Classtering: Joint Classification and Clustering with Mixture of Factor Analysers

Emanuele Sansone and **Andrea Passerini** and **Francesco G.B. De Natale**[1]

Abstract.

In this work we propose a novel parametric Bayesian model for the problem of semi-supervised classification and clustering. Standard approaches of semi-supervised classification can recognize classes but cannot find groups of data. On the other hand, semi-supervised clustering techniques are able to discover groups of data but cannot find the associations between clusters and classes. The proposed model can classify and cluster samples simultaneously, allowing the analysis of data in the presence of an unknown number of classes and/or an arbitrary number of clusters per class. Experiments on synthetic and real world data show that the proposed model compares favourably to state-of-the-art approaches for semi-supervised clustering and that the discovered clusters can help to enhance classification performance, even in cases where the cluster and the low density separation assumptions do not hold. We finally show that when applied to a challenging real-world problem of subgroup discovery in breast cancer, the method is capable of maximally exploiting the limited information available and identifying highly promising subgroups.

1 Introduction

Semi-supervised learning (SSL) is a well-known area of machine learning. The main idea is to exploit both labeled and unlabeled data to increase the performance of classification and clustering. This is motivated by the fact that labeled data are ususally expensive to collect and unlabeled data may aid to learning. The SSL field encompasses both semi-supervised classification and semi-supervised clustering [11]. In the former case, the goal is predicting the labels of unlabeled data based on few observed labeled samples, and smoothness assumptions are typically used in developing methods. The latter task aims at finding clusters in data subject to some given supervised constraints, defined usually as must- and cannot-link between instances.

Discriminative approaches, like Semi-Supervised SVM [7] or Laplacian SVM [6], provide among the best performance in semi-supervised classification. This is because they focus on minimizing an objective function based on classification error, by directly learning the mapping function between the sample and the class space. As a drawback, they cannot provide precise information about intra-class variabilities, since they do not estimate the class-conditional densities. On the other hand, generative approaches learn the joint probability density function over inputs and labels and, while usually not as accurate in classification [11], allow one to model both inter- and intra-class structures.

Concerning semi-supervised clustering, existing algorithms are able to discover the patterns of input data, but they strongly rely on the assumption that clusters have a direct correspondence with the structure of classes, the so-called cluster assumption [11]. However, there are many real-world situations where this assumption does not hold [16]. In fact, if labeled classes split up in different sub-clusters or if several classes cannot be distinguished leading to one larger cluster, then all existing approaches fail. As we will see later on in the experiments, the cluster assumption is not guaranteed also when feature dimensionality reduction is applied to the data.

Existing SSL approaches typically focus on either classification or clustering. However, many real-world applications requires to *jointly* address both tasks by classifying data and identifying groups within each class. Medicine is a paradigmatic example of this requirement. Many diseases are characterized by symptoms for which the discrimination between healthy and pathological cases is often hard, due to the lack of complete understanding of the pathology. Moreover, since the signs of each disease may assume multiple forms, discriminating between the healthy and pathological conditions is not sufficient, and identifying also the different forms of the disease becomes crucial [28].

Based on all these considerations, we introduce a unified generative framework based on mixture of factor analysers that jointly performs classification and reveals the hidden structure of data by estimating the modes and the factors of the class-conditional densities.[2] The framework only relies on the manifold assumption and is thus able to deal with cases where the cluster assumption is not valid. Experiments on synthetic and real world data show that the proposed model compares favourably to state-of-the-art approaches for semi-supervised clustering and that the discovered clusters can help to enhance classification performance. We show also that the proposed model is designed to exploit maximally the limited available information and that it is particularly suited to applications where the collection of new data is very expensive, like in the case of breast cancer samples.

The rest of the paper is organized as follows. At first, the propabilistic graphical model of the mixture of factor analysers in the semi-supervised setting is introduced, then a variational approximation to the log-likelihood function over training data is derived in order to make the posterior inference computationally and analytically tractable and to be able to predict the labels of new unseen data. Related works are then reviewed to highlight the main differences with the proposed method. After that, an extensive analysis of the results obtained in both semi-supervised classification and semi-supervised clustering is provided. Furthermore, the method is tested on a challenging real-world problem consisting in the identification of subgroups in breast cancer samples, obtaining significant results that confirm our claims. Finally, we briefly discuss the main findings

[1] University of Trento, Italy, email: e.sansone@unitn.it

[2] Code available at `https://github.com/emsansone/Classtering`

and highlight some possible directions for future work.

2 Probabilistic Model

We start by introducing a fully-supervised model and then extend it to the semi-supervised case. Given a set of i.i.d. observations $Y = \{\boldsymbol{y}_n\}_{n=1}^N$, where $\mathbf{y}_n \in \mathbb{R}^d$, and the respective set of labels $C = \{c_n\}_{n=1}^N$, where c_n specifies that $\mathbf{y}_n$ belongs to one among K predefined classes, the goal is to learn the underlying distribution generating the observations, namely the class-conditional densities. In particular, if we assume that the densities can be approximated by a Gaussian mixture and that high-dimensional data vectors lie approximately on a lower dimensional subspace, then we can model the data distribution as a mixture of factor analysers (MFA).

In the MFA model, if a factor analyser s_n is given (s_n is an indicator variable identifying one among S factors), then each sample $\boldsymbol{y}_n$ is described through the following linear relation

$$\boldsymbol{y}_n = \boldsymbol{\Lambda}_{s_n}\mathbf{x}_n + \boldsymbol{\mu}_{s_n} + \boldsymbol{\xi}$$

where $\mathbf{x}_n \in \mathbb{R}^k$ is a latent vector distributed according to a Gaussian density with zero-mean and covariance equal to the identity matrix, $\boldsymbol{\Lambda}_{s_n} \in \mathbb{R}^{d\times k}$ and $\boldsymbol{\mu}_{s_n} \in \mathbb{R}^d$ are respectively the factor loading matrix and the bias of factor analyser s_n, and $\boldsymbol{\xi} \in \mathbb{R}^d$ is the noise distributed according to a normal density with diagonal covariance matrix defined by $\boldsymbol{\Psi}$. From this, it is not difficult to show that each sample $\boldsymbol{y}_n$ can be generated by sampling a Gaussian density with mean value equal to $\boldsymbol{\mu}_{s_n}$ and covariance matrix equal to $\boldsymbol{\Lambda}_{s_n}\boldsymbol{\Lambda}_{s_n}^T + \boldsymbol{\Psi}$ [17]. As a consequence, the MFA model can be equivalently interpreted as a Gaussian mixture. In this case, the vector of the mixing proportions is defined by the latent vector $\boldsymbol{\pi} \in [0,1]^S$.

It is worth noting that $\boldsymbol{\Lambda}_{s_n}$ incorporates information about the local dimensionality of component s_n, while $\boldsymbol{\Psi}$ models the variability of data inside that component, namely the noise variance. Parameters $\boldsymbol{\mu}_{s_n}$ and $\boldsymbol{\Lambda}_{s_n}$ are treated as random variables, such that inference is performed by averaging over the ensemble of models and therefore model complexity is automatically taken into account.

The MFA model is an unsupervised method that simultaneously addresses the problem of clustering and the problem of local dimensionality reduction. Supervision can be incorporated into this model by introducing for each sample $\boldsymbol{y}_n$ a pair of independent latent variables $I_n \doteq (s_n, l_n)$, where s_n is the above-mentioned cluster indicator, while l_n is the class indicator.[3] I_n takes into account all $S \times K$ possible combinations between the two indicators. It is worth to say that these combinations are not equally probable. In fact, if we assume that a cluster is associated more likely to one class, then some combinations of clusters and classes tend to appear more often than others. The mixing proportions for variable l_n are therefore defined by the set of random vectors $B = \{\boldsymbol{\beta}_s\}_{s=1}^S$, where each K-dimensional $\boldsymbol{\beta}_s$ is governed by a Dirichlet prior. This means that estimating the distribution over B is equivalent to learning the probabilistic associations between clusters and classes.

The complete set of conditional distributions and priors of our model is summarized by the following relations:

$$p(I_n|\boldsymbol{\pi}, \{\boldsymbol{\beta}_s\}_{s=1}^S) \doteq \boldsymbol{\pi}(s_n)\boldsymbol{\beta}_{s_n}(l_n)$$

$$p(c_n|I_n) \doteq \delta(c_n - l_n)$$

[3] In our case, there is no distinction between l_n and c_n. Nevertheless, we keep these two variables separate. This is helpful for modelling scenarios with multiple and/or noisy labels. In these cases, l_n is the hidden true label, while c_n is the label provided by the annotator.

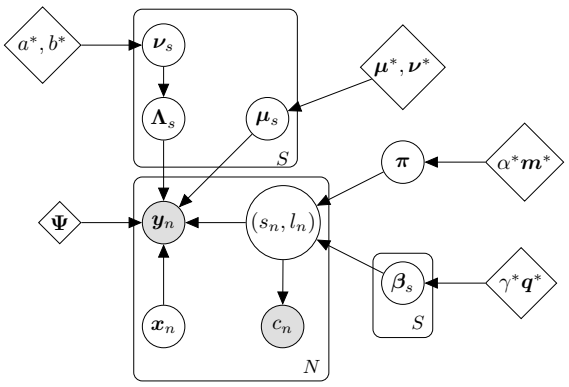


Figure 1. Representation of the supervised MFA model as a directed acyclic graph.

$$p(\boldsymbol{\Lambda}_s|\boldsymbol{\nu}_s) \doteq \prod_{j=1}^k \mathcal{N}(0, I/\boldsymbol{\nu}_s(j))$$

$$p(\boldsymbol{\nu}_s|a^*, b^*) \doteq \prod_{j=1}^k Gamma(\boldsymbol{\nu}_s(j)|a^*, b^*)$$

$$p(\boldsymbol{\mu}_s|\boldsymbol{\mu}^*, \boldsymbol{\nu}^*) \doteq \mathcal{N}(\boldsymbol{\mu}^*, diag(\boldsymbol{\nu}^*)^{-1})$$

$$p(\boldsymbol{\pi}|\alpha^*\boldsymbol{m}^*) \doteq Dir(\alpha^*\boldsymbol{m}^*)$$

$$p(\boldsymbol{\beta}_s|\gamma^*\boldsymbol{q}^*) \doteq Dir(\gamma^*\boldsymbol{q}^*)$$

$$p(\boldsymbol{x}_n) \doteq \mathcal{N}(0, I)$$

where I is the identity matrix and $\boldsymbol{\nu}_s$ is a k-dimensional vector whose elements govern the columns of $\boldsymbol{\Lambda}_s$. The mechanism known as automatic relevance determination (ARD) is used to improve the task of dimensionality reduction [9]. $a^*, b^*, \boldsymbol{\Psi}, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*, \gamma^*\boldsymbol{q}^*$ are the hyperparameters of the model. In Figure 1, we show the graphical representation of our probabilistic model.

In the next two sections, we see how to apply this probabilistic graphical model to the semi-supervised scenario. In particular, a variational approximation of the log-likelihood function over input data and labels is derived in order to make inference computationally tractable. The unlabeled data are therefore taken into account by simply adding their contribution to the estimated lower bound. Then, we show how to predict the labels of unseen data.

3 Variational Approximation

By defining $\mathcal{H} \doteq \{\boldsymbol{x}_n, I_n\}$ as the set of hidden variables and $\Theta \doteq \{\boldsymbol{\pi}, \{\boldsymbol{\beta}_s, \boldsymbol{\Lambda}_s, \boldsymbol{\mu}_s, \boldsymbol{\nu}_s\}_{s=1}^S\}$ as the set of parameters, we can express the log-likelihood function over Y and C as

$$\ln p(Y, C) = \ln \int d\Theta p(\Theta) \int d\mathcal{H} p(Y, C, \mathcal{H}|\Theta)$$

and by exploiting the conditional dependencies defined by the probabilistic graphical model we obtain that

$$\ln p(Y, C) = \ln \int d\Theta p(\Theta) \prod_{n=1}^N \sum_{s_n=1}^S \sum_{l_n=1}^K p(I_n|\Theta)p(c_n|I_n) \cdot \cdot \int d\boldsymbol{x}_n p(\boldsymbol{x}_n) p(\boldsymbol{y}_n|\Theta, \boldsymbol{x}_n, I_n, \boldsymbol{\Psi}) \quad (1)$$

Since the integrals in (1) are computationally and analytically intractable, we employ a standard approach to solve the Bayesian integration based on the variational approximation [2]. In practice, by introducing some auxiliary distributions for both the parameters and the hidden variables and by applying the Jensen's inequality, it is possible to obtain a lower bound on the log-likelihood over Y and C, namely

$$\begin{aligned}\ln p(Y,C) \geq & \int d\boldsymbol{\pi} q(\boldsymbol{\pi}) \ln \frac{p(\boldsymbol{\pi}|\alpha^* \boldsymbol{m}^*)}{q(\boldsymbol{\pi})} \\ & + \sum_{s=1}^{S} \int d\boldsymbol{\beta}_s q(\boldsymbol{\beta}_s) \ln \frac{p(\boldsymbol{\beta}_s|\gamma^* \boldsymbol{q}^*)}{q(\boldsymbol{\beta}_s)} + \sum_{s=1}^{S} \int d\boldsymbol{\nu}_s q(\boldsymbol{\nu}_s) \\ & \cdot \left[\ln \frac{p(\boldsymbol{\nu}_s|a^*, b^*)}{q(\boldsymbol{\nu}_s)} + \int d\tilde{\boldsymbol{\Lambda}}_s q(\tilde{\boldsymbol{\Lambda}}_s) \ln \frac{p(\tilde{\boldsymbol{\Lambda}}_s|\boldsymbol{\nu}_s, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*)}{q(\tilde{\boldsymbol{\Lambda}}_s)} \right] \\ & + \sum_{n=1}^{N} \sum_{s_n=1}^{S} \sum_{l_n=1}^{K} q(I_n) \Big[\int d\boldsymbol{\pi} q(\boldsymbol{\pi}) \int d\boldsymbol{\beta}_{s_n} q(\boldsymbol{\beta}_{s_n}) \\ & \cdot \ln \frac{p(I_n|\boldsymbol{\pi}, \boldsymbol{\beta}_{s_n})}{q(I_n)} + \int d\boldsymbol{x}_n q(\boldsymbol{x}_n|I_n) \ln \frac{p(\boldsymbol{x}_n)}{q(\boldsymbol{x}_n|I_n)} \\ & + \ln p(c_n|I_n) \int d\tilde{\boldsymbol{\Lambda}}_{s_n} q(\tilde{\boldsymbol{\Lambda}}_{s_n}) \int d\boldsymbol{x}_n q(\boldsymbol{x}_n|I_n) \\ & \cdot \ln p(\boldsymbol{y}_n|\tilde{\boldsymbol{\Lambda}}_{s_n}, \boldsymbol{x}_n, I_n, \boldsymbol{\Psi}) \Big] \doteq \mathcal{F}(\mathcal{Q}) \end{aligned} \tag{2}$$

where $\mathcal{Q}$ is the set of all auxiliary distributions, namely $q(\boldsymbol{\pi})$, $\{q(\boldsymbol{\beta}_s), q(\boldsymbol{\nu}_s), q(\tilde{\boldsymbol{\Lambda}}_s)\}_{s=1}^{S}$, $\{q(I_n), \boldsymbol{x}_n|I_n)\}_{n=1}^{N}$, and $\tilde{\boldsymbol{\Lambda}}_s$ represents the concatenation between $\boldsymbol{\Lambda}_s$ and $\boldsymbol{\mu}_s$. By maximizing the functional $\mathcal{F}$, the lower bound is guaranteed to monotonically increase [5] and can be used as an approximation of the log-likelihood function over Y and C. Furthermore, the functional $\mathcal{F}$ is used to compare models with different number of factor analysers in order to perform automatic model selection and choose the proper value of S.

The model can be further extended to perform semi-supervised classification by introducing the set of unlabeled observations $Y' = \{\boldsymbol{y}'_m\}_{m=1}^{N'}$ and by averaging over all possible labels. The extended log-likelihood function is therefore approximated following the same procedure in (2), namely

$$\begin{aligned}\ln p(Y,Y',C) \geq & \mathcal{F}(\mathcal{Q}) + \sum_{m=1}^{N'} \sum_{s_m=1}^{S} \sum_{l_m=1}^{K} q(I_m) \Big[\int d\boldsymbol{\pi} q(\boldsymbol{\pi}) \\ & \cdot \int d\boldsymbol{\beta}_{s_m} q(\boldsymbol{\beta}_{s_m}) \ln \frac{p(I_m|\boldsymbol{\pi}, \boldsymbol{\beta}_{s_m})}{q(I_m)} + \int d\boldsymbol{x}_m q(\boldsymbol{x}_m|I_m) \\ & \cdot \ln \frac{p(\boldsymbol{x}_m)}{q(\boldsymbol{x}_m|I_m)} + \int d\tilde{\boldsymbol{\Lambda}}_{s_m} q(\tilde{\boldsymbol{\Lambda}}_{s_m}) \int d\boldsymbol{x}_m q(\boldsymbol{x}_m|I_m) \\ & \cdot \ln p(\boldsymbol{y}'_m|\tilde{\boldsymbol{\Lambda}}_{s_m}, \boldsymbol{x}_m, I_m, \boldsymbol{\Psi}) \Big] \end{aligned} \tag{3}$$

which is equivalent to (2) except for the last three addends, that represent the contribution of unlabeled samples to the the lower bound.

4 Posterior Inference and Prediction

Posteriors over parameters and hidden variables are estimated by optimizing the functional in (3). The optimization is performed by taking the functional derivatives of (3) with respect to all auxiliary distributions $q(\cdot)$ and equating them to zero. Similarly, the hyperparameters of the model are estimated by simply taking the derivatives of the lower bound in (3) with respect to $a^*, b^*, \boldsymbol{\Psi}, \boldsymbol{\mu}^*, \boldsymbol{\nu}^*, \gamma^* \boldsymbol{q}^*$. This operation is equivalent to performing a maximum likelihood estimation, where the true log-likelihood function is replaced by its lower bound. Iterative updates of the auxiliary distributions and of the hyperparameters guarantee to monotonically and maximally increase the lower bound in (3), as shown in [5].

After the optimization is completed, the model can be used to predict the labels of new observed samples. In fact, if we define $D = Y \cup Y'$ as the set of data used for training the model and $Y'' = \{\boldsymbol{y}''_j\}_{j=1}^{M}$ as the set of test data, then the new labels can be estimated by maximizing the log-likelihood function conditioned on D and C. During the maximization, $\ln p(Y''|D,C)$ can be approximated by replacing the true parameter posterior with the estimated auxiliary distribution over the parameters, namely

$$\begin{aligned}\ln p(Y''|D,C) &= \ln \int d\Theta p(\Theta|D,C) \int d\mathcal{H} p(Y'', \mathcal{H}|\Theta, D, C) \\ &\approx \ln \int d\Theta q(\Theta) \int d\mathcal{H} p(Y'', \mathcal{H}|\Theta, D, C) \end{aligned} \tag{4}$$

Integrals in (4) are computationally intractable. Similarly to the (2) and (3) cases, we thus look for a tractable lower bound on $\ln p(Y''|D,C)$

$$\begin{aligned}\ln p(Y''|D,C) \geq & \sum_{j=1}^{M} \sum_{s_j=1}^{S} \sum_{l_j=1}^{K} q(I_j) \Big[\int d\boldsymbol{\pi} q(\boldsymbol{\pi}) \int d\boldsymbol{\beta}_{s_j} q(\boldsymbol{\beta}_{s_j}) \\ & \cdot \ln \frac{p(I_j|\boldsymbol{\pi}, \boldsymbol{\beta}_{s_j})}{q(I_j)} + \int d\boldsymbol{x}_j q(\boldsymbol{x}_j|I_j) \ln \frac{p(\boldsymbol{x}_j)}{q(\boldsymbol{x}_j|I_j)} \\ & + \int d\tilde{\boldsymbol{\Lambda}}_{s_j} q(\tilde{\boldsymbol{\Lambda}}_{s_j}) \int d\boldsymbol{x}_j q(\boldsymbol{x}_j|I_j) \ln p(\boldsymbol{y}''_j|\tilde{\boldsymbol{\Lambda}}_{s_j}, \boldsymbol{x}_j, I_j, \boldsymbol{\Psi}) \Big] \end{aligned} \tag{5}$$

Note that (5) is similar to the last three addends of (3). In this case, we are only interested in estimating the labels of test data and this is performed by taking the functional derivatives of (5) with respect to $q(I_j)$ for $j = 1, \ldots, M$.

5 Related Work

Mixture of Factor Analysers (MFA) has been extensively studied in the past. The model is targetted to the unsupervised learning setting, especially to perform model-based clustering in high-dimensional data [17]. The property of handling data in high dimensions is fundamental to distinguish it from the classical finite mixture models [25], like the mixture of Gaussians. Another interesting property is that the model can perform local dimensionality reduction. These aspects are particularly insightful for applications in computer vision, to perform density image estimation [34] and object tracking [35], or in biology, to cluster microarray data based on genes [24]. For a detailed overview of MFA and finite mixture models see [10]. The properties of MFA are promising also for the supervised and the semi-supervised learning setting and our work proposes an approach exactly in that regard.

Some other works based on finite mixture models are similar to ours, but differs for the kind of assumptions made. The work in [26] proposes a Gaussian mixture model that integrates the information about the presence/absence of labels to perform new class discovery. The model assumes that each cluster has a distribution over labels, but no information about the correspondences between classes and clusters is added to the generative model, thus making it dependent on the cluster assumption. In the experimental section we will see that this assumption is quite limiting for many cases. The

work in [27] proposes a finite mixture model for semisupervised classification. In their generative model, labeled samples are conditioned to unlabeled ones in order to ensure that, during the inference stage, the propagation of labels through the unlabeled samples respects the smoothness assumption. The authors apply the method also to the unsupervised learning setting, in particular to perform density estimation. Nevertheless, the experimental evaluation highlights the limitations of the method in this kind of setting, where the results are frequently worse than performance obtained by standard unsupervised techniques. In the context of semi-supervised clustering, the works in [21, 23, 22, 3] have addressed the problem of constraint propagation proposing solutions that fulfill both the constraints and the smoothness requirement. Like the other works in semi-supervised classification, they haven't considered that the problem of label/constraint propagation may be due to the violation of the cluster assumption. The recent work in [31] is probably the closest to ours. The method introduces a finite mixture model able to deal with an arbitrary number of clusters and classes. The learning is performed by optimizing an objective characterized by the log-likelihood function weighted by a term penalizing the violation of the must- and cannot-link constraints. Furthermore, a hard assignment between clusters and classes determines a partitioning of the feature space in which the majority of the constraints is satisfied. In our approach, instead, the assignment between clusters and classes is soft. This is essential for modelling the uncertainty of assignment due to the small amount of supervised information. Furthermore, the method is tested on datasets characterized by only few dozens of features.

The authors in [20] have recently proposed a unified framework that combines deep neural networks with generative models. The neural network learns an embedding of data and the generative models performs classification based on this new representation. The combination of these two parts is obtained by defininig a single probabilistic graphical model that permits to achieve good classification performance even when compared to discriminative approaches. Nevertheless, the framework is not designed to perform clustering and is based on the assumption that there exists a data representation for which the cluster assumption is valid. Furthermore, the use of a deep neural network requires generally large data sets for training, besides having to choose the proper architecture, making the framework not suitable to applications with limited number of samples.

6 Experimental Results

6.1 Data sets

In order to assess the performance of the proposed model and compare it with state-of-the-art approaches, we performed experiments on three artificial and three real world data sets. Table 1 summarizes their properties.

Table 1. Experimental data sets

Data sets	Classes	Features	Instances
G50C	2	50	550
CAKE	2	2	1000
TOES	2	2	1000
IRIS	3	4	150
USPS	3	256	1918
ISOLET	2	617	3119

The first synthetic data set, G50C, is inspired by [18]. Data are generated from two standard normal densities located in a 50-dimensional space, such that the Bayes error is 5%. In this case, each class is represented by only one Gaussian. In the second data set, CAKE, data are uniformly distributed according to a two-dimensional round shape. Two orthogonal decision functions are used to discriminate between the two classes in order to make them non-linearly separable. The Gaussian and the cluster assumptions do not hold in this case. The third data set, TOES, represents the case where class-conditional densities are characterised by multiple clusters. Samples are drawn independently from a two-dimensional density composed by five Gaussians, two for the first class and three for the second class. The two classes have the same prior, resulting into a balanced number of samples per class. The different number of clusters per class is useful to analyse how unlabeled data influence the decision boundary. Figures 3(a) and 4(a) show the representation of the CAKE and the TOES data sets respectively.

The real world data sets consist of two-class and multi-class problems from the UCI repository. The IRIS data set contains data belonging to three different classes of iris plants. One of the three classes is not linearly separable from the others. The second real world data set, USPS, represents a well-known benchmark for hadwritten digits recognition. In our experiments, we only used samples belonging to the categories of digits 3, 8 and 9, which are among the most difficult classes to recognize [15]. In order to deal with real-valued vectors, normalized histograms are used as feature descriptors. Finally, the ISOLET data set contains high-dimensional data for the spoken letter recognition task. In our case, the first three subsets of the whole collection were considered. Similarly to [6], we decided to classify the first 13 letters of the English alphabet from the last 13.

For the USPS and ISOLET data sets, we first apply a state-of-the-art technique for unsupervised dimensionality reduction, called t-SNE [14]. The motivation for the choice of t-SNE relies on the capability of visualizing high-dimensional data sets in a two or three-dimensional map without loosing too much information about the local and the global structure of data. Compared to other existing techniques, like Sammon mapping, Isomap and Locally Linear Embedding, t-SNE provides significantly better performance, especially in the data visualization task.

6.2 Semi-Supervised Clustering

For each data set, the number of labeled instances is varied in between 0 and 90 samples per class. For each of these configurations, 20 different data sets are generated by random sampling. To adhere to the problem of semi-supervised clustering, labeled samples are then converted into a balanced number of must- and cannot-link constraints following the same procedure of [1]. Performance are measured in terms of the normalized mutual information (NMI) using the true labels as gold standard.

We compare our method, called Classtering (CLSST for short), with four state-of-the-art approaches. The first method proposed in [33] is based on the integration of supervised constraints into a Gaussian mixture model (CGMM). The second method (MCCC) is the recent work proposed in [36], where the problem is formulated as a matrix completion task. The third method in [8] is based on an extension of the k-means algorithm (MPCK), where constraints and metric learning are incorporated into the objective function to enhance the performance. The last method is the semi-supervised kernel mean shift (SKMS) proposed in [1], where data are first mapped into a higher dimensional space and then clustered by the mean shift algorithm.

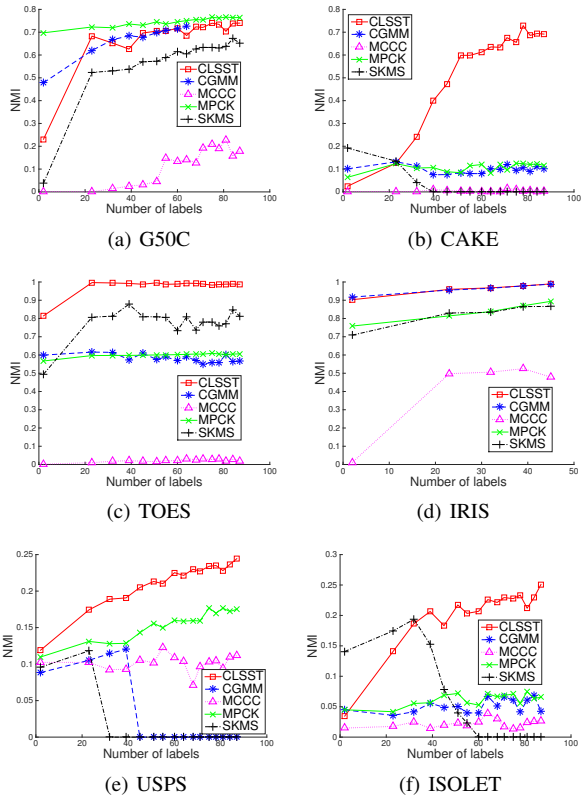


Figure 2. Experimental results for semi-supervised clustering (the higher the better).

For all competitors, the parameters are chosen from a finite grid set such that the best performance are always considered. In particular, the tradeoff parameter C for MCCC is chosen from the set $\{0.1, 1, 10, 100, 1000\}$, while the regularization parameter γ of SKMS is chosen from the range $\{10, 100, 1000\}$. It is important to mention that the number of clusters for MCCC, MPCK and CGMM is equal to four in CAKE and five in TOES, while it is chosen to be equal to the number of classes in all other data sets, as done in [1, 33, 8, 36]. It is also worth noting that our algorithm does not require to set any parameter manually, since all hyperparameters are learnt automatically during the training procedure.[4]

Figure 2 shows the results obtained over all data sets. CLSST clearly outperforms all competitors in all cases, except for the G50C data set. In this case, the data are generated from a distribution of two Gaussians with identity covariance matrices. Authors in [4] prove mathematically that the k-means algorithm is equivalent to performing an EM algorithm on a mixture of Gaussians under the assumption of identity covariance matrices and uniform mixture priors, which clearly motivates why MPCK, that is k-means-based, achieves very good performance. The gap with respect to the results obtained by CLSST on G50C are mainly due to the fact that, while in CLSST the parameters of the Gaussians are assumed to be random variables, in MPCK it is assumed that there only exists a unique combination of true parameters. It is worth mentioning that CLSST and CGMM are both algorithms based on Gaussian mixtures. In fact, when considering cases characterized by one cluster per class, namely the G50C and the IRIS data sets, the performance of both methods are almost equivalent. When the cluster assumption does not hold, viz in the CAKE data set, it is clearly visible that all methods, except CLSST, fail. The same holds when considering multiple clusters per class, i.e. the TOES data set. A representative example of the results obtained on CAKE and TOES can be visualized more intuitively in Figure 3 and Figure 4. A thorough analysis of the results obtained on the USPS and the ISOLET data sets performed by data visualization indicates that in both cases we have a superposition of two effects, namely the absence of validity of the cluster assumption and the presence of multiple clusters per class, which explains why results are qualitatively similar to those obtained on the CAKE and the TOES data sets.

[4] Except for the dimensionality of the latent variables x_n, which is always set to a low value.

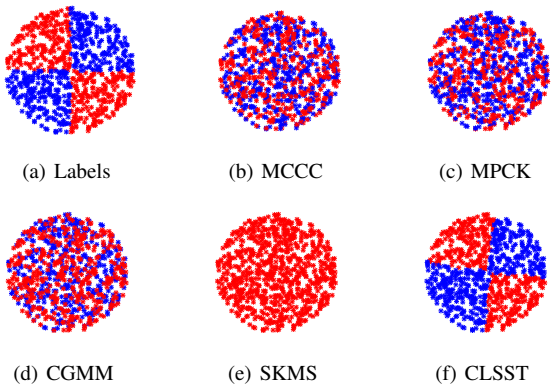


Figure 3. Estimated labels for semi-supervised clustering on CAKE data set (87 labels per class)

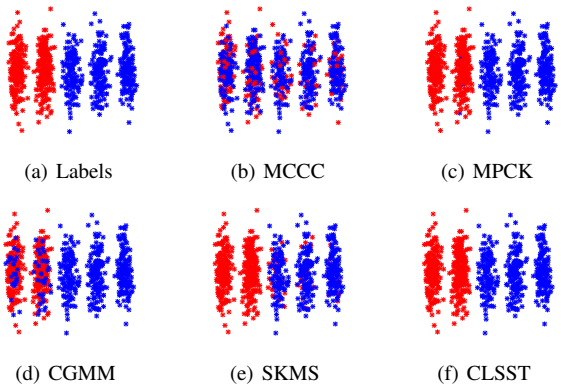
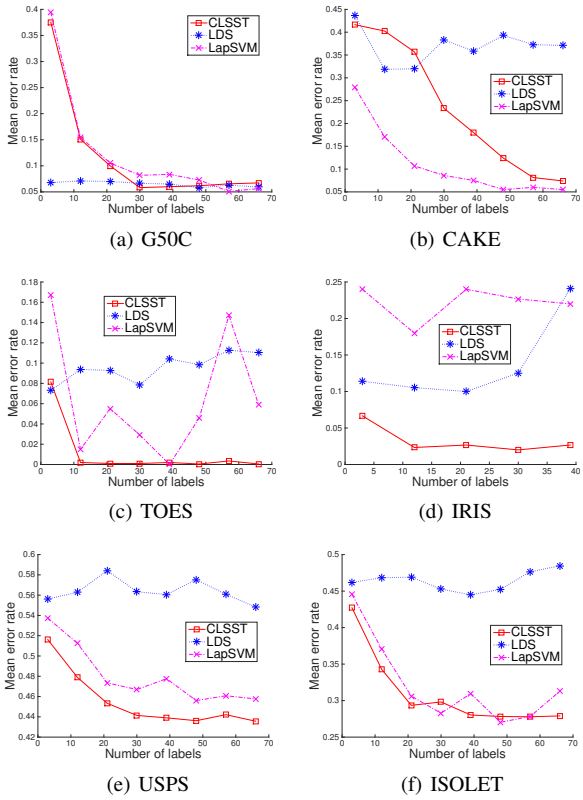


Figure 4. Estimated labels for semi-supervised clustering on TOES data set (87 labels per class)

6.3 Semi-Supervised Classification

For each data set, 5-fold cross-validation is used to split data into training and test parts. For each training split, the number of labeled instances is varied between 0 and 70 samples per class. 10 different data sets are then generated by random sampling. In these experiments, we provide estimates of the generalization error. Performance are measured in terms of the error rate, since all datasets have a balanced number of samples per class.

Our method is compared against two state-of-the-art approaches.



(a) G50C (b) CAKE

(c) TOES (d) IRIS

(e) USPS (f) ISOLET

Figure 5. Experimental results for semi-supervised classification (the lower the better).

The first method [12] is based on the low-density separation assumption (LDS), which is the equivalent supervised form of the cluster assumption. A nearest-neighbor graph is used to compute the kernel matrix of an SVM. The second method proposed in [6] is an extension of the SVM framework, namely the Laplacian SVM (LapSVM). In particular, a penalty term is added to the objective function to take into account the marginal distribution of unlabeled data.

For each training data set, hyperparameters for the two competitors are selected from a finite grid using an inner 3-fold cross-validation procedure. For LDS, ρ and C are respectively chosen from $\{1, 2, 4, 8\}$ and $\{0.1, 1, 10, 100\}$. In LapSVM, γ_A and γ_I are chosen from the same range, namely $\{0.005, 0.045, 05\}$. The σ value is chosen for both methods in the range $\{0.1, 1, 10\}$ in a transductive setting on the entire training set.

Figure 5 shows the results obtained over all data sets.

CLSST outperforms all other approaches in almost all cases, except for the G50C and the CAKE data sets. G50C is in fact the perfect scenario for approaches relying on the low density separation assumption, as it was previously seen in the clustering setting for the methods based on the cluster assumption. This motivates why LDS provides good performance even when the number of labeled samples is small.

In the CAKE data set, CLSST performs slightly worse than LapSVM. This is due to the fact that the Gaussian mixture density is not able to fit properly with the uniform distribution of unlabeled samples. The bias decreases as the number of labeled samples increases. In contrast, LapSVM is able to control the negative effect of the unlabeled samples even when the number of labeled data is small by assigning a higher value to the classification error term in the objective function. In all other cases, it is evident that CLSST achieves better performance than its competitors. This can be explained by the fact that our model is really flexible in estimating the class-conditional densities and the gained information about these distributions provides an effective way to fully exploit the unlabeled samples and increase the classification accuracy.

6.4 Subgroup Discovery in Breast Cancer

We finally tested our algorithm on a challenging real-world problem consisting in the identification of subgroups in breast cancer samples. A recent extensive study [13] analysed about 2,000 clinically annotated primary breast cancers collected from various sources and identified 10 novel subgroups with varying degrees of confidence. The authors used a subset of 997 samples as discovery set to identify clusters, and the remaining 995 ones as validation set to evaluate robustness of the detected clusters. Clustering was done with a joint latent variable model [32] on a set of 754 gene expression profiles. Reproducibility of clustering was measured in terms of in-group proportion (IGP) [19], which is the proportion of samples in a group whose nearest neighbours are also in the same group, after assigning samples in the validation set to the clusters in the discovery set.

Characterizing tumors in terms of subclasses is a crucial step in order to understand their behaviour and variability, and there is extensive literature addressing this task and proposing various classication schemes. Five "intrinsic" subtypes of human breast tumors have been identified in early studies [30] and termed Luminal A, Luminal B, HER2-Enriched (HER2-E), Basal-like and normal. The PAM50 gene is typically used [29] for gene expression-based subtyping in these five groups. Most of the 10 clusters identified in [13] contain samples belonging to multiple subtypes.

What we plan to investigate here is whether incorporating intrinsic subtype classification as class labeling can produce a clustering with improved generalization capability, as measured by IGP. We first reduce data to 50 features using PCA in order to alleviate the problem of redundant features and then apply CLSST to discover the clusters. In this particular setting, the algorithm discovers seven groups achieving an averaged IGP of 74.8%, with a minimum value of 57.5% and a maximum of 91.2%. After this, we investigate a second setting, where we run CLSST by fixing the number of clusters to ten, in order to have a fair comparison with the results reported in [32]. In this second configuration, the obtained IGP scores range from a minimum of 55.7% to a maximum of 92.7% with a mean value equal to 70.6%. In both settings, the performance are better than those obtained in [32], where the IGP values span from a minimum of 44.8% to a maximum of 82.4% with a mean value equal to 65.4%. The performance improvement is on average greater than 5%, indicating that our algorithm successfully exploits the supervised information in performing group discovery. Table 2 reports the complete set of results for [32] and for CLSST in the two settings, with clusters ordered by decreasing IGP value.

With this study we are not claiming that the clusters we found are more biologically relevant than those identified by the original method, as this would require in-depth analyses and extensive validations, which are out of the scope of this work. Nonetheless, we believe that the obtained results are promising and highlight the potential of the method in discovering structure in data.

Table 2. Results on breast cancer data set evaluated in terms of IGP measure. Clusters are ordered by decreasing IGP value.

Cluster	[13]	CLSST (fixed S)	CLSST (variable S)
1	0.8235	0.9266	0.9117
2	0.8099	0.8639	0.8377
3	0.7281	0.7899	0.7931
4	0.7091	0.6867	0.7730
5	0.6866	0.6842	0.7624
6	0.6455	0.6794	0.5833
7	0.6015	0.6780	0.5745
8	0.5818	0.6000	-
9	0.5072	0.5965	-
10	0.4481	0.5574	-

7 Discussion

In this work, a model based on mixture of factor analysers is proposed for both semi-supervised clustering and semi-supervised classification. Evaluation is performed on synthetic and real-world data sets. Results provide evidence about the effectiveness of the proposed model, especially when the cluster or the low density separation assumption does not hold. Furthermore, we have applied the proposed method to a challenging real-world problem consisting in the identification of subgroups in breast cancer samples and achieved promising results that enable future research in this direction. Other possible directions consist of extending the model to deal with multiple labels (e.g. difference classification schemes for tumors) and to perform active learning.

REFERENCES

[1] S. Anand, S. Mittal, O. Tuzel, and P. Meer, 'Semi-Supervised Kernel Mean Shift Clustering', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **36**(6), 1201–1215, (2014).

[2] H. Attias, 'Inferring Parameters and Structure of Latent Variable Models by Variational Bayes', in *Uncertainty in Artificial Intelligence*, pp. 21–30, (1999).

[3] M.S. Baghshah and S.B. Shouraki, 'Semi-Supervised Metric Learning Using Pairwise Constraints', in *International Joint Conference on Artificial Intelligence*, volume 9, pp. 1217–1222, (2009).

[4] S. Basu, A. Banerjee, and R. Mooney, 'Semi-Supervised Clustering by Seeding', in *International Conference in Machine Learning*, (2002).

[5] M.J. Beal, 'Variational Algorithms for Approximate Bayesian Inference', *Ph. D. Thesis*, (2003).

[6] M. Belkin, P. Niyogi, and V. Sindhwani, 'Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples', *Journal of Machine Learning Research*, **7**, 2399–2434, (2006).

[7] K. Bennett, A. Demiriz, et al., 'Semi-Supervised Support Vector Machines', *Neural Information Processing Systems*, 368–374, (1999).

[8] M. Bilenko, S. Basu, and R. Mooney, 'Integrating Constraints and Metric Learning in Semi-Supervised Clustering', in *International Conference in Machine Learning*, p. 11, (2004).

[9] CM Bishop, 'Variational Principal Components', in *International Conference on Artificial Neural Networks*, volume 1, pp. 509–514, (1999).

[10] C. Bouveyron and C. Brunet-Saumard, 'Model-based clustering of high-dimensional data: A review', *Computational Statistics & Data Analysis*, **71**, 52–78, (2014).

[11] O. Chapelle, B. Schölkopf, and A. Zien, 'Semi-Supervised Learning', (2010).

[12] O. Chapelle and A. Zien, 'Semi-Supervised Classification by Low Density Separation', *International Conference on Artificial Intelligence and Statistics*, 57, (2005).

[13] C. Curtis, S. Shah, S-F Chin, et al., 'The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups.', *Nature*, **486**(7403), 346–352, (2012).

[14] L. Van der Maaten and G. Hinton, 'Visualizing Data Using t-SNE', *Journal of Machine Learning Research*, **9**(2579-2605), 85, (2008).

[15] M. Diem, S. Fiel, A. Garz, M. Keglevic, F. Kleber, and R. Sablatnig, 'ICDAR 2013 Competition on Handwritten Digit Recognition (HDRC 2013)', in *International Conference on Document Analysis and Recognition*, pp. 1422–1427, (2013).

[16] I. Färber, S. Günnemann, H-P Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek, 'On Using Class-Labels in Evaluation of Clusterings', in *MULTICLUST: International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD*, p. 1, (2010).

[17] Z. Ghahramani, M.J. Beal, et al., 'Variational Inference for Bayesian Mixtures of Factor Analysers', in *Neural Information Processing Systems*, pp. 449–455, (1999).

[18] Y. Grandvalet and Y. Bengio, 'Semi-Supervised Learning by Entropy Minimization', in *Neural Information Processing Systems*, pp. 529–536, (2005).

[19] A. Kapp and R. Tibshirani, 'Are clusters found in one dataset present in another dataset?', *Biostatistics*, **8**(1), 9–31, (2007).

[20] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling, 'Semi-Supervised Learning with Deep Generative Models', in *Neural Information Processing Systems*, pp. 3581–3589, (2014).

[21] M.HC Law, A.P. Topchy, and A.K. Jain, 'Model-based Clustering With Probabilistic Constraints', in *International Conference on Data Mining, SIAM*, pp. 641–645, (2005).

[22] Z. Li, J. Liu, and X. Tang, 'Pairwise Constraint Propagation by Semidefinite Programming for Semi-Supervised Classification', in *International Conference in Machine Learning*, pp. 576–583, (2008).

[23] Z. Lu, 'Semi-Supervised Clustering with Pairwise Constraints: A Discriminative Approach', in *International Conference on Artificial Intelligence and Statistics*, pp. 299–306, (2007).

[24] G. McLachlan, R.W. Bean, and D. Peel, 'A mixture model-based approach to the clustering of microarray expression data', *Bioinformatics*, **18**(3), 413–422, (2002).

[25] G. McLachlan and D. Peel, *Finite mixture models*, John Wiley & Sons, 2004.

[26] D.J. Miller and J. Browning, 'A Mixture Model and EM-based Algorithm for Class Discovery, Robust Classification, and Outlier Rejection in Mixed Labeled/Unlabeled Data Sets', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **25**(11), 1468–1483, (2003).

[27] D.J. Miller, J. Raghuram, G. Kesidis, and C.M. Collins, 'Improved Generative Semisupervised Learning Based on Finely Grained Component-Conditional Class Labeling', *Neural Computation*, **24**(7), 1926–1966, (2012).

[28] V. Moskvina, N. Craddock, P. Holmans, et al., 'Gene-wide Analyses of Genome-wide Association Data Sets: Evidence for Multiple Common Risk Alleles for Schizophrenia and Bipolar Disorder and for Overlap in Genetic Risk', *Molecular Psychiatry*, **14**(3), 252–260, (2009).

[29] J. Parker, M. Mullins, M. Cheang, et al., 'Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes', *Journal of Clinical Oncology*, **27**(8), 1160–1167, (2009).

[30] C. Perou, T. Sørlie, M. Eisen, et al., 'Molecular Portraits of Human Breast Tumours', *Nature*, **406**(6797), 747–752, (2000).

[31] J. Raghuram, D.J. Miller, and G. Kesidis, 'Instance-Level Constraint-Based Semisupervised Learning With Imposed Space-Partitioning', *Neural Networks and Learning Systems, IEEE Transactions on*, **25**(8), 1520–1537, (2014).

[32] R. Shen, A. Olshen, and M. Ladanyi, 'Integrative Clustering of Multiple Genomic Data Types Using a Joint Latent Variable Model with Application to Breast and Lung Cancer Subtype Analysis', *Bioinformatics*, **25**(22), 2906–2912, (2009).

[33] N. Shental, A. Bar-hillel, T. Hertz, and D. Weinshall, 'Computing Gaussian Mixture Models with EM Using Equivalence Constraints', in *Neural Information Processing Systems*, pp. 465–472, (2004).

[34] J. Verbeek, 'Learning nonlinear image manifolds by global alignment of local linear models', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **28**(8), 1236–1250, (2006).

[35] M-H Yang and R. Li. Monocular tracking of 3d human motion with a coordinated mixture of factor analyzers, 2008. US Patent 7,450,736.

[36] S. Yi, L. Zhang, R. Jin, Q. Qian, and A. Jain, 'Semi-Supervised Clustering by Input Pattern Assisted Pairwise Similarity Matrix Completion', in *International Conference in Machine Learning*, pp. 1400–1408, (2013).