ATTRIBUTE SELECTION BY MULTIOBJECTIVE EVOLUTIONARY COMPUTATION APPLIED TO MORTALITY FROM INFECTION IN SEVERE BURNS PATIENTS

A. Jara, R. Martínez, D. Vigueras, G. Sánchez and F. Jiménez
Department of Communications and Information Engineering, University of Murcia, Murcia, Spain

Keywords: Attribute selection, Evolutionary computation, Multiobjective optimisation, Data mining.

Abstract: The problem of selecting variables in data-mining can be modelled as an optimisation problem involving multiple objectives which must be simultaneously optimised. This contribution proposes a multiple objective optimisation model for the problem of selecting variables applicable to the classification of mortality in patients from a hospital burns unit. The evolutionary multiobjective algorithm NSGA-II was adapted to resolve the proposed multiobjective optimisation model proposed and the results obtained were compared with those obtained with a battery of algorithms intended for selecting variables included in the Weka data-mining platform. The comparison underlines the efficacy and suitability of the proposed model and of the use of multiobjective evolutionary computation in this type of problem.

1 INTRODUCTION

The infection-related mortality rate in Intensive Care Units (in Spanish, UCI) exceeds 15,000 patients per year, while infection is the most common cause of reality in burns units. A rapid classification system is therefore of great interest for determining the most suitable treatment for ensuring recover. Studies carried out by the Spanish Society of Intensive, Critical and Coronary Medicine Units (in Spanish, SEMYCYUC) at national level (Society, 2007) and the Cornell Hospital Medical Center of New York in a study of 937 patients in their burns unit (Curreri et al., 1980) underline this interest.

Consequently, the ultimate aim was to implement and evaluate an easy to use system for classifying patients in a burns unit, to obtain the highest possible percentage of correct decisions (Jiménez et al., 2009) and to reduce the number of variables, thus making the system easier to understand by staff and restricting the number of clinical tests necessary to perform. This article will focus on the second of these objectives, using a multiobjective optimisation model to select the variables.

Evolutionary computation (Goldberg, 1989) has been successfully applied to optimise multiobjective problems (Deb, 2001),(Coello et al., 2002) and, particularly, to generate models for classifying patients such as those suffering leukaemia (Kumar et al., 2007) and for selecting variables (Pappa et al., 2002).

In this article is proposed an evolutionary focus for muliobjective optimisation to select variables in the context of mortality through infection among burns patients. A set of solutions denominated Pareto solutions are obtained in order to enable the user to choose the non-dominated solution that best suits, depending on the decision environment. The well known NSGA-II algorithm (Deb et al., 2002) is used to solve the optimisation problem for attribute selection. This algorithm is elitist and based on Pareto's concept, in which all the objectives are optimised simultaneously in the search for non-dominated solutions using, additionally, an explicit diversity mechanism.

To assess whether muliobjective evolutionary computation is suitable for the nature of the problem in hand, the results obtained with the multiobjective optimisation model proposed and solved by NSGA-II are compared with the results obtained using sixteen different techniques for attribute selection incorporated in the Weka platform of data mining (Hall et al., 2009).

The paper is structured as follows. Section 2 proposes a multiobjective optimisation model for selecting the attributes. Section 3 describes the basic components of the evolutionary NSGA-II multiobjective optimisation algorithm proposed. Section 4 describes

Jara A., Martínez R., Vigueras D., Sánchez G. and Jiménez F..
ATTRIBUTE SELECTION BY MULTIOBJECTIVE EVOLUTIONARY COMPUTATION APPLIED TO MORTALITY FROM INFECTION IN SEVERE BURNS PATIENTS .
DOI: 10.5220/0003126904670471
In *Proceedings of the International Conference on Health Informatics* (HEALTHINF-2011), pages 467-471
ISBN: 978-989-8425-34-8
Copyright © 2011 SCITEPRESS (Science and Technology Publications, Lda.)

the experiments carried out to validate the model, the results and a comparison with other techniques for attribute selection. Finally, Section 5 presents the conclusions and suggests possible future work.

2 MULTIOBJECTIVE OPTIMISATION MODEL FOR ATTRIBUTE SELECTION

The selection of variables concerns finding the smallest subset of variables in a data base to obtain the most accurate classification possible (Pappa et al., 2002). Described more formally, with X being the number of variables in an initial set T, the algorithm finds a subset P of Y variables from the set T, where Y $\leq$ X, with the aim of removing the irrelevant or redundant variables, and obtaining good accuracy in the classification (Aguilera et al., 2007). Therefore, the problem of attribute selection can be approached as a multiobjective optimisation problem (Deb, 2001), the solution of which comprise as set of solutions called non-dominated solutions (or Pareto solutions). Solution x dominates another solution y if (Deb, 2001):

- Solution x is not worse than y for any of the purposes in mind;
- Solution x is strictly better than y for at least one of the objectives.

For the variables selection problem in mind, two optimisation criteria have been considered: accuracy and compactness. To formulate these criteria the following quantitative measures have been defined .

Given a solution $x = \{ x_i \mid x_i \in T \}$:

- Accuracy. Based on the classification ratio $CR(x) = \frac{\Phi(x)}{N}$, where $\Phi(x)$ is the number of data correctly classified for a set of variables, x, by a given classification algorithm, and N are the total number of data.
- Compactness. By cardinality the $card(x)$ of the set x is established, that is, the number of variables used to construct the model.

In this way, the optimisation model proposed with the criteria defined is the following:

$$\begin{array}{ll} Maximize & CR(x) \\ Minimize & card(x) \end{array} \quad (1)$$

The objectives were to increase the accuracy of the model and to reduce the number of variable to the greatest extent possible. In some cases, such as will be presented, it was interesting to sacrifice accuracy slightly, when the number of variables were reduced significantly, in order to, simplify the model . Such as can be appreciated, the objectives in the optimization model 1 are contradictory since a lower number of significant variables means a lower classification rate and vice versa, that is the greater the number of variables the greater the classification rate. The solution to model 1 is a set of $m \leq X$ non-dominated solutions $C = \{x^k, k \in S\}$, $S = \{1, \ldots, X\}$, where each solution x^k of C represents the best collection of significant k variables. For example, for $X = 5$ (5 variables to be selected), a set of non-dominated solutions $C = \{x^3, x^5\}$ means that the Pareto front is composed of non-dominated solutions of 3 and 5 variables, respectively. The solutions with 1, 2 and 4 significant variables are not on the Pareto front and will therefore be dominated.

3 MULTIOBJECTIVE EVOLUTIONARY COMPUTATION FOR ATTRIBUTE SELECTION

Three elements can be distinguished in a variables selection algorithm (Aguilera et al., 2007).

- A search algorithm, which explores the space of the variables available.
- An evaluation function, which provides a measure of the fitness of the variables chosen. According to how this function is designed, the selection algorithms can be classified as filter models or embedded models. The former use measures that take into account the separation of classes based on information distance metrics, dependency metrics, etc., wile the latter use an estimate of the accuracy attained by a classification algorithm using selected variables.
- A fitness function that validates the subset of variables, which are finally chosen.

Evolutionary Computation has been used both for filter and embedded models. The work described here falls into the latter category since the accuracy and the simplicity of the cklassification obtained id one of the fundamental objectives. The NSGA-II (Deb et al., 2002) algorithm, the principal components of which are briefly described below, is used to resolve the problem described in 1.

Representation of Solutions. A binary codification of fixed length equal to the number of variables in the problem is used. In this way, a gene of value 1 in the

locus i of the chromosome means that the variable x_i has been selected, while 0 means that variable x_i has not been selected.

Initial Population. The initial population is generated randomly using a uniform distribution in the domain.

Suitability Function. The NSGA-II algorithm minimises the following two evaluation functions.

$$f_1(x) = -CR(x)$$
$$f_2(x) = \Phi(x)$$

where $CR(x)$ is the classification rate obtained using the algorithm C4.5 (Kotsiantis, 2007), and $\Phi(x)$ is the number of genes with a value 1 of the chromosome x. Specifically, classification rate $CR(x)$ is obtained from the classification ratio carried out for 25% of the cases of the knowledge base in a decision tree generated with the algorithm C4.5 constructed with 75% of the cases, that is, the rest of the knowledge base.

Genetic Operators. It is based on the uniform cross and the uniform mutation operators.

4 EXPERIMENTS AND RESULTS

This work is based on data from the Hospital Information System of an Intensive Care Unit (ICU) since 1999 to 2002, using the data (assessed by hospital staff) of 99 patients with different complications. The specialists consulted consider that the parameters described in Table 1, from the Electronic Health Record, are relevant for establishing the survival of patients with infections in an ICU.

For this study we selected the records of 99 patients to form part of the knowledge base (in the format used for the Weka platform in Table 1). As can be seen, the problem consists of 5 real entries, 12 discrete Boolean type entries and a Boolean output (prognosis of death). The NSGA-II algorithm was run 100 times with the parameters shown in Table 2

To evaluate the results is used the hypervolume (Deb, 2001) metric, which calculates the fraction of the space that is not dominated by any of the solutions obtained by the algorithm. This metric therefore estimates the distance of the solutions from the real Pareto front as the diversity of the same. So that it gives an objective measurement that permits the results obtained to be compared with those obtained by other algorithms. Table 3 shows the best, worst and mean values obtained with the 100 runs.

Table 1: Parameters of patients considered: Type T may be real (R) or Boolean (B).

Name	**Description**	**T**
Total	Area burnt %	R
Deep	Area of deep burns	R
SAPS II	General indicator of seriousness	R
Weight	Patient's weight	R
Age	Patient's age	R
Pneumonia	Pulmonary infection	B
Sex	Patient's sex	B
Inh	Use of inhibitors	B
Infect-Wound	Surgical infection	B
AIDS-Drugs	Drug consumption and HIV	B
Hepa-Co	Previous hepatic problems	B
Bacteremia	Presence of bacteria in blood	B
Cardiac-Co	Previous cardiopathies	B
Resp-Co	Respiratory problems	B
HBP	High blood pressure	B
Diabetes	Diabetic patient	B
Renal-Co	Renal problems	B
Death	Prognosis of death	B

Table 2: Parameters of NSGA-II algorithm.

Parameters of NSGA-II algorithm	
Size of population	200
Number of populations	500
Probability of uniform crossover P_c	0.3
Probability of uniform mutation P_M	0.01

Table 3: Best, mean and worst hypervolume values.

	Best	Mean	Worst
Hypervolume (ν)	0.986	0.911	0.869

Results were also contrasted with other algorithms for attribute selection, available on the Weka data mining platform. Such as presented in Table 4, the best solution contains non-dominated solutions of 1 and 6 variables, the classification rate with 6 variables being 100%.

Table 4: Set of non-dominated solutions obtained.

Variables	M	% Correct
Pulmonary infection	1	95.4 %
Burnt area, Weight, Pulmonary infection , HBP, Diabetes, Renal Problems	6	100 %

The attributes were first filtered with the method indicated in Table 5 to carry out the evaluation with Weka. The attributes selected by Weka were used to build a model with the algorithm C4.5, at the same way that has been carried out for the attributes selected with the NSGA-II solution. Thereby, the same

conditions are defined during all the evaluation. The evaluation has been made with 75% of the cases for training, and the remaining 25% of cases for test.

Table 5: Experiments using the algorithms from Weka.

Evaluator/method	%Hit	N Var
CfsSubsetEval		
BestFirst	74.62 %	5
ExhaustiveSearch	74.62 %	5
GeneticSearch	74.62%	5
GreedyStepwise	74.62 %	4
ClassifierSubsetEval.J48		
BestFirst	80.59 %	9
ExhaustiveSearch	76.11 %	7
GeneticSearch	71.64 %	5
GreedyStepwise	74.62 %	5
ConsistencySubsetEval		
BestFirst	74.62 %	10
ExhaustiveSearch	74.62 %	9
GeneticSearch	65.67 %	10
GreedyStepwise	74.62 %	3
FilteredSubsetEval		
BestFirst	74.62 %	4
ExhaustiveSearch	74.62 %	4
GeneticSearch	74.62 %	4
GreedyStepwise	74.62 %	2

Such as presented in Table 5, the best result obtained was 80.595 with 9 variables using the selector "BestFirst", the solutions obtained with our proposed method showed a better correctness rate and lower number of variables.

5 CONCLUSIONS AND FUTURE WORK

In this paper has been proposed an optimization multiobjetive model for attribute selection, particularly those related with mortality in burns unit patients suffering infections. The attributes selected are used to build a model, which classify the patients. This optimization multiobjetive model for attribute selection is based on NSGA-II algorithm, which has been adapted to the particularities of the problem. The results obtained present a clear improvement with respect to the 16 algorithms included in the Weka data mining platform. The results point to a 100% correct classification using only 6 of the 17 variables contained in the knowledge base, while the best case obtained with the algorithms included in the Weka platform was 80.59% using 9 variables. In other words we obtain a better classification rate with fewer variables.

Future work will be focused on compare the results obtained with the NSGA-II algorithm with other multiobjetive evolutionary models PAES, SPEA and ENORA. Use other standard data bases reported in the literature to validate the proposal, and establish a decision-making system for selecting the non-dominated solution, which best satisfies the decision maker requirements.

ACKNOWLEDGEMENTS

This work has been carried out in frames of: Programa de Ayuda a los Grupos de Excelencia de la Fundacin Sneca 04552/GERM/06, and the project MEC/FEDER TIN2009-14372-C03-01. Finally, the authors would like to thank Dr. Francisco Palacios Ortega for his collaboration.

REFERENCES

Aguilera, J., Chica, M., del Jesus, M., and Herrera, F. (2007). Un estudio sobre el uso de algoritmos genéticos multimodales para selección de características. In *Congreso Español sobre Metaheurísticas, Algoritmos Evolutivos y Bioinspirados (MAEB07)*, pages 485–492.

Coello, C., Veldhuizen, D., and Lamont, G. (2002). *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic/Plenum publishers.

Curreri, P. W., Luterman, Braun, A. J., and Shires, G. T. (1980). Burn injury. analysis of survival and hospitalization time for 937 patients. *Ann Surg*, 192(4):472–478.

Deb, K. (2001). *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast elitist multiobjective genetic algorithm: Nsgaii. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11.

Jiménez, F., Sánchez, G., Juárez, J., Alcaraz, J., and Sánchez, J. (2009). Fuzzy classification of mortality by infection of severe burnt patients using multiobjective evolutionary algorithms. In *Proceedings IWINAC in Lecture Notes Series*, pages 447–456.

Kotsiantis, S. (2007). Supervised machine learning: A review of classification techniques. *Informática*, 31:249–268.

Kumar, K., Sharath, S., D'Souza, G., and Sekaran, K. (2007). Memetic nsga. a multi-objective genetic algorithm for classification of microarray data. In *Proceedings of the 15th international Conference on Advanced Computing and Communications ADCOM. IEEE Computer Society*, pages 75–80.

Pappa, G., Freitas, A., and Kaestner, C. (2002). A multiobjective genetic algorithm for attribute selection. In *Proceedings of Fourth International Conference on Recent Advances in Soft Computing (RASC)*, pages 116–121.

Society, S. (2007). Intensive-critical medicine and coronary units and spanish society of emergency.generalized infection mortality could be 20 percentage off (in spanish).

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS