Li, Z., Li, H. & Shao, L. (2019). Improving attribute classification with imperfect pairwise constraints. In *Proceedings of The 19th International Conference on Electronic Business* (pp. 253-262). ICEB, Newcastle upon Tyne, UK, December 8-12.

Improving Attribute Classification with Imperfect Pairwise Constraints

(Full Paper)

Zequn Li*, Northumbria University, Newcastle Upon Tyne, UK, Zequn.li@northumbria.ac.uk
Honglei Li, Northumbria University, Newcastle Upon Tyne, UK, Honglei.li@northumbria.ac.uk
Ling Shao, Inception Institue of Artifical Intelligence, Abu Dhabi, UAE, ling.shao@ieee.org

ABSTRACT

Semantic attributes extracted from images could help to improve many interesting applications, including image classification, recommendation systems and online advertising. However, learning of such attributes requires a large well-labelled dataset which is usually difficult and expensive to collect and sometimes requires human domain experts to annotate. Partially labelled data, on the contrary, are relatively easy to obtain from social media websites or be annotated by less experienced people. However, a partially labelled dataset usually contains a lot of noisy data which are challenging for previous methods. In this paper, we propose a semi-supervised Random Forest algorithm that can handle a small well-labelled attribute dataset and large scale pairwise data at the same time for classifying grouped attributes. Results on two typical attribute datasets show that the proposed method outperforms the state-of-the-art attribute learner.

Keywords: Machine Learning, Pairwise data, Imperfect Data

*Corresponding author

INTRODUCTION

Semantic attribute learning attracted a considerable growth of interest from computer vision and multimedia researchers in the last few years. By using those mid-level features, the accuracy of high-level category classifiers could be improved, and the learned attributes can be adopted as semantic features for other multimedia applications. For example, using Computer Vision to recognize and describe fashion items at the semantic level is extremely helpful for analyzing personal fashion styles (Hu et al. 2015). However, the training of a semantic model needs a large set of well-labelled data (Bossard et al. 2012; Di et al. 2013; Hu et al. 2008) and as the fashion items increase every year, there is also a strong demand to update the current model with new data. The previous work mostly generated their datasets with the help of crowdsourcing (Chen et al. 2012; Liu et al. 2012) which is time consuming and expensive. This a need to design attribute learning algorithms that do not require many well-labelled data.

One possible solution is transfer learning (Pan and Yang 2010) or domain adaptation (Ben-David et al. 2010) converting well trained models from a previous well-labelled dataset to the new dataset with limited labelled data. However, the limitation of both transfer learning and domain adaptation lies in that it is required that the well-labelled and new dataset to be related. This requirement is easy to meet when only category level information is considered. In the real scenario, it's very common that the concentration is on the attribute level categorization which are mostly dataset specified (e.g. dressing style in clothes dataset or sociality in animal dataset) and not related. Therefore, we consider another possible solution of using unlabelled and partially labelled data which can be easily obtained from the Internet or quickly annotated by humans. Utilizing both well-labelled and unlabelled/partially labelled data is a typical semi-supervised learning scenario. However, partially labelled data usually contains a lot of noisy information, especially for abstract attributes such as the style of clothing or the design of fashion items.

There are many existing approaches on semi-supervised learning. Some consider it as a clustering problem without taking advantage of the well-labelled data at all (de Amorim 2012; Hu et al. 2008; Wagstaff et al. 2001; Zeng and Cheung 2012; Zhu et al. 2016) and others use a well labelled dataset along with totally unlabelled data (Liu et al. 2013) but ignore the weak relationship in the unlabelled data. A most related solution to our work was presented by Nguyen and Caruana (2008)and Yan et al. (2006) who used a well labelled dataset together with additional partially labelled data, but their methods do not work well with noisy data.

This paper aims to provide a solution to solve the problem of learning attributes from partially labelled noisy data together with well-labelled data. The contribution of our work is two-fold. Firstly, our model is built on a small set of well-labelled training data with a large amount of noisy pairwise constraint data. Secondly, the pairwise data is collected according to semantic groups (e.g. color or texture) and only two pairwise constraints (i.e. must-link, meaning a pair of samples must be in the same class, and cannot-link, meaning a pair of samples must come from different classes) are applied. The objective is to utilize such pairwise data to prove semantic attribute learning.

LITERATURE REVIEW

We review the related works in two aspects, semantic attribute analysis and semi-supervised learning.

Semantic Attribute Analysis

Attribute learning as a specific image classification prob- lem has gained lots of interest and resulted in many successful applications (Bossard et al. 2012; Chen et al. 2012; Di et al. 2013; Liu et al. 2012). Such approaches usually use low-level features to train many mid-level classifiers, and then the mid-level attribute information can be used to generate a more accurate high-level classifier or to analyze the semantic information of images. However, the training of mid-level classifiers relies on well-labelled data that are hard and expensive to collect. Furthermore, the selected at- tributes are mostly simple ones such as color or texture which do not contain much expert domain knowledge. Thus, the use of partially labelled data rather than well-label data makes the previous work more realistic.

Semi-supervised Learning With Pairwise Constraints

Semi-supervised learning is a class of machine learning techniques that makes use of unlabelled or partially labelled data for training. In this paper, we specify the partially labelled dataset as pairwise constraint data and analyze the existing work, which can handle pairwise constraint relations.

To utilize pairwise constraints, some researchers consider it as a clustering problem and develop constrained clustering algorithms based on K-means such as COP K- means (Wagstaff et al., 2001) and CMWK-Means (de Amorim, 2012). With a similar idea, a more efficient Semi-supervised Maximum Margin Clustering method was also introduced to handle the pairwise constraints (Zeng & Cheung, 2012; Hu et al., 2008). The drawback of these algorithms is that they make a strong assumption that the pairwise constraint data are accurate. However, in practice, pairwise information is usually collected from social media sites that makes it contain many noisy data. To avoid the influence of noisy data, the Constraint Propagation Random Forest which attempts to prevent the noise impact by an ensemble is formulated (Zhu et al., 2015). Compared to the previous algorithms, it performs better when dealing with noisy data. However, the limitation of these algorithms is that they only use the pairwise constraint data to do clustering rather than classification. To build attribute classifiers, a small set of well-labelled data should be utilized along with a large amount of pairwise data.

Taking labelled data into consideration, Liu et al. (2013) proposed a Random Forest based approach which trains trees with both labelled and unlabelled data, and as an ensemble method it is also robust for noisy data. However, this algorithm does not take the relation among unlabelled data into consideration, but only uses them as back- ground knowledge to find better splits. To use labelled data and pairwise information at the same time, Convex Pairwise Kernel Logistic Regression which builds a loss function with pairwise constraint information was introduced by Yan et al. (2006). Similar with this work, a Margin-based approach was also proposed (Nguyen and Caruana 2008). In this method, a regular multi-class SVM is extended with a pairwise optimization objective function. These two algorithms use both well-labelled and partially labelled data, but according to our experiments, they are not robust to noisy pairwise data and do not consider the different qualities be- tween well-labelled data and pairwise data. In addition, based on Spectral Kernel Learning, Shang et al. (2012) proposed a semi-supervised classification algorithm with enhanced spectral kernel under the squared loss (ESKS) which also takes labelled data and pairwise labelled data into consideration. However, the target dataset of ESKS is with large labelled data and small number of unlabelled or partially labelled data. That is different with the situation we considered in our work which the dataset contains small number of labelled data and large number of partially labelled data.

ROBUST RANDOM FOREST WITH PAIRWISE CONSTRAINTS

A fully supervised dataset usually includes a set of labelled training samples $L = \{xi \in X\}$ i=1...l and their labels $\{yi \in Y\}$ i=1...l where $X \subset {}'R$ is the feature space, Y = {1...k}(Chen et al. 2012) is the label set and l is the number of samples. In this paper, in addition to the labelled data, there are also partially labelled data, which include Must-Link $M = \{(x_i^\alpha, X_i^\beta) \| y_i^\alpha = y_i^\beta\}_{i=1\ldots m}$ and Cannot-Link $C = \{(x_i^\alpha, X_i^\beta) | y_i^\alpha \neq y_i^\beta\}_{i=1\ldots n}$. In our learning framework, we consider a more realistic situation where the partially labelled data containing many noisy information, which is different from the previous work proposed by Nguyen & Caruana (2008). The labelled dataset, on the other hand, is assumed the well-constructed one with fewer noisy data.

Random Forest With Supervised Learning

Random Forest is a widely accepted ensemble method to handle noisy data (Breiman 2001). It consists of a list of decision trees $\{t_1, t_2, \ldots, t_N\}$ independently trained by a random subset of the whole data and the results are generated by getting votes from all trees. Growing each decision tree involves data selection, node splitting and stopping criterion detection. According to a

predefined sub-sample rate r, the data selection phase is to randomly select a subset from the whole dataset (Ho 1998). Feeding the selected data into decision trees, each node splits the data into two parts with a splitting strategy. Considering both accuracy and efficiency, the Linear Combination Splits (Geurts et al. 2006) is used more frequently. The function is defined as

$$h(W,\theta) = \begin{cases} 0, W \cdot x < \theta \\ 1, otherwise \end{cases} \quad (1)$$

where W is the parameter with one or more non-zero elements to select features, and θ is the split threshold. According to the output of $h(W,\theta)$, all arriving samples are split into either the left or right child node. To find the optimal split parameter, the criterion ΔG is introduced and formulated as

$$\Delta G(R) = G(R) - \frac{R_l}{R} G(R_l) - \frac{R_r}{R} G(R_r) \quad (2)$$

where R refers to the current node, Rl, Rr represent the attempted split to left and right child nodes. The function G can be computed by many methods such as information gain and Gini impurity (Breiman 1984). In this paper, we choose Gini impurity $\sum_{i \neq j} p_i p_j$ due to its efficiency. The p in Gini impurity reflects the proportion of samples belonging to the same category, and it can be calculated as:

$$p_l = \frac{1}{|R|} \sum_{i=1}^{|R|} [y_i = 1] \quad (3)$$

where |R| is the number of samples in current node R and [yi = l] refers to indicator function, it equals to 1 when yi = l and otherwise equals to 0. The target of information gain is to maximize ΔG by selecting different W and θ. To make it more efficient, maximum attempting number m_{trv} is defined as a stopping criterion. After training, each tree in Random Forest gives a robability estimation of class l, $p_t(l|x)$ for a given test case x and the total probability of random forest is calculated by averaging

$$p(l \mid x) = \frac{1}{N} \sum_{i=1}^{n} p_i(l|x) \quad (4)$$

where N is the number of trees and $p_i(l|x)$ is obtained by calculating the ratio of class l getting votes from the leaves in the ith tree. The result of Random Forest is defined as

$$\hat{l} = \underset{l \in Y}{argmax}\, p(l|x) \quad (5)$$

Node Splitting With Pairwise Constraints

The conventional Random Forest described above only takes labelled data as input and the limited size of supervised data in our problem would lead to an obvious performance drop (Liu et al. 2013). To avoid this problem, we extend the current splitting strategy to take pairwise constraint data into training. According to equation 2, the target of split is to maximum the Gini impurity at each node that requires to obtain the proportion of well labelled data belonging to the same category. However, for partially labelled data, samples come in pairs, so when the Must-link or Cannot-link relation is broken by splitting, the gini index calculated by $\sum_{i \neq j} p_i p_i$ does not work. In this situation, we introduce a new method for both Must-link M pairs and Cannot-link C pairs.

Equation 6 obtains the number of samples in Must-link set M falling into the same node, and Equation 7 calculates the total number of samples from Must-link set M in the node R.

$$N^M(R) = 2 * |\{(x^\alpha, x^\beta) | x^\alpha \in R \wedge x^\beta \in R \wedge (x^\alpha, x^\beta) \in M\}| \quad (6)$$

$$N^M_{total}(R) = |\{x^\alpha | x^\alpha \in R \wedge (x^\alpha, x^\beta) \in M\}| + |\{x^\beta | x^\beta \in R \wedge (x^\alpha, x^\beta) \in M\}| \quad (7)$$

where R denotes the current tree node. Like Must-link, we obtain the same data from Cannot-link set C with Equation 8 and 9

$$N^C(R) = 2 * |\{(x^\alpha, x^\beta) | x^\alpha \in R \wedge x^\beta \in R \wedge (x^\alpha, x^\beta) \in C\}| \quad (8)$$

$$N_{total}^{C}(R) = |\{x^{\alpha} | x^{\alpha} \in R \wedge (x^{\alpha}, x^{\beta}) \in C\}| + |\{x^{\beta} | x^{\beta} \in R \wedge (x^{\alpha}, x^{\beta}) \in C\}| \quad (9)$$

Based on N and N_{total} defined above, we propose two estimation functions as follows:

$$E^{M} = -\log \frac{N^{M}(R)}{N_{total}^{M}(R)} \quad (10)$$

$$E^{C} = -\log \frac{N_{total}^{C}(R) - N^{C}(R)}{N_{total}^{C}(R)} \quad (11)$$

Equations 10 and 11 are the estimation functions for Must-link and Cannot-link sets respectively. These two equations are constructed by the ratio between successful splitting number and total samples. In addition to making it more robust for noisy data, we apply log function to it, which makes derivative of it smaller when the successful number is close to the total number.

According to Equation 2, we can propose a similar split criterion for pairwise constraint data evaluating the tree before and after splitting.

$$\Delta E(R) = E^{M}(R) - \frac{R_l^M}{R^M} E^{M}(R_l^M) - \frac{R_r^M}{R^M} E^{M}(R_r^M) + E^{C}(R) - \frac{R_l^C}{R^C} E^{C}(R_l^C) - \frac{R_r^C}{R^C} E^{C}(R_r^C) \quad (12)$$

The new target becomes to find a split to maximize Equation 12.

In addition, our algorithm considers a more complex condition that the pairwise data include a lot of noise information, when the procedure of tree construction closes to the leaf nodes, the total number of pairwise data $N_{total}^{M}(R)$ and $N_{total}^{C}(R)$ could be smaller and Equation 10 and 11 would too sensitive to noise data. In order to avoid this problem, we introduce a combined split strategy as follows:

$$\Delta C(R) = \begin{cases} \Delta G(R) + \alpha \Delta E(R), |L| < |M| + |N| \\ \Delta G(R), otherwise \end{cases} \quad (13)$$

Where |L|, |M|, |N| refers to the number of well-labelled, must-link and cannot-link data respectively, and α is the learning rate for pairwise data. In practice, we usually choose a small number for α which makes $\Delta E(R)$ only have a limited influence at tree construction.

In this paper, we assume a situation that the size of partially labelled data is much larger than that of the well-labelled data, but the accuracy is on the contrary. To calculate Equation 13 more efficiently, the samples with broken Must-link or satisfied Cannot-link will be removed from child nodes as shown in Equations 14 and 15.

$$M^{new}(R) = \{(x^{\alpha}, x^{\beta}) | x^{\alpha} \in R \wedge x^{\beta} \in R \wedge (x^{\alpha}, x^{\beta}) \in M\} \quad (14)$$

$$C^{new}(R) = \{(x^{\alpha}, x^{\beta}) | x^{\alpha} \in R \wedge x^{\beta} \in R \wedge (x^{\alpha}, x^{\beta}) \in C\} \quad (14)$$

The tree constructing procedure is shown in Algorithm 1.

Evaluation Of Trees With OOB Error

In our work, we consider partially labelled data containing a lot of noisy information, and the Random Forest, as an ensemble method, reduces the influence of noisy data by aggregating results from the individual trees. However, the splitting strategy described above considers all the data from both labelled and partially labelled datasets, which means the noisy data still have contributions to the result of each tree. Therefore, when the noise rate of partially labelled dataset is large, most of the trees in Random Forest will be affected by the wrong information and the result from Random Forest could be less accurate. Therefore, there need to be an evaluation method to select training samples.

The training of each tree in the Random Forest is independent and only uses a subset of training data. The samples that are not used for training are defined as the Out Of Bag (OOB) data (Breiman 2001) and the error calculated from it is known as an unbiased estimation of the accurate of trees. Therefore, to reduce the influence of noisy data in a partially labelled dataset, each tree is trained on only the labelled data or both labelled and partially labelled data. If the partially labelled dataset will be randomly generated for training new trees. With the help of this method, a subset containing too much noisy data will be ignored and a new one will be generated. The tree selecting procedure is shown in Algorithm 2.

EXPERIMENTS AND ANALYSIS

To test the performance of Pairwise Constraint Random Forest, we build two different attribute level datasets based on the Clothing with Attributes dataset (Farhadi et al. 2009) and Animal with Attributes dataset (Lampert et al. 2009). Considering the efficiency, the algorithm proposed in this paper is implemented based on the code provided online1.

Algorithm 1: Construct Random Forest Tree

Input: Well-labelled training data $X_l \in L$ in current node *R*, pairwise data $(x^\alpha, x^\beta) \in \{M \cap C\}$ in current node *R*, learning rate for pairwise data α, maximum splitting attempting number m_{try}

Output: The best cut hyperlane W and θ, the associated child node R_l and R_r

1 Romove unnecessary pairwise training data according to Eq 14 and 15;
2 if $x_l < (X^\alpha + X^\beta)/2$ **then**
3 Set split strategy $\delta C = \Delta G(R) + \alpha \Delta E(R)$
4 end
5 else
6 Set split strategy $\delta C = \Delta G(R)$
7 end
8 Initialize *W, θ*;
9 $n_{try} \leftarrow 0$;
10 $\delta C_{best} \leftarrow 0$;
11 repeat
12 $n_{try} = n_{try} + 1$
13 randomly select W_0 *and* θ_0;
14 **calculate** δC according to selected split strategy;
15 **if** $\delta C_{best} < \delta C$ **then**
16 $\delta C_{best} \leftarrow \delta C$;
17 $W \leftarrow W_0, \theta \leftarrow \theta_0$;
18 **end**
19 until $n_{try} > m_{try}$;
20 Generate R_l, R_r according to *W and θ*;
21 return W, *θ*, and R_l, R_r;

Dataset settings

Clothing with Attributes Dataset

The Clothing with Attribute dataset includes 11 different attribute groups and totally 36 attributes. According to the learning framework described above, a small labelled dataset and a large pairwise dataset should be generated. To make the pairwise dataset more realistic, we choose 5 major colors as the color group, 6 different patterns as the pattern group, 3 different lengths of sleeve

[1] https://github.com/karpathy/Random-Forest-Matlab

and 3 types of neckline shape as another two groups. We only keep 300 well-labelled data and randomly generate pairwise datasets with 800 and 1500 images. To simulate the noisy situation, the four attribute groups are manually added noisy data with 0%, 5%, 10%, 15% and 20% respectively.

Because the Clothing with Attribute dataset only provides the original images, we apply the pose estimation method by Yang and Ramanan (2013)to locate the clothing area. And with the help of VLFeat (Vedaldi and Fulkerson 2010) and scikit-learn (Pedregosa et al. 2011), low level features including SIFT features, Color Histogram features, Local Self-Similarity features and PHOG features are extracted. Then we use PCA to reduce the dimension and keep 99% information.

Algorithm 2: Select Random Forest Tree

Input: Well-labelled training data $X_l \in L$, pairwise data $(x^\alpha, x^\beta) \in \{M \cap C\}$ maximum training times n
Output: Random Forest Tree T

$x_j^i \leftarrow$ generate a new subset from X_l using bootstrap aggregation;
Training tree T_l with X_j^i;
$oobe_l \leftarrow$ Compute the OOB error of T_l;
$n_{try} \leftarrow 0$;
$best_{oobe} \leftarrow oobe_l, best_T \leftarrow T_l$;
repeat
  $n_{try} = n_{try} + 1$
  $(X^\alpha, X^\beta) \leftarrow$ generate a new subset from M and C using bootstrap aggregation;
  Training tree T with both $X_l^i \; and \; (X^\alpha, X^\beta)$;
  $oobe \leftarrow$ Compute the OOB error of T;
  If $best_{oobe} > oobe$ **then**
    $best_{oobe} \leftarrow oobe, best_T \leftarrow T$;
  end
until $n_{try} > n$;
return $best_T$;

Animal With Attributes Dataset

For the Animals with Attributes dataset, there are to- tally 85 different attributes in it, in our experiment, we randomly select 5 attributes as the color group, 4 attributes as the texture group, 5 attributes as the living place group and finally 5 attributes as the sociality group. Because this dataset is much bigger and much nosier than the previous one, we select 400 samples as the well-labelled dataset and 1000 and 3000 samples as pairwise data. Then the pairwise dataset is also manually added noisy information at 0%, 5%, 10%, 15%, 20%. As the low-level features are provided by authors (Lampert et al. 2009), we directly use these features and also apply PCA to them and just maintain 99% information.

Reduce The Influence Of Noisy Data

Random Forest uses a subset of training data to build each tree. The sub-sampling rate is the number of samples in a subset divided by that in the whole dataset (Breiman 2001; Liu et al. 2013). In traditional Random Forest, the sub- sampling rate is manually defined before training and the chance of samples to be selected into a subset is equal during training. After the sub-sampling, the data selected by each tree will all contribute to the evaluation for node splitting.

However, in Pairwise Constraints Random Forest, according to Algorithm 1 the pairwise data used in training could be ignored and the influence of them is limited. As we described in Section 3.3, to reduce the influence of noisy data, we evaluate each tree with OOB error and keep generating new subsets from the partially labelled dataset that means the randomly generated subsets containing too many noisy data will be deleted. As a result, even if we fix the sub-sampling rate before training, the chance to choose noisy data into a subset is less than the chance to choose normal data. In the experiment, because the dataset is not balanced, we choose different sub-sampling rates for labelled and unlabelled data as 0.9 and 0.6 respectively. Analyzing the procedure of Pairwise Constraint Random Forest working on both dataset, table 1 shows the average chance of different samples to be selected in all trees during training. The results indicate that the effect of noisy data is reduced by the methods we proposed in this paper.

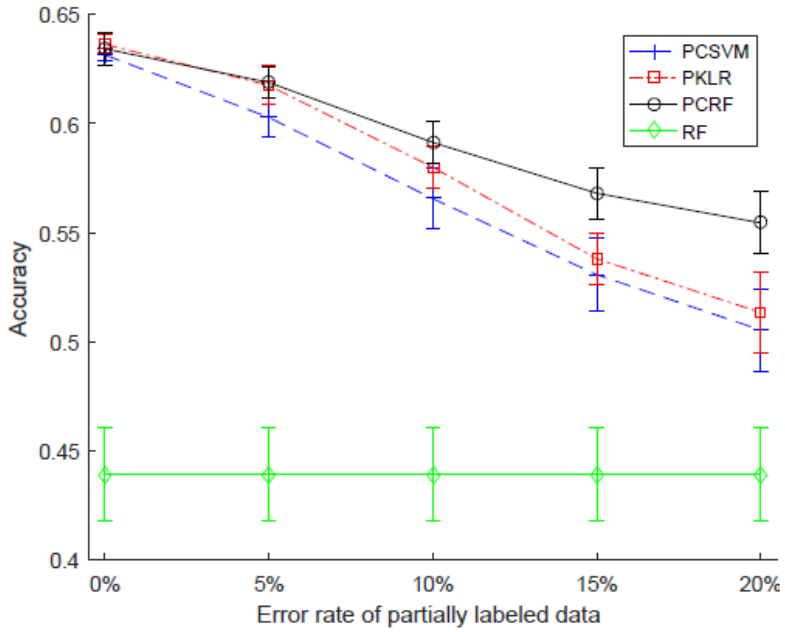
Table 1 Chance of different data to be selected for training					
Noise rate	0%	5%	10%	15%	20%
Well-labelled data	90%	90%	90%	90%	90%
Partially labelled data	60%	60.60%	61.87%	63.57%	66.63%
Noisy data	60%	48.51%	43.20%	39.37%	33.48%

Compared With Other Algorithms

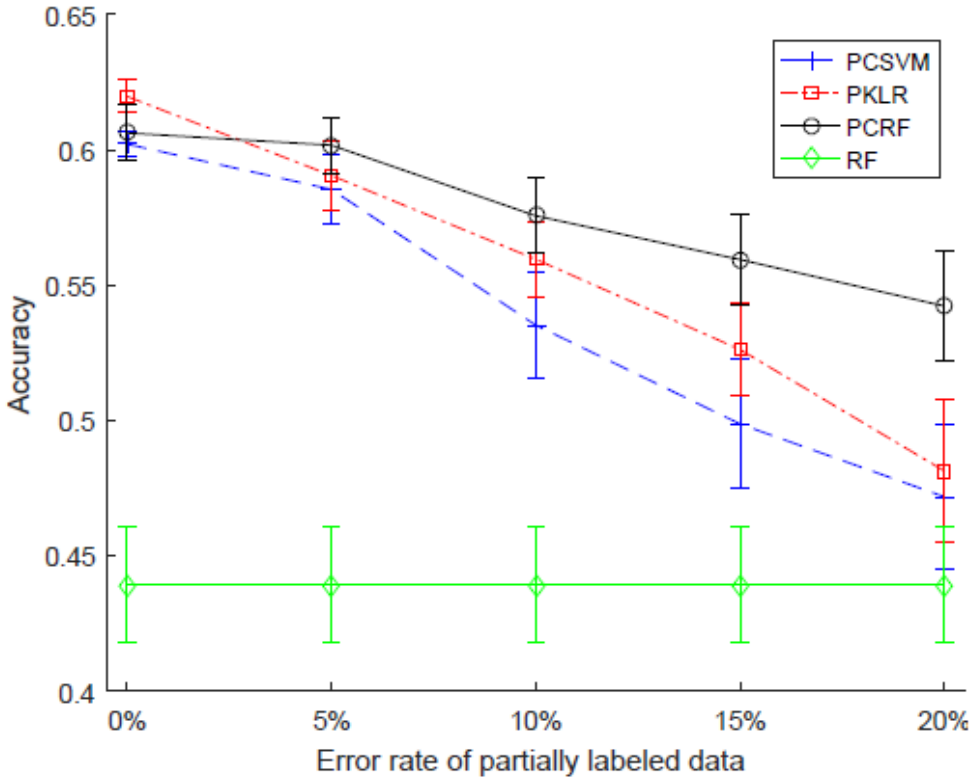
There are limited previous works on classification with a small well-labelled dataset and a large pairwise constraint dataset. As shown in Nguyen & Nam's research (2008), the PCSVM proposed by them reached the state-of-the-art performance and PKLR (Yan et al. 2006) works better when the data is limited. Therefore, we compare our pairwise constraint Random Forest with PCSVM and PKLR in the dataset proposed above. In addition, to illustrate the effect of partially labelled data, we also test a normal Random Forest training with only well labelled dataset.

To get the best performance of each algorithm, we apply 5-fold cross validation to select best parameters. For the pairwise constraint Random Forest we set the number of trees to 150 and the learning rate for partially labelled data is set to 0.01. The dataset used in our experiments contains only limited well-labelled data and a lot of noisy information, so the stability of algorithm is important. To get the variance of different methods, we apply 5-fold validation on testing and choose the average accuracy along with variance as the results. Figure 1 and 2 shows the average performance of all attribute groups with Pairwise Constraint RF against PCSVM, PKLR and the normal Random Forest. When the noise rate is smaller, all algorithms get the similar accuracy, which means all algorithms work well with accurate pairwise constraint data. However, when the noise increases in the pairwise data, Pairwise Constraint Random Forest is more robust and performs better than PCSVM and PKLR.

Compared the results we get from small set and large set partially data in Figure 1 and 2, Pairwise Constraint Random Forest can handle both small and large set of partially data, and the variances of results on both datasets are smaller than others. But the increasing number of partially labelled training data usually result in a more robust module.

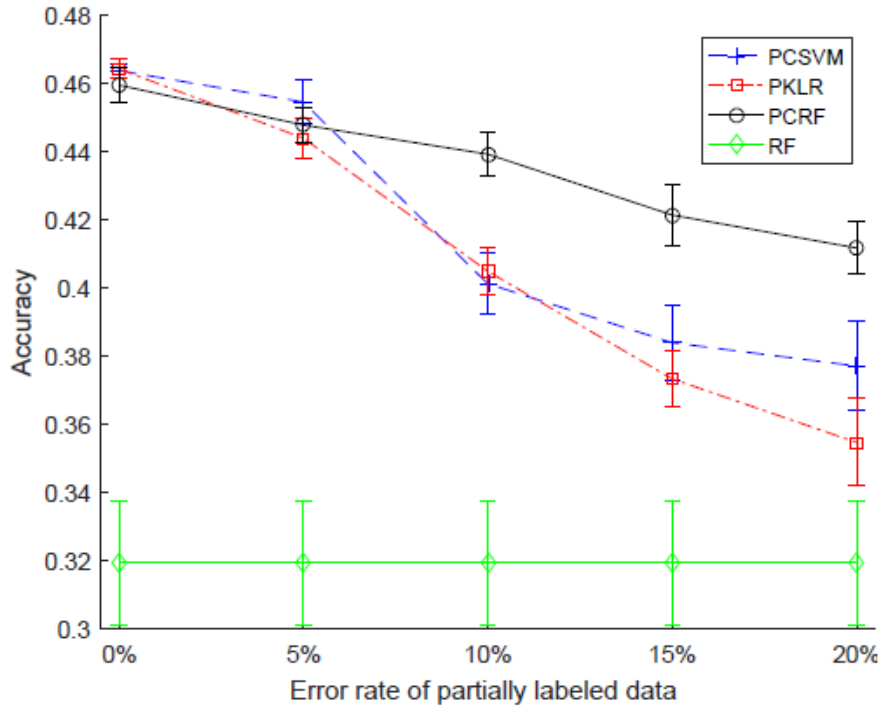


(a) Large set partially labelled training data

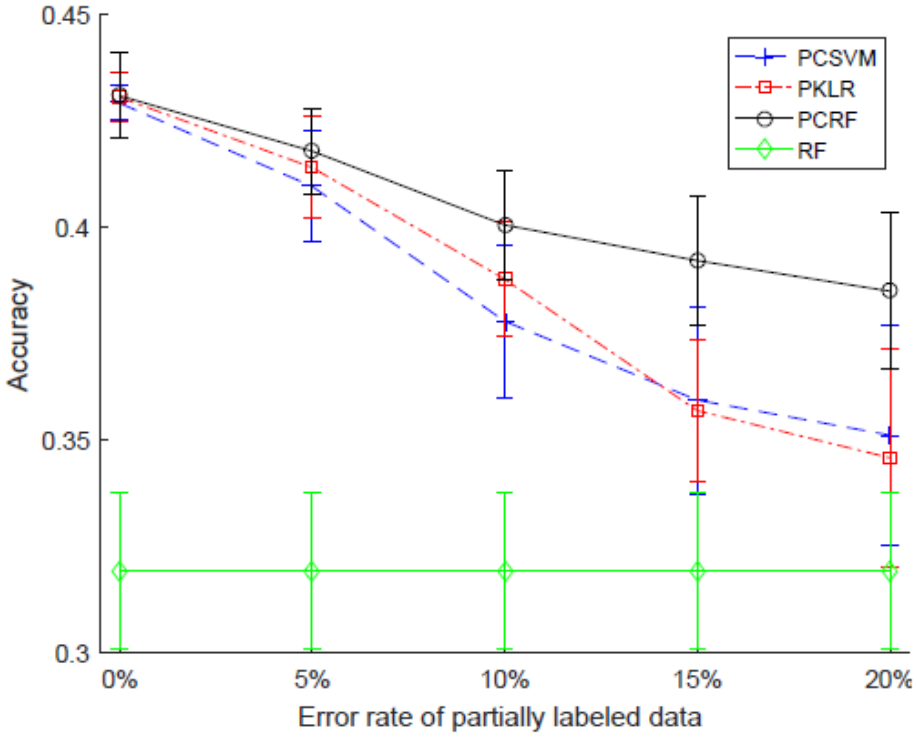


(b) Small set partially labelled training data

Figure 1. Results of Pairwise Constraint Random Forest on Clothing with Attribute dataset



(a) Large set partially labelled training data



(b) Small set partially labelled training data

Figure 2. Results of Pairwise Constraint Random Forest on Animal with Attribute dataset

CONCLUSION

In this paper, we addressed the problem of the lack of well-labelled data for training attribute classifiers, grouped related attributes and proposed a Pairwise Constraint Random Forest which handles well labelled data and imperfect pair- wise data at the same time. In the experiments, we extended two well-known attribute datasets with different noise rates and test the performance of Pairwise Constraint Random Forest on them. The results showed that the Pairwise Constraint Random Forest proposed in this paper works better than the previous methods on a large set of pairwise noisy data. In the future, we plan to use data collected from social media websites and extend this algorithm to work with different constraints.

REFERENCES

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A Theory of Learning from Different Domains, *Machine learning* , *79*(1-2) , 151-175.

Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., & Van Gool, L. (2012). Apparel classification with style. In *Asian conference on computer vision* (pp. 321-335). Springer, Berlin, Heidelberg.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth. International Group, 432, 151-166.

Breiman, L. (2001). Random Forests , *Machine learning ,45*(1), 5-32.

Chen, H., Gallagher, A., & Girod, B. (2012). Describing clothing by semantic attributes. In *European conference on computer vision* (pp. 609-623). Springer, Berlin, Heidelberg.

de Amorim, R. C. (2012). Constrained clustering with minkowski weighted k-means. In *2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI)* (pp. 13-17). IEEE.

Di, W., Wah, C., Bhardwaj, A., Piramuthu, R., & Sundaresan, N. (2013). Style finder: Fine-grained clothing style detection and retrieval. In *Proceedings of the IEEE Conference on computer vision and pattern recognition workshops* (pp. 8-13).

Farhadi, A., Endres, I., Hoiem, D., & Forsyth, D. (2009). Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1778-1785). IEEE.

Geurts, P., Ernst, D., and Wehenkel, L. (2006) . Extremely Randomized Trees, *Machine learning 63*(1), 3-42.

Ho, T. K. (1998). "The Random Subspace Method for Constructing Decision Forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*(8), 832-844.

Hu, Y., Wang, J., Yu, N., & Hua, X. S. (2008). Maximum margin clustering with pairwise constraints. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 253-262). IEEE.

Hu, Y., Yi, X., & Davis, L. S. (2015). Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 129-138).

Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 951-958). IEEE.

Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., & Yan, S. (2012). Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3330-3337). IEEE.

Liu, X., Song, M., Tao, D., Liu, Z., Zhang, L., Chen, C., & Bu, J. (2013). Semi-supervised node splitting for random forest construction. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 492-499).

Nguyen, N., & Caruana, R. (2008). Improving classification with pairwise constraints: a margin-based approach. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 113-124). Springer, Berlin, Heidelberg.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 1345-1359.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12, 2825-2830.

Shang, F., Jiao, L. C., & Liu, Y. (2012). Integrating spectral kernel learning and constraints in semi-supervised classification. *Neural processing letters*, 36(2), 101-115.

Vedaldi, A., & Fulkerson, B. (2010). VLFeat: An open and portable library of computer vision algorithms. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1469-1472).

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001). Constrained k-means clustering with background knowledge. In *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 577-584).

Yan, R., Zhang, J., Yang, J., & Hauptmann, A. G. (2006). A discriminative learning framework with pairwise constraints for video object classification. *IEEE transactions on pattern analysis and machine intelligence,* 28(4), 578-593.

Yang, Y., & Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12), 2878-2890.

Zeng, H., & Cheung, Y. M. (2011). Semi-supervised maximum margin clustering with pairwise constraints. *IEEE Transactions on Knowledge and Data Engineering*, 24(5), 926-939.

Zhu, X., Loy, C. C., & Gong, S. (2015). Constrained clustering with imperfect oracles. *IEEE transactions on neural networks and learning systems*, 27(6), 1345-1357..