International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Adaptive Classification of Imbalanced Data using ANN with Particle of Swarm Optimization

Nitesh Kumar[1], Dr. Shailja Sharma[2]

[1]PG Scholar, [2]Associate Professor
[1,2]Department of CSE, RNTU, Bhopal, Madhya Pradesh, India

How to cite this paper: Nitesh Kumar | Dr. Shailja Sharma "Adaptive Classification of Imbalanced Data using ANN with Particle of Swarm Optimization" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.166-170, **https://doi.org/10.31142/ijtsrd25255**

IJTSRD25255

ABSTRACT

Customary characterization calculations can be constrained in their execution on exceedingly uneven informational collections. A famous stream of work for countering the substance of class inelegance has been the use of an assorted of inspecting methodologies. In this correspondence, we center on planning alterations neural system to properly handle the issue of class irregularity. We consolidate distinctive "rebalance" heuristics in ANN demonstrating, including cost-delicate learning, and over-and under testing. These ANN-based systems are contrasted and different best in class approaches on an assortment of informational collections by utilizing different measurements, including G-mean, region under the collector working trademark curve, F-measure, and region under the exactness/review curve. Numerous regular strategies, which can be classified into testing, cost-delicate, or gathering, incorporate heuristic and task subordinate procedures. So as to accomplish a superior arrangement execution by detailing without heuristics and errand reliance, presently propose RBF based Network (RBF-NN). Its target work is the symphonious mean of different assessment criteria got from a perplexity grid, such criteria as affectability, positive prescient esteem, and others for negatives. This target capacity and its enhancement are reliably detailed on the system of CM-KLOGR, in light of least characterization mistake and summed up probabilistic plunge (MCE/GPD) learning. Because of the benefits of the consonant mean, CM-KLOGR, and MCE/GPD, RBF-NN improves the multifaceted exhibitions in a very much adjusted way. It shows the definition of RBF-NN and its adequacy through trials that nearly assessed RBF-NN utilizing benchmark imbalanced datasets.

KEYWORDS*: Imbalanced Classification, Neural Network, RBF, F-measure, Heuristic Search, MCE/GPD*

1. INTRODUCTION

In practical application, many datasets are imbalanced, i.e., some classes have much more instances than others. Imbalanced learning is common in many situations like information filtering and fraud detection. Datasets imbalance must be taken into consideration in classifier designing, otherwise the classifier may tend to be overwhelmed by the majority class and to ignore the minority class. Re-sampling technique is an effective approach to imbalance learning. Many re-sampling methods are used to reduce or eliminate the extent of datasets imbalance, such as over-sampling the minority class, under-sampling the majority class and the combination of both methods. But it showed that under-sampling can potentially remove certain important instances and lose some useful information, and over-sampling may lead to over fitting. Over-sampling methods also suffer from noise and outliers. Support Vector Machine (SVM) has been widely used in many application areas of machine learning. However, regular SVM is no longer suitable to imbalance-class especially when the datasets are extremely imbalanced. An effective approach to improve the performance of SVM used in imbalanced datasets is to bias the classifier so that it pays more attention to minority instances. This can be done by setting different misclassifying penalty. We proposed an over-sampling algorithm based on data density in previous work. However, this algorithm sometimes leads to over fitting. In this paper, an adaptive over-sampling algorithm with two smoothing methods to avoid over fitting is proposed. Compared with other oversampling algorithms and our previous work, this algorithm can synthesize samples more efficiently and eliminate the effects of noise. Contributions of this paper are as follows:

- This novel method can effectively eliminate the noise compared with most other sampling methods like RO and CM-KLR. Noise is recognized and no new samples are synthesized around it.
- Different number new samples are synthesized around each minority sample according to its level of learning difficulty. This level is related to the sample density information. To calculate the sample density, core-distance and reach ability-distance are used.
- To avoid over fitting, two smoothing methods are proposed. One is using a sigmoid function to smooth the disparity of new samples synthesized around each minority sample. The other is using linear interpolation method to tradeoff between our algorithm and CM-KLR algorithm.

2. TEXT PRE-PROCESSING

Classification is an important task of pattern recognition. A range of classification learning algorithms, such as decision tree, back propagation neural network, Bayesian network, nearest neighbor, support vector machines, and the newly reported associative classification, have been well developed and successfully applied to many application domains. However, imbalanced class distribution of a data set has encountered a serious difficulty to most classifier learning algorithms which assume a relatively balanced distribution. The imbalanced data is characterized as having many more instances of certain classes than others. As rare instances occur infrequently, classification rules that predict the small classes tend to be rare, undiscovered or ignored; consequently, test samples belonging to the small classes are misclassified more often than those belonging to the prevalent classes. In certain applications, the correct classification of samples in the small classes often has a greater value than the contrary case. For example, in a disease diagnostic problem where the disease cases are usually quite rare as compared with normal populations, the recognition goal is to detect people with diseases. Hence, a favorable classification model is one that provides a higher identification rate on the disease category. Imbalanced or skewed class distribution problem is therefore also referred to as small or rare class learning problem.

3. LITERATURE SURVEY

2017 [1], There have been many attempts to classify imbalanced data, since this classification is critical in a wide variety of applications related to the detection of anomalies, failures, and risks. Many conventional methods, which can be categorized into sampling, cost-sensitive, or ensemble, include heuristic and task dependent processes. In order to achieve a better classification performance by formulation without heuristics and task dependence, we propose confusion-matrix-based kernel logistic regression (CM-KLOGR).

2017 [2], BigData applications are emerging during the last years, and researchers from many disciplines are aware of the high advantages related to the knowledge extraction from this type of problem. However, traditional learning approaches cannot be directly applied due to scalability issues. To overcome this issue, the MapReduce framework has arisen as a "de facto" solution. Basically, it carries out a "divide-and conquer" distributed procedure in a fault-tolerant way to adapt for commodity hardware. Being still a recent discipline, few researches has been conducted on imbalanced classification for Big Data.

2014 [3], classifications of the data collected from students of polytechnic institute has been discussed. This data is pre-processed to remove unwanted and less meaningful attributes. These students are then classified into different categories like brilliant, average, weak using decision tree and naïve Bayesian algorithms. The processing is done using WEKA data mining tool. This paper also compares results of classification with respect to different performance parameters.

2014 [4], Rough set theory provides a useful mathematical concept to draw useful decisions from real life data involving vagueness, uncertainty and impreciseness and is therefore applied successfully in the field of pattern recognition, machine learning and knowledge discovery.

2012 [5], Re-sampling method is a popular and effective technique to imbalanced learning. However, most re-sampling methods ignore data density information and may lead to over fitting. A novel adaptive over-sampling technique based on data density (ASMOBD) is proposed in this paper. Compared with existing re-sampling algorithms, ASMOBD can adaptively synthesize different number of new samples around each minority sample according to its level of learning difficulty.

2012 [6], Classifier learning with data-sets that suffer from imbalanced class distributions is a challenging problem in data mining community. This issue occurs when the number of examples that represent one class is much lower than the ones of the other classes. Its presence in many real-world applications has brought along a growth of attention from researchers.

4. PROBLEM IDENTIFICATION AND OBJECTIVES

The identified problem in current research work is as follows

1. Unrelated data are classify for specific dataset
2. Inconsistency exists during classification process.
3. Due to low sampling irregular sampling rate generate
4. Uncertainty of classification

The desired objectives of proposed work is as follows

1. To improve precision and recall for related data are classify.
2. To improve predictive positivity for reduce inconsistency during classification process.
3. To improve F1-measure for regular sampling rate.
4. To improve harmonic mean for certainty of classification.

5. METHODOLOGY

The algorithm of RBF-NN is perform in three phases

Phase 1: Particle of Swarm Optimization

Step 1: Initialize

// Initialize all particles

Step 2: Repeat

for each particle i in S do

// update the particle best solution

if(fx_i) < f(pb_i) then

$pb_i = x_i$

end if

// update the global best solution

if (pbi) > f(gb)

$gb = pb_i$

end if

end for

Step 3:

// update particle's velocity and position

for each particle i in S do

for each dimension d in D do

$v_{i,d} = v_{i,d} + C_1 * Rnd(0,1) * [pb_{i,d} - x_{i,d}] + C_2 * Rnd(0,1) * [gb_d - x_{i,d}]$

$x_{i,d} = x_{i,d} + v_{i,d}$

end for

end for

Step 4:

//advance iteration

it = it+1

until it > *MAX_ITERATIONS*

Phase 2: Cuckoo Walk Optimization

Step 5: Objective function f(x), x= $(x_1, x_2,...,x_d)^T$

Step 6: Generate an initial population of n host nests x_i (i=1,2,..,n)

Step 7: while(t < MaxGeneration)
Get random value
Evaluate its quality / velocity F_i
Choose nest among n (say, j) randomly,
if (F_i>F_j)
Replace j by the new solution
end
A fraction (p_a) of worth nests are replaced by new random solutions
Keep the best solution
Rank the solution and find the current best
Pass the current best solution to the next generation
end while
return the best next
end

Phase 3: Neural Network

Step 8: Assign random weight to all the linkages

Step 9: Using the inputs and linkage find the activation rate of hidden nodes

Step 10: Using the activation rate of hidden nodes and linkages to output, find the activation rate of output nodes

Step 11: find the error rate of output node and recalibrate all the linkages between hidden nodes and output nodes.

Step 12: Using the weights and error found at output node, cascade down the error to hidden nodes

Step 13: Recalibrate the weights between hidden nodes and input nodes.

Step 14: Repeat the process till the convergence criterion is met.

Step 15: Using the final linkage weight score the activation rate of the output nodes.

6. RESULTS AND ANALYSIS

It is difficult to precisely evaluate the classification performance for imbalanced data, especially when the data is small. Because of the imbalance and the small number of instances, the distribution of instances often differs between the training, validation, and test sets. The difference in distribution makes performance evaluation imprecise, and consequently, it sometimes leads to improper setting. For this solution and the fair comparison of the classifiers, the following processes were designed to divide and feed datasets, to set the hyper-parameters, parameters, and cutoff, and to estimate the performance.

Table 1: Comparative analysis of sensitivity in between of CM-KLOGR [1] and RBF-NN (Proposed)

Dataset	CM-KLOGR [1]	RBF-NN (Proposed)
Breast	95.83	97.24
Haberman	75	78.18
Ecoli-pp	100	100
Ecoli-imu	50	53.21
Pop failures	100	100
Yeast-1 vs 7	100	100

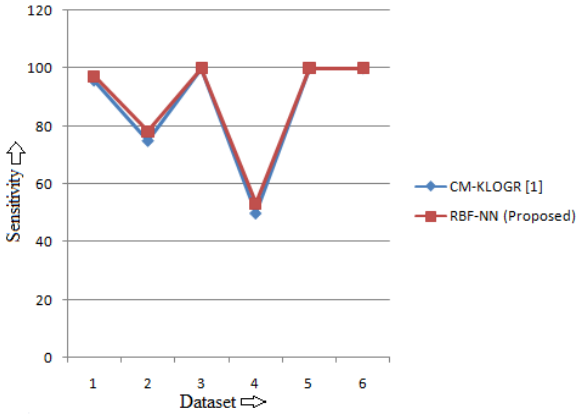

Figure 1: Graphical analysis of sensitivity in between of CM-KLOGR [1] and RBF-NN (Proposed)

In table 1, recall (sensitivity) of proposed work is improve as compare than CM-KLOGR[1] for all considerable dataset. Concretely speaking, RBF-NN perform best under low imbalance, tied for first place with RBF-NN under moderate imbalance, and performed best under high imbalance.

Table 2: Comparative analysis of specificity in between of CM-KLOGR [1] and RBF-NN (Proposed)

Dataset	CM-KLOGR [1]	RBF-NN (Proposed)
Breast	95.65	97.21
Haberman	82.61	85.72
Ecoli-pp	93.10	95.41
Ecoli-imu	96.67	98.96
Pop failures	95.92	97.13
Yeast-1 vs 7	88.37	91.08

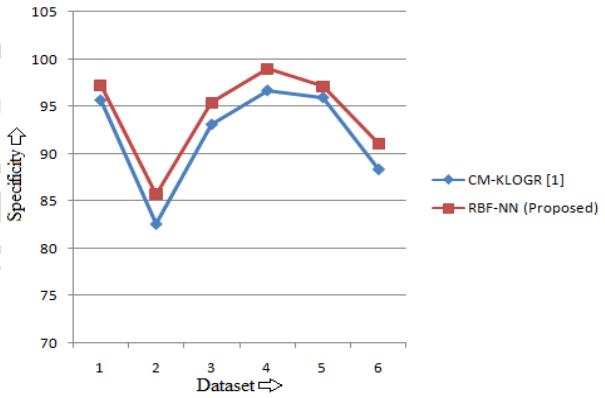

Figure 2: Graphical analysis of specificity in between of CM-KLOGR [1] and RBF-NN (Proposed)

In table 2, specificity of proposed work is improve as compare than CM-KLOGR[1] for all considerable dataset. Concretely speaking, RBF-NN perform best relevant items under predefined dataset, tied for first place with RBF-NN under moderate imbalance, and performed best under high imbalance.

Table 3: Comparative analysis of PPV in between of CM-KLOGR [1] and RBF-NN (Proposed)

Dataset	CM-KLOGR [1]	RBF-NN (Proposed)
Breast	92.00	93.76
Haberman	60.00	62.13
Ecoli-pp	71.43	73.66
Ecoli-imu	66.67	69.82
Pop failures	71.43	73.64
Yeast-1 vs 7	37.50	39.83

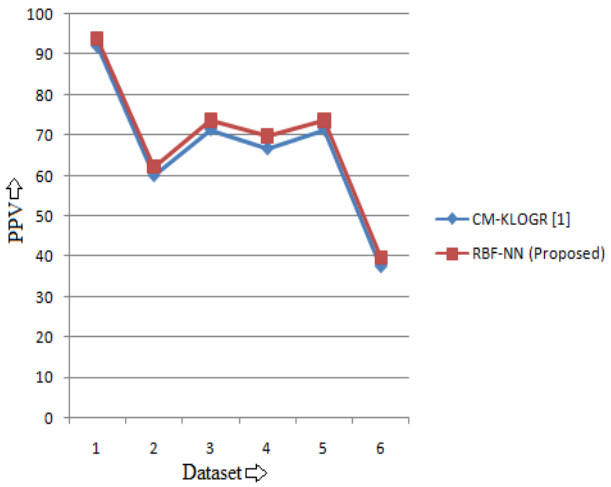

Figure 3: Graphical analysis of PPV in between of CM-KLOGR [1] and RBF-NN (Proposed)

In table 3, PPV of proposed work is improve as compare than CM-KLOGR[1] for all considerable dataset.

Table 4: Comparative analysis of NPV in between of CM-KLOGR [1] and RBF-NN (Proposed)

Dataset	CM-KLOGR [1]	RBF-NN (Proposed)
Breast	97.78	98.81
Haberman	90.48	93.08
Ecoli-pp	100.00	100.00
Ecoli-imu	93.55	95.14
Pop failures	100.00	100.00
Yeast-1 vs 7	100.00	100.00

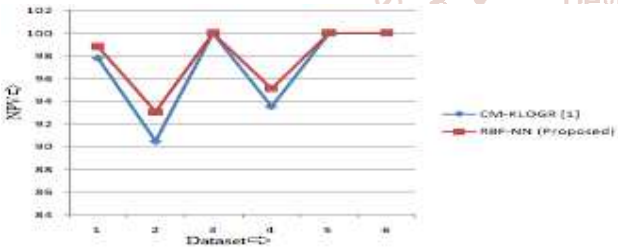

Figure 4: Graphical analysis of NPV in between of CM-KLOGR [1] and RBF-NN (Proposed)

In table 4, NPV of proposed work is improve as compare than CM-KLOGR[1] for all considerable dataset. Concretely speaking, RBF-NN perform best negativity of unrelated items under predefined dataset, tied for first place with RBF-NN under moderate imbalance, and performed best under high imbalance.

Table 5: Comparative analysis of HM in between of CM-KLOGR [1] and RBF-NN (Proposed)

Dataset	CM-KLOGR [1]	RBF-NN (Propose)
Breast	95.3061	96.06147
Haberman	75.2728	78.0031
Ecoli-pp	89.42544	90.8059
Ecoli-imu	71.4158	74.4601
Pop failures	90.0698	91.19927
Yeast-1 vs 7	69.0012	71.3393

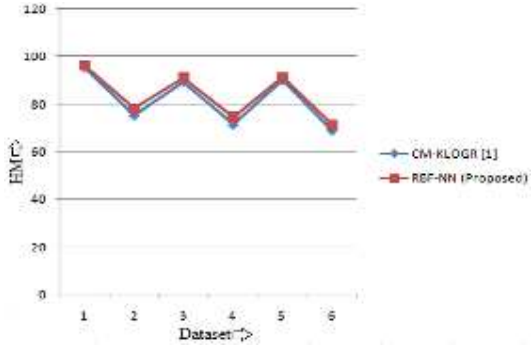

Figure5: Graphical analysis of HM in between of CM-KLOGR [1] and RBF-NN (Proposed)

7. CONCLUSIONS

In the experiments, RBF-NN outperformed CM-KLOGR and KLOGR with or without sampling, for several datasets under different settings of the weights on the evaluation criteria. It was confirmed that RBF-NN can increase the values of the evaluation criteria in a well balanced way, adaptively to their priorities.

1. Precision is improved by 1.9% and recall is improved by 1.47% for related data are classified.
2. Predictive positivity is improve for reduce inconsistency during classification process.
3. F1-measure is improved by 1.2% for regular sampling rate.

To compute density of each sample, reach ability-distance and core-distance need to be computed firstly, which consumes much more time than that of RBF-NN. Computation complexity reduction is another work we need to do in the future.

8. REFERENCES

[1] Miho Ohsaki, Peng Wang, Kenji Matsuda, Shigeru Katagiri, Hideyuki Watanabe, and Anca Ralescu, "Confusion-matrix-based Kernel Logistic Regression for Imbalanced Data Classification", IEEE Transactions on Knowledge and Data Engineering, 2017.

[2] Alberto Fernández, Sara del Río, Nitesh V. Chawla, Francisco Herrera, "An insight into imbalanced Big Data classification: outcomes and challenges", Springer journal of bigdata, 2017.

[3] Vaibhav P. Vasani1, Rajendra D. Gawali, "Classification and performance evaluation using data mining algorithms", International Journal of Innovative Research in Science, Engineering and Technology, 2014.

[4] Kaile Su, Huijing Huang, Xindong Wu, Shichao Zhang, "Rough Sets for Feature Selection and Classification: An Overview with Applications", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-3, Issue-5, November 2014.

[5] Senzhang Wang, Zhoujun Li, Wenhan Chao and Qinghua Cao, "Applying Adaptive Over-sampling Technique Based on Data Density and Cost-Sensitive SVM to Imbalanced Learning", IEEE World Congress on Computational Intelligence June, 2012.

[6] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince and Francisco Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging, Boosting, and Hybrid-Based Approaches",

IEEE Transactions on Systems, Man and Cybernetics—Part C: Applications and Reviews, Vol. 42, No. 4, July 2012.

[7] Nada M. A. Al Salami, "Mining High Speed Data Streams". UbiCC Journal, 2011.

[8] Dian Palupi Rini, Siti Mariyam Shamsuddin and Siti Sophiyati, "Particle Swarm Optimization: Technique, System and Challenges", International Journal of Computer Applications (0975 – 8887) Volume 14–No.1, January 2011.

[9] Amit Saxena, Leeladhar Kumar Gavel, Madan Madhaw Shrivas, "Online Streaming Feature Selection", 27th International Conference on Machine Learning, 2010.

[10] Yuchun Tang, Member, Yan-Qing Zhang, Nitesh V. Chawla and Sven Krasser, "SVMs Modeling for Highly Imbalanced Classification", IEEE Transaction on Systems, Man and Cybernetics, Vol. 39, NO. 1, Feb 2009.

[11] Haibo He and Edwardo A. Garcia, "Learning from Imbalanced Data", IEEE Transactions on Knowledge and Data Engineering, September 2009.

[12] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", International Multi Conference of Engineers and Computer Scientists, IMECS 2009, March, 2009.

[13] Haibo He, Yang Bai, Edwardo A. Garcia and Shutao Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning", IEEE Transaction of Data Mining, 2009.

[14] Swagatam Das, Ajith Abraham and Amit Konar, "Particle Swarm Optimization and Differential Evolution Algorithms: Technical Analysis, Applications and Hybridization Perspectives", Springer journal on knowledge engineering, 2008.

[15] "A logical framework for identifying quality knowledge from different data sources", International Conference on Decision Support Systems, 2006.

[16] "Database classification for multi-database mining", International Conference on Decision Support Systems, 2005.

[17] Volker Roth, "Probabilistic Discriminative Kernel Classifiers for Multi-class Problems", Springer-Verlag journal, 2001.

[18] R. Chen, K. Sivakumar and H. Kargupta "Collective Mining of Bayesian Networks from Distributed Heterogeneous Data", Kluwer Academic Publishers, 2001.

[19] Shigeru Katagiri, Biing-Hwang Juang and Chin-Hui Lee, "Pattern Recognition Using a Family of Design Algorithms Based Upon the Generalized Probabilistic Descent Method", IEEE Journal of Data Minig, 1998.

[20] I. Katakis, G. Tsoumakas, and I. Vlahavas. Tracking recurring contexts using ensemble classifiers: an application to email filtering. Knowledge and Information Systems, Pp 371–391, 2010.

[21] J. Kolter and M. Maloof. Using additive expert ensembles to cope with concept drift. In Proc. ICML, Pp 449–456, 2005.

[22] D. D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. Journal of Machine Learning Research, Pp 361–397, 2004.

[23] X. Li, P. S. Yu, B. Liu, and S.-K. Ng. Positive unlabeled learning for data stream classification. In Proc. SDM, Pp 257–268, 2009.

[24] M. M. Masud, Q. Chen, J. Gao, L. Khan, J. Han, and B. M.Thuraisingham. Classification and novel class detection of data streams in a dynamic feature space. In Proc. ECML PKDD, volume II, Pp 337–352, 2010.

[25] P. Zhang, X. Zhu, J. Tan, and L. Guo, "Classifier and Cluster Ensembles for Mining Concept Drifting Data Streams," Proc. 10th Int'l Conf. Data Mining, 2010.