International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Mahalanobis Distance-based Over-Sampling Technique

Mandar.M.Kulkarni and Madan.U.Kharat

Department of Computer Engineering, MET Bhujbal Knowledge City, Adgaon, Nashik, Maharashtra 422003

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

In data classification technique data is distributed among multiple classes. The varying structure of data distribution over multiple classes generates the skewness in data. The skewness in data represents the data imbalance. The imbalance dataset faces problem in data classification and hampers the classification accuracy. The major issue faced for minority class classification. Number of techniques has been proposed for balancing the dataset without hampering the classification accuracy of majority class. Adaptive Mahalanobis Distance-based Over-sampling (AMDO) is a over-sampling strategy. It works on mixed-type data sets. In the proposed approach the efficiency of AMDO technique is improved with the help of Principle Component Analysis (PCA) technique. This technique uses GSVD (Generalized Singular Value Decomposition) for mixed-type data. The experimental analysis will be performed on multiple multi-class imbalanced benchmarks datasets. The system performance is measured in terms of accuracy and execution time.

Keywords: *Imbalance data, data skewness, oversampling, Mahalanobis distance, hybrid data*

Introduction

The skewness in data represents the imbalance data. The skewness in data occurs due to rare events, unusual patterns and abnormal behavior of system generated data. Due to the imbalance data distribution over multiple classes the classes are divided in majority and minority classes. The majority class is the one having highest occurrence probability whereas minority class is the class having very few numbers of instances as compared to the majority class. The imbalance dataset faces problem in data classification and hampers the classification accuracy. The major issue faced due to minority class in classification process. The main purpose of data imbalance learning is to provide high classification accuracy for minority classes without hampering the accuracy of majority class. The data set having only two classes is called as binary class data. In such type of data if imbalance data occurs then one class represents the majority class and the other one represents the minority class. Lot of work has been done for balancing the binary class data. The re-sampling technique generates the dummy samples for minority class or the other solution for binary class imbalance data handling is to shift the classifier towards the minority class. The solution for binary class imbalance problem is not directly applicable for multi-class imbalance data handling where data set contains more than 2 classes and more than one class can be a minority class. In multiclass dataset the relation among classes in a dataset is not oblivious. The data boundaries of more than one class may overlap. If the two class solution is applied to the multiclass imbalance problem then it may suffer from low performance issue.

The solution for imbalance data handling is classified in mainly in two categories: data level and algorithmic level. The data level solution is pre-processing task. Before applying the mining algorithm, the dataset is balanced. Due to the data balancing the skewness in data is get reduced. The data level imbalance handling is classified in 2 categories: over sampling and under sampling. In over sampling the number of instances of minority class is increased up to the certain level to make the dataset balanced. The dummy instances are created and added to the dataset. In case of under sampling, the instances of majority class are removed to balance the dataset. Each technique has its own advantages and disadvantages. The oversampling may face a problem like: Overfitting and over-generalization, whereas in undersampling process, useful information may be loosed and may mislead the mining algorithms.

The other solution of balancing the dataset is algorithmic level. It is also called as model based solution. One class learning and ensemble learning are the way of dealing imbalance problem. Bagging, Boosting, SMOTE-Boost and Rare-Boost are the popular solutions for ensemble learning. In one class learning the training mechanism is modified so that classification accuracy of minority class can be increased. This method learns the model using single class.[3]

For uneven data distribution data balancing is the solution. For data balancing oversampling and under-sampling are two techniques in data level solution. The oversampling outperforms when there is a strong imbalance in dataset. The oversampling is a preprocessing task. After data preprocessing the balanced data is provided to the classifier. The learning process of classifier can be improved using balanced dataset, generated using oversampling. The under-sampling generates better results when dataset do not contain significant differences and low imbalance. In the following section II various oversampling techniques are discussed based on the technique studied the problem definition is formulated in section III. A new system is proposed to overcome the drawbacks of existing system in section IV.

Literature Survey

This work focuses on re-sampling techniques. The resampling is classified in 3 main sections: Oversampling, under sampling and hybrid. The easiest way of balancing data is simply duplicate the samples of minority class. This technique may cause overfitting. The alternative solution is to remove majority class instances. But this technique causes loss of information. The decision of sample selection, duplication or removal needs some criteria. The need of sample filtration is discussed in C.-X. Jian, J. Gao, and Y.-H. Ao[4].

The common solution for multiclass imbalance data handling is to convert the multiclass problem in to the two class sub-problems. It can be done in two ways: One versus One (OVO) and one-versus-all (OVA). Fern_andez et al.[5] had verified the re-sampling technique with various experiments and analyzed the good behavior in terms of costsensitive learning. An ensemble approach containing OVO and OVA is proposed by Z.-L. Zhang[6]. This technique also shows promising results over individual OVO and OVA technique.

Unlike OVO and OVA, Wang and Yao[7] proposes a new technique. This technique does not decompose data in binary subclass problem. It directly learns from whole dataset for multi-class imbalanced problems. It deals with two types of imbalance class problems: multi-minority and multi-majority cases. Three different ensemble methods are proposed based on the correlation analysis and performance pattern analysis.

Dynamic sampling procedure (DyS) is proposed by Lin et al. [8]. This is a neural network based system. It applies DyS with multilayer perceptrons (MLP). This system selects informative data dynamically and trains the MLP. It implies the iterative process and calculates the probability of sample being selected for training.

The another widely used approach is SMOTE[9]. This is oversampling technique. It selects minority class sample and finds its k nearest neighbors and finds the difference between them. A random value between [0,1] is selected and multiplied to the difference to calculate the new sample values. The SMOTE do not preserve the class covariance structure and increased the overlapping boundaries among multiple classes in case of multiclass imbalance problem.

To overcome the limitations of SMOTE, various updated versions of SMOTE are proposed. The solution includes: Safe-Level-SMOTE , Borderline-SMOTE , Smote + Tomek , SMOTE + ENN[10]. These techniques mainly deal with the overlapping of instances of multiple classes. These techniques generate better oversampling results for few datasets containing few minority classes. For other highly imbalanced datasets these algorithms leads to increase the false positive rate of the learning algorithms.

Like SMOTE, a growing ring self-organizing map(GRSOM) is a new criteria proposed by Chetchotsak et al.[11]. This is unsupervised algorithm. It executes two algorithms GRSOMO and GRSOMU. GRSOMO is an oversampling technique whereas GRSOMU is an undersampling technique.

RACOG and wRACOG are two probabilistic oversampling approaches proposed by Das et al. [12]. These two approaches use Gibbs sampling and joint probability distribution of data attributes to generate new attributes for minority classes. These sampling strategies are more beneficial as compared to the SMOTE in case of multiclass imbalance problem because these strategies uses class properties individually and mutual relation among classes.

Mahalanobis Distance-Based Over-Sampling Technique is recently proposed by Abdi and Hashemi[13]. This is a data level solution which generates the samples for minority classes. The MDO outperform in case of multiclass imbalance problem having class overlapping. MDO keep same Mahalanobis Distance from corresponding class means. And hence avoids the overlapping of classes. The proposed technique only works with numerical dataset.

To overcome the problem of oversampling of hybrid dataset a new technique is proposed by X.Yang , Q. Kuang , W. Zhang, and G. Zhang[1]. This technique applies oversampling on mixed type dataset. It uses Generalized Singular Value Decomposition for mixed type data. In sample generation process each attribute value of every sample need to be calculated. This process is time consuming in case of high dimensional dataset.

Problem Formulation

The existing work focuses mainly categorized in 2 sections: algorithmic level data balancing and data level data balancing. The algorithmic level solution is incorporated in data analysis algorithm and hence it takes high time in training phase and it degrades the efficiency of mining algorithms. The data-level solution is further classified in 2 sections: oversampling and under-sampling. When there is low imbalance in data then under-sampling is good technique whereas for strong imbalance in dataset oversampling is the ultimate solution.

Mahalanobis distance technique is a better solution for multiclass imbalance data handling solution. This is a oversampling technique. It preserves the class co-variance structure. This solution can also be applied over mixed type dataset but the efficiency of technique is inversely proportional to the number if attributes in a data. As the number of attributes increases the time required for calculating the sample value also increases. Dimensionality reduction is a technique that reduces the attributes in the dataset and improves the efficiency of processing. There is need of system that generates samples based on Mahalanobis distance along with dimensionality reduction technique to improve system efficiency on mixed type datasets.

Proposed Methodology

The proposed approach deals with mixed data-type. The system works on numerical values and hence initially the categorical data is converted in the numerical form using Heterogeneous Value Difference Metric HVDM [13] and Generalized Singular Value Decomposition (GSVD)[14] techniques. Using these techniques the system transfers the original mixed type instances in to the PC space.

The AMDO with dimensionality reduction generates synthetic samples by maintaining the same Mahalanobis distance by their corresponding class mean. This technique also optimizes the MDO performance by applying partially balancing the class distribution. The dimensionality reduction technique improves the system performance by reducing the system calculations.

A. Architecture

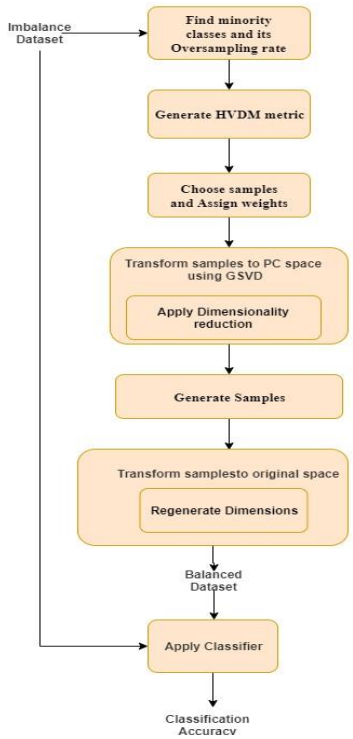

Fig 1: System Architecture

B. Preliminaries:

1. VDM distance: This is a Value Difference Metric distance calculated for nominal values. This is calculated as:

$$\sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|^{q} \quad (1)$$

Where,

$N_{a,x,c}$ = Number of instances that have value x for attribute a and class label c

C = Total number of classes q = Constant value (usually set to 1 or 2)

q = Constant value (usually set to 1 or 2)

2. Normalized Difference Distance:

The normalized difference for attribute a between two vector x and y is calculated as:

$$\text{Normalized} - \text{diff}_a(x, y) = \frac{|x-y|}{4\sigma_a} \quad (2)$$

Where, σ is the standard deviation value of attribute a

3. HVDM Distance:

This is a Heterogeneous Value Difference Metric distance. The distance between two vectors x and y can be calculated as:

$$\text{HVDM}\,(x,y) = \sqrt{\sum_{a=1}^{m} (d_a)^2 (x_a, y_a)} \quad (3)$$

Where, m = number of attributes da(x,y) = distance between two values of attribute a from vectors x and y. This cal be calculated as: da(x,y) = 1 if x and y is unknown da(x,y) = normalized-vdma(x,y) if a is nominal da(x,y) = normalized-diffa(x,y) if a is linear

4. GSVD: This is a Generalized Singular Value Decomposition. GSVD is used to transform the mixed type data to the PC space. The GSVD provides matrix decomposition of Xs using positive definite square matrices N and M. N and M are diagonal metrics of weights of rows and columns.

5. Mahalanobis distance based Sample Generation: The new samples can be generated based on ellipse contours by solving the following equation:

$$\frac{x1^2}{\alpha V1} + \frac{x2^2}{\alpha V2} + \cdots + \frac{xd^2}{\alpha Vd} = 1 \quad (4)$$

Where V is the vector of coefficients extracted from diagonal elements of covariance metrics structure of Xs and α is constant value. C. System Description: The system has imbalanced dataset as an input. The dataset contains mixed type of data. By analyzing the data and its structure minority classes are selected. For minority classes new samples are generated based on the

Mahalanobis distance. Using Mahalanobis distance the new instance is created within the eclipse contours. This preserved the class co-variance structure and reduces the overlapping entries of instances among multiple classes. Initially using dataset it is metadata is analyzed. and its attribute information is extracted. The Attribute contains numeric as well as nominal attributes. Based on the dataset metadata class distribution information is also extracted and minority classes are enlisted. Based on the instance count in minority classes and majority class the oversampling rate is defined using Partially Balanced Resampling algorithm. This algorithm finds the orate for minority classes. The dataset has numeric as well as nominal attributes. Thenominal attribute data cannot be directly processed hence the minority class data is converted in HVDM matrix for further processing. To generate new samples for minority class, few samples are selected from the class as a representative of that class based on the weight. The weight is assigned based on the count of nearest neighbors. The selected samples are used for further processing. The selected samples are transformed to the PC space after normalization of samples. For normalization mean and standard deviation is calculated initially and then data is normalized. At the time of conversion to the PC Space the dimensions are reduced using PCA dimensionality reduction technique. The new samples are generated in the PC space using ellipse contours based Mahalanobis distance function. The generated samples are then transformed to the original space and it is original dimensions are regenerated using PCA dimension restoration process. The above procedure is executed for all minority classes and balanced dataset is generated. The classification accuracy of the dataset is measured using c4.5 classifier. The accuracy comparison is done between original and updated datasets.

D. Algorithms

Algorithm 1: AMDO: Adaptive Mahalanobis Distance-based Over-sampling

Input: S: dataset,
A: Attribute Metadata
K1, K2: constants Output: Sup: updated dataset Processing:

1. Initialize variables: c: set of classes , p1: numeric attributes , p2: nominal attributes, m: total number of levels of each nominal attribute and D: Class distribution , n_{maj} : number of samples in majority class
2. Calculate Orate of class c1, c2,..,cn-1 using Partially Balanced Resampling algorithm
3. For s =1 to cn-1
4. Xs: Read samples of class s
5. For each sample i in Xs
6. Compute the K2 nearest neighbors of each Xs using HVDM matrix
7. Num(i) = number of nearest neighbors of selected sample i which has the same nearest neighbor i.
8. Weight(i) = num(i)/ K2
9. If Num(i)<K1
10. Remove i from Xs
11. End
12. End
13. Obtains Xsup = Transform P2 columns of Xs to m columns
14. Calculate means μ_s and standard deviation σ_s of samples Xsup
15. Normalize the sample set as:

Xsup = $\frac{Xsup-\mu_s}{\sigma s}$

16. Obtain matrix N and X from Xsup
17. Update Xsup = $N^{1/2}XsupM^{1/2}$
18. Compute V via GSVD of Xsup using N and M
19. Obtain diagonal vector of coefficients of matrix V
20. Reduce Dimensions from p1 to p11
21. If Orate if class s>0 then
22. For i=0 to Orate
23. Choose random sample from XsupV
24. Compute α
25. Generate p11 + m positive random numbers
26. $r1,..r_{p11+m}$ and make them sum to one
27. for k= 1 : p11 + m
28. calculate Xk = $\frac{Xk^2}{\alpha Vk}$
29. end
30. Add Xk to Xsnew
31. End
32. End
33. Regenerate Dimensions from p11 to p1
34. Xsnew = $\sigma_s(N^{-1/2}XsnewV^TM^{(-1/2)} + \mu)$
35. End
36. Update dataset Sup = S+Xsnew
37. Return Sup

Algorithm 2: Partially Balanced Resampling

Input: c: set of classes

D : class distribution Output: Orate of class c1, c2,..,cn-1 Processing:

1. Initialize n_{min} = number of samples in minority class, N: number of attributes
2. Maxit $= \frac{(n_{max}(cn-1))}{n_{min}}$
3. Initialize D* = D, N* = N 4. For i = 1 to maxit
5. N_{temp} = current size of c_{min}
6. P_{min}= n_{temp}/ N* as current minimum ratio
7. If $P^{min} <= \frac{2}{3c-1}$
8. Ntemp = Ntemp+ncmin
9. end
10. update D* and N*
11. end
12. calculate Orate by comparing D* and D
13. return Orate of class c1, c2,..,cn-1

Result and Discussions

The system is implemented in java using jdk 1.8 on widows system with 4GB ram.

A. Datasets:

The datasets are downloaded from UCI repository [15].Following table 1 represents the details of dataset:

Table 1 Dataset description

Sr. No.	Dataset	Number of Instances	Number of Attributes(numerical/ nominal)	Class Distributio n
1.	Dermatology	366	34(34/-)	112/61/72/ 49/52/20
2.	Contraceptive	1473	9(6/3)	629/333/5 11
3.	Thyroid	7,200	21(6/15)	166/368/6, 666
4.	Flare	1066	11(-/11)	331/239/2 11/147/95/ 43

B. Performance Measure:

The system performance is calculated using classification accuracy and time required for processing.

Accuracy: For accuracy evaluation C4.5 algorithm is used. The classification is applied on original as well as on reduced dataset and accuracy is compared

Time : The time required for processing is captured.

c. Implementation Status:

Initially the C4.5 algorithm is implemented and the accuracy of original dataset is calculated. Following table show the classification accuracy of imbalanced dataset.

Table 2 Classification results

Sr. No.	Dataset	AUC of Dataset	AUC as per Class Distribution
1.	Dermatology	94.53	0.97/0.78/0.94/0.87/0.96/0.86
2.	Contraceptive	51.25	0.60/0.36/0.45
3.	Thyroid	98.61	0.95/0.90/0.99
4.	Flare	74.29	1.0/0.46/0.55/0.99/0.33/0.14

In the above table II AUC of complete dataset as well as the AUC of each class in dataset is mentioned. Due to imbalanced dataset, the minority class has very low AUC value as compared to the majority class.

Partially Balanced Resampling is implemented and oversampling rate of each class is calculated. After complete implementation

1: time required for processing and 2: AUC and AUC as per Class Distribution will be calculated with modified balanced dataset.

Conclusions

The proposed system generates balanced dataset using Mahalanobis distance based oversampling technique. The proposed approach deals with mixed data-type. It generates synthetic samples by maintaining the same Mahalanobis distance by their corresponding class mean. This technique also optimizes the MDO performance by applying partially balancing the class distribution. The dimensionality reduction technique improves the system performance by reducing the system calculations. In future system can be implemented with ensemble learning to improve system efficiency and accuracy.

References

[1]. Xuebing Yang , Qiuming Kuang , Wensheng Zhang, and Guoping Zhang, "AMDO: An Over-Sampling Technique for Multi-Class Imbalanced Problems", IEEE Trans. Knowl. Data Eng., vol.30, no. 9, pp. 1672 - 1685 Sept. 2018

[2]. L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques," IEEE Trans. Knowl. Data Eng., vol. 28, no. 1, pp. 238-251, Jan. 2016.

[3]. L.-X. Duan, M.-Y. Xie, T.-B. Bai, and J.-J. Wang, "A new support vector data description method for machinery fault diagnosis with unbalanced datasets," Expert Syst. Appl., vol. 64, pp. 239-246, Dec. 2016.

[4]. C.-X. Jian, J. Gao, and Y.-H. Ao, "A new sampling method for classifying imbalanced data based on support vector machine ensemble," Neurocomputing, vol. 193, no. C, pp. 115-122, Jun. 2016.

[5]. A. Fernandez, V. Lopez, M. Galar, M. J. del Jesus, and F. Herrera, "Analysing the classification of imbalacned data-sets with multiple classes: Binarization techniques and ad-hoc approaches," Knowl.-Based Syst., vol. 42, pp. 97-110, Apr. 2013.

[6]. Z.-L. Zhang, B. Krawczyk, S. Garcia, A. Rosales-Perez, and F. Herrera, "Empowering ono-vs-one decomposition with ensemble learning for multi-class imbalanced data," Knowl.-Based Syst., vol. 106, no. C, pp. 251-263, Aug. 2016.

[7]. S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," IEEE Trans. Syst. Man Cybern. B, vol. 42, no. 4, pp. 1119-1130, Aug. 2012.

[8]. M.-L. Lin, K. Tang, and X. Yao, "Dynamic sampling approach to training neural networks for multiclass imbalance classification," IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 4, pp. 647-660, Apr. 2013.

[9]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol 16, no. 1, pp. 321-357, Jan. 2002.

[10]. G. E. A. P. A. Batista, R. C. Prati, andM. C.Monard, "A study of the behaviour of several methods for balancing machine learning training data," SIGKDDExplor. Newsl.,vol. 6, no. 1, pp. 20-29, Jun. 2004.

[11]. D. Chetchotsak, S. Pattanapairoj, and B. Arnonkijpanich, "Integrating new data balancing technique with committee networks for imbalanced data: GRSOM approach," Cogn. Neurodyn., vol. 9, no. 6, pp. 627-638, Dec. 2015.

[12]. B. Das, N. C. Krishnan, and D. J. Cook, "RACOG and wRACOG: Two probabilistic oversampling techniques," IEEE Trans. Knowl. Data Eng., vol. 27, no. 1, pp. 222-234, Jan. 2015.

[13]. D. R.Wilson and T. R.Martinez, "Improved heterogeneous distance functions," J.Artif. Intell. Res., vol. 6, no. 1, pp. 1-34, Jan. 1997.

[14]. M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco, "Multivariate analysis of mixed type data: The PCAmixdata R package," 2014. [Online]. Available: http://arxiv.org/abs/1411.4911

[15]. Dataset: http://mlr.cs.umass.edu/ml/datasets.html