

HYBRID APPROACHES FOR CLASSIFICATION UNDER INFORMATION ACQUISITION COST CONSTRAINT

Parag C. Pendharkar
School of Business Administration
Pennsylvania State University at Harrisburg
pxp19@psu.edu

Extended Abstract¹

The practical use of classification systems may be limited because the current classification systems do not allow decision makers to incorporate cost constraint. For example, in several financial applications (loan approval, credit scoring, etc.) an applicant is asked to submit a processing fee with the application (Mookerjee and Mannino 1997). The processing fee may be used to validate the information entered in the application. From an economic standpoint, it is important that the cost of validating the information not exceed the processing fee. Traditional classification systems do not allow the decision maker to incorporate information acquisition cost constraint. We term the problem of designing a classification system, where information acquisition costs are considered, as the problem of classification with information acquisition cost constraint (CIACC). The CIACC problem is a NP hard problem and is very difficult to solve to optimality.

Current computer-based medical diagnostic methods use neural networks, discriminant analysis and other machine learning approaches for medical diagnosis (Pendharkar et al. 1999). Most of these approaches do not allow the decision maker to incorporate the information acquisition cost constraints. For example, Turney (1995) argues that “the problem of cost-sensitive classification is medical diagnosis, where a doctor would like to balance the costs of various possible medical tests with the expected benefits of the tests for the patient.” Turney proposed a hybrid genetic algorithm and decision tree induction algorithm to develop a classification system that minimizes information acquisition and mis-classification costs. The objective of Turney's study was to simultaneously minimize information acquisition and mis-classification costs.

We believe that CIACC can be used for medical diagnosis for the following reasons:

1. The information acquisition cost constraint may be determined by the decision maker in light of the prescribed fixed fees of a given diagnostic related group.
2. The function classification function, $f()$, may be determined by the decision maker in light of the quality care objectives of the health care facility (maximize correct predictions, minimize mis-classification costs, or both).

The CIACC problem allows the decision maker to incorporate the information acquisition cost constraint. It is important to note that our approach is different from the one taken by Turney in that the classification function learned by our approach will always have information acquisition costs less than or equal to the maximum information acquisition cost. The algorithms that we use to solve the proposed knapsack classification problem are different from those used by Turney as well.

Since the CIACC problem is a NP hard problem, complete and exact methods for solving the NP hard problem have an exponential time complexity and solving time may become prohibitory for large size problems (Hao and Pannier 1998). For solving NP hard problems in practice, local search algorithms such as simulated annealing, tabu search, and genetic algorithms

¹**Keywords:** Economics of information, classification, artificial intelligence, simulated annealing, neural networks, tabu search.

are used (Hao and Pannier 1998). We propose and use hybrid simulated annealing and artificial neural network (SA-ANN), as well as tabu search and artificial neural network (TS-ANN) for solving CIACC problem.

We apply the proposed hybrid SA-ANN and TS-ANN procedures to a real life data set for prediction of the heart disease. The heart disease data set has been used in previous studies (King et al. 1994) and is publicly available. The data consists of 13 different independent variables and the information acquisition cost for each variable is available from Turney.

The total cost of using all the attributes is \$600.57. We create 10 data sets of 200 examples from the original set of 270 examples. We arbitrarily set the value of maximum allowable information acquisition cost to \$300.29 (50% of the total cost of all of the attributes) for our experiments. We use the hybrid SA-ANN and TS-ANN procedures to solve the CIACC problem for the 20 data sets. Tables 1 and 2 illustrate the results of our experiments using the SA-ANN and TS-ANN procedures respectively.

The results of SA-ANN and TS-ANN were compared with the results from an ANN using all 13 attributes (total information acquisition cost of \$600.57) and no difference of means was observed.

Table 1. The Results of SA-ANN Hybrid Procedure of the Heart Disease Data Sets

Experiment Number	Correct Classification	Cost	Solution Vector
1	140	\$ 198.40	[1110010000101]
2	132	\$ 16.50	[0010001000000]
3	138	\$ 198.40	[1110010000101]
4	142	\$ 211.90	[0010011000101]
5	137	\$ 234.77	[1101111100010]
6	110	\$ 131.87	[0001111100000]
7	117	\$ 130.87	[0000111100000]
8	104	\$ 221.17	[0111111000101]
9	112	\$ 207.70	[1100001100100]
10	128	\$ 298.30	[1100010000111]

Table 2. The Results of TS-ANN Hybrid Procedure of the Heart Disease Data Sets

Experiment Number	Correct Classification	Cost	Solution Vector
1	123	\$ 292.00	[1000001010110]
2	111	\$ 294.00	[1110001010110]
3	131	\$ 295.00	[1111001010110]
4	131	\$ 294.00	[1110001010110]
5	153	\$ 293.00	[1100001010110]
6	111	\$ 299.27	[1000101010110]
7	130	\$ 283.70	[1110010010110]
8	113	\$ 283.79	[1110010010110]
9	140	\$ 294.00	[1110001010110]
10	129	\$ 283.70	[1110010010110]

References

Hao, J. K. , and Pannier, J. “Simulated Annealing and Tabu Search for Constraint Solving,” *Fifth International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale, FL, 1998, pp. 1-15.

- King, R. D., Henry, R., Feng, C., and Sutherland, A. A. "Comparative Study of Classification Algorithms: Statistical, Machine Learning and Neural Network, Machine Intelligence," in *Machine Intelligence and Inductive Learning*, K. Furukawa, D. Michie, and S. Muggleton (eds.), Clarendon Press, Oxford, UK, 1994.
- Mookerjee, V. S., and Mannino, M. V. "Sequential Decision Models for Expert System Optimization," *IEEE Transactions on Knowledge and Data Engineering* (9), 1997, pp. 675-687.
- Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, N., and Benner, M. "Association, Statistical, Mathematical, and Neural Approaches for Mining Breast Cancer Patterns," *Expert Systems with Applications* (17), 1999, pp. 223-232.
- Turney, P. D. "Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm," *Journal of Artificial Intelligence Research* (2), 1995, pp. 369-409.

