IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-3, Issue-5 E-ISSN: 2347-2693

Analysis of Higher Education System Using Data Mining Techniques Using Efficient Machine Learning in Class Imbalance Data (EML-CID)

Dr.R.Periyasamy[1], P.Veeramuthu[2*]

[1]*Associate Professor, PG and Department of Computer Science, Nehru Memoriyal College, Puthanampatti*
[2] *Research Scholor, PG and Department of Computer Science, Nehru Memoriyal College, Puthanampatti*

Rpsamy62@gmail.com , er.veera86@gmail.com

www.ijcseonline.org

Received: May/02/2015 Revised: May/10//2015 Accepted: May/24/2015 Published: May/30/ 2015

***Abstract*—** In modern world a huge amount of data is available which can be used effectively to produce vital information. The information achieved can be used in the field of Medical science, Education, Business, Agriculture and so on. As huge amount of data is being collected and stored in the databases, traditional statistical techniques and database management tools are no longer adequate for analyzing this huge amount of data. Data Mining (sometimes called data or knowledge discovery) has become the area of growing significance because it helps in analyzing data from different perspectives and summarizing it into useful information. There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational environments. The data can be collected from various educational institutes that reside in their databases. The data can be personal or academic which can be used to understand students' behavior, to assist instructors, to improve teaching, to evaluate and improve e-learning systems, to improve curriculums and many other benefits. It neighbor, naive bayes, support vector machines and many others. Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for organization of syllabus, prediction regarding enrolment of students in a particular programme, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students and so on.

***Keywords*—**Datamining, Machine Learning, Class Imbalance

I. INTRODUCTION *(HEADING 1)*

The Machine Learning Algorithm is designed for neural network in prediction of complexity data. At first the neural network is familiarly developed by using Forward Selection and Back Propagation method to construct the network for the past decades to provide the efficient optimal solution for local minimum problem and computational cost. The neural network is then trained by the SVM Support Vector Machine or Super Vector Machine is consider as the traditional approach for neural network to avoid the local minimum problem and computational cost, the SVM act as a classifier and provide benefits by reducing Structural Risk. The SVM due to its simplicity and efficiency it is applied in various domain to reduce computational cost. Later The Extreme Machine Learning Algorithm is used in Neural Network as a Classifier, The EML algorithm is used to update weights from input layer to hidden layer and to the output layer. Later the EML is considered as the best algorithm in performance generalization for Neural Network. The ELM Algorithm is used in solving multiclass classification problems. The predicting accuracy of the ELM is more compared to the SVM algorithm. So the new algorithm is formed by the combination of SVM and ELM as ESVM effective and famous then the basic model ELM and SVM. The ELM is used to handle both the labelled data and unlabeled data where the labelled data is named as supervised data and the unlabeled data is named as unsupervised data. The Supervised data is used to ELM to form SS-ELM and the unsupervised data as US-ELM which are more familiar in solving multi class classification, class imbalance data. Since the SS-ELM and US-ELM consumes more computational cost. The unified framework is to be constructed with efficient and accurate prediction.

ELM (SS-ELM) and the unsupervised ELM (US-ELM) show the learning capability and computational efficiency of ELMs and can solve effectively in multi-class classification or multi-cluster clustering and unseen data. Moreover, it is shown in this paper that all the semi-supervised and unsupervised ELMs can actually be put into a unified framework as EML-CID. The EML-CID algorithm provides new perspectives for understanding the mechanism of Class Imbalance and Data Prediction which is the key concept in ELM theory.

II. SCOPE OF THE RESEARCH

The class imbalance problem occurs by presence of small size of minority class and class differences. The minority class allows the learner ability to reduce the discovering patterns. The increase of minority class helps the learner to predict interesting discovering patterns. The minority class is improved by using sampling method. Sampling means a change or alter in data or records of minority class. The Sampling is of three types they are Under Sampling, Over Sampling and combination of both. The under sampling is done by reducing the data from majority class, the over sampling is done by increasing the data from minority class. The other method is combined both where the records or removed from majority class and replicate or increase data in minority class. By using this methods the original data gets alter and leads reduces prediction in accuracy. Therefore the EML-CID is formed as a unified framework of Supervised and Unsupervised data to solve the Class Imbalance and Data Prediction Problem. The EML-CID aims in achieving are as follows

1. To investigate the use of open source data mining software to conflict fraud, with particular attention to fraud/non-fraudulent cases in Class classification system.

2. To identify the type of problems that arise when taking a data mining approach to fraud detection, particularly the class imbalance problem and Data Prediction.

3. To identify solutions to the class imbalance problem and show how they can be implemented through the use of open source software.

III. METHODOLOGY

In proposed method EML-CID that collect the information from the dataset, From the Dataset the EML-CID perform the clustering operation using the K-means clustering that implement to find the weightage of minority class samples, the weightage of is calculated based on the Euclidean distance from majority class samples. This classifier technique used to perform the imbalanced dataset.

Pseudo code for EML-CID:

```
Step 1: Input Dataset identify missing values and remove noisy data and redundant data.
Step 2: Group Data to form Clusters using KNN.
Step 3: Calculate Distance between each clusters using Manhattan Distance.
Step 4: Choose class of input dataset to form Majority Class and Minority Class.
Step 5: Identify the weightage for each records in each cluster or in each class
Step 6: Identify the Support Weightage Factor for each records.
Step 7: Choose the minority class and solve class imbalance by using the Threshold Selection weight.
Step 8: Repeat the process until all the minority class is solved by class imbalance problem
Step 9: The class Balanced dataset is used inn statistical analysis, fraud detection and for data prediction in open source software.
```

Algorithm for EML-CID:

I→Input
D_t→ Dataset
M_i → Minority Class
M_a→Majority Class
I_C→Input Class
C_b→Class Balance
KNN→ K nearest neighbor
W_i → Weightage of Individual Records
S_w → Support factor weight
T_{sw} → Threshold of Selection Weight
D_c→ Distance between each Clusters

Step 1: Input Dataset I→ D_t

Step 2: Form Clusters C= $\sum_{i=0}^{i=n} I(KNN)$

Step 3: Calculate Manhattan Distance

$$D_c = M_{d} = \sum_{Ci}^{Cn}\sum_{i=1}^{i=k} | xi - yi |^2$$

Step 4: Choose Input Class I_c in D_t , Compute M_i Minority Class and M_a Majority Class.

Step 5: Compute Weightage W_i = $\sum_{Ci}^{Cn} Wi(Ic)$ in each clusters or in each class.

Step 6: Identify the support weightage

$$S_w = \text{Avg}(\sum_{Ci}^{Cn} Wi(Ic))$$

Step 7: Calculate $T(S_w) = \text{Avg}(\sum_{Ci}^{Cn} Sw)$

Step 8: Choose the Minority Class and make Class Imbalance as Balance

$$C_b = \sum_{Ci}^{C->Mi} do\{if(Mi == Ma), ChooseT(Sw)$$

IV. ORIGINAL CONTRIBUTION

Imbalanced learning problems cover unequal distribution of data samples among different classes, where most of the samples belong to some classes and rest to the other classes. If such samples come only from two classes, the class having most of the samples is called the majority class and other the minority class. Learning from the imbalance data is of utmost important to the research community as it is present in many vital real-world classification problems, such as medical diagnosis information retrieval systems, detection of fraudulent telephone calls, detection of oil spills in radar images, data mining from direct marketing, and helicopter fault monitoring.

The primary goal of any classifier is to reduce its classification error, i.e., to maximize its overall accuracy. However, imbalance learning problems pose a great challenge to the classifier as it becomes very hard to learn the minority class samples. It is because the classifier learned from the imbalanced data tends to favor the majority class samples, resulting in a large classification error over the minority class samples. Imbalance that exists between the samples of two classes is usually known as between-class imbalance. The actual cause for the bad performance of conventional classifiers on the minority class samples is not necessarily related to only on the between-class imbalance. Classifiers' performance have been found to depreciate in the presence of within-class imbalance and small disjoints problems. Besides, the complexity of data samples is another factor for the classifiers' poor performance.

If the samples of the majority and minority classes have more than one concepts in which some concepts are rarer than others and the regions between some concepts of different classes overlap, then the imbalance problem becomes very severe. We propose a novel Hybrid Method named as EFFECTIVE MACHINE LEARNING-CLASS IMABALANCE DATA (EML-CID) whose goal is to alleviate the problems of imbalanced learning and generate the useful synthetic minority class samples to avoid fraud detection. The essences of the proposed method are: 1) selection of an appropriate subset of the original minority class samples, 2) assigning weights to the selected samples according to their importance in the data, and 3) using a clustering approach for generating the useful synthetic minority class samples.

The EML-CID Process is of 4 Phases

PHASE I: Dataset collection

A dataset collection of database contains the imbalanced dataset in various field such as medical diagnosis, facial recognition. The dataset is typically organized for student dataset to model aspects of reality in a way that supports processes requiring information. The dataset information is collected using and stored in this module that represent the user given input data set for the purpose of oversampling method. The input information contain the set of majority, minority, and synthetic sample.

PHASE II: Problem occurred in imbalanced dataset

The dataset contains the large amount of information. For example the normal dataset contains the True positive (TP), True negative (TN), False positive (FP), False negative (FN). So this information is known as imbalanced dataset in data mining. So the dataset contain the redundancy information and noisy data. The module represent the recognition of imbalanced dataset information is not accuracy and cannot determine the accurate prediction of minority sample dataset.

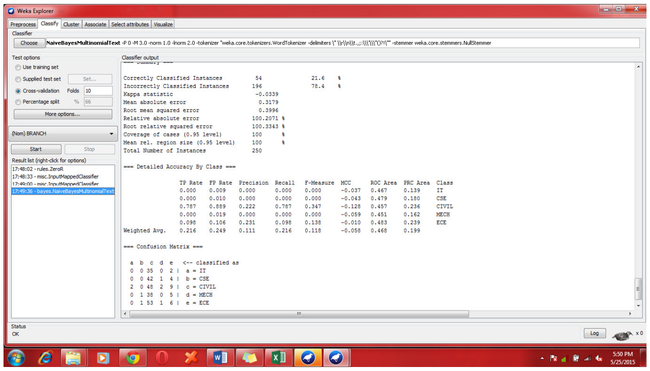
Phase III: Clustering the minority sample

The clustering using K-means clustering techniques in predicting the noisy minority class sample. Select an appropriate subset of the original minority class samples, assigning or calculate the weights to the selected samples according to their importance of the data and the calculated weightage based on the Euclidean distance from majority class samples. In based on weightage calculation to predict noisy data from minority sample set.

PHASE IV: Accuracy prediction of imbalanced dataset by EML-CID

The weight factor calculation is performed the selected sample set and using the K-means clustering operation that reduce the noisy minority class sample. The classifier techniques used to classify the separately on majority sample and minority sample of dataset and finally produce the oversampling dataset from the original dataset and that used to relate to the one record from another record dataset by EML-CID.

Input Dataset Accuracy Results:



Classifier

NaiveBayesMultinomialText -P 0 -M 3.0 -norm 1.0 -lnorm 2.0 -tokenizer "weka.core.tokenizers.WordTokenizer -delimiters \" \\r\\n\\t.,;:\\\'\\\"()?!\"" -stemmer weka.core.stemmers.NullStemmer

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) BRANCH

Result list (right-click for options)

17:48:02 - rules.ZeroR
17:48:33 - misc.InputMappedClassifier
17:49:00 - misc.InputMappedClassifier
17:49:36 - bayes.NaiveBayesMultinomialText

Classifier output

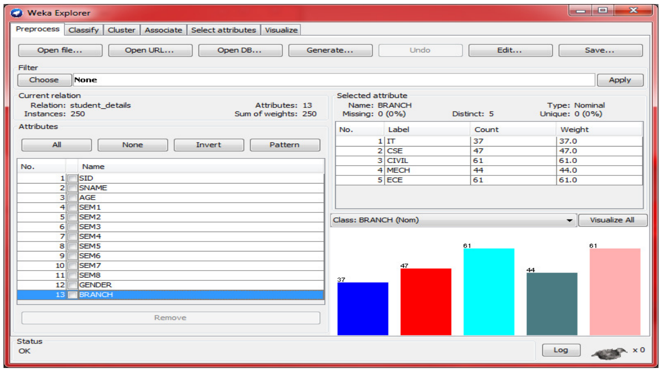
```
Correctly Classified Instances          54               21.6    %
Incorrectly Classified Instances       196               78.4    %
Kappa statistic                         -0.0339
Mean absolute error                      0.3179
Root mean squared error                  0.3996
Relative absolute error                100.2071 %
Root relative squared error            100.3343 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)     100      %
Total Number of Instances              250

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.009    0.000      0.000    0.000      -0.037   0.467     0.139     IT
                 0.000    0.010    0.000      0.000    0.000      -0.043   0.479     0.180     CSE
                 0.787    0.889    0.222      0.787    0.347      -0.128   0.457     0.236     CIVIL
                 0.000    0.019    0.000      0.000    0.000      -0.059   0.451     0.162     MECH
                 0.098    0.106    0.231      0.098    0.138      -0.010   0.483     0.239     ECE
Weighted Avg.    0.216    0.249    0.111      0.216    0.118      -0.058   0.468     0.199

=== Confusion Matrix ===

  a  b  c  d  e   <-- classified as
  0  0 35  0  2 |  a = IT
  0  0 42  1  4 |  b = CSE
  2  0 48  2  9 |  c = CIVIL
  0  1 38  0  5 |  d = MECH
  0  1 53  1  6 |  e = ECE
```

Status

OK

5:50 PM 5/25/2015

Class Imbalance Problem:



Filter: None

Current relation

Relation: student_details Attributes: 13

Instances: 250 Sum of weights: 250

Selected attribute

Name: BRANCH Type: Nominal

Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	IT	37	37.0
2	CSE	47	47.0
3	CIVIL	61	61.0
4	MECH	44	44.0
5	ECE	61	61.0

Attributes

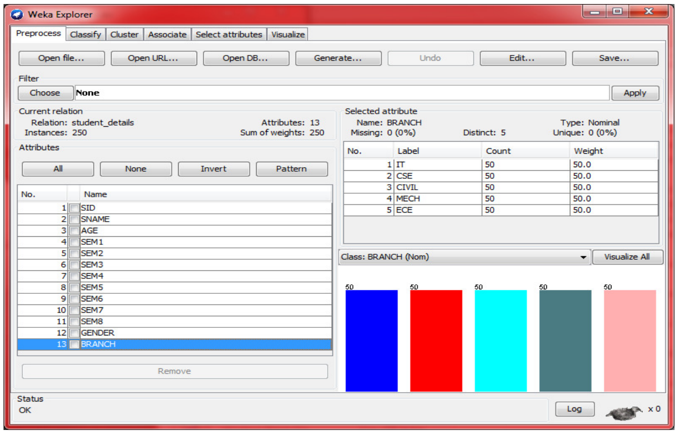
No.	Name
1	SID
2	SNAME
3	AGE
4	SEM1
5	SEM2
6	SEM3
7	SEM4
8	SEM5
9	SEM6
10	SEM7
11	SEM8
12	GENDER
13	BRANCH

Class: BRANCH (Nom)

Status

OK

Class Balance by EML-CID:



Filter: None

Current relation

Relation: student_details Attributes: 13

Instances: 250 Sum of weights: 250

Selected attribute

Name: BRANCH Type: Nominal

Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

No.	Label	Count	Weight
1	IT	50	50.0
2	CSE	50	50.0
3	CIVIL	50	50.0
4	MECH	50	50.0
5	ECE	50	50.0

Attributes

No.	Name
1	SID
2	SNAME
3	AGE
4	SEM1
5	SEM2
6	SEM3
7	SEM4
8	SEM5
9	SEM6
10	SEM7
11	SEM8
12	GENDER
13	BRANCH

Class: BRANCH (Nom)

Status

OK

V. Conclusion

The machine learning algorithm are more successful in overcoming the class imbalance problem. In this paper, a hybrid technique is to be introduced as Effective Machine Learning in Class Imbalance Data EML-CID algorithm, to extend the traditional ELMs beyond multi-class classification, multi clustering and Unseen data problems, the unified framework EML-CID is formed by the combination of Semi Supervised ELM (SS-ELM) and Unsupervised ELM (US-ELM). It leads to competitive results in Class Imbalance and Dataset prediction, and it requires significantly less training time. We test our algorithms on a variety of data sets including Student Dataset as Importance, and make comparisons by this experimental results. This favor the use of EML-CID for overcoming the class imbalance problem. Consideration should also be given to what performance metrics will be used to judge the performance of the methods and provide new paradigm into the Machine Learning Algorithm.

References

[1] Jung-Kyu Choi ; Dept. of Healthcare Eng., Chonbuk Nat. Univ., Jeonju, South Korea ; Chan Il Yoo ; Kyung-Ah Kim ; Yonggwan Won, “Study on Datamining Techinique for Foot Disease Prediction”, Published in: IT Convergence and Security (ICITCS), 2014 International Conference on Date of Conference: 28-30 Oct. 2014 Page(s): 1 – 4.

[2] Panah, O. ; Ayatollah Amoli Branch, Comput. Dept., Islamic Azad Univ., Amol, Iran ; Panah, A. ; Panah, A., “Evaluating the datamining techniques and their roles in increasing the search speed data in web”, Published in: Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on (Volume:9) Date of Conference: 9-11 July 2010.

[3] Erraguntla, M. ; Ramachandran, S. ; Chang-Nien Wu ; Mayer, R.J., “Avian Influenza Datamining Using Environment, Epidemiology, and Etiology Surveillance and Analysis Toolkit (E3SAT)”, Published in: System Sciences (HICSS), 2010 43rd Hawaii International Conference on Date of Conference: 5-8 Jan. 2010 Page(s): 1 – 7.

[4] Emanuel, A.W.R. ; Fac. of Inf. Technol., Maranatha Christian Univ., Bandung, Indonesia ; Wardoyo, R. ; Istiyanto, J.E. ; Mustofa, K., “Success factors of OSS projects from sourceforge using Datamining Association Rule”, Published in: Distributed Framework and Applications (DFmA), 2010 International Conference on Date of Conference: 2-3 Aug. 2010 Page(s): 1 – 8.

[5] Kulkarni, P., “Introduction to Reinforcement and Systemic Machine Learning”, Publisher : Wiley-IEEE Press Edition : 1 Pages : 1 – 21.

[6] Vidhate, D. ; Kulkarni, P., “Cooperative Machine Learning with Information Fusion for Dynamic Decision Making in Diagnostic Applications”, Published in: Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on (Volume:38 , Issue: 3) Page(s): 397 – 415 Date of Publication : May 2008.

[7] Guo Chen ; Nanjing Univ. of Aeronaut. & Astronaut., Nanjing ; Rongtao Hou, “A New Machine Double-Layer Learning Method and Its Application in Non-Linear Time Series Forecasting”, Published in: Mechatronics and Automation, 2007. ICMA 2007. International Conference on Date of Conference: 5-8 Aug. 2007 Page(s): 795 – 799.

[8] Zhanshan Ma ; Comput. Sci. Dept., Univ. of Idaho, Moscow, ID, “Cognitive ecology and social learning inspired machine learning: with particular reference to the evolving of resilient Airborne Networks (AN)”, Published in: Aerospace conference, 2009 IEEE Date of Conference: 7-14 March 2009 Page(s): 1 – 14.

[9] Ming Xue ; Changchun Inst. of Technol., Changchun, China ; Changjun Zhu, “A Study and Application on Machine Learning of Artificial Intellligence”, Published in: Artificial Intelligence, 2009. JCAI '09. International Joint Conference on Date of Conference: 25-26 April 2009 Page(s): 272 – 274.

[10] Stimpson, A.J. ; Dept. of Aeronaut. & Astronaut., Massachusetts Inst. of Technol., Cambridge, MA, USA ; Cummings, M.L., “Assessing Intervention Timing in Computer-Based Education Using Machine Learning Algorithms”, Published in: Access, IEEE (Volume:2) Page(s): 78 - 87 Date of Publication : 31 January 2014.

[11] Xiaofeng Liao ; Nat. Eng. Res. Center for Fundamental Software, Inst. of Software, Beijing, China ; Liping Ding ; Wang, Yongji, “Secure Machine Learning, a Brief Overview”, Published in: Secure Software Integration & Reliability Improvement Companion (SSIRI-C), 2011 5th International Conference on Date of Conference: 27-29 June 2011 Page(s): 26 – 29.

[12] Malik, A.M. ; Inst. of High Performance Comput., Singapore, Singapore, “Spatial Based Feature Generation for Machine Learning Based Optimization Compilation”, Published in: Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on Date of Conference: 12-14 Dec. 2010 Page(s): 925 – 930.

[13] Dong-Jun Yu ; Sch. of Comput. Sci. & Eng., Nanjing Univ. of Sci. & Technol., Nanjing, China ; Jun Hu ; Qian-Mu Li ; Zhen-Min Tang, “Constructing Query-Driven Dynamic Machine Learning Model With Application to Protein-Ligand Binding Sites Prediction”, Published in: NanoBioscience, IEEE Transactions on (Volume:14 , Issue: 1) Page(s): 45 – 58. Date of Publication : Jan. 2015.

[14] Kelly, D. ; National Coll. of Ireland, Dublin, Ireland ; Tangney, B.,” 'First Aid for You': getting to know your learning style using machine learning”, Published in: Advanced Learning Technologies, 2005. ICALT 2005. Fifth IEEE International Conference on Date of Conference: 5-8 July 2005 Page(s): 1 - 3.

[15] Batista, G. ; Inst. de Cienc. Mat. e de Comput., Univ. de Sao Paulo, Säo Carlos, Brazil ; Silva, D. ; Prati, R., “An Experimental Design to Evaluate Class Imbalance Treatment Methods”, Published in: Machine Learning and Applications (ICMLA), 2012 11th International Conference on (Volume:2) Date of Conference: 12-15 Dec. 2012 Page(s): 95 – 101.

[16] Shuo Wang ; Centre of Excellence for Res. in Comput. Intell. & Applic. (CERCIA), Univ. of Birmingham, Birmingham, UK ; Minku, L.L. ; Xin Yao, “Resampling-Based Ensemble Methods for Online Class Imbalance Learning”, Published in: Knowledge and Data Engineering, IEEE Transactions on (Volume:27 , Issue: 5) Page(s): 1356 - 1368 Date of Publication : 05 August 2014.

[17] Xiaoyuan Jing ; State Key Lab. of Software Eng., Wuhan Univ., Wuhan, China ; Chao Lan ; Min Li ; Yongfang Yao,” Class-imbalance learning based discriminant analysis”, Published in: Pattern Recognition (ACPR), 2011 First Asian Conference on Date of Conference: 28-28 Nov. 2011 Page(s): 545 – 549.

[18] Das, B. ; Sch. of Electr. Eng. & Comput. Sci., Washington State Univ., Pullman, WA, USA ; Krishnan, N.C. ; Cook, D.J., “Handling Class Overlap and Imbalance to Detect Prompt Situations in Smart Homes”, Published in: Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on Date of Conference: 7-10 Dec. 2013 Page(s): 266 – 273.

[19] Shuo Wang ; CERCIA, Univ. of Birmingham Birmingham, Birmingham, UK ; Minku, L.L. ; Xin Yao, “A learning framework for online class imbalance learning”, Published in: Computational Intelligence and Ensemble Learning (CIEL), 2013 IEEE Symposium on Date of Conference: 16-19 April 2013 Page(s): 36 – 45.

[20] Duhaney, J. ; Comput. & Electr. Eng. & Comput. Sci., Florida Atlantic Univ., Boca Raton, FL, USA ; Khoshgoftaar, T.M. ; Napolitano, A., “Studying the Effect of Class Imbalance in Ocean Turbine Fault Data on Reliable State Detection”, Published in: Machine Learning and Applications (ICMLA), 2012 11th International Conference on (Volume:1) Date of Conference: 12-15 Dec. 2012 Page(s): 268 – 275.

[21] Shuo Wang ; Sch. of Comput. Sci., Univ. of Birmingham, Birmingham, UK ; Minku, L.L. ; Xin Yao, “A multi-objective ensemble method for online class imbalance learning”, Published in: Neural Networks (IJCNN), 2014 International Joint Conference on Date of Conference: 6-11 July 2014 Page(s): 3311 – 3318.

[22] Shuo Wang ; Centre of Excellence for Res. in Comput. Intell. & Applic. (CERCIA), Univ. of Birmingham, Birmingham, UK ; Xin Yao, “Theoretical Study of the Relationship between Diversity and Single-Class Measures for Class Imbalance Learning”, Published in: Data Mining Workshops, 2009. ICDMW '09. IEEE International Conference on Date of Conference: 6-6 Dec. 2009 Page(s): 76 – 81.

[23] Orriols, A. ; Comput. Eng. Dept., Ramon Llull Univ., Barcelona ; Bernadó-Mansilla, E., “Class imbalance problem in UCS classifier system: fitness adaptation”, Published in: Evolutionary Computation, 2005. The 2005 IEEE Congress on (Volume:1) Date of Conference: 5-5 Sept. 2005 Page(s): 604 - 611 Vol.1.

[24] Jindaluang, W. ; Dept. of Comput. Sci., Chiang Mai Univ., Chiang Mai, Thailand ; Chouvatut, V. ; Kantabutra, S., “Under-sampling by algorithm with performance guaranteed for class-imbalance problem”, Published in: Computer Science and Engineering Conference (ICSEC), 2014 International Date of Conference: July 30 2014- Aug. 1 2014 Page(s): 215 – 221.

[25] Seliya, N. ; Comput. & Inf. Sci., Univ. of Michigan - Dearborn, Dearborn, MI ; Zhiwei Xu ; Khoshgoftaar, T.M., “Addressing Class Imbalance in Non-binary Classification Problems”, Published in: Tools with Artificial Intelligence, 2008. ICTAI '08. 20th IEEE International Conference on (Volume:1) Date of Conference: 3-5 Nov. 2008 Page(s): 460 – 466.