A Survey on Unstructured Data Classification using Uncertain Nearest Neighbor Decision Rule

NIJAGUNA GOLLARA SIDDAPPA, THIPPESWAMY KAMPALAPPA

Abstract— *Classification over numerous real-world datasets has a peculiar drawback called unstructured class problem. A dataset is said to be unstructured when majority class have more samples insignificantly than the minor class. Such drawbacks results in an ineffective performance of data classification techniques. Classification is a supervised learning method which acquires a training dataset to form its model for classifying unseen examples. Mostly such concerns are often arose in real world applications since day-to-day situations constitute majority class in abundant nature. Another reason for class unstructured problem is the limitations (e.g., cost, difficulty or privacy) on collecting instances of some classes. However, in unstructured data classification, the class boundary learned by standard machine learning algorithms can be severely skewed toward the target class. As a result, the false-negative rate can be excessively high. This review paper evaluates the researches done on the unstructured data classification using uncertain Nearest Neighbor (NN) decision rule and also auditing the major issues face by several techniques. The researchers can give the better solution for the current problems faced by using this procedure in the unstructured data classification.*

***Index Terms*: Classification, Data Imbalance, Data sets, k-Nearest Neighbor(KNN) , Uncertain nearest Neighbor (UNN), Multi Class Data Imbalance Data Set, Binary Class Data Imbalance Data Set**

1 Introduction

Enormous measures of real-world information are accumulated by various enterprises each day. While at the same time these enormous measures of information have made awesome potential for learning revelation, the measure of information extraction is now and then restricted by a typical issue among real-world datasets, which is the class imbalance issue, i.e. at the point when the quantity of information tests having a place with one class far outperforms the quantity of information tests having a place with every one of alternate classes. A few cases for such class unevenness issues would be diagnosis of uncommon infections [1], discovery of deceitful phone calls [2], network intrusion identification [3] and location of oil slicks in radar images [4]. Managing imbalance problem can be troublesome for classifiers as they tend to support the class that most examples has a place with [5]. Besides, the class with the minimum examples is typically the one of prime intrigue [6]. With the consistent expanding volume of content information from Internet, it is an imperative undertaking to sort these records into sensible and straightforward classes.

Text categorization plans to consequently put the pre-characterized names on known inconspicuous reports. It is a functioning examination zone in data recovery, machine learning and Natural Language Processing. A significant number of the standard characterization calculations for the most part expect that the preparation cases are equally conveyed among various classes. Be that as it may, as demonstrated in [7] imbalanced informational indexes frequently show up in numerous pragmatic applications. In an imbalanced dataset, the dominant part class is spoken to by an extensive bit of the considerable number of illustrations, while the other, the minority class has just a little level everything being equal. At the point when a content classifier experiences a lopsided report corpus, the order execution frequently diminishes. So as to enhance the execution of order calculations on uneven appropriation content corpora, a few specialists fall back on inspecting

procedures [8][9]. In any case, the expulsion of preparing reports in substantial classes may lose some critical data and dependably forfeits the order execution now and again. In this paper, study on the unverifiable NN choices administer has been done with a specific end goal to examination the execution and worries of a few philosophies. This procedure persuades the scientist's for additionally explore work in distributed computing.

This survey paper is composed as follows, the data imbalance problem are described in Section II. Section III describes the overview of the system model. Section IV survey several recent papers on uncertain NN decision rule. The conclusion is made in the section V.

2Data Imbalance Problem

If class dispersion among classes in dataset isn't uniform it is said to be an imbalance dataset [10]. In this condition there is no less than one class which is spoken to by just few illustrations (minority class), different classes make up whatever is left of dataset (majority class). Ongoing exploration in the machine learning demonstrated that utilizing an uneven conveyance of class cases in the learning procedure can leave learning calculations with execution predisposition. It implies that classifier gives high precision on the major share class however it gives poor exactness on the minority class. This is on the grounds that conventional preparing criteria, for example, the general achievement can be incredibly affected by the bigger number of cases from the dominant part class. As the minority classes assume a vital part in numerous genuine issues, as the precisely arranging cases from this class is likewise critical. Scientists have recognized information imbalance issue into two fundamental composes: Binary class information imbalance and multi class information imbalance [11].

2.1Binary Class Data Imbalance Dataset:

A binary dataset is said to have only two classes. On the off chance that in the parallel dataset there exists a class which is spoken to by just a couple of quantities of cases, at that point it is called binary class information imbalance issue. In binary class dataset zero class edges is by and large used to isolate two classes, so there is no compelling reason to recognize the limits of classes in dataset.

2.2 Multi Class Data Imbalance Dataset:

The dataset contains in excess of two classes is called multiclass dataset. Information imbalance issue makes extra overheads in multiclass dataset. Straightforward and proficient zero class edges can't be utilized as a part of multiclass dataset. Complex techniques like Static Search Selection or Dynamic Search Selection should be utilized. A few times to order dataset, multiclass issue is should have been partitioned into numerous paired class issues.

3Overview of System Model.

The normal approach incorporates allocating diverse misclassification expenses to erroneous class predictions [12] or creating enhanced learning criteria that are more delicate to the imbalance class circulations contrasted with the standard overall error rate or accuracy. Enhanced learning criteria incorporate the normal classification accuracy of the minority and larger part classes. In wide-ranging approaches, just couples of strategies have been proposed by specialists. One of the well-known strategies which take care of the information imbalance issue is accomplished by utilizing an approach namely NN decision rule. The figure 1 demonstrates the general structure of the model in imbalance dataset.

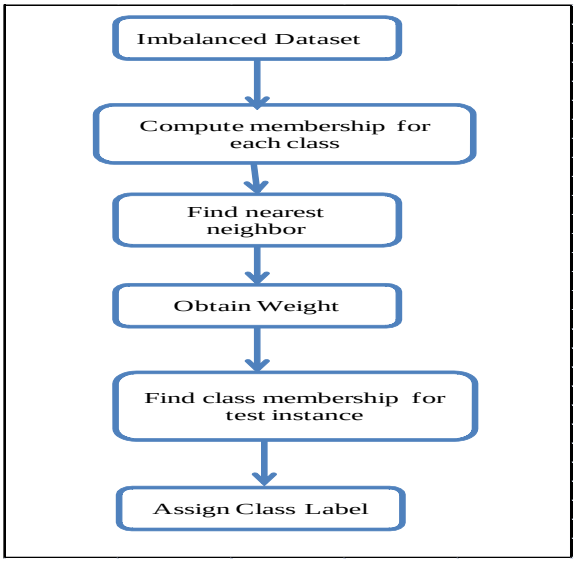

Figure 1: Structure of the model

3.1. Imbalance Dataset

Information level methodologies work by resampling the training instances with a specific

end goal to accomplish a more balanced dataset. This is finished by over-sampling the minority classes' examples, under sampling the larger part classes' examples, or applying half and half models which are a mix of over-testing and under sampling methods. Such strategies are considered as pre-processing approaches for managing the class imbalance issue. In conclusion, the effect of a resampling strategy on the classification task correspondingly relies upon the dataset. Now and again, picking the wrong resampling system may adversely influence the classification of unstructured data [13].

3.2. Computation of Membership Function

On thinking about a classification issue, if the earlier probabilities and the state restrictive densities of all classes are known, the Bayesian decision hypothesis delivers the ideal outcomes as in it limits the normal misclassification rate [14]. Be that as it may, in exact classification issues, the real likelihood conveyance of the populace is obscure. Under this situation, numerous non-Bayesian grouping methods, for example, clustering and discriminate examination, are composed in view of the idea of the closeness or separation in the element space that portrays the perceptions

3.3 Nearest Neighbor Algorithm

Algorithmic level strategies incorporate changing past machine learning calculations keeping in mind the end goal to manage the imbalance between classes straightforwardly; e.g. by appointing weights to training instances. A technique that has gotten a considerable measure of consideration in this prospect is k-Nearest Neighbor (k-NN) [15]. This is on the grounds that k-NN is a standout amongst the most productive and least complex classifiers in traditional machine learning undertakings. Be that as it may, k-NN's execution performance reduces when the dataset is imbalanced [16]. To defeat this downside is the K-Exemplar-based Nearest Neighbor calculation (ENN) was proposed. The Positive-one-sided Nearest Neighbor (PNN) [17] is another example situated strategy like ENN; however, it doesn't have a learning stage. As opposed to the example situated strategies, there are the distribution-oriented techniques, which depend on procuring helpful earlier information of the information circulation. The Class-Based Weighted k Nearest Neighbor is one of these techniques as it measures the examples in view of the computed misclassification rate of k-NN [18]. Informative k Nearest Neighbor-localized version (Ll-kNN) [19] and Class Conditional NearestNeighbor Distribution (CCNND) [20] are two different cases of circulation situated techniques.

3.4 Weight Calculation

The NN strategy is to dole out extensive weights to little classes and little weights to vast classes to limit the biasness of the classifier set out toward larger part class and evasion of minority class. These weights help in more exact classification of imbalance information. In the wake of discovering inquiry occasion q_u from NN technique, we discover weights from following condition:

$$W_j = \frac{1}{(N(C_j)/Min\{N(C_j)|j = 1,2\})^{1/p}}$$

$p$is a proponent and $p < 1$

$W_j$is the weight to be identified

$N(C_j)$ number of nearest neighbors of an object belongs to class$j$.

3.5. Assigning Class Label:

With the NN calculation, class names of the k learning occasions nearest to a testing instance which helps to decide the class name of the test example. Opposite separation weighting is to measure the vote of each neighbor as indicated by the backwards of its separation from the test occurrence. By taking the weighted normal of the neighbors closest to the test instance smoothes out the effect of disconnected boisterous preparing occurrences. Besides it lifts the heaviness of instance.

4Literature Review:

Several techniques are suggested by researchers in the imbalanced data structure by uncertain NN decision rule techniques. In this scenario, brief evaluations of some important contributions to the existing techniques are presented.

Author	J. Maillo, *et al.*, [21]	S. Saryazdi, *et al.*, [22]	Y. Xu, *et al.*, [23]
Methodology employed	Present a Iterative MapRedcue solution to perform an exact k-nearest neighbour classification based on Spark.	Neighbors' Progressive Competition (NPC) for classification of imbalanced datasets	Presented a K-Nearest Neighbour based Maximum margin and minimum Volume Hyper-sphere machine (KNN-M3VHM)
Dataset	Poker Hand, Susy, Higgs and ECBDL'14 Datasets	KEEL repository dataset	KEEL, BCI competition II datasets and UCI machine Learning repository
Advantage	KNN-IS is an exact parallel appraoch and obtained the same accuracy and very good achievements on runtimes	The NPC does not limit its decision criteria for every query sample to a present number of neighbors. Also, the method does not require any manually-set parameters	The significant advantage in speed since the KNN-M3VHMM solves two smaller sized optimization problems
Limitation	The proposed approach didnot tackle big datasets that contains missing values	The NPC method didnot concentrate on multi-class imbalanced datasets.	The method provides poor accuracy if the imbalanced datasets are not preprocessed by sampling methods
Performance measure	Map Runtime, Redcue Runtime and Speed UP	Geometric Means (GM) and Average Processing Time	Accuracy, Sensitivity, specificity and G-mean

5Conclusion

Uncertain rule based Classification has various applications in a wide assortment of mining and different applications, for example, distinguishing faces from pictures dataset, perceiving voice in information of discourse and so on. Given the measure of information that should be grouped, computerized order frameworks are exceptionally alluring. Classifiers order datasets as indicated by class marks. Classifiers perform well if dataset is adjusted. Dataset is called imbalance or unstructured in the event that at least one classes are introduced by just a couple of number of illustrations. In numerous applications class dissemination is imbalanced, and the minority class is by a wide margin of the essential intrigue. In these applications, commonly the motivation behind characterization learning is to effectively anticipate the minority class. Class imbalance has been accounted for to hamper the execution of standard characterization models, whose point is typically to advance the general exactness. This review paper gives an outline of imbalanced information in unverifiable NN choice govern and furthermore assesses the current techniques by methods for favourable position, restriction and execution measure. Moreover, studies assesses the real concerns looked by the current techniques in unstructured information arrangement. In any case, there is much work to be done on imbalanced information utilizing indeterminate NN choice administer for conveying better result. This review paper will assist the person who reads with understanding the cutting edge in imbalanced dataset and spur more important works.

References:

[1] N.N. Rahman, D.N Davis. "Addressing the Class Imbalance Problems in Medical Datasets", International Journal of Machine Learning andComputing, vol. 3, no. 2, pp. 224-228, 2013.

[2]J. Gavan, K. Paul, J. Richards, C. A. Dallas, H. Van Arkel, C. Herrington, and J. J. Wagner, "System and method for detecting and managing fraud." U.S. Patent No. 9, pp. 390-418, 2016.

[3]Xiao, Liyuan, Yetian Chen, and Carl K. Chang. "Bayesian model averaging of Bayesian network classifiers for intrusion detection." *Computer Software and Applications Conference Workshops (COMPSACW), 2014 IEEE 38th International*. IEEE, 2014.

[4]Guo, Yue, and Heng Zhen Zhang. "Oil spill detection using synthetic aperture radar images and feature selection in shape space." *International Journal of Applied Earth Observation and Geoinformation*, vol. 30, pp. 146-157, 2014.

[5]S. Ertekin, J. Huang, C. Lee Giles. "Adaptive Resampling with Active Learning", Technical Report, Pennsylvania State University, 2009.

[6]H. He, E.A. Garcia. "Learning from imbalanced data", IEEE Transactions

on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, 2009.

[7] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, and Y. Zhou,“A novel ensemble method for classifying imbalanced data,” *Pattern Recognition*, vol. 48, no. 5, pp. 1623-1637, 2015.

[8] Diaz, Fernando, Bhaskar Mitra, and Nick Craswell. "Query expansion with locally-trained word embeddings." *arXiv preprint arXiv:1605.07891*, 2016.

[9] Kuncheva, L. I., Arnaiz-González, Á., Díez-Pastor, J. F., & Gunn, I. A. “Instance Selection Improves Geometric Mean Accuracy: A Study on Imbalanced Data Classification.,” *arXiv preprint arXiv:1804.07155*, 2018.

[10] Urvesh Bhowan, Mark Johnston and Mengije Zhang“Developing New Fitness Functions in GeneticProgramming for Classification With Unbalanced Data”IEEE Transaction on system, man and cybernetics—partb, volume 42, pp 406-421, 2012.

[11] A. Orriols, “evolutionary rule based system for dataset,”in springer verlag soft comput., pp. 213-225, 2008.

[12] Le-Khac, N. A., O’Neill, M., Nicolau, M., & McDermott, J. “Improving fitness functions in genetic programming for classification on unbalanced credit card data,” In *European Conference on the Applications of Evolutionary Computation* (pp. 35-45). Springer, Cham, 2016.

[13] K.-J. Wang, B. Makond, K.-H. Chen, and K.-M. Wang, “A hybrid classifier combining SMOTE with PSO to estimate 5-year survivabilityof breast cancer patients,” *Appl. Soft Comput.*, vol. 20, pp. 15–24, Jul.2014.

[14] Chen, H. L., Yang, B., Wang, G., Liu, J., Xu, X., Wang, S. J., & Liu, D. Y. “A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method.,” *Knowledge-Based Systems*, *24*(8), 1348-1359, 2011.

[15] Li, Yuxuan, and Xiuzhen Zhang. "Improving k nearest neighbor with exemplar generalization for imbalanced classification." In Pacific-AsiaConference on Knowledge Discovery and Data Mining, pp. 321-332, 2011.

[16] Zhu, Yujin, Zhe Wang, and Daqi Gao. "Gravitational fixed radius nearestneighbor for imbalanced problem." Knowledge-Based Systems, vol. 90,pp. 224-238, 2015.

[17] Zhang, Xiuzhen, and Yuxuan Li. "A positive-biased nearest neighbouralgorithm for imbalanced classification." In Paci c-Asia Conference onKnowledge Discovery and Data Mining, Springer Berlin Heidelberg, pp.293-304, 2013.

[18]H. Dubey, V. Pudi, Class based weighted k-nearest neighbor over imbalance dataset, in: Advances in Knowledge Discovery and DataMining, Springer, 2013, pp. 305–316, 2013.

[19]Song, Yang, Jian Huang, Ding Zhou, Hongyuan Zha, and C. Lee Giles. "Iknn: l ormative k nearest neighbor pattern classification." In EuropeanConference on Principles of Data Mining and Knowledge Discovery, pp.248-264. Springer Berlin Heidelberg, 2007.

[20]E. Kriminger, J. Principe, C. Lakshminarayan, Nearest neighbordistributions for imbalanced classification, in: Proceedings of theInternational Joint Conference on Neural Networks, 2012, pp. 1–5, 2012.

[21] J. Maillo, S. Ramírez, I. Triguero, and F. Herrera, “kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data,” *Knowledge-Based Systems*, vol. 117, pp. 3-15, 2017.

[22] S. Saryazdi, B. Nikpour, and H. Nezamabadi-Pour, "NPC: Neighbors' progressive competition algorithm for classification of imbalanced data sets." *Intelligent Systems and Signal Processing (ICSPIS), 2017 3rd Iranian Conference on*. IEEE, 2017.

[23] Y. Xu, Y. Zhang, J. Zhao, Z. Yang, and X. Pan, “KNN-based maximum margin and minimum volume hyper-sphere machine for imbalanced data classification,” *International Journal of Machine Learning and Cybernetics*, pp. 1-12, 2017.