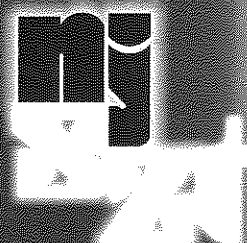


9



- Cultural Feature of Japan
- Controlling IT & Technology Innovations
- Linear Regression Model Selection
- Relationship between EAM and APM
- Evaluated Knowledge Representation
- Quality IS in Public Administration
- Allocating Time-Bound Tasks
- Tailoring Domain Ontology
- Measurement of Completed Projects
- Knowledge Intensive Testing
- eCompetences of ICT Professionals

G-I-K-O-F



SEFBIS Journal

Periodical of the Scientific and Educational Forum
on Business Information Systems and the IFP TC8
Enterprise Information System Working Group
HU ISSN 1788-2265

Editor in Chief:

RAFFAI, Maria (Hungary, Széchenyi University; Professor,
Honorary Secretary of IFIP, national representative in TC8)

Editorial Board:

CHROUST, Gerhard (Austria, Universität Linz)
DOBAY, Péter (Hungary, Pécs Univ. of Sciences)
GÁBOR, András (Hungary, Corvinus University of Bp)
LIN, Forrest (China, Chinese Institute of Electronics)
PRIES-HEJE, Jan (Denmark, Roskilde University)
RACSKÓ, Péter (Hungary, Corvinus University of Budapest)
TJOA, A Min (Austria, Technical University Vienna)
TRAVICA, Bob (Canada, University of Manitoba)
UCHIKI, Tetsuya (Japan, Saitama University)

Technical Editor: Tar, József, Publisher Palatia

Design: Perjés, András Gekko Design Studio

Publisher:

John von Neumann Computer Society – SEFBIS

Editor Responsible:

Alföldi, István managing director of NJSzT
Address: 1054 Budapest, Báthori u. 16.
Telephone/Fax: +36-1-472-2720; +36-1-472-2728
E-Mail: raffai@sze.hu
WebSite: <http://gikof.njszt.hu/bekoszo>
SEFBIS-Online: http://gikof.njszt.hu/sefbis_journal

Sponsors



Foundation Alexander

ISC Alapítvány



The SEFBIS Journal is published by the SEFBIS Special
Interest Group of John von Neumann Computer Society.

Content

Raffai, Mária:

Preface 1

Research Results

Uchiki Tetsuya:

The Cultural Feature of Japan 2

Kristóf, Péter:

Controlling IT and Technology 7

Láng, Blanka – Kovács, László:

Linear Regression Model Selection 15

ICT in Business

Mozsár, Livia – Michelberger, Pál:

The Relationship between EAM and APM 22

Szmodics, Péter:

Evaluated Knowledge Representation 28

Orbán, Anna:

Quality of IS in Public Administration 36

Case Studies

Szikora, Péter:

Allocating Time-Bound Tasks 42

Neusch, Gábor:

Tailoring Domain Ontology 51

Sinóros-Szabó, Laura:

Measurement of Completed Projects 59

Education

Weber, Christian:

Knowledge-Intense Testing 66

Dobay, Péter:

eCompetences of ICT Professionals 75

SEFBIS' Decisions 2014 84

Conferences

Events and Conferences Worldwide 85

Call for Papers WCC'2015 86

Call for Papers Big Data IS Conference 87

Gábor, András:

Report on ISBIS-OGIK Conference'2014 88

PÉTER KRISTÓF received the BSc degree in computer science from the Denis Gabor College while completing his studies at the Jönköping International Business School, Sweden, his MSc degree in business administration at the University of Pécs and another MSc degree in economics from the FernUniversität in Hagen, Germany. He is a PhD candidate at the Corvinus University of Budapest focusing on innovation management and business modelling of ICT enterprises. He is CEO of a software development company and the founder of a consultancy firm focusing on innovative projects and international business development. He co-founded Pannon Entrepreneur Meetup and is advisory board member at the National Council for Telecommunications and Informatics.



Linear Regression Model Selection using Improved Harmony Search Algorithm

BLANKA LÁNG – LÁSZLÓ KOVÁCS

Professor, Corvinus University of Budapest, blanka.lang@uni-corvinus.hu
Student, Corvinus University of Budapest, laszlo.kovacs20@stud.uni-corvinus.hu

ABSTRACT

The selection of independent variables in a linear regression model is a problem worth investigating. In this paper an improved harmony search metaheuristic algorithm (IHSRS) is presented for this problem. First, we investigate an existing genetic algorithm solution (GARS). GARS can tackle the problem successfully, but the solution can contain non significant variables - a variable is called non-significant, if when it changes by a unit, we can not clearly determine that the value of the dependent variable will increase or decrease. So we integrate a new criterion into the GARS to tackle this problem and find that this new criterion makes the GARS less effective. Second, we compare the performance of the improved GARS and IHSRS. We found that our solution is more robust and more effective.

Introduction

The selection of independent variables in a linear regression model (LRM) is a challenging problem. The number of the possible selections is $2^m - 1$ (m is the number of independent variables), if the empty set is excluded. The traditional solutions – forward and backward eliminations, Monte Carlo simulation etc. – can often converge to a local optimum and can not manage the effects of the multicollinearity phenomenon.

If a population based metaheuristic is used, the solution will much more likely converge to a global optimum, and the effects of the multicollinearity phenomenon will be managed. In this paper a former genetic algorithm solution (GARS) is investigated at first. We will prove that this kind of metaheuristic – although can converge much more likely to a global optimum than the traditional solutions,

can not fit properly to the linear regression selection problem, because the solution of this metaheuristic can contain non significant variables. First a new criterion is integrated into the former objective functions: the solution must include significant variables only. This improved GARS can tackle the problem successfully, but we would like to present another solution which has better runtime. We will prove that our harmony search metaheuristic fits better this problem due to its totally random nature, and its average solution time and the standard deviation of the solution times are significantly better.

LRM Selection Problem

Let n be the number of the observations, and let $X = \{X_1, X_2, \dots, X_m\}$ be the set of m possible independent variables, Y the dependent variable in a multivariate linear regression model. Let's suppose that this model describes the relationship between the independent and the dependent variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

where $\bar{X} \subseteq X$ is the set of $p \leq m$ independent variables – the regressors –, and

$$\beta = \{\beta_0, \beta_1, \dots, \beta_p\}$$

is the parameter set, ε is the disturbance term. Let's estimate the β set as

$$\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\}$$

by using the Ordinary Least Squares method. Once this is done, we need to choose the possible independent variables containing $\bar{X} \subseteq X$. We would like to choose this set, such that the

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_p X_p$$

model is optimal. In this paper similarly to the the authors of GARS – we choose AIC and BIC criteria to measure the degree of optimality of the model. In the literature there are a lot of AIC, BIC formulas. These criteria reward the decrease in the Sum of Squared Errors (SSE) of the model, and simultaneously penalize them based on the number of regressors. The lower the AIC or BIC values are, the better the model is. To compare GARS and IHSRS we chose the following formulas [5]: Akaike information criteria

$$AIC = \ln\left(\frac{SSE}{n}\right) \quad (1)$$

where n is the number of observations, p is the number of independent variables in the model, and

$$SSE = \sum_{i=1}^n (y_i - \bar{y}_i)^2 \quad (2)$$

where $\bar{y}_i$ is the i th element of $\bar{Y}$, and y_i is the i th element of the dependent variable. Bayesian-Schwarz Information Criterion (BIC or SBC):

$$SBC = \ln\left(\frac{SSE}{n}\right) + \frac{2p}{n} * \ln(\sqrt{n}) \quad (3)$$

In addition to the above we choose adjusted R-square indicator ($\bar{R}^2$):

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (4)$$

$$R^2 = 1 - \frac{SSE}{SST} \quad (5)$$

$$\bar{R}^2 = 1 - \frac{n-1}{n-p-1} * (1 - R^2) \quad (6)$$

where $\bar{y}$ is the average of the elements of the dependent variable. In a multivariate linear regression model multicollinearity can be interpreted as a type of redundancy. Let us suppose, that X_1 and X_2 are redundant regressors. Furthermore, if X_1 is in the model in itself, gives a value (AIC, BIC or $\bar{R}^2$) which is closer to the optimum than the value which X_2 gives, if that is in the model in itself, so X_1 would be better regressor for us than X_2 . But if X_1 and X_2 are both present in the model r , backward elimination, for example may prefer X_2 , and eliminate X_1 . In this case X_1 becomes a non-significant variable (a variable is called non-significant, if when it changes by a unit, we can not clearly determine that the value of the dependent variable will increase or decrease - at a given level of certainty, most commonly 95%), and the solution that includes X_2 is a local optimum, so we lost the better solution that includes X_1 "for ever".

The traditional solutions such as forward and backward eliminations, Monte Carlo simulation, etc. can not manage these effects of the multicollinearity phenomenon, so they can converge to a local optimum. The population based algorithms, such as genetic algorithm, harmony search metaheuristic, etc. can tackle this phenomenon, because these algorithms work with populations of individuals, so they can explore the complete search space simultaneously.

Genetic Algorithm for LRMs

To the best of our knowledge, there exists only one former metaheuristic procedure for the regression model selection problem, Minerva and Paterlini proposed a genetic algorithm solution, called GARS [4]. GARS has been tested on two real datasets. Unfortunately, we could find only one of them, so we

can work only with "Body Fat Measurement" dataset [1]. In this dataset the dependent variable is the percentage of the body fat, and the 16 dependent variables are: age, weight, height, chest and neck circumference, etc. This set consists of 252 observations.

The authors gave detailed empirical results, but unfortunately these results were insufficient for us to compare their algorithm with our solution. For example there is no information on the sequence order of the data when training and validation sets had been defined, and also, there is no information on the errors in the dataset – and if they were corrected by the authors. [1]. We do not know the names of the selected independent variables, only their numbers, so we could not identify them. We do not know, for example the name of the third variable. Therefore we could not use the proposed empirical results, so if we would like to compare the two solutions, then we need to reconstruct the GARS algorithm according to the article. In the Figure 1 the reconstruction of GARS is shown.

In the Minerva – Paterlini article the dataset was partitioned into three disjoint subsets: a *training set* (40% of the data), a *validation set* (40% of the data), and the *test set* (20% of the data). The regression coefficients ($\hat{\beta}_i$) were estimated using the training set. The statistical criteria – AIC, BIC values – were calculated using the validation set. Adjusted R^2 was computed using the test set.

First the algorithm parameters that are required to solve the optimization problem are specified. Unfortunately we could not use the values given in the article, because the authors did not give the readers all necessary information. For example we did not know exactly, what the mutation probability means: the probability of modifying each genes in an individual or the probability of modifying an individual. Additionally the authors did not explain some decisions: why is the crossover point just at the eight gene of the individual, and why the generation number equals to 1000, etc. So we specified the missing decision variables in such a way that results in the best AIC, BIC and $\bar{R}^2$ values. The GA individuals are vectors, a string of binary cells. If the i -th cell has the value 1, than X_i is in the model, other-

wise it is not. We evaluate each solution with respect to their objective functions, and we would like to minimize these AIC, BIC values and maximize the $\bar{R}^2$ value. Single point crossover operator, mutation operator, and elitism selection operator is used. The population size is 100, mutation probability 2%, the elit rate is 1%. When we changed the fix crossover point to a random one, the rate of the stable convergences increased significantly – from 40% to 60% –, so we chose random crossover point.

After the initialization of the parameters we initialize the first population. It is filled with randomly generated solution vectors – the individuals – and sorted by the values of their objective function, AIC or BIC or $\bar{R}^2$. Then the members of the new population are selected from the old population with respect to the elit rate, and for the residual spaces the individuals are generated with crossover operators. Afterwards, each gene is modified by the given mutation probability. At this point we can load the new population.

New and new populations are generated as long as the termination criterion will be satisfied. The original GARS solution has a remarkable drawback: the solutions of this metaheuristic may include non significant variables (at significance level of 5%). If we integrate a new criterion into the former objective function – the solution must include only significant variables – we find that this new criterion makes the GARS less effective. The average runtime necessary to achieve the same stabil convergence increases significantly. The standard deviation of the runtimes also highlights the drawbacks of this new objective function in the GARS. All of this can be seen in Table 1. The termination criterion is chosen in this way: the algorithms are stopped when we can regard the results as the global optimum. We consider a result as the global optimum when the last 100 AIC, BIC and $\bar{R}^2$ values do not change – do not increase, and do not decrease. We can consider these results as the global optimum, because in this case if we examine the next 8000 generations, the results stay the same. While we stop the improved GARS when the last 200 AIC, BIC and $\bar{R}^2$ values do not change. This modification is necessary to obtain as stabil convergences as the original algorithm

❖ Linear Regression Model Selection

obtains. Because in this case if we examine the next 8000 generations, the results stay the same. On the other hand, according to our examinations, if we use the termination criterion from the original GARS, the results change in the next 8000 generations.

AIC, BIC and $\bar{R}^2$ values were computed as the average of 30 runs, as in the cited paper. The final AIC value is -2.013 and BIC value is -1.987, and we got only one regressor in both cases: the 1st variable, the "density". The final $\bar{R}^2$ value is 0.9829 and

we got more regressors. In the case of the original GARS we got 2 insignificant variables: 2 (age) and 11 (thigh circumference).

These algorithms converge to a global optimum due to its population based nature - according to our experience the AIC, BIC and $\bar{R}^2$ values will not change if we stop the algorithm anytime later.

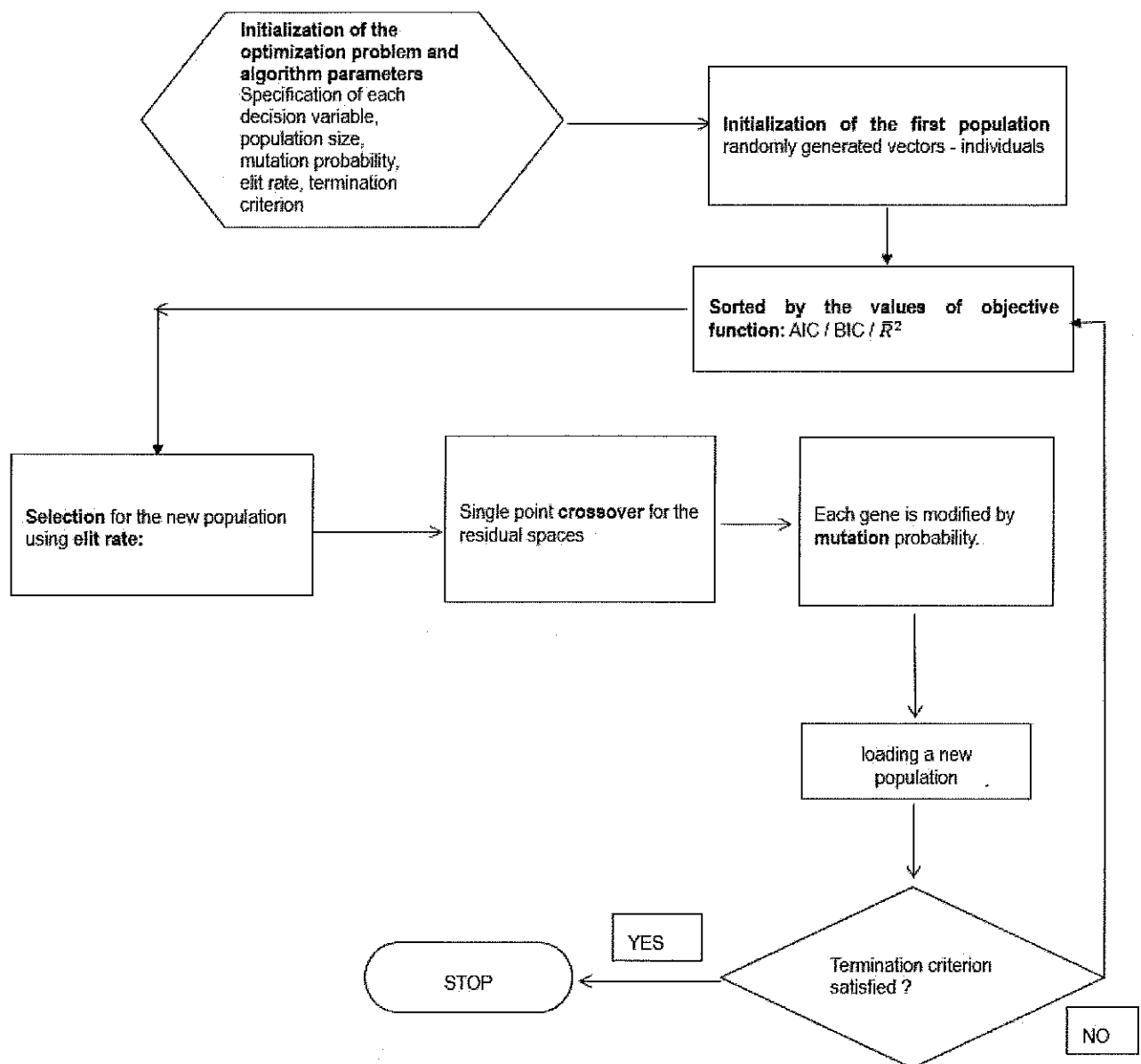


Figure 1. Reconstruction of GARS algorithm

Table 1. empirical results, improved GARS vs original GARS

	Variables:1		average solution time (sec)	standard deviation of solution times (sec)	adjusted R-square
	AIC	BIC			
original GARS	-2.013	-1.987	5.9209	2.750122	0.9829 Variables: 1, 2, 3, 5, 6, 8, 11, 12, 15
improved GARS	-2.013	-1.987	11.26783	8.646739	0.9822 Variables: 1, 3, 5, 6, 8, 12, 15

Where Variables 1: Density, 2: Age, 3: Weight, 5: Adiposity index, 6: Fat Free Weight, 8: Chest circumference, 11: Thigh circumference, 12: Knee circumference, 15: Forearm circumference

This improved GARS can tackle the “selection of independent variables in a linear regression model” problem successfully, can manage the effects of the multicollinearity phenomenon and the problem of the significant variables. However, we would like to get another solution which has a better runtime, because of the worrying increase in both the average runtime (91% increase), and the standard deviation of the runtimes (215% increase) of the improved GARS. We concluded that we should use another kind of metaheuristic, which nature and behavior fits better the problem. In an arbitrary population the rate of the “good solutions” – solutions which consist of only significant variables – is very low. If a genetic algorithm is used, when the crossover and mutation operators are applied, they can “spoil” these “good solutions” with a high probability, and the “spoiled solutions” can enter into the new population instantly. In our opinion harmony search metaheuristic, for example, would be a better choice, because of its strong random nature and its more strict selection rule.

Harmony Search Algorithm

Harmony search metaheuristic was developed in 2005 by Lee and Geem [2] in an analogy with music improvisation process, where music players improvise to obtain better harmony. In the harmony memory, (called repertoire) there are vectors (called melodies in the language of music). The aesthetic value of a melody is represented by the fitness function value. Namely, the higher this value is, the higher the quality of the melody is. With improvisations we would like to maximize (or minimize) the value of this function, to find the global optimum - in the language of music: a harmony. We applied this algorithm to our problem. A HS melody is a vector,

in our case a string of binary cells. If the i th cell has value 1, then X_i is in the model, otherwise not. Each solution is evaluated with respect to their objective functions, we would like to minimize these AIC, BIC and $\bar{R}^2$ values. First, the algorithm parameters that are required to solve the optimization problem are specified. We specified the decision variables in such a way to obtain the best AIC, BIC and $\bar{R}^2$ results. The harmony memory size is 5. The improvisation process is driven by two parameters – they are probability values between 0 and 1: HMCR (harmony consideration rate), PAR (pitch adjusting rate). In our case both equal to 0.7. After the initialization of the parameters we initialize the first population. It is filled with randomly generated solution vectors –melodies– and sorted by the values of the objective function, AIC or BIC or $\bar{R}^2$. Then the new melody improvisation phase follows. According to the harmony consideration rate we choose a melody from the memory, or we generate a totally new, random melody with probability $1 - \text{HMCR}$. According to the pitch adjusting rate, the melody, selected from the memory, is modified with probability PAR.

Afterwards, when the new melody is better than the worst in the repertoire, the worst one will be replaced by the better one, otherwise we through away the new one. A new melody is better for us than the old one, if its AIC/BIC value lower than the old one's value or its $\bar{R}^2$ value is higher than the old one's value. It is also required that the solution includes only significant variables. If a melody contains a non-significant variable, we through it away. In the HM, there are only melodies with significant variables. According to the applied replacement strategy, the quality of the population is increasing step by step. New and new melodies are generated as long as the termination criterion will be satisfied.

❖ Linear Regression Model Selection

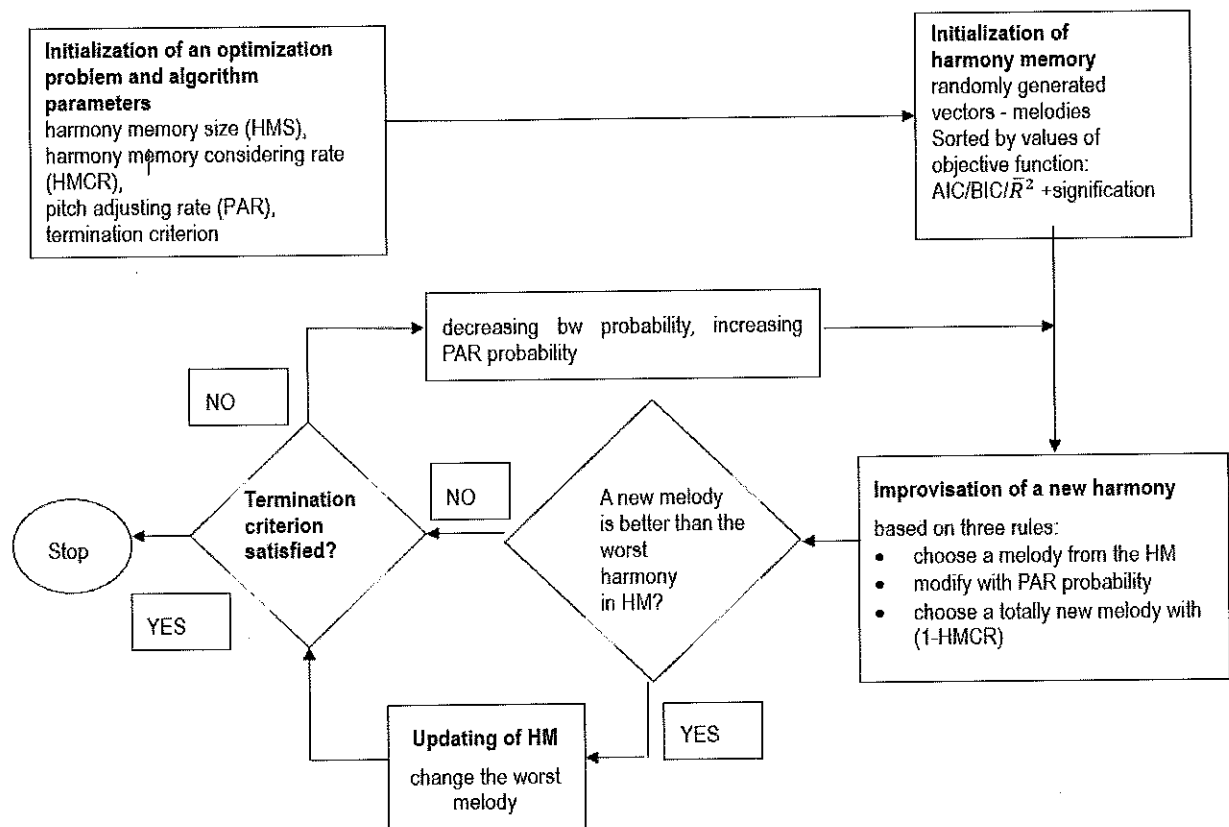


Figure 2. IHSRS algorithm

At this point our algorithm is tested, but we were not satisfied with the results. So we applied a “finetuning” of the HS. The HS was improved in 2007 by Mahdavi et al. [3]. In the improved harmony search (IHS) algorithm the only but very important novelty is the way of adjusting two parameters: PAR and *bw*. To improve the performance of the HS algorithm IHS algorithm uses variables PAR - the rate of pitch adjustment, and *bw* an arbitrary distance bandwidth, the measure of the modification: $x_{new} = x_{old} + bw * r$ x_{new} is the new melody, x_{old} is the melody chosen from the memory. *bw* is the bandwidth, the size of the searching space.

r is a uniformly distributed random number in the (-1,1) interval. But in our original algorithm there was no *bw* parameter. So we interpreted *bw* as a mutation probability of a pitch: during the run of the algorithm *bw* value is decreasing from 60% and the PAR value is increasing from 70% depending on the generation number. According to our experience, the finetuning greatly increased the performance of our algorithm: the standard deviation of the running time significantly improved from 0.107 to 0.035. In the Figure 2 the flow chart of IHSRS is shown.

Table 2. empirical results, improved GARS vs. IHSRS

	Variables:1		average solution time (sec)	standard deviation of solution times (sec)	Adjusted R-square
	AIC	BIC			
IHSRS	-2.013	-1.987	0.9682	0.035311	0.9822 Variables: 1, 3, 5, 6, 8, 12, 15
improved GARS	-2.013	-1.987	11.26783	8.646739	0.9822 Variables: 1, 3, 5, 6, 8, 12, 15

For comparison, the termination criterion is determined the following way: both algorithms are stopped when we can consider their results as the global optimum. The termination criterion of the improved GARS is defined earlier. The termination criterion of the IHSRS is set this way: we consider a result as the global optimum when the last 1500 AIC, BIC and $\bar{R}^2$ values do not change – do not increase, and do not decrease. We can consider these results as the global optimum, because in this case if we examine the next 8000 generations, the results do not change. Note that the conspicuous difference between the two termination criteria (200 and 1500) is due to the different structure and behaviour of the two metaheuristics. For example, new populations are created by different operators, and the sizes of the two populations are conspicuously different (5 and 100), and so on. Both algorithms provide the same AIC, BIC and $\bar{R}^2$ values - due to the termination criteria. The average solution time and the standard deviation of the solution times of the improved HSRS are significantly better than the improved GARS value ($\alpha = 5\%$), based on 30 runs of each algorithm.

Results

The paper presents an improved harmony search algorithm for the selection of independent variables in a linear regression model. First an existing genetic algorithm solution (GARS) was investigated. We proved that although the GARS can tackle the "selection of independent variables in a linear regression model" problem successfully, but the solution can contain non significant variables.

So a new criterion was integrated into the GARS to tackle this problem. Then, the improved GARS could tackle this problem successfully, could manage the effects of the multicollinearity phenomenon and the problem of the significant variables, but we found that this new criterion makes the GARS less effective. So our conclusion was that we should use another kind of metaheuristic, which nature and behavior fits better to the problem. It was shown that improved harmony search metaheuristic (IHSRS), is a better choice, because of its strong randomizing ability. Secondly, the performances of the improved GARS and IHSRS were compared. We found, that the average runtime and the standard deviation of the runtimes of our solution was better, which makes it more robust and more effective. Computational results were presented for a real world dataset.

References

- [1] Johnson R. W., (1996). Fitting Percentage of Body Fat to Simple Body Measurements. *Journal of Statistics Education* 4(1), 265-266.
- [2] Lee, K. S., Geem, Z. W. (2005). A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice. *Computer Methods in Applied Mechanics and Engineering*, 194, 3902–3933.
- [3] Mahdavi, M., Fesanghary, M., Damangir, E. (2007). An improved harmony search algorithm for solving optimization problems. *Applied Mathematics and Computation*, 188, 1567-1579.
- [4] Minerva, T., Paterlini, S. (2010). Regression Model Selection Using Genetic Algorithms. *Proceedings of the 11th WSEAS International Conference on RECENT Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing*, 19-27.
- [5] Ramanathan, R. (1992). *Introductory to econometrics with applications* (Vol. 4). Dryden Press.

BLANKA LÁNG received her MSc degree in mathematics from the Eötvös Loránd University in 1995, and another MSc degrees, one in informatics also from the Eötvös Loránd University in 1999 and another in economics from the University of Pécs in 2005. Her PhD degree in economics were obtained from the University of Pécs in 2011. Her thesis deals with the resource constrained project scheduling problem with discounted cash flows – a new metaheuristic. Currently she works as a senior lecturer at the Corvinus University at the Computer Science Department. Her research interests include metaheuristic algorithms and web development.





LÁSZLÓ KOVÁCS is a BSc student in Business Information Systems at Corvinus University of Budapest. He is to receive his degree in 2015. As a student, he actively participated in the Scientific Students' Associations Conferences at Corvinus University in the years 2013 and 2014. His research involving metaheuristics was awarded both times. He works as a teaching assistant at the Department of Statistics in Corvinus University. He teaches seminars in Statistics for BSc students.

The Relationship between Enterprise Architectural Management and Application Portfolio Management in Large Companies

ALICE LÍVIA MOZSÁR KOVÁCSNÉ – PÁL MICHELBERGER

Application Portfolio Manager, IBM CIO mozsar.livia@kgk.uni-obuda.hu

Associate professor, Vice dean for research, Óbuda University, michelberger.pal@kgk.uni-obuda.hu

ABSTRACT

Architectural management has many points to connect with the different areas in a large organization. Some of these are: Requirements Management, Project Portfolio Management, Strategy Management, Financial Management, Compliance Management, Risk Management, Process Management, Security Management, Domain Management and Application Portfolio Management (APM). This article focuses on enterprise architecture and application portfolio management relationship how the business units and brands can generate long term value from the strong and effective cooperation of these two areas in long time.

Introduction

The goal of the article is to document and explore the definition for enterprise information architecture, architecture process and to show the importance of enterprise architecture management (EAM). Another goal is to discover the importance of the relationship and connection between architectural management and application portfolio management in large companies worldwide. Large companies need to manage many thousands of business processes and information systems day by day. One of the key criteria for quick response in the marketplace to have a well managed and built by structural way information across organization. Enterprise architecture management helps support the business to reach strategic goals and to deliver the solutions to reduce the redundancies between business processes, and also to reduce the IT costs. Architectural management has many points to connect with the different areas in an organization.

Some of these are: Requirements Management, Project Portfolio Mgmt, Strategy Mgmt, Financial Mgmt, Compliance Mgmt, Risk Mgmt, Process Mgmt, Security Mgmt, Domain Mgmt and Application Portfolio Mgmt.

The method of the research is secondary research. The article integrates the theory and practices from this discipline and gives a brief overview about the principles of architectural management and application portfolio management while integrating the daily tasks and processes into one model.

Definitions: EIA and EIP

The objective of information management is to support the corporate information and data management with data sharing cross process integrations inside the company and improve the business values –customer satisfaction, employee productivity, cost avoidance, cost reduction, business