Partially Supervised Graph Embedding for Positive Unlabelled Feature Selection

Yufei Han and Yun Shen
Symantec Research Labs
Yufei_Han@symantec.com, Yun_Shen@symantec.com

Abstract

Selecting discriminative features in positive unlabelled (PU) learning tasks is a challenging problem due to lack of negative class information. Traditional supervised and semi-supervised feature selection methods are not able to be applied directly in this scenario, and unsupervised feature selection algorithms are designed to handle unlabelled data while neglecting the available information from positive class. To leverage the partially observed positive class information, we propose to encode the weakly supervised information in PU learning tasks into pairwise constraints between training instances. Violation of pairwise constraints are measured and incorporated into a partially supervised graph embedding model. Extensive experiments on different benchmark databases and a real-world cyber security application demonstrate the effectiveness of our algorithm.

1 Introduction and Related Work

Feature selection (FS) technology identifies the best subset of features from the input to form a compact but informative data representation. It helps in understanding data, reducing computation requirement, reducing the effect of curse of dimensionality and building robust machine learning models for classification, clustering, and other tasks. In general, there are three categories of feature selection methods: supervised feature selection, unsupervised feature selection and semi-supervised feature selection.

Supervised feature selection, e.g. Fisher score[Duda and P.E., 2001], LASSO [Tibshirani, 1996], Relief [Kononenko *et al.*, 1997], robust regression (RFS) [Nie *et al.*, 2010], trace ratio [Nie *et al.*, 2008], evaluates the importance of feature subsets based on correlation between labels and features. When data are unlabelled, unsupervised learning feature selection identifies the feature subsets best recovering low-dimensional graph embedding [Yan *et al.*, 2007] of the affinity graph of the given data. The graph embedding well preserves the relationship among data points in original high dimensional feature space, providing an biased but effective descriptor of underlying class distribution. Representative unsupervised feature selection methods include Laplacian Score [He *et al.*, 2005], Multi-Cluster Feature Selection (MCFS) [Cai *et al.*, 2010], Feature Selection via Joint Embedding Learning and Sparse Regression (JELSR)[Chenping Hou and Wu, 2011], Unsupervised Discriminative Feature Selection (UDFS) [Yang *et al.*, 2011], and Non-negative Discriminative Feature Selection (NDFS) [Li *et al.*, 2012]. Semi-supervised feature selection is designed to tackle with the predicament of certain practical data mining applications where labelled samples are extremely rare while unlabelled data are abundant. In this situation, supervised methods are prone to the overfitting issue given limited labelled data. Although the unsupervised approaches can handle unlabelled data, they neglect discriminative information carried by labelled samples. Previous efforts in semi-supervised feature selection [Zhang *et al.*, 2012] employ label propagation to inject label information into the graph embedding representation in order to improve the estimate of underlying class distribution.

Recent years have witnessed a challenging variant of semi-supervised learning, namely positive-unlabelled (PU) learning [Elkan and Noto, 2008]. It frequently arises in various real-world applications, e.g. text classification, biomedical informatics, pattern recognition, and recommendation system. In these cases, only limited positively labelled training samples are available. The rest unlabelled data set is a mixture of both positive and negative samples. In terms of feature selection, supervised and traditional semi-supervised methods can not be applied in PU learning scenarios directly since they require labelled data from every class. It is an interesting but challenging topic to identify informative features using only positive and unlabelled training samples. Pioneering research work in this field can be found in PRFS [S.Sundararajan and Keerthi, 2011] and apuCFS [Borja Calvo and A.Lozano, 2009]. PRFS introduces a pairwise ranking based SVM method to encourage the positive examples to score higher than the unlabelled examples. AUC value of the rank SVM classifier built on each feature measures how informative the corresponding feature is. apuCFS inherits the theoretical framework of correlation filtering selection (CFS). It searches for the best possible feature subset that are strongly correlated with the class, while weakly correlated with each other. Since the exact search is NP-hard, greedy forward filtering and backward elimination is employed to select the feature subset heuristically. Without explicitly labelled negative samples, both methods search for features that can separate positively labelled samples and the rest unlabelled data as much as possible. They describe the inter-class separation structure based on distributional characteristics of the available labelled positive samples. Given extremely limited number of labelled positive samples, the estimated distributional attributes can be severely biased. Therefore, it is still an open issue to extract information about the inter-class separation in an efficient way.

In this paper, we proposed a novel PU learning algorithm to attack the open problem, named Positive Unlabelled Feature Selection (PUFS). It is designed to integrate robust partially supervised graph embedding and sparse regression into a joint optimisation framework. Our contributions are threefold. Firstly, we encode the weakly supervised information in PU learning tasks into pairwise constraints between training instances. These constraints are imposed on graph embedding of the entire training samples to make it consistent with underlying class distribution. Secondly, we propose to measure violation of the pairwise constraints generated from the unlabelled data using a robust metric. It is defined to suppress the adverse effects of the potentially mislabeled constraints due to ambiguous affinity relation. Finally we cast feature selection into a regression procedure, which considers correlation among features and is able to evaluate joint feature combinations.

2 Notation and Background

2.1 Notations

Let $X \in R^{n \times m}$ denote the training data set. Each row $X_i \in R^m$ in X is a m-dimensional training instance. $X_{i,j}$ denote the j-th dimension of X_i. $Y = \{Y_1, Y_2 ..., Y_n\} \in \{0, 1\}$ denotes the true labels of x_i where $Y_i = 0$ if X_i is positive and $Y_i = 1$ if X_i is negative. In PU learning, Y is partially presented. We use $P = \{X_i^p\}$ to denote labelled positive training data and $U = \{X_j^u\}$ to denote unlabelled training data. In our work, we use K-Nearest Neighbouring (KNN) affinity graph to describe local affinity structure around each training sample, following the work in spectral clustering [Shi and Malik, 2000]. Similarity relationship between training data points is captured by an affinity graph $G = (V, E, S)$. V represents training data X. Each edge $e_{i,j} \in E$ with a non-negative weight $S_{i,j} \in S$ measures the similarity between data points X_i and X_j. $S_{i,j}$ is defined as in Eq.1:

$$S_{i,j} = \begin{cases} \exp\left(-\frac{\|X_i - X_j\|^2}{\theta^2}\right), X_i \in N_k(X_j), X_j \in N_k(X_i) \\ 0, \text{otherwise} \end{cases} \tag{1}$$

where $N_k(X_i)$ is the set of K-nearest neighbours of X_i. Furthermore we use $L = I - D^{-\frac{1}{2}} S D^{\frac{1}{2}}$ to denote the normalised graph laplacian of G, where D is a diagonal matrix with $D_{i,i} = \sum_{j=1}^{n} S_{i,j}$.

2.2 Preliminaries

Graph embedding of an affinity graph G is a feature dimension reduction technique [Yan *et al.*, 2007]. It is defined as a low-dimensional vector representation to preserve data similarity relations between vertex in G. The graph preserving criterion is given as follows:

$$\min_{W, W^T W = I} Tr(W^T X^T L X W) \tag{2}$$

where $XW \in R^{n \times d}$ is the linear graph embedding of training instances, which projects original training samples into a low-dimensional manifold. Similar data samples are tuned to stay close on the graph embedding space, which strengthens the data similarity structure. $W \in R^{m \times d}$ stores the linear projection coefficients, mapping data points from original feature space to d-dimensional graph embedding space. For binary classification, W is reduced to a m-dimensional vector and the resultant graph embedding is a scalar value. The linear graph embedding is frequently used for feature extraction and feature selection. The linear form is helpful for measuring features' contribution in classification and clustering. Therefore, we borrow the basic idea of linear graph embedding in our work.

We use two types of pairwise constrains in this paper. Must-link constraints M specify that two samples should be assigned into one class, and Cannot-link constraints C specify that two samples should be assigned into different classes. Previously, such pairwise constraints are used popularly in weakly supervised learning tasks: an expected classifier should minimise intra-class differences and maximise inter-class differences simultaneously. Recently, constrained spectral clustering technology also employs the pairwise constraint as complementary regularisation terms [Chatel *et al.*, 2014] to separate different classes apart.

3 Positive-Unlabelled Feature Selection

3.1 Robust Positive-Unlabelled Graph Embedding

We establish a robust partially supervised graph embedding procedure characterising latent class distribution for PU feature selection. To this end, there are two fundamental problems to solve. Firstly, we need to incorporate the weakly one-class supervised information into the graph embedding in a feasible way. Furthermore, since only limited positive samples are present, any further estimation about class distribution can be noisy. Mislabelling data points either as positive or negative class is unavoidable. Therefore, it is important to suppress the impacts of the resultant noisy pairwise constraints.

In order to strengthen class separation structure, we encode the partially supervised information by introducing pairwise link (must-link and cannot-link) based regularisation terms into the standard graph embedding, in order to strengthen class separation structure. Our design is motivated from Marginal Fisher Analysis (MFA) [Yan *et al.*, 2007]. Our purpose is to use the pairwise links to evaluate intra-class compactness and inter-class separability of latent class distribution directly. A graph embedding precisely representing information about class distribution should maximise both intra-class compactness and inter-class separability simultaneously. Another popular choice to insert label information into graph embedding can be found in graph based semi-supervised learning methods, such as label propagation [Zhu *et al.*, 2003]. In these methods, graph embedding is treated as pseudo labels of training data. For labelled data points, their pseudo labels are made consistent with the given labels by minimising least square error between the pseudo and true labels. However, given limited number of labelled points, label information propagated from the labelled points to the distant unlabelled ones can be ambiguous [Chapelle *et al.*, 2006]. It adds adverse effects to feature selection. By comparison, pairwise constraints strengthen the intra-class and inter-class distributional structure inside data, which provides a sensitive metric to evaluate features' discriminating power [Yan *et al.*, 2007].

In positive unlabelled learning framework, we firstly construct a plausible negative sample set based on data similarity structure, in order to measure inter-class separability. We calculate the average similarity ω_i^u of each unlabelled data point X_i^u to the positively labelled samples X^p:

$$\omega_i^u = \frac{1}{|P|}\sum_{j=1}^{|P|} S_{X_i^u, X_j^p} \tag{3}$$

We rank the unlabelled training data samples in descending order of the average similarity scores. R unlabelled samples with the lowest average similarity values are selected as the most plausible negative samples, denoted by $PN = \{X_i^{PN}\}_{i=1}^{R}$. Must-link constraints M and cannot-link constraints C are constructed based on positively labelled samples P and the plausible negative samples PN. The measurement of violation of the pairwise constraints is defined as Eq.4. They are used as regularisation to graph embedding, tuning the low-dimensional representation consistent with the weakly supervised information.

$$\begin{aligned} D(W) = &\sum_{i,j\in M} \|exp(-\frac{(X_iW - X_jW)^2}{{\theta_M}^2}) - 1\|^2 \\ &+ \sum_{i,j\in C} \|exp(-\frac{(X_iW - X_jW)^2}{{\theta_C}^2})\|^2 \end{aligned} \tag{4}$$

θ_M and θ_C are the variance parameters of the Gaussian functions. W is the unitary projection coefficient vector. Instead of using fisher criterion directly, we cast the pairwise link based regularisation to a regression model with respect to W. The benefits are two-folds. Firstly, measuring intra- and inter-class difference using euclidean distance needs no prior assumption on the data distribution of each class. In contrast, fisher discriminative analysis assumes the data in each class follow Gaussian distribution, which is not satisfied in most real-world data sets. Thus, our definition is more general for extracting discriminative description of class distribution. Secondly, the Gaussian function maps the infinite sum of euclidean distances monotonically to a bounded range $[0, 1]$. In this way, violation of must-link and cannot link constraints can be measured as a regression procedure with bounded target. Minimising the regression error D drives the graph embedding to be consistent with the pairwise constraints.

The mislabelling issue is coupled with the definition of violation measurement. Ambiguity of data affinity relations introduces mislabelled links into the must-link and cannot-link based constraints. Such ambiguity arises from irrelevant and noisy features, or intrinsically non-linear classification boundary of the training data samples. Especially, it is difficult to decide whether two samples should be grouped into one class, if their affinity level lies in the middle and ambiguous range between extremely strong and weak linkage. These noisy pairwise constraints can add adverse effects to graph embedding learning. In this work, we propose to utilise correntropy induced metric (CIM) [Weifeng *et al.*, 2007] as a robust violation measurement for the potentially noisy pairwise constraint, as illustrated in Eq.5:

$$\begin{aligned} \hat{D}(W) = &-\sum_{i,j\in M} G_\delta(exp(-\frac{(X_iW - X_jW)^2}{{\theta_M}^2}) - 1) \\ &- \sum_{i,j\in C} G_\delta(exp(-\frac{(X_iW - X_jW)^2}{{\theta_C}^2})) \end{aligned} \tag{5}$$

where $G_\delta(x)$ is the Gaussian kernel with δ as its variance. Mean square error used in the Eq.4 increases quadratically with large violation of the pairwise constraints. It is thus prone to mis-labelled constraints. Differently, CIM based metric has a close-to-constant penalty for violation, which avoids overweighting mislabelled pairwise constraints that are consistently violated during learning process. Although $\hat{D}$ has a non-linear form and is difficult to minimise directly, we will present an efficient gradient descent based solution in the next section. By comparison, handling inter-class separability usually requires to solve a more complex optimisation problem, such as generalised rayleigh quotient [Yan *et al.*, 2007] or trace ratio [Wang *et al.*, 2014]. Combining Eq.2 and Eq.5 , we can rewrite the robust positive unlabelled graph embedding as Eq.6.

$$\begin{aligned} W = &\underset{W}{argmin}\, Tr(W^T X^T LXW) + \alpha\hat{D}(W) \\ &\text{s.t. } W^T W = I \end{aligned} \tag{6}$$

where α is the penalty parameter balancing the impacts of the regularisation terms. We impose the orthogonality constraint on W, preventing W from arbitrary scaling. Note we don't enforce the projection XW to be orthogonal as in linear graph embedding

3.2 The Objective Function and Optimisation Algorithm

We cast the positive unlabelled feature selection procedure to a partially supervised subspace learning task. The linear projection coefficients W are the basis of feature importance evaluation. To locate informative feature subsets, L_1 norm based constraint is imposed on W to control the capacity of W's entries. It is equivalent to applying Laplacian prior on W and forcing W be sparsely valued. Therefore the graph embedding is estimated using only a small set of features, which have non-zero entries with large magnitudes in W. These features are identified as the best feature subset recovering discriminative class distribution information carried within the graph embedding. By adding the L_1 norm constraint, the objective function of the proposed PUFS algorithm is given by

$$\min_{W} Tr(W^T X^T LXW) + \alpha\hat{D}(W) + \beta\|W\|_1 + \gamma\|W^T W - I\|^2 \tag{7}$$

where $\|W\|_1$ is the L_1 norm of W. It is worth noting that we adopt normalised laplacian L for simplicity. Other more sophisticated forms of Laplacian matrix can be also utilised. $\gamma > 0$ is the penalty parameter to control the orthogonality condition on W. It should be sufficiently large to guarantee the orthogonality. By defining the objective function (Eq. 7), we embed both selection of informative features and reweighing noisy constraints into the learning process and solve the two problems jointly.

We propose to solve the non-linear optimisation problem as an alternative update procedure using the half-quadratic technique [Yuan and Hu, 2009]. Based on the theory of convex conjugated functions [Yuan and Hu, 2009], we can derive the proposition forming the base of the solution:

Proposition 1 There exists a convex function φ of $G_\delta(x)$ such that

$$G_\delta(x) = \max_{g}(g\frac{\|x\|^2}{\delta^2} - \varphi(g)), \tag{8}$$

and for a fixed x, the maximum is reached at $g = -G_\delta(x)$ [Yuan and Hu, 2009]. Substituting Eq.8 into Eq.7, we can derive the augmented objective function with the auxiliary variable g as follows:

$$\begin{aligned} g_M^{t+1} &= G_\delta(\exp(\frac{-(X_iW^t - X_jW^t)^2}{{\theta_M}^2}) - 1) \\ g_C^{t+1} &= G_\delta(\exp(\frac{-(X_iW^t - X_jW^t)^2}{{\theta_C}^2})) \\ \hat{D}(W) &= \sum_{i,j\in M}\|\exp(\frac{-(X_iW^t - X_jW^t)^2}{{\theta_M}^2}) - 1\|^2 g_M^{t+1} \\ &\quad + \sum_{i,j\in C}\|\exp(\frac{-(X_iW^t - X_jW^t)^2}{{\theta_C}^2})\|^2 g_C^{t+1} \end{aligned} \tag{9}$$

Minimising $\hat{D}$ with respect to W in Eq.9 is a weighted regression task. The auxiliary variables $g_{i,j\in M}^{t+1}$ and $g_{i,j\in C}^{t+1}$ are used to underweight the potentially mislabelled pairwise constraints that are consistently violated during iterative optimisation. Since L_1 norm based regularisation is non-smooth, we further approximate it using a smooth penalty, namely reweighed squared L_2 norm, as suggested by [Candes and Tao, 2005]. The smoothed objective function is given by Eq.10:

$$\begin{aligned} W^{t+1} &= \underset{W}{argmin}\, Tr(W^T X^T LXW) + \alpha\hat{D}(W) \\ &\quad + \beta W^T\Phi^t W + \gamma\|W^T W - I\|^2 \end{aligned} \tag{10}$$

where Φ^t is a diagonal matrix. Its diagonal entry $\Phi_{i,i}^t = \frac{1}{\|W_i^t\|^2}$ is calculated using W estimated in the precedent iteration. Solving Eq.10 is then performed using gradient descent. We initialise the gradient descent procedure randomly. The descent is stopped after maximum iterations or when the Frobenius norm of the gradient vector is less than a given threshold. In all experiments of our work, the gradient descent converges within maximum 15 iterations. Reweighed L_2-norm based smoothing is commonly used in compressed sensing. It provides a linear rate of convergence for L_1 minimisation problem under the restricted isometry property [Candes and Tao, 2005]. Though this property is not satisfied in this work, it doesn't change feature selection results according to empirical results.

Table 1: Database description

Dataset	$\|P\|$	$\|N\|$	# of features
USPS	750	750	241
COIL	750	750	241
G241	748	752	241
BGP	5000	5209	19

4 Experiments

We first perform the experiments to verify the effectiveness of PUFS on three semi-supervised learning benchmark datasets[1] - USPS, COIL and G241[Chapelle *et al.*, 2006]. Additionally, we conduct experiments on a real-world cyber security dataset - BGP hijacking events [Vervier *et al.*, 2015] - to demonstrate the practical usability and value of PUFS. The characteristics of four datasets are summarised in Table 1.

To illustrate verify the merits of PUFS, we compare the proposed PUFS with the other five state-of-the-art feature selection algorithms, including RFS [Nie *et al.*, 2010], JELSR [Chenping Hou and Wu, 2011], NDFS [Li *et al.*, 2012], Pairwise Ranking based Feature Selection (PRFS) [S.Sundararajan and Keerthi, 2011] and apuCFS [Borja Calvo and A.Lozano, 2009]. Firstly, the supervised method, RFS, is built using the potentially noisy negative samples. Although $L_{2,1}$ norm based metric used in RFS improves its robustness against potentially mislabelled negative data, the amount of labelled data is still limited and it doesn't consider data affinity structure as a complementary regularisation term. Thus the overfitting issue can deform the performance of feature selection. JELSR and NDFS are unsupervised feature selection methods following a similar joint learning framework of spectral graph embedding and sparse regression. They neglects the supervised information. PRFS and apuCFS select features based on the distributional attributes of positive samples estimated from the labelled positive samples. If labelled positive data are limited, the biases of the estimated distributional attributes can deteriorate their performances. In contrast, the proposed PUFS selects features not only separating positive samples from plausible negative samples, but also requires the selected features to preserve as much as possible the affinity relation between training samples. This design reduces the overfitting risk. Therefore, PUFS should be able to locate more powerful feature subsets than the these opponents in PU learning scenarios.

The number of plausible negative samples PN is chosen as the same to that of P in order to make a balanced labelled data

[1] They are available from http://www.kyb.tuebingen.mpg.de/ssl-book.

Table 2: Classfication Accuracy on Real-World BGP Hijacking Dataset.

	FP=0.05		FP=0.1		FP=0.15		AUC	
TPR	5 feat.	10 feat.	5 feat.	10 feat.	5 feat.	10 feat.	5 feat.	10 feat.
PUFS	**0.6628**	**0.7912**	**0.8227**	**0.8696**	**0.8589**	**0.9274**	**0.8589**	**0.9274**
RFS	0.2315	0.5725	0.3489	0.5320	0.6085	0.6744	0.3647	0.81
JELSR	0.3251	0.2732	0.3281	0.4495	0.5225	0.6559	0.8019	0.8015
NDFS	0.3296	0.3311	0.3353	0.4495	0.5225	0.6559	0.5225	0.6559
PRFS	0.4811	0.6053	0.5074	0.6605	0.6170	0.7065	0.8207	0.8795
apuCFS	0.2771	0.5009	0.3021	0.6071	0.4605	0.7354	0.8098	0.8850
Baseline	0.4895	0.4895	0.683	0.683	0.7027	0.7027	0.7027	0.7027

set for all the feature selection methods. For NDFS, JELSR and PUFS, the size of neighbourhood (k) of KNN affinity graph is specified to be 10 for all datasets. For fair comparison, we tune the regularisation parameters of all the feature selection algorithms using cross-validation before launching the experimental analysis. The best results are reported for all the algorithms. In the proposed PUFS, δ and γ in Eq.7 are fixed at 10^3 and 10^5 for all datasets, providing consistent results. We determine α, θ_M and θ_C in Eq.4 by grid search and finally fix them as 1, 10 and 40 in the experiments respectively. We also study the sensitiveness of the three parameters by varying each parameter around the chosen value, from 80% to 120% of its absolute value. The AUC values of the proposed PUFS at all FP levels are stable within the specified ranges. For space reasons, we don't illustrate sensitivity analysis in figures. We randomly select 80% of the entire data as training data, and the rest 20% as testing data. For each partition, all six feature selection algorithms are performed on the training data and select N best features. In the training data set, we choose randomly 10% of the positive training samples as labelled data P and treat the rest as unlabelled data U. This is designed to simulate the real world PU learning scenario, such as BGP hijacking events, where positively labelled samples are extremely limited.

To evaluate feature subsets selected by different feature selection algorithms, a linear support vector machine (SVM), is built using 5-fold cross-validation on the test data set with these feature subsets. Average ROC curve is derived from the cross-validation test. Area-Under-Curve (AUC) value of the average ROC is used as the overall metric to evaluate the classification accuracy. We also extract true positive rate (TPR) on the average ROC curve given the false positive rate (FPR) fixed at 5%, 10% and 15% as local and finer metrics of feature selection algorithms. We focus on low FPR range since FP level is an important criterion to evaluate practical usability of a classification system. Too many false alarms make the decision output of the system untrusty and useless. The aforementioned process is repeated 15 times for any give N. AUC and TPR values in each round are averaged to measure final classification performance.

4.1 Classification Accuracy

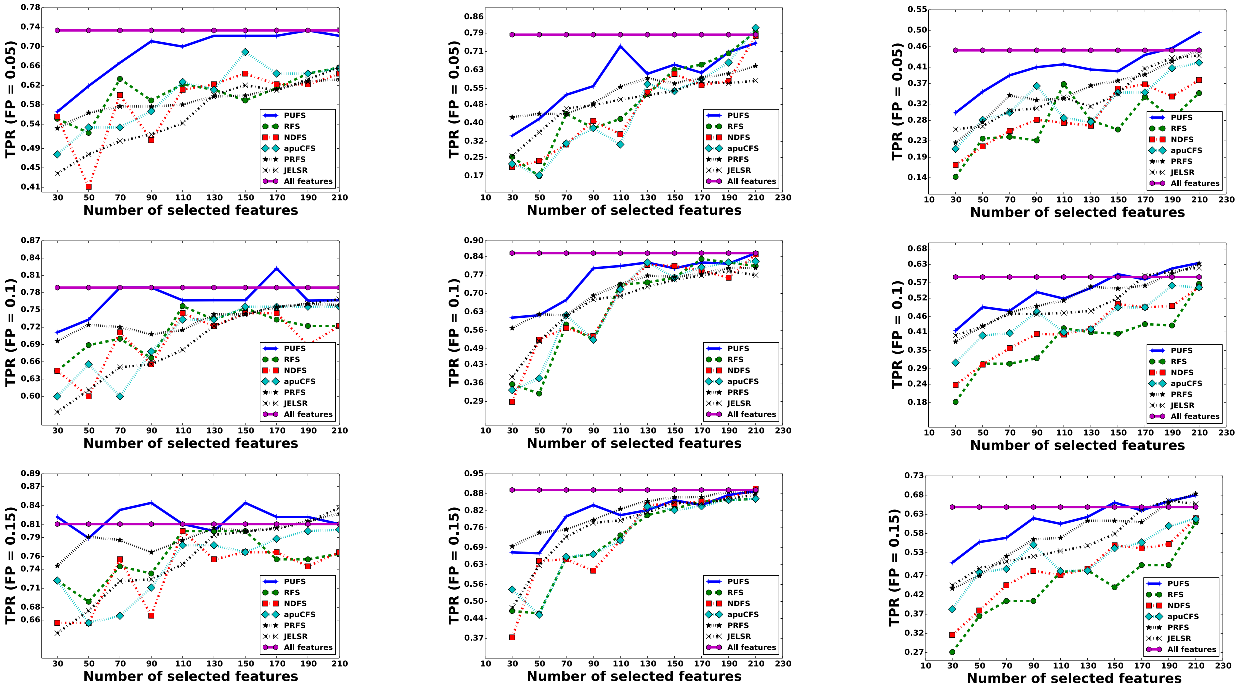
We set the number of selected features N as {30,50,70,90,110,130,150,170,190,210} for all the three benchmark data sets. Figure.1a, Figure.1b and Figure.1c, illustrate true positive rate metrics on the three public benchmark datasets. Figure.3 shows the overall AUC metric of each algorithm derived with different feature subsets. From the results, it is clear that the proposed PUFS is superior to the other feature selection methods. In general, It achieves consistently better classification performances, while selecting a smaller set of discriminative features on all 3 benchmark datasets. On USPS and G241 datasets, the proposed PUFS has higher TPR cross all FP levels and consistently higher AUC values no matter how many features are selected for classification use. On both of the datasets, the proposed PUFS manages to identify the discriminative feature subsets that perform closely or even better than the baseline with all features used for classification. On COIL dataset, the proposed PUFS presents better performances when FPR is lower than 15% and comparable to the PRFS when FPR is 15%. Similarly, AUC values of the proposed PUFS are close to that of PRFS, while distinctively better than the rest feature selection methods in the comparison.

PRFS performs the best among the feature selection methods except the proposed PUFS. The pairwise rank SVM used in PRFS separates the available positive samples and the unlabelled samples. This design helps to explicitly highlight the features that are potentially informative for binary classification, though the unlabelled set contains both negative and positive samples. Superior performances of PUFS and PRFS indicate a principle of positive unlabelled feature selection: information about interclass separation is important for identifying discriminating features. Different from PRFS, the proposed PUFS selects the most plausible negative samples from the unlabelled set to extract more stable inter-class separation structure and suppress the potential noise in the plausible negative samples with robust statistics. As a result, the proposed PUFS performs better over PRFS.

4.2 Experiment on Real World BGP Hijacking Dataset

BGP hijacking detection [Vervier *et al.*, 2015] is intrinsically a PU learning problem. Due to large variance of BGP announcing mechanisms from different Autonomy Systems, only a small number of suspicious BGP announcements can be manually labelled by security experts. Our BGP dataset contains 10,209 real world BGP announcements collected between April and October 2014. Each sample is labelled by security experts as either benign or malicious (labelled as either normal activity or BGP hijacking event).

We set the number of selected features from $\{3, 5, 9, 11\}$ and follow the same parameter settings in Section 4.1. Table.2 summarises classification performances of all in-



(a) USPS

(b) COIL

(c) G241

Figure 2: Classification Accuracy on USPS, COIL and G241.

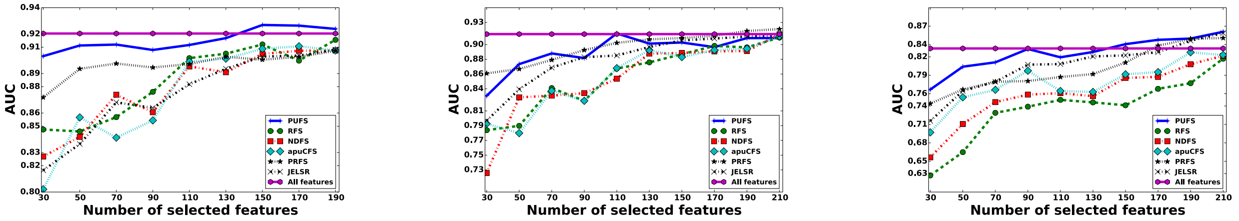


Figure 3: AUC Scores on USPS, COIL and G241.

volved algorithms using different feature subsets. Given each fixed size of feature subsets, the features selected by the proposed PUFS achieve superior classification precision at all FP levels over the other feature selection algorithms. Especially when the number of selected features is limited, e.g. only 5 features selected. For example, we set FP to 5% to limit the false alarm rate of hijacking event detection. At this FP level, the proposed PUFS achieves almost 2 times larger TPR than the other feature selection methods. Compared with the baseline using the all features, the proposed PUFS performs better with much smaller feature sets. This observation indicates the validity and superior performance of the proposed PUFS in practical PU learning tasks in real-world scenarios.

5 Conclusion

The proposed PUFS algorithm is defined as a partially supervised subspace learning procedure. Discriminative information about class distribution, namely intra-class compactness and inter-class separability, is incorporated into a robust and smooth objective function to conduct joint feature selection. We demonstrate the advantage of the proposed PUFS by extensive experiments on both standard benchmark databases and a real-world cyber security application. Our approach consistently outperforms state-of-the-art unsupervised and semi-supervised feature selection methods under positive-unlabelled learning scenarios. Although our approach is designed originally to handle PU learning problems. It can be also applied without further modification in multi-class partially supervised classification, where limited labelled samples come from only a subset of the classes.

Acknowledgments

We thank Pierre-Antoine Vervier from Symantec Research Labs and Xiangliang Zhang from King Abdullah University of Science and Technology gracefully for helpful discussions.

References

[Borja Calvo and A.Lozano, 2009] Pedro Larranage Borja Calvo and Jose A.Lozano. Feature subset selection from positive and unlabelled examples. *Pattern Recognition Letters*, 30:1027–1036, 2009.

[Cai *et al.*, 2010] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In *KDD*, pages 333–342, 2010.

[Candes and Tao, 2005] Emmanuel Candes and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51:4203–4215, 2005.

[Chapelle *et al.*, 2006] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[Chatel *et al.*, 2014] David Chatel, Pascal Denis, and Marc Tommasi. Fast gaussian pairwise constrained spectral clustering. In *ECML/PKDD*, pages 242–257, 2014.

[Chenping Hou and Wu, 2011] Dongyun Yi Chenping Hou, Feiping Nie and Yi Wu. Feature selection via joint embedding learning and sparse regression. In *IJCAI*, pages 1324–1329, 2011.

[Duda and P.E., 2001] Hart Duda, R.O. and Stork D.G. P.E. *Pattern Classification*. Wiley New York, 2001.

[Elkan and Noto, 2008] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabelled data. In *KDD*, pages 213–220, 2008.

[He *et al.*, 2005] Xiaofei He, Deng Cai, and Niyogi P. Laplacian score for feature selection. In *NIPS*, 2005.

[Kononenko *et al.*, 1997] Igor Kononenko, Edvard Simec, and Marko Robnik-Sikonja. Overcoming the myopia of inductive learning algorithms with relieff. *Journal of Applied Intelligence*, 7, 1997.

[Li *et al.*, 2012] Zechao Li, Yang Yi, Jing Liu, Xiaofang Zhou, and Hanqing Lu. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI*, pages 1026–1032, 2012.

[Nie *et al.*, 2008] Feiping Nie, Shiming Xiang, Yangqing Jia, Changshui Zhang, and Shuicheng Yan. Trace ratio criterion for feature selection. In *AAAI*, pages 671–676, 2008.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint l2,1-norms minimization. In *NIPS*, pages 1813–1821, 2010.

[Shi and Malik, 2000] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22, 2000.

[S.Sundararajan and Keerthi, 2011] Priyanka Garg S.Sundararajan and S.Sathiya Keerthi. Pairwise ranking based approach to learning with positive and unlabelled examples. In *CIKM*, pages 663–672, 2011.

[Tibshirani, 1996] Robert Tibshirani. Regression shrinkage and selection via lasso. *J R Stat Soc Series B*, 58, 1996.

[Vervier *et al.*, 2015] Pierre-Antoine Vervier, Olivier Thonnard, and Marc Dacier. Mind your blocks: On the stealthiness of malicious bgp hijacks. In *NDSS*, 2015.

[Wang *et al.*, 2014] Xiang Wang, Buyue Qian, and Ian Davidson. On constrained spectral clustering and its applications. *Data Mining and Knowledge Discovery*, 28, 2014.

[Weifeng *et al.*, 2007] Liu Weifeng, Puskal.P.Pokharel, and Jose.C.Principe. Correntropy:properties and applications in non-gaussian signal processing. *IEEE Transactions on Signal Processing*, 55:5286–5298, 2007.

[Yan *et al.*, 2007] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A generative framework for dimensionality reduction. *IEEE TPAMI*, 29, 2007.

[Yang *et al.*, 2011] Yi Yang, Hengtao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. L2,1-norm regularized discriminative feature selection for unsupervised learning. In *IJCAI*, pages 1589–1594, 2011.

[Yuan and Hu, 2009] Xiaotong Yuan and Baogang Hu. Robust feature extraction via information theoretic learning. In *ICML*, pages 1193–1200, 2009.

[Zhang *et al.*, 2012] Zhihong Zhang, Edwin R.Hancock, and Xiao Bai. Hypergraph spectra for semi-supervised feature selection. In *ECML/PKDD*, pages 207–222, 2012.

[Zhu *et al.*, 2003] Xiaojin Zhu, Zoubin Ghahramani, and John Laffery. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.