problem associated with class imbalance is the unavailability of informative training data for machine learning. Very less number of features related to minority class possesses challenges for classifier during the process of learning from such dataset and hence produces the results with weak learning models where minority features are misclassified.
3. Most of traditional machine learning algorithm uses greedy search and divide and conquer rule during the learning process. But these rules perform inadequately for a minor class of imbalance dataset. Therefore further researchers are looking up this matter very seriously for development of an enhanced version of traditional machine learning algorithm which can deal with the imbalanced type of dataset.



**Figure 1: Imbalanced Dataset**

**1.3 Sampling methods**

Sampling is the most common approach when dealing with class imbalance. The process of sampling is applied to the dataset in order to convert it as a balanced class distribution. Sampling is further classified into under sampling and oversampling. Under sampling is the process, where a feature of majority class is reduced and oversampling adds new artificial features to the minority class. Combination of over and under sampling can also be applied to balance data [4].

**1.3.1 Oversampling Methods**

It is a process of adding artificial features to minority either randomly or by some calculation. Newly added samples help to balance of the dataset but can increase the risk of overfitting.

**1.3.2 Random Oversampling**

Random oversampling is a very familiar way of balancing the dataset. In this technique, features are added randomly by selection of the set of examples from minority class. Although, this approach increases features of a minority class and hence balance the dataset the extra headache of overfitting and an increase in the time of training data are some of the drawbacks of random oversampling.

**1.3.3 Synthetic Minority Over-Sampling (SMOTE)**

SMOTE increases the features of minority class [12] by adding new artificial features similar to it. KNN algorithm is used by smote for adding new features to the minority class, by choosing its neighbor randomly depending on the amount of oversampling needed. This is proved to be a better sampling approach because here data points similar to the original features are added to a minority class. Many other algorithms derived from the SMOTE, but it gives poor results with high dimensional data as the calculation time of SMOTE increases thus decreasing its performance.

**1.3.4 Borderline SMOTE**

Borderline SMOTE [13] is an oversampling algorithm inspired by SMOTE. it has two versions, borderline SMOTE1, and borderline SMOTE2. These methods add features only near the borderline using KNN. Borderline SMOTE2 is advance than Borderline SMOTE1, as it selects both positive and negative nearest neighbors. It has been reported that Borderline SMOTE performs better than SMOTE.

**1.3.4 Adaptive Synthetic (ADASYS)**

It is also derived from SMOTE, using the weighted distribution of minority class. Here data points are added using KNN method depending upon majority NN. It cannot deal with outlier well and thus not perform well with noisy data.

**1.3.5 Random Under-Sampling**

Random undersampling is simplest among all the methods of reducing the majority class randomly such that minority class becomes equal or near to equal with majority class. But major drawback with this type of sampling is a huge information loss caused by deleting lots of features from the dataset. The features discarded from the dataset might contain useful information for building pattern, the absence of which can generate week models.

**Hybridization Pre-processing and Resampling Techniques (HPRT): A Solution for Imbalance and Messy Dataset**

**2.1 Objective**

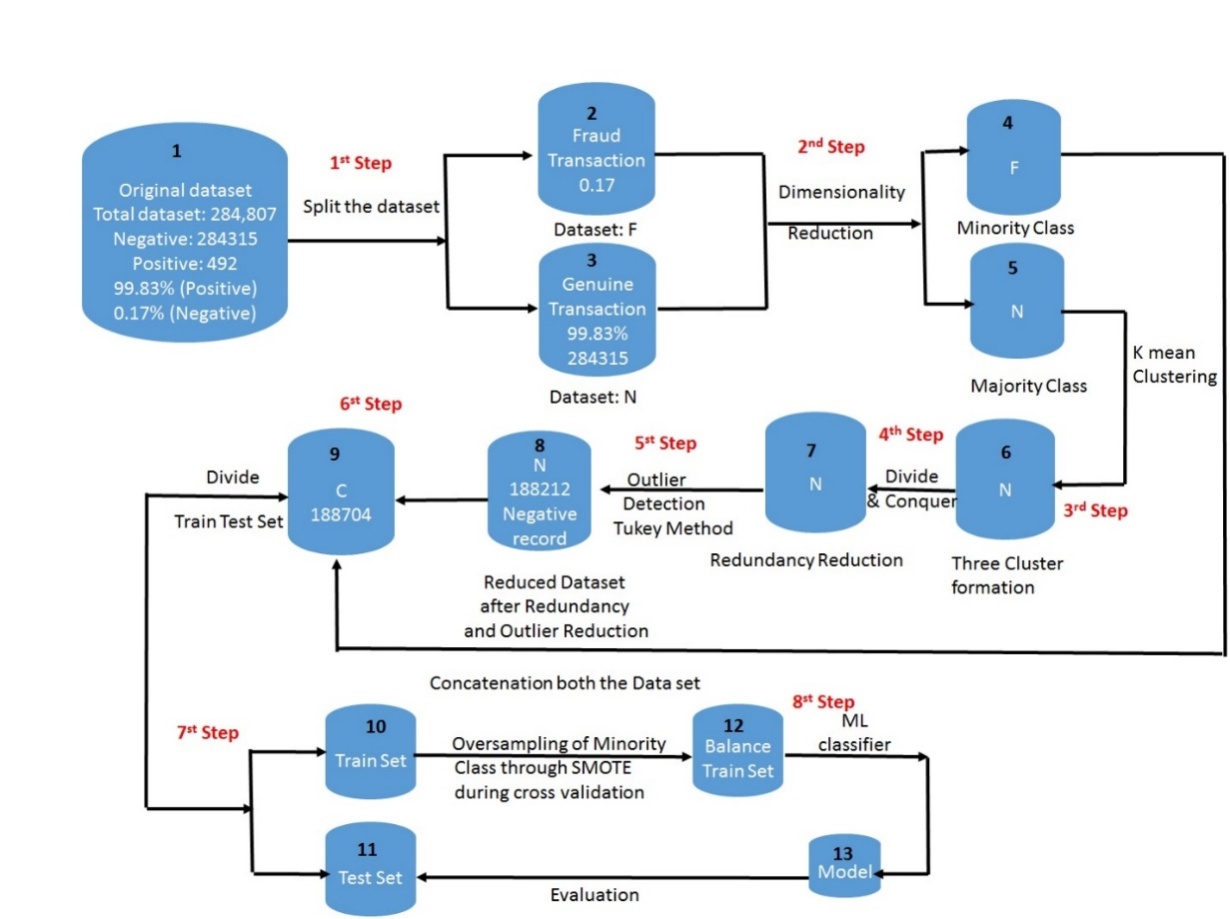
Popular resampling methods, Random undersampling, Random Oversampling and SMOTE oversampling, cause certain challenges during its application to an imbalanced dataset. Therefore, the goal of this research is to develop an enhanced and hybrid machine learning resampling approach, which can act as a perfect solution for reducing the complexity of big data because of its imbalance nature.

**2.2 Dataset Description**

Dataset had captured from university libre de Braxelleswebsite[14], during research collaboration of machine learning group and world line on credit card fraud detection. It contains details of European card holder (2013),), credit card transaction for two days. It is very large dataset consisting almost 300,000 transactions out of which only 0.17% is fraud cases. The Dataset is highly complex and imbalanced, containing very less fraudulent class compare to highly distribution genuine class. The aim is to detect credit card frauds such as to reduce fraud rate so that genuine customers do not have to pay for the item which is not purchased by them. Dataset is consist of numerical data only in altogether 30 features, in which 28 features has been transformed through PCA. Much information about the dataset is not obtained due to the security issues. ‘Class' is the dependent variable, consisting of two features -1 (fraud) and 0 (genuine).

**2.3Hybridization Pre-processing and Resampling Techniques (HPRT): Overview**

The HPRT, algorithm proposed during this work is a new hybridization pre-processing and resampling technique (HPRT) – a single solution solves two major issues - (a) the algorithm cleans the dataset (b) it balances the dataset. Our algorithm is designed to remove redundancy and outliers from a dataset thus cleaning it, at the same time reducing majority sample while increasing minority sample and balances the dataset. It is a pre-processing algorithm, contained several steps, which automatically detect the intensity of imbalance dataset and then convert it into several steps. The key point of this algorithm is a resampling of the dataset while reducing the majority class through redundancy and outlier’s reduction because almost all real word dataset contains duplicate as well as outliers records. Therefore eliminating this from dataset will clean the data as well as to some extent help to balance it without any information loss. Next step of the algorithm will check and amplify the minor class by adding a synthetic feature to it which brings normal distribution between both the classes. HPRT enhances the performance of machine learning classification algorithms by reducing the complexity of a dataset. Several experiments were conducted on credit card transaction dataset, selected as case study during current research and observe positive impact of HPRT on the performances of four different classification algorithms.



**Figure 2: Workflow of HPRT for reducing complexity while cleaning of an imbalanced credit card transaction dataset**

**2.4 Method & Steps incorporated: A New Hybridization Preprocessing and Resampling Technique (HPRT)**

1. HPRT accepts original imbalance dataset as an input. The dataset than undergoes through several steps, wherein each step complexity of the large dataset has been tried to reduce.

2. In Figure 2 it is clearly seen, an algorithm accepts, big or large dataset as an input and then divide it into two subparts, such that 1st part contains all the features from majority class.

3. Another part contains all the features from minority class. In this way two subsets are derived from the original dataset are N and F where N is consist of only majority features and F is consist of minority features.

4 & 5. In the next step, we individually apply PCA algorithm on both the subset to form P1and P2. This is a first phase where the complexity of the dataset is reduced through the dimensionality reduction.

6. In the next step, we will take only subset P1 {V1, V2, V3,….Vn} and divide it into various clusters with the help of K-mean unsupervised machine learning algorithms. K-means is capable of dividing observations in datasets into K different clusters. Clusters can be defined as the entities of similar group i.e. features in one cluster are identical that the features in the other clusters. With the help of K-mean clustering algorithm P1 {V1, V2, V3….Vn} is divided into P1{ki, kj,…kn}.The cluster is developed to tie similar items in one group so that the number of comparisons will decrease to an enormous extent and which is helpful in reducing the processing time of big and large dataset.

7. Redundancy is checked in each clusters using divide and conquer rule. Duplicate record detected is immediately discarded from the cluster.

8. After the removal of unwanted redundant data from the clusters P1{ki, kj, …..kn}, algorithm proceed with the detection of outliers. IQR method is used for the detection of extreme outliers. The threshold is calculated by multiply 1QR with 1.5, upper bound and lower bound is calculated. A feature is considered an outlier if its value is less than lower bound or more than the upper bound. All the outlier is stored in temporary list O and discarded at last from the cluster. At this stage, our clusters P1 {V1, V2, …Vn} is free from outliers as well as duplicates data and then the cluster is again converted into a data frame. Half of our intension to reduce the majority class had been achieved through these steps. Hence this reduction in the majority class is achieved without any type of information loss.

9. In the next step, we concatenate both subset P1 {V1, V2, V3….Vn} and P2 {V1, V2, V3….Vn} to form D {V1, V2, V3….Vn}.

10 & 11. In this step we divide dataset into training and test set, where training set contain 80% of data and test set contain 20% of data.

12. In the next step SMOTE over sampling is used to balance minority class of the dataset. In this process features in minority class are increased by adding artificial sample to the minority class with the help of the KNN algorithm and Euclidean Distance. SMOTE oversampling is performed during cross-validation to avoid overfitting because, smote if applied earlier there can be a possibility of adding features exactly the same in the validation set, hence causing the problem of data leakage. To avoid this, validation set must be excluded first with other training set and then during cross validation over sampling should be applied on rest of the dataset. This will avoid over fitting as well as data leakage problem because during testing phase validation set will be completely unseen.

13. Model is constructed on a balanced dataset through machine learning algorithm and evaluated further with test data.

**2.5 Technique: A New Hybridization Preprocessing and Resampling Technique (HPRT)**

**Input:** Imbalanced Big Dataset C {V1, V2, V3, …,Vn}

**Output:** Predictive Classification Model Constructed on Balanced Dataset **Reducing the complexity of a dataset through Redundancy and Outlier Reduction from the majority class**

* **Step 1:** Load dataset C, divide it into two parts (N, F)

F Ꞓ C **// F consist only Minority Class**

N Ꞓ C**// N consist only Majority Class**

* **Step 2:**Reduce the dimensionality of both N and F separately through the application of PCA
* **Standardize dataset N and F so that values of all the features V1, V2, V3,..Vnexists within 0-1without changing its original meaning.**

N = Standardize (N {V1, V2, V3,..Vn })

F = Standardize (F {V1, V2, V3,..Vn})

* **Calculatecovariance matrix for given set of data**  
  V1 = Cov (N {{V1, V2, V3,..Vn })   
  V2 = Cov (F {{V1, V2, V3,..Vn }
* **Calculate Eigen values and Eigen vectors for matrix V1 and V2**  
  E1, E2 = Eigen (V1)  
  E3, E4 = Eigen (V2)  
  Z1 = Select K1, Eigen values having K, largest Eigen values (E1, E2)  
  Z2 = Select K2, Eigen values having K, largest Eigen values (E3, E4)
* **Project data by multiplying the original matrix with its transpose.**  
  P1 = Z1T . N  
  P2 = Z2T . N

Where,  
P1= N {V1, V2, V3, …Vn}  
P2 = F {V1, V2, V3, …Vn}

* **Step 3:** Apply K- Means clustering algorithm on P1 to divide it into kn number of cluster.
  + P1= P1 {k1, k2,…kn}
* Select centroid for P1 {C1, C2, … Cn} for each of the cluster P1{k1, k2, ….. kn} in P1.
* Calculate the distance between each vector in a cluster P1 {k1, k2, ….. kn } and search for closest centroid.
* Evaluate new centroid for each cluster k1, k2, …..kn applying divide and conquer rule.
* Repeat above process 2, 3, 4 till centroid values for P1 {k1, k2,…kn} becomes constant.
* **Step 4:**Dropping redundant data from each of the clusters in P1 {k1, k2, ….. kn}
* Comparison for detection of redundant data points within cluster P1{k1, k2, ….. kn}

For i = 1 to kn

for j = 1 to number of data points X in ki

Search for redundant records using divide and conquer rule and store in R.

R = Duplicate features Xi //Achieved P1{k1, k2, …..kn} after redundancy reduction

Discard R form the clusters

end for

end for

* **Step 5:**Dropping extreme outliers feature X from each cluster P1 {k1, k2, ….kn} using IQR
* Sort each cluster P1{k1, k2, ….. kn} in ascending order
* Calculate inter quartile range within each of cluster

P1 { k1 { X1, X2, x3, …Xn}, k2{ X1,X2,X3,…,Xn},…kn { X1, X2, X3, ……Xn}}  
 for i = 1 to kn  
for j = 1 to Xn number of data point  
//Calculate 25 %( q1) and 75% (q2) of data in each cluster ki  
iqr = q2 – q1  
Threshold = iqr \* 1.5

l = q1 – threshold //Calculating upper threshold in each cluster ki  
u = q2 – threshold//Calculating lower threshold in each cluster ki  
for each data point Xj in cluster ki   
Calculating upper bound (u) and lower bound (l)  
end for  
for j = 1 to Xn number of data point  
If Xj< l or Xj> u  
O = Xj //Store extreme outliers in O

//Achieved P1 {k1, k2, …..kn} after dropping outliers

Drop O  
 end for  
 end for  
end for

* **Step 6**: Almost 30%-40% feature reduced from P1 {ki, kj, …..kn} after removal of redundancy and outliers without losing much information. P1 {ki, kj, …..kn} is converted to data frame again in its original format such as P1 { V1, V2, V3, ……Vn}.

D Ξ P1 + P2

D is dataset after concatenation of P1{V1, V2, V3, ……Vn} and P2 { V1, V2, V3, ……Vn}

**Application of SMOTE Over Sampling to Majority Class to balance the Dataset**

* **Step 7**: Split D {V1, V2, V3, ……Vn} into X1 and Y1 such that X1 is consist of an entire set of data excluding target class and Y1 is consist of target class only.

X1 = D. Drop {target class}

Y1 = D {target class}

* **Step 8**: Using SMOTE during cross validation for adding synthetic features in minority class

For train in split XTrain, YTrain

P = SMOTE (XTrain, YTrain)

M = P.Classifier (XTrain,[train] ,YTrain[train])

Where, M is a model constructed using P(training set containing artificial data point for majority class) with the help of ML Classifier.

**2.6 Comparative Studies to Evaluate HRPT Performance**

Several experiment conducted to evaluate the performance of HPRT with number of machine learning algorithm during building model by considering sample set of credit card transaction dataset for fraud prediction. Result evaluated confirm that HPRT boost up the performance of traditional machine learning algorithm.

**Table 1: Comparative Analysis of Combinational Effect of HPRT with several Machine Learning Classification Model for Credit Card Fraud Detection**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Matrices & Measures** | **Logistic Regression** | **KNN** | **Neural Net** | **Decision Tree** |
| Accuracy | 0.98% | 0.99% | 99% | 0.99% |
| Precision | 0.11% | 0.48% | 100% | 0.38% |
| Recall | 0.92% | 0.91% | 85% | 0.92% |
| F1-Score | 0.21% | 0.58% | 89% | 0.48% |
| Misclassification | 0.018% | 0.014% | 0.02% | 0.0016% |

**2.7 Claim**

HPRT, a new Hybridization Preprocessing and Resampling Technique, has been develop to solves several issues of messy data, such as it reduces redundancy and outliers from a dataset and automatically balances the dataset, thus reducing its complexity and enhancing the performance of ML classifier. Several outcomes of this research are as follow:

1. HPRT, a new Hybridization Preprocessing and Resampling Technique, act as a perfect single data management solution having a capability of cleaning the data simultaneously while reducing the complexity of an imbalanced dataset.
2. HRPT, act as preprocessing algorithm , able to enhance the performances of several traditional machine learning algorithms , which has been proved through the result of several experiments performed during study.
3. HPRT automates the process of resampling for big datasets by detecting its level of imbalance where redundancy and outliers also removes automatically.

Therefore I claim that the work discussed here is an enhanced hybrid technique of undersampling and oversampling to remove the difficulties possess by them. It is a preprocessing activity for enhancing the performance of machine learning algorithms with no important information loss.

**References:**

[1]ViashaluGanganwar,“Overview of classification algorithms for imbalanced datasets”, *International Journal of Emerging Technology and Advanced Engineering*, Vol.2, Issued: 4th April 20s12. Accessed:   
<https://pdfs.semanticscholar.org/239b/2210b3fbc1f4b8246437a88a668bf9a0d2c0.pdf>

[2]K.Madasamy , et al, Data Imbalance and Classifiers: Impact and Solutions from a Big Data Perspective”, *International Journal of Computational Intelligence Research*, ISSN 0973-1873, Vol.13, 2017, Accessed:<https://www.ripublication.com/ijcir17/ijcirv13n9_09.pdf>

[3] AkilaSomasundaram, et al,“Data Imbalance: Effects and Solutions for Classification of Large and Highly Imbalanced Data”, Available: January 2016. Accessed:   
<https://www.researchgate.net/publication/320895020_Data_Imbalance_Effects_and_Solutions_for_Classification_of_Large_and_Highly_Imbalanced_Data>

# [4]Nitesh , et al., “Data Mining for Imbalanced Datasets: An Overview”, *Data Mining and Knowledge Discovery Handbook,*Online: 7th July 2017. Accessed: <https://link.springer.com/chapter/10.1007/978-0-387-09823-4_45>

[5]Nitesh V. Chawla, et al, W. Philip Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique”, *Journal of Artificial Intelligence Research 16* (2002),   
<https://arxiv.org/pdf/1106.1813.pdf>

[6]Hui Han , et al, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, *Department of Automation, Tsinghua University, Beijing; Department of Statistics, Central University of Finance and Economics, Beijing.* Accessed:  
<https://sci2s.ugr.es/keel/keel-dataset/pdfs/2005-Han-LNCS.pdf>

[7]Jingke Xi, “Outlier detection algorithm in data mining’, *IEEE Xplore*, 20-22 Dec. 2008. Accessed: <https://ieeexplore.ieee.org/document/4739542>

# [9]Mrs. Radha R, et al. , “REMOVAL OF REDUNDANT AND IRRELEVANT DATA FROM TRAINING DATASETS USING SPEEDY FEATURE SELECTION METHOD”, *International Journal of Computer Science and Mobile Computing*, Vol.5, July 2016. Accessed: <https://www.ijcsmc.com/docs/papers/July2016/V5I7201670.pdf>

# [9]Ruchi Sharma , et al, ,“Improving the Accuracy and Reducing the Redundancy in Data Mining”, *International Journal of Engineering Science and Computing,* Vol.6, May 2016*.* Accessed: <http://ijesc.org/upload/7d1f711274530fdef9dc9a284e85ea20.Improving%20the%20Accuracy%20and%20Reducing%20the%20Redundancy%20in%20Data%20Mining.pdf>