Optimization Of Feature Weight The Voting Feature Intervals 5 Algorithm Using Partical Swarm Optimization Algorithm

Eka Hayana Hasibuan1, Herman Mawengkang2, and Syahril Efendi3

1Faculty of Computer Science and Informatics Technology, University Of Sumatera Utara, Indonesia,
2Faculty of Mathematic Science, University of Sumatera Utara, Indonesia
3Faculty of Computer Science and Informatics Technology, University Of Sumatera Utara, Indonesia
eka_hayana@yahoo.com, hayana_eka@students.usu.ac.id

Abstract. The use of Partical Swarm Optimization Algorithm in this research is to optimize the feature weights on the Voting Feature Interval 5 algorithm so that we can find the model of using PSO algorithm with VFI 5. Optimization of feature weight on Diabetes or Dyspepsia data is considered important because it is very closely related to the livelihood of many people, so if there is any inaccuracy in determining the most dominant feature weight in the data will cause death. Increased accuracy by using PSO Algorithm ie fold 1 from 92.31% to 96.15% increase accuracy of 3.8%, accuracy of fold 2 on Algorithm VFI5 of 92.52% as well as generated on PSO Algorithm means accuracy fixed, then in fold 3 increase accuracy of 85.19% Increased to 96.29% Accuracy increased by 11%. The total accuracy of all three trials increased by 14%. In general the Partical Swarm Optimization algorithm has succeeded in increasing the accuracy to several fold, therefore it can be concluded the PSO algorithm is well used in optimizing the VFI5 Classification Algorithm. Keywords: Partical Swarm Optimization, Voting Feature Intervals, fold, Accuracy

1. Introduction

Research conducted [1] using the classification algorithm Voting Feature Intervals 5 (VFI5) the average accuracy of Pulmonary TB data model is 83%. The feature weight used in the study is one for each feature. The feature weight used in the study is one for each feature. Accuracy of the study is still predicting many wrong data. In a previous study [2] using a genetic algorithm to find the weight of the K-NN feature and the result was applied to the VFI5 classification algorithm, and the accuracy obtained from 95% to 99%, therefore in this study the accuracy of the previous study [1] Can increase. Pre-existing research [3] using the Voting Feature Intervals 5 (VFI5) Classification algorithm on feature data of DM Disease (Diabetes Mellitus) versus Dyspepsia (Dys). This research uses 3-fold cross validation that is divided into first iteration, second iteration and third iteration, each with combination of different training and test data. The first iteration yielded an accuracy of 92.31% with two predictive errors in the 41st DM patient and 80th dyspepsia patient, a second iteration accuracy of 92.59% with 2 predictive errors in the 21st DM patient and the dyspeptic patient to- 67 and a third iteration accuracy of 85.19% with 4 predictive errors in 14th DM patients, 51st dyspepsia patients, 52nd dyspepsia patients, 57th dyspepsiapatients. Theresults of the feature hose on the training process and the final normalization in this study are similar. Subsequent research [4] is Optimization of Voting Feature Intervals 5 Algorithm with Genetic Algorithm on Tuberculosis of Lung Data results in
improved accuracy by using genetic algorithm ie in fold 1 of 76% -95% accuracy increase of 19%. Improved accuracy in fold 2 from 84% -95% increase in accuracy 11%. Improved accuracy in fold 3 from 89% -95% increase in accuracy of 6%. In general, genetic algorithms have improved accuracy for each fold. Therefore it can be concluded that a good genetic algorithm is used in optimizing the VFI5 Classification Algorithm.

Particle Swarm Optimization (PSO) is a frequently applied method for solving optimization problems in nonlinear cases. PSO produces more optimal weight as an alternative Back Propagation (BP) in the ANN training [5]. PSO provides optimal results in clusters center as an alternative to KMeans on the problem of pattern classification or pattern classification [6]. The combination of PSO with Support Vector Machine (SVM) on project scheduling optimization results in more performance [7]. The combination of PSO with Statistical Clustering (SC) on the feature classification problem shows that both algorithms can select features with much smaller numbers and achieve better classification performance [8]. The PSO algorithm is also used to optimize the Power of Communication Network with excellent results because the PSO combines three advantages of having good search performance, continuous variable control and discrete [9].

The formulation of the problem in this research is to find the accuracy of the optimal feature weight. The purpose of this research is to get the most optimum feature weight accuracy on VFI5 classification algorithm using PSO algorithm so as to find PSO making model with VFI5. Limitation of the problem in this research is weight data feature used is data of DM symptom symptom that is dizziness, nausea, vomiting, weakness, fever, heartburn, shortness of breath, chest pain, diarrhea, cough [3].

2. Voting Feature Intervals 5

The VFI5 classification algorithm represents a concept that describes the concept of interfacial hose. This algorithm was developed by GülşenDemiröz and Halil Altay Güvenir in 1997. It has been argued that the VFI5 algorithm is included in a supervised algorithm, ie an algorithm that has a target of data classes and is non-incremental meaning that all training data are processed simultaneously [10]. VFI algorithm continues to grow until the final version of the VFI5 algorithm. The process of classifying new instances is based on the feature vote. Each of these features gives its value to the votes among the classes. The class that receives the highest vote value is then specified as the predicted class. The advantage of the VFI5 algorithm is that the algorithm is robust against the irrelevant features but is able to deliver good results on the existing real world dataset. VFI5 is also capable of eliminating the unfavorable effects of features that are irrelevant to its voting mechanism [2].

2.1 Training

This training process aims to find the model that will be used for the classification process. In this training process will be generated hose on each feature. A hose represents the set of values of a given feature. The end point on the hose should be known to produce a specific feature hose. The process of finding different end points for linear features and nominal features. For linear features the values have a sequence and are comparable in rank. To determine the end point of this feature by finding the maximum and minimum values for that feature for each class.

2.2 Prediction

If ef is known then the interval can be found. The interval can store training instances of some classes. The classes in an interval are represented by the votes of the 2 classes at that interval. For each class c,
feature f gives the same vote with interval_class_vote [f, I, c]. The notation represents the vote feature f given for class c. Each feature f gathers its votes in a vector <feature_vote [f, C1], ..., feature_vote [f, Cj], ..., feature_vote [f, Ck]>, where feature_vote [f, Cj] is a vote feature f for class Cj and k is the number of classes. Then d vector vote, where d is the number of features, summed to get the total vector vote <vote [C1], ..., vote [Ck]>.

The class with the biggest vote is predicted as the class from the test instance e. Pseudocode prediction algorithm VFI5 can be seen in figure 1.

### 3. Partial Swarm Optimization

Algorithm Particle Swarm Optimization (PSO) is widely used to solve optimization problems as well as feature selection issues [11]. Optimization is the process of adjusting to input or device characteristics, mathematical processes, or experiments to find the minimum or maximum output of results. The input consists of a variable, a process or function known as a cost function, an objective function or a function and an output capability is a cost or a goal, if the process is an experiment, then a variable is the physical input for the experiment [12]. PSO is a population-based method like GA, but its basic concept is cooperation, not competition [13] [14]. A mathematical formula that describes the position and velocity of the particles on a certain dimension of space:

\[
X_i(t) = x_i1(t), x_i2(t), ..., x_iN(t) \quad ..........(1)
\]

\[
V_i(t) = v_i1(t), v_i2(t), ..., v_iN(t) \quad ..........(2)
\]

Where : X = particle position V = particle velocity I = particle index T = the iteration of the t N = size of the space dimension Here is a mathematical model that describes the mechanism Updating the particle status of Kennedy and Eberhart 1995:

\[
V_i(t) = V_i(t-1) + c1r1(XiL - X_i(t-1)) + c2r2(XG - X_i(t-1)) \quad ..........(3)
\]

\[
X_i(t) = V_i(t) + X_i(t-1) \quad ..........(4)
\]

Where = 1, 2, ..., represents the local best of the i-particles. Whereas = 1, 2, ..., represent the global best of the whole herd. While c1 and c2 are a constant constant positive constantsReferred to as learning factor. Then r1 and r2 are a random numberWhich ranges from 0 to 1. Equation (1) is used to calculate the velocity of a new particle based on its previous velocity, the distance between the current position with the best particle position (local best), and the distance between the current position and the best position of the flock (global best). Then the particle flies toward the new position based on equation (2). After the PSO algorithm is run with a certain number of iterations to reach the criteria of dismissal, it will get a solution located in the global best.

### 4. Research methodology

The research methodology used can be seen in Figure 2.

4.1. Data Sharing and Training Data sharing using 3 fold cross validation, same with previous training. The data will be divided into 2 for training data and 1 for many test data. The training was conducted using the VFI5 algorithm, where the specified trainer data was used as input of the VFI5 algorithm in the training process. Symptoms of disease as a feature of each patient, while DM and Dyspepsia is a class of data.
4.2. Normalization / Hose and Features

The lapse of each feature is obtained from the training process. Before determining the lapse do normalization on the training process. Hose contains a vote value for each class on each symptom.

4.3. Classification Classification is divided into 3 stages: first stage determination of vote value in test data, second stage summing value of each instances, and determine prediction class.

4.4. PSO optimization The process that will be run to optimize the feature weight with several stages as below:
   a. Determining the number of particles (particle size), the range of particles is 10 to 40 particles.
   b. Determines maximum speed (MaxVelocity), Maximum velocity (v) is set for particle displacement. If the velocity V is between -10 s / d to 10, then the maximum speed is 20.
   c. Learning factor Learning factors (c1 and c2) generally have a value of 2. Different The problem will also be different in value and the range is 0-2.
   d. Stop condition Includes maximum number of iterations of PSO Algorithm and error conditions Reach a minimum. The stop condition depends on the optimized problem.
   e. Inertia weight Inertia weight used is w = 1

4.5. Accuracy

The final stage of this research method is to calculate the accuracy of the results obtained at the data processing stage using VFI5 algorithm. Accuracy can be calculated by:

\[
\text{Accuracy} = \frac{\sum \text{data classification}}{\sum \text{total data}} \times 100\%
\]

5. Result and Discussion

5.1. 3- Fold Cross Validation The data is processed using Voting Feature Intervals (VFI5) algorithm. The first step is the training process, where each data sought the minimum and maximum value up to get the value of his vote weight.

| Table 1. Distribution of training and testing data |
|-----------------|-----------------|
| Data Training   | Data Testing    |
| Subset I        | Subset III      |
| Subset II       | Subset II       |
| Subset III      | Subset I        |
| Subset III      | Subset III      |

| Table 2. Accuracy of each fold Using VFI5 |
|-----------------|-----------------|
| Accuracy VFI5%  |                 |
| Fold 1          | 92.30           |
| Fold 2          | 92.59           |
| Fold 3          | 85.19           |
| Avrg            | 90.03           |
5.2. Experiment to find the optimum feature VFI5 weight with PSO algorithm

The representation of particles used is real number encoding and decoding. Real Number Encoding is used to represent the weight of each feature to be optimized. Decoder process is done by multiplying the weight of each feature with the feature vote of each class in each fold to find the vote value of each class. After searching the vote value of each class counts the percentage of each vote for each class, the process is a decode process presented in Appendix 1, Table 3 for weighting VFI5 and Table 5 for weighting with PSO Algorithm.

5.2.1. Representation of Particles

This experiment aims to determine the parameters of the most optimum PSO algorithm, which consists of testing particle size parameters, learning rate, inertia to obtain the best global best match until the calculation process stalled.

5.2.2. Determination of the optimum parameters In this experiment,

The relationship of PSO algorithm parameters such as population, inertia, personal learning rate with global learning rate is very influential.

5.2.3. . First Experiment

In the first experiment using parameter W = 1, Max Iteration = 1000, C1 = 1.0, C2 = 1.0. Global Best earned 2.69. Predictive errors are in the class, so in Table 4 the DM ID ID values on features with greater values in dyspepsia ID are nausea, fever, heartburn, chest pain, diarrhea, and cough. The second error in VFI5 is in the 80 dyspepsia ID data predicted into the DM class on features with a larger vote value on ID DM is dizziness, weakness, shortness of breath, diarrhea and cough. In Appendix 2, comparing data of normalization data after using PSO, it is found that the 80th data ID with dyspepsia class is correctly classified into the actual class ie dyspepsia. The result of the calculation on the fold to-one is 92.31% with 1 prediction error. 5.2.4. The Second Experiment In the second experiment using parameters W = 1, Max Iteration = 1000, C1 = 1.5, C2 = 1.5. Global Best obtained by 0. Data prediction errors are in the 21st ID and 67th ID and are presented in Appendix 3 of Table 5. The first incorrect data vote prediction value in the 21st ID of DM, where ID DM is predicted to ID Dyspepsia because of this DM ID that features dizziness, nausea, vomiting and weakness is actually the main characteristic of DM disease, but ID DM also features with a higher vote value on features of fever, heartburn, diarrhea and cough. Error in the ID of the 21st DM can be said that this ID has DM disease that has been complicated with dyspepsia. Accuracy in the second experiment was 92.59% with 2 data prediction errors.

5.2.5. The Third Experiment In the third experiment use parameter W = 1, Max Iteration = 1000, C1 = 2.0, C2 = 2.0. Global Best earned 3,704. Error Dyspepsia vote ID value on feature with bigger value in ID DM. The data are presented in Appendix 4 of Table 6. Prediction errors on classes in the VFI5 algorithm are not always the same as PSO algorithm results. The accuracy of the third experiment was 96.29% with one prediction error. From the above experiments, each of the accuracies presented in
### Table 3. Accuracy of all experiments

| Accuracy VF15 + PSO % |  |
|----------------------|---|
| Fold 1               | 96.15 |
| Fold 2               | 92.59 |
| Fold 3               | 96.29 |
| Avrg                | 95.01 |

5.3. Optimization Results

In each experiment the parameter size of the PSO algorithm produces better weight to produce better accuracy. In this research we get the pattern of weight of each feature after optimized with PSO algorithm. This study shows that if the population size is greater than the size of the generation the weights are more diverse, and the determination of the parameters greatly affects the search on global best.

![Grafik Peningkatan Akurasi](image)

Figure 1. Graph of Increased accuracy

From all experiment in this research, it is found that the value of hose result from training generally have the same pattern between the first iteration to the third iteration that is for the features of dizziness, nausea, vomiting and weakness are common symptoms of ID DM, while for fever feature, Heartburn, shortness of breath, chest pain, diarrhea and cough are symptoms of dyspepsia ID. But in the second iteration, the results of the training value are quite similar to the chest pain feature so that the value of the interval is not much different between DM class and dyspepsia class which means that for chest pain feature can be owned by both ID. The results of the training hose can be seen in Appendix 5. In Figure 3 shows an improved graph of experimental accuracy using the VF15 algorithm and the algorithm to optimize the PSO. The accuracy of the previous VF15 algorithm in fold 1 is 96.15% with 1 prediction error of 3.84% accuracy, in the second fold the accuracy of the VF15 algorithm is 92.59%, but no increase in the PSO calculation. The calculation on the third-fold accuracy generated by the VF15 algorithm of 85.19% increased to 96.29% and an increase of 11%.

6. Conclusions

This research has implemented the Optimization of Voting Feature Algorithm Interval5 (VF15) using PSO Algorithm. In general, the Partial Swarm Optimization algorithm has improved accuracy for several folds, therefore it can be concluded that the PSO algorithm is well used in optimizing the VF15 Classification Algorithm. Increased accuracy by using PSO Algorithm ie fold 1 from 92.31% to 96.15% increase accuracy of 3.8%, accuracy of fold 2 on Algorithm VF15 of 92.52% as well as
generated on PSO Algorithm means accuracy fixed, then in fold 3 increase accuracy of 85.19% Increased to 96.29% Accuracy increased by 11%. The total accuracy of all three trials is 14%.

7. References

[1] Rosyid, Abdul 2008. Diagnosis of Pulmonary Tuberculosis Using VFI5, Bogor: Department of Computer Science, Bogor Agricultural University.
[2] Guvenir HA, Emeksziz N, 2000. An expert system for the differential diagnosis of Erythematop-squamous diseases. www.elsevier.com/locate/eswa
[3] Hayana, Eka 2010. Diabetes Mellitus Disease Prediction Using Algorithm Classification Voting Feature Intervals 5, Bogor: Department of Computer Science, Bogor Agricultural University.
[4] Kristin, Ervina. 2010. Optimization of Voting Feature Intervals Algorithm 5 Using Genetic Algorithm On Pulmonary Tuberculosis Data, Bogor Department of Computer Science, Bogor Agricultural University.
[5] D. Dutta, R. Roy and K. Choudhury, Training artificial neural network using Particle swarm.International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no 3, pp. 430- 434, 2013.
[6] M. Nayak, GPAC-APSO Clustering using modified s-transform for data mining. International Journal Of Engineering Science & Advanced Technology, vol. 2, no.1, pp. 38 - 48,2012
[7] H. Yang, C-H. Hsiao and L-Y Chuang, Linearly Decreasing Weight Particle Swarm Optimization with Accelerated Strategy for ata Clustering. International Journal of Computer Science, vol. 37, no. 3, pp. 3-9, 2010.
[8] Lane, M.C., Xue, B. & Liu, I. 2013. PSO based selection of spectral features for remotely sensed image classification. International Journal on Artificial Intelligence Tools 22 (4): 214218.
[9] Wang, Yuhuai, Wang Qihui, Zhang Huixi, An Kang, Ye Xia and Sun Yaping 2016, Improved Particle Swarm Optimization Algorithm for Optimization of Power Communication Network. International Journal of Grid and Distributed Computing Vol. 9, No. 1 (2016), pp.225-236 http://dx.doi.org/10.14257/ijgdc.2016.9.1.23
[10] Demiroz G, Guvenir HA, 1997. Classification by Voting Feature Intervals. http://www.cs.ucf.edu/~ecl/papers/miros97classification.pdf.
[11] Y., Wang, G., Chen, H., Dong, H., Zhu, X., & Wang, S. Liu, "An Improved Particle Swarm Optimization for Feature Selection," Journal of Bionic Engineering, pp. 8(2), 191–200. doi:10.1016/S1672-6529(11)60020-6, 2011.
[12] R. L., &Haupt, S. E. Haupt, Practical Genetic Algorithms. United State of America: A John Wiley & Sons Inc Publication, 2004.
[13] Z.-L. Gaing, “A particle swarm optimization approach for optimum design of PID controller in AVR system,” IEEE Trans. Energy Conversion, vol.19, pp. 384-391, June 2004.
[14] Kennedy, J. and Eberhart, C., “Particle Swarm Optimization,” Proceedings of the 1995 IEEEInternational Conference on Neural Networks, Australia, pp. 1942 -1948, 1995.