Rare Class Mining: Progress and Prospect

Shuli Han, Bo Yuan, Wenhuang Liu

Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, P.R. China
E-mail: hansl08@mails.tsinghua.edu.cn, {yuanb, liuwh}@sz.tsinghua.edu.cn

Abstract: Rare class problems exist extensively in real-world applications across a wide range of domains. The extreme scarcity of the target class challenges traditional machine learning algorithms focusing on the overall classification accuracy. As a result, purposefully designed techniques are required for effectively solving the rare class mining problem. This paper presents a systematic review of the major representative approaches to rare class mining and related topics and gives a summary of the important research directions.

Key Words: Rare Class Mining, Rare Category Detection, Feature Selection

1 INTRODUCTION

The challenging issue of rare class mining is inevitable in many real-world data mining applications, such as network intrusion detection, video surveillance [1, 2], oil spills detection in satellite radar images [3], diagnoses of rare medical conditions, text categorization [4, 5], and so on. All these applications share a common characteristic: samples from one class are extremely rare, while the number of samples belonging to other classes is sufficiently large. Furthermore, the correct detection of the rare samples is of significantly greater importance than the correct classification of the majority samples.

For example, in the network intrusion detection domain, there are hundreds of thousands of access requests every day. Among all these requests, the number of malicious connections is, in most cases, very small compared to the number of normal connections. Obviously, building a good model that can effectively detect future attacks is crucial so that the system can respond promptly in case of network intrusions.

Samples from a rare class are sometimes referred to as rare events or rare objects in the literature. The major challenge comes from the fact that the rarely occurring samples are usually overwhelmed by the majority class samples so that they are much harder to be identified. Firstly, traditional machine learning algorithms usually aim at achieving the lowest overall misclassification rate, which creates an inherent bias in favor of the majority classes because the rare class has less impact on accuracy. Secondly, noisy data may look similar to the rare objects and hence are difficult to distinguish. Due to these issues, the rare class mining problem has attracted more and more attention from the research community.

In this paper, we will discuss various issues associated with rare class mining and give a systematic review of different techniques that have been proposed for effectively mining rare events. Note that the domain of rare class mining is not clearly defined in the literature, which is often mixed up with the imbalanced dataset problem. Weiss [6] presents a survey of mining with both imbalanced datasets and rare class, claiming that these two areas are closely connected, whereas in this paper we try to clarify the essential differences between them. Moreover, the emerging area of rare category detection is also discussed along with the prospects of rare class mining.

2 METHODS FOR RARE CLASS PROBLEMS

The rare class problem has been investigated from different aspects and the focus of this paper is on supervised learning methods for rare class analysis.

2.1 Evaluation Metrics

Evaluation metric plays an important role in data mining. It is not only used to evaluate the performance of a classifier but in many algorithms also guides the learning process. Although accuracy is a widely used evaluation metric, it does not work well for rare class mining issue due to its strong bias against the rare class. For example, in a classification problem where the rare class accounts for only 0.1% of the training set, a classifier that predicts every sample as the majority class can still achieve a seemingly satisfactory overall accuracy of 99.9%. However, it holds no significance for the rare class mining application as nothing about the rare class has been learnt by the classifier. Therefore, for rare class mining, the overall accuracy is not a meaningful metric. Instead, *Precision* and *Recall* are the two preferred metrics. Usually, the rare class is denoted as positive samples. According to this definition, *Precision* is the percentage of true positive samples among all samples that are identified as positive and *Recall* is the percentage of true positive samples that are correctly predicted. *Precision* measures the exactness while *Recall* measures the completeness and they are related directly to the objective of the rare class mining problem.

Since *Recall* and *Precision* are two conflicting metrics, an acceptable tradeoff depending on specific applications is often pursued. Two popular metrics seeking a good balance between *Recall* and *Precision* are GMPR [7] and F-measure [4, 8-11]. GMPR is defined as the square root of the product of *Precision* and *Recall*, while F-measure, a metric widely used in the Information Retrieval community, is defined as:

$$F_\beta = \frac{(1+\beta^2)(Precision \bullet Recall)}{\beta^2 Precision + Recall} \quad (1)$$

where β measures the importance of *Precision* vs. *Recall*. These two metrics achieve high values only if both *Precision* and *Recall* take high values. Additionally, other metrics have been proposed, such as sum of recalls, geometric mean of recalls, information score, and so on. Joshi [7] compares these metrics, and concludes that GMPR and F-measure are the most suitable metrics while F-measure is more favorable for tougher rare class problems.

2.2 Sampling-Based Methods

Sampling is one of the common data preprocessing techniques. The idea of sampling is to purposefully manipulate the distribution of samples so that the rare class could be well represented in the training set. Originally, sampling is a widely used approach to handle the class imbalance problem. Recently, a lot of studies discuss the sampling technique for dealing with rare class mining. It has been shown that sampling is an effective method for mining rare events [12].

2.2.1. Sampling Methods

The basic sampling methods include under-sampling and over-sampling. Under-sampling randomly discards the majority class samples while over-sampling randomly duplicates the minority class samples in order to modify the class distributions [13]. Although these two methods do alleviate the rare class problem to some extent, they also bring in some issues. In random under-sampling, some potentially useful majority samples may be left out, resulting in information loss and a less than optimal model. Also, in random over-sampling, the size of the training set is increased significantly, increasing the computational complexity. Moreover, since over-sampling makes exact copies of rare class samples, adding no new information to the dataset, it may cause the over-fitting problem.

Since the basic versions of sampling do not work well in practice, some heuristic sampling methods have been proposed. Kubat and Matwin [14] propose a novel sampling method called One-Sided Selection. The core idea is to only eliminate special samples of the majority class and keep all rare class samples. Samples to be discarded are those noisy, redundant or borderline ones close to the boundary separating the positive and negative regions. In the algorithm, the concept of Tomek links is introduced to recognize the noisy and borderline samples. The paper also shows that only when one of the classes is prohibitively rare could the algorithm be applied.

Liu et al. [15] propose to make effective use of the majority class by multi-sampling. Two under-sampling methods, EasyEnsemble and BalanceCascade, are presented and the main idea is to create multiple subsets from the majority class, and use AdaBoost to train a classifier based on each subset together with the rare class dataset. Finally, the outputs of these classifiers are combined. EasyEnsemble samples from the majority class with replacement while in BalanceCascade, samples that are correctly classified by previous classifiers are discarded before subsequent sampling. Empirical results suggest that, compared to EasyEnsemble, BalanceCascade is more efficient on highly skewed dataset. Chawla et al. [16] propose an approach called Synthetic Minority Over-sampling Technique (SMOTE) in which the rare class is over-sampled by creating new synthetic rare class samples according to each rare class sample and its k nearest neighbors. Each new sample is generated in the direction of some or all of the nearest neighbors. Experimental results show that SMOTE can improve the accuracy of classifiers on many rare class problems and the combination of SMOTE and under-sampling performs better than pure under-sampling.

2.2.2. Sampling Rate

While conducting sampling, there is an issue of how to determine the proper sampling rate, which directly affects the class distribution ratio. Given an imbalanced dataset problem, it is intuitive that a balanced distribution may yield the best or approximately best performance. However, it has been shown that the often used 'even distribution' is not optimal when dealing with rare events [17]. Instead, a ratio of 2:1 or even 3:1 in favor of the majority class often results in superior classification performance.

2.2.3. Other Sampling Related Questions

Seiffert et al. [12] show that sampling techniques are domain-dependent and which sampling technique is the best choice depends on the specific application. As to the negative impact of sampling techniques, Kubat and Matwin [14] show that although the classification performance of a classifier on the rare class may increase along with the adoption of sampling techniques, its performance on the majority class and its overall accuracy may drop to some extent.

2.3 Cost-Sensitive Learning

Cost-sensitive learning is a widely used technique in data mining, which assigns different levels of misclassification penalty to each class. Cost-sensitive technique has been incorporated into classification algorithms by taking into account the cost information and trying to optimize the overall cost during the learning process. Recently, this technique has been applied to the rare class problem in which a higher cost is given to the misclassification of rare objects compared to the majority class. In [18], a cost/benefit sensitive algorithm named Statistical Online Cost Sensitive Classification (STOCS) is proposed to classify rare events in online data and the results show that STOCS outperforms many other well-known cost-insensitive online algorithms.

However, it is often difficult to set the cost information in practice. Although it is well known that a false negative prediction is more risky than a false positive prediction, how to make a quantitative analysis between these two risks may require prior knowledge and domain experts' involvement. In practice, it is suggested to vary the cost ratio till a satisfactory objective function value is obtained [6]. As to multi-class problems, Sun et al. [19] apply Genetic Algorithms to search the optimal misclassification cost for each class.

2.4 Algorithms for Rare Class Mining

This section gives a review of some machine learning algorithms proposed or specifically modified for the rare class mining problem.

2.4.1. Boosting Algorithms

Boosting is a powerful sequential ensemble learning algorithm that can improve the performance of weak base learners [20]. In Boosting, a series of basic classifiers are built based on the weighted distributions of the original training set. At the end of each iteration, the weight of each training sample is adaptively changed based on the training error of the current classifier. By doing so, later classifiers are forced to put more emphasis on learning samples that are misclassified by former classifiers.

Boosting can be viewed as a generalized sampling method, as it changes the distribution of the original dataset. Since Boosting focuses on samples that are difficult to classify, it is a good choice to apply Boosting to detecting the rare class. However, since the standard Boosting algorithm treats the two kinds of errors (false positive and false negative) equally, the majority class may still dominate the training set after successive Boosting iterations.

In order to solve this issue, RareBoost [8] updates the weights of positive samples and negative samples in a different way. It allows the algorithm to focus on both *Recall* and *Precision* equally. AdaCost [21] is another variant of AdaBoost, which adopts the cost-sensitive technique. It imposes different costs for the two types of errors to update the distribution of the training set in order to reduce the cumulative misclassification cost. Chawla et al. [22] propose SMOTEBoost in which SMOTE is applied in each Boosting iteration. The newly created synthetic samples for the rare class are added into the training set to train a weak classifier, which are discarded after the classifier is built. The SMOTE procedure in each iteration makes every classifier learn more from the rare class, and thus broadens the decision regions for the rare class.

The core idea of the above algorithms is the same, which is to adaptively alter the distribution of the original dataset so that classifiers can focus on the samples that are difficult to classify. RareBoost and AdaCost change the distribution by applying modified weight updating mechanism while SMOTEBoost generates synthetic samples for the rare class. In terms of the weight updating mechanism, the standard Boosting focuses on all misclassified samples equally while RareBoost treats false-positive samples and false-negative samples differently. In the meantime, AdaCost updates the weights differently for all four types of classification outputs.

In addition to the weight updating mechanism, the effect of base classifiers has also been studied [10]. By analyzing the key components of Boosting: the accuracy metric, the ensemble voting process, the weight updating mechanism and the base learner, it is shown that, for the rare class mining problem, the performance of Boosting is critically dependent on the abilities of its base learner.

2.4.2. Rule-Based Algorithms

Traditional rule-induction techniques often fail to perform well in the rare class classification and some modified algorithms are proposed.

Joshi et al. [9] point out that existing sequential covering techniques may fail to effectively detect the rare class because they try to achieve high *Recall* rates and high *Precision* rates simultaneously but may face two problems: splintered false positives and small disjuncts caused by sparse target samples. In order to solve these issues, a two-phase rule-induction approach PNrule is introduced. In the first phase, rules that have high support and reasonable accuracy are discovered, which may contain both positive as well as negative samples. In the second phase, rules able to remove the false positive samples are developed to increase the accuracy. By doing so, PNrule is especially suitable for rare classes mining.

Emerging Patterns (EPs) are a new type of patterns introduced in [23]. They refer to itemsets whose supports in one class are significantly higher than in others, which can capture significant multi-attribute contrasts between classes. EPRC [24] is the first EP-based rare class classification approach, which aims to increase the discriminating power of EPs through three stages: generating new undiscovered rare class EPs, pruning low-support EPs and increasing the supports of rare class EPs. Alhammady and Ramamohanarao [25] propose a method called EPDT, which employs EPs to enhance the Decision Trees algorithm. It uses the rare class EPs to create new non-existing rare class samples and to over-sample the most important ones. The increase of both the rare class population and the training set size helps bias the decisions towards the rare class. DEP [26] is also a novel approach to mining EPs by dividing the majority class into subsets so that unseen rare class EPs could be discovered. It also defines a strength function to evaluate the rare class EPs in order to minimize the effect of noisy EPs.

2.4.3. Division-Based Algorithms

In some classification cases, subclass division within class is a quite normal procedure. Japkowicz [27] provides a general framework for combining unsupervised and supervised learning in classification tasks where division is implemented via clustering. Through sub-division, a complex concept may become much simpler and the imbalance level could also be weakened. Wu et al. [28] inherit the idea of sub-division by developing a method for rare class mining using local clustering. For the majority classes, local clustering is employed within each class, and for the rare class, over-sampling is adopted. The algorithm adjusts the over-sampling parameter to fit in with the clustering result so that the rare class size is approximate to the average size of the partitioned majority class.

2.5 Summary

In this section, we discuss approaches to address the rare class mining issue. These exists no universal methods for this issue and which approach should be adopted is dependent on

the specific application. In general, sampling and cost-sensitive are the most widely used techniques. In theory, cost-sensitive learning is preferable because no information is lost. However, in practice, sampling techniques are usually used as the cost information is difficult to quantify. Furthermore, how to select the best sampling technique and parameter combination is still an important research topic. In addition, modifying the traditional algorithms directly towards efficient rare class mining is also a good choice.

3 RELATED TOPICS

3.1. Rare Class vs. Imbalanced Dataset

Every dataset with uneven class distributions can be regarded as an imbalanced dataset. In this sense, rare class mining is an extreme case of the imbalanced classification problem where the minority class only accounts for 5% or even much less of the dataset. Weiss [6] demonstrates that rare class mining and imbalanced dataset classification face many common challenges and can benefit from similar remediation techniques. However, rare class mining could be much tougher than imbalanced classification as their major objectives are quite different. Class imbalance problems mainly focus on the imbalanced distributions among classes while rare class mining implies that the minority objects are abnormal and require special attention. Hence, the imbalanced classification problem still focuses on the overall accuracy and the accuracy of each class while rare class mining puts more emphasis on the detection of the rare class. In this case, the ROC curve or the G-mean metric, which is defined as the square root of accuracies of each class [19], is preferable for class imbalance problems whereas in rare class mining more targeted evaluation metrics should be used, as described in Section 2.1.

3.2. Rare Category Detection

In previous sections, we have reviewed various classification based techniques for rare class mining, which require a dataset with labeled samples for all the classes. However, in some application domains such as image processing and text categorization, labels are often difficult and time-consuming to obtain. As a result, the active learning model is proposed, which aims at labeling the most informative samples to minimize the cost of obtaining labeled data [29]. The motivation of active learning is to achieve high accuracies using as few labeled samples as possible. However, the fact that the majority classes always dominate the dataset is still a bottleneck in reducing the sample complexity of active learning techniques [30].

In order to solve this issue, the concept of rare category detection base on active learning has been proposed. Pelleg and Moore [31] present a solution by defining a mixture model to fit the data distribution. Samples that do not fit well are to be labeled and the model is improved after the labeling by oracle, then the cycle repeats. He and Carbonell [32] propose a nearest-neighbor-based active learning method named NNDM, which makes use of nearest neighbors to measure the local density for every sample and queries the label of samples that have the maximum changes in local density. A similar idea is also introduced in [33]. While NNDM requires the number of rare classes and the prior probability of each class as the inputs, SEDER (Semi-parametric Density Estimation based Rare category detection) [34] requires no prior information of the dataset. As in real-world applications it is often difficult to know exactly the number of classes in the dataset and the prior probability of each class, especially in extremely rare cases, SEDER is more suitable for real-world applications.

4 CONCLUSION

Classification based rare class mining methods can produce models that are easy to understand but are only applicable to well defined problems in which rare and common patterns are known in advance. However, the ability of detecting novel patterns is crucial. Take the network intrusion detection task as an example. It is impossible to pre-define the behaviors of all kinds of intrusions but whenever a possible intrusion happens, the system is still expected to raise the alarm adaptively. In such cases, techniques in related domains such as novel information detection [35] and some recently proposed ideas such as learning to rank algorithms [36] are likely to be helpful.

In the meantime, very limited work has been done on feature selection techniques in the rare class mining domain. Lee and Stolfo [37] discuss the feature extraction metrics for intrusion detection, and a few feature selection metrics are evaluated in the text categorization domain [38, 39]. Tang and Liu [4] investigate how different feature selection metrics affect different classifiers in text classification. However, none of the above work focuses directly on the issue of how traditional feature selection methods can be better adapted to fit the rare class problem, which is an important research direction worth thorough investigation.

REFERENCES

[1] G. Medioni, I. Cohen, F. Brémond, S. Hongeng, and R. Nevatia. Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8): 873-889, 2001.

[2] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. *Proc. CVPR'04*, Washington, DC, 2004, Vol.2, pp.819-826.

[3] M. Kubat, R. C. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine Learning*, 30(2): 195-215, 1998.

[4] L. Tang and H. Liu. Bias analysis in text classification for highly skewed data. *Proc. ICDM'05*, Houston, Texas, USA, 2005, pp.781-784.

[5] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. *Proc. SIGIR'94*, Dublin, Ireland, 1994, pp. 3-12.

[6] G. M. Weiss. Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, 6(1): 7-19, 2004.

[7] M. V. Joshi. On evaluating performance of classifiers for rare classes. *Proc. ICDM'02*, Maebashi, Japan, 2002, pp. 641-644.

[8] M. Joshi, V. Kumar, and R. Agarwal. Evaluating boosting algorithms to classify rare classes: comparison and improvements. *Proc. ICDM'01*, San Jose, 2001, pp.257-264.

[9] M. V. Joshi, R. C. Agarwal, and V. Kumar. Mining needle in a haystack: classifying rare classes via two-phase rule induction. *Proc. ACM SIGMOD'01*, Santa Barbara, California, 2001, pp. 91-102.

[10] M. V. Joshi, R. C. Agarwal, and V. Kumar. Predicting rare classes: Can boosting make any weak learner strong?. *Pro. SIGKDD'02*, Edmonton, Alberta, Canada, 2002, pp. 297-306.

[11] G. M. Weiss and H. Hirsh. Learning to predict extremely rare events. In: N. Japkowicz (Ed.), *Learning from Imbalanced Data Sets: Papers from the AAAI workshop*, vol. WS-00-05, AAAI Press, 2000, pp. 64-68.

[12] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Mining data with rare events: A case study. *Proc. ICTAI'07*, Patras, Greece, 2007, Vol.2, pp.132-139.

[13] P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.

[14] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One-sided selection. *Proc. ICML'97*, Nashville, Tennessee, 1997, pp. 179-186.

[15] X. Y. Liu, J. Wu, and Z. H. Zhou. Exploratory under-sampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 39(2): 539-550, 2009.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1): 321-357, 2002.

[17] T. M. Khoshgoftaar, C. Seiffert, J. V. Hulse, A. Napolitano, and A. Folleco. Learning with Limited Minority Class Data. *Proc. ICMLA'07*, Cincinnati, Ohio, 2007, pp.348-353.

[18] J. H. Zhao, X. Li, and Z. Y. Dong. Online rare events detection. In: Zhi-Hua Zhou, Hang Li, and Qiang Yang (Eds.), *Advances in Knowledge Discovery and Data Mining*, Springer, 2007, pp.1114-1121.

[19] Y. Sun, M. S. Kamel, and Y. Wang. Boosting for learning multiple classes with imbalanced class distribution. *Proc. ICDM'06*, Hong Kong, China, 2006, pp. 592-602.

[20] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Science*, 55(1): 119-139, 1997.

[21] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan. AdaCost: Misclassification cost-sensitive boosting. *Proc. ICML 1999*, Bled, Slovenia, 1999, pp. 97-105.

[22] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer. SMOTEBoost: Improving prediction of the minority class in boosting. *Proc. PKDD 2003*, Cavtat Dubrovnik, Croatia, 2003, pp. 107-119.

[23] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. *Proc. ACM KDD'99*, San Diego, CA, 1999, pp. 43-52.

[24] H. Alhammady and K. Ramamohanarao. The application of emerging patterns for improving the quality of rare-class classification. *Proc. PAKDD 2004*, Sydney, Australia, 2004, Vol. 3056, LNCS, pp. 207-211.

[25] H. Alhammady and K. Ramamohanarao. Using emerging patterns and decision trees in rare-class classification. *Proc. ICDM'04*, Brighton, UK, 2004, pp. 315-318.

[26] H. Alhammady. A novel approach for mining emerging patterns in rare-class datasets. In: T. Sobh (Ed.), *Innovations and Advanced Techniques in Computer and Information Sciences and Engineering*, Springer, 2007, pp. 207-211.

[27] N. Japkowicz. Supervised learning with unsupervised output separation. *Proc. ASC 2002*, Banff, Canada, 2002, pp. 321-325.

[28] J. Wu, H. Xiong, P. Wu, and J. Chen. Local decomposition for rare class analysis. *Proc. KDD 2007*, San Jose, California, 2007, pp. 814-823.

[29] B. Settles. Active Learning Literature Survey. *Computer Sciences Technical Report 1648*, University of Wisconsin-Madison, 2009.

[30] S. Dasgupta. Coarse sample complexity bounds for active learning. In: Y. Weiss, B. Schölkopf, and J. Platt (Eds.), *Advances in Neural Information Processing Systems 18*, MIT Press, 2006, pp.235-242.

[31] D. Pelleg and A. Moore. Active learning for anomaly and rare-category detection. In: L. K. Saul, Y. Weiss, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, MIT Press, 2005, pp.1073-1080.

[32] J. He and J. Carbonell. Nearest-neighbor-based active learning for rare category detection. In: J. C. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), *Advances in Neural Information Processing Systems 20*, MIT Press, 2008, pp. 633–640.

[33] J. He and J. Carbonell. Rare class discovery based on active learning. *Proc. ISAIM 2008*, Fort Lauderdale, Florida, 2008.

[34] J. He and J. Carbonell. Prior-free rare category detection. *Proc. SDM 2009*, Sparks, Nevada, 2009, pp.155-163.

[35] M. Markou and S. Singh. Novelty detection: A review—parts 1 and 2. *Signal Processing*, 83(12): 2481-2521, 2003.

[36] O. Dekel, C. Manning, and Y. Singer. Log-linear models for label ranking. In: S. Thrun, L. K. Saul, and B. Schölkopf (Eds.), *Advances in Neural Information Processing Systems 16*, MIT Press, 2004, pp.497-504.

[37] W. Lee and S. J. Stolfo. A framework for constructing features and models for intrusion detection systems. *ACM Transactions on Information and System Security*, 3(4): 227-261, 2000.

[38] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. *Proc. ICML*, Nashville, TN, 1997, pp. 412-420.

[39] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3: 1289-1305, 2003.