Combining labelled and unlabelled data in the design of pattern classification systems

Bogdan Gabrys and Lina Petrakieva
Applied Computational Intelligence Research Unit, Computing and Information Systems,
School of Information and Communication Technologies,
Faculty of Communications, Engineering & Science, University of Paisley, UK,
Phone: +44141 848 3752, Fax +44 141 848 3542,
email: Bogdan.Gabrys@paisley.ac.uk, Lina.Petrakieva@paisley.ac.uk

ABSTRACT: There has been much interest in applying techniques that incorporate knowledge from unlabelled data into a supervised learning system but less effort has been made to compare the effectiveness of different approaches on real world problems and to analyse the behaviour of the learning system when using different amount of unlabelled data. In this paper an analysis of the performance of supervised methods enforced by unlabelled data and some semi-supervised approaches using different ratios of labelled to unlabelled samples is presented. The experimental results show that when supported by unlabelled samples much less labelled data is generally required to build a classifier without compromising the classification performance. If only a very limited amount of labelled data is available the results show high variability and the performance of the final classifier is more dependant on how reliable the labelled data samples are rather than use of additional unlabelled data. Semi-supervised clustering utilising both labelled and unlabelled data have been shown to offer most significant improvements when natural clusters are present in the considered problem.

KEYWORDS: Combined learning methods, Supervised learning, Unsupervised Learning, Semi-supervised Clustering, Learning from Labelled and Unlabelled Data, Pattern Classification

I. INTRODUCTION

Learning system is a system that makes decisions based on the accumulated experience contained in available solved cases. Two major problems arise from this statement: Do we always have solved cases? and; How much labelled data do we need to be sure that the system will have acceptable performance?. Based on the availability of the labelled training data we can divide learning methods into two major groups – supervised learning (available labelled data) and unsupervised learning. Though supervised approaches based on fully labelled training sets can lead to constructing very well performing classification systems, in real-world problems labelling the data can be both time consuming and expensive. On the other hand, unlabelled data is often readily available but pure unsupervised (clustering) techniques very rarely result in building accurate classifiers.

It is therefore not surprising that there has been much interest in hybrid techniques that can learn both from labelled and unlabelled data [1]-[4], [7]-[18], [20]. The most frequently cited motivation for such combination is a hope that a better performing classifier could be constructed in comparison to the case when only a limited labelled data were to be used. And though there have been a number of different methods proposed, which use techniques from diverse fields, they can be categorised into one of the following three major approaches spanning a spectrum of methods between fully supervised and fully unsupervised learning:

1) A set of labelled data is used for designing an initial classifier, which is then used for labelling of the remaining unlabelled data. Once this is done a classifier is constructed on the basis of both the original and newly labelled data [4],[13]-[16]. In [4], a self organising map (SOM) neural network is first used for generating a classifier by clustering labelled data only and subsequent labelling of the unlabelled data by applying it to the generated model. The extended labelled set is then used for training a multilayer perceptron classifier. In [2],[10],[13]-[16], few versions of ‘co-training’ algorithm, which has been especially popular for document classification tasks, are presented. The basic version of co-training algorithm uses two classifiers (e.g. naïve Bayes classifiers) trained using different mutually exclusive subsets of the input features from the labelled samples. The labelling is then carried out sequentially by each classifier choosing the unlabelled data sample that can be classified with the highest confidence and adding it to the current pool of labelled data. Combination of co-training with Expectation Maximization (EM) algorithm and a version not requiring the

feature split are discussed in [16]. Another variant using two different classifiers within the same paradigm is discussed in [10]. In this case there is a problem of selecting the right classifiers, as the diversity of the classifiers is crucial for the performance of the classifiers' fusion [19].

2) A data model is generated from all available data, which is usually accomplished by applying a data density estimation procedure or clustering algorithm. The labels are then subsequently used for labelling whole clusters of data or estimating class conditional densities which involves labelling of the unlabelled data dependant on their relative placement in the data space with respect to the original labelled data [9][12]. Any of a large number of clustering algorithms could be used in the first stage. Labelling of the samples is usually based on counting the number of labelled samples representing specific classes within each of the clusters. The probabilistic framework utilising data density estimation based on a mixture of gaussians or Parzen windows has also been used for learning from labelled and unlabelled data. The general approach to dealing with missing data within EM algorithm is discussed in [9]. While in [12], a combination of labelled and unlabelled data is accomplished with Parzen windows used for estimation of class conditional distribution and a genetic algorithm (GA) employed for maximizing a posteriori classification of the labelled patterns.

3) Semi or partially supervised clustering in which both labelled and unlabelled data are processed at the same time [7],[8],[18]. In this approach, falling somewhere between 1) and 2), the clustering process is not only based on a suitably chosen similarity measure but is also guided/constrained by the labelled data. A general fuzzy min-max (GFMM) neural network is an example from this group [7],[8]. Both labelled and unlabelled samples are processed in an iterative manner for adaptation and labelling of hyperbox fuzzy set based clusters. In [18], a partially supervised fuzzy clustering based on optimisation of an objective function is proposed. The use of labels is facilitated by suitably modifying a standard objective function of fuzzy ISODATA clustering algorithm. A somewhat different method falling into this group is presented in [3], where a user acting at the meta level can control the process of clustering the documents by adding constraints and label-like information.

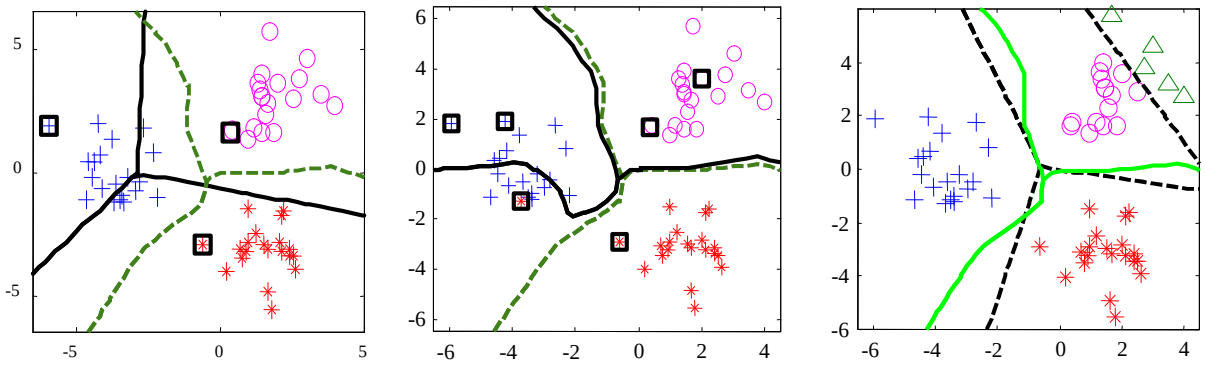

Figure 1. Illustration of generated decision boundaries by employing different algorithms combining labelled and unlabelled data: a) LEFT: Solid line – decision boundary based on using labelled data only, Dashed line – decision boundary with the labelled data supplemented by unlabelled data and dynamic labelling approach; b) MIDDLE: The influence of mislabelled sample on the generated decision boundaries – Solid line – decision boundary generated on the basis of static labelling of the unlabelled data set, Dashed line – decision boundary generated when using clustering algorithm; c) RIGHT: The boundaries generated by standard clustering and automated labelling of the samples in a cluster, Dashed, black line – decision boundary generated using semi-supervised clustering.

In all of the above discussed methods the use of additional unlabelled data has been shown to offer improvements in comparison to the classifiers generated only on the basis of limited labelled data set. However, in some of them a number of potential problems have also been noted.

Some of these problems and potential for using unlabelled data are illustrated in Fig. 1 representing a relatively simple case with three clearly separable clusters of data. In Fig. 1a and b each of the clusters represents a class. In Fig 1a case with only three labelled samples (depicted as squares) is shown. It can be seen that if only the labelled data is used the decision boundary (solid line) is far from optimal. Labelling in a dynamic fashion (i.e. as suggested in co-training etc.) or using a clustering algorithm would clearly be beneficial as illustrated by a much better decision boundary shown as dashed line. However, if different data samples were labelled even in this case the solution could be much better. This very simple example is indicative of a much more serious problem when the very limited labelled data is not representative of the underlying distribution or as illustrated in Fig. 1b when there are noisy or mislabelled samples in the labelled data set. The third case shown in Fig. 1c represents a problem of disproportional representation of two different classes especially when they are not clearly separable as in the previous cases. It is quite easy in such cases to

discard a minority class (represented by triangles in Fig. 1c) if the overwhelming labelled cases are from the majority class (represented by circles). A semi supervised clustering algorithms could be quite successfully used in such a case while standard clustering methods would have difficulties in distinguishing between the two classes since they would normally be treated as one cluster.

One additional problem with a vast majority of the reported results in the literature is that it is quite difficult to compare how effective different proposed methods are at using the unlabelled data since they usually concentrate on one specific problem with a one small set of labelled data. It is often argued that in real problems like document classification, a limited set of labelled documents is given and can only be supplemented by varying amount of unlabelled data. Nevertheless, or because of this fact, it is not clear whether the improved performance of the classifier supplemented by unlabelled data is mainly due to the representativeness of the original labelled set or to the proposed method for handling both kinds of data.

Therefore one of the main goals of this investigation was to carry out a systematic analysis of the performance of various algorithms representing all of the major approaches mentioned earlier in the introduction and described in the following sections

The remaining of this paper is organised as follows. In the second section a formal problem statement with the required notations to be used in the rest of the paper is provided. The third section will use the introduced notation for formal description of five different approaches to handling labelled and unlabelled data in pattern classification problems. This will be followed by experimental results and comparative analysis for four different, non-trivial classification data sets including two highly overlapping synthetic data sets and two well known data sets obtained from the repository of machine learning databases. Finally the conclusions will be presented.

II. PROBLEM STATEMENT AND NOTATION

Let $D=\{L,U\}$ be the training data set with $L=\{\mathbf{x}_i,t_i\}$, *i=1..M,* representing a set of M labelled samples and $U=\{\mathbf{x}_j,0\}$, *j=1..N,* representing a set of N unlabelled samples where $\mathbf{x}=(x_1,x_2,...,x_n)\in R^n$ is an n-dimensional feature vector and $t\in\{1,...,P\}$ 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 ***x*** into one of the *P* classes, which can be formally written as:

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

where C_D is a classifier *C* designed on the basis of the data set *D*.

However, depending on the ratio $r=M/(M+N)$ 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 the following sections the hybrid methods for coping with cases for $r\in(0,1)$, which pose serious problems for the standard classifier building approaches, will be discussed. The benefits/limitations of using unlabelled data for different values of *r* will be analysed in the section presenting experimental results.

III. METHODS FOR HANDLING LABELLED AND UNLABELLED DATA

Given the problem statement and notation introduced in the previous section five different approaches to generating classifier models given a set of labelled and unlabelled data will now be formally described.

1. The first and the most obvious way of dealing with the above problem is to build a classifier C_L using just the labelled subset *L* from *D* and completely ignoring *U*. The classification process from Eq. 1 in this case becomes:

   $$t=C_L(\mathbf{x}) \tag{2}$$

   In the experimental section this basic approach will be compared to the following four approaches, which attempt to utilise the unlabelled data in the process of building the final classifier.

2. 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 in the following way:

   $$\underset{j=1}{\overset{N}{\forall}}\mathbf{x}_j\in U \quad W=\{\mathbf{x}_j,t_j=C_L(\mathbf{x}_j)\} \tag{3}$$

where W is the newly labelled set U and subsequently redesigning the classifier using both the original L and the newly labelled W data sets. In result the Eq. 2 can be rewritten in the following way:

$$t = C_{L\cup W}(\mathbf{x}) \tag{4}$$

3. The next 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 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. Formally this iterative labelling and classifier redesign process can be described in the following steps:

   a) Given L and U initialise $U' = U$ and $W' = \{\varnothing\}$ where U' represents a current set of unlabelled data and W' represents a current set of newly labelled data.

   b) Design a classifier $C_{L\cup W'}$. Among all $\mathbf{x}_i \in U'$ find such $\mathbf{x}_j$ which can be the most confidently classified using the classifier $C_{L\cup W'}$ and add it to the current set of newly labelled data

   $$W' = W' \cup \{\mathbf{x}_j, t_j = C_{L\cup W'}(\mathbf{x}_j)\} \tag{5}$$

   Note: the definition of the most confidently classified sample is dependant on the type of classifier used and can refer to the shortest distance in case of nearest neighbour classifier, the highest classification probability for classifiers generating probabilistic outputs, the highest degree of class membership for classifiers generating fuzzy outputs etc.

   c) Remove $\mathbf{x}_j$ from the current unlabelled data set

   $$U' = U' - \{\mathbf{x}_j, 0\} \tag{6}$$

   d) If all the unlabelled data samples have been labelled (i.e. $U' = \{\varnothing\}$) go to e) otherwise go to b)

   e) Given L and the newly labelled set W' design the final classifier for which the Eq. 2 can be rewritten as:

   $$t = C_{L\cup W'}(\mathbf{x}) \tag{7}$$

4. The above two approaches can be thought of as using the unlabelled data for tuning an initial classifier C_L. As discussed in the introduction quite the opposite approach is based on initially discarding the labels and building a data model. In the following description clustering of data has been adopted for generating such data models. 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.

   Let S_l, l=1..k denote data clusters, $|S_l|$ - the l-th cluster cardinality (i.e. the number of samples in the l-th cluster), g_{lj}, j=1..P - the number of labelled samples from class j in the l-th cluster, t_{s_l} - the label of the samples from cluster S_l.

   Given the above notation and initialising an index b=1 to be used in the first step of the algorithm, the cluster labelling process can be formally described in the following steps:

   a) For all k clusters:

   i) if $\sum_{j=1}^{P} g_{lj} \neq 0$ (i.e. there are labelled samples in the cluster S_l)

   i.i) Find the label (index) of the most representative class t_{s_l}

   $$t_{s_l} = \arg\max_{j\in\{1..P\}}(g_{lj}) \tag{8}$$

   i.ii) relabel all the samples in the cluster S_l with this majority label and construct a labelled subset W_b'' as:

   $$\forall_{i=1}^{|S_l|} \mathbf{x}_i \in S_l \qquad W_b'' = \{\mathbf{x}_i, t_{S_l}\} \tag{9}$$

   i.iii) and for consistency in numbering of the labelled subsets W_b'' to be used in the next step of the algorithm increase the index b by 1: b=b+1

b) After step a) k clusters can be divided into z labelled clusters (S_i, t_{S_i}), i=1..z (and associated with them labelled subsets W_i'') and *(k-z)* unlabelled clusters $(S_j, 0)$, j=1..*z-k* (containing only unlabelled samples). The labelling of the unlabelled clusters can now be carried out on the basis of a suitably chosen cluster similarity measure Δ with Δ_{ij} representing the similarity values between the i-th labelled and the j-th unlabelled cluster in the following way:

i) For all unlabelled clusters S_j

i.i) Find the index $m \in \{1..z\}$ of the labelled cluster which is the most similar to the j-th unlabelled cluster

$$m = \arg\max_{i \in \{1..z\}} (\Delta_{ij}) \tag{10}$$

i.ii) label all the samples in cluster S_j with the label t_{S_m} and construct a labelled subset W''_{j+z} as shown in Eq. 9.

Note: In case when the clusters are represented by a point prototype the Euclidean distance between cluster prototypes could be used as the similarity measure Δ where the clusters with the shortest distance between them can be judged as the most similar. Various other non-vector cluster similarity measure discussed in [21] could also be used.

c) Given a newly labelled set $W'' = W_1'' \cup W_2'' \cup ... \cup W_k''$ construct a final classifier for which the Eq. 2 can be rewritten as:

$$t = C_{W''}(\mathbf{x}) \tag{11}$$

5. The final examined approach is a Semi-supervised Clustering where initial clusters are split until there is an overwhelming presence of one type of labelled samples in each of newly created sub-clusters. In contrast to the standard clustering used in the previous approach the labels are actively used for guiding the clustering process. In result the algorithm is more robust in a 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.

Starting with a relatively small number of clusters k the splitting of the clusters (if necessary) is based on examining whether: a) there are conflicts within a cluster (i.e. presence of labelled samples coming from different classes) and b) there are any labelled samples of the minority classes in the other clusters.

Let S_l, l=1..k denote data clusters and g_{lj}, j=1..P - the number of labelled samples from class j in the l-th cluster.

The splitting of the clusters can now be formally described as:

a) For all k clusters

i) if $\sum_{j=1}^{P} g_{lj} \neq 0$ (i.e. there are labelled samples in the cluster S_l)

i.i) Find the number of samples representing the majority class in S_l

$$g_{lm} = \max_{j=1}^{P} (g_{lj}) \tag{12}$$

i.ii) If the ratio of the labelled samples of a class to the total number of labelled samples in cluster S_l is lower than a user defined parameter $\Theta \in [0,1]$ which can be expressed as:

$$g_{lm} \Big/ \sum_{j=1}^{P} g_{lj} < \Theta \tag{13}$$

this class is referred to as a minority class. If there are no samples from minority classes represented in other clusters then the cluster S_l is split into two clusters, otherwise the minority class is ignored.

i.iii) If there is still more than one type of class labels in the cluster S_l then this cluster is split into two clusters. In case of hierarchical clustering splitting means that one just moves down the hierarchy of clusters and the sub-clusters can be examined in turn.

b) Once there are no clusters that need to be split the labelling of clusters and generation of the labelled set of samples can be carried out as in the previous section concerning the Majority Clustering method.

IV. EXPERIMENTAL RESULTS

While the descriptions of the general approaches in the previous section have been kept on a fairly general level illustrating a possibility of using different classifiers, clustering algorithms, cluster similarity measures etc., the simulation results reported in this section have been obtained for very specific settings which will now be summarised.

The nearest neighbour (NN) and pseudo-fisher support vector (PFSV) classifiers implemented in [5] have been used as the base classifiers for labelling and testing purposes as described in section III. While the NN classifier has been used for all five approaches, the PFSV classifier was only used for Static labelling method (section III.2).

A complete-linkage hierarchical clustering has been used for Majority Clustering (section III.4) and Semi-supervised Clustering (section III.5) with the shortest Euclidean distance adopted for the cluster similarity measure as described in section III.4. The parameter Θ used in the Semi-supervised Clustering has been set to 0.3.

The following four well known data sets representing non-trivial classification problems have been used in the experiments:

a) Normal mixtures data set – an artificial, 2-dimensional data set. The training data consists of 2 classes with 125 points in each class. Each of the two classes has bimodal distribution and the classes were chosen in such a way as to allow the best-possible error rate of about 8%. The training set and an independent testing set of 1000 samples drawn from the same distribution are available at http://www.stats.ox.ac.uk/~ripley/PRNN. The reported results are for this independent testing set.

b) Cone-torus data set – an artificial, 2-dimensional data set . The training data set consists of three classes with 400 data points generated from three differently shaped distributions: a cone, half a torus, and a normal distribution. The prior probabilities for the three classes are 0.25, 0.25 and 0.5. The training data and a separate testing set consisting of further 400 samples drawn from the same distribution are available at http://www.bangor.ac.uk/~mas00a/. The reported results are for this independent testing set.

c) Iris data set – a 4-dimensional data set representing a problem of classifying Iris plants taken from the Repository of Machine Learning Databases. The training set consists of 150 data samples with 50 samples from each of the 3 classes. The reported results have been obtained by using 10-fold cross validation procedure.

d) Glass data set – a 10-dimensional data set representing a problem of classifying of different types of glass. The training set consists of 214 data samples representing 6 classes. The reported results have been obtained by using 5-fold cross-validation procedure. The 5-fold cross-validation procedure has been used due to the fact that one of the classes has only 9 samples.

The experiments have been performed for different ratios r of labelled data to the total number of data samples ranging from virtually unlabelled sets only (1% of labelled data) to the fully labelled data sets (100% of labelled). The specific levels for which the experiments have been conducted were: 0,1,2,5% and then every 5% up to 100%. At each level the experiments have been repeated for 50 different randomly selected subsets to be used as labelled data. The same sets of labelled samples have been used in all the experiments with different classification methods. In this way we hoped to gain a better understanding of whether the selection of the labelled samples or the method for handling both types of data is more important.

The results for all four data sets for some levels of labelled data and six different methods of generating classifiers from labelled and unlabelled data are shown in Tables I, II, III and IV.

Fig. 2 shows a typical change in the mean classification performance and variance dependant on the subset of the labelled data used. Very similar patterns of change have been observed for all considered data sets. As also noted in [1], we can see that a specific subset picked as labelled has a great influence on the performance of the system if only a very limited amount of labelled data is used. The unlabelled data in such cases cannot be used efficiently and the whole process is dominated by how reliable the labelled samples are. The benefits of the unlabelled data and combined approaches can only be realised when sufficient level of labelled data representing an underlying distribution is available to compensate for noisy and mislabelled samples. Once such sufficient level of labelled data is reached the use of the combined approaches can provide a performance that is comparable with the classifiers trained using much higher number of labelled samples. This is shown in Fig. 2 by relatively stable performance from the moment when only about 10-20% of labelled data is used. It can also be observed for the Glass data set (highlighted cells in Table I) where the combined approaches using only 60% of labelled data have better performance than the labelled only NN

using 80% of labelled data. Similar effect can be observed for the Iris data set in Table III (highlighted cells). However, other results (e.g. for the Cone-torus data set shown in Table II) do not suggest a uniformly beneficial effects of using additional unlabelled data and the labelled only approach performs equally well (or bad).

The benefits of using Semi-supervised Clustering, in problems where natural clusters of data exist, is illustrated in Fig. 3 and Table IV. From the level of 25-30% of labelled data onwards a significant improvement can be seen in comparison to the approaches based on the labelled data only or using a static or dynamic labelling. The observed benefits of using Majority Clustering and the Semi-supervised Clustering are due to their ability to reduce the influence of noisy data and find smoother decision boundaries especially in cases of overlapping classes. This ability is dependant on the suitable choice of the number of clusters in the Majority Clustering case and the parameter Θ for the presented version of Semi-supervised Clustering. In general the Semi-supervised clustering has shown to be more robust due to its ability to adjust the number of clusters irrespective of the number of clusters with which the algorithm is initialised. On the other hand the Majority Clustering, while being able to produce good results, is highly dependant on the suitable selection of the number of clusters, which is related to the cluster validity problems.

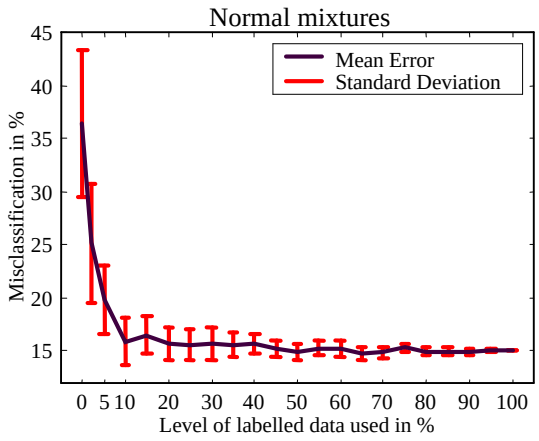

Figure 2: Normal mixtures dataset - Mean classification error and standard deviation for static NN based on different subsets and levels of the labelled data .

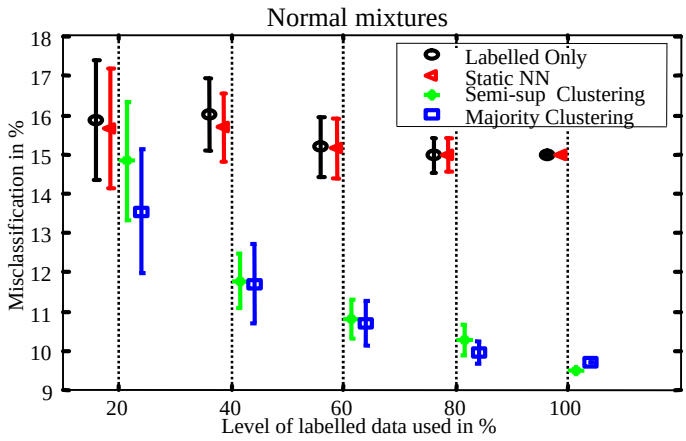

Figure 3: Normal mixtures dataset - Comparison of Static NN, Labelled only NN, Majority Clustering and Semi-supervised clustering algorithms.

% of labelled data	Dynamic NN		Static NN		PFSVC		Labelled Only NN		Semi-supervised Clustering		Majority Clustering	
0	28.62	(7.84)	18.02	(7.96)	18.50	(8.30)	17.88	(7.98)	18.00	(8.02)	18.00	(8.02)
2	21.54	(8.74)	14.18	(7.13)	14.49	(7.46)	14.17	(6.94)	14.23	(6.99)	14.23	(6.99)
5	13.52	(6.77)	10.41	(6.07)	10.52	(6.11)	10.59	(5.69)	10.31	(5.93)	10.31	(5.93)
10	7.98	(5.36)	7.10	(4.37)	7.25	(4.52)	7.47	(4.35)	7.01	(4.33)	7.01	(4.33)
20	4.44	(4.01)	4.49	(3.45)	4.37	(3.42)	5.37	(3.63)	4.54	(3.57)	4.54	(3.57)
40	2.24	(2.26)	2.15	(2.06)	1.87	(2.06)	3.12	(2.81)	2.26	(2.15)	2.26	(2.15)
60	1.58	(1.83)	1.56	(1.75)	1.15	(1.54)	2.37	(2.12)	1.57	(1.82)	1.57	(1.82)
80	1.18	(1.57)	1.16	(1.43)	0.71	(1.17)	1.63	(1.65)	1.16	(1.43)	1.16	(1.43)
100	0.92	(1.13)	0.92	(1.13)	0.47	(0.93)	0.92	(1.13)	0.92	(1.13)	0.92	(1.13)

Table I: Glass data set - Misclassification rate in % and its standard deviation (shown in brackets).

% of labelled data	Dynamic NN		Static NN		PFSVC		Labelled Only NN		Semi-supervised Clustering		Majority Clustering	
0	39.92	(14.04)	35.10	(8.74)	35.16	(9.67)	35.19	(8.71)	35.01	(8.99)	35.01	(8.99)
2	33.37	(13.03)	27.31	(6.46)	30.14	(10.22)	27.22	(6.88)	27.05	(6.20)	27.08	(6.26)
5	24.12	(5.70)	21.25	(3.89)	21.62	(3.97)	21.03	(4.11)	21.13	(3.59)	21.12	(3.56)
10	21.48	(3.28)	20.09	(3.51)	19.08	(2.90)	19.15	(3.26)	20.14	(3.17)	19.89	(3.24)
20	19.44	(3.01)	18.38	(2.51)	17.17	(2.25)	17.45	(2.54)	18.57	(2.36)	18.28	(2.21)
40	18.01	(1.99)	17.44	(1.45)	15.92	(1.43)	16.81	(1.52)	17.38	(1.46)	17.18	(1.86)
60	17.13	(1.20)	16.89	(1.16)	15.96	(1.12)	16.18	(1.43)	16.16	(1.30)	15.95	(1.24)
80	16.02	(0.94)	15.92	(0.92)	15.37	(0.92)	15.78	(1.02)	14.78	(0.88)	15.47	(1.13)
100	15.25	(0.00)	15.25	(0.00)	15.75	(0.00)	15.25	(0.00)	13.25	(0.00)	14.46	(0.80)

Table II: Cone-torus data set - Misclassification rate in % and its standard deviation (shown in brackets).

% of labelled data	Dynamic NN	Static NN	PFSVC	Labelled Only NN	Semi-supervised Clustering	Majority Clustering
0	20.33 (12.90)	14.13 (10.77)	14.13 (10.77)	13.43 (10.89)	13.97 (11.45)	14.57 (11.24)
2	18.83 (13.56)	13.73 (9.91)	13.77 (9.94)	13.30 (10.39)	13.23 (10.41)	13.17 (10.60)
5	8.00 (8.32)	8.73 (6.99)	8.33 (7.10)	8.60 (7.40)	8.27 (7.27)	8.47 (7.10)
10	5.93 (5.89)	5.07 (4.74)	4.77 (5.40)	6.50 (5.65)	6.47 (6.05)	6.77 (6.03)
20	4.80 (4.49)	5.07 (4.74)	4.77 (5.40)	5.13 (5.58)	4.53 (4.42)	5.20 (4.78)
40	4.10 (4.05)	4.47 (4.29)	4.73 (5.03)	5.10 (5.04)	4.50 (4.48)	4.47 (4.23)
60	3.63 (4.05)	3.83 (4.20)	4.33 (4.99)	4.60 (4.79)	3.67 (3.88)	3.77 (4.20)
80	3.83 (4.25)	3.97 (4.23)	4.50 (5.13)	4.40 (4.70)	3.40 (3.60)	3.43 (3.90)
100	4.00 (4.43)	4.00 (4.43)	4.67 (5.22)	4.00 (4.43)	2.67 (3.27)	2.17 (3.13)

Table III: Iris data set - Misclassification rate in % and its standard deviation (shown in brackets).

% of labelled data	Dynamic NN	Static NN	PFSVC	Labelled Only NN	Semi-supervised Clustering	Majority Clustering
0	43.68 (9.01)	36.44 (13.74)	36.42 (13.82)	36.22 (13.92)	36.76 (11.35)	36.80 (11.11)
2	36.59 (11.41)	25.17 (11.15)	25.85 (11.55)	25.04 (11.37)	25.56 (9.40)	26.01 (9.89)
5	26.36 (9.23)	19.81 (6.34)	18.69 (6.69)	19.56 (6.28)	19.73 (6.24)	19.04 (6.01)
10	18.95 (6.08)	15.89 (4.39)	14.25 (4.20)	15.72 (4.46)	17.70 (5.01)	16.38 (4.09)
20	16.09 (3.36)	15.67 (3.06)	13.64 (2.92)	15.88 (3.05)	14.84 (3.01)	13.55 (3.17)
40	15.87 (1.64)	15.69 (1.73)	13.78 (1.68)	16.02 (1.84)	11.77 (1.38)	11.71 (2.03)
60	15.28 (1.44)	15.16 (1.52)	12.98 (1.60)	15.19 (1.52)	10.81 (0.97)	10.69 (1.12)
80	15.04 (0.83)	14.99 (0.83)	13.00 (0.98)	14.98 (0.89)	10.28 (0.78)	9.96 (0.60)
100	15.00 (0.00)	15.00 (0.00)	12.80 (0.00)	15.00 (0.00)	9.50 (0.00)	9.70 (0.00)

Table IV: Normal mixtures data set - Misclassification rate in % and its standard deviation (shown in brackets).

V. CONCLUSIONS

The purpose of this paper was to conduct an experimental analysis of various approaches of handling labelled and unlabelled data in the process of constructing pattern classification systems.

All the performed tests and comparisons have confirmed that combined methods can be cost effective in a sense that less labelled data is required to obtain the performance comparable with the pure supervised approaches.

It was also found that if only a very limited amount of labelled data is available the results show high variability and the performance of the final classifier is more dependant on how reliable the labelled data samples are rather than use of additional unlabelled data. Semi-supervised clustering utilising both labelled and unlabelled data have been shown to offer most significant improvements when natural clusters are present in the considered problem.

Our future work will concentrate on a suitable selection of the samples to be labelled, rather than choosing them randomly, which (as the preliminary results suggest) could significantly reduce the variance of the final solutions and potentially improve the mean classification performance.

REFERENCES

[1] K.P. Bennett, A. Demiriz, 1998, "Semi-Supervised Support Vector Machines", *Proc. of Neural Information Processing Systems*, pp. 368-374.

[2] A. Blum, T. Mitchell, 1998, "Combining Labelled and Unlabelled Data with Co-Training", *Proc. of COLT'1998*, pp. 92-100;

[3] R. Caruana, D. Cohn, A. McCallum, 2000, "Semi-Supervised Clustering with User Feedback", *Machines that Learn*, Snowbird, Utah, USA;

[4] R. Dara, S.C. Kremer, D.A. Stacey, 2002, "Clustering Unlabelled Data with SOMs Improves Classification of Labelled Real-world Data", *Proc. of the 2002 IEEE World Congress on Computational Intelligence*, pp. 2237-2242;

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

[6] B. Everitt, 1981, *Cluster Analysis*, 2nd ed., Halsted Press;

[7] B. Gabrys, 2002, "Neuro-Fuzzy Approach to Processing Inputs with Missing Values in Pattern Recognition Problems", accepted to the *International Journal of Approximate Reasoning;*

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

[9] Z. Ghahramani, M.I. Jordan, 1994, "Supervised Learning from incomplete data via an EM Approach", *Advances in Neural Information Processing Systems*, pp. 120-127;

[10] S. Goldman, Y. Zhou, 2000, "Enhancing Supervised Learning with Unlabelled Data", *Proc. of the 17th ICML'2000*, pp. 327-334;

[11] A.F. Gomez-Skarmeta, F. Jimenez, M. Valdes, J.A. Botia, A.M. Padilla, 2002, "Towards a Modelling Framework for Integrating Hybrid Techniques", *Proc. of the 2002 IEEE World Congress on Computational Intelligence*, pp. 985-990;

[12] R. Kothari, V.Jain, 2002, "Learning from Labeled and Unlabeled Data", *Proc. of the 2002 IEEE World Congress on Computational Intelligence*, pp.

[13] T. Mithell, 1999, "The Role of Unlabelled Data in Supervised Learning", *Proc. of the Sixth International Colloquium Cognitive Science*;

[14] K. Nigam, A.K. McCallum, S. Thrun, T. Mitchell, 2000, "Text Classification from Labelled and Unlabelled Documents using EM", *Machine Learning*, pp. 103-134;

[15] K. Nigam, R. Ghani, 2000, "Understanding the Behaviour of Co-training", *Proc. of KDD-2000 Workshop on Text Mining*;

[16] K. Nigam, R. Ghani, 2000, "Analyzing the Effectiveness and Applicability of Co-training", Ninth International Conference on Information and Knowledge Management, *Proc. of CIKM'2000*, pp. 86-93;

[17] J. Park, H. Yae, 2002, "Analysis of Active Feature Selection in Optic Nerve Data Using Labelled Fuzzy C-Means Clustering", *Proc. of the 2002 IEEE World Congress on Computational Intelligence*, pp. 1580-1585;

[18] W. Pedrycz, J. Waletzky, 1997, "Fuzzy Clustering With Partial Supervision", *IEEE Trans. On System, Man, and Cybernetics, Part B: Cybernetics*, vol. 27, no. 5, pp. 787-795;

[19] D. Ruta, B. Gabrys , 2001, "Analysis of the correlation between majority voting error and the diversity measures in multiple classifier systems", *Proc. of the SOCO/ISFI'2001 Conference*, Abstract page 50, Paper no. #1824-025;

[20] M. Seeger, 2001, "Learning with Labelled and Unlabelled Data", *Technical Report*, Edinburgh University;

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