Adv. Math. Fin. App., 2022, **7(**2), P. 437-446

Advances in Mathematical Finance & Applications

www.amfa.iau-arak.ac.ir
Print ISSN: 2538-5569
Online ISSN: 2645-4610
Doi: 10.22034/AMFA.2021.1927097.1580

Applied-Research Paper

Insurance Claim Classification: Genetic Programming Approach

Alireza Bahiraie[a]*, Farbod Khanizadeh[b], Farzan Khamesian[b]

[a]*Faculty of Mathematics, Statistics & Computer Science, Semnan University 35131-19111, Semnan, Iran*
[b]*Insurance Research Centre (IRC), Tehran 1998758513, Iran*

ARTICLE INFO

Article history:
Received 2021-04-03
Accepted 2021-09-07

Keywords:
Genetic Programming
Supervised Learning
Classification

ABSTRACT

In this study we provide insurance companies with a tool to classify the risk level and predict the possibility of future claims. The support vector machine (SVM) and genetic programming (GP) are two approaches used for the analysis. Basically, in Iran insurance industry there is no systematic strategy to evaluate the car body insurance policy. Companies refer mainly to the world experience and employ it to rate the premium. An insurance claim dataset provided by an Iranian insurance company with a sample size of 37904 is considered for programming and analysis. According to the structure of the dataset, a supervised learning algorithm was used to describe the underlying relationships between variables. The model accuracy is over 90% and the outcomes indicate that car type, car plate, car color and car age were the main four factors contributing in prediction of claims.

1 Introduction

The insurance industry is one of the foremost basic monetary segments impacting GDP and the economy of each nation. Moreover, risk is the main concept every insurance company would face [1, 2]. Within this strategy the risk is transferred form a policyholder to the insurance company in exchange for some amount of money called premium. Therefore, it is important to assess the risk as accurate as possible in order to set a fair premium. In technical term, assessing the risk of policyholders is called underwriting which is the base of issuing insurance policies [3, 4]. Obviously risk levels differ from person to person and consequently every individual should be treated with respect to the corresponding risk impose. Therefore, scientific assessment of the risk of customers is a need for insurance companies. This will help both insurers and customers feel satisfied. On one hand insurance companies become profitable and on the other hand customers pay based on their risk level which means individuals with lower risk will pay less amount of money as their premium. This keeps both cautious, low-risk customers and insurers satisfied. Unfortunately, at the moment, Iran insurance industry suffers from a structured and organized approach in rate making. This creates the main gap in insurance industry. Basically, everything is copied form experiences achieved in other countries. This exacerbates the situation in the field of car body insurance, which is contribute extensively to the industry portfolio. Car Body Insurance (CBI) is a type of policy compensates for damages resulting from accidents that occur to the insured. With the development in the field of data mining, tools and algorithms can be employed to assess

* Corresponding author: Tel: +98-23-31535711
E-mail address: *alireza.bahiraie@semnan.ac.ir*

and predict the risk of insured. There are number of researches investigating the risk analysis for car insurance through data mining techniques. Variables such as age, gender, marital status, driving records and driving experience were typically used to assess the individual risk [5, 6, 7]. Given the dataset from a major British insurer, the results of the liner regression in [8]. Provide prediction on claim costs for automobile insurance. The article points out age, area of residence, group (vehicle type), and no-claim discount as four independent variables that would significantly affect the result of claim costs. An unsupervised clustering was used in for thirteen variables [9]. The variables were divided into categorical and interval groups and the minimum size of cluster was set to 1000.
The proposed clustering used to classify potential customers into a risk group and determine their expected claim costs over 12 month period. For the clustering model, 90% of the risk classes had deviation within 20% while for the heuristic model it was only 72%. The use of clustering algorithm can be found in more articles [10, 11, 12]. Some articles point out the driver characteristics as the pillar of their study to conduct classification. This is due to the development of telematics and emergence of the usage-based insurance (UBI). One can find analysis mainly based on the behaviours of drivers [13, 14, 15]. In [13], the model leads to an AUC of 0.61 which is able to improve the predictive performance of a traditional approach by 3.58. Logistic regression and XGBoost we provided in [14] to predict accident claims with accuracy of 0.66 and 0.83 respectively. Furthermore, Poisson regression is employed as claim frequency model. As mentioned earlier, no extensive investigation in Iran was devoted to car body insurance. In fact, the main concern of articles cover third party liability. This is due to the nature of third party liability field which is mandatory by law and therefore insurers try to assess the risk more carefully. However, almost all accidents involve two parties; driver at fault and the cautious driver, which both may suffer losses and are covered by their third party and body insurance policy respectively. Moreover, studies in Iran follow the similar strategy as ones in abroad in the sense that they take into account combination of individual, car and behavioural features [16, 17]. The model accuracy in [16] was 0.87.
This can be a reasonable approach in case of having suitable data. However, in the Iran insurance industry the only reliable variables for car insurance are traits associated with cars. First of all, customers are not obliged to provide their personal information and even in case that they need to fill in the required information there is no strict checking mechanism to detect wrong data [18]. Therefor analysing features other than the automobile characteristics can yield misleading outcomes. Thus the research motivation and main contributions of this paper are: (1) the very first study focusing on the most reliable variables in car body insurance; (2) an algorithm suitable for typical datasets collected in Iran insurance companies; (3) a systematic applied approach to assess high and low risk customers. Concerning number (2) it should be noted that it is very common to encounter datasets in which there are too missing values and duplicate rows. Therefore, to get the proper data set so that reasonable outcomes are achieved, one needs to omit plenty of observations. This can leave the researcher with not too many records. In this case a proper strategy and algorithm should be taken in order to reach to the desirable outcome.

2 Methodology

In today's world of technology personal computers have become more powerful and ubiquitous, they are called upon to perform and solve an ever increasing number of tasks and problems. In order to solve these problems or perform tasks a computer must execute a computer program. The act of writing such a computer program can be very complex and difficult. Under the pressure of an ever-growing need for computer programs, genetic programming/evolutionary computing emerged as a promising method in which the computer itself performs this often daunting task of writing software and helps in decreasing complexity. Genetic programming is a collection of methods for the automatic generation of computer programs that solve carefully specified problems, via the core, but highly abstracted principles of natural selection. Genetic programming has automatically produced the result that are competitive to human intelligence and performance. The study of genetic programming investigates the efficacy and potential of using genetic algorithms for multiscale materials modelling and addresses some of the challenges involved in designing competent algorithms that solve hard problems quickly, reliably and accurately. We propose the use of genetic programming (GP), a genetic algorithm that evolves computer programs, for bridging simulation methods across multiple scales of time and/or length [19]. It is the compound breeding of (initially random), computer programs where only the relatively more successful individuals pass on genetic materials (program and program fragments) to the next generation. One of the earliest practitioners of the GP methodology [20], applied evolutionary algorithms to the problem of discovering finite-state automata. Lisp is a famous language of genetic programming which is based upon two principles: 1. it is often easier to write an algorithm that can measure the amount of success a given program has at solving a problem, than to actually write the successful program itself. 2. A less-than-fully-successful program often contains components, which, appropriated in the right way, could contribute to designs for more successful programs.

2.1 Genetic Programming (GP)

The high economic and social costs within failures have encouraged investigation for better understanding and prediction capability [21]. One can see a major shift in the methodology following the presence of Neural networks as the primary heuristic strategies to the problem of corporate finance in the 1990s [2]. Further studies have provided more comprehensive surveys on prediction methods such as GA [3]. After this, a relatively new technique for prediction, Genetic programming with more accurate classification model for prediction was introduced. Genetic programming (GP) is a search methodology belonging to the family of evolutionary computation. GP can be considered as an extension of Genetic algorithms (GAs) [4]. GA is a stochastic search technique that can search large and complicated spaces from the ideas of natural genetics and evolutionary principle. They have been demonstrated to be effective and robust in searching very large spaces in a wide range of applications. GA has been applied in wide range of financial fields such as trading system, stock selection, etc. GP is a GA applied to a population of computer programs (CP). While a GA usually operates on (coded) strings of numbers, a GP has to operate on CP. GP allows, in comparison with GA, the optimization of much more complicated structures and can therefore be applied to a greater diversity of problems [5].

While classification can be considered as a classification problem, we provide necessary description of GP with emphasis on its application in classification role [5]. The Darwinian theory of evolution inspired genetic programming models. According to the most common implementations, a population of candidate solutions is maintained, and after a generation is accomplished, the population is fitted better for a given problem. Genetic programming uses tree-like individuals that can represent mathematical expressions. Following such a GP individual is shown in Fig. 1. Three genetic operators are mostly used in these algorithms: reproduction, crossover, and mutation. First, the reproduction operator simply chooses an individual in the current population and copies it without any changes into the new population. In the second step, two parent individuals are selected and a sub-tree is picked on each one. Then crossover swaps the nodes and their relative sub-trees from one parent to the other. If a condition is violated the too-large offspring is simply replaced by one of the parents. There are other parameters that specify the frequency in which internal or external points are selected as crossover points. Fig. 2 and Fig. 3 show an example of crossover operators.

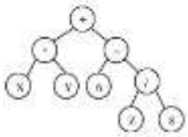

Fig. 1: Tree representation of the program (expression): (X*Y) + 6 - (Z/8)

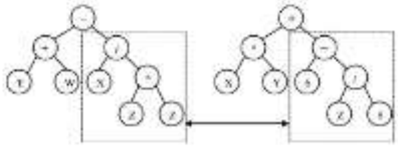

Fig. 2: Representation of crossover (parents)

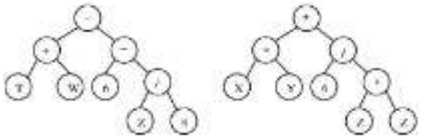

Fig. 3: Representation of crossover (children)

The mutation operator can be applied to either a function node or a terminal node, where in the tree is randomly selected. If the chosen node is a terminal node it is simply replaced by another terminal and if it is a function and point mutation is to be performed, it is replaced by a new function with the same parity. When tree mutation is to be carried out, a new function node is chosen, and a new randomly generated sub-tree substitutes the original node together with its relative sub-tree. A depth ramp is used to set bounds on size when generating the replacement sub-tree. Naturally, it is to check that this replacement does not violate the depth limit. If this happens, mutation just reproduces the original tree into the new generation [6].

Further parameters specify the probability with which internal or external points are selected as mutation points. To obtain the best fitness function for all classification problems and in order to apply a particular fitness function, the learning algorithms must convert the value returned by the evolved model into “1” or “0” using the 0/1 Rounding Threshold. If the value returned by the evolved model is equal to or greater than the rounding threshold, then the record is classified as “1”and “0” otherwise. There are many varieties of fitness function such as number of hits, sensitivity/specificity, relative squared error (RSE), mean squared error (MSE), that can be applied for evaluating performance of generated classification rules. We used “number of hits” as fitness function because of its simplicity and efficiency, which is based on the number of samples correctly classified. More formally, the fitness f_i of an individual program corresponds to the number of hits and is evaluated by $f_i = h$ where h is the number of fitness cases correctly evaluated or number of hits. So, for this fitness function, maximum fitness f_{max} is given by $f_{max} = n$ where n is the number of fitness cases. An example of mutation operator is shown in Fig. 4.

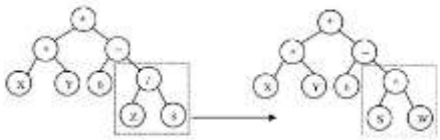

Fig. 4: Representation of mutation

Its counterpart with fitness measure f_i as “raw fitness”, rf_i and complements it with a parsimony term.

Parsimony pressure puts a little pressure on the size of the evolving solutions, allowing the discovery of more compact models. Thus, raw maximum fitness $rf_{max} = n$ and the overall fitness fpp_i is evaluated by:

$$fpp_i = rf_i \times (1 + \frac{1}{5000} \times \frac{S_{\max} - S_i}{S_{\max} - S_{\min}}),$$

where S_i is the size of the program and S_{max} and S_{min} represent minimum and maximum of program population respectively. Maximum and minimum of program sizes are evaluated by the formulas:

$$\begin{cases} S_{\max} = G(h+t) \\ S_{\min} = G \end{cases}$$

where G is the number of genes, and h and t are the head and tail sizes which are 0 and 1 respectively in classification problems. Thus, when

$$rf_i = rf_{\max} \qquad and \qquad S_i = S_{\min}$$

with the condition

$$fpp_{\max} = 1.0002 \times rf_{\max}$$

Once fitness function is defined, classification problem becomes a search problem of the best solution in the search space of all the possible solutions, that is to say an optimization of the fitness function for which optimization techniques can be used. Each rule is constituted by a logical combination of these ratios. The combination determines a class description, which is used to construct the classification rule.

2.2 Data

The employed data set is provided by an insurance company and contains 37904 records from 2013 to 2018. Age, Type (i.e. car type), Plate and Colour are the four independent features and the dependent variable is a binary categorical variable taking zero and one for no claim and filing a claim respectively as discribed in Table 1.

Table 1: Model Outcome Applied on the Imbalanced Data Set.

Category	Definition
Type:	Peugeot (Persia), Peugeot (RD), Peugeot206 (Type II), Peugeot206 (Type III), Peugeot206 (Type V), Peugeot206 (Type VI), Peugeot 206 hatchback, Peugeot 405, Peykan, Pride
Age:	before 2011, after 2011
Plate:	private, public

To test the procedure and accuracy of the results, the data set is splitted into two sets, training and testing sets which are splitted into 70:30 ratio which means 70% of the data is taken to train the model and the rest 30% is considered as the test set. As can be seen in Fig. 5, four categorical independent variables are divided into different classes. Note that we are only considering cars which are made assembled in Iran. Moreover, age is divided into two classes namely before and after 2011. This is due to the law which came into force by 2011 and obliged Iranian car factories to use ABS in their products. The main challenge that the authors faced is the imbalanced data set. Regarding imbalanced issue, 87% of the data set belongs to no claim class and the rest 13% contains claim class. Since this invalidates the high accuracy of the model due to the occurrence of data imbalance, we will provide a proper weight for the minority class so that the classifier can learn equally from all classes.

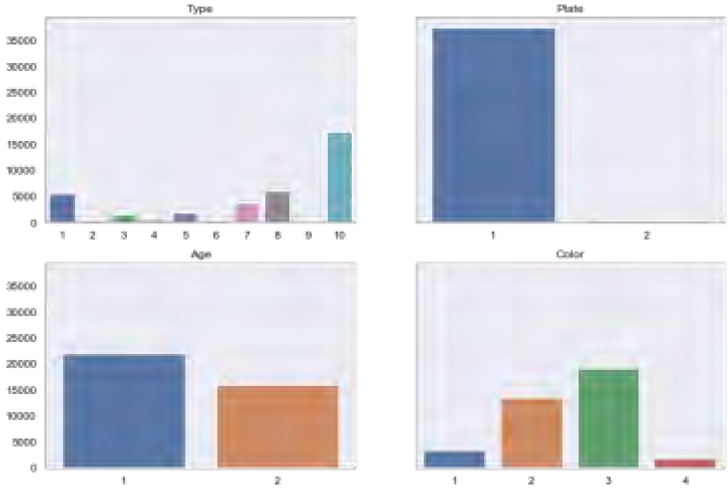

Fig 5: The bar charts illustrate the frequency of car type, car plate, and car colour and car age

Fig 6: Highly imbalanced data set between the binary classes (claim)

Table 2: Model outcome applied on the imbalanced data set.

Model	Train / Test Set	Accuracy
Kernel SVM	Train Set= 26532	0.91(rbf) 0.87 (poly) 0.81 (linear)
	Test Set= 11372	0.88 (rbf) 0.86 (poly) 0.77 (linear)
Genetic Programming	Train Set= 26532	0.96.8
	Test Set= 11372	0.92

3 Empirical Illustrative Results

As mentioned earlier we overcome the issue of imbalanced data by assigning the weight to the class with minor frequency. Notice that the challenge of imbalanced dataset is very common when analysing data in the field of auto insurance. The main problem is that the outcome cannot be very reliable and promising. Basically, the obtained high accuracy can't be generalized into new observations which is called bias error in classification. As reflected in the table, we obtain high accuracy. However, accuracy is not the best metric to evaluate the model performance when encountering highly imbalanced data set. In fact in the case we have skewed classes the suitable measure is considered to F1-score which is equal to zero. So we associate the higher weight to the minority class in order to get reliable results over the accuracy. The genetic programming decision trees are reported as claims were reported are indicated as 1 and non-reported as 0. Thus, a car will have a higher failure probability and classified as part of the failing group if its score is higher than the cut-off point for each approach. In this study, based on the identified factors by past studies and availability, 4 indices have been built using data. Significance of mean differences for each group are tested and the indices reflect different aspects of car structure and performance. Data are partitioned into two sets: the training set and the test set. The training set contains the known claims used during the evolution process to find an explicit classification rule able to separate an instance of a class of claimed insurance and from that of a non-claimed class.

In contrast, the test set is used to evaluate the generalization ability of the rule identified. Genetic programming process we employed using the Mahalanobis D^2 measure to select the variables that produced the greatest degree of effectiveness for separation of the two groups. This will create a more stable and well-balanced model. Subsequently, we tested the selected variables using Genetic Programming (GP) to illustrate that this new transformation will produce statistically more accurate predictions and can be used as an alternative to common values. Final regressions with different significant variables are obtained as significant indicators for each procedure. To test the accuracy, one value v_i for each company will serve as the cut-off point, and decisions will be made according to the following rules;

$\forall\ V_i \leq Cut-Off \Rightarrow non-claimed$

$\forall\ V_i > Cut-Off \Rightarrow claimed$

In the final regressions, different variables are identified as significant indicators from the selected list for each procedure.

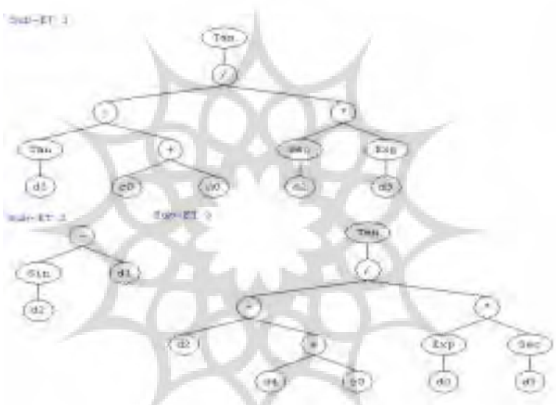

Fig. 7: The best GP model obtained

Fig. 7 shows the best GP model obtained for our approach. The model divided into three sub-trees, with each tree representing a gene, meaning that the model is a chromosome consisting of tree genes. The sum of the returns of the sub-trees for a firm are compared with the “Rounding Threshold” to determine the class of the insurance claim. Significant variables in each process presented in Table 3.

Table 3: Predictors used by Gp

T1C0	-4.22037
T3C0	4.36542
d0	Car type
d2	Car plate type
d3	Car Color
d4	Year production
d5	Car brand

The representation of a solution to the problem provided by the GP algorithm is in the form of a decision sub-tree. Decision trees for each approach consist of 3 sub-trees which are presented as T1, T2 and T3. Each node represented by constant value, independent variable or combination operator. The constant values in each sub-tree T1, T2 and T3, labelled as "C" and independent variables as "d". For example, the T1C1, represents the constant value in the first sub-tree which is "-5.57389". Each node which is assigned by operator is a function node taking one of the values from the set +, -, *, ^, EXP, etc. Operators that used in our study are shown in Table 4.

Table 4: Function nodes reported in decision trees in figures 2 and 3

Item	**Function**	**Item**	**Function**
+	Addition	Exp	Exponential
-	Subtraction	E	e
*	Multiplication	Pi	Π
/	Division	Log	Logarithmic
Sin	Sine	Asin	Arcsine
Cos	Cosine	Acos	Arccosine
Tan	Tangent	Atan	Arctangent
Cot	Cotangent	Acot	Arc cotangent
Sec	Secant	Asinh	Arcsine hyperbolic
Csc	Cosecant	Acosh	Arccosine hyperbolic
Sinh	Sine hyperbolic	Atanh	Arctangent hyperbolic
Cosh	Cosine hyperbolic	Acoth	Arc cotangent hyperbolic
Acsch	Arc cosecant hyperbolic	Csch	Cosecant hyperbolic
Asech	Arc secant hyperbolic	Sech	Secant hyperbolic
4RT	X^1/4	X3	X^3
5RT	X^1/5	X4	X^4

Table 5: Possible classification response

Symbol	Actual	Prediction
11	1:Claimed	1:Claimed
10	1:Claimed	0:non-Claimed
01	0:non-Claimed	1:Claimed
00	0:non-Claimed	0:non-Claimed

Since there are two possible classes of events, genetic programming process and algorithm considers "0.5" as benchmark for two-class classification decision-making problems. Thus, the benchmark value of 0.5 is used for possible probability whether a car is claimed or non-claimed through the genetic programming decision tree.

3.1 Misclassification test

The error rate is explored through a misclassification test as an alternative method. Within this method, a penalty number assigned for making any specific type of mistake. Penalty numbers represent cost of error rates and an average cost of misclassification is the weighted average of all errors. In Table 5, possible classifications and misclassifications are shown. Table 6 shows comparison accuracy for each classification model with respect to different data representations.

Table 6: Comparison accuracy of classifiers

	SVM (rbf)		Genetic Programming	
	Training	Testing	Training	Testing
11	18349	7613	18992	8102
10	1480	693	478	565
00	5796	1310	6691	2360
01	907	391	371	345
Total Accuracy	91%	88%	96.8%	92%

Table 6 exhibits the summarized accuracy level for SVM and GP procedures; clearly, the results improved under the genetic programming procedure. Moreover, the most contributing variables are illustrated in Figure 8:

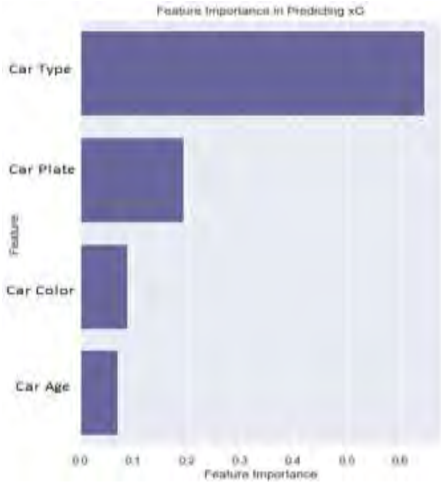

Fig 8: The most (car type) and the least (car age) contributing features

4 Discussion

In summary, the description results highlight the following evidence that, under the transformation process, better classification accuracy results are achieved. It is not only the pattern of influenced factors variation that is alternatively favourable to claims but also that of the non-claimed. However, Car type possesses the largest weight and second, third and fourth positions belong to car plate, car colour, car age respectively. It is worth to notice that car colour offers a very low impact on the occurrence of loss. This is due to the structure of available cars in Iran. Basically, the three dominant colours are white, silver and black which even globally do not affect the probability of an accident considerably. Finally, this paper comes up with an applied method to help insurance companies analyse their customers in more scientific way with current data structure. However, they need to think of a mechanism so that more features regarding drivers and their behaviours can be collected. This definitely increase the accuracy of the analysis and can yield fair premium

References

[1] Acharya, V., Pedersen, L., *Asset pricing with liquidity risk*, Journal of Financial Economics, 2005, **77**, P. 375-410, Doi:10.1016/j.jfineco.2004.06.007

[2] Bahiraie, A. and Alipour, M., *Jump OpVaR with operational risk*, International Journal of Computing Science and Mathematics, 2020, **12**(3), P.213-225, Doi:10.1504/IJCSM.2019.10025236

[3] Bahiraie, A., Azhar, A.K.M. and Ibrahim, A. *A new dynamic geometric approach for empirical analysis of financial ratios and bankruptcy*, Journal of Industrial and Management Optimization, 2011, **7**(4), P.947–965, Doi:10.3934/jimo.2011.7.947

[4] Bahiraie, A., Abbasi, B., Omidi, F., Hamzah, N.A. and Yaakub, A.H., *Continuous time portfolio optimization*, International Journal of Nonlinear Analysis and Applications, 2015, **6**(2), P. 103–112. Doi:10.22075/IJNAA.2015.258

[5] Bahiraie, A., and Alipour, M., *Option Pricing Accumulated with Operational Risk,* Advances in Mathematical Finance and Applications, 2020, **4**, P. 437-448, Doi:10.22034/amfa.2020.1897687.1409

[6] Black, F., Scholes, M., *The Pricing of Options and Corporate Liabilities*, Journal of Political

Economy, 1973, **81**, P. 637-654, Doi:org/10.1086/260062

[7] Chavez-Demoulin, V., Embrechts, P., Neslehova, J., *Quantitative models for operational risk: Extremes, dependence and aggregation*, Journal of Banking and Finance, 2006, **30**(206), P.2635-2658. Doi: 10.1016/j.jbankfin.2005.11.008

[8] Chorafas, D., *Operational Risk Control with Basel II: Basic Principles and Capital Requirements*, Butterworth-Heinemann, Oxford, 2004, Doi:10.1007/978-3-642-15923-7

[9] Cont, R., Tankov, P., *Financial Modelling with Jump Processes, second printing*. Chapman and Hall/CRC Press, London, 2004, Doi.org/10.1201/9780203485217

[10] Eshaghi, M. and Askari, G., *Hyper-Rational Choice and Economic Behaviour*, Advances in Mathematical Finance and Applications, 2018, **3**(3), P. 69-76, Doi:10.22034/AMFA.2018.544950

[11] Elandt, R., *The folded normal distribution: two methods of estimating parameters from moments*, Journal of Technimetrics, 1961, **3**, P.551-562, Doi:10.2307/ 1266561

[12] Hanson, F. B., Westman, J. J, *Jump-Diffusion Stock Return Models in Finance: Stochastic Process Density with Uniform-Jump Amplitude*, working paper, University of Illinois at Chicago, 2002, Doi:10.1016/j.finmar.2013.05.002

[13] Hull, J. and White, A., *The impact of default risk on the prices of options and other derivative securities*, Journal of Banking &Finance, 1995, **19**, P. 299-322, Doi: 10.1016/0378-4266(94)00050

[14] Izadikhah, M., *Improving the Banks Shareholder Long Term Values by Using Data Envelopment Analysis Model*, Advances in Mathematical Finance and Applications, 2018, **3**, P. 27-41, Doi:10.22034/amfa.2018.540829

[15] Jackson, P., Maude, D. J., Perraudin, W., *Bank capital and value at risk*, Journal of Derivatives, 1997, **4**, P.73-90, Doi: 10.3905/jod.1997.407972

[16] Khanizadeh, F., Khamesian, F., Bahiraie, A., *Customer Segmentation for Life Insurance in Iran Using K-means Clustering*, International Journal of Nonlinear Analysis and Applications, 2021, **1**(12), P.633-642. Doi:10.22075/IJNAA.2021.22324.2350.

[17] Klein. P., *Pricing Black-Scholes options with correlated credit risk*, Journal of Banking and Finance, 1996, **20**, P.1211-1129, Doi:10.1016/0378-4266(95)00052

[18] Merton, R. C., *Option pricing when underlying stock returns are discontinuous*, Journal of Financial Economics. 1976, **3**, P.125-144, Doi:10.1016/0304-405 (76)90022-2

[19] Mitra, S., *Operational risk of option hedging*, Journal of Economic Modelling, 2013, **33**, P.194-203, Doi.org/10.1016/j.econmod.2013.04.031

[20] Sette S, Boullart L. *Genetic programming: principles and applications.* Engineering applications of artificial intelligence, 2001, **14**, P.727-736. Doi:10.1016/s0952-1976(02)00013-1

[21] Su, X., Wang, W., *Pricing options with credit risk in a reduced form model*, Journal of the Korean Statistical Society, 2012, **41**, P.437-444, Doi:10.1016/j.jkss.2012.01.006

[22] Yeo AC, Smith KA, Willis RJ, Brooks M., *Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry*, Intelligent Systems in Accounting, Finance and Management. 2001,**10**, P. 39-50. Doi:10.1002/isaf.196