Efficient Evolutionary Data Mining Algorithms Applied to the Insurance Fraud Prediction

Jenn-Long Liu, Chien-Liang Chen, and Hsing-Hui Yang

***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.**

***Index Terms*— Evolutionary data mining, genetic algorithm, insurance fraud prediction, momentum-type particle swarm optimization.**

I. 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.

II. CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining) is a data mining process model that describes commonly used approaches for expert data miners use to solve problems. CRISP-DM was conceived in late 1996 by SPSS (then ISL), NCR and DaimlerChrysler (then Daimler-Benz). Also, it is the leading methodology used by data miners. CRISP-DM breaks the processes of data mining into six major phases as follows.

A. Business Understanding

This is mainly on the understanding of business project objectives and requirements, its conversion to a data mining problem definition, and the design of a preliminary plan.

B. Data Understanding

This phase collects an initial data and then gets itself familiarized with many activities to be able to identify its quality problems, develop its first insights, or detect some interesting subsets to form hypotheses for the yet-revealed information.

C. Data Preparation

This includes activities to construct the final dataset based upon the original data. It is likely to be repetitiously and randomly performed. It includes table, record and attribute selection, transformation, and the cleaning of data to be used as modeling tools.

D. Modeling

Here the parameters are calibrated to optimal values, and different modeling techniques are selected and put to use. Techniques used for the same data mining problem are often with specific requirements on data form, which makes it necessary to often go back to the data preparation phase.

E. Evaluation

Up to this phase, a model with high quality data analysis is built. Thoroughly evaluating the model and reviewing the performed steps in the construction of a model is a must in its achievement of business objectives. Some important, yet undecided business issue can determine a key objective. A decision based on data mining should be made.

Manuscript received April 5, 2012; revised May 5 2012. This work was supported in part by National Science Council of Republic of China under Grant Number NSC 100-2221-E-214-040.

Authors are with the Information Management Department, I-Shou University, Kaohsiung 84001, Taiwan (e-mail: jlliu@isu.edu.tw; muffin.chen@gmail.com; nancyyang@ms.aidc.com.tw).

F. Deployment

The completion of a model is often not the final goal though its purpose is to decipher more information from the data. Information from the original data will need to be further organized and then turned to a form that can be of use to the customer. This often includes the application of functioning models in an organization's decision making processes. This phase can be both simple and complex, depending on the requirements. It is often is the customer rather than the data analyst who carries this phase out. It is important for the customer to realize actions need to be carried out to the use of the created models.

III. 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 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.

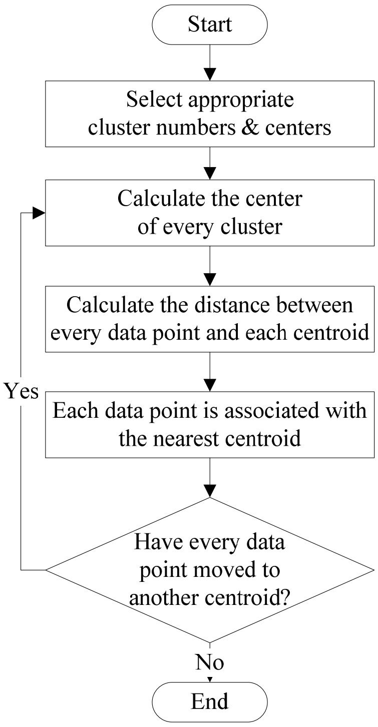

Start
Select appropriate cluster numbers & centers
Calculate the center of every cluster
Calculate the distance between every data point and each centroid
Each data point is associated with the nearest centroid
Have every data point moved to another centroid?
Yes
No
End

Fig. 1. Flowchart of K-means algorithm

A. 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.

B. 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. Generally, the pseudo-code of the GA is shown as follows:

```
Procedure: The Hybrid Genetic Algorithm
Begin
Create initial population randomly;
do {
      Choose a pair of parents from population;
       /* REPRODUCTION */
      children=CROSSOVER(parent1, parent2);
      MUTATION(children);
      Parents ← Children
   } while (stopping criterion not satisfied);
End;
```

Therefore, the flowchart of GA can be depicted in Fig. 2.

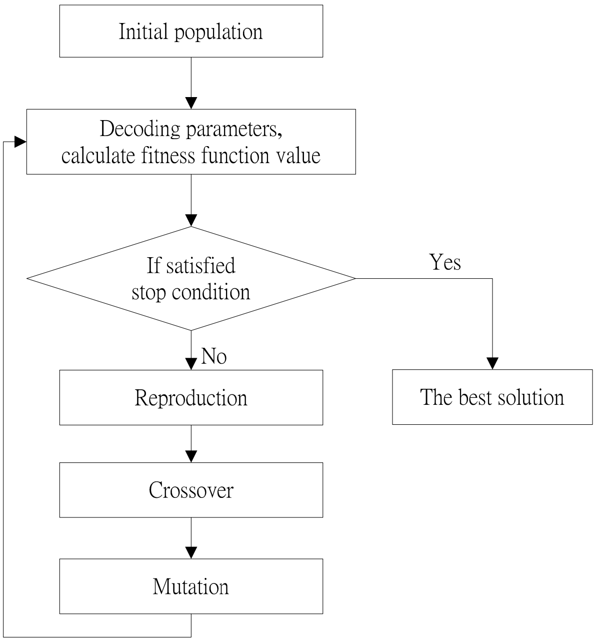

Initial population
Decoding parameters, calculate fitness function value
If satisfied stop condition
Yes
No
The best solution
Reproduction
Crossover
Mutation

Fig. 2. Flowchart of GA algorithm

C. 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 [11]. The equation is expressed as follows:

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

$$\vec{x}_i^{k+1} = \vec{x}_i^{k} + \vec{v}_i^{k+1}, \; i = 1,2,...,N_{particle} \quad (2)$$

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. The flowchart of PSO is depicted in Fig. 3.

D. 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 ith 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 ith particle is denoted by Pbesti. The term Gbesti 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:

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

$$\vec{x}_i^{k+1} = \vec{x}_i^{k} + \vec{v}_i^{k+1}, \; i = 1,2,...,N_{particle} \quad (4)$$

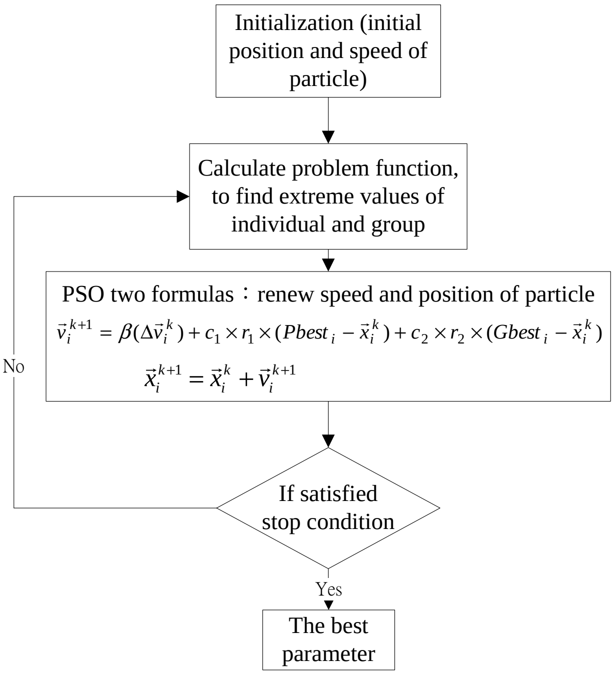
where c_1 and c_2 are the cognitive and social learning rates, respectively. The random function r_1 and r_2 are uniformly distributed in the range [0, 1]. The value of β is a positive number ($0 \le \beta < 1$) termed the momentum constant, which controls the rate of change in velocity vector. Equation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

E. C4.5 Algorithm

To evaluate the algorithmic performance of our presented two EvoDM algorithms, this paper also applied two existed software, C4.5 and Naïve Bayes algorithms, to the computation of the insurance fraud prediction. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. C4.5 constructs a complete decision tree first. Then, on each internal node, it prunes the decision tree according to the defined Predicted Error Rate. The decision trees generated by C4.5 can be used for classification. C4.5 is often referred to as a statistical classifier [13].

F. Naïve Bayes Algorithm

Naive Bayes algorithm is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. The main operating principle of Naive Bayesian classifier, is to learn and memorize the central concept of these training samples by classifying the training samples according to the selected properties. Then, apply the learned categorizing concept to the unclassified data objects and execute the category forecast, to gain the target of the test example.



Initialization (initial position and speed of particle)
Calculate problem function, to find extreme values of individual and group
PSO two formulas：renew speed and position of particle
$\vec{v}_i^{k+1} = \beta(\Delta\vec{v}_i^{k}) + c_1 \times r_1 \times (Pbest_i - \vec{x}_i^{k}) + c_2 \times r_2 \times (Gbest_i - \vec{x}_i^{k})$
$\vec{x}_i^{k+1} = \vec{x}_i^{k} + \vec{v}_i^{k+1}$
No
If satisfied stop condition
Yes
The best parameter

Fig. 3. Flowchart of PSO algorithm

IV. 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. 4 and 5.

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 (*Cactual*) for n training data to determine the optimal weights $(\vec{w})$ for each attributes as follows:

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

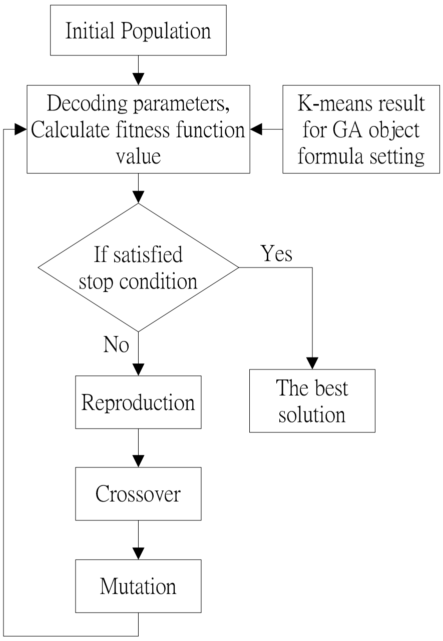

Initial Population
Decoding parameters, Calculate fitness function value
K-means result for GA object formula setting
If satisfied stop condition
Yes
No
Reproduction
The best solution
Crossover
Mutation

Fig. 4. Flowchart of GA-Kmeans algorithm

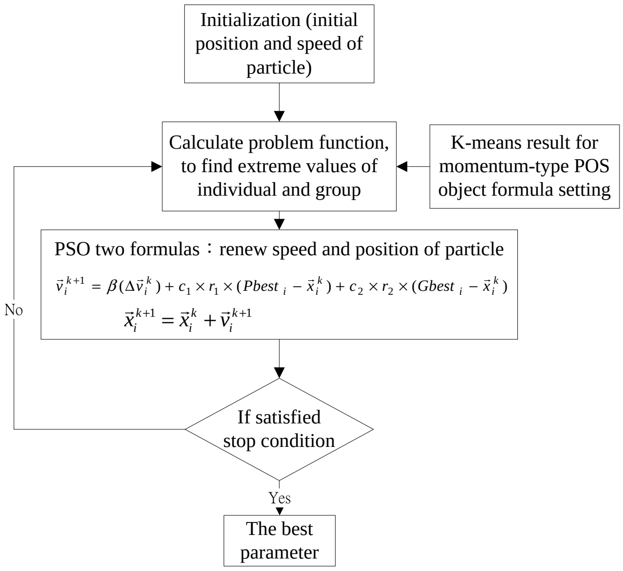

Initialization (initial position and speed of particle)
Calculate problem function, to find extreme values of individual and group
K-means result for momentum-type POS object formula setting
PSO two formulas：renew speed and position of particle
$\vec{v}_i^{k+1} = \beta(\Delta\vec{v}_i^k) + c_1 \times r_1 \times (Pbest_i - \vec{x}_i^k) + c_2 \times r_2 \times (Gbest_i - \vec{x}_i^k)$
$\vec{x}_i^{k+1} = \vec{x}_i^k + \vec{v}_i^{k+1}$
No
If satisfied stop condition
Yes
The best parameter

Fig. 5. Flowchart of MPSO-Kmeans algorithm

V. RESULTS & DISCUSSION

A. 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. Age means the age of the claimer. Gender means the claimer's gender. "Claim amount" means the amount of the claim, and "tickets" stands for the amount of the tickets the claimer that has received before. "Claim times" represents the number of times that the claimer has claimed before. Accompanied with attorney shows whether the claimer is accompanied with an attorney or not. The data types of age, claim amount, tickets and claim times are all numeric. The value of gender is male or female. The value of accompanied with attorney is lawyer's name or none. The partial datasets of original and optimized insurance claim was listed in Tables I and II, respectively. The normalization formulas are presented below.

(1) Age: younger than 20 years old is 0, 20-40 years old is (age-20)/20, 40-60 years old is 1, 60-70 years old is 1-(age-60)/10, older than 70 years old is 0.
(2) Gender: male is 1, female is 0.
(3) Claim amount: =Max(1-claim amount/5000,0).
(4) Tickets: 0 ticket is 1, 1 ticket is 0.6, over 2 tickets is 0.
(5) Claim times: none is 1, one time is 0.5, over 2 times is 0.
(6) Accompanied with attorney: none is 1, others is 0.
(7) Outcome: approved is 0, fraud is 1.

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].

TABLE I: PARTIAL DATA OF ORIGINAL INSURANCE FRAUD DATASET.

Instance	Age	Gender	Claim Amount	Tickets	Claim	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 II: 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
…							

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

Table III 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 IV 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 III: COMPARISON OF PREDICTION RESULTS OF CASE 1.

Algorithm Data set	Clustering (K-means only)	Evolutionary Data Mining Algorithms	
		GA-Kmeans	MPSO-Kmeans
Training set	35.62%	85.20%	85.20%
Test set	37.90%	86.32%	86.32%

TABLE IV: 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

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

Table V 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 VI 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. Accordingly, the presented two EvoDM algorithms not only can achieve high accuracy of prediction, but also they can determine the significant attributes automatically from all attributes based on the evaluated weights. The attribute information is most useful for a manager or a staff member who has the authority to make a right decision with agreement or not when a client submits the settlement of claims involving insurance cases.

TABLE V: 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%	97.60%	97.60%
Test set	89.72%	96.50%	96.50%

TABLE VI: 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

D. Confusion Matrix

Table VII lists the confusion matrix of four different algorithms for training set. The overall accuracy of using the four algorithms was very high (over 96%). Although the accuracy of C4.5 is 98.5% high, it cannot classify any fraud case. Naïve Bayes correctly predicts 12 fraud cases. Both of two EvoDM classify one more correct fraud case than Naïve Bayes. Table VIII lists the confusion matrix of four different algorithms for test set. The accuracies of all four algorithms are over 96%. C4.5 can not correctly predict any fraud case. The accuracy of Naïve Bayes is little higher than EvoDM. The correct prediction of fraud case with EvoDM is 5 cases and with Naïve Bayes is 3 cases.

TABLE VII: CONFUSION MATRIX OF C4.5, NAÏVE BAYES, AND EVO-DM ALGORITHMS FOR TRAINING SET.

Algorithm	C4.5		Naïve Bayes		GA-Kmeans		MPSO-Kmeans	
	a	b	a	b	a	b	a	b
a=approved	3940	0	3896	44	3891	49	3891	49
b=fraud	60	0	48	12	47	13	47	13
accuracy	98.5%		96.8%		97.6%		97.6%	

TABLE VIII: CONFUSION MATRIX OF C4.5, NAÏVE BAYES, AND EVO-DM ALGORITHMS FOR TEST SET.

Algorithm	C4.5		Naïve Bayes		GA-Kmeans		MPSO-Kmeans	
	a	b	a	b	a	b	a	b
a=approved	978	0	965	13	960	18	960	18
b=fraud	22	0	19	3	17	5	17	5
accuracy	97.8%		96.8%		96.5%		96.5%	

VI. 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 86.32% 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 96.5% while the accuracy obtained using K-means algorithm was 89.72%. 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.

The main purpose of the insurance fraud prediction is to find out the fraud cases correctly. Normally, the probability of fraud cases is so small that even if misjudgment of fraud cases occurs, the accuracy is still high. As listed in Table VII and VIII, even C4.5 algorithm can't predict every fraud case correctly, the accuracy of prediction is still higher than 97.8%. Although GA-Kmeans and MPSO-Kmeans are not the best in prediction accuracy, they can find more fraud cases than the others.

References

[1] W. H. Au, K. C. C. Chan, and X. Yao. "A Novel Evolutionary Data Mining Algorithm with Applications to Churn Prediction," *IEEE Transactions on Evolutionary Computation*, vol. 7, pp. 532-545, Dec. 2003.

[2] A. Brabazon, and P. Keenan, "A Hybrid Genetic Model for the Prediction of Corporate Failure," *Computational Management Science*. vol. 1, no. 3, pp. 293-310, Oct. 2004.

[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," in *Proc. IEEE Int. Conf. on Neural Networks (Perth, Australia)*, IEEE Service Center, Piscataway, NJ. vol. 4, Nov. 1995, pp. 1942-1948.

[7] P. C. Lin, and J. S. Chen, "A Genetic-Based Hybrid Approach to Corporate Failure Prediction," *International Journal of Electronic Finance*. vol. 2, no. 2, pp. 241-255, Mar. 2008.

[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*. vol. 39, no. 3, pp. 287-305, Apr. 2007.

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

[10] Y. Shi, and R. Eberhart, "A Modified Particle Swarm Optimizer," in *Proc. of IEEE International Conference on Evolutionary Computation (ICEC)*, pp. 69-73, May 1998.

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

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

[13] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.