Computación y Sistemas
ISSN: 1405-5546
computacion-y-sistemas@cic.ipn.mx
Instituto Politécnico Nacional
México

Cruz-Barbosa, Raul; Vellido, Alfredo
Generative Manifold Learning for the Exploration of Partially Labeled Data
Computación y Sistemas, vol. 17, núm. 4, octubre-diciembre, 2013, pp. 641-653
Instituto Politécnico Nacional
Distrito Federal, México

Generative Manifold Learning for the Exploration of Partially Labeled Data

Raúl Cruz-Barbosa[1] and Alfredo Vellido[2]

[1]Instituto de Computación, Universidad Tecnológica de la Mixteca, Huajuapan, Oaxaca, Mexico

[2]Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, Barcelona, Spain

rcruz@mixteco.utm.mx, avellido@lsi.upc.edu

Abstract. In many real-world application problems, the availability of data labels for supervised learning is rather limited and incompletely labeled datasets are commonplace in some of the currently most active areas of research. A manifold learning model, namely Generative Topographic Mapping (GTM), is the basis of the methods developed in the thesis reported in this paper. A variant of GTM that uses a graph approximation to the geodesic metric is first defined. This model is capable of representing data of convoluted geometries. The standard GTM is here modified to prioritize neighbourhood relationships along the generated manifold. This is accomplished by penalizing the possible divergences between the Euclidean distances from the data points to the model prototypes and the corresponding geodesic distances along the manifold. The resulting Geodesic GTM (Geo-GTM) model is shown to improve the continuity and trustworthiness of the representation generated by the model, as well as to behave robustly in the presence of noise. We then proceed to define a novel semi-supervised model, SS-Geo-GTM, that extends Geo-GTM to deal with semi-supervised problems. In SS-Geo-GTM, the model prototypes obtained from Geo-GTM are linked by the nearest neighbour to the data manifold. The resulting proximity graph is used as the basis for a class label propagation algorithm. The performance of SS-Geo-GTM is experimentally assessed via accuracy and Matthews correlation coefficient, comparing positively with an Euclidean distance-based counterpart and the alternative Laplacian Eigenmaps and semi-supervised Gaussian mixture models.

Keywords. Semi-supervised learning, Clustering, Generative Topographic Mapping, Exploratory Data Analysis.

Aprendizaje generativo de variedades para la exploración de datos parcialmente etiquetados

Resumen. En muchos problemas aplicados del mundo real, la disponibilidad de etiquetas de los datos para el aprendizaje supervisado es bastante limitada y los conjuntos de datos etiquetados incompletamente son habituales en algunas de las áreas de investigación actualmente más activas . Un modelo de aprendizaje de variedades, el Mapeo Topográfico Generativo (GTM como acrónimo del nombre en inglés), es la base de los métodos desarrollados en la tesis reportada en este artículo. Se define en primer lugar una extensión de GTM que utiliza una aproximación de grafos para la métrica geodésica. Este modelo es capaz de representar datos de geometría intrincada. El GTM estándar se modifica aquí para priorizar relaciones de vecindad a lo largo de la variedad generada. Esto se logra penalizando las divergencias posibles entre las distancias euclideanas de los puntos de datos a los prototipos del modelo y las distancias geodésicas correspondientes a lo largo de la variedad. Se muestra aquí que el modelo GTM geodésico (Geo-GTM) resultante mejora la continuidad y la fiabilidad de la representación generada por el modelo, al igual que se comporta robustamente en presencia de ruido. Después, procedemos a definir un modelo semi-supervisado novedoso, SS-Geo-GTM, que extiende Geo-GTM para tratar problemas semi-supervisados. En SS-Geo-GTM, los prototipos del modelo obtenidos de Geo-GTM son vinculados mediante el vecino más cercano a la variedad de datos. El grafo de proximidad resultante se utiliza como la base para un algoritmo de propagación de etiquetas de clase. El rendimiento de SS-Geo-GTM se evalua experimentalmente a través de las medidas de exactitud y el coeficiente de correlación de Matthews, comparando positivamente con una contraparte basada

Computación y Sistemas Vol. 17 No. 4, 2013 pp. 641-653
ISSN 1405-5546
DOI: 10.13053/CyS-17-4-2013-014

en la distancia euclideana y con los modelos alternativos de Eigenmapas Laplacianos y mezclas de Gaussianas semi-supervisadas.

Palabras clave. Aprendizaje semi-supervisado, agrupamiento, mapeo topográfico generativo, análisis exploratorio de datos.

1 Introduction

Labeling aspects of reality is one of the most standard tasks performed by the human brain and, therefore, of natural learning. When dividing the existing reality into different categories, humans seamlessly perform a classification task that can be improved over time through learning.

In the realm of non-natural, or machine learning, the task of unraveling the relationship between the observed data and their corresponding class labels can be seen as the modeling of the mapping between a set of data inputs and a set of discrete data targets. This is understood as supervised learning.

Unfortunately, in many real applications class labels are either completely or partially unavailable. The first case scenario is that of unsupervised learning, where the most common task to be performed is that of data clustering, which aims to discover the "true" group structure of multivariate data [32]. The second case is less frequently considered but far more common than what one might expect: quite often, only a reduced number of class labels is readily available and even that can be difficult and/or expensive to obtain.

In such context, unsupervised models are an adequate tool for a first exploratory approach. The available class labels can then be used to refine the unsupervised procedure. This becomes a task on the interface between supervised and unsupervised models: semi-supervised learning (SSL [17, 59, 7]). This type of learning is commonly understood as a way to improve supervised tasks (usually with few available labeled samples) with the use of unlabeled samples [47, 15, 33, 26, 44]. In this paper, the approach is a less typical one: improving and refining unsupervised learning by using class labeled data.

From several categories of SSL methods [17], we are specifically interested in graph-based methods based on generative models. In graph-based methods, the nodes of a graph come to represent the observed data points, while its edges are assigned the pairwise distances between the incident nodes.

The way the distance between two data points is computed can be seen as an approximation of the (true) geodesic distance (which is computationally intractable [11, 38]) between the two points with respect to the overall data manifold [10]. Therefore, we present a semi-supervised approach, inspired by that two-stage SSL method proposed in [31]. It is based on geodesic generative topograhic mapping (Geo-GTM: [20]), which is an extension of the statistically principled Generative Topographic Mapping (GTM: [14]).

In our proposal, the prototypes obtained from Geo-GTM are inserted and linked by the nearest neighbour to the data manifold. The resulting graph is considered as a proximity graph for which an ad hoc version of label propagation algorithm (LP: [58]) is defined. This becomes semi-supervised Geo-GTM (SS-Geo-GTM: [21, 22]), a model that uses the information derived from Geo-GTM training to accomplish the semi-supervised task (see Fig. 1). A detailed description of SS-Geo-GTM can be found in [21], whereas a practical application to a problem in the field of neuro-oncology, using human brain tumour datasets, is described in [22].

Unlike the aforementioned SS-Geo-GTM publications, here we compare its performance, via accuracy and Matthews correlation coefficient [40, 28], with that of two alternative semi-supervised techniques: Laplacian Eigenmaps [8] and semi-supervised Gaussian mixture models [41].

2 Semi-Supervised and Generative Manifold Learning

Semi-supervised learning is an emergent discipline that incorporates prior knowledge into supervised or unsupervised methods (classification and clustering, mainly). The need for SSL, understood as learning from a combination of both labeled and unlabeled data, rises naturally in cases for which there exists a large supply of unlabeled data but a limited one of labeled data (bearing in mind that in many practical domains it can be very difficult and/or expensive to generate the labeled data). When SSL is used for classification, the main goal is to improve the classification accuracy aided by unlabeled data.

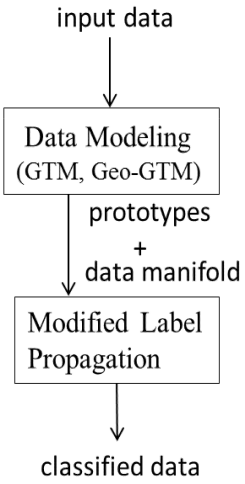

Fig. 1. Schematic GTM-based semi-supervised procedure

SSL for classification has become popular over the past few years. Some of the proposed methods include: co-training [15], in which there are two kinds (views) of information for training – about examples and the availability of both labeled and unlabeled data (some extensions of co-training and its applications can be found in [43, 27, 19, 55, 56, 34]); Transductive Support Vector Machines (TSVM, [33]), in which transduction follows Vapnik's principle: when trying to solve some problems, one should not solve a more difficult problem as an intermediate step (some extensions of TSVM or Semi-supervised SVMs and their applications are reported in [18, 23, 52, 53, 16, 50, 51, 45]); and Expectation-Maximization (EM), within the Maximum Likelihood framework, to incorporate unlabeled data into the training processes [26, 44, 3, 42].

On the other hand, semi-supervised clustering (SSC) uses class labels or pairwise constraints (specifying wether two instances should be in same or different clusters) on some examples to aid unsupervised clustering [5, 13, 6, 37, 4, 48, 29, 54, 2]. SSC is useful when knowledge of the relevant categories of a problem is incomplete. When it happens, SSC can group data using the categories in the initial labeled data as well as extend and modify the existing set of categories as needed to reflect other regularities in the data.

At the present time, there is a tendency to consider as "standard" SSL methods [17] only those which use it for classification tasks (as it is defined in [47]). However, SSC should be considered a more general SSL setting when the number and nature of the classes are not known in advance but have to be inferred from the data. For a more complete literature survey of SSL for classification and clustering tasks, readers can consult [17, 59, 7, 57] and the references therein.

As mentioned in the previous section, the approach in this work is a less typical one: improving and refining unsupervised learning by using class labeled data. Most used unsupervised learning tasks are: dimensionality reduction and clustering analysis. The non-linear dimensionality reduction problem of manifold learning can be expressed as the recovery of meaningful low-dimensional structures hidden in high-dimensional data [46, 49, 38]. This recovery should allow us to extract useful information and discover meaningful features, patterns and rules from data.

When the manifold assumption is taken up for clustering analysis, one important question is how to incorporate intrinsic geometric information of multivariate data in the corresponding clustering method. Identifying the underlying manifolds defining the data is of critical importance for their understanding. Methods such as ISOMAP [49] and Curvilinear Distance Analysis [39], for instance, use the geodesic distance as a basis for generating the data manifold. ISOMAP, in fact, can be seen as an instance of Multi-Dimensional Scaling (MDS) in which the Euclidean distance is replaced by the geodesic one. This metric measures similarity along the embedded manifold, instead of doing it through the embedding space. In doing so, it may help to avoid some of the distortions (such as breaches of topology preservation) that the use of a standard metric such as the Euclidean distance may introduce when learning the manifold, due to its excessive folding (that is, undesired manifold curvature effects). The otherwise computationally intractable geodesic metric can be approximated by graph distances [11], so that instead of finding the minimum arc-length between two data items on a manifold, we find the length of the shortest path between them, where such path is built by connecting the closest successive data items. Here, this is accomplished using the K-rule (for other alternative approaches see [38]). A weighted graph is then constructed by using the data (vertices) and the set of allowed connections

(edges). If the resulting graph is disconnected, some edges are added using a minimum spanning tree procedure in order to connect it. Finally, the distance matrix of the weighted undirected graph is obtained by repeatedly applying Dijkstra's algorithm [25], which computes the shortest path between all data items.

2.1 Standard and Geodesic Generative Topographic Mapping

The standard GTM is a generative non-linear latent variable model that, in its original definition, was intended for modelling continuous, intrinsically low-dimensional data distributions, embedded in high-dimensional spaces. It can also be understood both as a sound probabilistic alternative to the well-known and widely used Self-Organizing Maps (SOM: [36]) and as a constrained mixture of distributions model. Its constraints make it less flexible than general mixtures of distributions, but such renounce to flexibility is compensated by computational expediency and by data visualization capabilities akin to those of the SOM, which general mixture models lack. Like SOM, GTM is used for unsupervised clustering and visualization.

The GTM is a model of the manifold learning family defined as a mapping from a low dimensional latent space onto the multivariate space where observed data reside. The mapping is carried through by a number of basis functions generating a constrained mixture density distribution. It is defined as a generalized linear regression model:

$$\mathbf{y}(\mathbf{u};\mathbf{w}) = \phi(\mathbf{u})\mathbf{W} \tag{1}$$

where ϕ are M basis functions $\phi(\mathbf{u}) = (\phi_1(\mathbf{u}), ..., \phi_{\mathbf{M}}(\mathbf{u}))$. For continuous data of dimension D, spherically symmetric Gaussians

$$\phi_m(\mathbf{u}) = \exp\left\{-1/2\sigma^2\|\mathbf{u} - \mu_m\|^2\right\} \tag{2}$$

are an obvious choice of basis function, with centres μ_m and common width σ; $\mathbf{W}$ is a $M \times D$ matrix of adaptive weights w_{md} that defines the mapping, and **u** is a point in latent space. To avoid computational intractability a regular grid of K points $\mathbf{u}_k$ can be sampled from the latent space. Each of them, which can be considered as the representative of a data cluster, has a fixed prior probability $p(\mathbf{u}_k) = 1/K$ and is mapped, using Eq. 1, into a low dimensional manifold non-linearly embedded in the data space. This latent space grid is similar in design and purpose to that of the visualization space of the SOM. A probability distribution for the multivariate data $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$ can then be defined, leading to the following expression for the log-likelihood:

$$L(\mathbf{W},\beta|\mathbf{X}) = \sum_{n=1}^{N} \ln\left\{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\beta/2\|\mathbf{y}_k - \mathbf{x}_n\|^2\right\}\right\}, \tag{3}$$

where $\mathbf{y}_k$, usually known as *reference* or *prototype vectors*, are obtained for each $\mathbf{u}_k$ using Eq. 1; and β is the inverse of the noise variance, which accounts for the fact that data points might not strictly lie on the low dimensional embedded manifold generated by the GTM.

The Expectation-Maximization (EM) algorithm [24] is an straightforward alternative to obtain the maximum likelihood estimates of the adaptive parameters of the model, which are the adaptive matrix of weights $\mathbf{W}$ and β. In the E-step of the EM algorithm, the mapping is inverted and the responsibilities z_{kn} (the posterior probability of cluster *k* membership for each data point $\mathbf{x}_n$) can be directly computed as

$$\begin{aligned} z_{kn} =& p(\mathbf{u}_k|\mathbf{x}_n,\mathbf{W},\beta) \\ =& \frac{p(\mathbf{x}_n|\mathbf{u}_k,\mathbf{W},\beta)p(\mathbf{u}_k)}{\sum_{k'} p(\mathbf{x}_n|\mathbf{u}_{k'},\mathbf{W},\beta)p(\mathbf{u}_{k'})}, \end{aligned} \tag{4}$$

where $p(\mathbf{x}_n|\mathbf{u}_k,\mathbf{W},\beta) = \mathcal{N}(\mathbf{y}(\mathbf{u}_k,\mathbf{W}),\beta)$.

Standard GTM is optimized by minimization of an error that is a function of Euclidean distances, making it vulnerable to *continuity* and *trustworthiness* problems, especially for datasets of convoluted geometry. Such data may require plenty of folding from the GTM model, resulting in an unduly entangled embedded manifold that would hamper both the visualization of the data and the definition of clusters the model is meant to provide. Following an idea proposed in [1], the learning procedure of GTM is here modified by penalizing the divergences between the Euclidean distances from the data points to the model prototypes and the corresponding approximated geodesic distances along the manifold. By doing so, we prioritize neighbourhood relationships

between points along the generated manifold, which makes the model more robust to the presence of off-manifold noise.

The Geo-GTM model is an extension of GTM that favours the similarity of points along the learned manifold, while penalizing the similarity of points that are not contiguous in the manifold, even if close in terms of the Euclidean distance. This is achieved by modifying the standard calculation of the responsibilities in Eq. 4 proportionally to the discrepancy between the geodesic (approximated by the graph) and the Euclidean distances. Such discrepancy is made operational through the definition of the exponential distribution, as in [1]:

$$\mathcal{E}(d_g|d_e,\alpha) = \frac{1}{\alpha}\exp\left\{-\frac{d_g(\mathbf{x}_n,\mathbf{y}_m) - d_e(\mathbf{x}_n,\mathbf{y}_m)}{\alpha}\right\}, \tag{5}$$

where $d_e(\mathbf{x}_n,\mathbf{y}_m)$ and $d_g(\mathbf{x}_n,\mathbf{y}_m)$ are, in turn, the Euclidean and graph distances between data point $\mathbf{x}_n$ and the GTM prototype $\mathbf{y}_m$. Responsibilities are redefined as:

$$\begin{aligned} z_{mn}^{geo} &= p(\mathbf{u}_m|\mathbf{x}_n,\mathbf{W},\beta) \\ &= \frac{p'(\mathbf{x}_n|\mathbf{u}_m,\mathbf{W},\beta)p(\mathbf{u}_m)}{\sum_{m'} p'(\mathbf{x}_n|\mathbf{u}_{m'},\mathbf{W},\beta)p(\mathbf{u}_{m'})}, \end{aligned} \tag{6}$$

where

$$\begin{aligned} &p'(\mathbf{x}_n|\mathbf{u}_m,\mathbf{W},\beta) \\ &= \mathcal{N}(\mathbf{y}(\mathbf{u}_m,\mathbf{W}),\beta)\mathcal{E}(d_g(\mathbf{x}_n,\mathbf{y}_m)^2|d_e(\mathbf{x}_n,\mathbf{y}_m)^2,1). \end{aligned} \tag{7}$$

As for standard GTM, Geo-GTM provides data visualization capabilities that the alternative Manifold Finite Gaussian Mixtures model proposed in [1] lacks.

3 Semi-Supervised Geodesic Generative Topographic Mapping

In many of the databases generated in some of the currently most active areas of research, such as, for instance, biomedicine, bioinformatics, or web mining, class labels are either completely or partially unavailable. As was stated in section 2, SSL methods can be developed to assist either classification or clustering tasks mainly. The former task is the purpose of the models described in this section, but using a clustering method as a basis. That is, this section specifically concerns graph-based methods that use, as a basis, generative unsupervised models for clustering and visualization.

3.1 SS-Geo-GTM

The basic idea underlying the proposed semi-supervised approach is that neighbouring points are most likely to share their label and that these labels are best propagated through neighbouring nodes according to proximity. Assuming that the Geo-GTM prototypes and the corresponding constructed data manifold can be seen as a proximity graph, we modify an existing label propagation algorithm [58] to account for the information provided by the trained Geo-GTM. The result is the proposed semi-supervised Geo-GTM (SS-Geo-GTM, for short). An schematic procedure of SS-Geo-GTM is shown in Fig. 1.

The LP method is adapted to Geo-GTM as follows. A label vector $\mathbf{L}_m \in [0,1]^k$ is first associated to each Geo-GTM prototype $\mathbf{y}_m$. These label vectors can be considered as nodes in a proximity graph. The weights of the edges are derived from the graph distances d_g between prototypes. For this, the prototypes are inserted and linked to the graph through the nearest data point. It is important to note that, in this process, empty clusters (that is, those associated to a given prototype $\mathbf{y}_m$, to which no data point is assigned, or, in other words, those that do not bear a maximum of responsibility z_{mn}^{geo} for any data point n) are omitted. The edge weight between nodes m and m' is calculated as

$$w_{mm'} = \exp(-\frac{d_g^2(m,m')}{\sigma^2}), \tag{8}$$

where the σ parameter defines the level of sparseness in the graph for label information. One possible choice for the value of this parameter is the minimal inter-prototype distance. An alternative choice is introduced and defined in [21] as the main reference inter-prototype (MRIP) distance, which is the graph distance $d_g(\mathbf{y}_{m1},\mathbf{y}_{m2})$ between the two non-contiguous prototypes $\mathbf{y}_{m1},\mathbf{y}_{m2}$ of highest Cumulative Responsibility (i.e., the sum of responsibilities over all data items in X, for each cluster m, $CR_m = \sum_{n=1}^{N} z_{mn}^{geo}$). The prototypes with highest CR are considered as the most representative in the dataset. As illustration, the process of computing the graph distances between prototypes is shown in Fig. 2 (right), using the *Dalí* set described in the results section.

Following [31], the available label information of $\mathbf{x}_n \in X$ with class attribution $c(\mathbf{x}_n) = C_t \in \{C_1,\ldots,C_k\}$ will be used to fix the label vectors

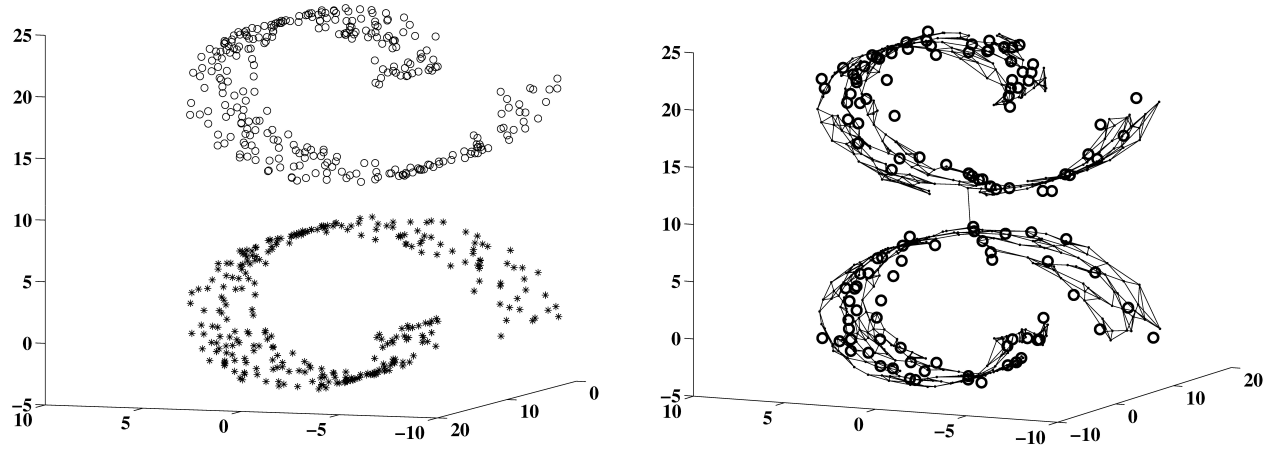

Fig. 2. (Left): The artificial 3-D *Dalí* dataset, where the two contiguous fragments are assumed to correspond to different classes, identified with different symbols. (Right): Results of the Geo-GTM modeling of *Dalí*. The prototypes are represented by '$\circ$' symbols. The graph constructed using 4-nearest neighbours is represented by lines connecting the data points, which are, in turn, represented by '$\cdot$' symbols. Figure taken from [21]

of the prototypes to which they are assigned ($\mathbf{x}_n$ is assigned to $\mathbf{y}_m$ through $\mathbf{u}_m = \arg\max_{\mathbf{u}_i} z_{in}^{geo}$), so that $L_{m,j} = 1$ if $j = t$, and $L_{m,j} = 0$ otherwise. Unlabeled prototypes will then update their label by propagation according to

$$\mathbf{L}_m^{new} = \frac{\sum_{m'} w_{mm'} \mathbf{L}_{m'}}{\sum_{m'} w_{mm'}}, \tag{9}$$

until no further changes occur in the label updating. Subsequently, unlabeled data items are labeled by assignment to the class more represented on the label vector of the prototype $\mathbf{y}_m$ bearing the highest responsibility for them, according to $c(\mathbf{x}_n) = \arg\max_{C_j \in \{C_1,\ldots,C_k\}} L_{m,j}$. The same methodology is used to build a semi-supervised version of a standard GTM model (SS-GTM).

A summary of a GTM-based semi-supervised algorithm is presented in Fig. 3, where the first five steps correspond to pre-processing addressed to prepare suitable data structures which will be used in the remainder steps.

3.2 Results

Geo-GTM, GTM, SS-Geo-GTM, and SS-GTM were initialized following a procedure described in [14]. All of these models were implemented in MATLAB®. The latent grid for GTM and Geo-GTM was fixed to a square layout of approximately $(N/2)^{1/2} \times (N/2)^{1/2}$, where N is the number of points in the data set.

Five datasets were selected for the reported experiments, where three of them (*Dalí*, *Oil-Flow* and *Digit1*) can be represented by low-dimensional manifolds, that is, the manifold assumption is hold. For the other sets, *Iris* and *g241c*, it is known in advance [17] that this assumption is not hold. The characteristics of these datasets are as follows:

— The first one is the artificial 3-D *Dalí* set. It consists of two groups of 300 data points each that are images of the functions $\mathbf{x}_1 = (t\cos(t), t_2, t\sin(t))$ and $\mathbf{x}_2 = (t\cos(t), t_2, -t\sin(t) + 20)$, where t and t_2 follow $\mathcal{U}(\pi, 3\pi)$ and $\mathcal{U}(0, 10)$, respectively.

— The second is the more complex *Oil-Flow* set, available online[1], which simulates measurements in an oil pipe corresponding to three possible configurations (classes). It consists of 1,000 items described by 12 attributes.

— The third one is the *Digit1* set, which consists of artificial writings (images) of the digit "1" initially developed in [30]. The original images are transformed in such way that

[1] http://research.microsoft.com/∼cmbishop/PRML/webdatasets/datasets.htm

1. Create a connected graph by inserting and linking the M prototypes provided by GTM models to the nearest neighbour of the data manifold. Here, the nodes are all prototypes.
2. Compute the (graph) distance among prototypes using the constructed graph in step 1.
3. Compute the edge weights, w_{ij}, between nodes i,j using Eq. 8.
4. Define a $(l+u)\times C$ label matrix L, whose i^{th} row represents the label probability distribution of data point x_i.
5. Define a $M\times C$ prototypes label matrix L', whose i^{th} row represents the label probability distribution of node (prototype) y_i. Here, the available label information of $x_n \epsilon X$ (given by L) with class attribution $c(x_n) = C_t$ is used to fix the label vectors of the prototypes to which they are assigned, so that $L'_{m,j} = 1$ if $j = t$, and $L'_{m,j} = 0$ otherwise. The initialization of unlabeled nodes is not relevant.
6. Propagate L' as in Eq. 9.
7. Row-normalize L' as $L'_{ij} = L'_{ij} / \Sigma_k L'_{ik}$.
8. Clamp the labeled data. Repeat from step 6 until L' converges.
9. Unlabeled data points in L are labeled as $c(x_n) = \arg\max_{cj \epsilon \{c1,...,ck\}} L'_{m,j}$.

Fig. 3. GTM-based semi-supervised algorithm summary

they are rescaled, noise is added, and some dimensions are masked [17]. The final data consists of 1500 241-dimensional samples, which correspond to two classes.

— The fourth set is the well-known *Iris* data, available from the UCI repository, which consists of 150 4-dimensional items representing several measurements of Iris flowers, which belong to 3 different classes.

— The fifth set is *g241c*, where 241-dimensional 750 points were drawn from each of two unit-variance isotropic Gaussians, the centers of which had a distance of 2.5 in a random direction. The class label of a point represents the Gaussian it was drawn from. *Digit1* and *g241c* are also available online[2].

The central goal of the experiments is the comparison of the performances of SS-Geo-GTM, SS-GTM, Laplacian Eigenmaps (LapEM [9]) and semi-supervised Gaussian mixture Model (SS-GMM [41]) in terms of classification accuracy and Matthews correlation coefficient (MCC) [40] for datasets where the manifold assumption is or not hold.

[2]http://olivier.chapelle.cc/ssl-book/benchmarks.html

We hypothesize that SS-GTM will yield lower rates of classification accuracy in the semi-supervised task than its geodesic distance-based counterpart, especially for datasets of convoluted geometry such as *Dalí* and *Oil-Flow*. LapEM was implemented in MATLAB®and for SS-GMM a recent open source implementation for R environment [12] was used.

Through several experiments shown in [21] is concluded that the choice of the MRIP as a value for σ is appropriate. Class labels were available for all data points in the original five datasets. In order to evaluate the models in a semi-supervised setting, labels were therefore randomly removed (thus becoming missing values) in every run of the experiments. In this setting, we evaluate the models in the most extreme semi-supervised setting, that is, when the class label is only available for a single input sample for each class and the remaining samples are considered as unlabeled data. In the next step, the label availability condition is relaxed, and the models are evaluated in the presence of higher ratios of labels as well as in the presence of noise.

All datasets are first modeled using GTM and Geo-GTM. SS-GTM and SS-Geo-GTM are then built on top of these. The semi-supervised

Table 1. Classification accuracy (mean ± std) and MCC results for a one label/class setting

Accuracy results					
Method	*Dalí*	*Oil-Flow*	*Digit1*	*Iris*	*g241c*
SS-Geo-GTM	99.52 ± 2.38	77.65 ± 7.61	77.99 ± 10.48	88.79 ± 7.86	51.50 ± 4.20
SS-GTM	90.64 ± 7.80	36.75 ± 2.78	52.28 ± 4.55	85.78 ± 8.10	57.49 ± 9.23
LapEM	54.61 ± 2.84	63.85 ± 10.52	50.41 ± 1.63	50.64 ± 3.94	50.33 ± 1.89
SS-GMM	100 ± 0	–	–	86.01 ± 18.86	49.55 ± 1.63
MCC results					
Method	*Dalí*	*Oil-Flow*	*Digit1*	*Iris*	*g241c*
SS-Geo-GTM	0.990	**0.711**	**0.561**	**0.842**	0.031
SS-GTM	0.813	0.051	0.048	0.791	**0.150**
LapEM	0.093	0.469	0.008	0.265	0.007
SS-GMM	**1.0**	–	–	0.791	-0.009

performance of the models is measured as the average percentage of correctly classified input samples over one hundred runs (accuracy) and in terms of the Matthews correlation coefficient for multi-class problem [28]. MCC is of common use in the bioinformatics field as a performance measure when the analyzed datasets are class-unbalanced. Both accuracy and MCC measures can be naturally extended from the binary to the multi-class context [28] and their definition is as follows.

Let us assume a classification problem with $\mathcal{S}$ samples and G classes, and two functions defined as $tc, pc : \mathcal{S} \rightarrow \{1, \ldots, G\}$, where $tc(s)$ and $pc(s)$ return the true and the predicted class of s, respectively. The corresponding square confusion matrix C is:

$$C_{ij} = |\{s \in \mathcal{S} : tc(s) = i \; AND \; pc(s) = j\}|, \quad (10)$$

in which the ij-th entry of C is the number of elements of true class i that have been assigned to class j by the classifier. Then, the confusion matrix notation can be used to define both the accuracy and the MCC as:

$$accuracy = \frac{\sum_{k=1}^{G} C_{kk}}{\sum_{i,j=1}^{G} C_{ij}}, \quad (11)$$

$$MCC = \frac{\sum_{k,l,m=1}^{G} C_{kk}C_{ml} - C_{lk}C_{km}}{covXX \cdot covYY}. \quad (12)$$

where

$$covXX = \sqrt{\sum_{k=1}^{G} \left[\left(\sum_{l=1}^{G} C_{lk} \right) \left(\sum_{f,g=1 f \neq k}^{G} C_{gf} \right) \right]}$$

and

$$covYY = \sqrt{\sum_{k=1}^{G} \left[\left(\sum_{l=1}^{G} C_{kl} \right) \left(\sum_{f,g=1 f \neq k}^{G} C_{fg} \right) \right]}.$$

MCC takes values in the interval $[-1, 1]$, where 1 means complete correlation (perfect classification), 0 means no correlation (all samples have been classified to be of only one class) and -1 indicates a negative correlation (extreme misclassification case). The MCC measure was originally extended to the multi-class problem in [28]. Recently in [35], MCC was recommended as an optimal tool for practical tasks, since it presents a good trade-off among discriminatory ability, consistency and coherent behaviors with varying number of classes, unbalanced datasets and randomization.

For the first experiment (only a single randomly selected input sample per class is kept labeled and the remaining samples are considered as unlabeled data), the results are shown in Table 1. SS-Geo-GTM significantly outperforms SS-GTM, LapEM and SS-GMM for almost all data sets in terms of accuracy and MCC measures and, most notoriously, for the data sets of more convoluted geometry (*Oil-Flow* and *Digit1*). The differences with SS-GTM are less notorious for the less convoluted *Iris* data set. LapEM yields a very poor behaviour in this setting. For the linearly separable *Dalí* set, the results for SS-GMM are similar to those of SS-Geo-GTM, but when data presents more difficulties as convoluted geometric properties or data points are not linearly separable SS-GMM can not face these problems. For instance, unstable results are presented for the *Iris* set as suggested for the corresponding standard deviation. Additionally, the SS-GMM model does not run for *Oil-Flow* and *Digit1* sets as indicated by '–' symbol in Table 1. For the two-class non-linearly

Table 2. Classification accuracy (mean ± std) in the presence of increasing levels of uninformative noise

Dataset	noise level	model	Percent of available labels				
			2	4	6	8	10
Dalí	0.1	**SS-Geo**	**100±0**	**100±0**	**100±0**	**100±0**	**100±0**
		SS-GTM	96.29±3.37	98.15±1.97	99.09±1.0	99.31±0.99	99.28±0.89
		LapEM	*75.48±6.56*	*75.73±10.38*	*94.48±4.66*	*98.07±2.02*	*98.50±1.96*
		SS-GMM	100±0	100±0	100±0	100±0	100±0
	0.3	**SS-Geo**	**99.83±1.11**	**100±0**	**100±0**	**100±0**	**100±0**
		SS-GTM	95.57±4.0	98.11±1.45	98.56±0.83	98.77±0.75	98.88±0.69
		LapEM	*74.47±5.27*	*75.11±11.11*	*95.55±4.82*	*99.03±1.96*	*99.54±1.12*
		SS-GMM	100±0	100±0	100±0	100±0	100±0
	0.5	**SS-Geo**	**99.04±3.16**	**100±0**	**100±0**	**100±0**	**100±0**
		SS-GTM	96.52±3.09	98.05±2.16	98.99±1.40	99.31±1.06	99.39±0.78
		LapEM	*77.67±6.79*	*76.56±10.30*	*95.06±4.53*	*97.49±2.76*	*98.87±1.61*
		SS-GMM	100±0	100±0.02	100±0	100±0.02	100±0.02
	1.0	**SS-Geo**	**95.14±5.52**	**97.75±2.94**	**98.71±1.98**	**99.23±0.73**	**99.28±0.92**
		SS-GTM	96.12±3.79	98.36±1.53	98.66±1.21	99.04±0.45	99.06±0.35
		LapEM	*73.86±6.07*	*70.73±10.57*	*92.15±5.34*	*97.23±3.09*	*98.93±1.39*
		SS-GMM	99.34±0.33	99.35±0.35	99.40±0.38	99.33±0.35	99.39±0.34
	2.0	**SS-Geo**	**94.78±3.66**	**96.45±1.63**	**96.96±0.67**	**97.11±0.58**	**97.19±0.48**
		SS-GTM	92.96±3.0	94.28±1.96	94.73±1.75	95.45±1.01	95.36±1.07
		LapEM	*74.02±5.72*	*72.11±11.66*	*90.00±5.91*	*94.54±3.37*	*95.99±1.86*
		SS-GMM	97.27±0.34	97.22±0.36	97.17±0.36	97.07±0.39	97.13±0.37
Oil-Flow	0.01	**SS-Geo**	**88.13±4.05**	**93.87±2.71**	**95.63±2.24**	**96.87±1.45**	**97.26±1.18**
		SS-GTM	55.54±11.94	70.66±5.84	77.14±4.65	80.25±3.58	84.15±3.39
		LapEM	*81.35±5.67*	*88.17±3.41*	*91.80±2.67*	*93.20±2.30*	*94.77±1.70*
		SS-GMM	–	–	–	–	–
	0.03	**SS-Geo**	**88.60±4.06**	**93.34±2.94**	**95.46±1.94**	**96.31±1.64**	**96.98±1.23**
		SS-GTM	55.14±10.71	71.54±6.00	77.26±4.53	81.40±3.63	82.60±3.24
		LapEM	*79.79±7.18*	*90.50±3.72*	*94.00±2.72*	*95.91±1.98*	*96.59±1.13*
		SS-GMM	–	–	–	–	–
	0.05	**SS-Geo**	**90.10±4.38**	**94.94±2.49**	**96.34±1.93**	**97.42±1.69**	**97.84±1.23**
		SS-GTM	53.39±11.81	70.52±7.42	75.79±4.77	81.32±4.52	83.84±4.34
		LapEM	*78.26±7.82*	*92.04±2.81*	*94.86±2.22*	*95.79±1.68*	*96.62±1.37*
		SS-GMM	–	–	–	–	–
	0.1	**SS-Geo**	**60.40±12.81**	**81.48±8.91**	**88.95±4.89**	**91.19±3.59**	**92.49±2.59**
		SS-GTM	49.88±10.11	70.30±8.63	78.20±4.48	82.68±4.50	85.08±4.23
		LapEM	*66.78±11.12*	*87.81±4.79*	*92.50±2.95*	*94.23±2.23*	*95.42±1.78*
		SS-GMM	49.28±3.85	49.23±3.96	49.78±3.72	48.86±3.95	50.26±3.80
	0.2	**SS-Geo**	**59.89±11.38**	**75.76±6.16**	**79.50±5.03**	**83.0±3.78**	**85.41±2.63**
		SS-GTM	44.94±9.92	56.18±10.59	66.01±7.04	72.31±5.55	75.37±4.27
		LapEM	*63.75±7.44*	*77.32±4.55*	*82.22±3.31*	*85.47±2.15*	*86.58±1.84*
		SS-GMM	49.79±3.27	50.37±3.49	50.51±3.48	51.36±3.38	51.36±4.14

separable (very overlapped data points) *g241c* set almost all models behave as at random as expected given the extreme semi-supervised setting.

We now gauge and compare the robustness of the analyzed methods in the presence of noise in some illustrative experiments for the easy separable *Dalí* set and the more complex 12-dimensional *Oil-Flow* set. For this, Gaussian noise of zero mean and increasing standard deviation was added to: a noise-free version of the *Dalí* set (added noise from σ = 0.1 to σ = 2.0) and the most difficult dataset, *Oil-Flow* (added noise from σ = 0.01 to σ = 0.2). The noise scale magnitude is in correspondence with the data scale. We also analyze the evolution of the performance of these models as the percentage of available labels for each dataset is increased from 2% to 10%. Given that these datasets are not class-unbalanced and that the accuracy results are consistent with those of the MCC, as shown in Table 1, in the following experiments only the average classification accuracy is presented.

These new results are shown in Table 2, where boldface, italics and normal fonts were used only to visually follow the alternately results. Here,

the geodesic variant SS-Geo-GTM consistently outperforms SS-GTM (and LapEM) across data sets and noise levels, with few exceptions. The robustness of the semi-supervised procedure for SS-GTM is surprisingly good, though. For the more complex *Oil-Flow* set, both models deteriorate significantly at high noise levels. Overall, these results indicate that the resilience of the models is mostly due to the inclusion of the geodesic metric and not to the semi-supervised procedure itself. It is worth noting that, in general, the results for LapEM only become comparable as the percentage of available labels increases. As in Table 1, SS-GMM results are similar to those of SS-Geo-GTM for the *Dalí* set across noise levels. For *Oil-Flow* set, SS-GMM only can be run for the highest noise levels, where a poor performance is obtained. Furthermore, no benefit is shown when the percentage of label availability is increased.

4 Conclusions and Future Work

The ultimate goal of this thesis was the development of novel generative manifold learning methods for the exploration of partially labeled data. The first contribution is the definition of Geo-GTM as a principled extension of GTM to uncover underlying structures in convoluted datasets. The second one is the definition of SS-Geo-GTM as a principled extension of Geo-GTM to semi-supervised problems. Through several experiments, the performance of SS-Geo-GTM has been assessed, in terms of classification accuracy and MCC, and it has been shown to be consistently better than that of the semi-supervised version of the standard GTM, even in the presence of high levels of noise. Its performance has also been compared to that of LapEM and SS-GMM in several datasets. It has been shown that SS-Geo-GTM significantly outperforms LapEM for all data sets and noise levels, with few exceptions. Furthermore, it has been reported that for datasets which present convoluted geometric properties or when data points are not linearly separable, SS-GMM obtains poor performance compared to SS-Geo-GTM.

As future work, a semi-supervised extension of Geo-GTM using pairwise constraints should be defined in order to deal with semi-supervised clustering tasks.

Acknowledgements

R. Cruz-Barbosa acknowledges the Mexican Secretariat of Public Education (SEP-PROMEP program) for his PhD grant.

References

1. **Archambeau, C. & Verleysen, M.** (**2005**). Manifold constrained finite gaussian mixtures. In **Cabestany, J., Prieto, A., & Sandoval, D. F.**, editors, *Procs. of IWANN*, volume LNCS 3512. Springer-Verlag, 820–828.
2. **Baghshah, M. S. & Shouraki, S. B.** (**2010**). Kernel-based metric learning for semi-supervised clustering. *Neurocomputing*, 73, 1352–1361.
3. **Baluja, S.** (**1998**). Probabilistic modeling for face orientation discrimination: learning from labeled and unlabeled data. In *Neural Information Processing Systems*. 854–860.
4. **Basu, S.** (**2005**). *Semi-supervised Clustering: Probabilistic Models, Algorithms and Experiments*. Ph.D. thesis, The University of Texas at Austin, U.S.A.
5. **Basu, S., Banerjee, A., & Mooney, R. J.** (**2002**). Semi-supervised clustering by seeding. In *Proc. of the 19th International Conference on Machine Learning*.
6. **Basu, S., Bilenko, M., & Mooney, R. J.** (**2004**). A probabilistic framework for semi-supervised clustering. In *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'04*. 59–68.
7. **Basu, S., Davidson, I., & Wagstaff, K.**, editors (**2009**). *Constrained Clustering: Advances in Algorithms, Theory and Applications*. Chapman & Hall/CRC Press.
8. **Belkin, M. & Niyogi, P.** (**2002**). Using manifold structure for partially labelled classification. In *Advances in Neural Information Processing Systems (NIPS) 15*.
9. **Belkin, M. & Niyogi, P.** (**2003**). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6), 1373–1396.
10. **Belkin, M. & Niyogi, P.** (**2004**). Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56, 209–239.
11. **Bernstein, M., de Silva, V., Langford, J., & Tenenbaum, J.** (**2000**). Graph approximations to geodesics on embedded manifolds. Technical report, Stanford University, CA, U.S.A.

12. **Biecek, P., Szczurek, E., Vingron, M., & Tiuryn, J.** (**2012**). The R package bgmm: mixture modeling with uncertain knowledge. *Journal of Statistical Software*, 47(3), 1–31.
13. **Bilenko, M., Basu, S., & Mooney, R. J.** (**2004**). Integrating constraints and metric learning in semi-supervised clustering. In *Proc. of the 21st International Conference on Machine Learning*.
14. **Bishop, C. M., Svensén, M., & Williams, C. K. I.** (**1998**). The Generative Topographic Mapping. *Neural Computation*, 10(1), 215–234.
15. **Blum, A. & Mitchell, T.** (**1998**). Combining labeled and unlabeled data with co-training. In *Procs. of the 11th Annual Conference on Computational Learning Theory (COLT 98)*. 92–100.
16. **Chapelle, O., Chi, M., & Zien, A.** (**2006**). A continuation method for semi-supervised SVMs. In *Proc. of the 23rd International Conference on Machine Learning (ICML 2006)*.
17. **Chapelle, O., Schölkopf, B., & Zien, A.**, editors (**2006**). *Semi-Supervised Learning*. The MIT Press.
18. **Chapelle, O., Sindhwani, V., & Keerthi, S. S.** (**2008**). Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, 9, 203–233.
19. **Chawla, N. & Karakoulas, G.** (**2005**). Learning from labeled and unlabeled data: an empirical study across techniques and domains. *Journal of Artificial Intelligence Research*, 23, 331–366.
20. **Cruz-Barbosa, R. & Vellido, A.** (**2008**). Geodesic Generative Topographic Mapping. In **Geffner, H., Prada, R., Alexandre, I., & David, N.**, editors, *Proceedings of the 11th Ibero-American Conference on Artificial Intelligence (IBERAMIA 2008)*, volume 5290 of $LNAI$. Springer, 113–122.
21. **Cruz-Barbosa, R. & Vellido, A.** (**2010**). Semi-supervised geodesic generative topographic mapping. *Pattern Recogn Lett*, 31, 202–209.
22. **Cruz-Barbosa, R. & Vellido, A.** (**2011**). Semi-supervised analysis of human brain tumours from partially labeled MRS information, using manifold learning models. *Int J Neural Syst*, 21, 17–29.
23. **De Bie, T. & Cristianini, N.** (**2004**). Convex methods for transduction. In **Thrun, S., Saul, L., & Schölkopf, B.**, editors, *Proc. of Advances in Neural Information Processing Systems 16*. MIT Press.
24. **Dempster, A. P., Laird, N. M., & Rubin, D. B.** (**1977**). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39, 1–38.
25. **Dijkstra, E. W.** (**1959**). A note on two problems in connection with graphs. *Numerische Mathematik*, 1, 269–271.
26. **Ghahramani, Z. & Jordan, M. I.** (**1994**). Supervised learning from incomplete data via the EM approach. In *Advances in Neural Information Processing Systems 6*. 120–127.
27. **Goldman, S. & Zhou, Y.** (**2000**). Enhancing supervised learning with unlabeled data. In *Proc. of the 17th International Conference on Machine Learning*. Morgan Kaufmann, 327–334.
28. **Gorodkin, J.** (**2004**). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28, 367–374.
29. **Grira, N., Crucianu, M., & Boujemaa, N.** (**2008**). Active semi-supervised fuzzy clustering. *Pattern Recognition*, 41, 1834–1844.
30. **Hein, M. & Audibert, J. Y.** (**2005**). Intrinsic dimensionality estimation of submanifolds in R^d. In *Proceedings of the 22nd International Conference on Machine Learning*. 289–296.
31. **Herrmann, L. & Ultsch, A.** (**2007**). Label propagation for semi-supervised learning in self-organizing maps. In *Procs. of the 6th WSOM 2007*.
32. **Jain, A. K. & Dubes, R. C.** (**1998**). *Algorithms for Clustering Data*. Prentice Hall, New Jersey.
33. **Joachims, T.** (**1999**). Transductive inference for text classification using support vector machines. In *Procs. of the 16th International Conference on Machine Learning (ICML-99)*. 200–209.
34. **Jones, R.** (**2005**). Learning to extract entities from labeled and unlabeled text. Doctoral Dissertation CMU-LTI-05-191, Carnegie Mellon University.
35. **Jurman, G., Riccadonna, S., & Furlanello, C.** (**2012**). A comparison of MCC and CEN error measures in multi-class prediction. *PLoS ONE*, 7(8), e41882.
36. **Kohonen, T.** (**1995**). *Self-Organizing Maps*. Springer-Verlag, Berlin.
37. **Kulis, B., Basu, S., Dhillon, I. S., & Mooney, R. J.** (**2005**). Semi-supervised graph clustering: a kernel approach. In *Proc. of the 22nd International Conference on Machine Learning*. 457–464.
38. **Lee, J. & Verleysen, M.** (**2007**). *Nonlinear Dimensionality Reduction*. Springer.
39. **Lee, J. A., Lendasse, A., & Verleysen, M.** (**2002**). Curvilinear distance analysis versus isomap. In *Procs. of European Symposium on Artificial Neural Networks (ESANN)*. 185–192.

40. **Matthews, B.** (**1975**). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta - Protein Structure*, 405, 442–451.
41. **Miller, D. J. & Uyar, H. S.** (**1997**). A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Neural Information Processing Systems (NIPS) 9*. 571–577.
42. **Nigam, K.** (**2001**). Using unlabeled data to improve text classification. Doctoral Dissertation CMU-CS-01-126, Carnegie Mellon University.
43. **Nigam, K. & Ghani, R.** (**2000**). Analyzing the effectiveness and applicability of co-training. In *Ninth International Conference on Information and Knowledge Management*. 86–93.
44. **Nigam, K., McCallum, A., Thrun, S., & Mitchell, T.** (**2000**). Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39, 103–134.
45. **Pozdnoukhov, A. & Kanevski, M.** (**2008**). Geokernels: Modeling of spatial data on geomanifolds. In *Proc. of the 16th European Symposium on Artificial Neural Networks (ESANN 2008)*. 277–282.
46. **Roweis, S. T. & Saul, L. K.** (**2000**). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, 2323–2326.
47. **Seeger, M.** (**2000**). Learning with labeled and unlabeled data. Technical report, Institute for ANC, Edinburgh, UK.
48. **Tang, W., Xiong, H., Zhong, S., & Wu, J.** (**2007**). Enhancing semi-supervised clustering: a feature projection perspective. In *Proc. of the 13th ACM SIGKDD International conference on Knowledge discovery and data mining, KDD'07*. 707–716.
49. **Tenenbaum, J. B., de Silva, V., & Langford, J. C.** (**2000**). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2319–2323.
50. **Wang, J. & Shen, X.** (**2007**). Large margin semi-supervised learning. *Journal of Machine Learning Research*, 8, 1867–1891.
51. **Weston, J., Leslie, C., Ie, E., Zhou, D., Elisseeff, A., & Noble, W. S.** (**2005**). Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21, 3241–3247.
52. **Xu, L. & Schuurmans, D.** (**2005**). Unsupervised and semi-supervised multi-class support vector machines. In *Proc. of the 20th National Conference on Artificial Intelligence (AAAI 2005)*.
53. **Xu, Z., Jin, R., Zhu, J., King, I., & Lyu, M.** (**2008**). Efficient convex relaxation for transductive support vector machine. In **Platt, J. C., Koller, D., Singer, Y., & Roweis, S.**, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, 1641–1648.
54. **Yin, X., Chen, S., Hu, E., & Zhang, D.** (**2010**). Semi-supervised clustering with metric learning: An adaptive kernel method. *Pattern Recognition*, 43, 1320–1333.
55. **Zhou, Y. & Goldman, S.** (**2004**). Democratic co-learning. In *Proc. of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004)*. 594–602.
56. **Zhou, Z. H. & Li, M.** (**2005**). Tri-training: exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11), 1529–1541.
57. **Zhu, X.** (**2007**). Semi-supervised learning literature survey. Technical Report TR 1530, University of Wisconsin – Madison.
58. **Zhu, X. & Ghahramani, Z.** (**2002**). Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02-107, Carnegie Mellon University.
59. **Zhu, X. & Goldberg, A. B.** (**2009**). *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers.

Raúl Cruz-Barbosa received his B.S. and M. Sc degree from the computer science department of the Autonomous University of Puebla (Mexico) in 1999 and 2002, respectively. He completed his PhD in Artificial Intelligence at Technical University of Catalonia (Spain) in 2009. He is an associate professor at Technological University of the Mixteca Region (UTM), since 1999, where he has been coordinator of both the applied computing technologies master program and the pattern recognition group. His research interests are related to machine learning (semi-supervised learning, mainly) and digital image processing techniques applied to pattern recognition as well as computer aided detection and diagnosis (of medical images). He is a member of the Mexican National System of Researchers (SNI) of the National Council for Science and Technology (CONACYT).

Alfredo Vellido received his degree in Physics from the Department of Electronics and Automatic Control of the University of the Basque Country (Spain) in 1996. He completed his PhD in Neural Computation at Liverpool John Moores University (UK) in 2000. Following a Ramón y Cajal research fellowship, he is currently assistant professor at Technical University of Catalonia in Barcelona, Spain. Research interests include, but are not limited to, pattern recognition, machine learning and data mining, as well as their application in biomedicine, business, ecology, and e-learning, on which subjects he has published widely. He is currently member of the IEEE Systems, Man & Cybernetics Society Spanish Chapter and the Task Force on Medical Data Analysis in the IEEE-CIS Data Mining Technical Committee.

Article received on 23/12/2011; accepted on 18/06/2013.