A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis

Julien Ah-Pine, Edmundo-Pavel Soriano-Morales

A Study of Synthetic Oversampling for Twitter Imbalanced Sentiment Analysis

Julien Ah-Pine and Edmundo Pavel Soriano Morales

University of Lyon, ERIC Lab
5 avenue Pierre Mendès France
69676 Bron Cedex, France
{julien.ah-pine,edmundo.soriano-morales}@univ-lyon2.fr

Abstract. The majority of Twitter sentiment analysis systems implicitly assume that the class distribution is balanced while in practice it is usually skewed. We argue that Twitter opinion mining using learning methods should be addressed in the framework of imbalanced learning. In this work, we present a study of synthetic oversampling techniques for tweet-polarity classification. The experiments we conducted on three publicly available datasets show that these methods can improve the recognition of the minority class as well as the geometric mean criterion.

Key words: Synthetic sampling, Sentiment analysis, Social media.

1 Introduction

Micro-blogging services are communication tools that are massively used by people to instantaneously share their opinions about any kinds of topics. These opinions are of interest for companies or individuals, like politicians, as they allow them to monitor their online reputation. Twitter has been the most popular micro-blogging service with more than 500 million tweets per day in 2013[1]. Thus, sentiment analysis of tweets[2] has received a lot of attention both from academia and industry during the last years.

In this paper, we focus on tweets polarity classification using supervised learning methods. This task is challenging in several respects. Firstly, tweets are limited to 140 characters and they contain irregular lexical units and syntactic patterns. Hence, these data are noisy, sparse and high-dimensional which makes the learning process difficult. Moreover, tweets expressing an opinion about a given topic usually present a skewed polarity distribution. In this case, any classifier would be biased towards the majority class.

In order to cope with these challenges, we propose to use synthetic oversampling techniques. These procedures are designed to deal with the class imbalance issue. We show that not only they enable reducing the bias towards the majority

[1] http://www.internetlivestats.com/twitter-statistics/
[2] Short informal messages in a more general perspective.

class, but they also alleviate the data sparsity burden commonly encountered in text mining.

The rest of the paper is organized as follows. In section 2, we discuss some related works in order to position and motivate our proposal. In section 3, we present our approach based on three synthetic oversampling methods and two supervised learning methods. Then, in section 4, we detail the experiments we conducted on three datasets including two different languages and we discuss the obtained results as well. We conclude the paper in section 5.

2 Related Works

2.1 Twitter Sentiment Analysis

Twitter sentiment analysis has received a growing interest starting from 2009 [5, 19]. In this work, we focus on polarity detection which aims at predicting the opinion of a tweet as positive or negative. Supervised learning techniques are the mainstream approaches in this case. Due to the characteristics of Twitter data, systems usually used for sentiment analysis (see [14] for a survey of this field) do not perform well. In order to improve classifiers' performance for tweets opinion mining, most of research works have proposed to extract features/lexicons which are specific to this type of data and/or leverage external resources [5, 19, 11, 22, 10, 20, 15]. In contrast, we apply a corpus-based approach with no particular feature engineering.

2.2 Imbalanced Sentiment Analysis

The class imbalance problem in binary classification occurs when the sizes of the classes differ greatly. In this case, any classifier is biased toward the majority class (see [9] for a survey of the domain). For example, in the datasets we examined, near 70% of the tweets of the datasets we experimented with are negative. If a naïve classifier always assigns the negative polarity to any tweet, it will give an overall accuracy of 70% but without recovering any positive tweet, which is not satisfying.

Imbalanced learning for sentiment analysis has been studied by several researchers in different learning settings [12, 13, 17, 25]. However, we found very few papers that directly address imbalanced sentiment analysis for Twitter data [16, 6]. The methods that are proposed in the two latter works are similar to cost-sensitive approaches. In our case, we rather use sampling techniques.

3 The Proposed Approach

3.1 Vector Space Representation and Neighborhood

Tweets contain slang words and irregular expressions. Thus, linguistic analyses by conventional NLP tools often give poor performances on such texts. To

circumvent these difficulties, and also to deal with different languages, we rely on a vectorial representation of tweets based on a bag-of-words approach. We denote by $\mathcal{F}$ the resulting feature space, $\mathbf{x} \in \mathcal{F}$ is a vector representing a tweet and its coordinates are its words' frequency. In what follows, we use $\mathbb{P}$ and $\mathbb{N}$ to designate the subsets of tweets with the minority and the majority class labels respectively ($|\mathbb{P}| < |\mathbb{N}|$).

In order to compare tweets, we use the cosine similarity function. Note that all pairwise proximity measures lie between 0 and 1 since the coordinates of vectors are non-negative. Let $\mathbf{x}$ be any tweet in $\mathbb{P}$ then its neighborhood is denoted $\mathbb{NN}(\mathbf{x})$ and it consists of the k nearest neighbors.

3.2 Synthetic Oversampling

To face the skewed class distribution problem, one straightforward approach is to balance the training set so that $|\mathbb{P}| = |\mathbb{N}|$. Undersampling the majority class or oversampling the minority class are two possible strategies. Since the data are very sparse, undersampling the majority class is sub-optimal as we may lose meaningful examples in the learning process. Therefore, oversampling the minority class seems a better solution. In this case, synthetic oversampling creates new examples in $\mathbb{P}$ by taking convex combinations of existing points.

We recall three popular synthetic oversampling methods: SMOTE [2], Borderline-SMOTE [7] and ADASYN [8]. Their general procedure can be cast as follows:

1. Select an original tweet $\mathbf{x}$ according to a probability distribution over $\mathbb{P}$.
2. Determine $\mathbb{NN}(\mathbf{x})$.
3. Select a neighbor $\mathbf{x}'$ according to a probability distribution over $\mathbb{NN}(\mathbf{x})$.
4. Create a synthetic example $\mathbf{y}$ as follows:

$$\mathbf{y} = \mathbf{x} + \alpha(\mathbf{x}' - \mathbf{x}) \tag{1}$$

   where α is a random value in $[0, 1]$.
5. Repeat 1-4 until the desired number of new examples is reached.
6. Append the set of synthetic points to $\mathbb{P}$.

Note that $\mathbf{y}$ lies in the line segment joining $\mathbf{x}$ and $\mathbf{x}'$. It is important to notice that $\mathbf{y}$ belongs to the subspace spanned by the union of the underlying subspaces of $\mathbf{x}$ and $\mathbf{x}'$. Therefore, synthetic examples are less sparse than original ones.

The main differences between the three oversampling methods concern the random selection of $\mathbf{x} \in \mathbb{P}$ in step 1. SMOTE assumes a uniform distribution over $\mathbb{P}$ whereas Borderline-SMOTE assumes a uniform distribution over $\mathbb{B}$, a subset of $\mathbb{P}$. $\mathbb{B}$ consists of tweets in $\mathbb{P}$ whose neighborhoods contain a majority of points in $\mathbb{N}$. These items lie in subspaces where the decision boundary is prone to errors. Thereby, it is expected that oversampling in these parts of the space improves the classifier performances. Regarding ADASYN, it assumes a non uniform distribution over $\mathbb{P}$. It can be seen as a smoothed version of Borderline-SMOTE: the noisier the neighborhood of $\mathbf{x}$, the more synthetic points around $\mathbf{x}$. In other words, the probability to select $\mathbf{x}$ in step 1 is proportional to the number of points of $\mathbb{N}$ contained in $\mathbb{NN}(\mathbf{x})$.

4 Experiments

4.1 Datasets and Data Representation

We assess the approach introduced previously on three publicly available Twitter datasets. The first two are OMD "Obama-McCain Debate" [21] and HCR "Health Care Reform" [23]. The third one is IW "Imagiweb" and concerns tweets in French, posted during the 2012 french presidential election [24]. We chose political tweets because they present a particularly skewed class label distribution.

Concerning the vectorial representation of tweets, we used unigrams of words and we only removed the hapax.

We give the descriptive statistics[3] of these datasets below:

- OMD: 1906 tweets (710 positive, 1196 negative) and 1569 features;
- HCR: 1922 tweets (541 positive, 1381 negative) and 2066 features;
- IW: 4519 tweets (1092 positive, 3427 negative) and 3918 features.

4.2 Supervised Learning Methods

We experimented with two different learning models: decision trees and the $l1$ penalized logistic regression.

Decision trees are well-known symbolic learning techniques and offer the advantages of coping with high-dimensional data as well as providing human-readable outputs. In this work, we used CART [1], which builds a binary classification tree based on the Gini index splitting criterion. The R package `rpart` was used and the default parameters values specified in `rpart.control` were applied.

The $l1$ penalized logistic regression [18] is also an appropriate supervised learning for high-dimensional data since it implicitly performs feature selection. Moreover, this method has proven to provide competitive results in text classification [4]. We used the `glmnet` R package [3] and in particular the function `cv.glmnet` which allows us to select the mixing parameter λ based on the error observed during training phase.

4.3 Assessment Measures

We use several performance criteria: overall accuracy (OA), F1-measures of the positive and negative classes (F-$\mathbb{P}$ and F-$\mathbb{N}$ respectively). OA evaluates the overall performance of a classifier but it does not properly account for the performances on $\mathbb{P}$ as compared to $\mathbb{N}$ because of the skewed distribution of class labels. Hence, we also use a popular criterion for imbalanced learning: the geometric mean (GM) of both class accuracy rates. Unlike OA, GM is independent of the class distribution (see [9, Chapter 8] for an overview of this topic). Thus we argue that GM should also be a default evaluation criterion in Twitter sentiment analysis tasks.

[3] We removed tweets that were labeled as neutral since we are only concerned with polarity detection.

4.4 Experiments Setting and Results

It is important to note that we are not interested in comparing the results of decision trees against $l1$ penalized logistic regression. Our purpose is rather to illustrate that synthetic oversampling can improve the performances of learning methods on Twitter imbalanced-polarity detection tasks.

We tested the two learning models on the three collections with different relatively balanced training sets. In what follows, τ is a variable taking its values in $\{0, 1/4, 1/2, 3/4, 1\}$ which measures how much the training set is balanced with respect to the initial distribution. In fact, $\tau = 0$ is when no oversampling was carried out and we used the initial imbalanced training set (this is our baseline); $\tau = 1/4$ means we generated $\lfloor(|\mathbb{N}| - |\mathbb{P}|)/4\rfloor$ positive synthetic examples; ...; and $\tau = 1$ means we exactly sampled $|\mathbb{N}| - |\mathbb{P}|$ new positive items in order to have a perfectly balanced training set. The neighborhood was set to $k = 20$ nearest neighbors[4]. The results we obtained using a 5 fold cross-validation are plotted in Figure 1 for decision trees and in Figure 2 for $l1$ penalized logistic regression.

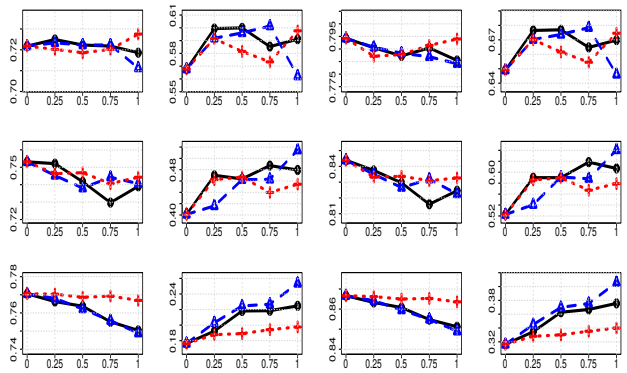

Fig. 1: Results for decision trees (CART). Solid line with circles refers to SMOTE, dashed line with triangles refers to Borderline-SMOTE and dotted line with plus signs refers to ADASYN. From left to right: plots of OA, F-$\mathbb{P}$, F-$\mathbb{N}$ and GM measures. From top to bottom: plots for OMD, HCR and IW benchmarks. The x-axis refers to τ going from initial imbalanced ($\tau = 0$) to fully balanced ($\tau = 1$) training sets.

Our main findings are the following:

– For both decision tree and $l1$ penalized logistic regression, we note quite the same trends: oversampling generally improves the results. Indeed, All

[4] We also tested with $k = 10, 30$ but the trends were similar and the results comparable.

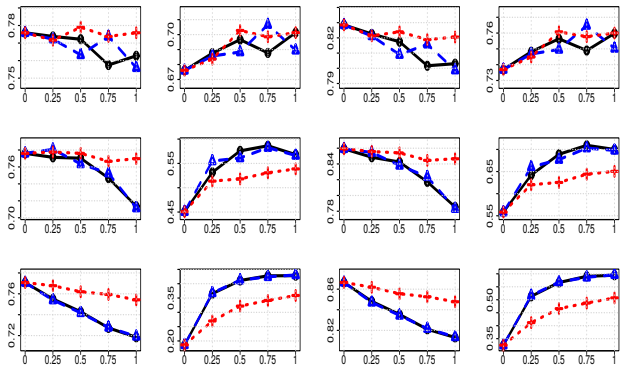

Fig. 2: Results for $l1$ penalized logistic regression. Same legend as in Figure 1.

three sampling methods globally improve the GM measure[5]. Thereby, our approach allows alleviating the class imbalance problem effectively. For OMD, when $\tau = 1$, the most important gains for GM measures are given by ADASYN (1st row, 4th column in the figures). Regarding HCR and IW, Borderline-SMOTE performs the best but SMOTE often provides comparable results (2nd and 3rd rows respectively and 4th column in the figures).

– All three oversampling strategies generally boosts F-$\mathbb{P}$ values[6]. The minority class is thus better recognized. However, this is at the expense of a reduction of F-$\mathbb{N}$ values. Nonetheless, since the increasing rate of F-$\mathbb{P}$ is generally much larger than the decreasing rate of F-$\mathbb{N}$, we note the overall increase of GM values as highlighted previously.
– For all three sampling techniques, the OA measure tends to diminish as the training set is more and more balanced. In fact, since the class distribution in the test set is skewed towards $\mathbb{N}$, the errors on true negative tweets have more impact on OA than the correct detection of true positive tweets. This illustrates again the fact that OA is not a criterion that properly accounts for imbalanced data.
– We cannot conclude on which of the three oversampling strategies is the best. However, we can make the following remarks:
  • SMOTE and Borderline-SMOTE have quite the same behaviours for the HCR and IW collections. F-$\mathbb{P}$ measures are greater than ADASYN whereas F-$\mathbb{N}$ values are lower. Both methods allows a much better recognition of the minority class but in doing so they make more mistakes when detecting the majority class.
  • In contrast, ADASYN presents peculiar properties. The increase of GM values are lower than for the two other methods but this oversampling

[5] The only exception is observed for OMD when using a fully balanced training sets ($\tau = 1$) generated by Borderline-SMOTE with CART as shown in Figure 1.
[6] Except the same particular case mentioned previously.

technique shows more stable OA values and even better ones in some cases. For the OMD dataset specifically, this approach not only provides among the best performances for the GM criterion but it also allows improving the OA measures unlike the other methods.

5 Conclusion

Twitter sentiment analysis is confronted with the class imbalance problem and it is important to take this aspect into account when designing opinion mining systems based on machine learning.

A way to address this challenge is to use synthetic oversampling which aims at balancing the training set in a meaningful way. Three state-of-the-art methods have been examined in that regard. We conducted experiments on political-tweets polarity classification using three datasets and in two different languages. The obtained results show that our proposal makes it possible to deal with the skewed class distribution issue by providing better recognition of the minority class as well as obtaining large increases of the overall geometric mean criterion.

In future work, we intend to extend our study to multiclass sentiment analysis and also to examine the use of synthetic oversampling methods in other NLP tasks as a general approach to cope with the sparsity problem.

Acknowledgment This work was partly supported by the french national projects Imagiweb ANR-2012-CORD-002-01 and Request PIA/FSN.

References

1. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC press (1984)
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. J. Artif. Int. Res. 16(1) (2002)
3. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33(1) (2010)
4. Genkin, A., Lewis, D.D., Madigan, D.: Large-scale bayesian logistic regression for text categorization. Technometrics 49 (2007)
5. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification using Distant Supervision. Tech. rep., Stanford University (2009), `https://sites.google.com/site/twittersentimenthelp/home`
6. Hamdan, H., Bellot, P., Bechet, F.: Lsislif: Feature extraction and label weighting for sentiment analysis in twitter. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (2015)
7. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: A new over-sampling method in imbalanced data sets learning. In: Advances in Intelligent Computing, Lecture Notes in Computer Science, vol. 3644 (2005)
8. He, H., Bai, Y., Garcia, E., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence) (2008)

9. He, H., Ma, Y.: Imbalanced Learning: Foundations, Algorithms, and Applications. Wiley-IEEE Press (2013)
10. Kiritchenko, S., Zhu, X., Mohammad, S.M.: Sentiment analysis of short informal texts. Journal of Artificial Intelligence Research (JAIR) 50 (2014)
11. Kouloumpis, E., Wilson, T., Moore, J.D.: Twitter sentiment analysis: The good the bad and the omg! In: Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011 (2011)
12. Li, S., Wang, Z., Zhou, G., Lee, S.Y.M.: Semi-supervised learning for imbalanced sentiment classification. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three. IJCAI'11 (2011)
13. Li, S., Zhou, G., Wang, Z., Lee, S.Y.M., Wang, R.: Imbalanced sentiment classification. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. CIKM '11 (2011)
14. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Aggarwal, C.C., Zhai, C. (eds.) Mining Text Data. Springer US (2012)
15. Martínez-Cámara, E., Martín-Valdivia, M.T., Urena-López, L.A., Montejo-Ráez, A.R.: Sentiment analysis in twitter. Natural Language Engineering 20(01) (2014)
16. Miura, Y., Sakaki, S., Hattori, K., Ohkuma, T.: Teamx: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (2014)
17. Mountassir, A., Benbrahim, H., Berrada, I.: An empirical study to address the problem of unbalanced data sets in sentiment classification. In: Systems, Man, and Cybernetics (SMC), IEEE International Conference on (2012)
18. Ng, A.Y.: Feature selection, l1 vs. l2 regularization, and rotational invariance. In: Proceedings of the Twenty-first International Conference on Machine Learning. ICML '04 (2004)
19. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: LREC. vol. 10 (2010)
20. Saif, H., He, Y., Fernandez, M., Alani, H.: Semantic patterns for sentiment analysis of twitter. In: Proceedings of the 13th International Semantic Web Conference - Part II. ISWC '14 (2014)
21. Shamma, D.A., Kennedy, L., Churchill, E.F.: Tweet the debates: Understanding community annotation of uncollected sources. In: Proceedings of the First SIGMM Workshop on Social Media. WSM '09 (2009)
22. Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., Gordon, J.: Empirical study of machine learning based approach for opinion mining in tweets. In: Advances in Artificial Intelligence (2012)
23. Speriosu, M., Sudan, N., Upadhyay, S., Baldridge, J.: Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the First Workshop on Unsupervised Learning in NLP. EMNLP'11 (2011)
24. Velcin, J., Kim, Y., Brun, C., Dormagen, J., SanJuan, E., Khouas, L., Peradotto, A., Bonnevay, S., Roux, C., Boyadjian, J., Molina, A., Neihouser, M.: Investigating the image of entities in social media: Dataset design and first results. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014) (2014)
25. Xu, R., Chen, T., Xia, Y., Lu, Q., Liu, B., Wang, X.: Word embedding composition for data imbalances in sentiment and emotion classification. Cognitive Computation 7(2) (2015)