5. Optimization of C4.5 Algorithm-Based Particle Swarm Optimization for Breast Cancer Diagnosis

by Much Aziz Muslim

Submission date: 23-Jul-2019 01:21PM (UTC+0700)
Submission ID: 1154276213
File name: ased_Particle_Swarm_Optimization_for_Breast_Cancer_Diagnosis.pdf (678.12K)
Word count: 2896
Character count: 15371

Journal of Physics: Conference Series

Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis

Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis

M A Muslim[1*], S H Rukmana[1], E Sugiharti[1], B Prasetiyo[1] and S Alimah[2]

[1]Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia
[2]Department of Biology, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Indonesia

*Corresponding author: a212muslim@yahoo.com

Abstract. Data mining has become a basic methodology for computational applications in the field of medical domains. Data mining can be applied in the health field such as for diagnosis of breast cancer, heart disease, diabetes and others. Breast cancer is most common in women, with more than one million cases and nearly 600,000 deaths occurring worldwide each year. The most effective way to reduce breast cancer deaths was by early diagnosis. This study aims to determine the level of breast cancer diagnosis. This research data uses Wisconsin Breast Cancer dataset (WBC) from UCI machine learning. The method used in this research is the algorithm C4.5 and Particle Swarm Optimization (PSO) as a feature option and to optimize the algorithm. C4.5. Ten-fold cross-validation is used as a validation method and a confusion matrix. The result of this research is C4.5 algorithm. The particle swarm optimization C4.5 algorithm has increased by 0.88%.

1. Introduction

Data mining has become a basic methodology for computational applications in the field of medical domains. Data mining can be applied in the field of health such as diagnosing breast cancer, heart disease, diabetes and others [1]. Data mining has various techniques such as estimation, classification, association, and clustering. Among the various algorithms, classification algorithm plays an important role in predictive analysis. Classification aims to divide the object assigned only to one of the categories called class [2].

Utilization of data mining can be done in various fields, for example for Clustering Student Scholarship Applicants [3], Optimization of Classification of Student Final Project [4]. In the field of health such as for Prediction of Pregnancy Hypertension with Decision Tree Technique [5], Identification of Tuberculosis (Tb) Disease in Humans using Naïve Bayesian Method [6].

One of the most powerful and widely used techniques for classification and prediction is decision tree [7]. Decision tree is a frequently used classification algorithm and has a simple structure as well as easy to be interpreted [8]. Decision Tree transforms a very large fact into a decision tree presenting the rules [9]. The C4.5 algorithm proves its performance in predicting with best results in terms of accuracy and minimum execution time [10]. Many researchers have tried to apply the machine learning algorithm to diagnose breast cancer.

Breast cancer is the most common cancer happens to women in both developed and developing countries. Breast cancer is a disease in which there is an excessive growth or uncontrolled development of breast tissue cells. Breast cancer is considered the most common invasive cancer in women, with more than one million cases and nearly 600,000 deaths occurring around the world each year [12]. The most effective way to reduce deaths from breast cancer is by early diagnosis [13].

The C4.5 algorithm has weaknesses in handling large data, including: (1) empty branch, nodes with zero value or near zero value do not contribute to generate rules or help to build classes for classification tasks but make bigger and more complex tree sizes, (2) insignificant branch, insignificant branch not only reduce the usefulness of the decision tree but also bring overfitting problems, (3) Overfitting occurs when the algorithm model takes data with unusual characteristics (noise) [5].

Data quality such as noise and overfitting data can affect the performance of classification algorithms. Feature selection is commonly used in machine learning when it involves attributes of high-dimensional and noise datasets. Feature Selection is the process of selecting relevant features, or a subset of feature candidates [13]. Feature selection search locally. Metaheuristic optimization can find solutions in full search space and use global search capabilities that significantly improve the ability to find high-quality solutions within a reasonable timeframe [14]. Improved algorithmic accuracy is required, for example through the application of Discretization and Bagging Techniques to Improve Classification Accuracy in Algorithm C4.5 [15].

One of metaheuristic optimization for feature selection is Particle Swarm Optimization (PSO). PSO has proven to be more competitive than genetic algorithms in some cases, especially in the area of optimization [16]. In this study, a combination of PSO-based C4.5 algorithms is proposed to improve the accuracy of breast cancer diagnoses and to overcome weaknesses in the C4.5 algorithm using PSO metaheuristic optimization for feature selection and to optimize C4.5 algorithm accuracy. Based on the description above, it is necessary to improve the method of diagnosing breast cancer accurately.

2. Methods

In this research would be conducted analysis of comparation and fusion of two classification methods of data mining. The method used was the C4.5 algorithm and particle swarm optimization. The first step in this research was to measure the accuracy of C4.5 algorithm. The next step was to measure the accuracy of C4.5 algorithm based on particle swarm optimization. Particle swarm optimization as feature selection and to optimize the accuracy of C4.5 algorithm, then compare which algorithm gives better accuracy. At this stages conducted the steps of the method used. Flowchart of C4.5 algorithm optimized using particle swarm optimization was shown in Figure 1.

At preprocessing stage was done initial processing of data. In the data of Wisconsin breast cancer, there were 699 records consisting of 11 attributes with 10 attributes of numerical type and 1 categorical type. In this research was done pre-processing in accordance with KDD process that was data cleaning, data selection, and data transformation.

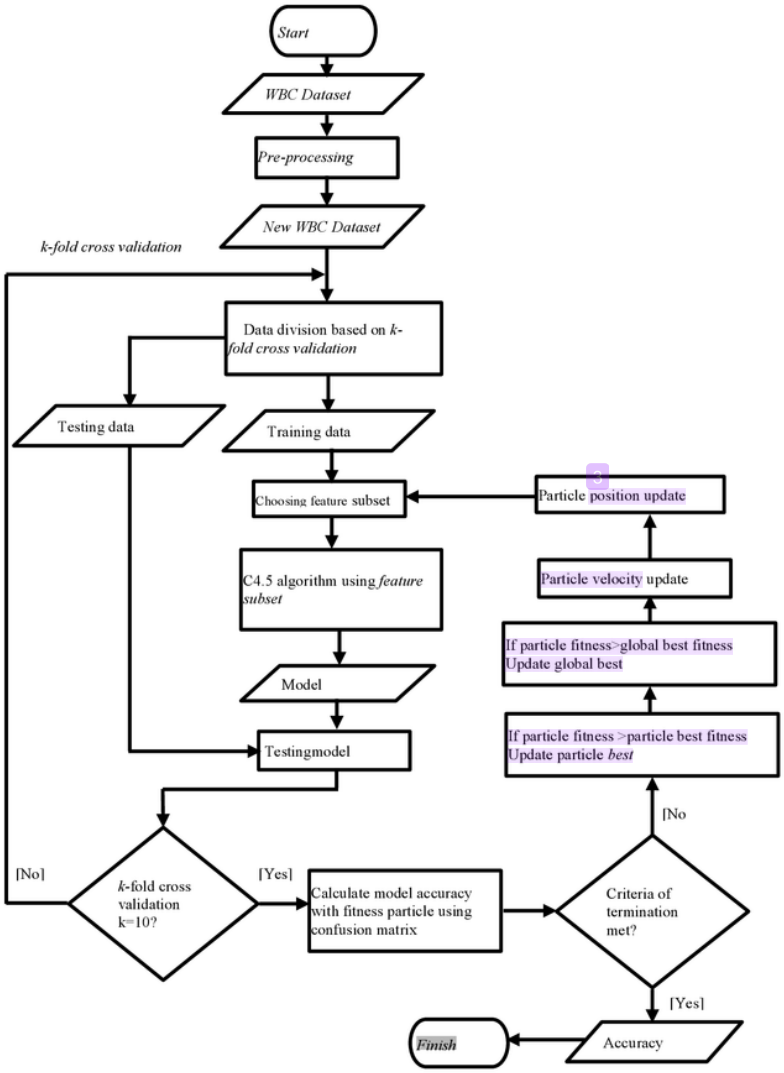

Figure 1. Flowchart of C4.5 Algorithm Based on Particle Swarm Optimization

a. Data cleaning

At this stage was done cleaning on incomplete, empty, or null data, data containing noise, and inconsistent data. There were 16 missing value data on bare nuclei attribute. There were several ways of missing value handling, among others ignoring tuples, filling missing value manually, using global constants to fill missing value, using measures of central tendency for attributes (eg, mean or median), using mean or median attributes for all samples included in the class which was the same as the tuple given, and using the value that was most likely to be filled in the lost value [16]. Handling of missing value using average in this study reduced the level of accuracy. Therefore, the handling of missing value in this study was done by reducing the data object so that the amount of wisconsin breast cancer dataset which was originally 699 records became 683 records. The detail of data to be cleaned was shown in Table 1.

Table 1. Data cleaning

Data type	Number of breast cancer data
Initial data	699
Incomplete data	16
Number of clean data	683

b. Data selection

At this stage data selection would be done to reduce irrelevant and redundant data. In dataset of wisconsin breast cancer was done the process of elimination on the attribute of sample code number due to the attribute included into nominal or ordinal feature that was categorical types and qualitative value. This value was actually a symbolic value, it was impossible to perform arithmetical operations as in numerical type so that only 10 attributes were used with 9 attributes as predictor variables and 1 attribute as destination / target variable. The attribute details were shown in Table 2.

Table 2. Research Attributes

No	Name of Attributes	Information	Values
1.	clump thickness	This attribute determined whether the cell was laminated or not because benign cells tended to have only one layer (monolayer) whereas malignant cells tended to have multiple layers (multilayer).	1-10
2.	uniformity of cell size	This attribute determined the consistency of cell size.	1-10
3.	uniformity of cell shape	This attribute determined the similarity of cell shape.	1-10
4.	marginal adhesion	This attribute determined whether cells were together or not because malignant cells tended to lose this ability.	1-10
5.	single epithelial cell size	This attribute determined whether the epithelial cell tended to enlarge or not.	1-10
6.	bare nuclei	This attribute determined whether the cell was surrounded by cytoplasm (the rest of the cell) or not.	1-10
7.	bland chromatin	This attribute determined the texture level of the chromatin cell.	1-10
8.	normal nucleoli	This attribute determined the shape of nucleoli.	1-10
9.	Mitoses	This attribute determined how many cancer cells divided, splited or multiplied.	1-10
10.	Class	This attributes determined whether the tumour was benign or malignant.	2 and 4

c. Data transformation

At this stage would be conducted transformation data. The data of class value had formats 2 and 4, this format was changed namely 2 into benign and 4 for malignant.

After the pre-processing stage was completed, then the data was divided based on tenfold cross validation. Tenfold cross-validation divided data into 10 sets, the size of data set divided by 10 then 9 sets of data for training and 1 set of data for testing then the step was repeated up to 10 times iteration. Training data was used to build the model while testing data was used to validate the model.

Later, data training was used for the modelling of C4.5 algorithm based particle swarm optimization. Particle swarm optimization gave weight to each attribute and produced the best solution (fitness) then done the calculation of C4.5 algorithm. The steps to generate fitness were as follows.

1. Calculate the best solution of particle i on iteration t.
$$p_i^t = \{p_{i1}^t, p_{i2}^t, \dots, p_{iD}^t\} \quad (1)$$
2. Calculate the best solution of p_i^t in the population on iteration t.
$$p_g^t = \{p_{g1}^t, p_{g2}^t, \dots, p_{gD}^t\} \quad (2)$$
3. Calculate particle velocity.
$$v_{id}^t = w * v_{id}^{t-1} + c_1 r_1 \left(p_{id}^t - x_{id}^t\right) + c_2 r_2 \left(p_{gd}^t - x_{id}^t\right), \text{d= 1,2, ...D} \quad (3)$$
4. Calculate the new position.
$$x_{id}^{t+1} = x_{id}^t + v_{id}^t, \text{d} = 1,2,\dots \text{D} \quad (4)$$

The basic process of the PSO algorithm was given as follows.

1. Initialization: randomly generated initial particles.
2. Fitness: fitness size of each particle in the population.
3. Update: calculated the velocity of each particle with equation (3).
4. Construction: for each particle, moved to the next position according to equation (4).
5. Termination: stopped the algorithm if the termination criterion was met, and returned to step 2 (fitness) was declared.

Iteration was stopped if the number of iteration reached the maximum number of predefined iterations.

Thereafter was doing the modelling of C4.5 algorithm with attribute that has been given weight with the following steps [16].

a. Split attribute

The attribute that has the best value was selected as the split attribute for the given tuple.

1) Calculate info(D) or called also entropy.
$$Info\ (D) = -\sum_{i=1}^{m} pi * log_2(pi) \quad (5)$$
2) Calculate $info_A$(D) or also called information gain.
$$Info_A D = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * info\ (D_j) \quad (6)$$
3) Calculate gain *(A)*.
$$Gain\ (A) = Info\ (D) - Info_A D \quad (7)$$
4) Calculate split info.
$$SplitInfo_A D = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} * log_2 \frac{|D_j|}{|D|} \quad (8)$$
5) Calculate gain ratio.
$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A D} \quad (9)$$
6) Attribute with a maximum gain ratio was selected as a split attribute (roots).

b. Repeat the process for each branch until all the cases on the branch had the same class.

c. The recursive partition stopped if it met one of the following termination conditions.

1) All tuples in D partition (represented on N node) had the same class.
2) There was no left attribute where tuples could be further partitioned.
3) There was tuple for a particular branch, ie the D partition was empty.

3. Result and discussion

The result of this research aimed to compare C4.5 algorithm with C4.5 algorithm optimized in feature selection with particle swarm optimization. The modelling used C4.5 algorithm would produce a model of decision tree. This decision tree would then go through the stage of accuracy testing using confusion matrix (Figure 2).

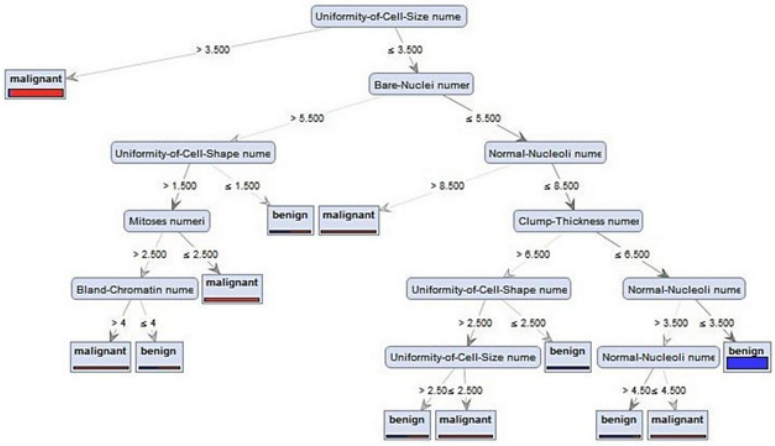

Figure 2. Model of C4.5 algorithm decision tree

Based on the modelling, a method evaluation was performed using confusion matrix which yielded accuracy of 95.61% (Table 3).

Table 3. Confusion matrix of C4.5 algorithm

	true benign	true malignant
pred. benign	426	12
pred. malignant	18	227

The accuracy of the C4.5 algorithm based particle swarm optimization was evaluated using confusion matrix that produced accuracy of 96.49% (Table 4).

Table 4. Confusion matrix of C4.5 algorithm based particle swarm optimization

	true benign	true malignant
pred. Benign	429	9
pred. malignant	15	230

4. Conclusion

The new method which integrates the C4.5 algorithm and particle swarm optimization algorithms in this study proved can improve the accuracy of breast cancer diagnosis. Particle swarm optimization is applied as feature selection and to optimize the accuracy of C4.5 algorithm. Based on the results of the research shows the accuracy of C4.5 classification algorithm equal to 95.61%, while for accuracy of C4.5 based on particle swarm optimization equal to 96.49% so that it can increase accuracy equal to 0.88%. Based on the research, it can be

concluded that C4.5 algorithm based particle swarm optimization can improve the accuracy of C4.5 algorithm.

References

[1] Daniel L T 2005 *Discovering Knowledge in Data: An Introduction to Data Mining* (New Jersey: John Wiley & Sons, Inc.)
[2] Bramer M 2007 *Principles of Data mining* (London: Springer)
[3] Defiyanti S, Jajuli M and Rohmawati N 2017 *Sci. J. Inform.* **4** 27
[4] Somantri O, Wiyono S and Dairoh D 2016 *Sci. J. Inform.* **3** 34
[5] Muzakir A and Wulandari R A 2016 *Sci. J. Inform.* **3** 19
[6] Trihartati S A and Adi C 2016 *Sci. J. Inform.* **3** 99
[7] Perveen S 2016 *Procedia Comp. Sci.* **82** 115
[8] Mantas C J and Abellán J 2014 *Expert Syst. W. App.* **41** 4625
[9] Boukenze B, Haqiq A and Mousannif H 2016 *IJDMS* **8** 1
[10] Salama G I, Abdelhalim M and Zeid M A E 2012 *Breast Cancer (WDBC)* **32** 2
[11] Gupta S, Kumar D and Sharma A 2011 *IJCSE* **2** 188
[12] Mazid M M, Ali S and Tickle K S 2010 *Proceedings of the 9th WSEAS international conference on Artificial intelligence, knowledge engineering and data bases (pp. 296-301) World Scientific and Engineering Academy and Society (WSEAS).*
[13] Wijaya K P and Muslim M A 2016 *Prosiding Seminar Nasional Ilmu Komputer* (pp. 22-27) Semarang
[14] Wahono R S and Suryana N 2013 *IJSEIA* **7** 153
[15] Muslim M A, Sugiharti E, Prasetiyo B and Alimah S 2017 s *Lontar Komputer: Jurnal Ilmiah Teknologi Informasi* **8** 135
[16] Sousa T, Silva A and Neves A 2004 *Parallel Comput.* **30** 767
[17] Jiawei H, Micheline K and Jian P 2012 *Data Mining: Concepts and Techniques 3rd Edition* (Elsevier)

5. Optimization of C4.5 Algorithm-Based Particle Swarm Optimization for Breast Cancer Diagnosis

ORIGINALITY REPORT

26%	%	26%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	M A Muslim, A J Herowati, E Sugiharti, B Prasetiyo. "Application of the pessimistic pruning to increase the accuracy of C4.5 algorithm in diagnosing chronic kidney disease", Journal of Physics: Conference Series, 2018 Publication	7%
2	Dwi Meylitasari Br. Tarigan, Dian Palupi Rini, Sukemi. "Particle Swarm Optimization – Based on Decision Tree of C4.5 Algorithm for Upper Respiratory Tract Infections (URTI) Prediction", Journal of Physics: Conference Series, 2019 Publication	5%
3	Wahono, Romi Satria, Nanna Suryana, and Sabrina Ahmad. "Metaheuristic Optimization based Feature Selection for Software Defect Prediction", Journal of Software, 2014. Publication	2%
4	D Alighiri, W T Eden, K I Supardi, Masturi, A. Purwinarko. "Potential Development Essential Oil Production of Central Java, Indonesia",	2%

Journal of Physics: Conference Series, 2017
Publication

---

5 S. Dick, A. Tappenden, Curtis Badke, O. Olarewaju. "A Novel Granular Neural Network Architecture", NAFIPS 2007 - 2007 Annual Meeting of the North American Fuzzy Information Processing Society, 2007 — 1%
Publication

---

6 Sampurna Mandal, Supratim Bhattacharya, Jayanta Poray. "Towards a decision support system by the study of cell malfunctions for breast cancer", 2016 International Conference on Computer, Electrical & Communication Engineering (ICCECE), 2016 — 1%
Publication

---

7 M. Nadjib Bustan, M. Arif Tiro, Adiatma. "Modeling of Breast Cancer Diagnosis Classification Based on Hospital Medical Records", Journal of Physics: Conference Series, 2018 — 1%
Publication

---

8 S Prabakaran, Shilpa Mitra. "Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning", Journal of Physics: Conference Series, 2018 — 1%
Publication

---

9 E V Kotelnikov, V R Milov. "Comparison of rule

induction, decision trees and formal concept analysis approaches for classification", Journal of Physics: Conference Series, 2018 — 1%
Publication

10 Sri Wahyuni. "Implementation of Data Mining to Analyze Drug Cases Using C4.5 Decision Tree", Journal of Physics: Conference Series, 2018 — 1%
Publication

11 Much Aziz Muslim, Aldi Nurzahputra, Budi Prasetiyo. "Improving accuracy of C4.5 algorithm using split feature reduction model and bagging ensemble for credit card risk prediction", 2018 International Conference on Information and Communications Technology (ICOIACT), 2018 — 1%
Publication

12 Katarína Močarníková, Michal Greguš. "Chapter 8 Conceptualization of Predictive Analytics by Literature Review", Springer Science and Business Media LLC, 2020 — 1%
Publication

13 Wasiur Rhmann. "Cross project defect prediction using hybrid search based algorithms", International Journal of Information Technology, 2018 — <1%
Publication

14 G. Serpen. "A novel algorithm for classification

	of SPECT images of a human heart", Proceedings Ninth IEEE Symposium on Computer-Based Medical Systems CBMS-96, 1996 Publication	<1%
15	Subham Sadhukhan, Nityasree Upadhyay, Prerana Chakraborty. "Chapter 12 Breast Cancer Diagnosis Using Image Processing and Machine Learning", Springer Science and Business Media LLC, 2020 Publication	<1%
16	John A. Newby, C. Vyvyan Howard. "Environmental influences in cancer aetiology", Journal of Nutritional & Environmental Medicine, 2009 Publication	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		