Semi-Supervised Learning from a Translation Model Between Data Distributions

Henry Anaya-Sánchez, José Martínez-Sotoca, Adolfo Martínez-Usó
Institute of New Imaging Technologies, Universitat Jaume I, Spain
Department of Languages and Computer Systems, Universitat Jaume I, Spain
henry.anaya@alumail.uji.es, {sotoca,auso}@lsi.uji.es

Abstract

In this paper, we introduce a probabilistic classification model to address the task of semi-supervised learning. The major novelty of our proposal stems from measuring distributional relationships between the labeled and unlabeled data. This is achieved from a stochastic translation model between data distributions that is estimated from a mixture model. The proposed classifier is defined from the combination of both the translation model and a kernel logistic regression on labeled data. Experimental results obtained over synthetic and real-world data sets validate the usefulness of our proposal.

1 Introduction

In the last years, the task of semi-supervised classification has attracted a considerable amount of research in machine learning and pattern recognition [Singh *et al.*, 2008]. Broadly, this task consists of learning a classifier from a training set composed of both labeled and unlabeled data. The motivation of semi-supervised classification stems from the use of unlabeled data to help build a better classifier from the labeled data. This is of great interest in many real-world applications [Cherniavsky *et al.*, 2010; Bandos *et al.*, 2006], mainly in those in which the acquisition of labeled data is quite expensive and time consuming, whereas a large amount of unlabeled data is far easier to obtain.

There are actually two different semi-supervised learning settings, namely transductive and inductive semi-supervised learning [Zhu and Goldberg, 2009]. In the transdutive setting, the goal is to predict only the labels of the unlabeled data in the training set. In addition, the inductive semi-supervised learning is aimed at devising a good classifier on future data, beyond the training set.

We focus our research on the inductive setting. That is, given a training set $\mathcal{T} \subseteq \mathcal{X}$ that includes a set of labeled instances $\mathcal{L}$ ($\mathcal{L} \subseteq \mathcal{T}$), the goal is to train a classifier in order to predict the labels for the instances in $\mathcal{X}$. Specifically, in this paper we address the task in which the labels are binary by introducing a new probabilistic classification model.

The novelty of our proposal consists of measuring distributional relationships between the labeled data and the instances in $\mathcal{X}$. To this end, the approach relies on a stochastic translation model between distributions from a mixture, which is also introduced in this paper. The proposed method combines the stochastic translation model with a kernel logistic regression on labeled data to define the probabilistic classifier.

The proposal can be contextualized into the class of those stochastic (generative) semi-supervised methods that estimate conditional prediction models. Traditionally, these methods have been focused on estimating structured models such as conditional random fields [Mann and Mccallum, 2008; Dillon *et al.*, 2010], and they have been mainly applied to discrete data such as texts.

On the other hand, this work can be seen as an extension to the inductive setting of the classifier derived from the semi-supervised multi-task learning framework presented in [Liu *et al.*, 2009]. In that work, authors rely on t-step Markov transition probabilities between the training points to learn their conditional predictive model. In our case, we consider arbitrary Markov chains on mixture distributions to setup the stochastic translation model. This allows us both to directly apply our approach to the inductive setting, and to base the method on local and global knowledge from the data.

The rest of this paper is organized as follows. In Section 2, we present the proposed probabilistic model for semi-supervised classification. Section 3 describes some experiments for validating the performance on both synthetic and real-world data sets. Finally, in Section 4 we provide some conclusions and future work.

2 The Probabilistic Classification Model

In order to address our semi-supervised classification task, we rely on the latent structure of data $\mathcal{X}$ given by a mixture model:

$$p(x) = \sum_{i=1}^{m} p(x|g_i)p(g_i) \tag{1}$$

where $x \in \mathcal{X}$, $\forall i \in \{1, \ldots, m\}$ g_i represents a probability distribution, $p(g_i)$ is the prior for g_i, and $p(x|g_i)$ represents the probability of generating data point x from g_i. Such a mixture can be obtained from the training set $\mathcal{T}$ by typically applying an *Expectation-Maximization* algorithm, *Latent Dirichlet Allocation* in the case of discrete data [Blei *et al.*, 2003], or a *Dirichlet Process Mixture Model* [Rasmussen,

2000]. In our experiments, we consider both Dirichlet processes and *data spectroscopy* [Shi *et al.*, 2008] to learn this mixture.

Let $\mathcal{G}$ denotes the set of probability distributions $\{g_1, \ldots, g_m\}$ from the mixture shown in Equation 1. In this work, the distributions in $\mathcal{G}$ are aimed at measuring distributional relationships between the labeled data in $\mathcal{L}$ and the data points in $\mathcal{X}$. In this way, for all $x' \in \mathcal{L}$ and all $x \in \mathcal{X}$ we can define a posterior probability for x' given x as:

$$p(x'|x) \propto \sum_{g_i \in \mathcal{G}} \sum_{g_j \in \mathcal{G}} p(x'|g_i) t(g_i|g_j) p(g_j|x) \tag{2}$$

where $p(g_i|x) = p(x|g_i)p(g_i)/p(x)$ (from Bayes' theorem), and $t(g_i|g_j)$ represents some posterior probability for g_i given g_j.[1]

In this paper, we define $t(g_i|g_j)$ as the probability of transforming or translating g_j into g_i by regarding the overall structure of distributions from the mixture shown in Equation 1. We refer to $\{t(g_i|g_j)\}_{g_i,g_j \in \mathcal{G}}$ as the (stochastic) translation model between mixture distributions.

Accordingly, $p(x'|x)$ can be thought of as a measure of how likely x can be transformed into x'. Hence, a straightforward definition for the posterior probability of class $y \in \mathcal{Y}$ given $x \in \mathcal{X}$ can be determined by:

$$\begin{aligned} p(y|x) &\propto \sum_{x' \in \mathcal{L}} p^*(y|x') p(x'|x) \\ &= \sum_{x' \in \mathcal{L}} \sum_{g_i \in \mathcal{G}} \sum_{g_j \in \mathcal{G}} p^*(y|x') p(x'|g_i) t(g_i|g_j) p(g_j|x) \end{aligned} \tag{3}$$

where $p^*(y|x')$ is an estimation of the probability of including the labeled point x' in class y.

Thus, in our proposal each data point $x \in \mathcal{X}$ can be classified as belonging to class $y^*(x)$ according to the rule:

$$y^*(x) = \arg\max_{y \in \mathcal{Y}} p(y|x) \tag{4}$$

where $p(y|x)$ is defined as in Equation 3.

The next subsections are devoted to describe the estimation of both (i) the stochastic translation model between mixture distributions (i.e., $\{t(g_i|g_j)\}_{g_i,g_j \in \mathcal{G}}$), and (ii) the class posteriors conditioned on the labeled points (i.e., $\{p^*(y|x')\}_{y \in \mathcal{Y}, x' \in \mathcal{L}}$).

2.1 Translation Between Mixture Distributions

For estimating the translation model $\{t(g_i|g_j)\}_{g_i,g_j \in \mathcal{G}}$, we consider Markov chains between the mixture distributions.

In a generative model of Markov chains between distributions, the generation of a chain $\langle g_{i_1} g_{i_2} \ldots g_{i_k} \rangle$ starting with distribution g_{i_1} stems from the model:

$$p(\langle g_{i_1} g_{i_2} \ldots g_{i_k} \rangle) = (1-\alpha) \prod_{l=2}^{k} (\alpha p(g_{i_l}|g_{i_{l-1}})) \tag{5}$$

where α is the probability of adding a new distribution to the chain being generated, and $p(g_b|g_a)$ typically represents the probability of generating the distribution g_b immediately after g_a in a chain.

Currently, we propose two approaches for estimating the conditional probability $p(g_b|g_a)$ from the training set $\mathcal{T}$. The first one simply relies on the chain rule to define the conditional densities as follows:

$$p(g_b|g_a) \propto \sum_{x \in \mathcal{T}} p(g_b|x) p(x|g_a) \tag{6}$$

On the other hand, the second approach estimates $p(g_b|g_a)$ from a given metric h between distributions in the following manner:

$$p(g_b|g_a) \propto \exp\left(-\frac{1}{2} \left(\frac{h(g_b, g_a)}{h_0} \right)^2 \right) \tag{7}$$

where h_0 is a given distribution width. Notice that different from the first approach, in this case we compare two probability distribution without considering their context in the modeling of data $\mathcal{X}$ (i.e., their priors are disregarded).

The rationale here is to rely on the method based on the chain rule when both (a) the underlying distribution of the training set $\mathcal{T}$ approaches the true underlying distribution of actual data $\mathcal{X}$, and (b) the distributions reveal some chaining effect to conform the class structures. Otherwise, we consider the distance-based method to estimate the conditional probabilities $p(g_a|g_b)$.

Overall, $\{p(g_b|g_a)\}_{g_a,g_b \in \mathcal{G}}$ can be seen as a translation model [Berger and Lafferty, 1999] that expresses the likelihood of translating distribution g_a into g_b in one translation step. In this way, a distribution chain of length k starting at distribution g_a and ending at g_b can be seen as a translation sequence that translates g_a into g_b in k steps.

Consequently, the overall probability of translating a distribution g_j into g_i can be defined as the probability of generating an arbitrary chain starting at g_j and ending at g_i, and hence:

$$t(g_i|g_j) = \sum_{k=2}^{\infty} \left(\sum_{i_2, \ldots, i_{k-1} \in \{1, \ldots, m\}} p(\langle g_j g_{i_2} \ldots g_{i_{k-1}} g_i \rangle) \right) \tag{8}$$

This definition can be summarized in the following closed form [Lafferty and Zhai, 2001]:

$$t(g_i|g_j) = \left((1-\alpha)(I - \alpha P)^{-1} \right)_{i,j} \tag{9}$$

where I is the $m \times m$ identity matrix, and P is a $m \times m$ matrix whose element $P_{i,j}$ is defined as $p(g_i|g_j)$.

2.2 Class Posteriors Conditioned on Labeled Points

In this work, the estimation of the conditional probabilities $\{p^*(y|x')\}_{y \in \mathcal{Y}, x' \in \mathcal{L}}$ is based on *Kernel Logistic Regression* [Roth, 2001].

Typically, given a set of kernel functions $\{K_1, \ldots, K_m\}$ with domain X and a discrete random variable Y (taking values in $\{-1, 1\}$), a kernel-based logistic regression model computes the posterior probability of a value $y \in Y$ conditioned on $x \in X$ as follows:

$$p^*(y|x) = \frac{1}{1 + e^{-y(\beta_0 + \sum_{i=1}^{m} \beta_i K_i(x))}} \tag{10}$$

[1] The values $p(x|g_i)$, $p(g_i)$ and $p(x)$ are defined from Equation 1. Both $x \in \mathcal{X}$ and $x' \in \mathcal{L}$, in $p(x|g)$ and $p(x'|g)$ respectively, are drawn according to the same distribution.

where $\forall i \in \{0, \ldots, m\}$ β_i is a scalar coefficient.

For estimating the probability of including the labeled point $x' \in \mathcal{L}$ in the class $y \in \mathcal{Y}$, we follow Equation 10. However, we properly rely on the mixture distributions $g_1, \ldots, g_m$ instead of using traditional kernel functions in our estimation. Note that this is feasible since each distribution corresponds to a probability density function.

Therefore, we define $p^*(y|x')$ as the parameterized regression:

$$p^*\left(y|x', \beta = \langle \beta_0, \ldots, \beta_m \rangle\right) = \frac{1}{1 + e^{-y(\beta_0 + \sum_{i=1}^{m} \beta_i p(x'|g_i))}} \tag{11}$$

Several optimization methods can be applied to estimate vector β. In [Roth, 2001], it has been shown that kernel logistic regression can be learned in the primal using Newton's method. Thus, in this work we estimate β by considering a Newton-like method for maximizing the log-posterior:

$$\ell(\beta) = \sum_{x \in \mathcal{L}} \left(\log \sum_{x' \in \mathcal{L}} p^*(y(x')|x', \beta) p(x'|x) \right) \tag{12}$$

where $y(x')$ is the class label associated to the labeled point $x' \in \mathcal{L}$.

Starting from an initial value of β, namely $\beta(0)$, the Newton-like method iteratively approaches the value of vector β until convergence by using the following updating equation in the kth iteration:

$$\beta(k) = \beta(k-1) - \gamma H^{-1} g \tag{13}$$

where the scalar value γ is the learning step ($\gamma > 0$), and H and g represent the Hessian matrix and the gradient vector respectively of the regularized $\ell(\beta)$ at $\beta = \beta(k-1)$.

Notice that the overall formulation of our methodology can be seen as a combination of two models: (i) the model of labeled data conditioned on the domain data $\{p(x'|x)\}_{x \in \mathcal{X}, x' \in \mathcal{L}}$, and (ii) the model of class posteriors conditioned on the labeled data $\{p^*(y|x')\}_{y \in \mathcal{Y}, x' \in \mathcal{L}}$. Actually, in the learning mechanism parameterization only affects the second model (i.e. only a component of the overall model) since the estimation of the probabilities $p(x'|x)$ is performed in an unsupervised manner. Thus, the overall model can be considered as a partially parameterized one.

3 Experiments

In order to validate the proposal here presented, we consider both synthetic and real-world evaluation data sets. The performance of the classification is evaluated in terms of the *Accuracy* measure, defined as the ratio of the number of correctly classified data over the total number of data being tested.

For each data set $\mathcal{X}$, the experimentation is carried out considering different training and test sets, and also different labeled sets from each training set. Specifically, for each data set $\mathcal{X}$, and given a size l for the labeled data set, we consider 25 different triplets of training, labeled and test sets, which are generated as follows.

Firstly, we randomly sample a uniform partition $\mathcal{X} = \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_5$. Then, for each $i \in \{1, \ldots, 5\}$, we define both a training and a test set from $\mathcal{X}$ as $\mathcal{X} \setminus \mathcal{X}_i$ and $\mathcal{X}_i$ respectively. Finally, from each training set $\mathcal{X} \setminus \mathcal{X}_i$, we randomly sample 5 labeled data sets $\mathcal{L}_{i_1}, \ldots, \mathcal{L}_{i_5}$, each one of size l. We report the results for different values of l in each data set $\mathcal{X}$ by averaging the accuracy on their corresponding 25 test sets.

Our experimentation is mainly focused on the following issues:

i. Measure the impact of combining the kernel logistic regression with the proposed translation model between distributions. With this aim, we compare our model with a Kernel Logistic Classifier (KLC) [Roth, 2001].

ii. Compare our approach with existing most related work. That is, a version of the Semi-Supervised Single Task Learning (SS−STL) classifier derived from [Liu *et al.*, 2009]. This is also useful for measuring the impact of using the translation model between distributions, instead of t-step Markov transition probabilities between training points.

To ensure a fair comparison, the underlying kernels for both KLC and SS−STL were given by the mixture distributions learned from the training sets.

3.1 Synthetic Data Sets

As for synthetic data, we consider the two toy data sets represented in left column of Figure 1. These data sets are uniformly divided into 2 classes. The first data set (*moons*) consists of 600 points in the real plane, whereas the second one (*p-moons*) comprises 1000 data points.

In this case, for generating the mixture models we regard the unsupervised mixture generation given by Dirichlet processes as defined in [Rasmussen, 2000]. Thus, for each training set we regard a gaussian mixture to carry out the semi-supervised classification. The second column of Figure 1 shows one of the training samples generated from the databases together with the respective gaussian mixture distributions learned from the sample. It is worth mentioning that the mixture generation process produced some spatial confusion according to the actual class distributions in some training sets.

In tables 1 and 2, we show the results obtained over *moons* and *p-moons* data sets respectively. We consider two versions of our proposal, namely T−Chain and T−Distance, obtained from the use of the chain rule and the distance-based approaches respectively for estimating the translation model between distributions (see equations 6 and 7).

For the version T−Distance, we rely on the well-known Hellinger distance as the metric between distributions. The distribution width h_0 was conveniently defined as the third part of the average distance between each gaussian and its nearest (gaussian) neighbor in the mixture model. The parameter α in the translation model can be seen as a confidence on the chaining effect of the mixture distributions. Thus, we consider a large α for version T−Chain ($\alpha = 0.99$), and a small α for T−Distance ($\alpha = 0.10$). Since SS−STL is a transductive method, the results reported for this method on the test data correspond to the base kernel classifier embedded in its own methodology.

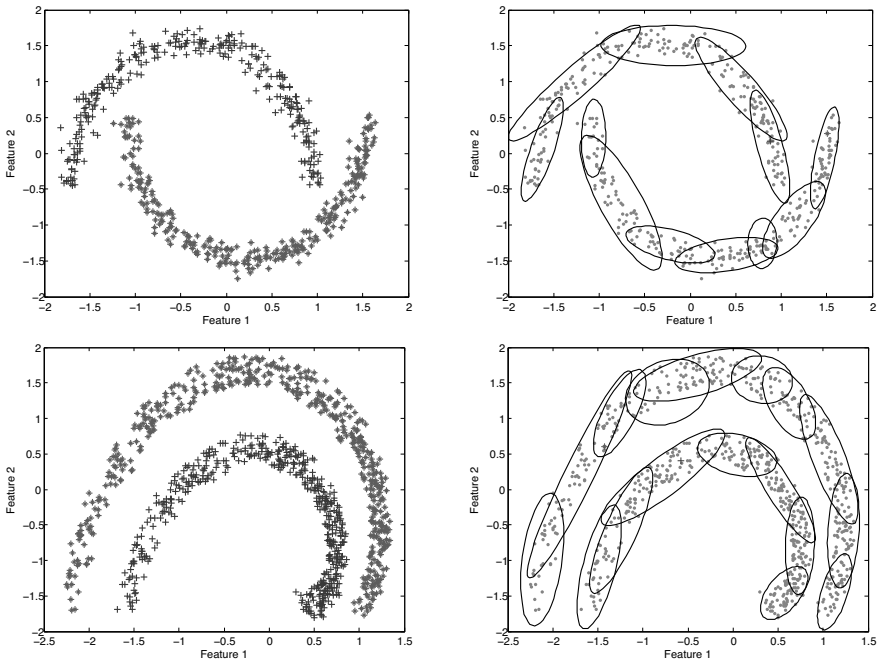

Figure 1: Synthetic data sets: *moons* (first row) and *p-moons* (second row).

As it can be seen, our two versions consistently outperforms both KLC and SS−STL. Particularly, version T−Chain performs largely better than T−Distance. This was expected on these data sets since we perceived from Figure 1 that the mixture model learned from $\mathcal{T}$ accurately approaches the actual underlying distribution of data $\mathcal{X}$, and there is a certain overlapping degree between gaussians that generates subsets of the same class with a high likelihood. It can be noticed that a distance-based criterion can produce some uncertainty on the conditional probability between gaussians in the extremes of the moons.

Currently, learning from very few labeled samples is a challenging issue in semi-supervised classification. In these data sets, it can be appreciated that T−Chain obtains a small number of misclassified data despite of the size of the labeled data set. Notice also that in the case of *p-moons* this version correctly classifies the entire dataset from l0 labeled instances. We estimate that when the number of labeled samples is large enough so that the randomly selected labeled points tend to spatially cover the class distributions, the performance of all of these methods will be similar.

3.2 Real-World Data Sets

For the case of real-world data sets, we test our approach using two binary databases from the UCI repository, namely Ionosphere and Wisconsin Breast Cancer (WBC). The Ionosphere database is a collection of radar data obtained by a system in Goose Bay, Labrador. This database comprises 351 instances described by 34 continuous attributes. On the other hand, WBC database includes 683 samples, each one consisting of 9 features taken from fine needle aspirates from a patient's breast. All 9 features were graded on an integer scale from 1 to 10.

Since these databases consists of a small number of samples according to their dimensionality, we avoid the use of Dirichlet processes to generate the mixture models (currently, they need a very large number of iterations to converge). Instead, we consider data spectroscopy [Shi *et al.*, 2008] to efficiently produce a mixture of gaussian distributions from the training sets.

Our aim was to generate a number of gaussians from each mixture so that the entropy value of the classes conditioned on each gaussian distribution was small. This caused a relative large separation between the gaussian distributions within each mixture learned, which in turn impacts negatively on the chaining effect between the distributions. In this way, the results obtained by T−Chain do not significantly improve those obtained by KLC and SS−STL. Note that this problem is overcome by T−Distance.

Figures 2 and 3 show the results obtained over Ionosphere and WBC respectively. In these figures we show in the left and right columns the average accuracy values obtained over the unlabeled samples in the training sets and the test samples respectively.

Several observations can be made by analyzing these figures. Firstly, it can be seen that T−Distance has a good performance on unseen data (i.e., on the test data sets). This corroborates the usefulness of our proposal for the inductive semi-supervised task, which is the main purpose of this work. Indeed, T−Distance significantly outperforms both KLC and SS−STL on test data.

Secondly, it can be appreciated that despite SS−STL performs the best on the unlabeled data from WBC, this method is not appropriate for inductive learning. It suffer from some

Table 1: Performance on the unlabeled samples (U) and test data (T) of *moons*.

	l=2		*l*=5		*l*=10		*l*=20	
	U	T	U	T	U	T	U	T
KLC	0.6649	0.6727	0.7865	0.7990	0.9151	0.9107	0.9644	0.9587
SS-STL	0.6829	0.6843	0.7929	0.8013	0.9140	0.8987	0.9722	0.9637
T-Distance	0.7163	0.7193	0.8648	0.8707	0.9343	0.9363	0.9883	0.9897
T-Chain	0.9607	0.9610	0.9933	0.9937	0.9813	0.9810	0.9979	0.9980

Table 2: Performance on the unlabeled samples (U) and test data (T) of *p-moons*.

	l=2		*l*=5		*l*=10		*l*=20	
	U	T	U	T	U	T	U	T
KLC	0.6578	0.6555	0.8034	0.8070	0.8596	0.8549	0.9521	0.9584
SS-STL	0.6618	0.6591	0.7942	0.7956	0.8753	0.8708	0.9608	0.9624
T-Distance	0.6996	0.7081	0.8354	0.8380	0.9081	0.9081	0.9873	0.9902
T-Chain	0.9894	0.9897	0.9887	0.9922	1.0000	1.0000	1.0000	1.0000

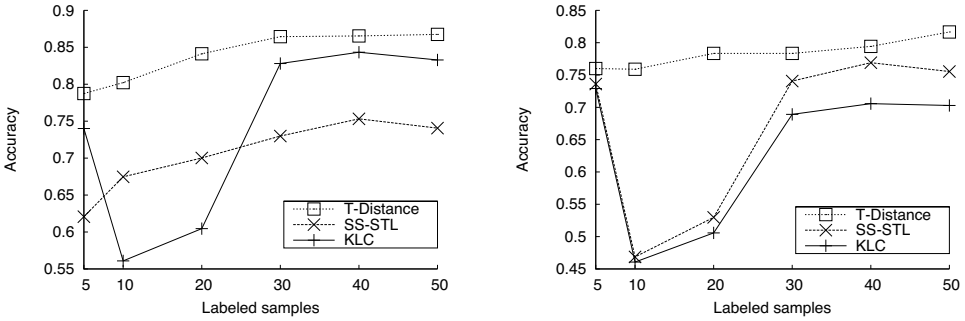

Figure 2: Performance on the unlabeled samples (left) and test data (right) from Ionosphere.

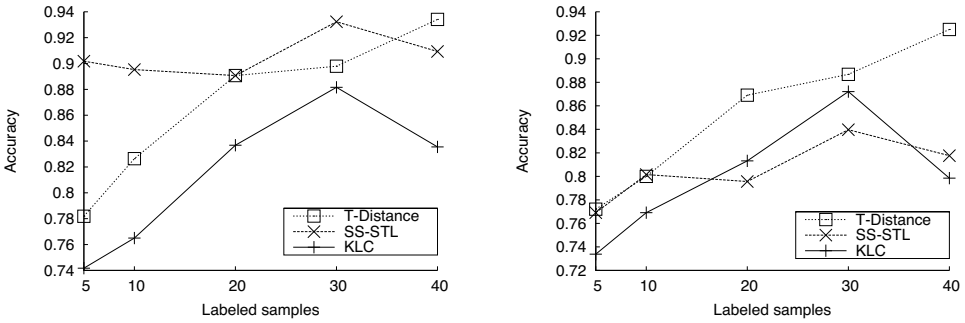

Figure 3: Performance on the unlabeled samples (left) and test data (right) from WBC.

overfitting, which can be explained from both (i) the unstable behavior of it accuracy curve on unseen data and (ii) the similarity of this curve with that of KLC (i.e., SS−STL performs similar to a pure supervised classifier). On the other hand, it can be clearly observed that our approach obtains stable and similar results on both transductive and inductive semi-supervised learning.

The performance improvement over KLC validates the combination of logistic regression with the proposed translation model, whereas the improvement over SS−STL measures the positive impact of translating from Markov chains on mixture distributions instead of using t-step Markov transition probabilities between the training points for inductive learning.

4 Conclusions

In this paper, a new semi-supervised probabilistic classification model has been introduced. The proposal exploits the local and global context of the data by measuring distributional relationships between the labeled and unlabeled data. This is carried out from a stochastic translation model between data distributions, which allows our method to be directly applied to unseen data within the semi-supervised learning task.

The proposed learning mechanism combines both the stochastic translation model and a kernel logistic regression on labeled data. Experimentally, we tested two approaches derived from the methodology on synthetic and real-world data sets. The obtained results corroborates the usefulness of our proposal for the semi-supervised classification task.

Future works include extending this methodology to the multiclass setting. This can be easily achieved since the learning mechanism relies on logistic regression, which is suitable for the multiclass learning problem.

Acknowledgments

This work has been supported by the Spanish Ministry of Science and Innovation under projects Consolider Ingenio 2010 CSD2007-00018 and TIN2008-01825/TIN, and by the project P1-1B2009-45 of Fundació Caixa-Castelló.

References

[Bandos *et al.*, 2006] T Bandos, D Zhou, and G Camps-Valls. Semi-supervised hyperspectral image classification with graphs. In *IEEE International Conference on Geoscience and Remote Sensing Symposium*, pages 3883–3886, 2006.

[Berger and Lafferty, 1999] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229. ACM, 1999.

[Blei *et al.*, 2003] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[Cherniavsky *et al.*, 2010] N. Cherniavsky, I Laptev, J Sivic, and A. Zisserman. Semi-supervised learning of facial attributes in video. In *First International Workshop on Parts and Attributes, in conjunction with European Conference on Computer Vision*, 2010.

[Dillon *et al.*, 2010] J.V. Dillon, K. Balasubramanian, G. Lebanon, Y. Chen, J.W. Vaughan, R. Srivastava, C.E. Koksal, S. Liu, C. Ling, D. Stehlé, et al. Asymptotic analysis of generative semi-supervised learning. In *Proc. of the International Conference on Machine Learning*, 2010.

[Lafferty and Zhai, 2001] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 111–119, 2001.

[Liu *et al.*, 2009] Q. Liu, X. Liao, H. Li, J.R. Stack, and L. Carin. Semisupervised multitask learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):1074–1086, 2009.

[Mann and Mccallum, 2008] G.S. Mann and A. Mccallum. Generalized expectation criteria for semi-supervised learning of conditional random fields. In *Proc. ACL, pages 870–878*, 2008.

[Rasmussen, 2000] C.E. Rasmussen. The infinite Gaussian mixture model. *Advances in neural information processing systems*, 12:554–560, 2000.

[Roth, 2001] V. Roth. Probabilistic discriminative kernel classifiers for multi-class problems. In *Pattern Recognition: 23rd DAGM Symposium, Munich, Germany, September 12-14, 2001. Proceedings*, page 246. Springer, 2001.

[Shi *et al.*, 2008] T. Shi, M. Belkin, and B. Yu. Data spectroscopy: Learning mixture models using eigenspaces of convolution operators. In *Proceedings of the 25th International Conference on Machine Learning*, pages 936–943. ACM, 2008.

[Singh *et al.*, 2008] A. Singh, R. D. Nowak, and X Xhu. Unlabeled data: Nowit helps, now it doesn't. In *Advances in Neural Information Processing Systems. Proceedings*, pages 1513–1520, 2008.

[Zhu and Goldberg, 2009] X. Zhu and A.B. Goldberg. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1):1–130, 2009.