Selective Sampling for Combined Learning from Labelled and Unlabelled Data

Lina Petrakieva, Bogdan Gabrys
Applied Computational Intelligence Research Unit,
School of Information and Communication Technologies,
University of Paisley, UK
Phone: +44-141-848 3284, +44-141-848 3752,
E-mail: [petr-ci0,gabr-ci0]@wpmail.paisley.ac.uk

Abstract: *This paper examines the problem of selecting a suitable subset of data to be labelled when building pattern classifiers from labelled and unlabelled data. The selection of representative set is guided by a clustering information and various options of allocating a number of samples within clusters and their distributions are investigated. The experimental results show that hybrid methods like Semi-supervised clustering with selective sampling can result in building a classifier which requires much less labelled data in order to achieve a comparable classification performance to classifiers built only on the basis of labelled data.*

Keywords: Combined Learning, Labelled and Unlabelled Data, Clustering, Selective Sampling.

1 Introduction

In many domains labelled data is difficult or expensive to obtain as it may require manual work or be computationally expensive. Therefore it is not surprising that there has been much interest in applying techniques that incorporate knowledge from unlabelled data into a supervised learning system [4, 5, 6, 9, 11, 12, 13, 14]. The task is to minimize the overall cost of the classification process, which depends both on the classifier accuracy and the cost of obtaining labelled data as discussed in [2].

In a number of publications discussing hybrid methods for coping with labelled and unlabelled data, the use of additional unlabelled data has been shown to offer improvements in comparison to classifiers generated only on the basis of labelled data [4, 5, 6, 9, 11, 13]. However it was not clear whether the improved performance of the classifiers supplemented by unlabelled data was mainly due to representativeness of the original labelled set or to the specific combined methods.

In order to answer this question we had conducted extensive experimental analysis of different combined approaches (shortly summarised in Section 4) for different ratios of labelled to unlabelled samples [5]. One of the main findings of that study was that the final classification performance depends more on the specific labelled subset used rather than on the combined classification method. Additionally, as a result of random selection of samples to be labelled from the initial pool of unlabelled data, a high variability of the classifier performance was commonly observed.

In this paper a continuation of the previous study is presented with a focus on methods for selecting samples to be labelled rather than choosing them randomly. It is hoped that a suitable selection of samples to be labelled could reduce the variance of the final solutions and improve mean classification performance obtained from random selection.

The rest of the paper is organised as follows: In Section 2 a general problem statement is given. This is followed by description of two selective sampling methods in Section 3 and a summary of the investigated combined methods in Section 4. Section 5 presents the experimental results. And finally conclusions are given in the last section.

2 General Problem Statement

Let $D = \{L, U\}$ be the training data set with $L = \{(\mathbf{x}_i, t_i) \mid i = 1 \dots M\}$, representing a set of M labelled samples and $U = \{(\mathbf{x}_j, 0) \mid j = 1 \dots N\}$, representing a set of N unlabelled samples where $\mathbf{x} = (x_1, x_2, \dots, x_n) \in R^n$ is an n-dimensional feature vector and $t \in \{1, \dots, p\}$ is a class label representing one of p classes with 0 used to denote an unlabelled sample. As in the conventional cases of designing a classifier on the basis of a training data set the main goal is to find a function transforming a feature vector $\mathbf{x}$ into one of the p classes, which can be formally written as:

$$C_D : \mathbf{x} \rightarrow t \qquad \text{or} \qquad t = C_D(\mathbf{x}) \tag{1}$$

where C_D is a classifier C designed on the basis of the data set D. However, depending on the ratio

$$r = \frac{M}{(M + N)} \tag{2}$$

of the labelled samples to the total number of samples in D the problem ranges from the pure supervised learning for $r = 1$ to the pure unsupervised learning for $r = 0$. In [5] various hybrid methods for coping with cases for $r \in (0, 1)$ with random selection of the labelled samples, in order to obtain L have been examined.

But since the use of specific labelled subset has been found to be of crucial importance our main effort in this study will concentrate on selecting representative samples to be labelled rather than creating sophisticated methods for coping with both types of data.

3 Selective Sampling Methods

In the context of pattern classification systems selective sampling techniques have been most frequently used in active learning approaches [8], where samples for labelling are selected in a dynamic manner (one at a time). In the research presented here the preliminary selection techniques will be examined. In contrast to the active selection, the preliminary selection operates on the basis of selecting whole batches of data to be labelled (i.e. all M samples forming the labelled subset L).

Trying to find representative samples when working with unlabelled data means that one has to make decisions based only on clustering information. If the clusters are already available one needs just to select the samples from the clusters. However, the immediate question is: How many samples and from which clusters? The following two distinctive approaches of allocating the number of samples per cluster have been investigated: a) *proportional* allocation - samples for labelling are allocated proportionally to the cardinality of the cluster which means more samples for bigger clusters and some of the small clusters may have no samples selected; b) *consecutive* allocation - samples for labelling are allocated uniformly disregarding clusters' sizes. Furthermore, the actual selection of the samples to be labelled per cluster have been done in two ways - a) by selecting cluster prototypes and b) by trying to describe a cluster by selecting samples close to its boundary. A short description of both methods is presented bellow.

Cluster Mean Selection - The subset of samples to be labelled is created from the prototypes of the clusters of the dataset. The prototypes are selected as the closest samples to the means of the clusters. If there are more than one sample per cluster to be selected the clusters are divided into subclusters and their prototypes are selected. In case when the clusters are predefined the number of data points for labelling per cluster has to be calculated as discussed above. Then the process of selecting the samples can be applied to each cluster separately. If there are no defined clusters the whole dataset can be considered as one cluster or the dataset can be divided into b clusters where b

is the number of samples to be selected.

Cluster Boundary Selection - The process of selection begins with a set of randomly picked b samples. Then the algorithm is optimising this initial set by removing from it the samples that are too close to each other and by selecting outermost samples. Thus by maximizing the minimum distance between the selected data points the algorithm is selecting them around the boundary of the cluster. If there are many samples to be selected the method is placing some of them at the boundary and when they become too close to each other it is selecting the rest of the samples spread within the cluster.

4 Combined Learning Methods

Once the labelled subset has been generated, either using random selection or the approaches discussed in the previous section, one of a number of combined learning algorithms can be used to design a classifier. A short description of the combined approaches used in the experiments and formally defined in [5] follows.

The first most obvious way of dealing with situations when labelled and unlabelled data is available is to ignore the unlabelled data (referred to as Labelled Only method) and build the classifier C_L using just the labelled subset L from D completely ignoring U. The classification process from the Eq.(1) in this case becomes:

$$t = C_L(\mathbf{x}) \tag{3}$$

Labelled Only method is implemented for comparison purposes. However, since using the unlabelled data can be advantageous we used various combined learning methods falling into one of the following three major groups (see [5] for more details):

Pre-labelling approaches - a) Static methods - The first of the considered approaches to utilising the unlabelled data, referred to as Static Labelling approach in the later sections, is based on generating an initial classifier on the basis of the labelled data only C_L and labelling the remaining unlabelled data U by applying the initial classifier so that W is the newly labelled set U. Finally the classifier is redesigned using both the original L and the newly labelled W data sets; b) Dynamic methods - This approach is a modification of the above whereas an initial classifier is generated on the basis of the labelled data only C_L but the unlabelled data U are iteratively labelled one sample at a time. The newly labelled sample is added to the pool of labelled data and the classifier is redesigned at each step. The samples which can be most confidently classified are chosen first. The process is continued until all unlabelled samples have been labelled and the final classifier obtained. This will be referred to as a Dynamic Labelling approach.

Post-labelling approaches - Majority Clustering - The considered method is based on clustering all the data and using the labelled data for labelling the whole clusters by applying the majority principle i.e. the label of the cluster is assigned on the basis of the largest number of samples from a given class represented in the cluster. We will refer to this method as the Majority Clustering method.

Semi-supervised approaches - Semi-supervised Clustering - The detailed algorithm is presented in [5]. In contrast to the standard clustering the labels are actively used for guiding the clustering process. In result the algorithm is more robust in the sense of the number of created clusters and their sizes which to a large extent is dependant on the relative placement of the labelled samples in the input space. The idea is to split the initial clusters until there is an overwhelming presence of one type of labelled samples in each of newly created sub-clusters. After splitting the clusters the labelling process is carried out using Majority Clustering method.

5 Experimental Results and Analysis

Methods described in Section 4 are used in selective sampling experiments. As described in Section 3, both proportional and consecutive distribution as well as Mean and Boundary Selection methods are used. The nearest neighbour (NN) and pseudo-fisher support vector (PFSV) classifiers implemented in [3] have been used as the base classifiers for labelling and testing purposes. A complete-linkage hierarchical clustering has been used for Majority Clustering ([5]) and Semi-supervised Clustering ([5]) with the shortest Euclidean distance adopted for the cluster similarity measure. The user defined parameter Θ used in the Semi-supervised Clustering has been set to 0.3.

Due to the space limitations only selected results are presented for one artificial (Normal mixtures) and one real (Glass) datasets. Normal mixtures dataset (available at www.stats.ox.ac.uk/ ∼ripley/PRNN) represents a two dimensional problem with two highly overlapping classes. The training set consists of 250 samples and a separate testing set has 1000 samples. This dataset has been constructed in such a way as to allow the best possible performance of around 8%. The Glass dataset [1] consists of 214, 10 dimensional samples representing 6 classes of different glass types found at the crime scenes and used during forensic investigations. Testing for Glass dataset is performed using 5-fold cross-validation.

The experiments have been performed for different ratios r of labelled data to the total number of data samples ranging from virtually unlabelled sets ($\sim 0\%$ of labelled data is one sample per class for ”random per class” selections and 3 samples for selective sampling methods) to the fully labelled data sets (100% of labelled data). For the random selection the experiments have been repeated many times at each level for collecting reliable statistical information. The same sets of labelled samples have been used in all the experiments with different classification methods. Two types of random selection have been performed - a) random selection but ensuring that at least one sample per class is selected (referred to as ”random per class”) and b) completely random selection. The results for random selection are compared with the results for selective sampling methods.

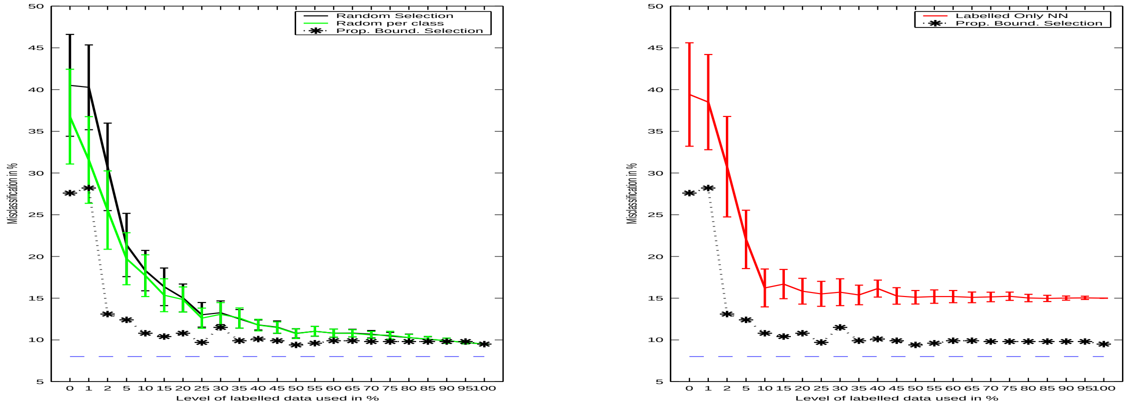

Figure 1: Normal mixtures dataset - Left: Semi-supervised clustering - mean classification error and standard deviation of Random selection method compared to Random selection per class and to Proportional Boundary selection; Right: Proportional Boundary selection v Labelled Only NN.

As illustrated in the left parts of Fig.1 and Fig.2 the combined methods using selective sampling have shown an improved performance in comparison to completely random selection methods. This is especially evident for small values of r. However, it can also be noted (Fig.2 left) that the prior information about the number of classes used in the ”random per class” selection method for the Glass dataset resulted in much better performance for small r than when using selective sampling where no information about the number of classes is used. This is common feature in multiclass problems

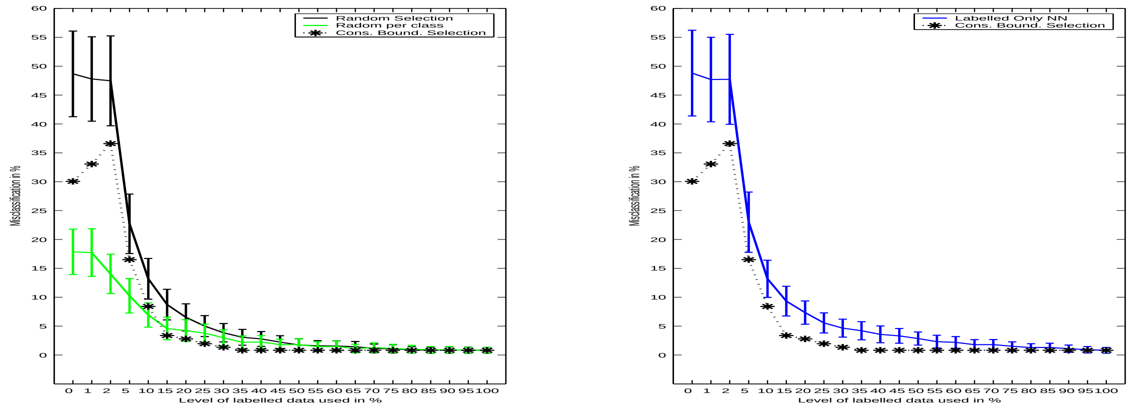

Figure 2: Glass dataset - Left: Semi-supervised clustering - mean classification error and standard deviation of Random selection method compared to Random selection per class and to Proportional Boundary selection; Right: Proportional Boundary selection v Labelled Only NN.

with uneven distribution (prior class probabilities) of samples from different classes. The right parts of Fig. 1 and Fig. 2 illustrate the better performance when using selective sampling together with Semi-supervised clustering in comparison to classifiers generated on the basis of labelled data only selected randomly. In all the cases a high classification errors are observed when only a very limited number of labelled data is used (small r). No consistent significant difference have been noted when comparing the Boundary Selection with Mean Selection and/or consecutive and proportional allocation methods. The results depend on suitable choice of the number of clusters for different levels of labelled data. In general, the better results have been obtained when using smaller number of clusters for small r and increased number of clusters with an increase of available labelled samples.

In the patterns of change illustrated in Fig.1 and Fig. 2 the level after which there is no significant improvement of the classifier performance is referred to as a *sufficient* level (SLLS - Sufficient Level of Labelled Samples). This level is different for different datasets. In general, the more complex the dataset distribution, the more labelled samples the algorithm needs to describe it so the SLLS will be at a higher ratio r, i.e. when more labelled samples are used. This level indicates that generally much less labelled data is needed for constructing a reliable classifier when unlabelled data is used in addition. When using selective sampling methods the stable performance related to the the SLLS is often achieved at lower values of r compared to the random sampling methods.

6 Conclusions

The random sampling methods analysed in [5] show that selection of a representative labelled subset is more important than combining learning from labelled and unlabelled data. Therefore in this paper we have concentrated our investigations on selective sampling methods. The preliminary results presented here indicate an improvement of both the mean classifier performance and reduction of the classification variance when using selective sampling methods in comparison to random selection of samples to be labelled.

A distinct disadvantage of the discussed methods is that they assume preliminary selection. The algorithms used that way cannot take advantage of any available class information in contrast to the active learning approaches. Therefore our future research will extend to active learning as an alternative to overcoming the disadvantages of the preliminary selection methods presented in this paper.

References

[1] C.L. Blake, C.J. Merz, UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998;

[2] R.O.Duda, P.E.Hart, D.G.Stork, *Pattern Classification*, 2^{nd} edition, A Wiley-Interscience Publication, 2001;

[3] R.P.W. Duin, Pattern Recognition Tools for Matlab, ftp://ftp.ph.tn.tudelft.nl/pub/bob/prtools/, 2000;

[4] B. Gabrys, A. Bargiela, "General Fuzzy Min-Max Neural Network for Clustering and Classification", *IEEE Transactions on Neural Networks*, Vol. 11, No. 3, pp. 769-783, 2000;

[5] Bogdan Gabrys, Lina Petrakieva, "Combining labelled and unlabelled data in the design of pattern classification systems", *European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, ISBN 3-89653-919-1, pp.441-449, Albufeira, Portugal, 2002;

[6] S. Goldman, Y. Zhou, "Enhancing Supervised Learning with Unlabeled Data", In *Proceedings of the Seventeenth ICML*, June 2000;

[7] T. Hofmann, J.M.Buhmann, "Active Data Clustering", *NIPS*, pp.528-534, 1997;

[8] V.S. Iyengar, C. Apte, T. Zhang, "Active Learning using Adaptive Resampling", *ACM SIGKDD*, 2000;

[9] A. Klose, R. Kruse, "Enabling Neuro-Fuzzy Classification to Learn from Partially Labeled Data", *IEEE World Congress on Computational Intelligence, IEEE International Conference on Fuzzy Systems*, Honolulu, HI, USA, pp.4232-4238, 2002;

[10] R. Kothari, V. Jain, "Learning from Labeled and Unlabeled Data", *IEEE World Congress on Computational Intelligence, IEEE International Joint Conference on Neural Networks*, Honolulu, HI, USA, pp.1468-1474, 2002;

[11] J. Larsen, A. Szymkowiak, L.K. Hansen, "Probabilistic Hierarchical Clustering with Labeled and Unlabeled Data", *International Journal of Knowledge-Based Intelligent Engineering Systems*, vol. 6, no. 1, pp. 56-62, January 2002;

[12] T.M. Mitchell, "The Role of Unlabeled Data in Supervised Learning", *Proceedings of the Sixth International Colloquium on Cognitive Science*, Spain, 1999;

[13] K. Nigam, R. Ghani, "Analyzing the Effectiveness and Applicability of Co-training", In *Ninth International CIKM*, pp. 86-93, 2000;

[14] W. Pedrycz, J. Waletzky, "Fuzzy Clustering with Partial Supervision", *IEEE Transactions on Systems, Man and Cybernetics - Part B:Cybernetics*, Vol.27, No 5, pp.787-795, October 1997;

[15] S. Theodoridis, K. Koutroumbas, *Pattern Recognition*, Academic Press, 1999.