ISSN 2250-3153

Minority Oversampling Technique for Imbalanced Data

Date Shital Maruti
Department of Computer Engineering
Matoshri College of Engineering
Nashik, India
shitaldate@gmail.com

Abstract— Today, solving imbalanced problems is difficult task as it contains an unequal distribution of data samples among different classes and poses a challenge to any classifier becoming hard to learn the minority class samples. Unequal distribution of data samples among many classes confuses supervised learning based classifier as it makes challenging to learn minority class samples.Generating synthetic minority class samples tries to balance the sample distribution between minority and majority classes.

To handle imbalanced learning problem, proposed work finds minority samples which are difficult to learn and computes Euclidean distance between nearest majority class samples. Using clustering approach and weighted minority class samples it generates synthetic samples for oversampling purpose. Proposed work will evaluate this approach on real & artificial datasets.

Keywords— *Imbalanced learning, undersampling, oversampling, synthetic sample generation, clustering.*

INTRODUCTION

The class imbalance problem is one of the new problems that emerged when machine learning matured from an embryonic science to an applied technology, amply used in the worlds of business, industry and scientific research. Although practitioners might already have known about this problem early, it made its appearance in the machine learning or data mining research circles about a decade ago. Its importance grew as more and more researchers realized that their data sets were imbalanced and that this imbalance caused suboptimal classification performance.

The actual cause for the bad performance of conventional classifiers on the minority class samples is not necessarily related to only on the between two imbalance classes. Besides, the complexity of data samples is another factor for the classifiers. If the samples of the majority and minority classes have more than one concepts than others and the regions between some concepts of different classes overlap, then the imbalance problem becomes very severe [13], [14]. Some of the approaches to deal with imbalanced learning problems are based on the oversampling methods.

A novel synthetic oversampling method, i.e., Majority Weighted Minority Oversampling Technique (MWMOTE), whose goal is generate the useful minority class samples.

LITERATURE SURVEY

Most of these works fall under four different categories:

Two different sampling methods exist in the literature. They are -under sampling and oversampling.

Under sampling methods work by reducing the majority class samples. This reduction can be done randomly in which case it is called random under sampling [24] or it can be done by using some statistical knowledge in which case it is called informed undersampling [25].

Oversampling methods add samples by generating the minority class samples. The generated samples add essential information to the original data set that may help to improve classifiers' performance. Depending on the technique of how synthetic samples will be generated, various methods exist in the literature such as Synthetic Minority Oversampling Technique (SMOTE) [15], Borderline-SMOTE [16], and Adaptive Synthetic Sampling Technique (ADASYN) [17]. Some sampling methods first use clustering to partition the data set and then apply undersampling and/or oversampling on different partitions' data [29], [30], [31]. A cluster-based oversampling method was proposed in [29], which randomly oversampled both the minority class and majority class samples in such a way that all clusters became the same size. In [30], clustering was applied within each large class for producing subclasses with relatively balanced sizes and random oversampling was applied on the minority class samples.

Chawla [31] proposed a cluster-based algorithm, called local sampling, in which the Hellinger distance measure [32] is used first for partitioning the original data set.

Thus, boosting and oversampling together provide a good option for efficiently learning imbalanced data [18], [35], [36], [37]. Oversampling is lot more useful than undersampling and oversampling dramatically improves classifiers performance even for complex data [12]. Both oversampling and undersampling are effective methods.Other methods besides the sampling-based strategy also work comparably well.

RELATED WORK

Researchers works have been done for handling the imbalanced learning problems. Most of these works fall under different categories.

Sampling methods have been shown to be very successful in recent years. Researchers are interested in sampling methods. Details of work performed on the other categories can be found[19]. In imbalanced learning, sampling

methods focus on balancing and the distribution between the majority class and the minority class samples.Although it is impossible to predict what true class distribution should be [20], [11] it is observed that classifiers learn well from a balanced distribution than from an imbalanced one [21], [22], [23]. This kind of oversampling sometimes creates very specific rules, leading to over fitting [9]. However, synthetic oversampling methods add samples by generating the synthetic minority class samples. The generated samples add essential in- formation to the original data set that may help to improve classifiers' performance. Depending on the technique of how synthetic samples will be generated, various methods exist in the literature such as Synthetic Minority Oversampling Technique (SMOTE) [15], Borderline-SMOTE [16], and Adaptive Synthetic Sampling Technique (ADASYN) [17]. Some sampling methods first use clustering to partition the data set and then apply under sampling and/or over sampling on different partitions' data [29], [30], [31]. A cluster-based oversampling method was proposed in [29], which randomly oversampled both the minority class and majority class samples in such a way that all clusters became the same size. In [30], clustering was applied within each large class for producing subclasses with relatively balanced sizes and random oversampling was applied on the minority class samples. Cieslak and Chawla [31] proposed a cluster-based algorithm, called local sampling, in which the Hellinger distance measure [32] is used first for partitioning the original data set. A sampling method is then applied to each partition and finally data of all partitions are merged to create the new data set. An oversampling method may create additional bias to classifiers, which may decrease classifiers' performance on the majority class samples. To prevent such degradation, an ensemble method, for example, boosting [33], [34], can be integrated with the oversampling method. While oversampling focuses on improving classifiers' performance on the minority class samples, boosting iteratively focuses on the hard-to-learn majority class samples. Thus, boosting and oversampling together provide a good option for efficiently learning imbalanced data [18], [35], [36], [37].Other methods besides the sampling-based strategy also work comparably well. While comparing oversampling and undersampling one natural observation favoring oversampling is that undersampling may remove essential information from the original data, while oversampling does not suffer from this problem. Regardless of the method used to counter the imbalance problem, factors such as the uncertainty of the true distribution between the samples of the minority class and majority class samples, complexity of data, and noise in data may pose a limit on the classifiers' performance [13], [14], [38].

PROBLEM DEFINITION

Synthetic oversampling methods shows to be very successful in dealing with imbalance data ,However, users study finds out some insufficiencies and inappropriateness of the existing methods that may occur in many different scenarios. Researchers describe them in this section. One main objective of the synthetic oversampling methods, for example, Borderline-SMOTE [16], is to identify the border-line minority class samples.

MATHAMATICAL MODEL

A. *Set Theory*

Let I be the closed system which belong to the,

I(Smaj,Smin,N,k1,k2,k3)where,

Input:

S_{maj}: Set of majority class samples
S_{min}: Set of minority class samples
N: Number of synthetic samples to be generated
k1: Number of neighbors used for predicting noisy minority class samples
k2: Number of majority neighbors used for constructing informative minority set
k3: Number of minority neighbors used for constructing informative minority set
M: no of clusters.

B. *Functions*

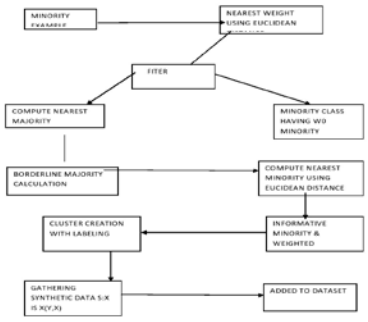
Function (F1)
F1(X)=generates nearest neighbour set.
F1(x)={x1,x2,x3…….xn}->{NN(x1),NN(x2)……....NN(x)}
Function (F2)
F2(NNi(xi))->it accepts the nearest neighbor and remove class having no minority.
F2(NNi(xi))->{Sminf = Smin – {xi € Smin } NN(xi)}
Function (F3):This function compute nearest minority set
F3(xi)- >min(x1),Nmin(x2)…min(xi)}Sminf
Function (F4):it contains Borderline majority
F4(xi)->{Ux € Sminf ,Nmaj(x)}
Function(F5):This function accepts majority example and compute nearest minority end informative minority,informative weight
F5(y1')->{Nmin (y1),Nmin(y1')}
F(xi € Smin)-> Sw(xi)->∑,€ Smaj Iw(yi)
Function (F6):it compute cluster assigns label to it
F6(xi)->Sp(xi)->Sw(xi)/∑ xi € Smin Sw(zi)->L1,L2,L3…….Lm.
Function (F7): it generates synsthetic data
F7->Somin->Somin U {S}where S->x + α * (y-x)
Function (F8): it displays oversampled minority set Somin.

IMPLEMENTATION DETAILS

A.System Architecture

This section introduces system architecture. The block diagram of system is shown in Fig 1 which gives the details of the system components.

Fig.1. Architecture Diagram.



B. **Algorithmic Strategy**

1) For each minority example xi € Smin, compute the nearest neighbor set, NN(xi).

NN(xi) consists of the nearest k1 neighbors of xi according to euclidean distance.

2) Construct filtered minority set, Sminf by removing those minority class samples which have no minority example in their neighborhood:

Sminf= Smin –{xi € Smin: NN(xi) contains no minority example}

3) For each xi € Sminf, compute the nearest majority set, Nmaj(xi). Nmaj(xi) consists of the nearest k2 majority samples from xi according to euclidean distance.

4) Find the borderline majority set, Sbmaj, as the union of all Nmaj(xi)s, i.e.,

Sbmaj = Ụ x Sminf Nmaj(xi)

5) For each majority example yi € Sbmaj, compute the nearest minority set, Nmin (yi). Nmin (yi) consists of the nearest k3 minority examples from yi according to euclidean distance.

6) Find the informative minority set, Simin, as the union of all Nmin(yi)s, i.e., Simin= Ụ yi € Sbmaj Nmin(yi)

7) For each yi € Sbmaj and for each xi € Simin, compute the information weight, Iw(yi;xi).

8) For each xi € Simin, compute the selection weight Sw(xi) as Sw(xi)=∑yi€Sbmaj Iw(yi;xi)

9) Convert each Sw(xi) into selection probability Sp(xi) according to Sp(xi)=Sw (xi)=∑zi € Simin Sw(zi)

10) Find the clusters of Smin. Let, M clusters are formed which are L1;L2;...;LM.

11) Initialize the set, Somin =Smin.

12) Do for j = 1...N.

a) Select a sample x from Simin according to probability distribution {Sp(xi)}. Let, x is a member of the cluster Lk;1 ≤ k ≤ M.

b) Select another sample y, at random, from the members of the cluster Lk.

c) Generate one synthetic data, s, according to

s = x +α (y – x),

where α is a random number in the range [0,1].

d) Add s to Somin: Somin = Somin Ụ{s}.

13) End

Loop End

CONCLUSION

The proposed over-sampling method generates useful synthetic samples for the classification of imbalanced data. Besides, the proposed over-sampling method is basically compatible with basic classification algorithms and the existing over-sampling methods.The result can be improved by increasing accuracy by the consideration of various distance formulae.

Acknowledgment

Inspiration and guidance are invaluable in every aspect of life, especially in the field of education, which I have received from our respected H.O.D. and my project guide Dr.V.H.Patil who has guided me throughout the project work and gave earnest co-operation whenever required. I would like to express sincere gratitude towards her.

I would also like to thanks all my friends who co-operated me in first two phases of seminar like information gathering and giving advice to choose correct topic. At last, I would like to take this opportunity to convey thanks to all my staff members, who directly or indirectly encouraged and helped me to complete my work on time and contributed their valuable time in helping me to achieve success in the work of project.

References

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955.

[1] P.M. M urphy and D.W. Aha, “UCI Repository of Machine Learning Databases,” Dept. of Information and Computer Science, Univ. of California, Irvine, CA, 1994.

[2] D. Lewis and J. Catlett, “Heterogeneous Uncertainty Sampling for Supervised Learning,” Proc. Int’l Conf. Machine Learning, pp. 148- 156, 1994.

[3] T.E. Fawcett and F. Provost, “Adaptive Fraud Detection,” Data Mining and Knowledge Discovery, vol. 3, no. 1, pp. 291-316, 1997.

[4] M. Kubat, R.C. Holte, and S. Matwin, “Machine Learning for the Detection of Oil Spills in Satellite Radar Images,” Machine Learning, vol. 30, no. 2/3, pp. 195-215, 1998.

[5] C.X. Ling and C. Li, “Data Mining for Direct Marketing: Problems and Solutions,” Proc. Int’l Conf. Knowledge Discovery and Data Mining, pp. 73-79, 1998.

[6] N. Japkowicz, C. Myers, and M. Gluck, “A Novelty Detection Approach to Classification,” Proc. 14th Joint Conf. Artificial Intelligence, pp. 518-523, 1995.

[7] S. Clearwater and E. Stern, “A Rule-Learning Program in High Energy Physics Event Classification,” Computer Physics Comm., vol. 67, no. 2, pp. 159-182, 1991.

[8] G.M. Weiss, “Mining with Rarity: A Unifying Framework,” ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 7-19, 2004.

[9] R.C. Holte, L. Acker, and B.W. Porter, “Concept Learning and the Problem of Small Disjuncts,” Proc. Int’l Joint Conf. Artificial Intelligence, pp. 813-818, 1989.

[10] J.R. Quinlan, “Induction of Decision Trees,” Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.

[11] F. Provost, “Machine Learning from Imbalanced Data Sets 101,” Proc. Learning from Imbalanced Data Sets: Papers from the Am. Assoc. Artificial Intelligence Workshop, 2000.

[12] N. Japkowicz and S. Stephen, “The Class Imbalance Problem: A Systematic Study,” Intelligent Data Analysis, vol. 6, no. 5, pp. 429- 449, 2002.

[13] R.C. Prati, G.E.A.P.A. Batista, and M.C. Monard, “Class Imbal- ances versus Class Overlapping: An Analysis of a Learning System Behavior,” Proc. Mexican Int’l Conf. Artificial Intelligence, pp. 312-321, 2004.

[14] T. Jo and N. Japkowicz, “Class Imbalances versus Small Disjuncts,” ACM SIGKDD Exploration Newsletter, vol. 6, no. 1, pp. 40-49, 2004.

[15] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, “SMOTE: Synthetic Minority oversampling Technique,” J. Artifi- cial Intelligence Research, vol. 16, pp. 321-357, 2002.

[16] H. Han, W.Y. Wang, and B.H. Mao, “Borderline-SMOTE: A New Oversampling Method in Imbalanced Data Sets Learning,” Proc. Int’l Conf. Intelligent Computing, pp. 878-887, 2005.

[17] H. He, Y. Bai, E.A. Garcia, and S. Li, “ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,” Proc. Int’l Joint Conf. Neural Networks, pp. 1322-1328, 2008.

[18] S. Chen, H. He, and E.A. Garcia, “RAMOBoost: Ranked Minority Oversampling in Boosting,” IEEE Trans. Neural Networks, vol. 21, no. 20, pp. 1624-1642, Oct. 2010.

[19] H. He and E.A. Garcia, “Learning from Imbalanced Data,” IEEE Trans. Knowledge Data Eng., vol. 21, no. 9, pp. 1263-1284, Sept. 2009.

[20] N.V. Chawla, D.A. Cieslak, L.O. Hall, and A. Joshi, “Automati- cally Countering Imbalance and Its Empirical Relationship to Cost,” Data Mining and Knowledge Discovery, vol. 17, no. 2, pp. 225- 252, 2008.

[21] G.M. Weiss and F. Provost, “The Effect of Class Distribution on Classifier Learning: An Empirical Study,” Technical Report ML- TR-43, Dept. of Computer Science, Rutgers Univ., 2001.

[22] J. Laurikkala, “Improving Identification of Difficult Small Classes by Balancing Class Distribution,” Proc. Conf. AI in Medicine in Europe: Artificial Intelligence Medicine, pp. 63-66, 2001.

[23] A. Estabrooks, T. Jo, and N. Japkowicz, “A Multiple Resampling Method for Learning from Imbalanced Data Sets,” Computational Intelligence, vol. 20, pp. 18-36, 2004.

[24] J. Zhang and I. Mani, “KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction,” Proc. Int’l Conf. Machine Learning, Workshop Learning from Imbal- anced Data Sets, 2003.

[25] X.Y. Liu, J. Wu, and Z.H. Zhou, “Exploratory Under Sampling for Class Imbalance Learning,” Proc. Int’l Conf. Data Mining, pp. 965- 969, 2006.

[26] I. Tomek, “Two Modifications of CNN,” IEEE Trans. System, Man, Cybernetics, vol. SMC-6, no. 11, pp. 769-772, Nov. 1976.

[27] M. Kubat and S. Matwin, “Addressing the Curse of Imbalanced Training Sets: One-Sided Selection,” Proc. Int’l Conf. Machine Learning, pp. 179-186, 1997.