An approach to improve imbalanced learning from oversampling the data of minority class

DT228A - DA
Han Tang - D16129273
Sources of Data: the UCL Machine Learning Repository
Affiliation

BACKGROUND, CONTEXT AND SCOPE

Classification is an important task as part of Machine Learning and plays an important role in pattern recognition. Many algorithms are introduced to deal with classification tasks. However, dealing with imbalanced datasets, which contain far more samples in the majority class than the minorities, is one of the greatest challenges in data mining and machine learning, as classifiers trained from imbalanced datasets would become over-fitted to the majority class and under-fitted to the minorities.

At the same time, many real-life data mining classifiers are facing the problem that their training datasets are imbalanced, such as text classification (A. Sun, Lim, & Liu, 2009), cancerous tissues identification (Mazurowskia et al., 2008), transaction fraud detection (Chan & Stolfo, 1998), etc. The unpleasant consequences and the frequency of the imbalanced learning problem make it a concerning topic for data mining in the usage of industry and research.

For the understanding of community, imbalanced learning commonly refers to the datasets exhibit significant, or even extreme imbalances, on the order of 100: 1, 1,000: 1, or even 10,000:1, as such. In other words, one class usually severely outrepresents the others in imbalanced learning (He & Shen, 2007)(Kubat, Holte, & Matwin, 1998)(Pearson, Gonye, & Schwaber, 2003).

Imbalances could not only exist in binary datasets but also could in datasets with various classes(Y. Sun, Kamel, & Wang, 2006)(Chen, Bao-Liang Lu, & Kwok, 2006)(Zhi-Hua Zhou & Xu-Ying Liu, 2006). The imbalance of subclass within a class also raises concerns.

PROBLEM DESCRIPTION

Innovative approaches are applied to solve the problem of learning from imbalanced data, which include sampling methods, cost-sensitive methods, kernel-based methods and active learning methods. Data resampling and adjusted cost-sensitive algorithms are the most frequently used methods among them. Though they respectively have their drawbacks.

Most resampling methods either inevitably bring noise or increase the risk of overfitting (Oversampling) or have risks of dropping potentially useful information (Undersampling). On the other hand, cost-sensitive learning can decrease the risk of overfitting or underfitting compares to modifying the training data, though the cost-sensitive learning needs to be specific for different machine learning algorithms.

This research means to introduce a new data undersampling technique, aims to adjust the distribution of skewed datasets by removing some data from the overrepresent class, and preserve as much information as possible at the same time.

LITERATURE REVIEW

It would naturally come into one's mind that imbalanced data refers to the distribution is imbalanced between two classes, though many imbalanced datasets' target features contain various values. This research would still focus on imbalanced datasets containing only two classes, as multiclass imbalanced data would impose more complexity. Classical methods to deal with the imbalanced data problems are compared and discussed in this review, which could mainly be classified as resampling methods, cost-sensitive methods and algorithm-level methods. Data-level methods would be the focus of this literature review as it is more commonly researched compare to the other two methods. Cost-sensitive methods and algorithm-level method would still be introduced briefly. This review would also discuss the pros and cons of each method.

Sampling Methods for Imbalanced Learning

Sampling methods apply some techniques to modify the distribution of imbalanced datasets, as classifiers trained from balanced datasets outperform those from imbalanced datasets, according to a set of studies(Weiss & Provost, 2001)(Estabrooks, Jo, & Japkowicz, 2004).

Random Oversampling and Undersampling. The mechanic of random oversampling is to simply increase the instances of the underrepresented class by randomly sending the same samples to classifiers multiple times. Similarly, random undersampling is to randomly remove samples from the over-represented class from the training dataset. Though random sampling could balance the datasets, it could bring additional problems to the training process, such as increasing the risk of either overfitting or underfitting(Mease, Wyner, & Buja, 2007).

Informed Undersampling. Two algorithms applied by Liu(Liu, Wu, & Zhou, 2009) have shown good results in informed undersampling, which are Easy Ensemble and Balance Cascade. Easy Ensemble is an ensemble learning system of multiple subsets of the whole data. Each subset composed of a subset of the majority class and the minority class. Balance Cascade applies a supervised learning approach to systematically choose the samples of the majority class to be undersampled.

Another successful and well-applied technique of informed undersampling is to use the K Nearest Neighbor (KNN) classifier to undersample the class of majority. Multiple methods of undersampling based on KNN is applied in the work of Zhang (Zhang & Mani, 2003). It turns out the method of selecting the majority class examples whose average distances to the three closest minority classes examples is the largest return the most competitive results of imbalanced learning.

Synthetic Sampling with Data Generation. Synthetic minority oversampling technique (SMOTE) is a more practical and powerful method to generate samples of minority class compares to the random oversampling method(Chawla, Bowyer, Hall, & Kegelmeyer, 2002). The SMOTE algorithm creates new samples based on the feature space similarity within the class of minority. Each new sample is created from multiplying the corresponding feature vector of a selected one of the K-Nearest neighbours with a random number between [0, 1]. Though this method proved to be capable of improving imbalanced learning effectively, its drawbacks such as over-generalisation and variance are still significant(Wang & Japkowicz, 2004).

Other sampling methods. Other approaches of oversampling also applied, based on the existed SMOTE algorithm, intend to solve the problems occurred by SMOTE, which are the inconsideration of neighbour samples between classes and the incurrence of overlapping between classes because of it. Various methods are proposed to solve the problems such as Borderline-SMOTE(Han, Wang, & Mao, 2005) and Adaptive Synthetic Sampling (ADA-SYN) algorithms(Haibo He, Yang Bai, Garcia, & Shutao Li, 2008).

Sampling with Data cleaning techniques("Two Modifications of CNN," 1976) and Cluster-based oversampling (CBO) methods(Jo & Japkowicz, 2004) are also proved to be capable of improving imbalanced learning effectively, according to studies.

State-of-the-art approaches of Resampling data. Many research on the topic of data resampling in recent years are also worth attention. In 2018, Georgios(Douzas, Bacao, & Last, 2018) introduced a variation of SMOTE which combines it with the K-means clustering algorithm. This method focuses more on the within-class imbalance compares to the existed SMOTE algorithm. On the other hand, this method also avoids noise generation effectively.

Another study by Li in 2018(L. Sun, Song, Hua, Shen, & Song, 2018) introduced a value-aware resampling method to tackle the problem of imbalanced learning. He assumes that the samples within a class are of different values to train a classifier. Thus he designs an algorithm to assign values for each sample and train the classifiers to focus more on those samples of higher value. This method of resampling was found performs better compare to existed method on 13 imbalanced learning tasks.

Cost-sensitive Methods and Algorithm-level Methods

Cost-sensitive methods consider the costs of misclassifying samples(Elkan, 2001). Though it performs better than resampling methods in some specific applications(Zhi-Hua Zhou & Xu-Ying Liu, 2006)(McCarthy, Zabar, & Weiss, 2005), its problem-specific characteristic makes it less versatile.

Algorithm-level methods are less applied for they are always bound to specific classifiers(Douzas et al., 2018)

Gaps in research

This research intends to compare existed resampling technique and aims at creating a new variation to amend existed approaches.

RESEARCH QUESTION

"Is the F1-score of the classifiers trained from imbalanced data resampled by Borderline-SMOTE method statistical significantly higher than it of the classifiers trained from imbalanced data resampled by ADA-SYN algorithm?"

HYPOTHESIS

H0: The F1-score of the classifier trained from imbalanced data resampled by Borderline-SMOTE method is not statistically significantly higher than it of the classifier trained from imbalanced data resampled by ADA-SYN algorithm.

H1: The F1-score of the classifier trained from imbalanced data resampled by Borderline-SMOTE method is statistically significantly higher than it of the classifier trained from imbalanced data resampled by ADA-SYN algorithm.

The objective of the research is to compare the effectiveness of some existed data resampling methods under the context of imbalanced learning problem. A new method of data resampling would be created in further research.

The research methodologies used are quantitative. The effectiveness of each resampling method will be measured by the F1-scores of classifiers. The results will show the difference of distribution in the result performances of models trained from imbalanced data resampled by different methods.

Secondary research is ongoing to complete a comprehensive literature review of the previous research on solving the imbalanced learning problem. The major work is to research given datasets.

DESIGN AND IMPLEMENTATION

Solution Description

1. Document research will be conducted first to identify and compare existed methods of dealing with the imbalanced learning problem. Several state-of-the-art methods will be identified and reproduced.

2. The identified methods will be reproduced on 12 imbalanced datasets from the UCL Machine Learning Repository, which are breast_disease, ecoli, glass, Haberman, heart, iris, libra, liver_disorder, Pima, segment, vehicle and wine. Same machine learning algorithms (Decision tree, Neural network, Support Vector Machine, Logistic Regression and

Naive Bayes Classifier) will be trained from the imbalanced datasets with selected data resampling methods. The data mining process will be programmed in Python. 10 fold cross-validation will be used as train-test partition to prevent the bias introduced by variances within datasets.

3. Further document research will be conducted to create a new method of oversampling data of the minority class. This method will be applied to the 12 imbalanced datasets to compare the performance of it with those generated before.

4. F1-scores of each machine learning algorithm of each resampling method of each dataset will be generated. The performances divided by resampling methods will be compared to prove or reject the hypothesis.

EVALUATION OF DESIGN

The performances of all the trained classifiers will be measured by F1-score. For 12 datasets of 5 types of algorithms, there would be 60 F1-scores for each data resampling method.

The hypothesis will be accepted if independent t-test finds out that there are statistically significant differences between the F1-score arrays ($p < 0.01$, we are 99% confident there are statistically significant differences between the two F1-score arrays).

The findings can be related to the research question as we can compare the performances of the same classifiers trained from the same datasets, but of different data resampling methods. If there are universal differences between the performances of them over many problems and many algorithms, we can say the selected data resampling method can influence the performance of imbalanced learning.

BIBLIOGRAPHY

References

Chan, P. K. & Stolfo, S. J. (1998). Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. *KDD*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi:doi.org/10.1613/jair.953

Chen, K., Bao-Liang Lu, & Kwok, J. T. (2006). Efficient classification of multi-label and imbalanced data using min-max modular classifiers. In *The 2006 ieee international joint conference on neural network proceedings* (pp. 1770–1775). doi:10.1109/IJCNN.2006.246893

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. *Information Sciences*, *465*, 1–20. doi:https://doi.org/10.1016/j.ins.2018.06.056

Elkan, C. (2001). The foundations of cost-sensitive learning. *Proc. Intfffdfffdfffdl Joint Conf. Artificial Intelligence*, 973–978.

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, *20*(1), 18–36. doi:doi.org/10.1111/j.0824-7935.2004.t01-1-00228.x

Haibo He, Yang Bai, Garcia, E. A., & Shutao Li. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 ieee international joint conference on neural networks (ieee world congress on computational intelligence)* (pp. 1322–1328). doi:10.1109/IJCNN.2008.4633969

Han, H., Wang, W.-Y., & Mao, B.-H. (2005). Borderline-smote: A new over-sampling method in imbalanced data sets learning. In D.-S. Huang, X.-P. Zhang, & G.-B. Huang (Eds.), *Advances in intelligent computing* (pp. 878–887). Berlin, Heidelberg: Springer Berlin Heidelberg.

He, H. & Shen, X. (2007). A ranked subspace learning method for gene expression data classification. (Vol. 1, pp. 358–364).

Jo, T. & Japkowicz, N. [Nathalie]. (2004). Class imbalances versus small disjuncts. *SIGKDD Explor. Newsl. 6*(1), 40–49. doi:10.1145/1007730.1007737

Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, *30*(2), 195–215. doi:10.1023/A:1007452223027

Liu, X., Wu, J., & Zhou, Z. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539–550. doi:10.1109/TSMCB.2008.2007853

Mazurowskia, M. A., Habasa, P. A., Zuradaa, J. M., Lob, J. Y., Bakerb, J. A., & Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, *21*, 427–436. doi:doi.org/10.1016/j.neunet.2007.12.031

McCarthy, K., Zabar, B., & Weiss, G. (2005). Does cost-sensitive learning beat sampling for classifying rare classes? In *Proceedings of the 1st international workshop on utility-based data mining* (pp. 69–77). UBDM '05. Chicago, Illinois: ACM. doi:10.1145/1089827.1089836

Mease, D., Wyner, A. J., & Buja, A. (2007). Boosted classification trees and class probability/quantile estimation. *J. Mach. Learn. Res. 8*, 409–439. Retrieved from http://dl.acm.org/citation.cfm?id=1248659.1248675

Pearson, R. K., Gonye, G. E., & Schwaber, J. S. (2003). Imbalanced clustering of microarray time-series. *Workshop on Learning from Imbalanced Dataset II, ICML.*

Sun, A., Lim, E.-p., & Liu, Y. (2009). On strategies for imbalanced text classification using svm: A comparative study. *Decision Support System*, *48*, 191–201. doi:doi: 10.1016/j.dss.2009.07.011

Sun, L., Song, J., Hua, C., Shen, C., & Song, M. (2018). Value-aware resampling and loss for imbalanced classification. (pp. 1–6). doi:10.1145/3207677.3278084

Sun, Y., Kamel, M. S., & Wang, Y. (2006). Boosting for learning multiple classes with imbalanced class distribution. In *Sixth international conference on data mining (icdm'06)* (pp. 592–602). doi:10.1109/ICDM.2006.29

Two modifications of cnn. (1976). *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-6*(11), 769–772. doi:10.1109/TSMC.1976.4309452

Wang, B. & Japkowicz, N. [N.]. (2004). Imbalanced data set learning with synthetic samples. *Proc. IRIS Machine Learning Workshop*.

Weiss, G. M. & Provost, F. (2001). The effect of class distribution on classifier learning: An empirical study. *Technical Report ML-TR-44, Department of Computer Science, Rutgers University*.

Zhang, J. & Mani, I. (2003). KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In *Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets*.

Zhi-Hua Zhou & Xu-Ying Liu. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, *18*(1), 63–77. doi:10.1109/TKDE.2006.17

ACTIVITIES

There would be 17 weeks in total for this research.

The first 3 weeks are for document research.

4 weeks are for comparing the effectiveness of several data resampling methods and design a new data resampling method.

4 weeks are for applying the new data resampling method and producing classifiers of this method.

The last 6 weeks are for finishing the dissertation.