World Congress on Intelligent Control and Automation
June 25 - 27, 2008, Chongqing, China

TCM-RF：Hedging the predictions of Random Forest

Huazhen Wang, Fan Yang, Wujie Wu and Chengde Lin

Department of Automation
Xiamen University
Xiamen, 361005, P.R.China

{hzwang, yang, wjwu, cdlin}@xmu.edu.cn

Abstract- **The output of traditional classifier is point prediction without giving any confidence of it. To the contrary, Transductive Confidence Machine (TCM), which is a novel framework that provides a prediction result coupled with its accurate confidence. This method also can hedge the prediction in which the predicting accuracy will be controlled by predefined confidence level. In the framework of TCM, the efficiency of prediction depends on the strangeness function of samples. This paper incorporates Random forests (RF) into the framework of TCM and proposes new TCM algorithm named TCM-RF, in which the strangeness obtained by RF will be used to implement the confidence prediction. Compared with traditional TCM algorithms, our method benefits from the more precise and robust strangeness measure and takes advantage of random forest. Experiments indicate its effectiveness and robustness. In addition, our study demonstrated that using ensemble strategies to define sample strangeness may be a more principled way than using a single classifier. On the other hand, it also shows that the paradigm of hedging prediction can be applied to an ensemble classifier.**

Index Terms- confidence machine; transductive confidence machine; random forests; hedging prediction

1 Introduction

Most machine-learning classifiers output predictions for new instances without indicating how reliable the predictions are. The application of these classifiers is limited in the domains where incorrect predictions have serious consequences. It is better to couple each prediction with a confidence value [1]. Thus, given the prediction for an instance and corresponding confidence value, a system can decide whether it is safe to classify. Even though progress has been made towards designing reliable confidence machine, most of existing reliable instance classifications has at least one unanswered disadvantages [2].

This paper introduces an effective paradigm to demonstrate reliable instance classification executed by transductive inference learning. It is transductive confidence machine (TCM) that proposed by Gammerman and Vovk[3,4]. The exploiters advanced a welcome preference for formal relationship among Kolmogorov complexity, universal Turing Machines (TMs) and strict minimum message length (MML). They assumed the transductive prediction as a randomness test which returns p values closely associated with the property of the i.i.d distribution governing all of the examples. TCM extends to produce qualified region prediction complemented with a confidence level. Moreover, the error calibration of TCM is controlled by prior significance level. When evaluating the performance of a learning algorithm, it is important to measure error calibration as well as its accuracy. This has been a somewhat neglected aspect of evaluation.

It is a remarkable fact that error calibration is guaranteed regardless of which sample strangeness measure uses. However, the quality of region predictions, and hence the TCM's efficiency, is dependent on the strangeness measurement[5]. It is a general empirical fact that the accuracy and reliability of TCMs are in line with the error rate of the particular classifier plugged into the TCM framework. This issue has been discussed by several authors and several types of classifiers have been used[6], such as (1) support vector machine (TCM-SVM), (2) k-nearest neighbors (TCM-KNN), (3)nearest centroid (TCM-NC), (4) kernel perceptron (TCM-KP), (5) naive Bayes (TCM-NB), and (6) linear discriminant (TCM-LDC). The implementations of these methods are determined by the nature of these classifiers. So TCM-SVM and TCM-KP mainly consider binary classification tasks, TCM-KNN and TCM-KNC is the simplest mathematical realization, and TCM-NB and TCM-LDC is suitable for transductive regression. Indeed, the above methods have demonstrated their applicability and advantages over inductive learning, but there is still much infeasibility. For non-linear datasets, it is especially challenging to TCM-LDC. TCM-KNN and TCM-NC have difficulties with dispersed datasets. TCM-SVM is so processing intensive that suffers from large datasets. TCM-KP is only practicable to relatively noise-free data. In short, there are many restrictions on data qualities when applying these methods to real world data. The difficulties in essence lie in the strangeness measure, which remains an unanswered question.

In addition to the problem of data qualities, there is one more issue to concern. There is always a gap between the definitions of sample strangeness and the real randomness. Finding a precise measure of strangeness assists a good measure of randomness. When designing a measure of strangeness, a conformal transformation to the randomness is

The work is supported by The National High Technology Research and Development Program of China (863 Program No.2006AA01Z129), "National 985 Project" Scientific Innovation Project II

978-1-4244-2114-5/08/$25.00 © 2008 IEEE.

needed. But noise and different parameter settings will influence the value of strangeness measure. Empirical studies on TCMs family have verified this point: parameters settings have great impact on the performance. So, close attention should be paid to robustness of strangeness measure.

Taking into above account, we propose a new algorithm called TCM-RF, which plugs random forest (RF) into TCM, namely, utilizes RF dissimilarity generated by RF proximity matrix to define sample strangeness. This method mainly highlights ideas in two respects: First, RF dissimilarity is not a distance in the Euclidean space but in a "strange" space defined by a collection of trees. It has been proven to be a credible correlation between pairs of samples and giving interesting views of data. Moreover, dissimilarity between pairs of examples is invariant under any permutation of the indices of these examples, which guarantees it a feasible strangeness measure. Second, benefiting from random forest, RF dissimilarity is robust to mixed variable types (categorical, continuous, and semi-continuous) and missing, noisy data. There is an alternative viewpoint that TCM-RF defines the strangeness measure basing on an ensemble (RF) model, not a single model (KNN, KN and SVM). It may provide "a more principled way of designing well measures of strangeness '' [7], because ensemble methods guarantee lower error than average error of individual classifiers [8]. Up to now, RF has not been plugged into the framework of TCM.

The main purpose of this paper is to investigate and demonstrate the efficiency and advantages of TCM-RF. The rest of this paper is organized as follows: Section 2 reviews relevant research related to TCMs, and introduces the principle of hedging predictions of many classifiers, and its realization by transductive inference. Section 3 describes the RF dissimilarity and our new method TCM-RF. Section 4 tests our algorithm on eight datasets. Section 5 concludes the efficiency, calibration and robustness of TCM-RF and discusses the future work.

II OVERVIEW OF HEDGING PREDICTION

2.1 Transductive Confidence Learning

Since training data presented to a machine-learning classifier are finite, Gammerman applied difficult mathematical concepts e.g. algorithmic randomness and Martin-Löf randomness tests to explore the transductive inference learning. Owing to the reliability analysis of prediction and intuitive application of transductive learning, they named it transductive confidence machine (TCM) [3]. Given a training set $T = \{(x_1, y_1),..., (x_{l+1}, y)\} \in X \times Y$, and an unlabeled test instance x_{l+1}, each possible label $y \in Y$ is tried as a label for x_{l+1}. The symbol z will be used as a compact notation for $X \times Y$, so we get the extended sequence:

$$(x_1, y_1),...,(x_{l+1}, y) = z_1,..., z_{l+1} \tag{1}$$

In each try, TCM measures how likely it is that the resulting sequence is generated by the underlying i.i.d distribution P. According to Kolmogorov, an i.i.d distribution sequence means an algorithmically random sequence. Thus, he provided a universal randomness definition to find how random or non-random a specific sequence is, although it is not computable. Martin-Löf extended Kolmogorov's definition of randomness to show its connection with statistical hypothesis tests for sequence randomness level. This extension allows constructing a randomness test, e.g. P-value that can be used in practice.

2.2 The approximate calculation of P-Value

In order to use randomness test and construct P-value in practice, a strangeness measure with each element in the extended training sequence (denoted α_i) is defined:

$$s(z_1,..., z_{l+1}) = \alpha_1,...,\alpha_{l+1} \tag{2}$$

$s(\cdot)$ must be a symmetric function, namely, if we change the order of $z_1,..., z_{l+1}$, the order of $\alpha_1,...,\alpha_{l+1}$ will change in the same way. The sample strangeness is a substantive comparative measure according to all data in sequence (1) of conformability to the underlying i.i.d distribution P.we use the P-value to approximate the conformity of z_{l+1}. It takes advantage of the fact that since the distribution is i.i.d, all permutations of the sequence (2) have the same probability of occurring. So α_{l+1} can take any place in the sequence (2) with the same probability. Thus the probability that α_{l+1} is among the j largest α occurs with probability of at most $\frac{j}{l+1}$. Then P-value, e.g., nonconformity of z_{l+1}, under the label y, is defined as:

$$p = \frac{\#\{i = 1,..., l+1 : \alpha_i \geq \alpha_{l+1}\}}{l+1} \tag{3}$$

If P-value is close to its lower bound $\frac{1}{l+1}$, then example z_{l+1} is very nonconforming with the i.i.d assumption. The closer the p-value is to upper bound 1, the more typical example z_{l+1} is. Hence, P-value indicates how likely tried label y for the unlabeled x_{l+1} is in fact the true label.

2.3 Hedging prediction

TCMs extends to be an efficient way to hedge the predictions produced by many other traditional machine-learning methods i.e., to complement them with measures of their accuracy and reliability. Appropriately chosen, these measures are valid and informative. We specify our understanding of this procedure in Figure1.

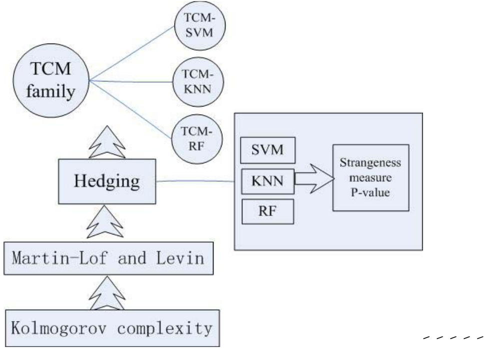

Figure 1 Framework of hedging predictions

According to figure 1, SVM, KNN, NN and any other algorithm (such as RF) can be plugged into the framework of hedging predictions, and join the family of TCMs. But notice that outcome predictions don't come from the mechanism of the original algorithm, but the transductive learning scheme. It is the hedging techniques that make use of the underlying algorithms to give reliable predictions.

TCM-RF: HEDGING PREDICTIONS OF RANDON FORESTS

3.1 Random Forest and RF dissimilarity

Beriman's random forest is one of the most successful ensemble methods. It applies Bagging [9] and Randomization [10] technique to grow many classification trees, which with the largest extent possible without pruning. To classify a new sample, it input the sample down each of the trees in the forest. Each tree gives a classification, and the forest chooses the classification having the most votes over all the trees in the forest.

All the tasks of supervised and unsupervised learning need a suitable measure of dissimilarity or similarity. Obviously, the two issues cannot be well dealt with using a Euclidean measure.

RF naturally leads to a dissimilarity measure between samples in a "strange" space. After a RF is grown, since an individual tree is unpruned, the terminal nodes will contain only a small number of observations. The data are run down each tree, If instance i and j both land in the same terminal node, the proximity between i and j is increased by one. At the end of the run, the proximities are divided by twice the number of trees in the run and similarity between an instance and itself set equal to one. The proximity formed a $N \times N$ matrix [prox (i, j)], it shows that this matrix is symmetric, positive definite and bounded above by 1, with the diagonal elements equal to 1.

The proximity, equivalent to the dissimilarity, is an inherent measure between instances regardless of labels, that is, when changing the label of an instance, its dissimilarity hold invariable. Moreover this dissimilarity distinguishes and outperforms Euclidean measure as the following: Benefiting from robustness of random forest, RF dissimilarity is very robust to noisy, miss data and fixed variables. Furthermore, Y. Qi et al [11] demonstrated its robustness to parameter settings of RF. Thus it can be used as a good input of multi-dimensional scaling or mapping, as well as measure of strangeness.

3.2 Using RF dissimilarity as strangeness measure

Here we make use of the RF proximity matrix to define a new strangeness measure of TCMs, and we expect to find a new way of designing a strangeness measure by using an ensemble method. This work can also be viewed as a hedging prediction of random forest.

To make its function the same as the Euclidean measure and define a similar equation of TCM-KNN, we adopt dissimilarity e.g., 1-prox (i, j), to denote difference between instances.

We put all together P_{ij}^t, which is the proximity from instance x_i to its alike samples j in class t, and P_{ij}^{-t} with proximity from instance x_i to its alike samples j not in class t. then we give

$$P_i^t = \sum_{cl(k)=t} [1 - prox(i, j)]^2 / nclass^t$$

$nclass^t$ refers to the number of the rest of the data in class t. Similarly, $nclass^{-t}$ denotes the number of the data not in class t, let

$$P_i^{-t} = \sum_{cl(k)\neq t} [1 - prox(i, j)]^2 / nclass^{-t}$$

Strangeness measure can be defined as:

$$\alpha_i = \frac{p_i^t}{p_i^{-t}} \qquad (4)$$

The process of our new TCM-RF algorithm is depicted followly:

```
TCM-RF Algorithm
Input: Training set $T = ((x_1, y_1),...,(x_l, y_l))$ and a new unlabeled example $x_{l+1}$.
Output: The set of P-values $\{p^1{}_{l+1},...,p^m{}_{l+1}\}$ when T is a m-class data-set.
for i = 1 to m do
    Assign label i to $x_{l+1}$;
    Construct a classifier RFi using $T \bigcup \{x_{l+1}, i\}$ and output the sample proximity matrix [prox (i, j)].
    Compute strangeness sequence $\{\alpha_1,...,\alpha_l,\alpha^i{}_{l+1}\}$ of all observed examples using [prox (i, j)] ( $\alpha_{l+1}^i$ is the strangeness of $x_{l+1}$ when assigned label i );
    Compute the randomness level, e.g., $p^i{}_{l+1}$ of sequence (1) by (3).
end
```

Given a significance level ε, the P-values above the level ε in the set $\{p^1{}_{l+1},..., p^m{}_{l+1}\}$ make up of a prediction region Γ^ε. To apperceive how effective the prediction region is, we use the following key statistics: (1)certain prediction, percentage of prediction regions with only one label. (2)uncertain prediction, percentage of regions with two or more labels which indicate that all these labels are likely to be correct. (3)empty prediction, percentage of regions that is empty. (4) corrective prediction, percentage of regions which contain the true label and distinguish with traditional accuracy rate made by RF, SVM et al.

When it comes to avoiding empty prediction, TCM-RF outputs an alternative prediction, named forced point prediction, which selects the label with highest P-value in the prediction region Γ^ε. Meanwhile, TCM-RF outputs as credibility the largest P-value, and outputs as confidence one minus the second largest P- value. The credibility measure gives us the credibility of our prediction, and the confidence shows how suitable our prediction.

Ⅳ Experiments and Discussions

4.1 Experimental Setup

In this section we present an experimental evaluation of our approach. This section is divided into three sections: First, we demonstrate the efficiency of TCM-RF; second, we compare classification performances with 2 typical TCMs (TCM-SVM and TCM-KNN); third, an important issue called robustness is discussed.

We apply offline learning, because most practical problems have at least some offline aspects and it is better for demonstrating efficiency by use a larger training set. For conventional comparisons, we use 4 datasets in S.Vanderlooy's technique report[6]which summarizes all existing TCMs. Additionally, to illustrate the advantage of TCM-RF, we employ 4 UCI datasets, including satellite, isolet, soybean, covertype, Some details of the data used in experiment are included in Table 1, which contains information on the number of instances (n), number of class (c), number of attributes (a), and number of numeric (num) and nominal (nom).

Table1 Datasets used in the experiments

name	*n*	*c*	*a*	*num*	*nom*
liver	345	2	7	7	0
pima	768	2	8	8	0
sonar	208	2	60	60	0
house votes	435	2	16	0	16
satellite	6435	6	60	60	0
isolet	300	26	618	618	0
soybean	683	19	35	0	35
covertype	500	3	54	10	44

We perform TCM-RF by applying a ten-fold cross validation process. We report the average performance of all experiments.

4.2 The efficiency analysis of TCM-RF

Given a significance level ε, the efficiency can be laid out. We make the number of trees *ntree* equal to1000 and the number of variables to split on at each node *mtry* be the default $\sqrt{a}$ (a is the number of attributes).On figure 2, we demonstrate TCM-RF efficient curves according with the significance level ε ranging from 0.01 to 1. Limited by the space of paper, we select four graphs of experimental results on pima (continuous variables), soybean (categorical variables), covertype (mixed variables) and liver (poor data quality).

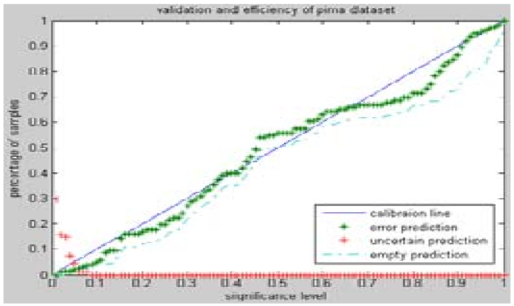

Figure 2(a) the calibration and efficiency on pima

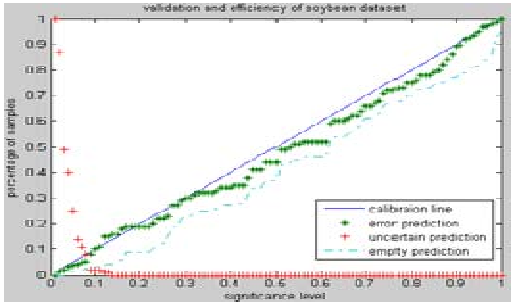

Figure 2(b) the calibration and efficiency on soybean

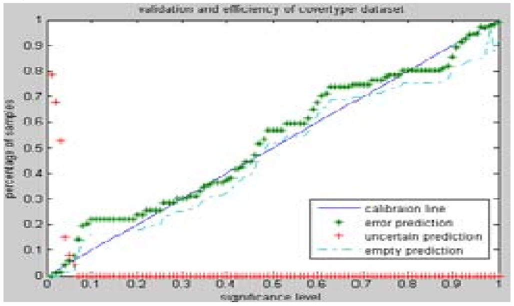

Figure 2(c) the calibration and efficiency on covertype

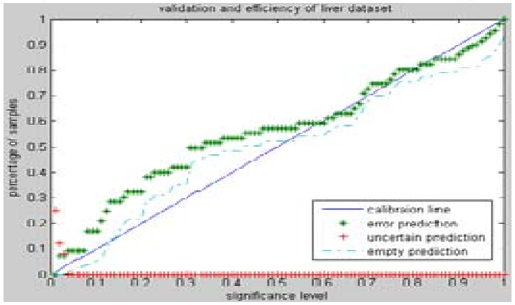

Figure 2(d) the calibration and efficiency on liver

It shows that TCM-RF is well-calibrated up to neglectable statistical fluctuations; the empirical error line can hardly be distinguished from the error calibration line. It allows controlling the number of errors prior to classification. In addition, it is noticed that the percentage of uncertain predictions monotonically decreases with higher significance levels. How fast this decline goes to zero depends on the efficiency of the classifier plugged into the TCM framework.

Percentages of error with a predefined sinigicance level illustrate the calibration of TCM-RF. Some interest points are extracted in table 2 for impressive purpose.

Table 2 corrective prediction under predefined significance level

Significance level	liver	pima	sonar	vote
0.05	0.92	1	0.96	0.98
0.10	0.94	0.92	0.93	1

0.15	0.96	0.98	0.92	1
0.20	0.90	0.98	0.95	1
Significance level	satellite	isolet	soybean	covertype
0.05	0.98	0.88	0.96	0.89
0.10	0.95	0.93	0.94	0.90
0.15	0.96	0.93	0.96	0.91
0.20	0.95	0.92	0.97	0.93

Table 2 demonstrates that TCM-RF ensures relative high accuracy when controlling a low risk of error. It is important in many domains to measure the risk of misclassification, and if possible, to ensure low risk of error.

4.3 Comparison of prediction performance of TCMs

We first compare the percentage of certain prediction because it reflects the efficiency of prediction and is most useful for TCMs. For the convenience of comparison, we apply standard TCM-KNN and TCM-SVM algorithm on the website: http://www.clrc.rhul.ac.uk/research/confidencemachineoverview.htm, which are offered by Gammerman . The performances of sonar dataset are given for comparison in table 3.

Table 3 Comparison of certain prediction on sonar

Percentage of confidence >=	TCM-RF	TCM-KNN	TCM-SVM
99%	53.15%	44.23%	23.07%
95%	77.89%	73.07%	48.07%
90%	86.74%	80.76%	71.15%

We then turn to forced point prediction. Table 4 shows the comparison of TCM-RF, TCM-SVM, and TCM-KNN.

Table 4 the comparisons of forced point accurate prediction

model	liver	pima	sonar	vote
TCM-RF	***66%***	***86%***	84%	***95%***
TCM-KNN	61%	85%	83%	91%
TCM-SVM	51%	77%	96%	84%
level	satellite	isolet	soybean	covertype
TCM-RF	***84%***	82%	***93%***	***83%***
TCM-KNN	82%	70%	89%	74%
TCM-SVM	74%	89%	77%	675

It is clear that TCM-RF performs well at most of the datasets and is especially robust to datasets with categorical and mixed variable. TCM especially outperforms TCM-KNN for high-dimensional dataset (isolet). Furthermore, TCM-RF outperforms TCM-SVM for noisy data (covertype).

4.4 Robustness analysis of TCM-RF

A common way to validate an approach is to ensure robustness, that is, the approach must produce consistent results independent of the initial parameter settings. Empirical studies show the parameters adjustments of TCMs have great impacts on TCMs. Normalization of examples affects TCM-KNN greatly. As for TCM-SVM, not only the normalization but the type and parameters of kernel functions are important. Thus, the empirical and non-theoretically alteration hints a potential instability.

As is mentioned above, RF dissimilarity is more robust than many kinds of dissimilarities. Breiman [10] offered the theoretical proof that if $ntree$ is large (500 is enough), the Strong Law of Large Numbers convinces the RF dissimilarity be robust to the parameter settings. To demonstrate this point, we set up different parameters for TCM-RF, with $ntree = 500 ,1000 ,5000$ and $mtry = 1,..., \sqrt{a}$ (a is the number of attributes). Mean and standard deviation of forced accuracy are reported.

We compare the fluctuation by the normalization of TCM-KNN and the affection by the type of kernel for TCM-SVM). The results are summed up in table 5.

Table 5 the robust comparison of TCMs for sonar

sonar			
TCM-KNN	Without normalization	Attributes normalization	Examples normalization
accuracy	82.69%	88.46%	86.54%
TCM-SVM	Simple dot product	Radial basis function	Binomial coefficient polynomial
accuracy	63.46%	48.08%	96.15%
TCM-RF	Mean	standard deviation	
	84.92%	3.52%	

Our study indicates that TCM-RF shows a comparatively trivial fluctuation with the change of parameter settings. The robustness comes from the advantages of RF, which is an ensemble method. We believe that it may provide a more principled way of designing sample strangeness measure.

V Conclusions and Future work

In this paper, we illustrate that TCM-RF is a more effective and robust transductive confidence machine. Its efficiency is demonstrated by its rapid descending uncertain curve and high percentage of certain predictions. It is very robust to the type of variables and parameters settings.

For TCM-RF, RF is an ensemble method that guarantees lower error than the average individual error. Using ensemble strategies to define sample strangeness may be a more principled way than using a single classifier. On the other hand, it shows that the paradigm of hedging prediction can also be applied to an ensemble classifier.

We have the following main directions for future research. We plan to apply TCM-RF for further evaluation and assessment with our interest anomaly detection, microarray data and finance data. We also will continue exploring some other ensemble methods which can be applied to the framework of TCM.

REFERENCE

[1] S. Vanderlooy, I.G. Sprinkhuizen-Kuyper, and E.N. Smirnov. Reliable classifiers in ROC space. In Proceedings of the 15th BENELEARN Machine Learning Conference, 2006:27–36.

[2] S. Vanderlooy, I.G. Sprinkhuizen-Kuyper. An Overview of Algorithmic Randomness and its Application to Reliable Instance Classification. Technical Report MICC-IKAT 07-02, Universiteit Maastricht ,2006.

[3] C. Saunders, A. Gammerman, and V. Vovk. Transduction with confidence and credibility. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999:722–726.

[4] V. Vovk, A. Gammerman, and C. Saunders. Machine-Learning Applications of Algorithmic Randomness. In Proceedings of Sixteenth International Conference on Machine Learning, 1999:444-453.

[5] T. Bellotti, Z.Y. Luo, A. Gammerman, et al. Qualified predictions for microarray and proteomics pattern diagnostics with confidence machines. International Journal of Neural Systems, 2005, 15 (4) :247–258

[6] S. Vanderlooy, V. Laurens, D. Maaten, et, al. An off-Line Learning with Transductive Confidence Machines: an Empirical Evaluation. Technical Report, MICC-IKAT 07-03, Universiteit Maastricht, Maastricht, The Netherlands ,2007.

[7] A. Gammerman, V. Vovk. Hedging predictions in machine learning. Computer Journal, 2007, 50(2), 151–163.

[8] G. Brown, J. Wyatt, R. Harris, et al. Diversity Creation Methods: A Survey and Categorization. Information Fusion , Elsevier, 2005: 5:20.

[9] L. Breiman. Bagging Predictors. Machine Learning, 1996, 24 (2) :123-140.

[10] L. Breiman. Random forests. Machine Learning, 2001,45(3):5-32.

[11] Y. Qi, S. J. Klein, J. Z. Bar. Random forest similarity for protein–protein interaction prediction from multiple. Proceedings of the 10th Annual Pacific Symposium on Biocomputing, 2005:531-542.