다수 분류기를 이용한 메타레벨 데이터마이닝

김형관[1], 신성우[1,2]
[1] 한국과학기술원 테크노경영대학원
[2] 삼성 SDS

Metalevel Data Mining through Multiple Classifier Fusion

Steven H. Kim[1] and *Sung Woo Shin*[1,2]
[1] Graduate School of Management, Korea Advanced Institute of Science and Technology, Seoul, Korea
[2] Samsung SDS, Seoul, Korea

E-mail:
skim@msd.kaist.ac.kr
shinswoo@{kaist.ac | samsung.co}.kr

ABSTRACT

This paper explores the utility of a new classifier fusion approach to discrimination. Multiple classifier fusion, a popular approach in the field of pattern recognition, uses estimates of each individual classifier's local accuracy on training data sets. In this paper we investigate the effectiveness of fusion methods compared to individual algorithms, including the artificial neural network and k-nearest neighbor techniques. Moreover, we propose an efficient meta-classifier architecture based on an approximation of the posterior Bayes probabilities for learning the oracle.

Key words: Classifier fusion, consensus theory, meta-learning, feature weighting

INTRODUCTION

The method of classifier fusion is based on consensus theory. The latter is a well-established field involving procedures with the goal of combining single probability distributions to summarize estimates from multiple experts with the assumption that the experts make decisions based on Bayesian decision theory (Winkler, 1981; Benedikson and Swain, 1992). There are two types of combination of multiple classifiers: classifier fusion and classifier selection. In classifier fusion, individual classifiers are applied in parallel and their predicted outputs are combined in some manner to achieve a "consensus". Classifier selection attempts to predict which individual classifier is most likely to be correct for a given sample. Well-known classifier fusion algorithms include the majority vote, the Borda count, weighted voting, fuzzy integral, and the Dempster-Shafer throry.

In this paper an advanced fusion topic, meta-learning on the behavior of base classifiers, was investigated in terms of overall classification accuracy and compared against the representative fusion technique of majority voting. The approach is also compared against base classifiers relating to machine learning models including artificial neural network (ANN), k-nearest neighbor (kNN). Since two classifiers follow the Bayes posteriori probabilities (Ruck et al., 1990; Mitchell, 1997), it becomes possible to implement classifier fusion. The ideas are explored against the background of a practical application relating to financial fraud management.

CONSENSUS THEORY

Several consensus rules have been proposed including common linear opinion pool and the logarithmic opinion pool (Benedikson, et al., 1997; Cho and Kim, 1995). In contrast to the traditional opinion pool, we propose a veto-free, nonlinear consensus scheme as a form of the neural meta-classifier.

Type A. ***Linear opinion pool***

$$C_i(X) = \frac{1}{n}\sum_{k=1}^{n} p_k(\omega_i \mid X) \qquad C_i(X) = \sum_{k=1}^{n} \lambda_k p_k(\omega_i \mid X)$$

where

$p_k(\omega_i \mid X)$: classifier-specific posterior prob. given input X
λ_k : classifier-specific weight (relative expertise)
n : number of classifiers
ω_i : information class
C_i : the consensus probability for the class ω_i, $1 \le i \le c$

Type B. ***Logarithmic opinion pool***

$$C_i(X) = \prod_{k=1}^{n} p_k(\omega_i \mid X)^{\lambda_k}$$

- Treats the classifiers independently.
- Zeros in the logarithmic opinion pool are vetos.
- Externally Bayesian.
- Computational complexity is higher than for linear opinion pool.

Type C. ***Proposed approach***

$$C_i(X) = f\{\sum_{l=1}^{h} \lambda_{il} f\{\sum_{k=1}^{n} \lambda_{lk} p_k(\omega_i \mid X)\}\}$$

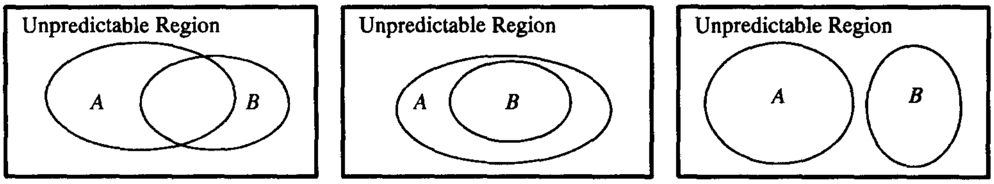
where λ_{lk} : weight from the kth base classifier's posteriori prob. for the class ω_i to the meta-classifier's lth hidden neuron
λ_{il} : weight from the lth hidden neuron to the ith class output
f : nonlinear mapping function such as sigmoid

META-LEARNING MODEL AS A CLASS COMBINER

The purpose of combining multiple classifiers is to achieve higher accuracy. The integrative process has a number of alternative names in the literature - *classifier ensembles, mixture of experts, consensus aggregation, committees of experts.*

Meta-learning might be viewed as a multilateral version of pure classifier fusion. In this paper, the elementary techniques representing basic learning related to a backpropagation MLP (BP) and an extended kNN. In addition, BP was used as a meta-learner (or meta-classifier) module. The training data for the meta-learner are based on predictions of a set of base classifiers. In other words, the meta-learning rule is composed of the actual classification and the elementary classifier's behavior (Chan and Stolfo, 1993; Wolpert, 1992; Kim and Shin, 1999). The aim of a meta-learner as a class combiner is to extend the predictable region of the base classifiers by learning the relationship between these predictions and the actual classification (see Figure 1). A comprehensive training scheme for a meta-learner is described in Figures 2 and 3. To compare the performance of meta-learning with the majority voter which needs an odd number of classifiers, we designed five base classifiers including ANN and kNN models. backpropagation MLP

Classifier Fusion	**Classifier Selection**	**Meta-Learning**
➢ Majority voting	➢ Gating network	➢ Stacked-generalization
➢ Weighted voting	➢ DCS-LA	➢ Class Combiner (CC)
➢ Sugeno's Fuzzy integral	➢ Behavior Knowledge Space	➢ CAC (Class Attribute Combiner)
➢ Dempster-Shafer theory	➢ Classifier Ranking	
➢ Borda count		
➢ *Linear opinion pool*		
➢ Nonlinear opinion pool		



(a) (b) (c)

<Figure 1> Potential regions of competence corresponding to two distinct classifiers. The symbol *A* denotes the region of competence for model A; and *B* for model B. Intuitively, the desired performance of a meta-learner should be the *oracle*: the union of *A* and *B*. In diagram (b), model A is a superset of B; consequently the predictive performance of a meta-learner should not exceed that of model A. For a meta-learner, the situation in diagrams (a) and (c) are the most interesting. In practical situations, fully disjoint regions as depicted in diagram (c) are not likely to occur. This study focused on meta-learning for the situation in diagram (a).

Class-combiner

$$T = \{(p_1(\omega_1 \mid X),\ldots, p_1(\omega_c \mid X),\ldots, p_n(\omega_1 \mid X),\ldots, p_n(\omega_c \mid X), O_1(X),\ldots, O_c(X)) \mid X \in V\}$$

Class-attribute-combiner

$$T = \{(p_1(\omega_1 \mid X),\ldots, p_1(\omega_c \mid X),\ldots, p_n(\omega_1 \mid X),\ldots, p_n(\omega_c \mid X), X^s, O_1(X),\ldots, O_c(X)) \mid X \in V, X^s \subset X\}$$

where T: training set, $p_n(\omega_c \mid X)$: *n*th classifier's posterior probability for the class ω_c given an feature-vector *X*, $O_1(X),\ldots, O_c(X)$: 1-of-c encoding of the actual class of X, X^s: feature-subset, V : validation set (training for the meta-classifier)

< Figure 2> Training set structure of the meta-classifier.

EXPERIMENTAL RESULTS

Data sets. For the experiments, we employed two data sets related to credit card fraud from the UCI Machine Learning Repository. This resource contains datasets which serve as standard yardsticks for evaluating algorithms developed by the machine learning community. The applications involved the realm of finance. The application involved credit card approval based on customer demographics and usage patterns. The datasets were divided into disjoint three segments: first one for training base classifiers, second one for training meta-learner, and the last one for evaluating performance of all classifiers.

Results. During the experiment, three separate trials were conducted using different partitions. For the two datasets, the overall performance of CC and CAC is promising: simple fusion (MV) yielded a better accuracy than individuals, and CC and CAC outperformed MV. To guide independent predictions from base classifiers, mutually exclusive feature-subset approach for ANN and feature-weighted representation for kNN were used (see Table 1 for detail).

CONCLUSION

This paper has presented a feasibility study of classifier fusion methods compared to individual algorithms for classification. In summary, the classifier fusion approach - including majority voting and meta classifier - can improve the overall classification accuracy. The computational experiments indicated that learning techniques can be combined synergistically to yield performance beyond that of the component modules. The fusion approach appears to be applicable to many other business domains where neural networks and other data mining tasks are currently employed: bankruptcy prediction, credit/bond rating, fraud/churn classification, consumer-choice prediction, and so on.

REFERENCES

[Benediktsson, J. A., and Swain, P. H.] (1992). "Consensus theoretic classification methods", *IEEE Trans. System, Man, and Cybernetics*, vol. 22, pp. 688-704.

[Benediktsson, J. A., et al.] (1997). "Parallel consensual neural network", *IEEE Trans. Neural Networks*, vol. 8, pp. 54-64.

[Chan, P. K., and Stolfo, S. J.] (1993). "Meta-learning for multistrategy and parallel learning", *Internaltional Workshop on Multistrategy Learning*, pp. 150-165.

[Cho, S.-B., and Kim, J. H.] (1995). "Multiple Network Fusion using Fuzzy Logic", *IEEE Trans. Neural Networks*, vol. 6, pp. 497-501.

[Jacobs, R., Jordan, M., Nowlan, S., and Hinton, G.] (1991). "Adaptive mixture of local experts", *Neural Computation*, vol. 3, pp. 79-87.

[Kim, S. H., Shin, S. W.] (1999). "Hybrid learning architecture for advanced data mining: An application to binary classification for fraud management", *Journal of Information Technology Application*, vol. 1, pp. 173-211.

[Lam L., and Suen, C. Y.] (1997). "Application of majority voting to pattern recognition: An analysis of its behavior and performance", *IEEE Trnas. System, Man, and Cybernetics*, vol. 27, pp. 553-568.

[Mitchell, T. M.] (1997). *Machine Learning*, NY: McGraw-Hill.

[Ruck, D. W., Rogers, et. al.] (1990). "The multilayer perceptron as an approximation to a Bayes optimal discrimination function", *IEEE Trans. Neural Networks*, vol. 1, pp. 296-298.

[Rumelhart, D. E., Hinton, G. E. and Williams, R. J.] (1986). "Learning Internal Representations by Error Propagation", in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, vol. 1, pp. 318-362, Cambridge, MA: MIT Press.

[Winkler, R. L.] (1981), "Combining probability distributions from dependent information sources", *Management Science*, vol. 27, pp. 479-488.

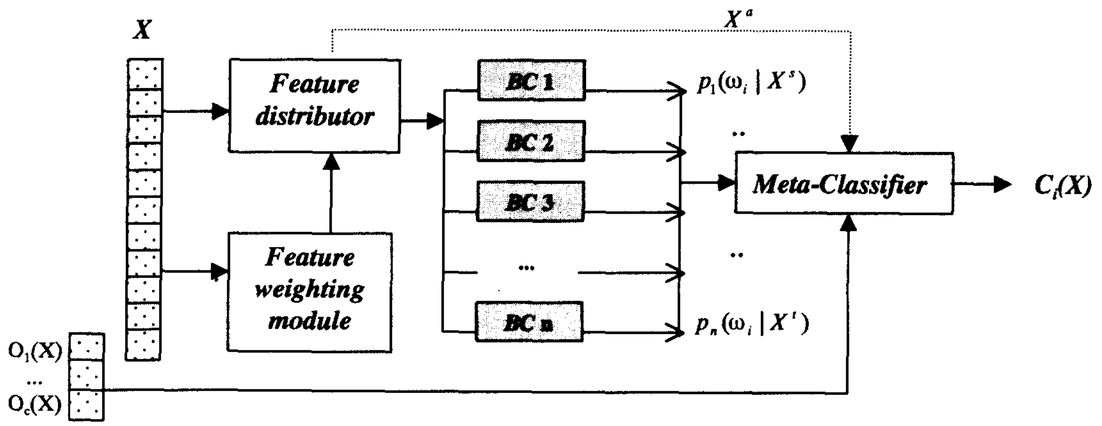
[Wolpert, D.] (1992). "Stacked generalization", *Neural Networks*, vol. 5, pp. 241-259.

< Table 1> Average performance of the two types of fusion against base classifiers. For the second type of fusion, in spite of the inferior performances of base classifiers - ANN5 and ANN7, the overall performance of the meta-classifiers is not inferior. Thus multiple weak classifiers with independent errors have greater potential than dependent strong classifiers.

TYPE 1 FUSION	**Accuracy (%)**								
	Individual model					**Fusion approaches**			
	ANN1	**ANN2**	**ANN3**	**kNN1**	**kNN2**	**CC**	**CAC**	**MV**	***Oracle***
Australian (230/690)	82.5	81.7	82.4	84.5	85.5	**86.1**	**86.5**	**85.9**	**94.7**
German (333/1000)	73.8	72.3	71.5	73.5	75.3	**76.5**	**76.5**	**75.6**	**88.2**

TYPE 2 FUSION	**ANN4**	**ANN5**	**ANN6**	**ANN7**	**kNN1**	**CC**	**CAC**	**MV**	***Oracle***
Australian (230/690)	83.5	67.4	81.3	65.7	84.5	**86.3**	**87.0**	**83.5**	**95.9**
German (333/1000)	71.7	72.8	69.7	66.8	73.5	**75.4**	**76.3**	**71.9**	**88.5**

ANN1, 2, 3, 4, 5, 6, 7: Artificial Neural Network with different feature-subset. **kNN1**: k-Nearest Neighbor. **kNN2**: kNN weighted by feature-relevancy. **CC**: Class Combiner. **CAC**: Class Attribute Combiner. **MV**: Majority Voter. ***Oracle***: theoretical upper bound for classifier fusion methods since it returns the correct classification if any of the base classifiers did so.



< Figure 3> The classifier fusion architecture with a meta-learning capability. Symbol *BC* stands for base classifier.