Proceedings of 2012 4th International Conference on Machine Learning and Computing
IPCSIT vol. 25 (2012) © (2012) IACSIT Press, Singapore

Application of Evolutionary Data Mining Algorithms to Insurance Fraud Prediction

Jenn-Long Liu[1+] and Chien-Liang Chen[2,3]

[1] Dept. of Information Management, I-Shou University, Kaohsiung 840, Taiwan
[2] Dept. of Information Management, Fortune Institute Technology, Kaohsiung 831, Taiwan
[3] Dept. of Information Engineering, I-Shou University, Kaohsiung 840, Taiwan

Abstract. This study proposes two kinds of Evolutionary Data Mining (EvoDM) algorithms to the insurance fraud prediction. One is GA-Kmeans by combining K-means algorithm with genetic algorithm (GA). The other is MPSO-Kmeans by combining K-means algorithm with Momentum-type Particle Swarm Optimization (MPSO). The dataset used in this study is composed of 6 attributes with 5000 instances for car insurance claim. These 5000 instances are divided into 4000 training data and 1000 test data. Two different initial cluster centers for each attribute are set by means of (a) selecting the centers randomly from the training set and (b) averaging all data of training set respectively. Thereafter, the proposed GA-Kmeans and MPSO-Kmeans are employed to determine the optimal weights and final cluster centers for attributes, and the accuracy of prediction for test set is computed based on the optimal weights and final cluster centers. Results show that the presented two EvoDM algorithms significantly enhance the accuracy of insurance fraud prediction when compared the results to that of pure K-means algorithm.

Keywords: Evolutionary Data Mining, Genetic Algorithm, Momentum-Type Particle Swarm Optimization, Insurance Fraud Prediction

1. Introduction

This study aims using two evolutionary data mining (EvoDM) algorithms to evaluate whether case is a insurance fraud or not. The insurance fraud is a behavior that the beneficiary makes up fake affairs to apply for compensation such that he/she can get illegal benefits to himself /herself or some other people. Generally, the characteristics of insurance fraud are that it is low cost and high profit and also it is an intelligent crime. Moreover, insurance fraud could be an international crime, and could happen in any kinds of insurance cases. Recently, there are more and more new types of insurance proposed on the markets such that how to detect possible fraud events for a manager/analyst of insurance company becomes more important than ever before.

This work proposes two kinds of EvoDM algorithms, which combines a clustering algorithm, K-means, with two evolutionary algorithms, Genetic Algorithm (GA) and Momentum Particle Swarm Optimization (MPSO). The two proposed EvoDM algorithms are termed GA-Kmeans and MPSO-Kmeans, respectively. This work conducts 5000 instances of insurance cases for data mining. The 5000 instances are divided into 4000 instances to be the training set and 1000 instances to be the test set. Furthermore, this work applies K-means, GA-Kmeans and MPSO-Kmeans algorithms to evaluate the fraud or not from the training set and also evaluate the accuracy of fraud prediction for the test set.

2. Literature Review

Data Mining is a crucial step in the Knowledge Discovery in Database (KDD) process that consists of applying data analysis and knowledge discovery algorithms to produce useful patterns (or rules) over the

[+] Jenn-Long Liu . Tel.: +886-7-657-7711#6551; fax: +886-7-657-8491.
E-mail address: jlliu@isu.edu.tw

datasets. Although the data mining has several different definitions from the scholars, its purpose is discovering useful knowledge and information from database. Generally, data mining technologies include (1) Associate Rules, (2) Classification, (3) Clustering Analysis, (4) Regression Analysis, (5) Particle Swarm Optimization and (6) Time Series Analysis, and so on [4, 12]. This work proposes two kinds of EvoDM algorithms, which combines a clustering algorithm, K-means, with two evolutionary algorithms, Genetic Algorithm (GA) and Momentum Particle Swarm Optimization (MPSO). The below introduce clustering analysis, GA, and MPSO.

2.1. Clustering Analysis

Clustering Analysis is a main method for exploring data mining and also is a common technique for statistical data analysis. It can be applied to machine learning, image analysis, pattern recognition, information retrieval, and bioinformatics. The K-means algorithm is the one of often used method in the clustering algorithms. When the number of clusters is fixed to k, K-means algorithm gives a formal definition as an optimization problem to specify k cluster centers and assign each instance to its belonging cluster with the smallest distance from the instance to assigned cluster [4]. The flowchart of K-means depicted in Fig. 1.

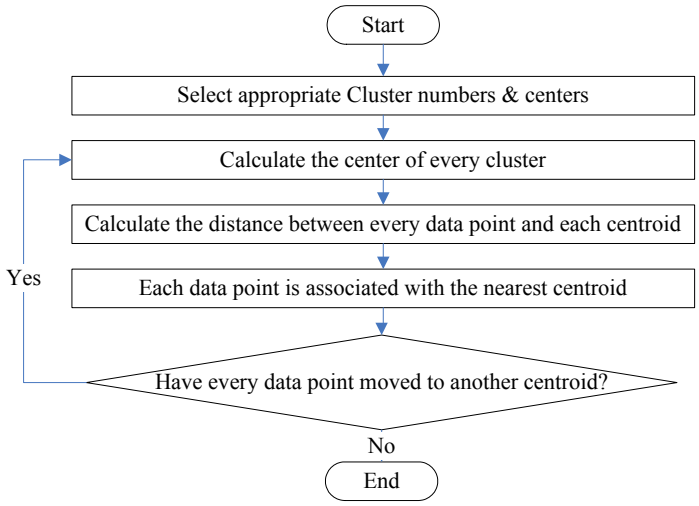

Fig. 1: Flowchart of K-means algorithm

2.2. Genetic Algorithm

Genetic Algorithm is a stochastic search algorithm which based on the Darwinian principal of natural selection and natural genetics. The selection is biased toward more highly fit individuals, so the average fitness of the population tends to improve from one generation to the next. In general, GA generates an optimal solution by means of using reproduction, crossover, and mutation operators [3, 9]. The fitness of the best individual is also expected to improve over time, and the best individual may be selected as a solution after several generations.

2.3. Particle Swarm Optimization

The PSO algorithm was first introduced by Kennedy and Eberharth [6] in 1995. The concept of PSO is that each individual in PSO flies in the search space with a velocity which is dynamically adjusted according to its own flying experience and its companions' flying experience. Each individual is treated as volume-less particle in the D-dimensional search space. Shi and Eberhart modified the original PSO in 1999 [12]. The equation is expressed as follows:

$$\vec{v}_i^{k+1} = w\vec{v}_i^{k} + c_1 \times r_1 \times (Pbest_i - \vec{x}_i^{k}) + c_2 \times r_2 \times (Gbest_i - \vec{x}_i^{k}) \tag{1}$$

$$\vec{x}_i^{k+1} = \vec{x}_i^{k} + \vec{v}_i^{k+1}, \ i = 1,2,..., \mathrm{N}_{particle} \tag{2}$$

where c_1 and c_2 are the cognitive and social learning rates, respectively. The random function r_1 and r_2 are uniformly distributed in the range [0, 1]. Equation (1) reveals that the large inertia weight promotes global exploration, whereas the small value promotes a local search.

2.4. Momentum-Type Particle Swarm Optimization

Liu and Lin proposed a MPSO in 2007 [8] for improving the computational efficiency and solution accuracy of Shi and Eberhart's PSO [10]. The original PSO developed by Kennedy and Eberhart [6] supposed that the *i*th particle flies over a hyperspace, with its position and velocity given by $\vec{x}_i$ and $\vec{v}_i$. The best previous position of the *i*th particle is denoted by Pbest*i*. The term Gbest*i* represents the best particle with the highest function value in the population. The Liu and Lin's MPSO proposed the next flying velocity and position of the particle *i* at iteration $k+1$ by using the following heuristic equations:

$$\vec{v}_i^{k+1} = \beta(\Delta\vec{v}_i^{k}) + c_1 \times r_1 \times (Pbest_i - \vec{x}_i^{k}) + c_2 \times r_2 \times (Gbest_i - \vec{x}_i^{k}) \tag{3}$$

$$\vec{x}_i^{k+1} = \vec{x}_i^{k} + \vec{v}_i^{k+1}, \ \ i = 1,2,...,\mathrm{N}_{particle} \tag{4}$$

where c_1 and c_2 are the cognitive and social learning rates, respectively. The random function r_1 and r_2 are uniformly distributed in the range [0, 1]. The value of β is a positive number ($0 \le \beta < 1$) termed the momentum constant, which controls the rate of change in velocity vector. Equation (3) allows each particle the ability of dynamic self-adaptation in the search space over time. That is, the *i*th particle can memorize the previous velocity variation state and automatically adjust the next velocity value during movement.

3. Evolutionary Data mining Algorithm

In the data mining field, clustering analysis is a very important technology for KDD. This study aims to find insurance fraud cluster optimization by EvoDM algorithms based on the K-means algorithm [4, 12]. In general, K-means algorithm is a popular method to solve this kind of clustering problem, but the drawback of it is that the accuracy of clustering results needs to be further improved. Therefore, the K-means clustering algorithm is combined genetic algorithms as hybrid genetic models [2, 7] to improve the accuracy of prediction. This study proposes two kinds of EvoDM algorithms as GA-based K-means and MPSO-based K-means which are termed GA-Kmeans and MPSO-Kmeans, respectively. The flowcharts of GA-Kmeans and MPSO-Kmeans are depicted in Figs. 2 and 3.

The objective function, $Obj(\vec{w})$, for GA-Kmeans and MPSO-Kmeans is specified by minimizing the clustering errors between classification results of prediction (Cpred) and original (Corig) for n training data to determine the optimal weights $(\vec{w})$ for each attributes as follows.

$$Obj(\vec{w}) = Min\left(\sum_{i=1}^{n} \left| \left(C_{pred}\right)_i - \left(C_{orig}\right)_i \right| \right) \tag{5}$$

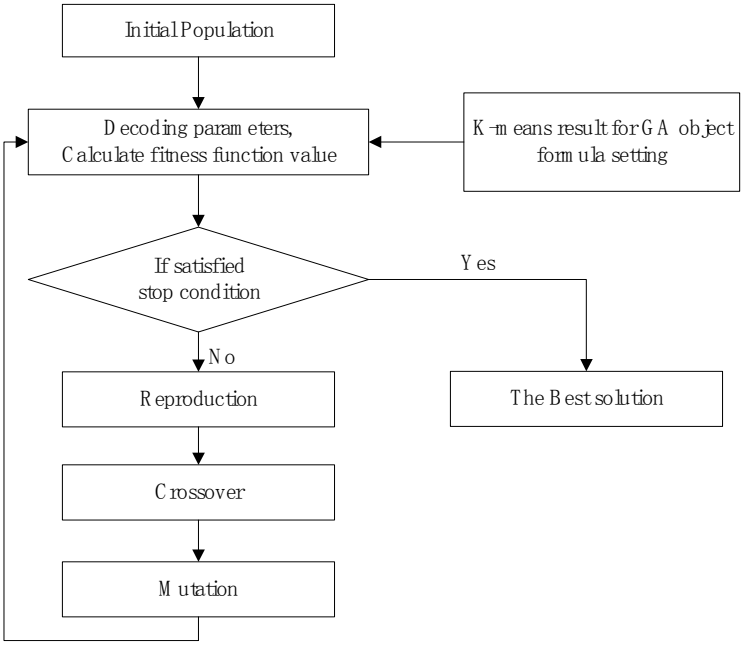

Fig. 2: Flowchart of GA-Kmeans algorithm

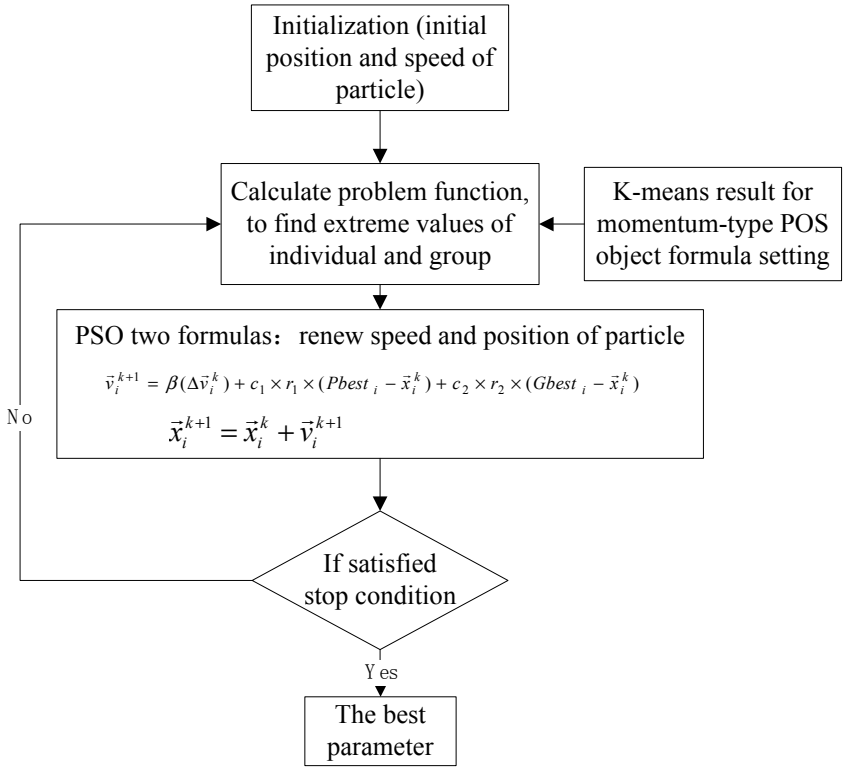

Fig. 3: Flowchart of MPSO-Kmeans algorithm

Table. 1: Partial data of original insurance fraud dataset.

Instance	age	gender	claim amount	tickets	claim times	attorney	outcome
1	54	male	2700	0	0	none	approved
2	39	male	1000	0	0	none	approved
3	18	female	1200	0	1	none	approved
4	42	female	1800	1	0	none	approved
5	18	male	5000	0	3	Gold	fraud
6	51	female	1900	1	0	none	approved
7	44	male	2300	0	0	none	approved
8	23	Female	4000	3	2	Smith	approved
9	34	Female	2500	0	0	none	approved
10	56	male	2500	0	0	none	approved
…							

Table. 2: Partial data of normalized insurance fraud dataset.

Instance	age	gender	claim amount	tickets	claim times	attorney	outcome
1	1	1	0.46	1	1	0	0
2	0.95	1	0.8	1	1	0	0
3	0	0	0.76	1	0.5	0	0
4	1	0	0.64	0.6	1	0	0
5	0	1	0	1	0	1	1
6	1	0	0.62	0.6	1	0	0
7	1	1	0.54	1	1	0	0
8	0.15	0	0.2	0	0	1	0
9	0.7	0	0.5	1	1	0	0
10	1	1	0.5	1	1	0	0
…							

4. Results & Discussion

4.1. Dataset Sample

This study uses 5000 instances of insurance claim with six variables [12]. The six variables are age, gender, claim amount, tickets, claim times, and accompanied with attorney. The partial datasets of original and optimized insurance claim was listed in Tables 1 and 2, respectively. The normalization formulas are presented in Ref. [12]. This work specified six weights (w_1, w_2, w_3, w_4, w_5, w_6) for applying GA-Kmeans and MPSO-Kmeans algorithms due to six attributes for the dataset. All values of $\vec{w}$ are specified in the range [0, 1].

4.2. Case 1: Initial Cluster Centers are Selected Randomly from Training Set

Table 3 lists the accuracy of using three different algorithms for Case 1 which the initial cluster centers are selected from training set randomly. The accuracy evaluated by GA-Kmeans is the same as that of MPSO-Kmeans. Also, it is clearly that the solutions obtained using the two EvoDM algorithms were better than that of K-means. Table 4 lists the optimal weights of six attributes computed by GA-Kmeans and MPSO-Kmeans. The attributes for claim amount, claim times and attorney were significant than other attributes for determining the clusters.

Table. 3: Comparison of prediction results of Case 1.

Algorithm / Data set	Clustering (K-means only)	Evolutionary Data Mining Algorithms	
		GA-Kmeans	MPSO-Kmeans
Training set	35.625%	85.20%	85.20%
Test set	37.90%	86.325%	86.325%

Table. 4 Optimal weights of Case 1 computed by presented EvoDM algorithms

Weights for 6 attributes	GA-Kmeans	MPSO-Kmeans
w_1(Age)	0.08937	0.06027
w_2 (Gender)	0.03081	0.1
w_3 (Claim Amount)	0.94993	0.46535
w_4 (Tickets)	0.00521	0.04573
w_5 (Claim times)	0.63839	0.67031
w_6 (Attorney)	0.54930	0.9

Table. 5: Comparison of prediction results of Case 2.

Algorithm / Data set	Clustering (K-means only)	Evolutionary Data Mining Algorithms	
		GA-Kmeans	MPSO-Kmeans
Training set	88.30%	96.50%	96.50%
Test set	89.725%	97.60%	97.60%

Table. 6: Optimal weights of Case 2 computed by presented EvoDM algorithms.

Weights for 6 attributes	GA-Kmeans	MPSO-Kmeans
w_1(Age)	0.09542	0.18947
w_2 (Gender)	0.40204	0.13705
w_3 (Claim Amount)	0.94579	0.9
w_4 (Tickets)	0.17894	0.26487
w_5 (Claim times)	0.09067	0.02102
w_6 (Attorney)	0.96118	0.69686

4.3. Case 2: Initial Cluster Centers are Determined by Averaging Training Set

Table 5 lists the accuracy of three different algorithms for Case 2 which the initial centers are obtained by averaging all training set for each attributes. The overall accuracy of using the three algorithms for the case was higher than that of the previous one. Computational results also showed that the accuracy of presented two EvoDM algorithms was better than that of K-means algorithm. Moreover, Table 6 lists the optimal weights of six attributes obtained using GA-Kmeans and MPSO-Kmeans algorithms. The attributes for claim amount and attorney were relatively significant than other attributes for determining the clusters.

5. Conclusion

This study introduced the K-means algorithm and two EvoDM algorithms including GA-Kmeans and MPSO-Kmeans algorithms to the insurance fraud prediction. The two EvoDM algorithms were hybrid by incorporating the K-means algorithm with GA and MPSO, respectively. Two initial cluster centers conditions were studied to check the robustness of the algorithms. From our computational results, the accuracy for test set prediction obtained using GA-Kmeans and MPSO-Kmeans algorithms was 86.325% for Case 1 which the initial cluster centers were selected from training set randomly, whereas the accuracy obtained using K-means algorithm was 37.9% only. From the weight distribution of Case 1, the attributes of claim amount, claim times and attorney showed the relatively important in judging the insurance fraud. Furthermore, this work made changes for the initial cluster centers, termed Case 2, by averaging all the data training set for each attributes. The accuracy for test set prediction obtained using GA-Kmeans and MPSO-Kmeans algorithms for Case 2 was significantly enhanced to 97.6% while the accuracy obtained using K-means algorithm was 89.725%. From the weight distribution of Case 2, the attributes of claim amount and attorney demonstrated relatively important in judging insurance fraud. Accordingly, the accuracy of insurance fraud prediction can be enhanced by using the presented two EvoDM algorithms.

6. Acknowledgements

This work was supported in part by grant NSC 100-2221-E-214-040 from the National Science Council of Republic of China.

7. References

[1] W. H. Au, K. C. C. Chan, X. Yao. A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction. *IEEE Transactions on Evolutionary Computation*. 2003, **7** (6): 532-545.

[2] A. Brabazon, and P. Keenan. A Hybrid Genetic Model for the Prediction of Corporate Failure. *Computational Management Science*. 2004, **1**, (3-4): 293-310.

[3] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison Wesley, 1989.

[4] J. Han, and M. Kamber. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[5] M. Kantardzic. *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons, 2002.

[6] J. Kennedy, and R. Eberhart. Particle Swarm Optimization, *Proc. IEEE Int. Conf. on Neural Networks* (Perth, Australia), IEEE Service Center, Piscataway, NJ. 1995, **4**: 1942-1948.

[7] P. C. Lin, and J. S. Chen. A Genetic-Based Hybrid Approach to Corporate Failure Prediction. *International Journal of Electronic Finance*. 2008, **2** (2): 241-255.

[8] J. L. Liu, and J. H. Lin. Evolutionary Computation of Unconstrained and Constrained Problems Using a Novel Momentum-type Particle Swarm Optimization. *Engineering Optimization*. 2007, **39** (3): 287-305.

[9] Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs*. 3rd ed., Springer-Verlag, 1999.

[10] Y. Shi, and R. Eberhart. A Modified Particle Swarm Optimization, in *Proc. of IEEE International Conference on Evolutionary Computation* (ICEC). 1998, pp. 69-72.

[11] Y. Shi, and R. Eberhart. Empirical study of particle swarm optimization, in *Proceedings of the 1999 Congress on Evolutionary Computation*. 1999, pp. 1945-1950.

[12] D. Olson, and Y. Shi. *Introduction to Business Data Mining*, McGraw-Hill Education, 2008.