Using a Genetic Algorithm Approach to Study the Impact of Imbalanced Corpora in Sentiment Analysis

Lohann C. Ferreira, Mariza Miola Dosciatti, Julio Cesar Nievola, Emerson Cabrera Paraiso

Pontifícia Universidade Católica do Paraná, Rua Imaculada Conceicao, 1155 - Curitiba, PR, Brazil
{lohann.ferreira, mariza.dosciatti, nievola, paraiso}@ppgia.pucpr.br

Abstract

The SVM classifier has been used in many methods to identify emotions in text due to their good generalization capability and robustness with high dimensionality data. However, most textual corpora usually subject to such methods are naturally imbalanced. As a consequence, the SVM, sensitive to imbalance data, assigns to most texts the majority class. In this article, we present a Genetic Algorithm based approach that aims to reduce the imbalance of the data in the context of emotions identification. This approach allowed us to study the impact of its application in a method of emotion identification in texts written in the Brazilian Portuguese. Experimentations showed us that balancing the corpus could be an alternative when using the SVM classifier for emotions identification, especially in a multiclass configuration.

Introduction

Although existing knowledge discovery and data engineering techniques have shown great success in many real-world applications, the problem of learning from imbalanced data still a challenge (He and Garcia 2009). A dataset is imbalanced if the classification categories are not approximately equally represented. Applications such as detecting fraud in banking operations or detecting network intrusions are examples of domains where imbalanced datasets occur. The automatic identification of emotions in texts, one of the tasks related to Sentiment Analysis, is a domain characterized by imbalanced datasets as well.

Sentiment Analysis is the field of research dedicated to study emotions and opinion mining. It is a challenging natural language processing or text mining problem. Due to its value for practical applications, there has been a growth of both research in academia and applications in the industry (Liu 2012). The research in the field started with sentiment and subjectivity classification, which treated the problem as a text classification problem (Liu 2010). There has been progress in research on polarity (positive and negative texts) and sentiment analysis, but less work has been done in automatic identification of emotions in text. Currently, researches and studies concerning emotions are divided in several and different areas, but the one which will be discussed and used in this paper is the simplest of them, named Basic (or Pure) Emotions. This concept is related to the innate emotions shared among all cultures worldwide. It was proposed in the 1970's by Paul Ekman and Wallace Friesen (Ekman and Firesen 1978) and defines six basic emotions: sadness, anger, happiness/joy, fear, disgust and surprise.

Most textual corpora used in Sentiment Analysis, such as newspaper articles or blog posts, are naturally imbalanced. In (Strapparava and Mihalcea 2008), Strappavara and Mihalcea used a corpus compounded by 8,761 posts, where 55% of them were labelled as containing the emotion *happiness* and 0.8% of them were labelled as *disgust*. Ghazi and colleagues used in their work (Ghazi, Inkpen, and Szpakowicz 2010) a weblog corpus containing 4,090 annotated sentences: 68% of them annotated as non-emotional ones and 3% of them annotated as belonging to the *fear* and *surprise* emotion.

The SVM (Support Vector Machine) classifier has been used in many methods to identify emotions in text due to its good generalization capability and robustness with high dimensionality data (Rangel and Rosso 2013). Most of these methods were tested using balanced data. However, as previously mentioned, most textual corpora usually subject to emotions identification are naturally imbalanced. As a consequence, the SVM classifier, sensible to imbalance data, assigns to most texts the majority class.

In this paper we present a Genetic Algorithm (GA) approach to balance the corpus of texts in order to investigate the impact of this action in the classification level when using a SVM classifier. We also present a SVM

based emotion identification method used to run some experiments.

The remainder of this paper is organized as follows. First, we briefly overviews imbalanced data in machine learning. The next section presents an SVM based method for emotion identification in texts. Then, we present the Genetic Algorithm approach to balance the corpus. Some experimental results and analysis are also given. Finally, we give some conclusions and discuss future work.

Imbalanced Data in Machine Learning

The issue with imbalance in the class distribution became more pronounced with the applications of the machine learning algorithms to the real world (Chawla 2005). Some researchers, as Weiss and Provost (Weiss and Provost 2003), studied the effect of class distribution on classifier learning. They state that the natural distribution is not the best one for learning a classifier. Many other researchers are working on this problem ((Batista, Prati, and Monard 2004), (Japkowicz and Stephen 2002) and (Khoshgoftaar, Hulse, and Napolitano 2010)). From these studies, two main approaches arise: the ones that deal the problem in the data level and the ones that deal the problem in the classifier (algorithm) level.

The first approach attempts to balance the distribution of the classes in the dataset, by under-sampling the majority instances or over-sampling the minority instances. The under-sampling method extracts a smaller set of majority instances while preserving all the minority instances. This method is suitable for large-scale applications where the number of majority samples is tremendous and lessening the training instances reduces the training time and makes the learning problem more tractable (Zhou and Liu 2006). However, one problem associated with under-sampling techniques is that we may lose informative instances from the discarded instances (Nguyen, Bouzerdoum, and Phung 2008). Over-sampling method increases the number of minority instances by oversampling them. The advantage is that no information is lost from the training samples because all instances are employed (Alejo et al. 2007). However, the minority instances are over-represented in the training set and moreover, adding training instance means increasing training time.

In the second approach, specific learning algorithms are modified or adjusted to learning from imbalanced data, such as cost-sensitive learning, one-class learning, and ensemble learning. He and Garcia published a comprehensive survey on learning from imbalanced data where they present a list of different algorithms (He and Garcia 2009).

A semi-supervised approach based on under-sampling and random subspace for sentiment classification was presented by Li and colleagues (Li et al. 2011). The approach first use under-sampling to generate multiple sets of balanced initial training data. Then, a semi-supervised learning method based on random subspace dynamically generates various subspaces in the iteration process to guarantee enough variation among the involved classifiers. Their approach is limited to positive and negative classes that are not enough in our multiclass (six emotions) domain.

In order to test our GA approach, we need to present a SVM based emotion identification method. The method is presented in the next section.

An SVM Emotion Identification Based Method in a Multiclass Configuration

The task of identifying emotions in texts is a multiclass problem: an instance may belong to one of the y_i classes, where $i > 2$. The SVM is originally binary. Strategies for solving multiclass problems using SVM have been developed and allow converting a multiclass problem in several binary sub-problems (Hsu and Lin 2002). Several researches (most of them actually) reduce the problem of emotion identification to a binary problem: polarity approach ((Liu 2010) and (Li et al. 2011)). The positive polarity represents the emotion *joy* and the negative polarity groups the emotions *sadness*, *anger*, *disgust* and *fear*. This approach simplifies the problem.

In our research, we are interested in identifying the basic emotions (sadness, anger, happiness/joy, fear, disgust and surprise) present in a text. Despite the imbalance on data, some emotions are very similar (or close) to each other, increasing the opportunity for confusion. This can be seen on Figure 1. It shows a wheel proposed by Robert Plutick (Plutchik 2001), where he presents his emotion model, containing eight emotions. The wheel is used to illustrate the different emotions compelling and nuanced. The idea is to show bipolar relations among opposite emotions, such as *joy* and *sadness* or *anger* and *fear*. It is possible to conclude that emotions vary in their degree of similarity to one another.

In our method, each text in the corpus needs to be preprocessed. After traditional natural language preprocessing tasks, such as removal of stopwords, we use the Bag-of-Words (Radovanovic and Ivanovic 2008) model to generate a list of words without repetition. Then, two filters designed to extract the best features of the data set and to reduce dimensionality are applied. The first filter removes all terms that have a few occurrences (under a threshold). The threshold varies according to the size of the corpus and considers that rare terms are irrelevant for classification of the text. The second filter implements the

Information Gain (Mitchell 1997), which is used to select the most representative terms of the feature set.

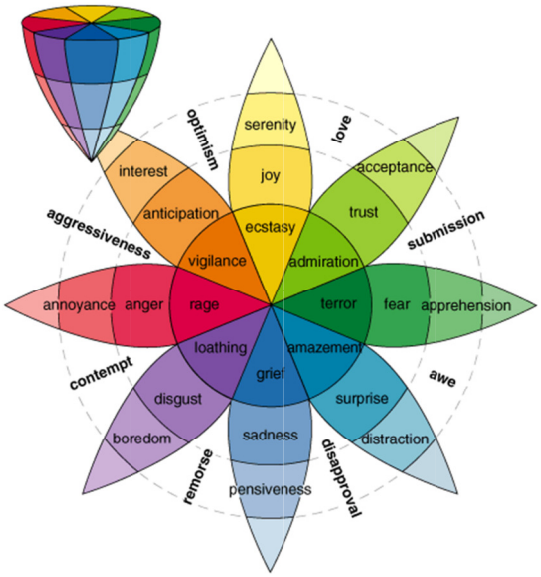

Figure 1. Plutick´s Emotion Model (Plutick 2001).

The Information Gain of a term is defined using equations 1 and 2. Equation 1 measures the Entropy: the mixing degree of each term in relation to classes, where the training set S can have c different classes and p_i is the proportion of data in S that belongs to class i.

$$Entropy(S) \equiv -\sum_{i=1}^{c} (p_i \log_2 p_i) \quad (1)$$

Equation 2 defines the Information Gain (IG) of a term t, where v are all possible values for each attribute (present in S in our case).

$$IG(S,t) = Entropy\ (S) - \sum \frac{|S_v|}{|S|} Entropy\ (S_v) \quad (2)$$

The vector that represents each text is obtained using TF-IDF (Term Frequency - Inverse Document Frequency) (Salton and Buckley 1988). The TF-IDF model was chosen to generate an efficient representation of textual data in a vector form. It is intended to reflect how important a term is to a document in a collection or corpus. This model assigns a weight $w_{t,d}$ for a term t in a document d according to equations 3 and 4. The value $tf_{t,d}$, is the number of times that the term t appears in the document d, N is the total number of documents that make up the data set and df_t is the number of documents containing the term t.

$$w_{t,d} = tf_{t,d} * idf_t \quad (3)$$

$$idf_t = \log \frac{N}{df_t} \quad (4)$$

The corpus is then ready for the learning phase. We used a supervised approach based on SVM. The SVM was configured using the kernel RBF (Radial Basis Function), setting up *gamma* to zero and *c* to one. We also used 10-fold cross validation.

It is very important to highlight that this method does not use any lexical resource (set of emotional words), such as WordNet Affect (Strapparava and Valitutti 2004). Several researches were done using lexical resources in order to improve emotions identification (Strapparava and Mihalcea 2008). In general, such approachs can improve the ability an algorithm has to identify emotions in texts. However, this kind of resource (lexical resource, domain-specific words, etc.) is not present to every language. In addition, lexical resources such as WordNet Affect may contain words that have different meaning in different cultures or contexts. This is the case for texts written in the Portuguese language, our main interest.

It is also important to note that this identification method is simple but complete enough to test the approach for balancing the corpus. More sophisticated methods already exist published in the literature (Chaffar and Inkpen 2011) and (Strapparava and Mihalcea 2008)).

The next section present the genetic approach used to study the impact of imbalanced data in Sentiment Analysis.

A Genetic Algorithm Approach to Balance a Corpus

The next paragraphs describe an algorithm to balance a corpus. The algorithm aims to increase the F-measure (F1) value for each class. The F1 can be seen as a weighted average of the precision and recall, where an F1 reaches its best value at 1 and worst score at 0. It combines precision and recall, been the harmonic mean of precision and recall (Van Rijsbergen 1979).

The algorithm works iteratively. Each iteration performs over-sampling and/or under-sampling over the training dataset, trains and tests the SVM classifier. The classifier is evaluated using the fitness function presented in equation 5, that calculates the geometric mean of F1 (Ai) for each class i.

$$Fitness = \sqrt[i]{A_1 * ... * A_i} \quad (5)$$

After preprocessing the corpus, we need a population of individuals. An individual is formed by n genes, where n is the number of instances present in the corpus. The

population is randomly generated with *k* individuals, where *k* is empirically set. An iterative process begins, using under-sampling and over-sampling to generate the population of candidates. The following three steps are thus repeated *k* times:

1. The individual *i* has *n* genes, where each gene is initially set to be 1.

2. For each gene of *i*, the under-sampling will set it to 0 if a randomly generated value *v* is smaller than 0.1 (i.e. the under-sampling is limited to 10% of genes), where *v* varies between 0 and 1.

3. The over-sampling step means randomly choose a gene to be incremented. It will be repeated *n/5* times, which means that approximately 20% of genes will be changed. A gene can be selected more than once.

Figure 2 shows how the initial individual evolves from step 1 to step 3, considering 20 genes.

Step 1: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
Step 2: [0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0]
Step 3: [0, 1, 1, 0, 1, 1, 1, 1, 2, 3, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1]

Figure 2. An Example of Individual After Under-Sampling and Over-Sampling.

There are some constants that must be defined (the over-sampling and under-sampling rate, for instance). They are related to the imbalance rate of the corpus. Our practical experience using news articles and social media texts suggest using 10% for under-sampling and 20% for over-sampling.

Finished the population generation, it has *k* individuals to be evolved.

A new population is generated based on the chromosome of an individual in a Cross Validation (10-fold) process. We use elitism to select the best individuals to take part of the next generation. Parents are selected using roulette wheel. Genetic operators (arithmetic crossover and a variation of mutation) are then applied. After randomly choosing two individuals, equation 6 calculates the individual of the next generation, where λ ($0 \leq \lambda \leq 1$) and *l* is the position of a gene. The process is repeated for each gene.

$$c_l^{\text{Child}} = \lambda c_l^1 + (l - \lambda) c_l^2 \quad (6)$$

For each gene *y*, it is randomly generated a value *p* between 0 and 1. If *p* is smaller than 0.05, then the equation 7 generates the new gene value.

$$y = \{ y \in N \mid y \in [Lim^{Inf}, Lim^{Sup}] \} \quad (7)$$

where:

$$Lim^{Inf} = \begin{cases} 0, & if \quad x - \sigma \leq 0 \\ x - \sigma, & if \quad x - \sigma > 0 \end{cases}$$

$$Lim^{Sup} = x \pm \sigma$$

In order to execute the mutation operator, σ ($\sigma > 0$) is defined. Empirically, σ was set to 3.

This process is done only in the training folds. The algorithm terminates when there is no change in results after 100 generations.

Experiments and Analysis

In this section, we apply the proposed approach to two different experiments. The first experiment evaluates how the identification method performs in a multiclass configuration (six classes or six emotions). The second experiment evaluates how the identification method performs in a binary configuration (positive and negative). In both experiments, we used the same corpus. The corpus contains newspaper headlines extracted from several Brazilian online news sites, such as www.globo.com. The corpus contains 1,312 labeled texts. It is annotated base on Eckman's emotions at the text level. Table 1 shows the distribution of each emotion in the corpus. The corpus was preprocessed setting up the Information Gain to 80% (using the top 80% sorted words).

Emotion	# of texts	% in the corpus
Joy	280	21.34
Disgust	226	17.23
Fear	160	12.20
Anger	168	12.80
Surprise	172	13.11
Sadness	306	23.32

Table 1. Distribution of Emotions in the Corpus.

In the first experiment, the corpus contains all available data (1,312 texts), with six classes/emotions. We ran the experiment using the original (imbalanced) corpus and the balanced corpus. The GA was started with 100 individuals, using the initialization approach shown in the previous section. Table 2 shows the obtained results.

Class	*Precision*		*Recall*		*F1*	
	Imbalanced	*Balanced*	*Imbalanced*	*Balanced*	*Imbalanced*	*Balanced*
Joy	0.461	0.599	0.868	0.704	0.602	0.647
Disgust	0.623	0.507	0.212	0.469	0.317	0.487
Fear	0.857	0.750	0.188	0.488	0.308	0.591
Anger	0.764	0.723	0.327	0.589	0.458	0.649
Surprise	0.831	0.733	0.285	0.576	0.424	0.645
Sadness	0.428	0.550	0.758	0.716	0.547	0.622
Accuracy (Imbalanced): 50.07%			Accuracy (Balanced): 60.82%			

Table 2. Multiclass Experiment Results.

The second experiment reduces the problem to a binary classification, where the positive class represents the emotion *joy* and the negative class represents the emotions *sadness*, *anger*, *disgust* and *fear*. In this case, texts labelled as *surprise* were not used, since it is difficult to classify them as positive or negative. Table 3 shows the number of texts for each class. As we can conclude, in this configuration, the classes are more imbalanced than the first one.

Polarity	**# of texts**	**% in the corpus**
Positive	280	24.56
Negative	860	75.44

Table 3. Number of Texts in the Polarity Experiment.

Table 4 shows the results for two classes.

Class	*Precision*		*Recall*		*F1*	
	Imbalanced	*Balanced*	*Imbalanced*	*Balanced*	*Imbalanced*	*Balanced*
Positive	0.933	0.754	0.050	0.646	0.095	0.696
Negative	0.764	0.890	0.999	0.931	0.865	0.910
Accuracy (Imbalanced): 76.58%			Accuracy (Balanced): 86.14%			

Table 4. Polarity Experiment Results.

To study the impact of balancing the corpus in Sentiment Analysis, we have defined that is more important to improve the F1 for each class, than improving accuracy. Thus, we studied the behavior of F1 in both experiments. We can conclude based on experiments, that balanced corpus, as expected, improved the SVM results. This is especially important when trying to identify Pure Emotions (six emotions) in texts. As we noted, most researchers work with balanced data, despite they know that their domain is commonly imbalanced (Ghazi, Inkpen, and Szpakowicz 2010).

Tables 2 and 4 show that using the GA approach, accuracy improved. Statistical tests (Wilcoxon and t-test) were done to certify that results are statistically valid. The impact is more important in the multiclass scenario (six class). We might note that recall decreases for some classes in the multiclass problem. This is due to the fact that, in order to improve the F1 measure, recall of majorities' classes should be reduced. On the opposite, recall of minorities' classes should be increased.

Conclusions and Future Work

Most textual corpora used in Sentiment Analysis, such as newspaper articles or blog posts, are naturally imbalanced. In this paper we presented a Genetic Algorithm approach to balance the corpus of texts in order to investigate the impact of this action in the classification level when using a SVM classifier. Results showed that balancing the corpus could be an alternative for emotion identification in texts (multi-emotion identification).

In the near future we intend to use different corpora, such as blog post or tweets. We planed to test this approach in different domains, such as, automatic bird species classification.

We are also interested in studying the impact of balancing the corpus using only under-sampling (EasyEnsemble or BalanceCascade approaches) (Liu, Wu, and Zhou 2006) or over-sampling (SMOTE algorithm) (Chawla et al. 2002). Different classifiers will also be tested.

References

Akbani, R., Kwek, S., and Japkowicz, N. 2004. Applying support vector machines to imbalanced datasets. In Proceedings of the 15th European Conference on Machine Learning, 39-50, Pisa, Italy.

Alejo, R., Garcia, V., Sotoca, J. M., Mollineda, R. A., and Senchez, J. S. 2007. Improving the performance of the RBF neural networks trained with imbalanced samples, Volume 4507 of Lecture Notes in Computer Science, 162-169.

Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. 2004. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. vol. 6, issue 1, 20-29.

Chaffar, S., and Inkpen, D. 2011. Using a Heterogeneous Dataset for Emotion Analysis in Text. 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, St. John's, Canada.

Chawla, N. V. 2005. Data Mining for Imbalanced Datasets: An Overview. Data Mining and Knowledge Discovery Handbook, 853-867.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, vol. 16.

Ekman, P., and Friesen, W. V. 1978. Facial Action Coding System. Palo Alto: Consulting Psychologists Press.

Ghazi, D., Inkpen, D., Szpakowicz, S. 2010. Hierarchical Approach to Emotion Recognition and Classification in Texts. Advances in Artificial Intelligence. 23rd Canadian Conference on Artificial Intelligence, Canada, 40-50.

He, H. and Garcia, E. A. 2009. Learning from Imbalanced Data. IEEE Transaction on Knowledge and Data Engineering. 21, 9, 1263-1284.

Hsu, C., and Lin, C. 2002. A comparison of methods for multi-class support vector machines. IEEE Transactions on Neural Networks, 13(2): 415-425.

Japkowicz, N., and Stephen, S. 2002. The class imbalance problem: A systematic study. Journal Intelligent Data Analysis, vol. 6. issues 5, 429-449.

Khoshgoftaar, T. M., Hulse, J. V., and Napolitano, A. 2010. Supervised neural network modeling: An empirical investigation into learning from imbalanced data with labeling errors. IEEE Trans. on Neural Networks vol. 21, issues 5, 813-830.

Li, S., Wang, Z., Zhou, G., and Lee, S. Y. M. 2011. Semi-supervised learning for imbalanced sentiment classification. In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence, Volume Three. AAAI Press, 1826-1831.

Liu, B. 2012. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. University of Illinois at Chicago. Morgan & Claypool Publishers.

Liu, B. 2010. Sentiment Analysis: A Multifaceted Problem. IEEE Intelligent Systems, vol. 25(3), 76-80.

Liu, X. Y., Wu, J., Zhou, Z. H. 2006 Exploratory Under Sampling for Class Imbalance Learning, In: International Conference IEEE on Data Mining.

Mitchell, T. 1997. Machine Learning. McGraw-Hill, New York.

Nguyen, G. H., Bouzerdoum, A. and Phung, S. 2008. A supervised learning approach for imbalanced data sets, Pattern Recognition, 1-4.

Plutchik, R. 2001. The Nature of Emotions. American Scientist 89: 344-350.

Radovanovic, M., and Ivanovic, M. 2008. Text Mining: Approaches and Applications. Novi Sad J. Math.V. 38, N. 3.

Rangel, F., and Rosso, P. 2013. On the Identification of Emotions and Authors' Gender in Facebook Comments on the Basis of their Writing Style. In proc.: Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI. ESSEM 2013 at XIII Conference of the Italian Association for Artificial Intelligence, 4-6.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management - Cornell University.

Strapparava, C., and Mihalcea R. 2008. Learning to Identify Emotions in Text. 23rd Annual ACM Symposium on Applied Computing, 1556-1560.

Strapparava, C., and Valitutti, A. 2004. WordNet-Affect: an affective extension of WordNet. In Proc. of 4th International Conference on Language Resources and Evaluation, Lisbon.

Van Rijsbergen, C. J. 1979. Information Retrieval (2nd ed.). Butterworth.

Weiss, G. and Provost, F. 2003. Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction, Volume 19, 315-354.

Zhou, Z. and Liu, X. 2006. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 16(1): 63-77.