AN OPTIMISTIC OVERSAMPLING USING K-MEAN AND NEAREST NEIGHBOR

Abstract

The popularity of Oversampling techniques is increasing day by day. In this world of big data, the data is generated with high velocity, volume and unstructured. The user interested data is picked from large pool of heterogeneous data and it is in improper format. The reasons for improper data are inconsistency, redundancy, imbalance, missing data, disjunct data, overlapping and so on. Among them imbalance is the biggest problem. The sampling methods are best to overcome imbalance problem. Many sampling techniques are been defined and implemented among them oversampling techniques shows good accuracy but with high complexity. In this work a new optimistic oversampling method is introduced KMNN with combination of K-Mean and Nearest Neighbor with less and best time complexity. The imbalance data is collected and from that weaker class (less number of required samples) is picked. On that weaker class K-Mean operation is performed which generate K clusters. From each cluster a center element is identified and from that center element a nearest neighbor is identified. The required oversampled elements are generated in between center of each cluster to the nearest neighbor in equal ratio. Now the class becomes balance with less time complexity than previous proposed methods.

Keywords: Imbalance, Big data, weaker class, stronger class, balance.

Authors

Shaheen Layaq
Department of Computer Science
Singareni Women's Degree & PG College
Kothagudem, Telangana, India

B. Manjula
Department of Computer Science,
Kakatiya University
Warangal, Telangana, India

I. INTRODUCTION

The big data analytics [1] are playing a vital role in processing large amount of data. To process or classify this large data it is not in proper format. Whenever the user interested data is been collected most of the times there are facing with the problem of imbalance [2-5]. An imbalance [6] issue is observed when one class has more elements known as stronger class and other with less number of elements known as weaker class. When user picks sample of interested data consisting of this variation leads to imbalance. If classification is done on this imbalance data it generates to in accurate result. As the classification accuracy depends on the count. The stronger class has more count than the weaker class the result shows good accuracy for stronger class. In real life applications the weaker class is more important than the stronger class. The weaker class can't be just ignored because of less count. To overcome it the count of the weaker class has to improve. The class count is increased perfectly by sampling methods.

The sampling methods are divided into three types oversampling, undersamping and hybrid sampling. The oversampling methods are gaining more popularity than other. During the process of oversampling the weaker class is considered and count of it is increased. Many oversampling methods are proposed to overcome imbalance problem. Some of the existing oversample methods are discussed here and there problems also are also considered. A new optimistic algorithm is proposed to overcome the problem.

II. LITERATURE REVIEW

Many researchers had contributed their efforts in proposing oversampling methods. Most popular methods are SMOTE, CIR and IDROS.

A work related to SMOTE[7]: Synthetic minority over-sampling technique was presented by N.V. Chawia. Here weaker class is considered and a random element is selected from it. New random samples are generated in between them. Total accuracy depends on initial random element and faced with the problems of noisy, outliers, boundary elements, unnecessary elements and time complexity is more.

The Class Imbalance Reduction (CIR)[8] is contributed by B. Kiran Kumar, J. Gyani and Narsimha. G [8]: The mean of the weaker class is calculated and only one nearest neighbor is selected. The required oversampled are generated in between center and first nearest neighbor. The generated random oversamples were found biased, overlapped , dense and time complexity is more..

An IDROS[9] innovative oversampling method for imbalanced data reduction was proposed by B.Manjula and Shaheen Layaq. They gave importance to the weaker class and oversampled it by performing two operations finding of center (mean) and from that center KNN are identified. Among that center and KNN oversampled elements were generated in equal ratio. It was found slow and has more time complexity.

III. PROPOSED METHOD

To overcome imbalance problem an optimistic oversampling method is proposed. The proposed optimistic oversampling method KMNN is mash up of k-Mean and Nearest Neighbor. The proposed framework is shown below Figure 1. The weaker class data is collected. On that weaker class K-Mean operation is performed which generate K clusters. From each cluster a center element is identified using mean and from that center element a first nearest neighbor is identified. The required oversampled elements are generated in between center of each cluster to the nearest neighbor in equal ratio. Now weaker and stronger classes are balanced. On this balanced dataset any classification can be performed which yields accurate result with less and best time complexity.

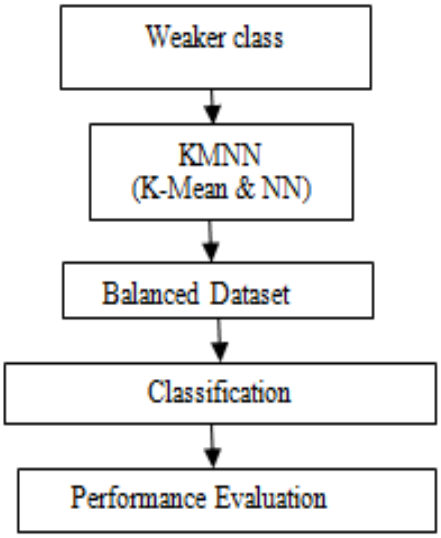

Figure 1: Proposed Framework

Suppose the weaker class contains 1000 elements, K (user specified value) is equal to 2 and required oversample elements are 20. First, K-Mean operation is performed which generate two cluster1 with 500 elements and cluster2 with 500 elements. Now consider cluster1, a center element is calculated and first nearest neighbor from center is spotted and ten random oversampled are generated from them (center and NN). Similarly, next ten oversamples are generated from cluster2 respectively. Now the class becomes balance with less time complexity than previous proposed methods. The detailed steps of KMNN algorithm are given below in Algorithm 1.

Algorithm 1: K-Mean and Nearest Neighbor

[The KMNN oversampling operation is performed to increase the count of weaker class using mash up of K-Mean and NN by which optimum and best time complexity is obtained]

Step 1: The weaker class of data is collected and denoted as W_{DS}.

Step 2: [The K-Mean operation is performed on the weaker class which generates K clusters]

$K_{Clusters}$ <- K-Mean(W_{DS})

Step 3: Required half oversamples are calculate
half <- oversample/2
Step 4: [for each k cluster center is calculated and stored in $Center_i$ .Nearest neighbor for $Cluster_i$ is calculated and stored in NN. Total half number of random oversamples are generated from $Cluster_i$ and added to $Cluster_i$]
For i 1 to k //loop repeats for K clusters
$Center_i$ <- Mean ($Cluster_i$)
For j 1 to count($Cluster_i$) // Nearest Neighbor is generated
NN <- Distance($Center_i$, D_i)
For s 1 to half // half required number of random oversample data is generated from $Cluster_i$
Random numbers R_s are generated and added to $Cluster_i$
$Cluster_i$ + (NN+ R_s* $Center_i$) //now $Cluster_i$ is oversampled with half number of data.
Step 5: End

Algorithm 1 Proposed KMNN Algorithm

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In terms of time complexity it is been analysis and found that KMNN algorithm has less time complexity than other which is shown in Table 1.

Table 1: Comparison Table

S.No	Oversampling Algorithm	Time Complexity
1	SMOTE	O(n)
2	CIR	O(n)
3	IDROS	O(n)
4	KMNN	O(log n)

V. CONCLUSION AND FUTURE WORK

Big data is facing a most challenging problem known as imbalance. Many works are been proposed to overcome it. Among many sampling methods oversampling methods stood better. The SMOTE, CIR and IDROS are few oversampling methods which are considered.

The time complexity required to make imbalance to balance using SMOTE, CIR and IDROS is more. To overcome it a new optimistic method is proposed KMNN. The KMNN it is mash up of K-Mean and NN. On weaker class K-Mean operation is performed which generate K clusters. From each cluster a center element is identified using mean and from that center element a first nearest neighbor from center is identified. The required random oversampled elements are generated in between center of each cluster to the nearest neighbor in equal ratio. Now weaker and stronger classes are balanced. On this balanced dataset any classification can be performed which yields accurate result with less time complexity. In future more innovative and optimistic methods can be combined to reduce still the time complexity.

REFERENCES

[1] Arockia Panimalar.S , Varnekha Shree.S, Veneshia Kathrine.A, “The 17 V’s Of Big Data” International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 09 , Sep-2017
[2] T.Munkhdalai, Oyun-Erdene Namsrai and K.H. Ryu,” Self-training in significance space of support vectors for imbalanced biomedical event data,” BMC Bioinformatics,Vol.16 (Suppl 7): s6, April 2015
[3] Z.Goa,L.Zhang,M.-Y.Chen,A.G.Hauptmann,H.Zhang and A.-N.Cai,”Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset,” Multimed. Tools Appl.,Vol. 68,Issue 3, 2014, pp. 641-657.
[4] S.Razakarivony and F.Jurie,” Vehicle detection in aerial imagery: a small target detection benchmark,” Journal of visual communication and image representation, vol. 34, 2016, pp.187-203.
[5] M.J.Siers and Md Zahidul Islam,”Software defect prediction using a cost sensitive decision forest and voting and a potential solution to the class imbalance problem,” in Information Systems, vol. 51, 2015, pp. 62-71.
[6] N.Japkowicz and S.Stephen,” The class imbalance problem: a systematic study,” Intelligent Data Analysis, vol. 6, issue 5, 2002, pp.429-449.
[7] N. V. Chawia, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, “Smote:Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321-357.
[8] K.B. Kiran, J.Gyani and G. Narsimha,” Class Imbalance Reduction (CIR): A Novel Approach to Software Defect Prediction in the Presence of Class Imbalance” *Symmetry* 2020, 12, 407.
[9] B.Manjula and Shaheen Layaq, “An Innovative Over Sampling Method for Imbalanced Data reduction”, Journal of Advanced Engineering Science,vol. 54, issue 01, 2022, pp. 891-899.

