Random Forest for Relational Classification with Application to Terrorist Profiling

Jian Xu
Department of Computer Science
Louisiana State University
Baton Rouge, LA 70803, USA
Email: jianxu@csc.lsu.edu

Jianhua Chen
Department of Computer Science
Louisiana State University
Baton Rouge, LA 70803, USA
Email: jianhua@csc.lsu.edu

Bin Li
Department of Experimental Statistics
Louisiana State University
Baton Rouge, LA 70803, USA
Email: bli@lsu.edu

***Abstract*—We study the problem of detecting and profiling terrorists using a combination of an ensemble classifier, namely random forest and relational information. Given a database for a set of individuals characterized by both "local" attributes such as age and criminal background, and "relational" information such as communications among a subset of the individuals, with a subset of the individuals labeled as terrorist or normal people, our task is to design a classifier that captures the patterns of terrorists and achieves good accuracy in predicting the labels of the remaining part of the database. In previous work, a hybrid approach was presented that iteratively applies a flat classifier (such as decision trees, fuzzy clustering) augmented with flattened relational attributes for learning and classification. In the current work, our approach is to use random forest as the "flat" classifier in the terrorist detection setting. Random forest is known to have advantage in handling tasks with high dimensionality in input data. This merit of random forest method is very useful for relational learning if the number of "flattened" relational attributes is quite large, which is indeed the case for the terrorist detection task. We report our experiments on a synthetic terrorist database that compare the prediction accuracy of random forest with two other "flat" classifiers, namely, ordinary decision tree and fuzzy clustering. The experimental results show that random forest classifier outperforms both ordinary decision tree classifier and fuzzy clustering.**

***Keywords*-random forest, relational learning, terrorist profiling**

I. Introduction

In this paper, we study the use of random forest method in a relational learning setting for terrorist detection and profiling. Suppose we have a database of individuals, each described by a set of attributes such as age, ethnic group, past criminal records, etc. Only a subset of the individuals are labeled as terrorists or normal people, whereas the class label (as far as being a terrorist or not is concerned) for many other is unknown. In addition to the attributes pertaining to each individual, the database also contains "relational information" that indicates the connections among the individuals. Examples of such relational information include records of events and the persons participating in such events together, frequencies of two-way or N-way communications among the individuals, etc. The objective of the learning method is to discover a good terrorist-detection model from the training data (those individuals with class label known, plus the relations among them) such that the model generalizes well in predicting the class labels of the test data.

It has been argued in our previous work [1] that in such settings, the learning method should best utilize both the local attributes characterizing each individual traits, and the *relational* information highlighting the connections and associations among the individuals in deriving the classification model. "Flat" (attribute-based) classifiers such as decision trees (DT), while simple and efficient to learn, would mostly ignore the relational information. On the other hand, the *simple relational classifier* RN proposed by Macskassy and Provost [2], based on the idea of"guilty by association", uses *only* relational information.

We proposed a hybrid approach [1] that combines the use of a flat classifier with iterative "guilty by association" using multiple (aggregated) relational attributes in addition to local attributes. We iteratively construct and apply a traditional flat classifier using both local attributes and "flattened" (aggregated) relational attributes. The value of a flattened relational attribute for an object x is estimated (through aggregation) from the labels of the neighbors of x in the current iteration. Multiple relational attributes could be used, in addition to local attributes. Considering both local attributes and relational attributes could be necessary in some applications, as the studies in relational learning indicate. Experiments were conducted [1] on a synthetic terrorist database using the Fuzzy C-means (FC) clustering method as the flat classifier. The results showed quite high classification accuracy.

In the current work, we report our experimental results on applying the random forest method as the flat classifier within the hybrid (flat classifier plus flattened relational attributes) relational learning framework. Random forest (RF) is an ensemble learning method that constructs multiple decision trees from samples of the training data set. In the decision tree induction process, the random forest method uses only a random subset of the available attributes when selecting the best split attribute of a node. The motivation for considering random forest for this relational learning and terrorist detection task mainly comes from the observation that random forest is shown [3] to work well when the

number of attributes is large, and the number of training data points is small, which is exactly the case for the terrorist detection task: We typically do not have many available training data points (not many known terrorists), while the number of relational attributes can be quite large.

The rest of this paper is organized as follows. In Section 2, we briefly describe the random forest learning method and discuss relational learning methods relevant to our study. In section 3, we recapture the hybrid iterative relational learning framework developed earlier [1]. Experimental results of applying random forest method within the hybrid framework to a synthetic terrorist database are presented in Section 4. We conclude in Section 5.

II. Random Forest and Relational Learning

We describe briefly random forest ensemble learning method, and some relational learning algorithms such as the simple relational classifiers RN and RN^* in this section.

A. Random Forest

Random forest, proposed by Leo Breiman [3], is an ensemble algorithm that combines many individual classification trees in the following way. For each tree (1) a bootstrap sample is drawn from the original sample; (2) an unpruned classification tree is built using the bootstrap sample of the data, and at each split the candidate set of variables is a random subset of all the variables. The class status of the response variable is predicted via the majority vote of the predictions of all the trees in the forest.

Random forest uses both *bagging* (bootstrap aggregation [4]), a successful approach for combining unstable learners such as a single classification tree, and random variable selection for tree building. In the random forest algorithm, each tree is unpruned (grown fully), so as to obtain low-bias trees; at the same time, bagging and random variable selection result in low correlation of the individual trees. Hence, the algorithm yields an ensemble that can achieve both low bias and low variance (from averaging over a large ensemble of low-bias, high-variance but low correlation trees). Studies (see e.g. [3]) have shown that the prediction accuracy of the ensemble is usually better than the one from an individual classification tree.

B. Relational Learning and Simple Relational Classifiers

Relational learning has attracted the interest of many researchers in recent years. A number of effective relational learning algorithms such as Probabilistic Relational Models [5] and Relational Bayesian Classifiers [6], have been developed. Relational classifiers such as these are based on the idea that the connections (relations) among the objects should be taken into consideration when performing classification. Thus, they search the relational space for relational neighborhood structures meaningful for classification. While the classification performance of relational classifiers has been demonstrated through various works, typically relational classifiers are quite complex. Thus it is desirable to find a good hybrid of flat and relational learning methods that utilizes all relevant information, yet is simple and efficient.

Macskassy and Provost [2] proposed a very simple relational classifier RN which is based on the notion of "guilty by association": the class label of an object x will be solely dependent on the weighted majority of the class labels of its neighbors. RN assigns the class label c to object x (which is currently labeled as "unknown") if the sum of weights of the neighbors of x belong to class c is maximal among all such weight sums of neighbors of x in a class. Note that for a currently labeled "unknown" object x, if the weight sum for its neighbors labeled as "unknown" of x turns out to be maximal, then the label of x will not be altered. This observation gives rise to an iterative version RN^* of RN, which applies RN iteratively from an initial set of objects with a subset labeled, until all objects are labeled or no more labels could be assigned to any remaining object. Macskassy and Provost [2] show that RN and RN^*, while simple, compete quite well against other more complex relational classifiers. Clearly, these simple relational classifiers are much simpler than their more sophisticated counterparts such as PRM [5]. However these simple classifiers do not consider multiple relations, which could be necessary in some applications in order to get good prediction accuracy.

In [7] and [8], Van Assche et. al. proposed to use random forest for relational learning in the Inductive Logic Programming (ILP) framework. Their method focuses on learning first-order logic formulas. In contrast, we use random forest as simple attribute-based classifier in this study and thus the learned decision trees can be seen as propositional logic formulas.

III. The Hybrid Iterative Learning Framework

In this section we briefly recapture the generic iterative learning algorithm [1], which is in the spirit of RN^* using flat (attribute-based) classifier and data with a combination of local and relational attributes.

The hybrid iterative algorithm starts with a set of partially labeled data set $D = \{\langle e_1, c_1\rangle, \ldots, \langle e_n, c_n\rangle\}$. Here each class label c_j could be a specific class c (like terrorist, non-terrorist) or "unknown". Each object e_j in the data is described by a set of m "local" attributes $\{A_1, \ldots, A_m\}$ plus a set of "relational" attributes $\{B_1, B_2, ..., B_k\}$. A relational attribute B_j is defined by applying an aggregation function f to some weighted relation w over the objects e_1, ..., e_n. For example, if w represents the relation "two-way communication" between two persons, i.e., $w(e, e')$ denotes the frequency of two-way communications between e and e', we can define a relational attribute B_1 such that $B_1(e)$ is the sum of $w(e, e')$ over all e' belonging to the class

"terrorist": the total number of two-way communications of e with known terrorists.

The algorithm applies an attribute-based classifier L iteratively thus generating learned models M_1, M_2, ..., until some termination criterion is met. The core of the algorithm works as follows:

1) $D_0 \leftarrow D$
2) **Repeat** until convergence
   - $M_{i+1} \leftarrow L(D_i)$
   - Apply M_{i+1} to D_i to generate D_{i+1}
   - Update B_1, ..., B_k values for data in D_{i+1}
   - $i \leftarrow i+1$

Applying L to the current dataset D_i would produce the classification model M_{i+1}. Then the learned model is applied to D_i to classify the data points still labeled as "unknown". Thus some of such data points may get its class label assigned in this iteration. This produces the dataset D_{i+1}. The newly assigned class label of an object x (say as "terrorist") may lead to the update of some relational attributes of x's relational neighbors.

Note that the hybrid learning algorithm is rather a generic algorithm framework, which has one "hyper-parameter" L, the flat learning algorithm. Different choices of L will actually produce different specific learning algorithms working in the same fashion. The choice of L is dependent on the classification task at hand, and in general should follow the principle of simplicity and efficiency. In this paper, the L is chosen to be the random forest ensemble learning algorithm.

IV. Applying Random Forest for Terrorist Detection

We applied an instance of the hybrid relational learning algorithm to the task of terrorist detection and profiling using a synthetic terrorist database. The specific flat classifier used is the random forest ensemble learning algorithm.

A. The Synthetic Terrorist Database and Selection of Relational Attributes

We obtained a synthetic terrorist database (see acknowledgments) from which the data for our experiments are extracted. The terrorist database is a relational database consisting of 87 tables, containing information about 204 individuals, their participation in various events, and the connections among them such as two-way and N-way communications. In total, there are 7128 two-way communication events in the database.

Given so many tables capturing various relations among the persons in the database, one could potentially define many relational attributes. After careful analysis of the relations, we extracted 7 relational attributes from the database. They include "total number of times of two-way communications with known terrorists", and "total number of times participating in exploitation of vulnerability or productivity events together with a known terrorist", etc. These 7 attributes have been used in [1] when we tested the performance of the hybrid learning method using the fuzzy C-means clustering as the flat classifier.

Since random forest has been shown to work well when many attributes are available in learning tasks, we further extracted 4 additional relational attributes. This would allow us to compare the performance (classification accuracy) of random forest using different number of attributes.

B. Experiment Setup

We conducted several types of experiments in evaluating the performance of RF-based relational learning method. By performance we mean here the classification accuracy (or error rate) of the learned model to test data, which is disjoint from the labeled training data. Comparisons of performance are made between RF-based, DT-based, and FC-based relational learning methods.

One study concerns the impact of the percentage of "seeds" (data points with class labels known) in the initial data set. Using the synthetic database (with the total number of objects equal to 204), we randomly picked (roughly) 25%, 50%, 75%, 85% of the data points as seeds, and use the remaining part of the data as *test data*. Then the learned decision trees in the random forest are used to predict the class label for the test data. So in the various figures in the following, the horizontal axis shows the number of training data points and the vertical axis corresponds to the classification error rate of the classifiers on training data or on test data. For example, in Fig. 2, the mark "t15n36" below the horizontal axis indicates that the number of seeds is 51 (15+36), with 15 terrorists and 36 non-terrorists labeled. This corresponds to 25% data for training and 75% data for testing.

A tuning parameter in random forest is the size of the candidate set of variables for split at each node, which is denoted as $mtry$. The following Figure 1 shows the averaged classification errors from multiple runs by the RF-based method with various values of $mtry$. From Figure 1, we see a U-shaped curve on classification error rates over different values of $mtry$. This shows a trade-off between the goodness of individual trees and correlation among them. On one hand, when $mtry$ is small (i.e. the candidate set for split is small), each tree is poorly fitted and less correlated to other trees. On the other hand, when $mtry$ is large, each tree is well-fitted and more correlated to other trees than the one with small $mtry$. Hence, we fix $mtry$ to be 5 from now on.

The experiments on the RF-based and DT-based relational learning are done by using the **randomForest** and **rpart** packages in **R**, a free software environment for statistical computing, respectively. They are run on the Lenovo X61 laptop under Windows Vista system. The experiments on the FC-based relational learning are done using a C program implementation on the FreeBSD/i386 system.

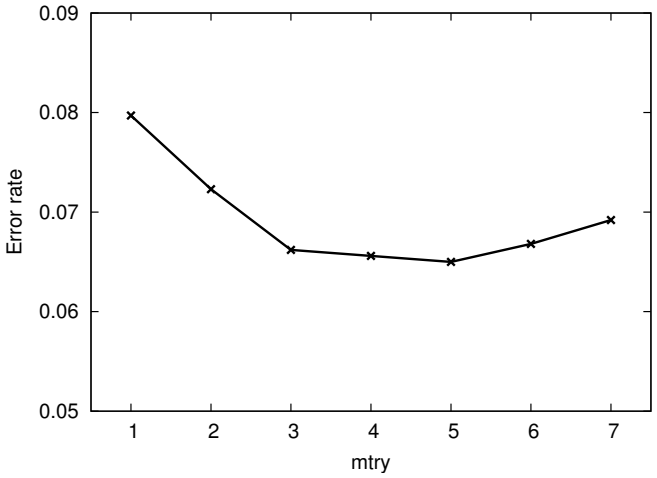

Figure 1. The impact of choice of $mtry$, total number of attributes = 7

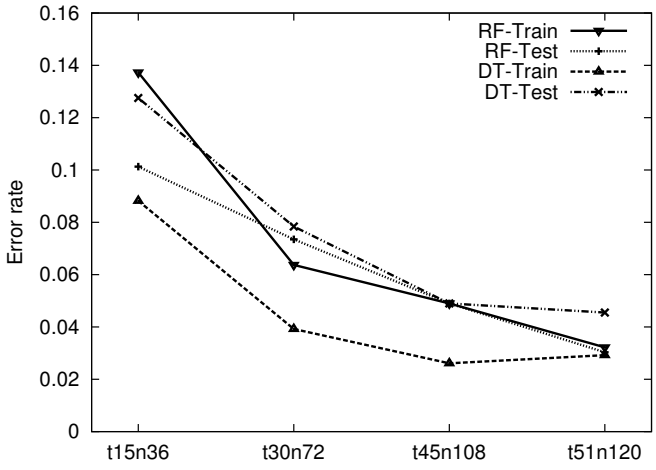

Figure 2. Performance of RF vs. DT, total number of attributes = 7

C. Comparing Random Forest with Decision Tree

Figure 2 and 3 shows the comparison of the error rates between the RF-based method and the DT-based method with seven and eleven attributes, respectively. Based on Figure 2 and 3, we have the following conclusions. (1) RF-based method has lower error rates (than DT-based method) on test data across all cases with different percentage of seeds. (2) Test error rates for RF-based method are decreasing monotonically in both figures. This implies that random forest is more stable than an individual tree. When the training sample changes slightly, an individual may change dramatically, which will further affect the prediction accuracy. (3) The training errors for DT-based method are substantially lower than its testing errors. This indicates that DT-based method tends to overfit the data. For RF-based method, the training and testing error curves are relatively close to each other (particularly in Figure 2), which implies the models are not overfitted.

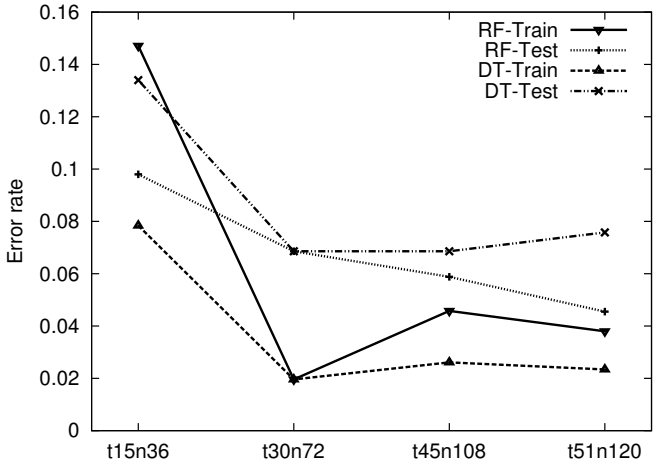

Figure 3. Performance of RF vs. DT, total number of attributes = 11

D. Comparing Random Forest with Fuzzy C-means

Figure 4 shows the comparison of the error rates between the RF-based method and the non-iterative FC-based method. We can see here that RF-based method performs better than FC-based method in term of its lower classification error rate and higher stability. We also compared the RF-based method with the iterative FC-based method using 50% of observations as training data set. The result in Figure 5 shows that RF-based method is comparable to the iterative FC-based method in accuracy, and the error rate from the RF-based method is good enough, hence we do not need iteration here. In case the training set is very small and iteration may be strongly suggested by the single-run results, random forest can be easily adapted in the iterative framework, and we expect the number of iterations needed by RF-based method to be much smaller than that by the FC-based approach. Another advantage of using RF-based method is that there are high quality and free implementations: the original Fortran code from L. Breiman and A. Cutler (http://www.math.usu.edu/∼adele/forests/cc_home.htm), and readily available in many statistical and machine learning software packages such as **R**.

V. Conclusions

We present our study of using random forest for relational learning and its application for terrorists detection and profiling. Random forest is known to work well when the number of attributes is large and the available data is limited. We report our experiments on a synthetic terrorist dataset comparing the performance of the random forest approach with ordinary decision trees and fuzzy clustering. The experimental results show that the random forest method performs better than both other methods. Random forest

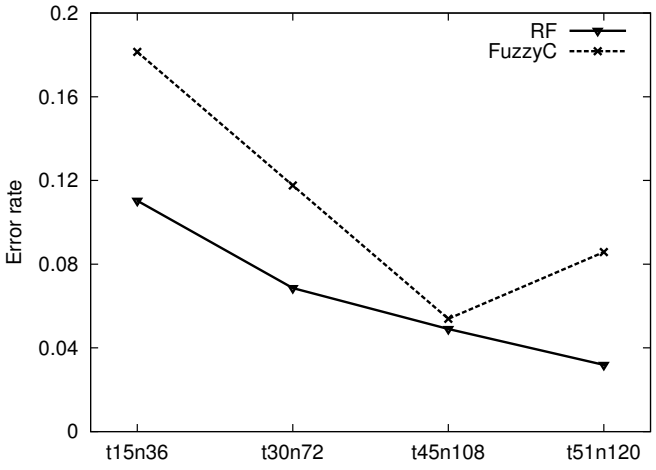

Figure 4. RF vs. FC without iteration, total number of attributes = 7

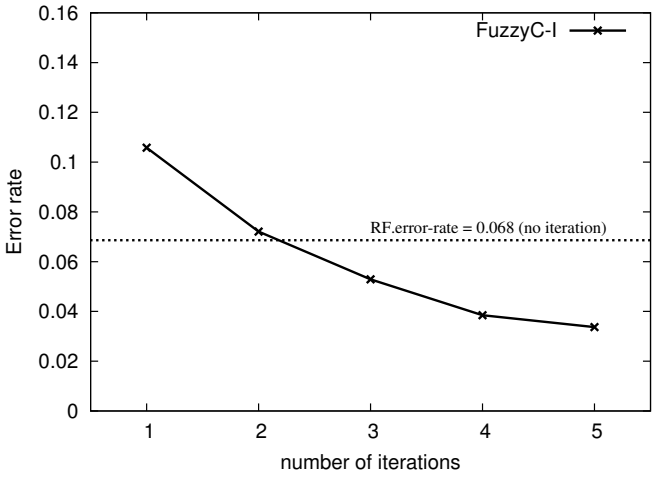

Figure 5. RF vs. iterative FC, total number of attributes = 7

tends to have lower classification error rate, higher stability, and it does not overfit the data.

ACKNOWLEDGMENT

This work is supported in part by NSF grant ITR-0326387, Air Force grant AFOSR FA9550-05-1-0454, and GAANN fellowship. We are grateful to Dr. A. Macskassy, Dr. S. Maripuri and Dr. E. Rickard for providing us the synthetic terrorist database.

REFERENCES

[1] J. Chen, J. Xu, P. Chen, G. Ding, R. F. Lax, and B. Marx, "Fuzzy clustering and iterative relational classification for terrorists profiling," in *the proceedings of the IEEE International Conference on Granular Computing*, Hang Zhou, China, 2008, pp. 142–147.

[2] S. Macskassy and F. Provost, "A simple relational classifier," in *Proceedings of 2nd Workshop on Multi-Relational Data Mining (MRDM)*, 2003.

[3] L. Breiman, "Random forests," *Machine Learning*, no. 45, pp. 5–32, 2001.

[4] ——, "Bagging predictors," *Machine Learning*, no. 24, pp. 123–140, 1996.

[5] B. Taskar, E. Segal, and D. Koller, "Probabilistic classification and clustering in relational data," in *17th International Joint Conference on Artificial Intelligence*, 2001, pp. 870–878.

[6] J. Neville, D. Jensen, B. Gallagher, and R. Fairgrieve, "Simple estimators for relational bayesian classifiers," Department of Computer Science, University of Massachusetts Amherst, Tech. Rep., 2003, technical Report 03-04.

[7] A. V. Assche, C. Vens, H. Blockeel, and S. Dzeroski, "A random forest approach to relational learning," in *Proceedings of the ICML Workshop on Statistical Relational Learning and its Connections*, 2004.

[8] ——, "First order random forests: Learning relational classifiers with complex aggregates," *Machine Learning*, no. 64, pp. 149–182, 2006.