Reference for this paper is:
Huang, T.M., Kecman, V., *Semi-supervised Learning from Unbalanced Labeled Data – An Improvement*, 'International Journal of Knowledge Based and Emergent Technologies Relied Intelligent Information and Engineering Systems', ISO Press, in print, 2005

Semi-supervised Learning from Unbalanced Labeled Data – An Improvement

Te Ming Huang, Vojislav Kecman

School of Engineering, The University of Auckland, Auckland, New Zealand
e-mail: v.kecman@auckland.ac.nz, huangjh@win.co.nz

Abstract. We present a possibly great improvement while performing semi-supervised learning tasks from training data sets when only a small fraction of the data pairs is labeled. In particular, we propose a novel decision strategy based on normalized model outputs. The paper compares performances of two popular semi-supervised approaches (Consistency Method and Harmonic Gaussian Model) on the unbalanced and balanced labeled data by using normalization of the models' outputs and without it. Experiments on text categorization problems suggest significant improvements in classification performances for models that use normalized outputs as a basis for final decision.

1. Introduction

Today, there are many learning from data paradigms, the most popular and the most used ones being classification and regression models [2]. They belong to the so-called *supervised* learning algorithms in which a learning machine attempts to learn the input-output relationship (dependency or function) $f(\mathbf{x})$ by using a training data set $X = \{[\mathbf{x}(i), y(i)] \in \Re^m \times \Re, i = 1,...,n\}$ consisting of n pairs $(\mathbf{x}_1, y_1)$, $(\mathbf{x}_2, y_2)$, …, $(\mathbf{x}_n, y_n)$, where the inputs $\mathbf{x}$ are m-dimensional vectors $\mathbf{x} \in \Re^m$ and the labels (or system responses) $y \in \Re$ are continuous values for regression tasks and discrete (e.g., Boolean) for classification problems. Another large group of standard learning algorithms are the ones dubbed as *unsupervised* ones when there are only raw data $\mathbf{x}_i \in \Re^m$ without the corresponding labels y_i (i.e., there is a 'no-teacher' in a shape of labels). The most popular, representative, algorithms belonging to this group are various clustering and (principal or independent) component analysis routines.

Recently, however, we are facing more and more instances in which the learning problems are characterized by the presence of (usually) a small percentage of labeled data only. In this novel setting, the task is to predict the labels (or the belonging to some class) of the unlabeled data points. This learning task belongs to the so-called *semi-supervised* or *transductive inference* problems. The cause for an appearance of the unlabeled data points is usually expensive, difficult and slow process of obtaining labeled data. Thus, labeling brings the costs and often it is not feasible. The typical areas where this happens is the speech processing (due to the slow transcription), text categorization (due to huge number of documents, slow reading by humans and their general lack of a capacity for a concentrated reading activity), web categorization,

and, finally, a bioinformatics area where it is usually both expensive and slow to label huge number of data produced.

Recently several approaches to the semi-supervised learning were proposed. Here, we present, compare and improve the two transductive approaches, namely, the harmonic Gaussian model introduced in [6] and consistency method for semi-supervised learning proposed in [5].

However, none of the methods successfully analyzes the possible problems connected with the so-called unbalanced labeled data, meaning with the situations when the number of labeled data differs very much between the classes. We propose the normalization of the classifier outputs before a final decision about the labeling is done.

Paper is organized as follows: In section 2 we present the basic forms of the two methods. Section 3 introduces the normalization step which improves the performance of both the consistency method and the harmonic Gaussian model faced with unbalanced labeling significantly. It also compares the effects of normalization with the results of both methods obtained and presented in [5]. Section 4 concludes the presentations here and proposes possible avenues for the further research in this novel area of semi-supervised learning.

2. Consistency Method Algorithm and Harmonic Gaussian Model

There exist a great variety of methods and approaches in semi-supervised learning. The powerful software SemiL for solving semi-supervised (transductive) problems, used within this study, is capable of using 12 different models for a semi-supervised learning (as suggested in [4]). Namely, it can solve the following variously shaped semi-supervised learning algorithms: both the hard label approach with the maximization of smoothness and the soft label approach with the maximization of smoothness, for all three types of models (i.e., Basic Model, Norm Constrained Model and Bound Constrained Model) and by using either Standard or Normalized Laplacian. Presenting all the variety of results would require much bigger space than it is allowed within the constrained space allotted here. That's why the presentation here will be focused on two basic models only, and on an introduction of a normalization step as the first possible significant stage in improving results to date.

Below we present Global consistency model from [5] which is a soft label approach with the maximization of smoothness that uses a normalized Laplacian without a norm constraint, as well as the Harmonic Gaussian method presented in [6] which is a hard label approach with the maximization of smoothness that uses a standard Laplacian also without a norm constraint.

2.1 Global consistency model

The presentation here follows the basic model proposed in [5] tightly.

Given a point set X as defined in the Introduction the first l points x_i are labeled, and the remaining points $x_u (l + 1 \leq u \leq n)$ are unlabeled. The goal is to predict the label of the unlabeled points.

Let F denote the set of $n \times c$ matrices with nonnegative entries. A matrix $\mathbf{F} = [\mathbf{F}_1^T, \ldots, \mathbf{F}_n^T]^T \in F$ corresponds to a classification on the dataset X by labeling each point x_i as a label $y_i = \arg\max_{j \leq c} F_{ij}$. We can understand $\mathbf{F}$ as a vectorial function $F : X \rightarrow R^c$ which assigns a vector $\mathbf{F}_i$ to each point x_i. Define an $n \times c$ matrix $\mathbf{Y} \in F$ with $Y_{ij} = 1$ if x_i is labeled as $y_i = j$ and $Y_{ij} = 0$ otherwise. Clearly, $\mathbf{Y}$ is consistent with the initial labels according the decision rule. The algorithm is as follows:

1. Form the affinity matrix $\mathbf{W}$ defined by $W_{ij} = \exp(-\| x_i - x_j \|^2 / 2\sigma^2)$ if $i \neq j$ and $W_{ii} = 0$.
2. Construct the matrix $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ in which $\mathbf{D}$ is a diagonal matrix with its (i, i)-element equal to the sum of the i-th row of $\mathbf{W}$.
3. Iterate $\mathbf{F}(t+1) = \alpha\mathbf{S}\mathbf{F}(t)+(1 - \alpha)\mathbf{Y}$ until convergence, where α is a parameter in (0, 1).
4. Let $\mathbf{F}^*$ denotes the limit of the sequence $\{F(t)\}$. Label each point x_i as a label $y_i = \arg\max_{j \leq c} F_{ij}^*$.

First, one calculates a pairwise relationship $\mathbf{W}$ on the dataset X with the diagonal elements being zero. In doing this, one can think of a graph $G = (V, E)$ defined on X, where the vertex set V is just X and the edges E are weighted by $\mathbf{W}$. In the second step, the weight matrix $\mathbf{W}$ of G is normalized symmetrically, which is necessary for the convergence of the following iteration. The first two steps are exactly the same as in spectral clustering [3]. Here, we did not solve the problem in an iterative way as shown above. Instead, we solve the corresponding equivalent system of linear equations $(\mathbf{I} - \alpha\mathbf{S})\ \mathbf{F}^* = \mathbf{Y}$ for $\mathbf{F}^*$ by using conjugate gradient method which is highly recommended approach for dealing with huge data set. Also, instead of using the complete graph we calculated the $\mathbf{W}$ matrix by using only 10 nearest neighbors. This step decreases the accuracy only slightly, but it increases the calculation speed significantly. Note that *self-reinforcement* is avoided since the diagonal elements of the affinity matrix are set to zero in the first step ($W_{ij} = 0$). The model labels each unlabeled point and assigns it to the class for which the corresponding $\mathbf{F}^*$ value is the biggest, as given in step 4above.

2.2 Harmonic Gaussian Model

The presentation here also follows the basic model proposed in [6] tightly. The algorithm is as follows:

1. Form the affinity matrix $\mathbf{W}$ defined by $W_{ij} = \exp(-\| x_i - x_j \|^2 / 2\sigma^2)$.
2. Construct the diagonal matrix $\mathbf{D}$ with its (i, i)-element equal to the sum of the i-th row of $\mathbf{W}$. Note that we can use $\mathbf{W}$ and $\mathbf{D}$ as given in section 2.1 above too.

3. Form the following two matrices $\mathbf{W} = \begin{bmatrix} \mathbf{W}_{ll} & \mathbf{W}_{lu} \\ \mathbf{W}_{ul} & \mathbf{W}_{uu} \end{bmatrix}$, $\mathbf{D} = \begin{bmatrix} \mathbf{D}_{ll} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{uu} \end{bmatrix}$ as well as the vector $\mathbf{f} = [\mathbf{f}_l \quad \mathbf{f}_u]^T$, where *l* stands for the labeled data points and *u* for the unlabeled ones.
4. Solve for $\mathbf{f}_u$ as follows $\mathbf{f}_u = (\mathbf{D}_{uu} - \mathbf{W}_{uu})^{-1} \mathbf{W}_{ul}\mathbf{f}_l$ which is the solution for the unlabeled data points.

More detailed description of the two basic models, namely, the global consistency model and the harmonic Gaussian model can be found in [5] and [6] respectively.

3. Performance of the Two Models and Possible Improvement

The extensive simulations on various data sets (as presented in [5]) have indicated that both models behave similarly and according to the expectations that with an increase in the number of labeled data points *l*, the overall models' accuracies improve too. There was just a slightly more superior performance of the consistency model from [5] in respect to the harmonic Gaussian model, when faced with a small number of unbalanced labeled data. At the same time, the later model performed much better for extremely small number of the labeled data as long as they are balanced (meaning there is a same number of the labeled points for all the classes. Here, an extremely small number meant 1 labeled data per each class only, in the text categorization problem from [5]).

Such a behavior needed a correct explanation and it asked for further investigations during which several phenomena have been observed. While working with balanced labeled data (meaning with the same number of labeled data per class) harmonic Gaussian method performed better than the consistency model. On the contrary, for a small number of unbalanced labeled data, the harmonic Gaussian model performed worse than the consistency one. This indicates a sensitivity of the former while working with the unbalanced labeled data.

At the same time a simulation shows that in the harmonic Gaussian method the mean value of the class with less labeled points is lower than for the classes with more labeled data. Recall that the final decision is made based on the maximum of the $\mathbf{F}^*$ values and obviously the elements of the class with less labeled data could be assigned to different class just due to the fact that the (mean) values of other classes are higher.

The causes of these phenomena can be understand by interpreting both algorithms in term of a lazy random walking which is determined by the transition probability matrix $\mathbf{P}=(1-\alpha)\mathbf{I}+\alpha\mathbf{D}^{-1}\mathbf{W}$ as shown in [7]. In this setting, one can find the expected number of steps for a random walk starting at some initial position x_i to reach x_j and then return. This expectation is often referred as the commute time between the two positions and denoted by $\mathbf{C}_{ij}$. It has been shown that

$$\mathbf{C}_{ij} \propto \mathbf{G}_{ii} + \mathbf{G}_{ij} - \mathbf{G}_{ii} - \mathbf{G}_{ji} \quad \text{if} \quad x_i \neq x_j$$

where **G** is the inverse of the matrix **D**-α**W**. If we now consider a binary classification that is given by **f**=(**D**-α**W**)$^{-1}$**y** then the classification is based on the comparison between $p_+(x_i) = \sum_{\{j|y_j=1\}} \mathbf{G}_{ij}$ and $p_-(x_i) = \sum_{\{j|y_j=-1\}} \mathbf{G}_{ij}$ [7]. This means, we are labeling an unlabeled point by summing up and comparing the commute times of this point to all the positive labeled points and to all the negative labeled points. With more positive labeled points, the mean of **f** will be greater than zero and vice verse. Similar phenomenal will occur in the global consistency method, but instead of using the commute time as a measure of distance, a normalized commute time obtaining from the inverse of the normalized Laplacian **(I-**α**S)** matrix is used. In the harmonic Gaussian methods, the original commute time is used, but instead of using the commute time between the labeled and the unlabeled points, it uses commute times between the point of interested to the rest of the unlabeled points. These commute times are weighted by $\mathbf{W}_{ul}\mathbf{f}_l$ first and then added together. Again, the mean of **f** will still be affected by the difference in number between positive label and negative label points. If we now consider solving a multi-class problem using several binary classifiers, then a binary classifier with less number of positive labeled points will be more disadvantageous then others, because the mean of its output **f** will be lower, i.e., the class with less labeled points will be disadvantageous. The figure bellow demonstrates the effect of unbalanced labeled data. It is clear that with unbalanced labeled data, the performance of the algorithm can deteriorate even with more labeled data available.

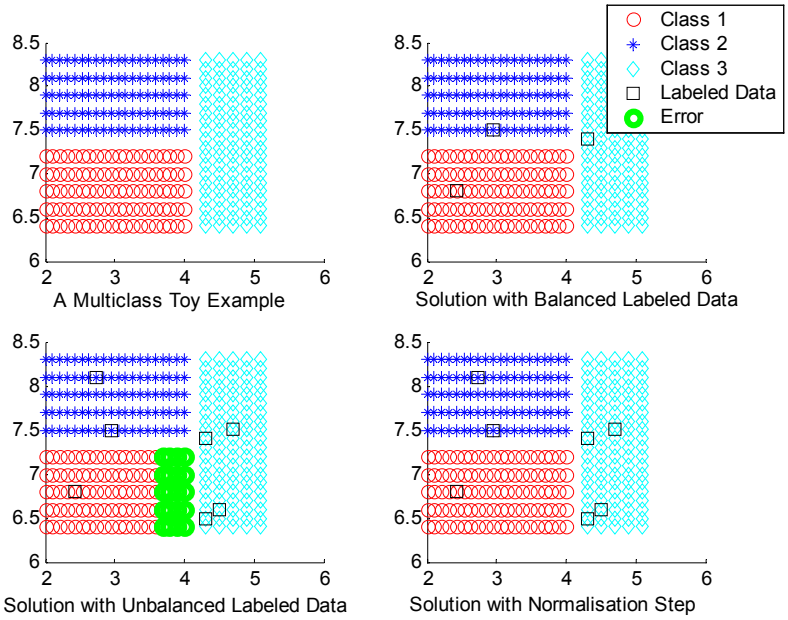

Fig. 1. A multi-class toy example for demonstrating the effect of unbalanced labeled data.

This led us to the introduction of a normalization step for the elements of the column vectors $\mathbf{F}_i^*$ bringing them to the vectors with a mean = 0, and with a standard

deviation = 1. Only now, after the normalization is performed, the algorithm searches for the maximal value along the rows of a matrix $\mathbf{F}^*$ and labels the unlabeled i-th data to the class j if $F^*_{ij} > F^*_{ik}$, $k = 1, c$, $k \neq j$.

The introduction of the normalization step improves the behavior of the algorithm significantly as it is shown in Fig. 2, where we compare performances of the two models without normalization as given in [5] to the performances of both models incorporating a normalization part.

Same as in [5], in the experiment here, we investigated the task of text classification using the 20-newsgroups dataset. The chosen topic was *rec* which contains *autos, motorcycles, baseball,* and *hockey* from the version 20-news-18828. The articles were processed by the Rainbow software package with the following options: (1) passing all words through the Porter stemmer before counting them; (2) tossing out any token which is on the stop list of the SMART system; (3) skipping any headers; (4) ignoring words that occur in 5 or fewer documents. No further preprocessing was done. Removing the empty documents, we obtained 3970 document vectors in a 8014-dimensional space. Finally the documents were normalized into TFIDF representation. The cosine distance between points was used here too. The mentioned procedure is the same as in [5] just in order to ensure the same experiment's setting for same data set.

We played with various widths of the Gaussian RBF and the results with a few σ-s are shown in Fig. 1. The results in [5] use $\sigma = 0.15$ for both harmonic Gaussian method and consistency method. The test errors shown are averaged over 100 trials. Samples were chosen so that they contain at least one labeled point for each class. Thus, the setting of the experiment is identical to the one in [5].

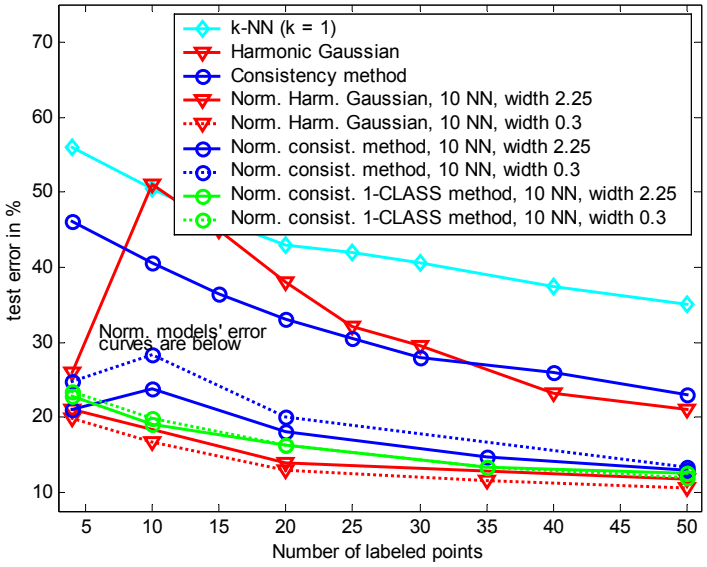

Fig. 2. The error rates of text classification with 3970 document vectors in an 8014-dimensional space for recreation data sets from version 20-news-18828. At least one labeled

data for each class must be labeled. The smallest number of labeled data here is therefore 4. The normalized model outputs outperform the algorithms without normalization

Several interesting phenomena can be observed in Fig. 2. First, the normalization improves the performances of both methods very significant ly. This can be observed easily by comparing the error rates between the models with and without normalization. The error rates of the consistency method for four labeled points drop from 46% to 22%. When 50 points are labeled, the error rates drop from around 22% to about 13% and similar improvements can be found on the harmonic Gaussian method.

The only exception is in the case of the later method when there are only four labeled points available. In this situation, the error rate of the harmonic Gaussian is already much lower than the consistency method's one, even without normalization and the improvement by normalization is not as significant as in other cases. This is a consequence of having balanced labeled data points from each class (1 in each class). Hence, the mean values of $\mathbf{F}^*$ along each column are closer to each other and there is no need for normalization.

In contrast, when the number of labeled points in each class is different (i.e., unbalanced which is the case whenever there is more than 4 labeled data for four classes and random labeling is used) the performance gain from normalization is more significant. The negative effect of unbalanced data can be observed from following the increase in error rate when working with ten data of labeled points and if normalization is not applied within the harmonic Gaussian method. Without normalization, the harmonic Gaussian method needs approximately forty unbalanced labeled points to match its very performance when having four balanced labeled points only. In contrast, the performance of the normalized model with ten unbalanced labeled data outperforms the result for the four balanced points. With a normalization step, the harmonic Gaussian method seems to be slightly better than the consistency method. This is not the case while working without the normalization. The best model for the text categorization data in our experiments is a harmonic Gaussian model with width equal to 0.3 which achieves an accuracy of 90% with only 50 labeled points out of 3970 of the total data points. For both methods with normalization of $\mathbf{F}^*$, models with smaller width parameter perform slightly better than with the larger widths. Finally, for a 3970 data, the learning run based on a conjugate gradient algorithm takes only about 25 seconds of a CPU time on a 2MHz laptop machine for 100 random tests runs.

4. Conclusions

The extensive simulations have shown that an introduction of a normalization step improves the behavior of both transductive inference models (namely, consistency method and harmonic Gaussian one) very significantly. In both methods, the normalization of $\mathbf{F}^*$ improves the performance up to fifty percents. However, the results are inconclusive, because many areas still need to be explored and more investigations are needed before final conclusions. For example, in this study we only investigate two basic models out of the twelve possible models mentioned earlier. Also, there are several parameters associated with these algorithms which can alter the overall per-

formance of the model, e.g., the parameter for constraining the norm of $\mathbf{F}^*$ (as suggested in [4]) can also have some impact on the performance of the models. This means that there may still be some space for improving the performance of the semi-supervised learning algorithms even further. In addition, the effects of a normalization step for other data set should also be further explored. The work presented here, can be treated as an initial step in this area only. It demonstrated that the way how the decisions are made from the output of these models can have a significant impact on the final classification performance. Our future work will go along the path of finding better decision strategies.

Acknowledgement

Our particular thank goes to Dr. Chan-Kyoo Park for all his support and communications on various guises of graph-based semi-supervised learning algorithms. We also thank Dr. Dengyong Zhou for introducing the first author to this area during his short stay at Max Planck Institute in Tübingen.

References

1. Huang, T. M., Kecman, V.: SemiL, Software for solving semi-supervised learning problems, [downloadable from: http://www.support-vector.ws/html/semil.html or from http://www.engineers.auckland.ac.nz/~vkec001], Auckland, (2004)
2. Kecman, V.: Learning and Soft Computing, Support Vector Machines, Neural Networks and Fuzzy Logic Systems, The MIT Press, Cambridge, MA, (2001)
3. Ng, A. Y., Jordan, M., Weiss, Y.: On Spectral Clustering: Analysis and an Algorithm, Advances in Neural Information Processing Systems **14**, (Eds.) Dietterich, T. G., Ghahramani, Z., MIT Press, Cambridge, Mass. (2002)
4. Park, C., Personal Communication, Tübingen, (2004)
5. Zhou, D., Bousquet, O., Lal, T. N., Weston, J., Schölkopf, B.: Learning with Local and Global Consistency, Advances in Neural Information Processing Systems **16**, (Eds.) Thrun, S., L. Saul and B. Schölkopf, MIT Press, Cambridge, Mass. (2004) 321-328
6. Zhu, X.-J., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using Gaussian fields and harmonic functions, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, (2003)
7. Zhou,D., Schölkopf, B.: A Requlariza tion Framework for Learning from Graph Data. Workshop on Statistical Relational Learning at Twenty-first International Conference on Machine Learning