Application of SsVGMM to Medical Data - Classification with Novelty Detection

Fan Yang, Jaymar Soriano, Takatomi Kubo, *Member, IEEE*
Kazushi Ikeda, *Senior Member, IEEE*

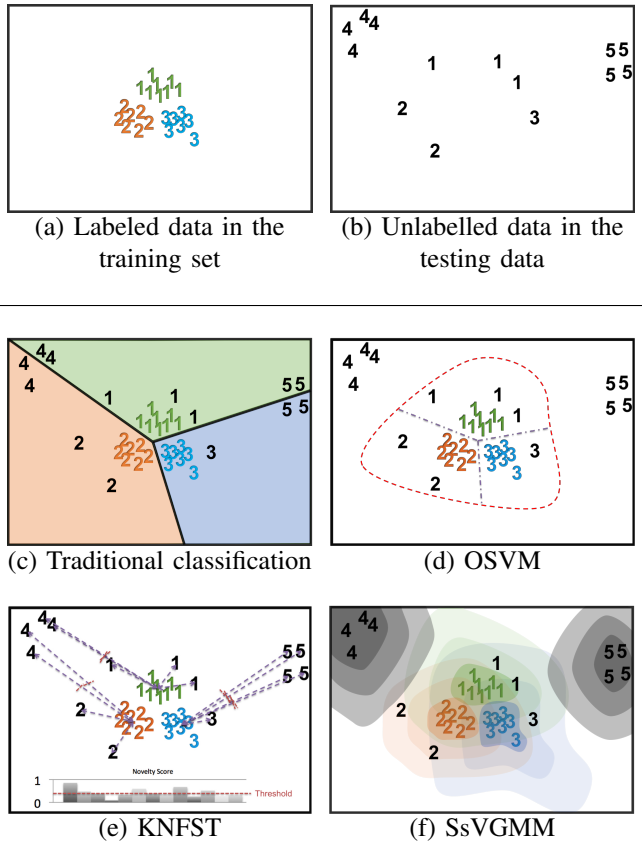
***Abstract*— There is a huge demand to apply classification in medical analysis. A traditional classifier requires having training samples from each class. However, in reality, it is possible that the testing set may include classes that are not in the training set. This inevitably causes an issue: data from a undefined class will be assigned to a predefined classes. To tackle this, we propose a semi-supervised variational Gaussian mixture model to perform multi-class classification with novelty detection. Comparing to some popular novelty detection methods, we demonstrate that it gets better performance on a thyroid disease data, by generating the distribution of predefined classes and undefined class, without explicitly setting a threshold.**

I. Introduction

Classification tasks are commonly performed in medical data analysis in which unknown labels (e.g., categories of disease) are predicted from observed features (e.g., clinical records). Although numerous classifiers have been developed, classification with novelty detection has attracted less attention. Using a traditional classifier, when the class of a testing data is absent in the training set (undefined class), this data will be miss-classified to one of classes represented in the training set (predefined classes). This usually results to high recall but low precision. For instance, the training set in Fig. 1(a) contains only samples from three predefined classes (i.e., class 1, class 2 and class 3); however, two other classes (i.e., class 4, class 5), are introduced in the testing set (Fig. 1(b)). When a traditional classifier is applied, decision boundaries will divide the feature space into three parts, thus, the data from previously undefined class will be assigned to one of the predefined classes (Fig. 1(c)). Despite the 100% recall for class 1, the precision is only around 43%.

Generally, to tackle this problem, novelty detection [1] is applied. Among various approaches, one-class support vector machine (OSVM) [2] is commonly used. In previous studies, OSVM has been applied to different medical data analysis [3-6]. OSVM generates soft boundary around predefined classes. By selecting a threshold to decide a global boundary, data either belongs to predefined classes or undefined class are determined (Fig. 1(d)). Nevertheless, how to select a proper threshold remains to be a challenge. Without specific knowledge of the observation domain, a fine-tuned OSVM may work in one application but fail in another. Furthermore, for multiple classes, OSVM performs novelty detection by considering all predefined classes globally. A global boundary, however, may not suitable for each class exactly, and it may affect the classification performance. There are novelty detection methods that consider each predefined class seperatelly, one example is Kernel Null Foley-Sammon transform (KNFST) [3]. By measuring the smallest distance between a test sample and observed classes in a projection space, the test sample gets a novelty score (Fig. 1(e)). Despite the novelty score is calculated locally, the threshold for setting the decision boundary is not specified for each class, and it may still affect the classification performance.

In this paper, we propose a semi-supervised variational Gaussian mixture model (SsVGMM), which simultaneously models the distributions of predefined classes and undefined class using Gaussian mixture model (GMM). From the probability density of each class, probabilistic boundaries are generated, without explicitly setting a threshold (Fig. 1(f)).



(a) Labeled data in the training set
(b) Unlabelled data in the testing data
(c) Traditional classification
(d) OSVM
(e) KNFST
(f) SsVGMM

Fig. 1: Classification with undefined class.

All authors are with the Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara 630-0192, Japan.
E-mail: {yang.fan.xv6, jaymar.soriano.jk2, takatomi-k, kazushi}@is.naist.jp

Using a two-dimensional synthetic data, we illustrate how probabilistic boundaries are generated and the classification with novelty detection is performed by SsVGMM. Using a thyroid disease data, we compare the performance of KNFST, OSVM and SsVGMM, and demonstrate the benefit of using SsVGMM to get probabilistic boundaries.

II. Semi-supervised Variational Gaussian Mixture Model

SsVGMM is a semi-supervised extension of variational mixture of Gaussians, taking advantage of inferring the number of Gaussian components from the data [4]. Given an initialised number of components K (theoretically could be infinite, we set $K = 100$ in practice), SsVGMM automatically shrinks K to a proper value based on the data density. This flexibility helps SsVGMM use a reasonable number of Gaussian components to approximate the distribution of each class. Moreover, through semi-supervised learning, both the labelled and unlabelled data are utilised in the model inference.

The implementation of SsVGMM is mainly based on the variational mixture of Gaussians, but a 'cannot-link' constraint [5] is introduced for the semi-supervised learning. Essentially, it is a semi-supervised clustering method without the 'must-link'. In the inference process, due to the 'cannot-link' constraint, samples with different labels cannot be assigned to the same Gaussian component, while samples with the same label can be assigned to different Gaussian components. Therefore, data with different labels can be separated, and data with the same label will not constrain the flexibility of Gaussian mixture model.

Given $X = \{x_1, ..., x_N\}$ $(X \in \mathbb{R}^{N \times D})$ as the combination of the training dataset X_{train} and testing dataset X_{test}. Observed labels are denoted by Y_{train} and the unknown labels are denoted by Y_{test} corresponding to X_{train} and X_{test}, respectively. The coordinate-ascent variational inference (CAVI) is used to optimize the model parameters, which mainly includes two steps: (i) VB Expectation-step and (ii) VB Maximization-step. Responsibilities of data assigned to Gaussian components, which are denoted by $\boldsymbol{r}$, are evaluated in (i). Other variational parameters, including the set of Gaussian means and covariances which are denoted by $\boldsymbol{\theta}$, are updated in (ii). An evidence lower bound $\mathcal{L}$ is used to decide when to terminate CAVI. The output is Y_{test}. In the inference process, the labels assigned to Gaussian components are just index without specific meaning. Therefore, referring to Y_{train}, we map predefined labels to these assigned Gaussian components. Some assigned Gaussian components do not get predefined labels, then they are all labeled as undefined class. The pseudocode of SsVGMM is shown in Algorithm 1.

Algorithm 1: SsVGMM

input : X_{train}, Y_{train}, X_{test}
output: Y_{test}

Combine $X \leftarrow [X_{train}, X_{test}]$.
Initialise variational parameters.
while $\mathcal{L}$ *has not converged* **do**
    VB E-step: evaluate $\boldsymbol{r}$ *using* X*;*
    Perform 'Cannot-link' constraint using Y_{train}*;*
    VB M-step: compute $\boldsymbol{\theta}$ *using* X*;*
    Compute $\mathcal{L}$.
end
Assign data to Gaussian component via maximum $\boldsymbol{r}$.
Map predefined label (from Y_{train}*) to assigned Gaussian components.*
The assigned Gaussian components without predefined label are labeled as undefined class.
Generate Y_{test}.

III. Experiments

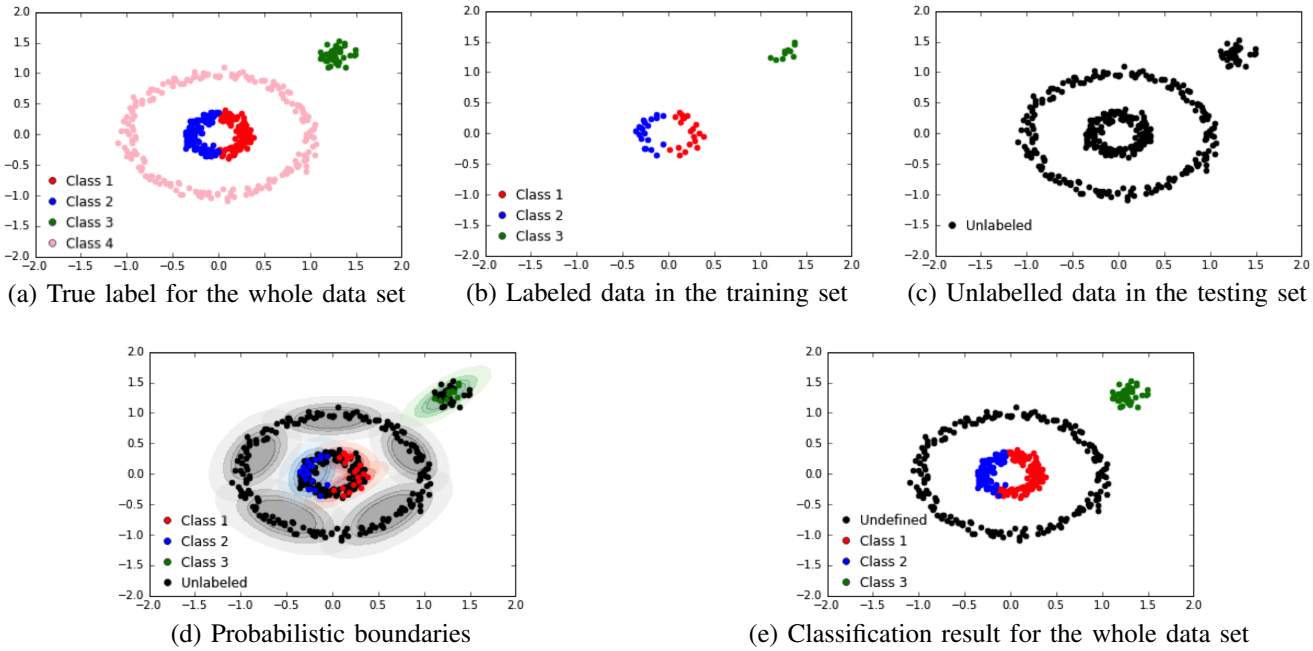
A. Synthetic data

We generated a total of 450 samples, from Class 1 (100 samples), Class 2 (100 samples), Class 3 (50 samples) and Class 4 (200 samples) (Fig. 2(a)). The training set only contains 20% of samples from Class 1 - 3, while the other 80% of them, together with 100% of Class 4, are in testing set (Fig. 2(b)-(c)).

SsVGMM utilizes samples from both training set and testing set to generate the probabilistic boundaries. Referring to the distribution of labelled and unlabelled samples, SsVGMM uses two Gaussian components to approximate the distribution of Class 1, and one Gaussian component each for Class 2 and Class 3 (Fig. 2(d)). Based on the pattern of the whole data set and observed labels, SsVGMM supposes these unlabelled samples from Class 4 do not belong to any predefined classes. Therefore, they are assigned to a undefined class, and five Gaussian components are used to approximate its distribution. The classification result is shown as Fig. 2(e), where we choose the label for each sample via the maximum probability of the probabilistic boundaries.

B. Thyroid disease data

A thyroid disease data [6] is used for comparing the performance of KNFST, OSVM and SsVGMM in multi-class classification with novelty detection. This data set includes three classes: normal thyroid, hyperthyroidism and hypothyroidism, and corresponding number of samples are 150, 35 and 30, respectively. Each sample consists of five diagnostic features, namely T3-resin uptake test, total serum thyroxine as measured by the isotropic displacement method, total serum triiodothyronine as measured by radioimmunoassay, basal thyroid-stimulating hormone (TSH) as measured by radioimmunoassay, and maximal absolute difference of TSH value after injection of 200 μg of thyrotropin-releasing



(a) True label for the whole data set (b) Labeled data in the training set (c) Unlabelled data in the testing set

(d) Probabilistic boundaries (e) Classification result for the whole data set

Fig. 2: Novelty detection on synthetic data by SsVGMM

hormone as compared to the basal value. All feature values are normalised to have zero mean and unit variance before performing classification. Although the classification is performed in the five-dimensional feature space, for ease of visualisation, we plot the classification results in a two-dimensional space using the first and second principal components obtained via PCA.

Suppose the labelled data only contains 75 samples of normal thyroid and 15 samples of hypothyroidism, while a batch of unlabelled records contain samples of hyperthyroidism (see Table I). KNFST, OSVM and SsVGMM are then applied and their results are compared against each other.

	Normal	Hyper	Hypo
Training	75	0	15
Testing	75	35	15

TABLE I: The training and testing set of Thyroid disease data

As the classification results show, all of methods are able to detect samples from undefined class. With fine-tuning, both KNFST and OSVM are able to detect the undefined class which was not represented in the training set. However, since their novelty detection boundary do not fully fit with each class, part of samples belonging to hypothyroidism were miss-classified as undefined class. In contrast, SsVGMM yields high precision and recall for predefined classes (normal and hypothyroidism), by generating probabilistic boundaries. In addition, SsVGMM gives the approximate distribution of the undefined class which can be used for further analysis. For instance, by further studying the distribution property, we may actually find that this undefined class is hyperthyroidism.

Nevertheless, the results should be interpreted with the following limitations in mind. First, in the given feature space, unlike this thyroid disease data, samples from different classes are heavily overlapping in some medical data, novelty detection will be significantly affected. Therefore, performing feature selection and transformation may be needed. Second, as the feature dimension increases, SsVGMM optimization is more likely to take much longer time to converge to a local optimum. It is recommended to perform dimension reduction first.

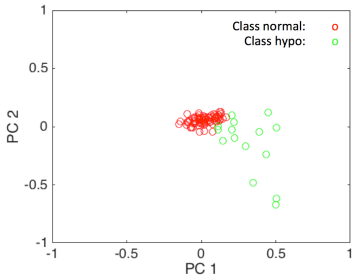
IV. Conclusion

To tackle the problem that the testing set may include classes that are not represented in the training set, SsVGMM was proposed to perform multi-class classification with novelty detection. SsVGMM simultaneously models the distributions of predefined classes and undefined class, generating probabilistic boundaries. As a benefit, we remedy the effect of the novelty detection to the multi-class classification, such that SsVGMM may give better performance on classification. We suppose that SsVGMM will find a wide range of application on medical data when multi-class classification with novelty detection is needed.

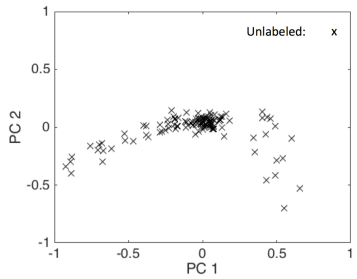
References

[1] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

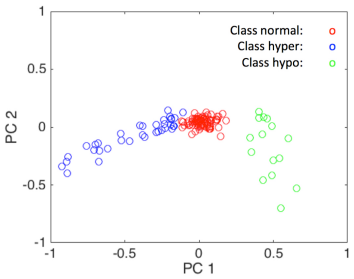
[2] H. J. Shin, D.-H. Eom, and S.-S. Kim, "One-class support vector machines - an application in machine fault detection and classification," *Computers & Industrial Engineering*, vol. 48, no. 2, pp. 395–408, 2005.



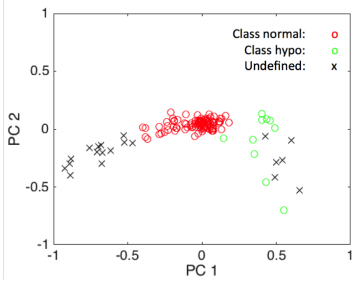
(a) Labeled data in the training set



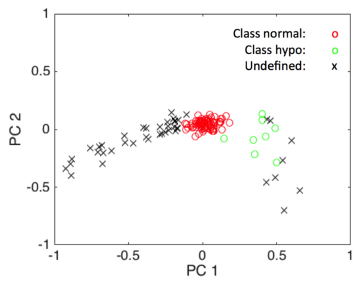
(b) Unlabelled data in the testing set



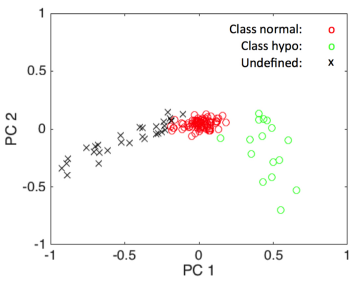
(c) True label of the testing set



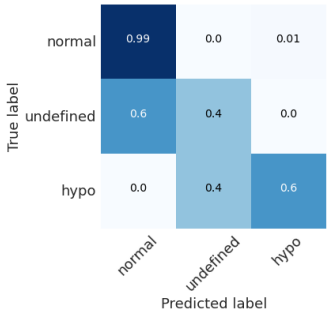
(d) Fine-tuned KNFST



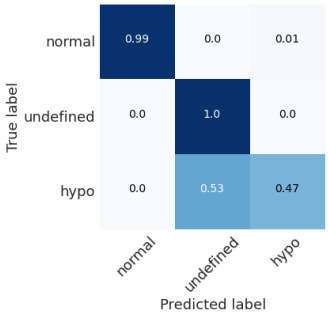
(e) Fine-tuned OSVM



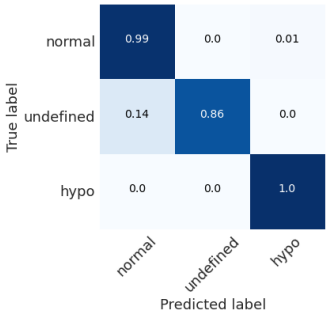
(f) SsVGMM



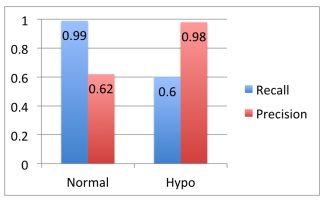
(g) Confusion matrix from KNFST



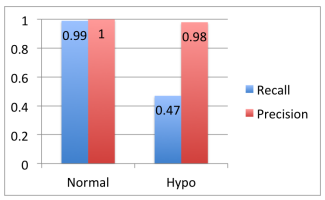
(h) Confusion matrix from OSVM



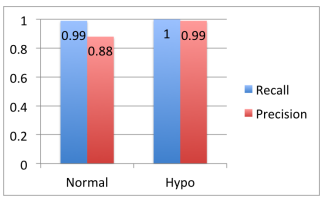
(i) Confusion matrix from SsVGMM



(j) Recall and precision (KNFST)



(k) Recall and precision (OSVM)



(l) Recall and precision (SsVGMM)

Fig. 3: Novelty detection on thyroid disease data

[3] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler, "Kernel null space methods for novelty detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3374–3381.

[4] C. Bishop, "Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn," *Springer, New York*, 2007.

[5] S. Basu, M. Bilenko, A. Banerjee, and R. J. Mooney, "Probabilistic semi-supervised clustering with constraints," *Semi-supervised learning*, pp. 71–98, 2006.

[6] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml