SEMI-SUPERVISED LEARNING USING A GRAPH-BASED PHASE FIELD MODEL FOR IMBALANCED DATA SET CLASSIFICATION

Aymen El Ghoul and Hichem Sahbi

CNRS TELECOM ParisTech

ABSTRACT

In this paper, we address the problem of semi-supervised learning for binary classification. This task is known to be challenging due to several issues including: the scarceness of labeled data, the large intra-class variability, and also the imbalanced class distributions.

Our learning approach is transductive and built upon a graph-based phase field model that handles imbalanced class distributions. This method is able to encourage or penalize the memberships of data to different classes according to an explicit a priori model that avoids biased classifications. Experiments, conducted on real-world benchmarks, show the good performance of our model compared to several state of the art semi-supervised learning algorithms.

Index Terms— Transductive Learning, Statistical Machine Learning, Graph-based Inference, Imbalanced-class Distributions, Image and Data Classification.

1. INTRODUCTION

There is an increasing interest in semi-supervised learning (SSL) that exploits both labeled and unlabeled data during inference [1, 2, 3]. This interest is mainly motivated by the lack of labeled data for different training and classification tasks. Most of SSL methods are either transductive or inductive. Transductive methods [4] aim to predict labels of test sets that coincide with unlabeled data while inductive methods [5] seek to learn an explicit decision rule, using both labeled and unlabeled data, in order to predict labels on unseen test data. Transductive methods can also be extended on unseen test data using, for instance, the Nystrom extension. In all these SSL methods, three main assumptions are considered [1]. Firstly, the smoothness assumption which assigns similar labels to samples belonging to the same dense area of the input space. Secondly, the cluster assumption which seeks to find a decision boundary in low density areas of the input space so samples belonging to separate clusters tend to have different labels. Finally, the manifold assumption which states that high-dimensional data are distributed in low-dimensional manifolds.

Considering $\mathcal{X}$ as the union of labeled and unlabeled data; in what follows, we seek to learn a decision criterion that splits $\mathcal{X}$ into a positive subset $\mathcal{R}$ (corresponding to the class of interest) and a negative subset $\bar{\mathcal{R}}$. We will derive this decision criterion by minimizing an energy function which mixes (i) a prior term E_P that models a prior knowledge about class distribution in $\mathcal{X}$, and (ii) a data term E_D which fits the decision criterion to labeled data. In many real-world applications, (e.g. handwritten digit recognition, image classification, text classification, etc.), the volume or the number of points in $\mathcal{R}$ is much smaller than in $\bar{\mathcal{R}}$ especially when the class of interest corresponds to a rare event. This imbalanced configuration of $\mathcal{R}$ and $\bar{\mathcal{R}}$ makes the binary classification problem hard as the contribution of the negative class becomes more dominant. This paper addresses, particularly, the problem of binary classification for imbalanced data which also rely on the manifold assumption.

A given set $\mathcal{X}$ is assumed to be distributed in a low-dimensional manifold denoted $\mathcal{M}$. SSL methods, which hold the manifold assumption [6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22], rely on the intrinsic properties of data and consider that the underlying classifier, that separates positive and negative data, is smooth along the tangent space of $\mathcal{M}$. Learning and classification algorithms operating on imbalanced class distributions are studied in the literature with two levels of contributions (see for instance [23, 24]). From the data point-of-view, over or under sampling strategies are proposed in order to re-balance the class distributions prior to apply usual learning algorithms. For more challenging data sets these algorithms hit two major limitations: an interesting limit is when the cardinality of the rare class is only one and much smaller compared to the cardinality of the dominant class. Therefore, training should be achieved multiple (K) times[1] on subsets including only one sample per class, and this is clearly insufficient to accurately capture the decision boundary. Furthermore, the learning process becomes prohibitive as K reaches very large values (for instance in [24], $K = 900$).

From the methodological point-of-view, advanced models, incorporating a priori knowledge about class distributions, enable to reduce the bias due to the dominant classes. These methods have shown successful results in many practical problems particularly text and image classification [23]. In

[1] with K being the ratio between the cardinalities of the dominant and the non-dominant classes.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

[25], the proposed graph-based model incorporates prior knowledge about the class distribution of the labeled data by re-weighting the contribution of positive and negative data in the learned classifier. In [17], authors incorporate an a prior term to model imbalanced classes on labeled data. The key idea is to minimize a distance between discrete probabilities of class memberships using balance priors only on the labeled data. Notice that both models in [25, 17] fail for highly imbalanced class distributions in the unlabeled data, especially when the amount of the labeled data is significantly small compared to the unlabeled ones.

In this paper, we introduce a novel semi-supervised learning algorithm based on a phase field model that tackles the imbalanced class problem. The phase field model has been studied in image processing and computer vision mainly for classification of 2D images [26, 27, 28, 29, 30, 31] and high dimensional data [9]. In [9], the proposed graph-based phase field model assigns the same energy to positive and negative classes. In other words, a given data point has the same probability of being labeled positively or negatively; however this method fails when the two classes are highly imbalanced.
To the best of our knowledge, none of the existing work uses the phase field model in order to tackle imbalanced class distributions for learning. Our work then extends the phase field model, in [9], with the following contributions:
-The construction of a new phase field prior on graphs, that overcomes the bias due to the dominant class;
- And a comprehensive analysis of the proposed model using an ansatz to emphasize the role of the balance prior.
Our proposed prior phase field energy includes two terms. The first one guarantees the smoothness of the learned classifier along the tangent space of $\mathcal{M}$ while the second term encourages the response of the classifier to be close to $+1$ inside $\mathcal{R}$ and -1 otherwise; this term assigns different energies to the phases $+1$ and -1, to correct the bias due to imbalance class distribution *both on labeled and unlabeled data*.

This paper is organized as follows. In Section 2, we first recall the basic 2D phase field model, used in [30], and we introduce our extended phase field model on graphs. In Section 3, we show the performance of our model on real-world benchmarks and comparison w.r.t related work. Finally, we conclude the paper in Section 4.

2. OUR GRAPH-BASED PHASE FIELD MODEL

In this section, we first recall the basic phase field model in [30] defined on 2D Euclidean spaces, then we introduce our prior for data classification. We extend this with a theoretical and an algorithmic analysis of our proposed model and its ability to handle imbalanced classes in $\mathcal{R}$ and $\bar{\mathcal{R}}$.

2.1. A 2D phase field model

A phase field ϕ is a real-valued function defined on an image domain $\Omega \subset \mathbb{R}^2$. A phase field determines a region by the map $\xi_z(\phi) = \{x \in \Omega : \phi(x) > z\}$ where z is a given threshold. The phase field energy is [30]

$$E(\phi) = \int_\Omega d^2x \left\{ \frac{\epsilon}{2}\partial\phi.\partial\phi + \lambda\left(\frac{\phi^4}{4} - \frac{\phi^2}{2}\right) + \alpha\left(\phi - \frac{\phi^3}{3}\right)\right\}. \quad (1)$$

If Eq. (1) is minimized subject to $\xi_z(\phi) = \mathcal{R}$, i.e. for a fixed region, then away from the boundary $\partial\mathcal{R}$, the minimizing function $\phi_\mathcal{R}$ assumes approximately the value $+1$ inside, and -1 outside $\mathcal{R}$ thanks to the ultralocal terms. The derivative term ensures the smoothness of $\phi_\mathcal{R}$, producing a narrow interface around $\partial\mathcal{R}$ and interpolating between -1 and $+1$.

2.2. A graph-based extension

In what follows, we extend the model in Eq. 1 to data in $\mathcal{X}$ lying on a given manifold $\mathcal{M}$. Considering a semi-supervised setting, we are given ℓ labeled samples $\mathcal{L} = \{(x_i, y_i)\}_{i=1}^{\ell}$, with y_i being the label of x_i, and u unlabeled samples $\mathcal{U} = \{x_i\}_{i=\ell+1}^{\ell+u}$. The whole training set is then defined as $\mathcal{X} = \mathcal{L} \cup \mathcal{U}$. Our goal is to infer the label of a given $x \in \mathcal{U}$ by optimizing a real-valued classifier defined on $\mathcal{X}$. Let $\phi = (\phi_1, \phi_2, \ldots, \phi_{\ell+u})^t$ be a vector with each entry equal to the value of that classifier on samples in $\{x_i\}_{i=1}^{\ell+u}$ where ^{t} denotes matrix transpose.
In the case of a manifold $\mathcal{M}$, we replace the gradient operator in Eq. 1 by its generalization $\partial\mathcal{M}$ as:

$$\int_\mathcal{M} d\mu(x)\partial_\mathcal{M}\phi(x).\partial_\mathcal{M}\phi(x) = -\int_\mathcal{M} d\mu(x)\phi(x)\partial^2_\mathcal{M}\phi(x), \quad (2)$$

where $\partial^2_\mathcal{M}$ and $d\mu$ denote respectively the Laplace-Beltrami operator and the volume measure defined on $\mathcal{M}$. In practice, the intrinsic geometric properties of $\mathcal{X}$ are described with a graph, denoted $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes nodes associated to samples in $\mathcal{X}$, and $\mathcal{E}$ denotes edges describing relationships (e.g. similarities) between nodes in $\mathcal{V}$. Considering a discrete approximation of $-\partial^2_\mathcal{M}$ using a Laplacian operator on graphs [32], we obtain a graph-based discrete form of our energy term in Eq.1 as

$$E_P(\phi) = \frac{\epsilon}{2\sum_{i,j=1}^{\ell+u} w_{ij}}\phi^t L\phi + \frac{1}{\sum_{i,j=1}^{\ell+u} w_{ij}}\sum_{i=1}^{\ell+u} d_i V(\phi_i), \quad (3)$$

here $\phi^t L\phi = \frac{1}{2}\sum_{i\sim j} w_{ij}(\phi_i - \phi_j)^2$, with $w_{ij} = \exp\{-\|x_i - x_j\|_2^2/2\sigma^2\}$ and L is the graph Laplacian. V is the ultralocal term given by $V(t) = \lambda\big(\frac{t^4}{4} - \frac{t^2}{2}\big) + \alpha\big(t - \frac{t^3}{3}\big)$, and d_i is a volume measure which is chosen as the degree of x_i (i.e., $d_i = \sum_{j\sim i} w_{ij}$). In the equation above, $\epsilon > 0$, $\lambda > 0$, $\alpha \in]-\lambda, \lambda[$ are the three parameters of the model and $1/\sum_{i,j=1}^{\ell+u} w_{ij}$ is a normalizing factor for a sparse similarity matrix $W = (w_{ij})_{ij}$; for a non-sparse matrix W one may use instead $1/(\ell+u)^2$ as a normalization.

If energy (3) is minimized for a given subset $\mathcal{R}$, then away from the hypersurface $\partial\mathcal{R}$, $\phi_\mathcal{R}$ is close to $+1$ inside,

and -1 outside $\mathcal{R}$; and it varies smoothly (depending on ϵ and λ) across the interface near $\partial\mathcal{R}$ thanks to the smoothing (i.e. Laplacian) term. The degree d_i captures the information about the density of points: d_i is high (resp. low) depending on whether x_i is located in high-density (resp. low-density) areas of the space. Thus, the ultralocal term V contributes significantly in high-density regions, and weakly in low-density regions; and this facilitates a smooth transition of ϕ along low density regions, guaranteed by the Laplacian term.

Notice that the formulation of our model yields to choose the threshold $z = \alpha/\lambda$ to be at the maximum of V (see Fig. 1). To guarantee two energy minima at -1 and $+1$ associated to the two classes, the inequality $\lambda > |\alpha|$ must be satisfied: $V'(\pm 1) = 0$ and $V''(\pm 1) > 0$ where $'$ and $''$ denote the first and second derivatives respectively. When setting $\alpha = 0$ and $\lambda = 1/\epsilon$ and choosing a uniform volume measure in front of V, we get the particular case of the phase field model proposed in [9]. This leads to $V(+1) = V(-1)$ corresponding to equiprobable phases ± 1, and hence the model E_P is suitable for balanced class distributions. In contrast, for significantly imbalanced data sets (which is the main scope of this work), one should select $\alpha \neq 0$ so that the two phases ± 1 will have different energies: $V(+1) - V(-1) = 4\alpha/3 \neq 0$. In the following subsection, we discuss the influence of α in order to handle imbalanced class distributions.

2.3. Model analysis

In order to simplify the analysis, and without loss of generality, we assume that the interface around the boundary $\partial\mathcal{R}$ is very narrow compared to the interior and exterior of $\mathcal{R}$. Following this assumption, $\mathcal{X}$ is then split into two disjoint regions $\mathcal{R}$ and $\bar{\mathcal{R}}$. In addition, we define an $ansatz$ $\tilde{\phi}_R$ to compute the total energy corresponding to the region $\mathcal{R}$: $\tilde{\phi}_{\mathcal{R}}(x) = 1$ if $x \in \mathcal{R}$ and -1 if $x \in \bar{\mathcal{R}}$. We use a shifted ultralocal term, $\tilde{V}(t) = V(t) - V(-1)$, to make calculation easier. Thus, the contribution of the smoothness and the ultralocal terms in $E_P(\tilde{\phi}_{\mathcal{R}})$ becomes

$$\left(\frac{4\epsilon}{2\sum_{i,j=1}^{\ell+u} w_{ij}} \sum_{x_i \in \mathcal{R}, x_j \in \bar{\mathcal{R}}} w_{ij}\right) + \frac{4\alpha}{3\sum_{i,j=1}^{\ell+u} w_{ij}} \mathbf{vol}(\mathcal{R}), \quad (4)$$

where $\mathbf{vol}(\mathcal{R}) = \sum_{x_i \in \mathcal{R}} d_i \ll +\infty$ is the volume of $\mathcal{R}$ and $\tilde{V}(1) = V(1) - V(-1) = 4\alpha/3$. Equation above shows, in addition to the usual smoothness term, the effect of the ultralocal term. It is clear that the sign of the parameter α significantly affects the behavior of the energy minimizer: if $\alpha > 0$ then the model prefers to reduce the volume of $\mathcal{R}$; if $\alpha < 0$ then the model prefers to increase the volume of $\mathcal{R}$. This behavior corresponds to the core contribution of our work that makes it possible to classify significantly imbalanced data sets. Again, Fig. 1 shows the behavior of V for three different values of α.

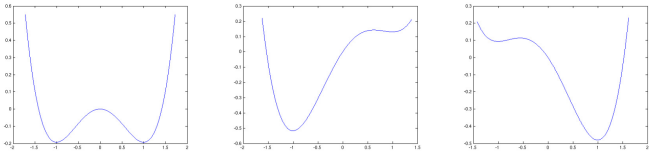

Fig. 1. Behavior of the ultralocal term V for different values of α and for fixed λ. From left to right: $\alpha = 0$, $\alpha > 0$ and $\alpha < 0$.

2.4. Our global energy minimization model

Our total energy function E is the sum of the prior energy term E_P, defined in Eq.3, and a data term E_D which links the phase field function ϕ to the labeled samples in $\mathcal{L}$, with $E_D(\phi) = \frac{1}{2\ell}(\phi - \mathbf{Y})^t U(\phi - \mathbf{Y})$ and $\mathbf{Y} = (y_1, \ldots, y_\ell, 0, \ldots, 0)^t$ is a vector of size $\ell + u$. U is a $(\ell + u) \times (\ell + u)$ diagonal matrix with entries $U_{ii} = 1$ if $i \in \{1, \ldots, \ell\}$ and 0 if $i \in \{\ell + 1, \ldots, \ell + u\}$.

To minimize the total energy $E(\phi) = E_P(\phi) + E_D(\phi)$, we differentiate E with respect to ϕ:

$$\begin{aligned} \frac{\delta E(\phi)}{\delta \phi} &= \frac{\epsilon}{\sum_{i,j=1}^{\ell+u} w_{ij}} L\phi + \frac{1}{\ell} U(\phi - \mathbf{Y}) \\ &+ \frac{1}{\sum_{i,j=1}^{\ell+u} w_{ij}} (d_1 V'(\phi_1), \ldots, d_{\ell+u} V'(\phi_{\ell+u}))^t, \end{aligned} \quad (5)$$

where $V'(t) = \lambda(t^3 - t) + \alpha(1 - t^2)$. In practice, we use a quasi-Newton minimization algorithm based on LBFGS to solve our non-linear optimization problem. This algorithm needs an initialization of the function ϕ. In all our experiments, the initial configuration of ϕ is $\phi^{(0)} = (y_1, \ldots, y_\ell, \alpha/\lambda, \ldots, \alpha/\lambda)^t$: for unknown labels, i.e. $i \in \{\ell+1, \ldots, \ell+u\}$, $\phi_i^{(0)}$ equals to the threshold α/λ located at the maximum of V (see Fig. 1). This initialization is neutral because the derivative of V evaluated at α/λ is 0 and so only the Laplacian term contributes to propagate labels at the first iteration of the algorithm.

The model E has 3 parameters ϵ, λ and α, and an application-dependent parameter σ. The latter is estimated empirically as follows: we evaluate a function $S(\sigma) = \sum_{i,j=1}^{\ell+u} w_{ij}(\sigma)$ for many values of σ, and select the value which corresponds to the inflection point of the curve $\log(S)$ against $\log(\sigma)$.

Now we define the essential steps of our SSL algorithm:

1. find σ and then construct the graph Laplacian, L, for a given data set $\mathcal{X} = \mathcal{L} \cup \mathcal{U}$ as described in Section 2.2 (we choose a neighborhood of 50 points to compute W);
2. choose $\epsilon \in [0.1, 1000]$ and $\lambda \in [0.1, 1000]$ using grid search with a logarithmic scale;
3. choose $\alpha \in]-\lambda, \lambda[$ as mentioned in Section 2.2;
4. set $\phi^{(0)}$, and then run the LBFGS minimization algorithm;
5. compute $y_i = \text{sign}(\phi_i - \alpha/\lambda)$, $\forall i \in \{\ell + 1, \ldots, \ell + u\}$.

3. EXPERIMENTS

We compare the performance of our SSL algorithm (based on the energy E) w.r.t several SSL algorithms described in [1];

	ℓ	Method	CER	ℓ	Method	CER
		MP	**11.9**		MP	**3.8**
USPS	10	IterLap	13.1	100	IterLap	4
		QC+CMN	13.6		LapRLS	4.7
		LapRLS	**54.5**		DiscreteReg	**9.6**
COIL	10	QC+CMN	59.7	100	MP	9.8
		LDS	61.9		QC+CMN	10
		RegBoost	**46.7**		RegBoost	**23.7**
BCI	10	IterLap	**46.7**	100	AdaBoost	24
		CHM	46.9		Assemble	28.7
		LDS	**27.2**		SGT	**23.1**
Text	10	SGT	29	100	LDS	23.2
		MP	29.7		RegBoost	23.3

Table 1. Classification errors, given in percentage, of the top ranked state-of-the-art algorithms on SSL benchmarks [1].

ℓ	Method	USPS	COIL	BCI	Text
	Best SoA	11.9	54.5	46.7	**27.2**
10	Our (E_0)	19.2	23.5	44.8	37.5
	Our (E)	**4.2**	**14**	**44.8**	29
	Best SoA	**3.8**	9.6	**23.7**	23.1
100	Our (E_0)	7.2	4.5	45.1	27.3
	Our (E)	**3.8**	**3**	44.2	**21.5**

Table 2. Classification errors, given in percentage, of our methods (E_0, E) and the best state-of-the-art on 4 SSL benchmarks [1] with 10 (and 100) labeled points only. "Best SoA" stands for the best state of the art method (among those reported in Table 1). Note that for USPS, COIL and Text $\ell + u = 1500$ while for BCI $\ell + u = 400$.

including boosting (Assemble [8], AdaBoost [33], ManifoldBoost [14], RegBoost [10]), measure propagation (MP) [17], Iterated Laplacian (iterLap) [22], etc. (see Tab 1) and our baseline (based on E_0, i.e. when $\alpha = 0$). We analyze the results on 4 standard benchmarks: 2 benchmarks with imbalanced class distributions, i.e., $|\mathcal{R}| \ll |\bar{\mathcal{R}}|$ ("USPS" for hand-written digit recognition and "Coil" for image classification) and 2 others with balanced class distributions, i.e., $|\mathcal{R}| \equiv |\bar{\mathcal{R}}|$ ("BCI" for brain computer interface and "Text" for newsgroup categorization).

All these benchmarks deal with binary classification excepting Coil that has 6 classes, so we carry out multiple classes using "one versus all" binary classifiers. For all these experiments, we follow up the protocols defined in [1, 10, 17]: for each dataset, we run experiments using 12 splits where each split contains 10 (or 100) randomly sampled labeled data and the remaining data are considered as unlabeled; the goal is then to predict the classes of the unlabeled data. We report performances using the CER (Classification Error Rate) averaged over the 12 splits for each dataset.

Table 1 shows CER performances which are *available and reported* in state of the art on the 4 benchmarks. These performances are shown for different sizes of labeled data ($\ell = 10$ and $\ell = 100$). The results reported in table 2, show that the performances of both models E_0 and E are comparable (and relatively close to the best state of the art methods) for balanced class distribution datasets (BCI and Text). On the imbalanced class databases (Coil and USPS), E_0 fails whereas E overtakes E_0 and shows a significant decrease of the CER compared to both E_0 and related work.

Finally, Fig. 2 shows the plots of CER against α for 3 different values of λ on the USPS benchmark (with $\ell = 100$). It is clear that the blue curve of CER (corresponding to the best choice of λ) has exactly one minimum which corresponds to the desired behavior of the model (discussed in Section 2.3) that reduces classification bias toward the dominant class (in contrast to model E_0, i.e., when $\alpha = 0$) without introducing a bias toward the rare class.

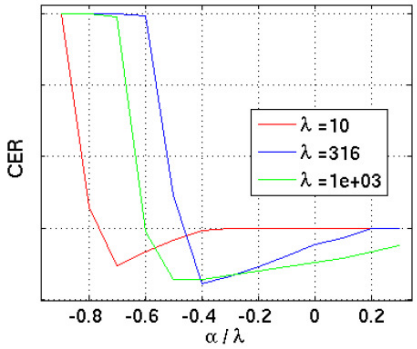

Fig. 2. This figure shows the evolution of the CER against α for 3 different values of λ on the USPS benchmark. The blue curve corresponds to the best value of λ. The best value of α corresponds to the minimum of CER. Note that these error rates are plotted against α/λ for better readability.

4. CONCLUSION

We introduced, in this paper, a graph-based phase field model for transductive semi-supervised learning. This model is able to deal with (significantly) imbalanced class distributions especially when labeled data in the class of interest are extremely scarce. Our main contribution is to tackle the problem of significantly imbalanced datasets via the incorporation of an additional prior term which encourages or penalizes the membership of data to different classes according to their class distribution.

Experiments and comparisons conducted on different standard benchmarks show that for balanced class distribution datasets, model E_0 is sufficiently performant while E is better than E_0 (as well as related SSL work) for imbalanced class distribution datasets.

Acknowledgement. This work was supported in part by grants from the Space Agency CNES (Centre National des Etudes Spatiales) under the VENISE project and from the Research Agency ANR (Agence Nationale de la Recherche) under the MLVIS project.

5. REFERENCES

[1] O. Chapelle, B. Scholkopf, and A. Zien(Eds), "Semi-supervised learning," *MIT Press, Cambridge, MA*, 2006.

[2] M. Seeger, "Learning with labeled and unlabeled data," *Tech. rep. School of informatics, Univ. of Edinburgh*, 2001.

[3] X. Zhu, "Semi-supervised learning literature survey," *Tech. rep., Department of Computer Sciences, University of Wisconsin*, 2005.

[4] V.N. Vapnik, "Statistical learning theory," *1st Edition, Wiley*, 1998.

[5] G. Haffari, "A survey on inductive semi-supervised learning," *Tech. rep., Department of Computer Sciences, Simon Fraser Univ*, 2006.

[6] M. Belkin and P. Niyogi, "Using manifold stucture for partially labeled classification," *in: Advances in Neural Information Processing Systems*, 2003.

[7] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[8] K.P. Bennett, A. Demiriz, and R. Maclin, "Exploiting unlabeled data in ensemble methods," *in: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002.

[9] A. L. Bertozzi and A. Flenner, "Diffuse interface models on graphs for classification of high dimensional data," *Multiscale Modeling and Simulation*, vol. 10, no. 3, pp. 1090–1118, 2012.

[10] K. Chen and S. Wang, "Semi-supervised learning via regularized boosting working on multiple semi-supervised assumptions," *IEEE Trans. Patt. Anal. Mach Intell*, vol. 33, no. 1, pp. 129–143, 2011.

[11] B. Geng, D. Tao, C. Xu, L. Yang, and X.-S. Hua, "Ensemble manifold regularization," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 34, no. 6, pp. 1227–1233, 2012.

[12] T. Joachims, "Transductive learning via spectral graph partitioning," *in: Proceedings of the International Conference on Machine Learning*, 2003.

[13] X. Liu, X. Yuan, S. Yan, and H. Jin, "Multi-class semi-supervised svms with positiveness exclusive regularization," *in: Proc. IEEE Int. Conf. Comp. Vis.*, 2011.

[14] N. Loeff, D. Forsyth, and D. Ramachandran, "Manifoldboost: stagewise function approximation for fully-, semi- and un-supervised learning," *in: Proceedings of the International Conference on Machine Learning*, 2008.

[15] F. Nie, H. Wang, H. Huang, and C.H.Q. Ding, "Unsupervised and semisupervised learning via l1-norm graph," *in: Proc. IEEE Int. Conf. Comp. Vis.*, 2011.

[16] H. Sahbi, P. Etyngier, J. Audibert, and R. Keriven, "Manifold learning using robust graph laplacian for interactive image retrieval," *in: Proc. IEEE Int. Conf. Comp. Vis. Patt. Rec.*, 2008.

[17] A. Subramanya and J. Bilmes, "Semi-supervised learning with measure propagation," *Journal of Machine Learning Research*, vol. 12, pp. 3311–3370, 2011.

[18] A.D. Szlam, M. Maggioni, and R.R. Coifman, "Regularization on graphs with function-adapted diffusion processes," *Journal of Machine Learning Research*, vol. 9, pp. 1711–1739, 2008.

[19] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," *in: Proceedings of the International Conference on Machine Learning*, 2003.

[20] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," *in: Proc. IEEE Int. Conf. Comp. Vis. Patt. Rec.*, 2012.

[21] S. Melacci and M. Belkin, "Laplacian support vector machines trained in the primal," *Journal of Machine Learning Research*, vol. 12, pp. 1149–1184, 2011.

[22] X. Zhou and M. Belkin, "Semi-supervised learning by higher order regularization," *Journal of Machine Learning Research - Proceedings Track 15*, pp. 892–900, 2011.

[23] H. He and E.A. Garcia, "Learning from imbalanced data," *IEEE Trans. on Knowl. and Data Eng*, vol. 21, no. 9, pp. 1263–1284, 2009.

[24] S. Li, Z. Wang, G. Zhou, and S.Y.M. Lee, "Semi-supervised learning for imbalanced sentiment classification," *in: Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, pp. 1826–1831, 2011.

[25] J.Wang, S.-F. Chang, X. Zhou, and S.-C.Wong, "Active microscopic cellular image annotation by superposable graph transduction with imbalanced labels," *in: Proc. IEEE Int. Conf. Comp. Vis. Patt. Rec.*, pp. 1–8, 2008.

[26] G. Aubert, J.F. Aujol, and L. Blanc-Feraud, "Detecting codimension two objects in an image with ginzburg-landau models," *Int. J. Comp. Vis*, vol. 65, no. 1-2, pp. 29–42, 2005.

[27] A .El-Ghoul, I.H. Jermyn, and J. Zerubia, "A phase field higher-order active contour model of directed networks," *in: 2nd IEEE Workshop NORDIA, at ICCV*, 2009.

[28] A. El-Ghoul, I.H. Jermyn, and J. Zerubia, "A theoretical and numerical study of a phase field higher-order active contour model of directed networks," *in: Proc. Asian Conf. Comp. Vis.*, 2010.

[29] H. Grossauer and O. Scherzer, "Using the complex ginzburg-landau equation for digital inpainting in 2d and 3d," *in: Proceedings of Scale-Space*, 2003.

[30] M. Rochery, I.H. Jermyn, and J. Zerubia, "Phase field models and higher-order active contours," *in: Proc. IEEE Int. Conf. Comp. Vis.*, 2005.

[31] C. Samson, L. Blanc-Feraud, G. Aubert, and J. Zerubia, "A variational model for image classification and restoration," *IEEE Trans. Patt. Anal. Mach. Intell*, vol. 22, no. 5, pp. 460–472, 2000.

[32] M. Belkin and P. Niyogi, "Towards a theoretical foundation for laplacian-based manifold methods," *Journal of Computer and System Sciences*, vol. 74, no. 8, pp. 1289–1308, 2008.

[33] Y. Freund and R.E. Schapire, "Experiments with a new boosting algorithm," *in: Proceedings of the International Conference on Machine Learning*, 1996.