Clustering-based Feature Selection in Semi-supervised Problems

Ianisse Quinzán, José M. Sotoca, Filiberto Pla
Institut de Noves Tecnologies de la Imatge, Dept. Llenguatges i Sistemes Informàtics
Universitat Jaume I
Castellón de la Plana, Spain
ianisseqs@yahoo.es, [sotoca,pla]@lsi.uji.es

Abstract— **In this contribution a feature selection method in semi-supervised problems is proposed. This method selects variables using a feature clustering strategy, using a combination of supervised and unsupervised feature distance measure, which is based on Conditional Mutual Information and Conditional Entropy. Real databases were analyzed with different ratios between labelled and unlabelled samples in the training set, showing the satisfactory behaviour of the proposed approach.**

Keywords: Semi-supervised learning, feature selection, information measures.

I. INTRODUCTION

With the advent of new technologies large amount of data with a high number of attributes can be collected automatically. The classification or labeling of samples by an expert can often be too expensive in time and sometimes even unfeasible.

When we have an empirical knowledge about the classes assigned to the samples in the training set, we say that the learning problem has a supervised nature. If the samples are not labeled, the learning problem is considered as unsupervised. In many application problems there is available a significant amount of unlabelled data, and only few labeled samples. The introduction of such few labels can improve the classification accuracy significantly [17]. We will refer to this problem as semi-supervised learning and it has recently received an increased interest in the pattern recognition and machine learning communities.

A challenge in semi-supervised learning is the feature selection problem. It is defined as follows: given a set of *n* features, select a subset of size *m* that leads to the smallest classification error. This problem involves a special difficulty especially when the size of labeled information is very small.

Regarding the strategy, filter methods select features subsets independently of the classifier. Wrapper methods use a classifier to evaluate some feature subsets considered; thus, these methods are computationally heavy and they are conditioned to the classifier chosen. In addition to wrapper and filtering methods, there is a third kind of techniques which can be employed to perform feature selection processes, denominated embedded methods [7].

Several works have developed strategies to solve the problem of feature selection in semi-supervised learning. In [11] [4] [12] a wrapper strategy is used. In [12] the authors propose a clustering method extends labels to unlabeled samples and evaluates the partition. In [11], for each step, the algorithm builds a new training set many times, which are utilized to make one selection. At the end of each step it adds the feature more frequently selected. In this method, the unlabeled information is not fully used. Moreover, the subset of attributes derived from the random training sets used may not be adequate, but once the feature is chosen, it will never be eliminated.

In [4], a genetic algorithm is applied to generate the candidate subset and evaluate the clustering of them. Given the high cost of the genetic part, this method is limited for databases with high dimensions (less than 20 attributes).

On the other hand, the works [3], [15] and [16] can be considered inside the filter approaches. In [3] an extension of a Logistic I-Relief is proposed. This method returns a weight for each feature through an iterative optimization algorithm to get a maximal margin. The margin of each data measures the difference between the distances of the nearest neighbors of the same class and the different class. The objective function includes the margins of unlabeled samples. This technique has the inconvenient that it is formalized and tested for problems with only two classes.

Others algorithms [15][16] using the spectral graph theory. The graph has a node for each sample, and an edge between two nodes if they are close. The approach [15] evaluates each feature vector, transforming it into a cluster and checking whether it is consistent with the class. In [16], a local geometrical structure and discriminate structure of data is captured with two graphs: within-class graph and between-class graph. The importance of the features is characterized by its degree of preserving the graph structures. The presence of a large amount of irrelevance and noise features often leads to inexact neighborhood mapping and produces that the method can fail.

In this paper, a filter method for feature selection is presented. A new hybrid method for semi-supervised problem is proposed, which combines supervised and unsupervised measures of information. This approach applies a strategy to obtain a feature subset through clustering techniques.

In section 2 the methodology of the filter feature selection method for semi-supervised problem is presented. In section 3 several experiments with real databases are shown. Finally, some conclusions are drawn in section 4.

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.211

IEEE computer society

II. FEATURE CLUSTERING AND SEMI-SUPERVISED DISTANCES

In semi-supervised learning, the number of reliable labeled samples is often small and semi-supervised methods try to utilize the large amount of available unlabeled data. If the result using a supervised method with these few objects is satisfactory, then the semi-supervised approach is not worthwhile.

We focus on the application problems where the unlabeled information can improve in a significant way the classification result of just using the limited labeled samples available. Therefore, the technique here presented aims to combining in a hybrid method supervised and unsupervised information for such critical cases.

Given the original set of features X, and a subset of features $\widetilde{\boldsymbol{X}}=\{\tilde{X}1;\ldots,\tilde{X}m\}$, m < n, where $\widetilde{\boldsymbol{X}}\subset$ X, it can easily be shown the following relation between these Mutual Information

$$I(\widetilde{\boldsymbol{X}};Y) \le I(\boldsymbol{X};Y) \qquad (1)$$

Note that the highest $I(\widetilde{\boldsymbol{X}};Y)$ leads to the subset of features that better represents the original set with respect to the relevant variable Y .

This is the underlying principle that has motivated different approaches for supervised feature selection, named also *Max-Dependency criterion* [10] using different optimization strategies [6], [1]. These works differ in the way they approximate two practical issues:

- The estimation of $I(\widetilde{\boldsymbol{X}};Y)$ and $I(\boldsymbol{X};\boldsymbol{Y})$ become very complex and computational expensive.
- The search strategy to overcome the combinatorial problem, finding the optimal subset of features.

Our criterion function is also related to the *Max-Dependency*, that is, maximizing $I(\widetilde{\boldsymbol{X}};Y)$, which is equivalent to minimizing $I(\boldsymbol{X};Y) - I(\widetilde{\boldsymbol{X}};Y) \ge 0$.

Given $\widetilde{\boldsymbol{X}}\subset\boldsymbol{X}$, the decrease in mutual information about a relevant variable Y can be expressed as conditional mutual information $I(\boldsymbol{X},Y/\widetilde{\boldsymbol{X}})$, and the problem can be viewed as a minimization of this expression. Thus, it can be shown that the decrease in mutual information $I(\boldsymbol{X};Y) - I(\widetilde{\boldsymbol{X}};Y)$ is upper-bounded by:

$$\begin{aligned} I(\boldsymbol{X};Y) - I(\widetilde{\boldsymbol{X}};Y) &= I(\boldsymbol{X},Y/\widetilde{\boldsymbol{X}}) \\ &\le \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i,Y/\tilde{X}_j) \end{aligned} \qquad (2)$$

The subtraction $I(\boldsymbol{X};Y)$ - $I(\widetilde{\boldsymbol{X}};Y)$ is related with the Conditional Entropy $H(\boldsymbol{X}/\widetilde{\boldsymbol{X}})$ as:

$$\begin{aligned} I(\boldsymbol{X},Y) - I(\widetilde{\boldsymbol{X}},Y) &\le H(\boldsymbol{X}/\widetilde{\boldsymbol{X}}) \\ &\le \sum_{i=1}^{n}\sum_{j=1}^{m} H(X_i/\tilde{X}_j) \end{aligned} \qquad (3)$$

Bounds (2) and (3) can be combined in a single expression using some weights α_1, α_2 for the unsupervised and supervised parts, respectively:

$$\begin{aligned} I(\boldsymbol{X};Y) - I(\widetilde{\boldsymbol{X}};Y) \le \alpha_1 \sum_{i=1}^{n}\sum_{j=1}^{m} H(X_i/\tilde{X}_j) \\ +\alpha_2 \sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i,Y/\tilde{X}_j) \end{aligned} \qquad (4)$$

In order to find the subset of $m < n, \widetilde{\boldsymbol{X}}\subset \boldsymbol{X}$, features that minimize the above expression, a clustering based strategy is proposed, as an approximation to use expression (4) as the criterion function to minimize, grouping the original set of features $\boldsymbol{X}$ into clusters $\widetilde{\boldsymbol{X}}$, and finally selecting a feature representative for each cluster.

Thus, Ward's agglomerative hierarchical clustering has been used as a clustering strategy [9], but using an adequate distance for the semi-supervised approach.

Ward's linkage method [13] has the property of producing minimum variance partitions. Therefore this method is also called minimum variance clustering because, the hierarchical grouping merges at each step the pair of clusters that minimize the increment intra-cluster variance. The intra-cluster feature similarities would tend to be small inside each cluster.

The distance measure proposed between each pair of features is hybrid, and has two parts, one supervised and other unsupervised. In essence, it is the sum of these two terms, weighted with the above mentioned coefficients. These terms are a uppers bound of $I(\boldsymbol{X};Y)$ - $I(\widetilde{\boldsymbol{X}};Y)$. The distance function proposed is a consequence of the bound in expression (4), by making the expression to be symmetric, thus:

$$\begin{aligned} &DNI_{hibrid}(X_i,X_j) \\ &\quad = \alpha_1 * (H(X_i/X_j) + H(X_j/X_i)) \\ &\quad + \alpha_2 * (I(X_i,Y/X_j) + I(X_j,Y/X_i)\,) \end{aligned} \qquad (5)$$

The first term (also known as the Mantaras' distance [8]) calculates the symmetrical conditional entropy of two random variables. This is an expression of the independent information that a feature has with respect to another one. Minimizing the conditional entropy we group in the same cluster the features that have higher Mutual Information between them and do not have much independent information of each other. When selecting one of the variables of a cluster and eliminate the others, we are discarding features with redundant information and low entropy within the same group.

The second term (it can be also proven that this term is also a distance metric) is based on Conditional Mutual Information $I(X_i,Y/X_j)$, and it can be interpreted as how much information variable X_i can predict about relevant variable Y that variable X_j cannot. That is, in terms of cost, what would be the cost if feature X_i is represented/substituted by feature X_j, with respect to predicting variable Y. This can be interpreted as the independency between variables X_i and X_j with respect to variable Y, that is, the amount of information they can predict about Y, but they do not share.

In the estimation of the unsupervised term, all instances in the training data set are used. The precision of the supervised term depends on the number of labeled samples.

Finally, for each cluster C_j, the representative feature $\tilde{X}_j$ will be selected as the one with the highest Mutual Information with respect to the relevant variable Y,

$$\tilde{X}_j = \max I(X_i, Y) \qquad \forall X_i \in C_j \qquad (6)$$

III. Experiments and results

The proposed method *Semi-Supervised Feature Clustering* (*ssfc*) has been tested in several semi-supervised real problems, and compared with WaLuMI [9], which is an unsupervised method based also on feature clustering, in this case, using all unlabelled data (*nsup*), and with a supervised version of the proposed method (*sup*) over the labeled samples. This supervised version uses only the second term of the expression (5), eliminating the unsupervised term. We denote *supT* to this supervised version applied over all the samples labeled, that is, labeling also all the unlabeled samples used in the previous algorithms. This will allow checking the accuracy when a full supervised learning is reached.

WaLuMI is very similar to our method; its' differences are in dissimilarity measure and the form of select a feature representative of each cluster. This is chosen using a strategy based in minimum variance. The relation between attributes, used to make the clusters, is a normalization of Mutual Information between each pair of them.

Weights α_1 and α_2 were estimated experimentally, choosing α_2=0,75 and α_1=0,25, which provided in general satisfactory results.

As a preliminary test, in this work only databases represented on discrete random variables have been used. For the experimentation the following datasets have been chosen of the UCI repository (*http://archive.ics.uci.edu/ml/*).

- Gisette is a big data in the UCI repository, with 5000 attributes and 13500 objects, 7000 of them labelled. It is a handwritten digit recognition problem; and the task is to discriminate between the four and the nine numbers. It has artificial attributes that are randomly generated, and the values are rather sparse, about 13% of the values are non-zero. In the experiments, only the first 500 features with the highest entropy were used.
- Optdigits problem is about the recognition of a handwritten number. The database has 5620 samples and 64 features. The 32x32 bitmaps are divided into no overlapping blocks of 4x4 and the number of pixels is counted in each block. This generates an input matrix of 8x8 (64 features) where each value is an integer in the range 0..16.
- In covtype database, the objective is predicting forest cover type from cartographic variables, with no remotely sensed data. This database has 54 features, 581012 objects and 7 classes.

A hyperspectral image called 92AV3C [9] has also been used, corresponding to a spectral image (145 x 145 pixels, 220 bands, and 17 classes) acquired with the AVIRIS, in June 1992, over the Indian Pine Test, in Northwestern Indiana. (http:/dynamo.ecn.purdue.edu/~biehl/MultiSpec) It has also pixels with an undetermined class. In this hyperspectral image several bands are discarded due to the effect of atmospheric absorption; 185 of the original 220 bands were used discarding the lowest signal-to-noise ratio bands.

For each database the methods were tested using different number of labeled samples 25, 50, 100, 500 and 1000. They were selected randomly preserving the prior probability of the classes. The process for each size of labelled samples was performed five times, and the averages of the accuracy for these five tests are used for comparison.

To validate the significance of the subsets of features obtained, a holdout scheme in training (50%) and test (50%) is applied with a K-Nearest Neighbour classifier [2] with K=3 (3-NN). This classifier only uses the spatial distributions of the samples without assumptions about the distributions of classes. The classification method is supervised and it is used a measure of goodness of the subset of features selected.

In covtype dataset, the validation process is too expensive in computational time (the training set has 290.506 elements), and the training set was reduced. Noise samples were edited with a method in [14] Edited Nearest Neighbours (ENN) with K =3; and next it was condensed using Condensed Nearest Neighbours (CNN) [5]. The final size of the training has 28.291 samples, in this case it was validated using the Nearest Neighbour (1-NN) classifier.

Tables 1 to 4 show the classification accuracy for all databases used and the different number (rows) of labeled sets in the training process. Only the subsets with less than 30 features are shown, because in all cases, with approximately 30 features the classifier reaches a stable performance. The columns show the accuracy of subsets with 5, 15 and 30 elements.

Figure 1 represents a graphic for gisette and optdigits datasets with 25, 100 and 1000 samples labeled. The x-axis represents the number of feature selected, whereas the y-axis shows the average classification accuracy for the 5 trials performed.

Note that the supervised method that uses all labelled samples (*supT*) always obtain better accuracy than the unsupervised method with all the samples (*nsup*), except in 92AV3C where both methods have similar performance. Notice also that, in general, jointly supervised and unsupervised information improve the results, particularly when the unsupervised version tend to perform poorly, and adding few labeled sample increase the accuracy in a significant way.

Thus, the proposed hybrid method (*ssfc*) and its supervised version (*sup*) are better when the number of labeled samples is increased. When the number of labeled samples is sufficiently large, the performance of *ssfc* and *sup* converge.

If the behavior of *sup* method with low number of labeled samples is not satisfactory, the application of *ssfc* method obtains better results and good subsets of features. This can be specially noticed in covtype and gisette datasets. Sometimes the subsets obtained by *sup* are worse than *nsup*

TABLE I. RESULTS IN COVTYPE WITH 25, 50, 100, 500 AND 1000 SAMPLES LABELED.

	5 feat		**15 feat**		**30 feat**	
	supT=54.95 ***nsup***=39.89		***supT***=91.73 ***nsup***=39.91		***supT***=93.57 ***nsup***=40.01	
	sup	***ssfc***	***sup***	***ssfc***	***sup***	***ssfc***
25	49.76	**69.54**	63.87	**93.47**	73.07	**93.47**
50	48.21	**69.46**	60.62	**93.47**	78.08	**93.47**
100	42.22	**59.75**	55.47	**93.47**	82.67	**93.47**
500	47.66	**70.58**	**93.47**	**93.47**	**93.47**	**93.47**
1000	50.39	**74.32**	**93.47**	**93.47**	**93.47**	**93.47**

TABLE II. RESULTS IN GISETTE WITH 25, 50, 100, 500 AND 1000 SAMPLES LABELED.

	5 feat		**15 feat**		**30 feat**	
	supT=57.83 ***nsup***=66.94		***supT***=86.43 ***nsup***=83.17		***supT***=91.46 ***nsup***=89.74	
	sup	***ssfc***	***sup***	***ssfc***	***sup***	***ssfc***
25	61.57	**69.87**	79.23	**86.13**	80.94	**91.09**
50	**72.21**	71.54	82.10	**87.13**	85.25	**90.57**
100	72.88	**74.67**	83.37	**84.00**	84.82	**89.37**
500	70.70	**75.30**	**86.06**	85.43	**89.31**	89.09
1000	70.65	**72.34**	**87.35**	85.92	**89.60**	89.18

TABLE III. RESULTS IN 92AV3C WITH 25, 50, 100, 500 AND 1000 SAMPLES LABELED.

	5 feat		**15 feat**		**30 feat**	
	supT=72.35 ***nsup***=76.90		***supT***=80.43 ***nsup***=81.34		***supT***=82.42 ***nsup***=76.98	
	sup	***ssfc***	***sup***	***ssfc***	***sup***	***Ssfc***
25	**64.25**	62.08	77.12	**77.48**	**80.53**	79.34
50	67.51	**69.25**	76.70	**78.62**	79.43	**80.28**
100	71.72	**72.38**	76.84	**79.44**	78.14	**80.39**
500	**72.92**	72.91	79.92	**81.04**	81.16	**81.73**
1000	**74.42**	73.11	**80.60**	80.28	81.80	**81.94**

TABLE IV. RESULTS IN OPTDIGITS WITH 25, 50, 100, 500 AND 1000 SAMPLES LABELED.

	5 feat		**15 feat**		**30 feat**	
	supT=68.14 ***nsup***=20.32		***supT***=95.48 ***nsup***=69.89		***supT***=98.07 ***nsup***=90.71	
	sup	***ssfc***	***sup***	***ssfc***	***sup***	***Ssfc***
25	**45.94**	41.67	**89.25**	87.85	96.75	**97.06**
50	**49.44**	46.92	90.03	**91.90**	97.16	**97.31**
100	53.52	**56.39**	93.56	**93.70**	**97.96**	97.88
500	**64.33**	62.46	**95.01**	93.77	98.02	**98.09**
1000	62.46	**65.64**	**95.06**	95.33	97.28	**98.18**

because they need more supervised information. In these cases, the unsupervised information improves the accuracy and the *ssfc* method is adequate (see gisette in figure 1).

Optdigits is a database where *sup* technique gets high-quality features for few labeled samples. Thus, in this case the *ssfc* has similar performance than *sup*. Nevertheless when the number of labeled samples is increased, *ssfc* and *sup* become similar to *supT*.

In 92AV3C database, *nsup* has similar classification accuracy than *supT*. In this situation, the supervised method and the hybrid method have similar performance, and they reach *supT*.

Covtype is a special database where a small numbers of features (7 of 54) achieve the total accuracy (93.47%). When these attributes are in the subset selected the outcome is maximum while in other cases not. The hybrid method finds them, even with few labeled samples.

In gisette database, *supT* method with few features (less than 13) is worse than the other methods. The reason is due to the fact that this dataset has noisy samples. If noisy data are previously eliminated, *supT* obtains the best accuracy. For example, eliminating noise with ENN method, the set of 5 attributes selected by *supT* reaches 73.98% accuracy. When the number of labeled samples is less than 100, *sup* shows poor results, while the proposed *ssfc* has better precision.

IV. CONCLUSIONS

In this paper, a filter feature selection technique based on information theory for semi-supervised problems has been proposed. The method utilizes a dissimilarity measure between each pair of features with two parts: a supervised and an unsupervised.

The proposed method has been tested in different databases of real problems, comparing its performance with an unsupervised method and a supervised version of the presented method, where only the labelled information is used. The proposed semi-supervised method obtains satisfactory subsets of features when there is not enough

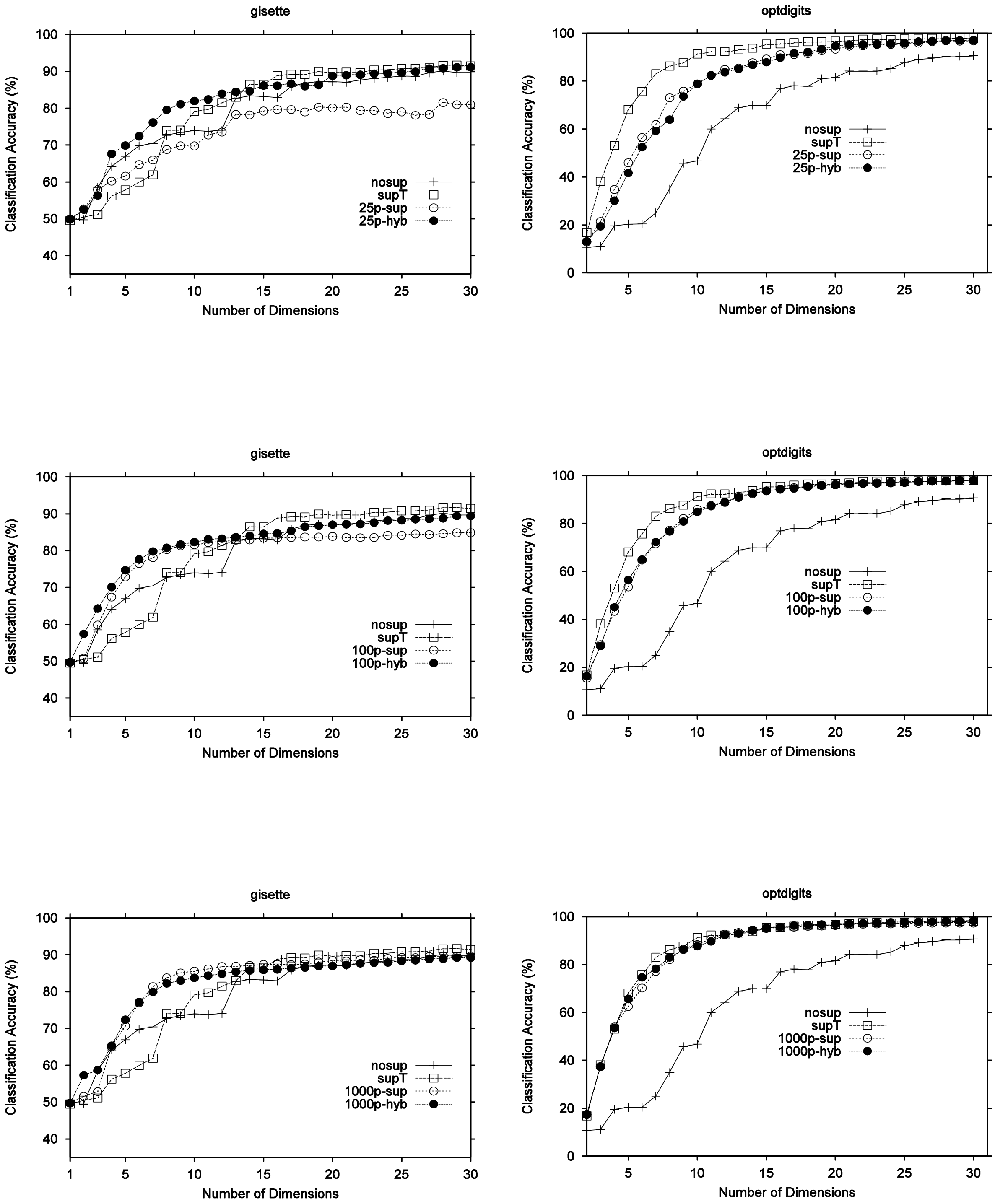

Figure 1. Results of classification accuracy with KNN for database gisette (left) and optdigits (right). We plot for 25, 100 and 1000 numbers of samples labeled (top to bottom).

labelled information. In these cases, using jointly the labelled and unlabeled information improves the selected features.

ACKNOWLEDGMENT

This work was supported by a Spanish Ministry of Science and Education under Project CSD2007-00018 and Spanish Agency of International Cooperation for the Development (AECID).

REFERENCES

[1] R. Battiti, “Using Mutual Information for selection features in supervised neural net learning”, IEEE Transactions Neural Network, vol. 5, no. 4, November 1994, pp. 537-550.

[2] T. Cover, and P. Hart, ”Nearest neighbor pattern classification”, IEEE Transactions on Information Theory, vol. 13, no. 1, January 1967, pp. 21-27.

[3] Y. Cheng, and Y. Cai, “Semi-Supervised Feature Selection Under Logistic I-RELIEF Framework”, 19th International Conference on Pattern Recognition, ICPR 2008, IEEE, Tampa, Florida, USA, December 2008, pp 1-4.

[4] J. Handl, and J. Knowles, “Semi-supervised feature selection via multiobjective optimization”, International Joint Conference on Neural Networks, IJCNN06, IEEE, Vancouver, Canada, July 2006, pp 3319-3326.

[5] P.E Hart, “The condensed nearest neighbour rule”. IEEE Transactions on Information Theory, Vol. 14,1968, pp 515-516.

[6] N. Kwak, and Chong-Ho Choi, ”Input Feature Selection for Classification Problems”, IEEE Transactions Neural Network, vol. 13, no. 1, January 2002, pp. 143-159.

[7] H. Liu and H. Motola, “Computational Methods of Feature Selection”. Chapman & Hall/CRC., New York, 2007.

[8] López de Mantaras, R. “ID3 revisited: a distance-based criterion for attributes selection”. In. Ras, Z. W. (Ed.), Methodologies for Intelligent Systems, 3. North-Holland, New York, pp. 342-350, 1989.

[9] A. Martínez-Usó, F. Pla, J. M. Sotoca, P. García-Sevilla: “Clustering-based multispectral band selection using mutual information”, 17th International Conference on Pattern Recognition, ICPR 2006, IEEE, Hong Kong, August 2006, pp 760-763

[10] H. Peng, F. Long and C. Ding, ”Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundance”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 27, no 8, August 2005, pp. 1226-1238.

[11] J. Ren, Z. Qiu, W. Fan, H. Cheng, and P. S. Yu, “Forward Semi-Supervised Feature Selection”, Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD08, Osaka, Japan, May 2008, T. Washo, E. Suzuki, K.M. Ting and A. Inokuchi (Eds) : Springer-Verlag LNAI 5012, pp 970-976.

[12] B. Wang, Y. Jia, and S. Yang, “Forward semi-supervised feature selection based on Relevant set correlation”, International Conference on Computer Science and Software Engineering (CSSE2008), IEEE, Wuhan, China, December 2008, pp 210-213.

[13] J. H. Ward, "Hierarchical grouping to optimize an objective function", American Statistical Association, vol. 58, no. 301, 1963, pp. 236-244.

[14] D.L. Wilson, “Asymptotic properties of nearest neighbour rules using edited data”, IEEE Transactions on Systems, Man and Cybernetics, Vol. 2, 1972, pp 408-421.

[15] Z. Zhao, and H. Liu, “Semi-supervised Feature Selection via Spectral Analysis”, SIAM International Conference on Data Mining, SIAM-07, SIAM, Minneapolis, Minnesota, USA, April 2007, pp 641-646.

[16] J. Zhao, K. Lu, and X. He, “Locality sensitive semi-supervised feature selection”, Elsevier Science Publishers, Neurocomputing Volume 71, June 2008, pp 1842-1849.

[17] X. Zhu, “Semi-supervised learning literature survey”, Computer Science TR 1530, University of Wisconsin – Madison, February, 2006