

Research Journals (https://www.gyanvihar.org/researchjournals/)

### CLUSTER-BASED CLASSIFICATION TO PROCESS IMBALANCED DATA USING SMOTE TECHNIQUE

Home / Volume 7, Issue 1, 2021 /

Cluster-Based Classification To Process Imbalanced Data Using SMOTE Technique

#### CLUSTER-BASED CLASSIFICATION TO PROCESS IMBALANCED DATA USING SMOTE TECHNIQUE

VOLUME 7, ISSUE 1, 2021

(HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/CATEGORY/INTERNATIONAL-JOURNAL-OF-CONVERGING-TECHNOLOGIES-MANAGEMENT/VOLUME-7-ISSUE-1-2021-INTERNATIONAL-JOURNAL-OF-CONVERGING-TECHNOLOGIES-MANAGEMENT/)

Arbind Kumar Chaurasia

Research Scholar(M.Tech), Department of Computer Science & Engineering

Suresh Gyan Vihar University, Jaipur, Rajasthan, India

chau.arbind@gmail.com (mailto:chau.arbind@gmail.com)

#### Abstract

Technical advances in the digitally linked and controlled period, big terabyte data were generated and even more. Data repositories have been tremendously upgraded because different organizations like the government generate data cooperates, caring for health in large quantities. Online generation, processing, collection and analysis of large amounts of information are under way. It demands that this data be translated into useful knowledge. This process is called data mining for extracting information from a large number of data. We refer to it as research phase in the exploration of information in the database process. The research presented in this paper focuses on issues in the detection of class imbalances and on the improved detection of minority class cases by application of mixed data sampling approaches. The following extensions will be applied to improve the class difference definition of class problems and to more correctly classify minority class instances.

Keywords- Cluster, Classification, Imbalance data, Analysis, Prediction

#### 1. INTRODUCTION

Information mining consists of the information, knowledge and machine learning intersection processes. The numerical data relationship study shows statistics. Statistics are used for data quantification and data classification. Artificial intelligence deals with the development of artificial computer intelligence. In this artificial machine the human intelligence is artificial, such that as humans, if the same thing occurs, they will benefit from their past experience and apply it in the future. Machine learning has algorithms that learn and predict accordingly from certain data.

Data mining consist of various approaches and techniques such as Classification, Clustering, and Prediction, Association rule learning, Anomaly detection, Regression, Time series analysis and Summarization.

Machine Learning is the field of study in computer science that gives capability to computers to learn by itself. Machine Learning is all about writing software that can learn from past experience without being explicitly programmed. In this computers are made intelligent like humans and developing "Ability to Learn" in computers. As humans learn from their past experience and they implement it in future just like machines are also made intelligent enough such that they learn from their past experience and if same problem arrives in future and they should be capable to solve that problem without any human intervention. Machine learning is all about building algorithms which receives input data and use statistical analysis to predict the output.

In today's global era Machine Learning is used by many web-based companies to boost their recommendation engines. Facebook decides what to show on our news-feed according to the past activity we had done on Facebook. Also Netflix, Amazon Prime suggest movies we might want to watch, this all recommendations are based on predictions that are made out of patterns from our past activities on Netflix and Amazon Prime.

Machine Learning algorithms are often categorized into two types -Supervised Learning Algorithm and Unsupervised Learning Algorithm.

#### 2. LITERATURE REVIEW

In his article on disjuncture, Holteet al. (1989) proposed. All examples in a dataset should be meaningful disjoints. Disjunct size can be defined as the number of examples of training correctly classified during the classification process. There is a misclassification problem if any dataset consists of small datasets, and a small breakdown can be prone to error.

In his article, Japkowicz (2000) explored the social inequality. He experimentally demonstrated that class disparity problems impair the classification results, as such applications include the identification of satellite radar oil spills, the identification of illegal telephone calls and the tracking of in-flight helicopter gearbox errors as they consist of classes where a class has more and a separate class less. The minority class is of primary interest to us but it will be more class oriented towards the class with more majority class instances and will not classify minority class instances properly.

Pratietal .( 2004) examined learning of skews and slight breakdowns. Class skews are a class discrepancy where one class has outnumbered other class instances and disjuncts instances that are categorized correctly from the training dataset. This paper shows that a small breakdown has higher error rates than the large breakdown, and how the small breakdowns and skews of class affect one another. For the solution paper the results are contrasted with pruned and unpruned trees and other resampling techniques are also used. It shows that the over-sample and under sample techniques can be used with both small disjunct and class imbalances.

Guoetal .( 2008) checked all available methods for class desequilibrium problems. First method examined is data sampling methods to adjust the class distribution of the oversamples and sub-samples, with the use of certain strategies increasing oversampling minority class instances and reducing the majority class instances. Another approach evaluated by authors is cost-delicate learning, in which decision taking costs are used to enhance the efficiency of the classification. Another technique is one class learning, in which the target class alone is used for model creation. Another technique includes bagging and boosting, which combine the results of several classifications.

#### 3. PROPOSED WORK

#### **3.1 COST SENSITIVE LEARNING**

Sun et al . (2006) addressed that if we find C (Min, Maj) to be the cost of misclassification of large class cases as a minority and C (Maj and Min) for any double-class problem, then we find C as the price for misclassification of minority class instances as a majority class and class disorder problems. Cost-sensitive learning thus reduces the classification coat to a minimum.

Garcia et al .( 2007) provides a summary of costly learning, because this kind of learning takes into account the costs of misclassifying instances. A cost matrix is constructed in this type of learning in which the numerical representation of the cost for the number of malclassified cases is specified.

#### **Table1: Cost Sensitive Matrix**

Actual	Predicted	
	Positive Class	Negative Class
Positive Class	True Positive(Tp) Or C(+,+)	False Negative(Fn) Or C(+,-)
Negative Class	False Positive(Fp) Or C(-,+)	True Negative(Tn) Or C(-,-)

#### The cost matrix of two classes is defined by Table 1.

Cost-sensitive learning takes into account cost-sensitive matrix and reduces misclassification costs by using the following:

- Change the weight of instances in the space of the data.
- Build a specific classification algorithm that adjusts its configuration to the weight of instances.
- Using the Bayes Risk principle, each instance will be graded more appropriately by the lowest risk level.

#### **3.2 SMOTE WITH ONE SIDED SELECTION**

SMOTE with one-way filtering is another alternative method to test the feasibility of the methodology we have suggested. In this SMOTE firstly, the weight of minority groups is over-sampled and then One Sided Selection is a strategy for the decline of majority class circumstances. A side-sided choice is a method of data cleaning that combines Tomek Links and the closest condensed person. First Tomek Links is used in One Side Selection, removing inconsistent instances such as noisy and limit instances. In the removal of redundant instances is the Condensed Nearest Neighbor technique.

The parameter used is random seed equal to zero for Tomek Links and the Condensed Nearest neighbor. Thus, one-faced filtering improves classification efficiency as it excludes distracting, minimal and redundant instances.

#### **3.3 PROPOSED METHODOLOGY**

In order to classify a majority or minority case, the proposed technology for balance between imbalanced data and the hybrid approach is proposed with the combination of oversampling and under sampling. The suggested process structure comprises the following steps:

- 1. Unbalanced evidence were broken down into preparation and evaluations.
- 2. The data collection for instruction is then then broken down into the groups of the majority and minority.
- 3. SMOTE and suggested under-sampling methods then refer to the minority class to the majority class.
- 4. If the majority and minority classes have been over-sampled, the majority and minority groups must combine into a new integrated training program.
- 5. The balanced data set of the SVM classifier is used to train the model and a test dataset is used on the

model to evaluate the classifier 's performance

The flowchart shows that we have an imbalanced dataset, which is split into testing and evaluation datasets. The Scikit-Learn learning machine has a feature in which the dataset can be divided into training and test results. The task allows the ratio parameter to be split into training and test data set with what ratio dataset. The dataset testing should have more instances or the same number of instances as the reference dataset. Nonetheless, we consider more cases in testing datasets than in the test dataset as the data pre-processing method in the training dataset is implemented and then the SVM model can be developed and the output of the SVM model test data set tested.

We have a way to split data into test and testing datasets in Scikit Learn machine learning application. We have the parameter test size equal to 0.2. In all cases, 80 percent of the datasets are segregated into a test data set and 20 percent of the dataset is evaluated.

Following the separation between preparation and evaluation results, we split workout results between majority and minority groups. The more numerous class is the majority class and the less numerical class is the minority class.

We will then implement our suggested cluster distance under sampling technique (CDBU), which eliminates noisy instances, boundary instances, outliers and unnecessary instances. In the majority level, our approach is the under sampling technique. Then we add an SMOTE (Sinthetic Minority Oversampling) strategy in the minority community. After the methodology of data collection and then a mixed data sample, we merge the majority and the minority groups. Then we construct the SVM model using a balanced dataset and analyze SVM models' output using a testing dataset using parameters such as sensitivity, precision, G-mean and AUC.

#### 4. CONCLUSION

The platform is used in Python with the machine learning method Scikit-Learn. Scikit-Learning Machine Learning Framework uses SVM as the classification algorithm for the analysis of the suggested hybrid solution. Different imbalances have been used in twelve data kits in UCI and NASA MDP repository. Compare the performance of the hybrid approach with existing hybrid SMOTE (Chawla et al., 2002), Random under sample SMOTE (Agrawal et al., 2015), Tomek Links SMOTE (Batista et al., 2004) and One Sided Selection SMOTE (Pristyantoet al., 2018) approaches. Sensitivity, precision, G-meaning and AUC are the performance metrics used.

#### 5. References

- Acuña, E. and Rodríguez, C. 2005. An empirical study of the effect of outliers on the misclassification error rate. *IEEE Transactions on Knowledge and Data Engineering***17**: 1-21.
- Agrawal, A., Viktor, H. and Paquet, E. 2015. SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Under sampling. *In: Proceedings of 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* held at Lisbon during November 12-14, 2015, pp. 226-234.
- Asuncion, A. and Newman, D. J. UCI Machine Learning Repository. *Journal of Intelligent Learning Systems and Applications***2**: 1-10.
- Barandela, R., Sánchez, J. S., García, V. and Rangel, E. 2003. Strategies for learning in class imbalance problems. *Elsevier Journal of Pattern Recognition***36**:849–851.

- Batista, G., Bazzan, B. and Monard, M. 2003. Balancing Training Data for Automated Annotation of Keywords: a Case Study. *In: Proceedings of II Brazilian Workshop on Bioinformatics* held at Brazil during December 3-5, 2003, pp. 35-43.
- Batista, G., Prati, R. C. and Monard, M. C. 2004. A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. ACMSIGKDD Explorations 6:20-29.
- Bunkhumpornpat, C., Sinapiromsaran, K. and Lursinsap, C. 2009. Safe-level-SMOTE: safe levelsynthetic minority over-sampling Technique for handling the class imbalanced problem. *In: Proceedings* of 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining held at Bangkok during April 27-30, 2009, pp. 475-482.
- Cao, L. and Zhai, Y. 2015. Imbalanced Data Classification Based on a Hybrid Resampling SVM Method.
  *In: Proceedings of 12<sup>th</sup> IEEE International Conference on Ubiquitous Intelligence and Computing* held at
  Beijing during August 10-14, 2015, pp. 1533-1536.
- Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research***16**: 321-357.
- Chen, M., Chen, L., Hsu, C., and Zeng, W. 2008. An information granulation based data mining approach for classifying imbalanced data. *Information Sciences***78**: 3214-3227.
- Debray, T. 2009. Classification of Imbalanced Data Sets. Master's Thesis in Artificial Intelligence submitted to Faculty of Humanities and Sciences, Maastricht University Netherlands.
- Das, B., Krishnan, N.C. and Cook, D.J. 2014. Handling Imbalanced and Overlapping Classes in Smart Environments Prompting Dataset. Studies in Big Data, Berlin pp. 199-219.
- Elhassan, A.T., Aljourf, M., Mohanna, F. and Shoukri, M. (2016). Classification of Imbalance Data using Tomek Link Combined with Random Under-sampling as a Data Reduction Method. *Global Journal of Technology and Engineering***7**:1-12.

- Estabrooks, A., Jo, T. and Japkowicz, N. 2004. Multiple Resampling Method for Learning from Imbalanced Data Computational Intelligence20: 18-36.
- Fagan, M. 1986. Advances in Software Inspections. IEEE Transaction on Software Engineering 12: 744-751.
- Farquad, M.A.H. and Bose, I. 2012. Preprocessing unbalanced data using support vector machine.

Decision Support Systems 53: 226-233.

#### SHARE :

f (http://www.facebook.com/sharer.php?
 u=https://www.gyanvihar.org/journals/index.php/2021/05/17/cluster-based-classification-to-process-imbalanced-data-using-smote-technique-2/&t=Cluster-based Classification to Process Imbalanced Data using SMOTE Technique) 
 ✓ (http://twitter.com/home/?status=Cluster-based Classification to Process Imbalanced Data using SMOTE Technique - https://www.gyanvihar.org/journals/index.php/2021/05/17/cluster-based-classification to Process Imbalanced Data using SMOTE Technique - https://www.gyanvihar.org/journals/index.php/2021/05/17/cluster-based-classification-to-process-imbalanced-data-using-smote-technique-2/)

Type here to search

#### RECENT POSTS

STUDY OF HYBRID FRP COMPOSITE MADE OF FLAX AND HEMP FIBERS WITH KEVLAR (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/05/17/STUDY-OF-HYBRID-FRP-COMPOSITE-MADE-OF-FLAX-AND-HEMP-FIBERS-WITH-KEVLAR-2/)

DOCKERIZING VIRTUAL WIRELESS SENSOR NETWORK(VWSN):ISSUES AND SOLUTIONS (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/05/17/DOCKERIZING-VIRTUAL-WIRELESS-SENSOR-NETWORKVWSNISSUES-AND-SOLUTIONS/)

CLUSTER-BASED CLASSIFICATION TO PROCESS IMBALANCED DATA USING SMOTE TECHNIQUE (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/05/17/CLUSTER-BASED-CLASSIFICATION-TO-PROCESS-IMBALANCED-DATA-USING-SMOTE-TECHNIQUE-2/)

ULTRAWIDE BAND ANTENNAS BANDWIDTH INCREASING TECHNIQUE (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/05/17/ULTRAWIDE-BAND-ANTENNAS-BANDWIDTH-INCREASING-TECHNIQUE/)

DISCRETE WAVELET TRANSFORM AND DECISION RULES SUPPORTED METHOD FOR DETECTION OF FAULTY EVENTS IN DISTRIBUTION NETWORK WITH RENEWABLE ENERGY (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/05/17/DISCRETE-WAVELET-

Q

TRANSFORM-AND-DECISION-RULES-SUPPORTED-METHOD-FOR-DETECTION-OF-FAULTY-EVENTS-IN-DISTRIBUTION-NETWORK-WITH-RENEWABLE-ENERGY/)

COMPARING DIFFERENT MODELS FOR CREDIT CARD FRAUD DETECTION (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/04/23/COMPARING-DIFFERENT-MODELS-FOR-CREDIT-CARD-FRAUD-DETECTION/)

STATIC LOAD ANALYSIS OF SIDE UNDERRUN PROTECTION DEVICE FOR HEAVY COMMERCIAL VEHICLES USING HYPERMESH TOOL

(HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/04/23/STATIC-LOAD-ANALYSIS-OF-SIDE-UNDERRUN-PROTECTION-DEVICE-FOR-HEAVY-COMMERCIAL-VEHICLES-USING-HYPERMESH-TOOL/)

COMPARATIVE STUDY OF EXISTING DOCKER AUDIT TOOLS FOR PERFORMING VULNERABILITY DISCOVERY AND AUDITING TASKS IN DOCKER. (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/04/23/COMPARATIVE-STUDY-OF-EXISTING-DOCKER-AUDIT-TOOLS-FOR-PERFORMING-VULNERABILITY-DISCOVERY-AND-AUDITING-TASKS-IN-DOCKER/)

SOFTWARE VULNERABILITY DEPENDENT ON MACHINE LEARNING (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/04/23/SOFTWARE-VULNERABILITY-DEPENDENT-ON-MACHINE-LEARNING/)

ANALYSIS OF SURFACE FINISHING PROCESSES-A REVIEW (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/04/23/ANALYSIS-OF-SURFACE-FINISHING-PROCESSES-A-REVIEW/)

REVIEW PAPER ON HOW TO RECOVER NODE FAILURE USING (MULTI-PROTOCOL LABEL SWITCHING (MPLS) NETWORK AND ENHANCE THE ENERGY OF SIGNAL (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/02/05/REVIEW-PAPER-ON-HOW-TO-RECOVER-NODE-FAILURE-USING-MULTI-PROTOCOL-LABEL-SWITCHING-MPLS-NETWORK-AND-ENHANCE-THE-ENERGY-OF-SIGNAL/)

TO EXPLORE THE ENERGY OF WIRELESS SENSOR NETWORKS TO SOLVE BOTTLENECK PROBLEM USING SWARM ROUTING APPROACH: A REVIEW (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/02/05/TO-EXPLORE-THE-ENERGY-OF-WIRELESS-SENSOR-NETWORKS-TO-SOLVE-BOTTLENECK-PROBLEM-USING-SWARM-ROUTING-APPROACH-A-REVIEW/)

SECURITY AND PRIVACY CHALLENGES IN INTERNET OF EVERYTHING (IOE) WITH SECURITY REQUIREMENTS

(HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/01/29/SECURITY-AND-PRIVACY-CHALLENGES-IN-INTERNET-OF-EVERYTHING-IOE-WITH-SECURITY-REQUIREMENTS/)

THE IMPACT OF EXTREMOPHILIC ENZYME: WIDELY APPLICATION IN INDUSTRIAL USES (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/01/24/THE-IMPACT-OF-EXTREMOPHILIC-ENZYME-WIDELY-APPLICATION-IN-INDUSTRIAL-USES/)

EVALUATION OF PHYSICO-CHEMICAL PARAMETERS OF DRINKING WATER AT VARIOUS SITES OF KOTA, RAJASTHAN

(HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/01/24/EVALUATION-OF-PHYSICO-CHEMICAL-PARAMETERS-OF-DRINKING-WATER-AT-VARIOUS-SITES-OF-KOTA-RAJASTHAN/)

PHARMACOLOGICAL POTENTIAL OF CURCUMA CAESIA (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/01/24/PHARMACOLOGICAL-POTENTIAL-OF-CURCUMA-CAESIA/)

PLANT-PARASITIC NEMATODES AND THEIR MANAGEMENT (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/01/24/PLANT-PARASITIC-NEMATODES-AND-THEIR-MANAGEMENT/)

ANALYTICAL AND KINETIC INVESTIGATION OF AZO DYE ARSENAZO(III) AS COPPER(II) DERIVATIVE BY USING SPECTROSCOPIC METHODS (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/01/24/ANALYTICAL-AND-KINETIC-INVESTIGATION-OF-AZO-DYE-ARSENAZOIII-AS-COPPERII-DERIVATIVE-BY-USING-SPECTROSCOPIC-METHODS/)

SOURCES AND BIOLOGICAL ACTIVITY OF COUMARINS: AN APPRAISAL (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/01/24/SOURCES-AND-BIOLOGICAL-ACTIVITY-OF-COUMARINS-AN-APPRAISAL/)

PRE- AND POST-MONSOON VARIATION IN PHYSICO-CHEMICAL CHARACTERISTICS IN GROUNDWATER QUALITY OF AMER TEHSIL IN JAIPUR, RAJASTHAN (HTTPS://WWW.GYANVIHAR.ORG/JOURNALS/INDEX.PHP/2021/01/23/PRE-AND-POST-MONSOON-VARIATION-IN-PHYSICO-CHEMICAL-CHARACTERISTICS-IN-GROUNDWATER-QUALITY-OF-AMER-TEHSIL-IN-JAIPUR-RAJASTHAN/)

\_\_\_\_\_



(HTTP://WWW.GYANVIHAR.ORG/RESEARCHJOURNALS/ENGG.PHP)

ISSN: 2394-9570

# SURESH GYAN VIHAR UNIVERSITY INTERNATIONAL JOURNAL OF ENVIRONMENT SCIENCE AND TECHNOLOGY



(HTTP://WWW.GYANVIHAR.ORG/RESEARCHJOURNALS/ENVIRMENTAL\_SCIENCE.PHP)



#### (HTTP://WWW.GYANVIHAR.ORG/RESEARCHJOURNALS/CTM\_JOURNALS.PHP)

# Suresh Gyan Vihar University Journal of Climate Change and Water



VOLUME 1 | MILE 1 | ANUARY MILE



(HTTP://WWW.GYANVIHAR.ORG/RESEARCHJOURNALS/INT\_JOU\_CLIMATE\_CHANGE\_WATER.P

6.06 J378 J478



## Suresh Gyan Vihar University International Journal of Economics and Management



VOLUME 1 | ISSUE 1 | JANUARY 2014



(HTTP://WWW.GYANVIHAR.ORG/RESEARCHJOURNALS/ECO\_MANG\_JOUR.PHP)

**ISSN-XXXX-XXXX** 

## SURESH GYAN VIHAR UNIVERSITY JOURNAL OF PHARMACEUTICAL RESEARCH & EDUCATION

### VOLUME-I | ISSUE-I | JULY 2016



(HTTP://WWW.GYANVIHAR.ORG/RESEARCHJOURNALS/PHARMACY.PHP)

COPYRIGHT © 2018 BY TEAM SGVU. ALL RIGHTS RESERVED.