A Comparative Study of Feature Selection Techniques for Bat Algorithm in Various Applications

Rozlini, Mohamed[1,*], *Munirah* Mohd Yusof[1], and *Noorhaniza* Wahidi[2]

[1]Department of Software Engineering, Universiti Tun Hussein Onn Malaysia (UTHM), 86400 Batu Pahat, Johor, Malaysia
[2]Department of Multimedia, Universiti Tun Hussein Onn Malaysia (UTHM), 86400 Batu Pahat, Johor, Malaysia

Abstract. Feature selection is a process to select the best feature among huge number of features in dataset, However, the problem in feature selection is to select a subset that give the better performs under some classifier. In producing better classification result, feature selection been applied in many of the classification works as part of preprocessing step; where only a subset of feature been used rather than the whole features from a particular dataset. This procedure not only can reduce the irrelevant features but in some cases able to increase classification performance due to finite sample size. In this study, Chi-Square (CH), Information Gain (IG) and Bat Algorithm (BA) are used to obtain the subset features on fourteen well-known dataset from various applications. To measure the performance of these selected features three benchmark classifier are used; k-Nearest Neighbor (kNN), Naïve Bayes (NB) and Decision Tree (DT). This paper then analyzes the performance of all classifiers with feature selection in term of accuracy, sensitivity, F-Measure and ROC. The objective of these study is to analyse the outperform feature selection techniques among conventional and heuristic techniques in various applications.

1 Introduction

Feature selection is a technique that has an ability to decrease the number of attribute by eliminates the least significant features [1]. However, the problem in feature selection in finding the optimum features. Most of the features in datasets that did not contribute to end result are unknown. Some unimportant or irrelevant features need to be diminished in order to reduce the classification complexity and time processing [1]. Feature selection becomes the important process in order to improve the classification performance. However, not all the feature selection technique reduces the same feature in dataset. For that reason, chosen the feature selection techniques is crucial where subset feature is needed for dimensionality reduction and gives better performance in classification.

Previously, some researcher likely to employ the Chi-Square (CH), Attribute evaluation (AE), Information Gain (IG), Correlation-based Attribute Evaluation (CB) and Symmetrical Uncertainty Attribute evaluation (SU) [2] that knows as conventional techniques. Today, several metaheuristic optimization algorithms seems become the potential technique to become feature selection techniques. The well-known metaheuristic optimization algorithms that widely such as Firefly Algorithm (FA), Genetic Algorithm (GA), Simulated Annealing (SA), Cuckoo Search (CS), Artificial Bee Colony (ABC), Bat Algorithm (BA) and Particle Swarm Optimization (PSO)[3]. Due to the strength of the combination of several metaheuristic algorithms, BA has become more powerful than PSO, GA and Harmony Search (HS) [4]. Thus, BA is seen as one of the possible solutions to resolve problem in data mining problems such as feature selection and classification.

The aim of this study is to find out the outperform feature selection technique by taking into consideration the capability of BA as a feature selection. The main contribution of this study are 1) run the experiment in 14 dataset from various application order to find out the outperform feature selection techniques between CH, IG and BA., 2) analyse the outperform feature selection techniques among conventional and heuristic techniques where kNN, NB and DT were applied in these experiments to evaluate the performance of selected features by using four performance measures accuracy, sensitivity, F-Measure and ROC area. The rest of this paper is organized to provide a brief explanation of BA in Bat Algorithm section. The following section will discuss about Related Work. While, in the next section will present the Methodology, Experimental Result and conclude the finding by Discussion and Conclusion in the last section.

2 Feature selection

Feature selection has been an active and fruitful field of research area in pattern recognition, machine learning, statistics and data mining communities [5]. It is a dimensionally reduction technique that main goal is to reduce irrelevant data and finding a features that increase classification accuracy. The main objective of feature

* Corresponding author: rozlini@uthm.edu.my

selection is to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information. It has been proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results [6,7].

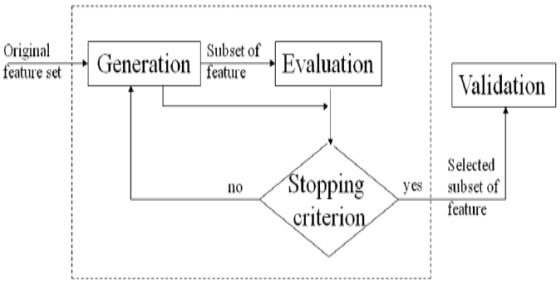

Fig. 1: Feature selection process [8].

There are four basic steps in a typical feature selection process as shown in Fig. 1[8].
The process of feature selection is as below;

- The generation procedure to generate the next candidate subset from original feature set
- The evaluation function to evaluate the subset to determine the relevancy towards the classification task using measure for instances distance, dependency, information and consistency
- Stopping criteria to decide when to stop. This is where it determine the relevant subset or optimal feature subset
- Validation procedure is to check whether the selected feature subset is valid

. There exists a several feature selection method that used by researcher. Some researcher trend to employ conventional method such as information gain and chi-square for instance [3,9,10]. In the other research, heuristic method such as genetic algorithm [11] ACO [12] and [13] in memetic feature selection, noisy data, spam email, binary variables; respectively.

Feature selection also involve as active field of research such as in pattern recognition, machine learning and data mining area [14,15]. Feature selection objective is to reduce irrelevant data and finding the most relevant features that would increase classification accuracy. It has been proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results [16].

A wrapper feature selection approach based on BA and Optimum Path Forest had been proposed by Nakamura [17]. This approach modeled a problem of feature selection as a binary based optimization technique. Six datasets been used in experiments that demonstrated that the proposed approach provides statistically significant more compact sets and in some cases it indeed improves the classification effectiveness.

Binary Bat Algorithm (BBA) was one of the inspired binary version feature selection that proposed to find the most significant feature in a search space [18]. BBA was proposed to associate each bat a set of binary coordinates that indicate whether that feature belongs to the final set of features or not. It combined the power of bat algorithm and Optimum Path Forest in finding the set of features that maximizes the accuracy of validating sets. It been proved that the proposed techniques can outperformed other well known techniques such as PSO, FFA and GSA.

From [19] proposed bio-inspired method called Bat Algorithm hybridized with a Naive Bayes classifier (BANB). Twelve benchmarks datasets from different domains been used in experiments to compared their performance measures with three well known feature selection techniques; GA, PSO and GPSO in term of the number of selected features from the original datasets. It shows that BANB significantly outperformed other algorithms in selecting significant number of features and lead to maintaining and improving classification accuracy.

Generally these studies found that feature selection techniques are capable to improve the performance of learning algorithms thru increasing the accuracy of the classifier by removing irrelevant attributes. Therefore with high quality features, it makes the classification process accurate, comprehensible and produces better results. For that reason this research will conduct the experiment that focuses to analyze the outperform techniques among conventional and heuristic techniques.

3 Experimental results

In this research, the experiment is setup to analyze the outperform feature selection techniques among conventional and heuristic techniques. The conventional techniques involve in this experiments are Chi-Square (CH) and Information Gain (IG). Meanwhile Bat algorithm (BA) as a heuristic method. This section discusses on the experimental results of 14 datasets from UCI repository in various application with selected attributes using CH, IG and BA. Three benchmark classifier; kNN, NB and DT DT were applied in these experiments to evaluate the performance of selected features by using four performance measures accuracy, sensitivity, F-Measure and ROC area.

Table 1 shows the characteristics of 14 datasets; number of attributes and instances. In general, selected features by CH and IG improve the classification performance for all classifier as shown in Table 2,3,5 until 15. From Table 2 and 3, ROC area for NB is a highest performance measures in Credit Approval and Ecoli with value 0.879 and 0.988, respectively. However the result for Hill Valley dataset is little bit different as shown in Table 4. Selected features by CH and IG only improve the classification performance for NB in all performance measures. But BA improves all the performance measures for KNN. Meanwhile, BA obtains the same result with CH and IG for DT in all

classification performance. The highest value is ROC area for KNN with value 0.59.

For Image Segmentation dataset, ROC area for NB, 0.946 is the highest value (see Table 5). Meanwhile, the ROC area with BA is 0.926. From Table 6, ROC area for NB, 0.943 is the highest value while ROC area by BA is 0.938. In Table 7, the highest value is ROC area for NB, 0.876. From Table 8 shows, ROC area for NB is the highest value for Plant Species dataset in CH, IG and Bat with value 0.996, 0.996 and 0.993, respectively. In Table 9, the highest value for ROC area for NB in CH, IG and Bat with value 0.965, 0.965 and 0.957, respectively. In Table 10 and 12, the highest value for ROC area in Automobile dataset is 0.806 and in Yeast dataset is 0.817.

In Table 11, the highest performance is ROC area for NB in CH, IG and Bat with value 0.71, 0.71 and 0.704, respectively. For Waveform dataset that shown in Table 13, the highest performance is ROC area for NB in CH and IG is 0.96 and BA is 0.954.From Table 14 and 15, the highest value is ROC area for NB is 0.935 and 0.878, respectively.

4 Discussion and conclusion

This study was used three feature selection techniques and 14 datasets in various applications from UCI repository. The experiment was testing using Weka and Matlab. CH and IG is consider as conventional features selection method while BA as heuristic features selection method. These features selection techniques are used to find the subset features from 14 datasets. Then, analyse the outperform feature selection techniques, it can be determine according to classification performance. There are three benchmark classifier; NB, KNN and DT with four performance measures accuracy, sensitivity, F-Measure and ROC area were used in this experiment

The results shows the outperform feature selection method is determine by ROC area. The outperform feature selection method is conventional techniques, CH and IG. However, BA has a potential to be outstanding feature selection method. The result from Table 4 show BA improves all the performance measures for KNN. And also, results from Table 5, 6, 8, 9, 11 and 13 shows the ROC area for BA is only slightly different with CH and IG.

Table 1. Datasets Characteristics

#	**Datasets**	**Number of Instances**	**Number of Attributes**
1.	Credit Approval	690	15
2.	Ecoli	336	8
3.	Hill_Valley	606	100
4.	Image Segmentation	210	19
5.	Libras Movement	360	90
6.	Steel Plates Faults	1941	27
7.	Plant Species	1600	64
8.	Urban Land	507	147
9.	Automobile	159	25
10.	Abalone	4177	8
11.	Yeast	1484	8
12.	Waveform	5000	21
13.	Ionosphere	351	34
14.	Water Treatment	523	38

Table 2. Results for classification performance for Credit Approval dataset

FS	**Irrelevant features**	**Precision**			**Recall**			**Fmeasure**			**ROC Area**		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
BA	8-11,14,15	0.628	0.609	0.632	0.596	0.61	0.632	0.59	0.61	0.632	0.68	0.604	0.625
CH	1,12	0.783	0.827	0.842	0.765	0.828	0.842	0.756	0.827	0.842	0.879	0.824	0.855
IG		0.783	0.827	0.842	0.765	0.828	0.842	0.756	0.827	0.842	0.879	0.824	0.855

Table 3. Results for classification performance for Ecoli dataset

FS	**Irrelevant features**	**Precision**			**Recall**			**Fmeasure**			**ROC Area**		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	1,2,4-6	0.694	0.704	0.665	0.735	0.711	0.753	0.712	0.707	0.7	0.936	0.831	0.879
CH	5	0.962	0.932	0.973	0.973	0.932	0.964	0.967	0.931	0.969	0.988	0.96	0.985
IG		0.962	0.932	0.973	0.973	0.932	0.964	0.967	0.931	0.969	0.988	0.96	0.985

Table 4. Results for classification performance for Hill Valley dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	2,8,20,21,23,27,28, 30-52, 55,57, 59,62,64,66,68,69,73-75 ,77-94,96,98-100	0.515	0.589	0.253	0.505	0.589	0.503	0.432	0.589	0.337	0.491	0.59	0.494
CH	1	0.519	0.584	0.253	0.507	0.584	0.503	0.434	0.584	0.337	0.492	0.585	0.494
IG		0.519	0.584	0.253	0.507	0.584	0.503	0.434	0.584	0.337	0.492	0.585	0.494

Table 5. Results for classification performance for Image Segmentation dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	1-3,5-9,14,16,19	0.619	0.771	0.781	0.619	0.771	0.776	0.587	0.769	0.777	0.926	0.87	0.902
CH	1,3,4,5	0.781	0.894	0.89	0.786	0.895	0.89	0.774	0.894	0.889	0.946	0.939	0.939
IG		0.781	0.894	0.89	0.786	0.895	0.89	0.774	0.894	0.889	0.946	0.939	0.939

Table 6. Results for classification performance for Libras Movement dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	1,3,5,7-9,11, 30,41,43,45 ,47-49,51,53-55, 57, 59,61,63,65,67,69,71,73,75,77,79,81,83,85,87 ,89	0.626	0.816	0.701	0.606	0.811	0.694	0.609	0.809	0.69	0.938	0.894	0.867
CH	43,47,51,55,59,57,61,63,65,67,69,71,73,75,77 ,79	0.647	0.865	0.709	0.631	0.861	0.706	0.635	0.859	0.702	0.943	0.921	0.866
IG		0.647	0.865	0.709	0.631	0.861	0.706	0.635	0.859	0.702	0.943	0.921	0.866

Table 7. Results for classification performance for Steel Plates Faults dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	1-5,8,11,12,15,	0.616	0.642	0.71	0.536	0.641	0.712	0.529	0.641	0.71	0.844	0.767	0.84
CH	10,21,16	0.663	0.707	0.749	0.6	0.708	0.75	0.582	0.707	0.749	0.876	0.809	0.859
IG	21,16	0.667	0.713	0.759	0.604	0.714	0.76	0.589	0.713	0.759	0.876	0.813	0.864

Table 8. Results for classification performance for Plant Species dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	2,3,5,6,8-10,14,16-19,21-26,30,41-45,48-51,55,57,63,64	0.785	0.644	0.412	0.774	0.639	0.398	0.776	0.636	0.4	0.993	0.818	0.728
CH	16,8,61,60,34	0.849	0.753	0.501	0.842	0.739	0.493	0.843	0.74	0.491	0.996	0.868	0.766
IG		0.849	0.753	0.501	0.842	0.739	0.493	0.843	0.74	0.491	0.996	0.868	0.766

Table 9. Results for classification performance for Urban Land dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	2,3,18-21,36-42,44,61,62,65,81-84,103,105,107,124,128,145,	0.761	0.751	0.77	0.75	0.738	0.768	0.752	0.738	0.767	0.957	0.845	0.884
CH	13,16-18,34,36,38,39,55,57,59,60,76,78,80,81,87,97-99,101,102,105,118-123,126,139-144,147	0.83	0.816	0.812	0.821	0.804	0.804	0.824	0.806	0.805	0.965	0.88	0.905
IG		0.83	0.816	0.812	0.821	0.804	0.804	0.824	0.806	0.805	0.965	0.88	0.905

Table 10. Results for classification performance for Automobile dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	1,9-13,16,21-25	0.486	0.717	0.641	0.453	0.698	0.648	0.437	0.706	0.64	0.765	0.814	0.858
CH	4,3,15,7,19,20,14,8	0.562	0.855	0.809	0.566	0.836	0.805	0.547	0.843	0.805	0.806	0.905	0.892
IG		0.562	0.855	0.809	0.566	0.836	0.805	0.547	0.843	0.805	0.806	0.905	0.892

Table 11. Results for classification performance for Abalone dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	3,7,8	0.182	0.19	0.212	0.233	0.192	0.219	0.201	0.191	0.215	0.704	0.549	0.604
CH	1	0.197	0.199	0.2	0.241	0.201	0.206	0.209	0.2	0.202	0.71	0.551	0.585
IG		0.197	0.199	0.2	0.241	0.201	0.206	0.209	0.2	0.202	0.71	0.551	0.585

Table 12. Results for classification performance for Yeast dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	5-8	0.517	0.487	0.496	0.53	0.486	0.515	0.519	0.486	0.504	0.79	0.659	0.711
CH	7	0.591	0.524	0.564	0.579	0.522	0.575	0.569	0.522	0.567	0.817	0.684	0.741
IG		0.591	0.524	0.564	0.579	0.522	0.575	0.569	0.522	0.567	0.817	0.684	0.741

Table 13. Results for classification performance for Waveform dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	3,7-9,13,17	0.831	0.745	0.762	0.814	0.746	0.762	0.808	0.745	0.762	0.954	0.809	0.856
CH	1,21	0.841	0.785	0.761	0.81	0.785	0.761	0.798	0.785	0.761	0.96	0.839	0.846
IG		0.841	0.785	0.761	0.81	0.785	0.761	0.798	0.785	0.761	0.96	0.839	0.846

Table 14. Results for classification performance for Ionosphere dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	3-5,8-11,13,15-17,19-21,23,26,28-31,33	0.788	0.896	0.862	0.789	0.895	0.863	0.788	0.892	0.862	0.881	0.861	0.854
CH	2	0.842	0.871	0.915	0.826	0.863	0.915	0.829	0.857	0.913	0.935	0.825	0.892
IG		0.842	0.871	0.915	0.826	0.863	0.915	0.829	0.857	0.913	0.935	0.825	0.892

Table 15. Results for classification performance for Water Treatment dataset

FS	Irrelevant features	Precision			Recall			Fmeasure			ROC Area		
		NB	KNN	DT	NB	KNN	DT	NB	KNN	DT	NB	KNN	DT
Bat	4,5,7,11-15,17-22,24-38	0.613	0.563	0.596	0.62	0.585	0.61	0.608	0.573	0.599	0.77	0.671	0.712
CH	2, 34, 36, 38	0.739	0.699	0.673	0.734	0.713	0.677	0.732	0.702	0.673	0.878	0.77	0.769
IG		0.739	0.699	0.673	0.734	0.713	0.677	0.732	0.702	0.673	0.878	0.77	0.769

Acknowledgments

This paper has been supported by, Short Term Grant, Universiti Tun Hussein Onn Malaysia (UTHM) (vot U539) for the financial support. This research also supported by GATES IT Solution Sdn. Bhd under its publication scheme.

References

1. R. Amirreza, N. Hossein, A hybrid feature selection approach based on ensemble method for high-dimensional data, 2nd Conference on Swarm Intelligence and Evolutionary Computation (CSIEC), 7-9 March, pp. 16-20, (2017)
2. Y. Zhang, A. Yang, C. Xiong, T. Wang, Z. Zhang, "Feature selection suing data envelopment analysis", Knowledge-Based Systems Journal, 64, pp. 70-80, April 2014, (2014)
3. Y. Sayes, I. Inza, , P. Larranaga, A review of feature selection techniques in bio-informatics, Bioinformatics vol. 23, no. 19, pp. 2507 – 2517, (2007)
4. X.S. Yang, A new metaheuristic Bat-inspired algorithm, in: J.R. Gonzalez, et al.(Eds.), Nature Inspired Cooperative Strategies for Optimization (NISCO 2010). Studies in Computational Intelligence, Springer Berlin, Springer, Berlin, August 2010, pp. 65–74, (2010)
5. J. Han, and M. Kamber, Data Mining; Concepts and Techniques. Morgan Kaufmann Publishers, 2000.
6. H. Almuallim, T. G. Dietterich, T. G. Learning boolean concepts in the presence of many irrelevant features. Artificial Intelligence, vol. 69, no. 1-2, pp. 279– 305, (1994)
7. D. Koller, M. Sahami, M. Toward optimal feature selection. In Proceedings of the Thirteenth International Conference on Machine Learning, pp. 284–292, (1996)
8. M. A. Hall, L. A. Smith, Feature Subset Selection: A Correlation Based Filter Approach, In 1997 International Conference on Neural Information Processing and Intelligent Information Systems, pp. 855- 858, (1997)
9. M. Y. Munirah, M. Rozlini, N. Wahid, A comparative analysis on feature selection techniques for medical datasets, APRN Journal of Engineering and Applied Sciences, vol 11, no 22, November 2016, (2016)
10. C. Shang, M. Li, S. Feng, Q. Jiang, J. Fan, Feature Selection via Maximizing Global Information Gain for Text Classification, *Knowledge-Based Systems*, vol. 54, 298-309, (2013)
11. J. H. Lee, J. R. Anaraki, C. W. Ahn, J. An, Eficient Classification System based on Fuzzy-Rough Feature Selection and Multitree Genetic Programming for intension Pattern Recognition using Brain Signal, Expert Systems with Applications, vol. 42, 1644-1651, (2015)
12. S. Kashef, H. Nezamabadi-pour, An Advanced ACO Algorithm for Feature Subset Selection, Neurocomputing vol. 147, 271279, (2015).
13. Y. Zhang, D. Gong, Y. Hu, W. Zhang,. Feature Selection Algorithm based on Bare Bones Particle Swarm Optimazation, Neurocomputing, vol. 148, pp.150-157, (2015)
14. Y. Shen-Lan, R. Gang, F. Yi-Ping, Multiple kernel learning based feature selection for process monitoring, 2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS), 24-16 May 2017, pp. 809-814, (2017)
15. B. Emel, S. Mustafa, Video classification based on ConvNet collaboration and feature selection, 25th Signal Processing and Communications Applications Conference (SIU), 15-18 May 2017, pp. 1-4, (2017)
16. D. Koller, and M. Sahami, "Toward optimal feature selection". In Proceedings of the Thirteenth International Conference on Machine Learning, pp. 284–292, (1996)
17. Rodrigues, L. Pereira and R.Y.M Nakamura, "Wrapper approach for feature selection based on bat algorithm and optimum-path forest", Expert Systems with Applications, vol.41, pp.2250–2258, (2014)
18. R.Y.M. Nakamura, L. Pereira, M. Acuckoo, K.A Costa, D. Rodrigues, J.P. Papa, X.S. Yang, "BBA: a binary bat algorithm for feature selection'. in 25th, SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 22–25 August, IEEE Publication, pp. 291-297, (2012)
19. A.M Taha, A. Mustapha and S.D. Chen, "Naïve Bayes-Guided bat algorithm for feature selection", The Scientific World Journal, vol 2013, http://dx.doi.org/10.1155/2013/325973 ,(2013)