VOL 4 (2020) NO 3

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Feature Selection Techniques for Selecting Proteins that Influence Mouse Down Syndrome Using Genetic Algorithms and Random Forests

Fiqhri Mulianda Putra[#,1], Fadhlal Khaliq Surado[#,2], Global Ilham Sampurno[#,3]

[#]*Department of Computer Science, Faculty of Mathematics and Natural Sciences, Bogor Agricultural University, 16680, Indonesia*
E-mail: [1]surado_fk@apps.ipb.ac.id, [2]fiqhri_mulianda@apps.ipb.ac.id, [3]global_ilham28@apps.ipb.ac.id

***Abstract*— Feature selection technique is a technique to reduce data dimensions which are widely used to find the set of features that best represent data. One area of science that often applies this technique is bioinformatics. An example of its application is the selection of significant proteins in the case of Down syndrome. To find out the most influential protein, experiments were carried out on normal mice with trisomy rats (down syndrome mice) totaling 1080 samepl and obtained 77 levels of protein expression. The analysis carried out was divided into three groups. Each group was searched for the most influential proteins using genetic algorithms with fitness calculations using random forest algorithms. The results of the protein selection of the three data groups indicate the relationship of the selected proteins to the improvement of learning ability and memory. The results of evaluating selected protein models show a high degree of accuracy, which is above 98.7% for each data group.**

***Keywords*— genetic algorithm; protein expression; random forest; feature selection; down syndrome mouse.**

I. INTRODUCTION

Feature selection is the process of selecting the best set of features that best represents the data of all the features that are in the data. Various algorithms for feature selection have been widely used in various studies. One of the advantages of feature selection according to Chandrashekar and Sahin [1] is that it helps improve the performance of prediction algorithms, both in terms of efficient use of resources and from increasing the accuracy of predictions. In addition, Fong et al. [2] mentions that choosing the relevant subset of features can provide convenience in understanding the data to be analyzed. Based on Guyon and Elisseeff [3], there are three main categories of feature selection algorithms, namely wrapper, filter, and embedded.

In its application in the real world, feature selection is often applied to the field of bioinformatics [4]. One example is the analysis of gene or protein expression levels [1]. Most features related to gene / protein expression have a high correlation with other features. That is, one of the few features with high correlation is enough to represent the whole data, because the other features do not provide additional information.

In this activity, the 77 protein expression profile data of rats taken from two types of mice, normal mice and trisomy mice, were analyzed for which proteins had a great influence. Trisomy rats themselves are mice with down syndrome (DS) which results in a lack of memory and learning ability. Kulan and Dag [5] propose the use of a wrapper model feature selection technique that uses a forward feature selection algorithm with a random forest algorithm as a determinant of the optimization of the selected features. The forward feature selection algorithm itself is a simple feature selection algorithm that starts from a set of empty features, then in each iteration one feature will be selected until the addition of features does not improve the performance of the model.

One disadvantage of forward feature selection is that any features that have been selected cannot be removed. This allows the forward feature selection algorithm not to evaluate the possibility of other feature combinations that produce more optimum results than the current feature combination. Therefore, this activity tries to apply a machine learning based approach, namely by using genetic algorithms. Genetic algorithm is one of the optimization algorithms that can be used in the problem of finding the optimal combination [6]. Based on this ability, in its application to the problem of feature selection, genetic algorithms are expected to be able to choose features, in the form of protein expression, which are significantly related to down syndrome in mice.

II. METHOD

A. Data

Data was obtained from the University of California Irvine Machine Learning Repository (UCI ML), entitled "Mice Protein Expression". The data contains information on the 77 protein expression levels obtained from the cortical nucleus fraction (part of the cerebrum) of rats. Rat samples consisted of 38 control mice and 34 Tsisdn trisomic mice. Mouse sampling was repeated 15 times for each mouse, resulting in 1080 samples. The data consists of 8 classes as listed in Table 1. Data representation can be seen in Fig. 1.

TABLE I
CLASS DESCRIPTION

Class name	Mouse type	Experiment	Treatment	total	Learning outcomes
c-CS-m	Control	*Context-Shock*	Memantine	150	Normal
c-CS-s	Control	*Context-Shock*	Saline	135	Normal
c-SC-m	Control	*Shock-Context*	Memantine	150	- *
c-SC-s	Control	*Shock-Context*	Saline	135	- *
t-CS-m	Trisomy	*Context-Shock*	Memantine	135	Success
t-CS-s	Trisomy	*Context-Shock*	Saline	105	Failed
t-SC-m	Trisomy	*Shock-Context*	Memantine	135	- ^
t-SC-s	Trisomy	*Shock-Context*	Saline	135	- ^

*) Normal conditions are not able to learn, ^) Trisomy conditions are not able to learn

There are two types of experiments performed, namely Context-Shock (CS) and Shock-Context (SC). CS experiments gave mice the opportunity to explore their environment, then the mice were given an electric shock. Meanwhile, the SC experiment directly gave the electric shock mouse, only after it was released in its environment. The results show normal CS experimental mice will not move when released in the same environment, whereas normal SC experimental mice continue to explore. This indicates that normal CS experimental mice study and remember their environmental conditions, while normal SC experimental mice do not. This is different in experimental trisomy mice who are unable to learn and remember environmental conditions, unlike normal mice. However, when injected with memantine fluid, CS trisomy experimental mouse is able to learn about the condition of its environment. Meanwhile, CS experimental trisomy mice injected with saline fluid were unable (failed) to learn about their environmental conditions.

Mouse	**Q1**	**P2**	**...**	**Q77**	**Class**
Mouse	0.504	0747	...	1,676	c-CS-m
Mouse2	0.515	0.689	...	1,755	c-CS-s
...	...	...	...	...	...
Mouse1080	0.509	0730	...	1,926	t-SC-s

Fig. 1 Data representation

B. Stages

The stages carried out in this activity can be seen in Fig 2. There are 4 main stages, including (1) preprocessing data, (2) data sharing, (3) feature selection process, and (4) evaluating feature selection results.

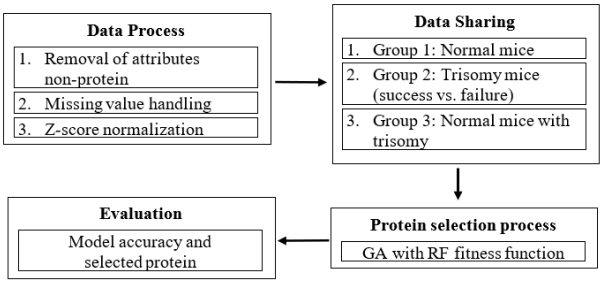

Fig. 2 Stages

C. Data Process

Preprocessing data follows Kulan and Dag [5] which consists of 3 things, (1) eliminating non-protein attributes, (2) checking and handling attributes that have missing values, (3) normalizing numeric attribute values. Non-protein attributes other than class attributes that do not represent the level of protein expression need to be removed, such as mouseID, genotype, treatment, and behavior. Attributes that have missing value are filled with the average value of the attribute corresponding to the class. Finally, numerical attributes will be normalized using Z-score normalization (equation 1).

$$Z = \frac{x - \mu}{\sigma} \quad (1)$$

D. Data Sharing

Data sharing follows Kulan and Dag [5] which divides into 3 data groups, among which are as follows.

1. Normal mouse group
   In this data group we want to obtain important proteins that influence the success / failure of the learning process in normal mice. The data classes used are c-CS-m, c-CS-s, c-SC-m, and c-SC-s.
2. Trisomy mouse group (success vs. failure)
   In this data group we want to obtain important proteins that influence the success of the trisomy rat learning process given memantine treatment. The data classes used are t-CS-m and t-CS-s.
3. Normal mouse group with trisomy mice
   In this data group we want to obtain important proteins that influence the failure of trisomy mice in studying their environment and compared to normal mice who have successfully studied their environment. The data classes used are t-CS-s, c-CS-m, and c-CS-s.

E. Protein Selection Process

The protein selection process is carried out on the 3 data groups mentioned earlier. This process applies a genetic algorithm to select important subset of features (proteins) based on the fitness value of each of the 3 data groups.

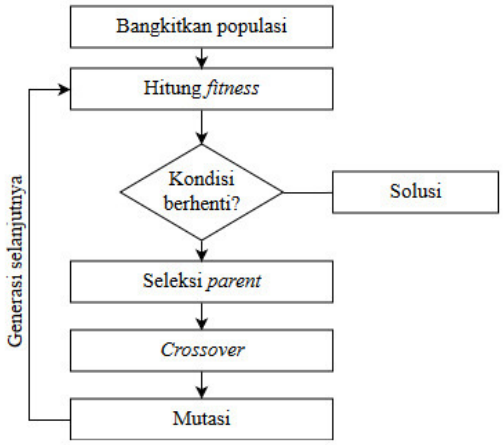

Fig. 2 Genetic Algorithm

Genetic algorithms are heuristic optimum value search algorithms that mimic the process of natural selection [7]. Algorithms begin by generating a set of populations consisting of several individuals (chromosomes). Each individual is represented by a binary number. This algorithm has 3 types of operators inspired by biological processes, such as selection, crossover, and mutation.

General steps contained in genetic algorithms as in Fig. 2. In its application to feature selection, individuals (chromosomes) represent whether or not a feature is selected. Fitness value is obtained from the measurement of model performance, namely the level of accuracy. Calculation of fitness values in each generation (iteration) of genetic algorithms is done by applying the random forest algorithm as a classifiermodel by calculating the accuracy of the selected feature models. The highest fitness value obtained from all generations is the solution of the feature selection process.

The parameters used in the genetic algorithm and random forest are as follows (Fig. 3).

	GA parameter
Max generation	100
Population size	20, 40
Crossover Opportunity	0.8
Mutation Opportunities	0.1

	RF parameter
ntree	500
Cross validation	5

Fig. 3 GA and RF parameters

F. Evaluation

The results of the selected proteins will be evaluated based on their accuracy. The selected protein models will be compared with the selection results compared to the Kulan and Dag experiments [5] which use the forward feature selection algorithm with the genetic algorithm as the feature selection algorithm and the random forest algorithm as the model classifier.

III. RESULTS AND DISCUSSION

A. The Results of Group 1 Protein Data Selection (Normal Rat)

The results of the selected proteins in group 1 data using the method proposed by the Kulan and Dag method [5] can be seen in Table II and Table III. The results of group 1 protein selection showed that proteins had a significant influence on the ability of normal mice given CS experiments (c-CS-m and c-CS-s) in studying their environmental conditions compared to normal SC experiment mice (c-SC-m and c-SC-s) who do not study environmental conditions.

TABLE II
GROUP 1 SELECTED PROTEIN

Population size	Selected protein	Model accuracy
20	DYRK1A, NR1, pCAMKII, pMEK, ELK, pPKCG, S6, RRP1	0.981
40	pCAMKII, TRKA, pPKCG, PSD95, SHH, H3AcK18	**0.987**

TABLE III
PROTEINS FROM GROUP 1 SELECTION [5]

Selected protein	Model accuracy
SOD1, Ubiquitin, pGSK3B, S6, CaNA, IL1B, BAX, pNR2A, BDNF, pJNK, pCFOS	0.963

Based on the results of experiments conducted, the proposed method is able to obtain the highest accuracy in experiments with a population size of 40. Based on the study Ahmed et al. [8], protein pCAMKII and pPKCG are closely related to normal mice in studying environmental conditions, especially when given memantine treatment. Meanwhile, PSD95 protein is related to NMDAR protein which influences memory and learning ability [9].

.

B. Results of Group 2 Protein Data Selection (Trisomy Rats)

The results of the selected proteins in group 2 data using the method proposed by the Kulan and Dag method [5] can be seen in Table IV and Table V. The results of group 2 protein selection showed that proteins had a significant influence on the success of trisomy rats given memantine treatment (t-CS-m) in studying their environmental conditions compared to trisomy mice given saline treatment (t-CS-s) which were not successful.

TABLE IV
PROTEINS FROM GROUP SELECTION 2

Population size	Selected protein	Model accuracy
20	DYRK1A, pCREB, pPKCAB, JNK, P38, pNUMB, ADARB1, ERBB4, CaNA	**0.987**
40	pCAMKII, pRSK, GSK3B, NR2B, TIAM1, P70S6, pGSK3B_Tyr216, BCL2, pS6, SYP	**0.987**

TABLE V
GROUP 2 SELECTED PROTEINS [5]

Selected protein	Model accuracy
BRAF, S6, CDK5, BDNF, pCREB, PKCA, SOD1, PSD95, pNR2A	0.946

Based on the results of the experiments conducted, the proposed method is able to obtain the highest accuracy in the trial population size of 20 and 40. Based on Ahmed et al. [8]

and Ahmed et al. [10], the expression of NR2B, TIAM1, GSK3B, and pS6 proteins increased when given memantine treatment, both of which could affect memory and learning processes. Based on Czabotar et al. [11] and Harada et al. [12], protein BCL2 and P70S6 have a role in the immune system.

C. Results of Group 3 Protein Data Selection (Normal Mice with Trisomy Mice)

The results of the selected proteins in group 3 data using the method proposed by the Kulan and Dag method [5] can be seen in Table VI and Table VIITABLE VII . The results of group 3 protein selection showed that proteins had a significant influence on the failure of trisomy rats in CS experiments treated with saline (t-CS-s) compared to those of normal CS experimental rats (c-CS-s and c-CS-m).

TABLE VI
PROTEIN FROM GROUP SELECTION 3

Population size	Selected protein	Model accuracy
20	DYRK1A, pNR2A, BRAF, TRKA, P38, NUMB, ADARB1, GluR3, GluR4, IL1B, pCFOS	0.969
40	BDNF, pCAMKII, pNR2B, APP, Tau, pCASP9, H3AcK18, H3MeK4	**0.987**

TABLE VII
GROUP 3 SELECTED PROTEINS [5]

Selected protein	Model accuracy
SOD1, Ubiquitin, pGSK3B, S6, CaNA, IL1B, BAX, pNR2A, BDNF, pJNK, pCFOS	0.892

Based on the results of the experiments conducted, the proposed method is able to obtain the highest accuracy in the trial population size of 40. Based on Ahmed et al. [10], BDNF protein, Tau, H3AcK18, and H3MeK4have a connection to the failure of the learning process and protein APP is a family of Hsa21 which if over-expressed can lead to failure of the learning process. Meanwhile, protein pCAMKII and pCASP9 are related to the normal learning process [10], [13].

IV. CONCLUSION

Based on the results obtained, the proposed method that uses genetic algorithm as a feature selection algorithm with a random forest algorithm as a classifier model is able to provide better results compared to the Kulan and Dag [5] methods that use the forward feature selection algorithm. Good results are shown with the accuracy of the selected protein models using the proposed method is higher. The selected proteins through a literature study have links to cases from each of the three data sharing groups.

REFERENCES

[1] Chandrashekar G, Sahin F. 2014. A survey on feature selection methods. Computers & Electrical Engineering. 40 (1): 16-28. doi: 10.1016 / j.compeleceng.2013.11.024.

[2] Fong S, Zhuang Y, Tang R, Yang XS, Deb S. 2013. Selecting optimal feature sets in high-dimensional data by swarm search. Journal of Applied Mathematics. 2013: 1-18. doi: 10.1155 / 2013/590614.

[3] Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of machine learning research, 3 (Mar), pp.1157-1182.

[4] Li J, Liu H. 2017. Challenges of feature selection for big data analytics. IEEE Intelligent Systems. 32 (2): 9-15. doi: 10.1109 / e.g.2017.38.

[5] Kulan H, Dag T. 2019. In silico identification of critical proteins associated with learning processes and immune systems for Down syndrome. PLoS ONE 14 (1): e0210954.https: // doi. org / 10.1371 / journal.pone.0210954

[6] Rajappa GP. 2012. Solving Combinatorial Optimization Problems Using Genetic Algorithms and Ant Colony Optimization [dissertation]. Knoxville (US): University of Tennessee.

[7] Mitchell M. 1996. An Introduction to Genetic Algorithms. Cambridge, MA: MIT Press. ISBN 9780585030944.

[8] Ahmed MM, Dhanasekaran AR, Block A, Tong S, Costa ACS, Gardiner KJ. Protein Profiles Associated With Context Fear Conditioning and Their Modulation by Memantine. Molecular Cellular Proteomics: MCP. 2014; 13 (4): 919–937. https://doi.org/10.1074/mcp.M113.035568 PMID: 24469516

[9] Newcomer JW, Farber NB, Olney JW. 2000. NMDA receptor function, memory, and brain aging. Dialogues in clinical neuroscience, 2 (3), 219-232.

[10] Ahmed MM, Dhanasekaran AR, Block A, Tong S, Costa ACS, Stasko M, et al. Protein dynamics associated with failed and rescued learning in the Ts65Dn mouse model of Down syndrome. In Cunto F, ed. PLoS ONE. 2015; 10 (3): e0119491. https://doi.org/10.1371/journal.pone.0119491 PMID: 25793384

[11] Czabotar PE, Lessene G, Strasser A, Adams JM. 2013. Control of apoptosis by the BCL-2 protein family: implications for physiology and therapy. Nature Reviews Molecular Cell Biology. 15 (1): 49-63. doi: 10.1038 / nrm3722.

[12] Harada H, Andersen JS, Mann M, Terada N, Korsmeyer SJ. 2001. P70s6 kinase signals cell survival as well as growth, inactivating the pro-apoptotic molecule BAD. Proceedings of the National Academy of Sciences. 98 (17): 9666-9670. doi: 10.1073 / pnas.171301998.

[13] Higuera C, Gardiner KJ, Cios KJ. Self-Organizing Feature Maps Identify Proteins Critical to Learning in a Mouse Model of Down Syndrome. PLoS ONE. 2015; 10 (6): e0129126. https://doi.org/10.1371/journal. pone.0129126 PMID: 26111164.