A Systematic Approach to Identify an Appropriate Classifier for Limited-Sized Data Sets

Alanoud Bin Dris
National Center for Cyber Security Technology, King Abdulaziz City for Science and Technology
Riyadh, KSA
abindris@kacst.edu.sa

Najla Alzakari
National Center for Cyber Security Technology, King Abdulaziz City for Science and Technology
Riyadh, KSA
nalzakari@kacst.edu.sa

Heba Kurdi
Computer Science Department
King Saud University
Riyadh, KSA
hkurdi@ksu.edu.sa

***Abstract*—Data size is a main issue in any data mining application, since limited size data results in a small training set that leads to a poor classification model and therefore a poor classification performance. Although, many real-life applications need a classifier that deals with limited size data sets appropriately. A considerable interest is focused on how to achieve a reasonable classification performance for small data sets. Current works focus on either enhancing classification algorithms or enlarging the data sets, these solutions have limitations such as increasing the computational time, or reaching data sets that do not reflect the actual population of the real data. However, this research looks at the problem from a different angel, it aims to address the data quantity issue by identifying the most appropriate classifier for small data sets using three well-known classifiers which are Decision tree (J48), Support Vector Machine (SVM) and Naïve Bayes. Extensive experiments are conducted to examine the performance in terms of four different measures which are accuracy, f-measure, sensitivity and specificity. We used six small data sets from UCI repository with different attributes and instances sizes. Results revealed that SVM accomplished the best performance along most of the used data sets.**

Keywords—classification algorithms, classification, data mining, performance, small data sets

I. Introduction

Digital data grows day after day which introduce huge benefits for different fields when they collect and use these data effectively [1]. These data are raw materials that can be converted to useful knowledge or even wisdoms [2]. However, traditional processing techniques become infeasible when they applied to the enormous data amount as well as human analysis may take long time to reach useful information [3]. Data Mining is an increasing area of interest for data processing where classification acts as one of the main data processing techniques [4].

Nowadays, data classification plays a key role for various competitive global market to help them discover interesting and hidden knowledge from their collected data [5],[6]. Typically, it can be applied in a wide variety of mining applications due to its strength in extracting relationship between a set of feature variables and a target variable of interest [7]. Classification goes through two main steps where we use previous known data -training data-to build a model as a first step -learning step-. Then we examine the resulted model accuracy, if it is acceptable; use this model to classify new data -classification step-[4]. However, classification performance is strongly depending on the characteristics of the used data set. Data set size considered the main factor that affects classification performance [8].

Generally, classification algorithms may not perform accurately with small data sets [9]. Although, there are many real-world situations with an urgent need to apply classification algorithms on their data, nevertheless, the available data set is not large enough. For instance, a new product release in an early stages lead to deal with small number of customers' data that needs classification and analysis for system improvement [5]. As well, some rare diseases, such as bladder cancer for which there are only few medical records [5],[10]. Moreover, factors such as cost and time limitations are reasons that increase the usability of using small data sets [11]. Since limited-sized data set remains an open problem and scientists are willing to overcome this issue, our motivation of this study is to address an important question concerning which classification algorithm is an appropriate for small data sets?

A challenging task is to provide a high-performance classification results that work in all cases regardless of the data set size [12]. Central cause behind this challenge relates to the limited training data resulted from small data sets, compared with larger training data produced from large data sets which leads to a stronger classification model [13].

Current works focus on either enhancing classification algorithms, so they can accurately classify limited size data sets [10], or attempting to extend the data set size by adding new features [5],[14], or by extending the data instances [13], [15]. However, most of these suggested solutions have their own limitations, either by increasing computational time significantly, or getting a larger data set that does not reflect the actual population of the real data. To the best of our knowledge, there are no studies conducted to identify an appropriate classifier for limited size data set. Consequently, our aim is to recommend a specific classifier for limited size data sets.

This study contributed by conducting an experimental comparison to identify an appropriate classification algorithm for limited size data using three well-known classification algorithms namely: Decision tree (J48) [16], Support Vector Machine (SVM) [17], and Naïve Bayes [18]. These three

978-1-7281-1244-2/19/$31.00 ©2019 IEEE

classifiers were applied on six small different data sets, and the results were analyzed based on four different performance measures which are accuracy, f-measure, sensitivity and specificity.

The organization of this paper is as follows. Section II reviews the existing solutions proposed to handle limited size data issue and discusses their limitations, Section III describes the proposed methodology, and Section IV shows the experimental results along with a discussion on the findings. Finally, section V concludes the report and addresses some open research directions for future work.

II. Related Work

Different research studies have been applied to reveal patterns and discover knowledge using classification algorithms. In the medicine field, different studies conducted in order to predict the diseases existence in an early stage [19],[20]. Moreover, schools and universities are always looking forward improving their students' performance and increase their achievements, to do so; many research studies have been conducted in education area [21],[22],[23]. Social media is another field where the classification is applied to get benefits from its rich content [24],[25],[26].

Additionally, there are a lot of research studies compared between classification algorithms in order to recommend one to be used in a specific applications [24],[27],[28]. However, these research studies did not consider the size of data sets that involved in their experiments. Currently, dealing with small size data sets has been a research focus for its significant importance. Various solutions have been proposed to overcome low classification performance with such data. Mainly, Case-Based Reasoning is a proposed solution to classify small data set, but it suffers from storage and complexity issues [29].

In this research, we have categorized the literature on small data sets into three main categories based on their proposed solution. First, studies on extending data sets features. Second, studies on extending data sets samples. Third, enhancements on classification algorithms.

A. Studies on Extending Data Sets Features

Several researches looked at the problem from different angle where they suggest extending data set size by increasing data dimensionality. One of these research studies learn the relationship between the data set features to generate new data attributes using the fuzzy rules. Each data point is assigned to a membership value from fuzzy rules and these values are used to extend data set features, this improvement can enhance the performance of the classification processes [14].

Extending data set features has been studied in another research to test the applicability of extraction features from small data sets, the similarity-based algorithms using fuzzy membership function have been applied on 7 different data sets that falls in the range of 18 to 768 instances. The proposed method works with higher performance compared with Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA), and Kernel Independent Component Analysis (KICA) [5].

Extending data set features may enhance the classification performance, but in case most of the observations are not available in the data set, this solution cannot fill the gaps between data samples. Therefore, the unavailability of data observations degrades the performance [30].

B. Studies on Extending Data Sets Samples

On the other hand, generation of virtual instances is a proposed idea; one research suggests generating new data based on the Gaussian distribution by utilizing the smoothness, which is a form of the prior knowledge, it states that if two inputs are close to each other, their outputs will be close as well, this method may lead to have data replicates. New samples are added to the oriinal samples and then they constructed the new training data from the whole set.

This solution can improve the accuracy in case of having low dimensionality, imbalanced and full of noise data set, where the classifiers will learn more from the new training data [15]. This solution has a drawback, extending the small sample data that contain noise with new instances may not reflect the actual distribution of the population [31].

Posterior Probability Neural Networks is proposed algorithm to fill the gaps between samples, this research suggests using SVM to compute posterior probabilities for samples, which will facilitate the process of deriving new learning samples to fill up the gaps between the original data samples. However, the research concluded by addressing that the influence of adding new samples needs further investigations [13].

C. Enhancements on Classification Algorithms

Neural Networks algorithms successful prediction depends on the availability of data. For limited size data set, computational intelligence techniques have been proposed to enhance the performance [10]. These techniques use Fuzzy ARTMAP neural networks [32], which can learn incrementally to perform the prediction and classification. However, there is a trade-off between the proposed improvement and the computational overhead where this solution increases the computational overhead significantly, which makes it unscalable solution [10].

Most of the studies mentioned above are trying to extract useful information from the available data. The following Table 1 summarizes these studies and shows additional information about data sets characteristics.

Table 1 shows that the data set can be considered as small if it has the following characteristics:

- The number of instances fall in the range from 18 to 768.
- The number of features approximately fall in the range from 3 to 30.

To summarize, almost all proposed studies try to extract the maximum possible information from available data to enhance the classification performance, and these solutions have different limitations as mentioned above. However, there is no universal optimal solution to improve the performance of classification processes for small size data sets [10]. To help in overcoming these situations, this research recommends a specific classifier for limited size data sets, by applying different classification algorithms on several limited size data sets, and the results discussed based on different performance evaluation measures.

TABLE 1
RELATED WORKS SUMMARIZATION

Category \ Criteria	Research study	Number of data sets	Number of samples	Number of features	Training-Testing data partitioning
Extending data set features	Extending Sample Information for Small Data Set Prediction [14]	1	30	4	10-fold cross-validation
	Extending Attribute Information for Small Data Set Classification [5]	7	18 – 768	3 – 30	Applied many times on different separations
Extending data set samples	A novel virtual sample generation method based on Gaussian distribution [15]	3	66 - 90	4	10-fold cross-validation
	A New Method to Assist Small Data Set Neural Network Learning [13]	1	200	7	Applied many times on different separations
Classification method enhancement	Extreme Data Mining: Inference from Small Data sets [10]	1	176	NA	20-fold cross-validation

III. METHODOLOGY

The study objective is to decide which classifier is an appropriate for limited size data sets? To address this goal, we compare the performance of three different methods which are Decision tree (J48), Support Vector Machine (SVM) and Naïve Bayes [33]. The methodology followed in this study is presented in [8].

This experiment was executed with Weka version 3.8 running under windows 10 operating system with CPU 2.70 GHz, Core i7 processor and 8.0 GB of memory (RAM). It can be tested on many different computer platforms where Weka is available. Weka is Waikato Environment for Knowledge Analysis (WEKA), an open source software introduced by Waikato University, New Zealand. It provides a wide-range collection of algorithms and data preprocessing tools useful for research, education and project. In addition, it offers a full access to the state-of the-art techniques in machine learning used for data mining tasks. In addition, Weka facilitates experimental comparison of algorithms performance using many evaluation criteria such as accuracy. Experiments can be done on multiple classifiers that are run using multiple data sets [34],[35].

A. Data Sets

Based on the analysis of the characteristics of data sets in related work section, we derived ranges for number of samples and features for small data set. Where the range of number of feature is between 3 and 30 and the range of number samples is around 18 to 768. We have chosen six relatively small data sets within these ranges. For this experiment, all selected data sets are taken from UCI machine learning repository [36], Table 2 shows brief description of each selected data set.

TABLE 2
DATA SET DESCRIPTION

Data set	Description	Number of features	Number of instances
Statlog (Heart)	This data set is a heart disease database.	13	270
Forest type mapping	This data set goal is to map different forest types using spectral data.	27	325
Thoracic Surgery Data	This data is dedicated to classification problem related to the post-operative life expectancy in the lung cancer patients: class 1 - death within one year after surgery, class 2 – survival.	17	470
Student Performance	This data set predicts student performance in secondary education (high school).	33	649
Blood Transfusion Service Center	This data set classifies whether a patient donated blood in March 2007.	5	748
German Credit Data	This data set classifies people described by a set of attributes as good or bad credit risks.	20	1000

B. Preprocessing

Data preprocessing step should be considered to enhance the accuracy of each classifier [37]. All the selected data sets are free from missing values. Moreover, because of the classifiers are affected by the class attribute either it is nominal or numeric, a numeric to nominal filter has been applied to the class feature.

C. 10-Cross Validation

In this study, we divided our data into training and testing using 10-cross validation. The advantage of such partitioning is that all instances in the labeled data have an opportunity to be treated as test instances [38].

D. Classification Evaluation Measures

Classification Accuracy, F-measure, classification Sensitivity and Specificity are used to evaluate the classification performance. The accuracy, sensitivity and specificity are three common performance measurements [33].

IV. RESULTS AND DISCUSSION

Tables 4, 5, and 6 show the accuracy, f-measure, sensitivity and specificity for all data sets, starting with Decision tree, SVM then Naïve Bayes. Results are visualized as plot graphs to simplify the comparison as shown in Fig. 1-4. Sensitivity and specificity are assigned to value of ”NA” for the forest-type mapping data set, as it predicts more than 2 classes, while sensitivity and specificity are binary measures.

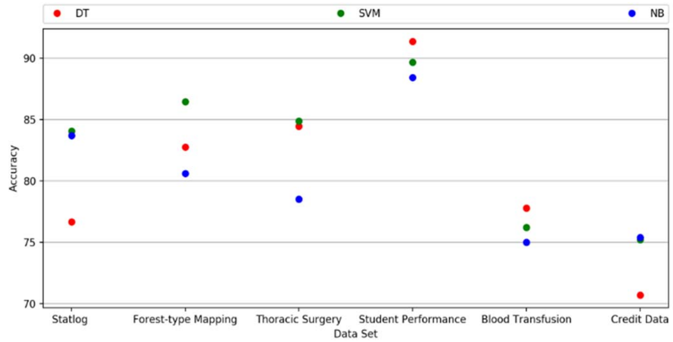

Fig. 1. Accuracy Level for each Classifier.

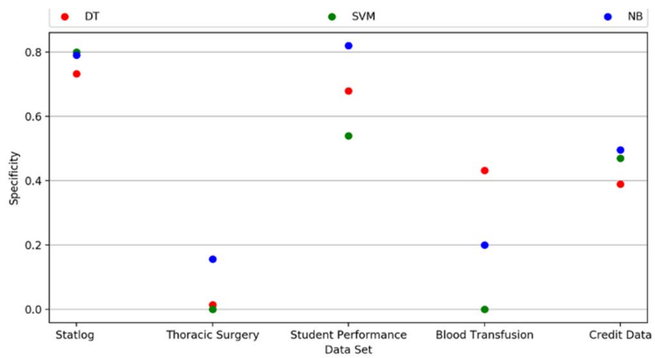

Fig.2. Speci• city Level for each Classifier.

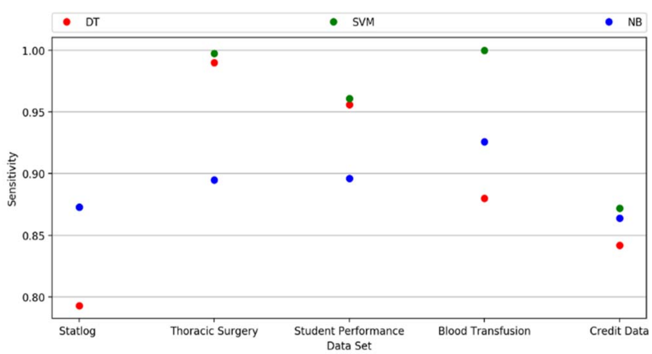

Fig. 3. Sensitivity Level for each Classifier.

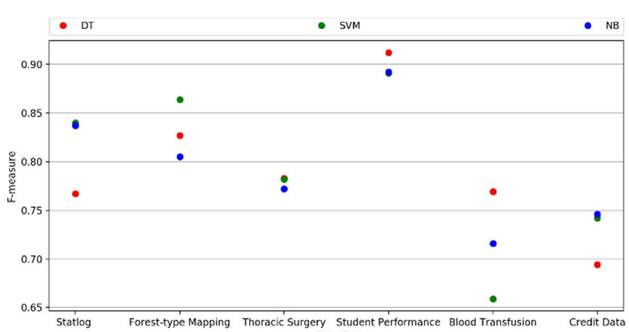

Fig. 4. F-Measure Level for each Classifier.

TABLE 4
J48 PERFORMANCE EVALUATION

Data set	J48			
	Accuracy	F-measure	Sensitivity	Specificity
Statlog (Heart)	76.667%	0.767	0.793	0.733
Forest Type Mapping	82.7692%	0.827	NA	NA
Thoracic Surgery Data	84.4681%	0.783	0.99	0.0142
Student Performance	**91.3731%**	0.912	0.956	0.68
Blood Transfusion Service Center	**77.80%**	0.769	0.88	0.432
German Credit Data	70.7%	0.694	0.842	0.39

TABLE 5
SVM PERFORMANCE EVALUATION

Data set	SVM			
	Accuracy	F-measure	Sensitivity	Specificity
Statlog (Heart)	**84.0741%**	0.840	0.873	0.8
Forest type mapping	**86.4615%**	0.864	NA	NA
Thoracic Surgery Data	**84.8936%**	0.782	0.9975	0
Student Performance	89.6764%	0.891	0.961	0.54
Blood Transfusion Service Center	76.2032%	0.659	1	0
German Credit Data	75.2%	0.742	0.872	0.47

TABLE 6
NAÏVE BAYES PERFORMANCE EVALUATION

Data set	Naïve Bayes			
	Accuracy	F-measure	Sensitivity	Specificity
Statlog (Heart)	83.7037%	0.837	0.873	0.791
Forest type mapping	80.6154%	0.805	NA	NA
Thoracic Surgery Data	78.5106%	0.772	0.895	0.157
Student Performance	88.4438%	0.892	0.896	0.82
Blood Transfusion Service Center	75.001%	0.716	0.926	0.20
German Credit Data	**75.4%**	0.746	0.864	0.496

Results show that Decision tree classifier (J48) reaches the highest accuracy 91.3731% among all classifiers within one data set only, Student Performance data set, out of the six cases. Also, it reaches a better accuracy 75.001% than SVM and Naïve Bayes with Blood Transfusion data set but this accuracy level does not consider as the optimal possible. For the other two measures sensitivity and specificity, J48 has a good sensitivity that is near to one in two data sets which are Thoracic Surgery Data and Student Performance while its specificity rate shows a weakness because it is far from zero in four cases out of five.

On the other hand, SVM achieves higher accuracies than the other two classifiers with three data sets which are Statlog (Heart), Forest-type mapping and Thoracic Surgery: 84.0741%, 86.4615% and 84.8936% respectively. SVM has a perfect sensitivity along Blood Transfusion Service Center data set that is equal to one while all other four cases have true positive rate that is near perfect. SVM specificity shows optimal values along two data sets, Thoracic Surgery Data and Blood Transfusion Service Center, that is equal to zero and good results along the other three cases.

For Naïve Bayes, it has approximately an equal accuracy with SVM among German Credit Data with a slight increase than the other two, yet this increase does not consider the optimal accuracy compared with SVM and Decision Tree results. Sensitivity shows high values among all data sets but not as good as previous two classifiers while Naïve Bayes specificity shows values that are far from zero in almost all data sets.

It is evident from the results that SVM is considered as the best option for small data sets because it has high classification accuracy along half of the data sets which means it shows higher number of correctly classified instances in three out of six data sets. The second choice for classifying small data sets is Decision tree algorithm. On the other hand, Naïve Bayes results showed the lowest classification accuracy in five cases among the three algorithms.

Consequently, SVM outperforms all other classifiers in terms of classification accuracy as well as in f-measure. Sensitivity shows the true positive rate while Specificity is the true negative rate. From the results, it can be noticed that the best case is when true positive rate reaches one which means all instances are classified correctly while true negative rate reaches zero means no instance has been classified wrongly. SVM has proven this in approximately two data sets cases, Blood Transfusion Service Center and Thoracic Surgery Data.

To sum up, based on the produced results we can conclude that SVM classifier is the first choice for relatively limited size data sets due to its nature where it tries to build a hyper-plane that separates classes from each other with the goal of maximizing the margin. Support vector points are the data points that used for constructing the hyper-plane. The increase of data samples or data features can accordingly increase the number of support vectors, and the complexity of the model will be increased as well. However, SVM may not produce high performance for all small data sets; as there are different influential factors that affect SVM classification like data distribution, noisy data and the overlapping between data points [33],[39],[40]. Naïve Bayes can predict the class of each data point by computing the membership probability of the data point for each class, and then it picks up the maximum probability. Applying Naïve Bayes on small data set might not represent the real probability for real class as the reliable estimation of class probability needs big data to represent the real probability for classes.

Additionally, Naïve Bayes algorithm assumes that there are no dependencies between attributes values in a given class and this is why it named Naïve, therefore, it ignores all dependences between the attributes. This assumption leads to a performance degradation especially with the limited data with relationships between its features. Small data sets do not represent the real population of data; therefore, decision tree may cause overfitting problem, it can generate complex trees that do not generalize the real data well. Moreover, small data leads to the generation of set of completely different trees [33].

V. CONCLUSION

To conclude, this study reviewed current work that aims in solving limited size data issue using different solutions that mainly focuses on extracting more information from the available data. However, adding new information may not reflect the real distribution of population. An enhancement on neural networks to make it suits small data set is proposed as well but it raises the computation overhead significantly.

To address this issue, our research goal is to reach a conclusion by answering the question which classification algorithm is an appropriate for small data sets? For this purpose, we have performed an extensive experiment to determine the most appropriate algorithm for limited size data sets that consist of 200 to 1000 instances. Three classifiers namely J48, SVM and Naïve Bayes were compared on the basis of correctly classified instances (i.e. accuracy), F-Measure, Sensitivity and Specificity.

Encouraging results were achieved by SVM where it has the highest classification accuracy performance. Furthermore, J48 showed acceptable results within two data sets. Naïve Bayes classifier showed less accuracy as compared to the previous two mentioned. For the sensitivity and specificity, SVM was near perfect in two data sets which adds a value for using this classifier. Moreover, the study indicates that SVM classification algorithm should be favored with limited size data sets applications when classification accuracy performance is important.

There are several future directions. For example, the inclusion of different evaluation measures into consideration such as the time requirement to compare the performance of each algorithm. In addition, enhancing SVM to increase its ability in classifying limited size data sets is an open research direction.

ACKNOWLEDGMENT

This work was supported by the Deanship of Scientific Research at King Saud University through research group no. RG-1438-002.

REFERENCES

[1] V. Ribeiro, A. Rocha, R. Peixoto, F. Portela, and M. F. Santos, "Importance of statistics for data mining and data science," in 2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), Aug 2017, pp. 156–163.
[2] J. M. Stanton, "Introduction to data science," 2013.
[3] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
[4] P.-N. Tan et al., Introduction to data mining. Pearson Education India, 2006.
[5] D.-C. Li and C.-W. Liu, "Extending attribute information for small data set classification," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 3, pp. 452–464, 2012.
[6] G. A. Marcoulides, "Discovering knowledge in data: an introduction to data mining," 2005.
[7] C. C. Aggarwal, Data classification: algorithms and applications. CRC Press, 2014.

[8] M. Sordo and Q. Zeng, "On sample size and classification accuracy: a performance comparison," Biological and medical data analysis, pp. 193–201, 2005.
[9] J. Prusa, T. M. Khoshgoftaar, and N. Seliya, "The effect of dataset size on training tweet sentiment classifiers," in Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on. IEEE, 2015, pp. 96–102.
[10] R. Andonie, "Extreme data mining: Inference from small datasets," International Journal of Computers Communications & Control, vol. 5, no. 3, pp. 280–291, 2010.
[11] D.-C. Li, Y.-C. Chang, M.-L. Su, and L.-S. Lin, "A learning method for small data sets with multimodality variables," in Grey Systems and Intelligent Services, 2013 IEEE International Conference on. IEEE, 2013, pp. 481–483.
[12] L. Mor´an-Fern´andez, V. Bol´on-Canedo, and A. Alonso-Betanzos, "Can classification performance be predicted by complexity measures? a study using microarray data," Knowledge and Information Systems, vol. 51, no. 3, pp. 1067–1090, 2017.
[13] R. Mao, H. Zhu, L. Zhang, and A. Chen, "A new method to assist small data set neural network learning," in Intelligent Systems Design and Applications, 2006. ISDA'06. Sixth International Conference on, vol. 1. IEEE, 2006, pp. 17–22.
[14] H.-Y. Chen, D.-C. Li, and L.-S. Lin, "Extending sample information for small data set prediction," in Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on. IEEE, 2016, pp. 710–714.
[15] J. Yang, X. Yu, Z.-Q. Xie, and J.-P. Zhang, "A novel virtual sample generation method based on gaussian distribution," Knowledge-Based Systems, vol. 24, no. 6, pp. 740–748, 2011.
[16] A. Trabelsi, Z. Elouedi, and E. Lefevre, "Handling uncertain attribute values in decision tree classifier using the belief function theory," in International Conference on Artificial Intelligence: Methodology, Systems, and Applications. Springer, 2016, pp. 26–35.
[17] D. Kumar and M. Thakur, "Weighted multicategory nonparallel planes svm classifiers," Neurocomputing, vol. 211, pp. 106–116, 2016.
[18] Z.-L. Xiang, X.-R. Yu, and D.-K. Kang, "Experimental analysis of na¨ıve bayes classifier based on an attribute weighting framework with smooth kernel density estimations," Applied Intelligence, vol. 44, no. 3, pp. 611–620, 2016.
[19] G. Fung and J. Stoeckel, "Svm feature selection for classification of spect images of alzheimer's disease using spatial information," Knowledge and Information Systems, vol. 11, no. 2, pp. 243–258, 2007.
[20] N. R. R. Nery Jr, D. B. Claro, and J. C. Lindow, "Prediction of leptospirosis cases using classification algorithms," IET Software, 2017.
[21] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.
[22] X. Chen, M. Vorvoreanu, and K. Madhavan, "Mining social media data for understanding students learning experiences," IEEE Transactions on Learning Technologies, vol. 7, no. 3, pp. 246–259, 2014.
[23] J.-L. Huang, Y.-F. Yeh, Y.-S. Hsu, and Y.-H. Tseng, "Critical factors of becoming secondary school teachers: Mining the process of teacher education by decision trees," JOURNAL OF RESEARCH IN EDUCATION SCIENCES, vol. 62, no. 2, pp. 89–123, 2017.
[24] S. B. Aher and L. Lobo, "Comparative study of classification algorithms," International Journal of Information Technology, vol. 5, no. 2, pp. 239–43, 2012.
[25] S. Wang, Z. Yan, X. Hu, S. Y. Philip, Z. Li, and B. Wang, "Cpb: a classification-based approach for burst time prediction in cascades," Knowledge and Information Systems, vol. 49, no. 1, pp. 243–271, 2016.
[26] T. Iju, S. Endo, K. Y. N. Toma, and Y. Akamine, "Estimating age on twitter using self-training semi-supervised svm," JOURNAL OF ROBOTICS NETWORKING AND ARTIFICIAL LIFE, vol. 3, no. 1, pp. 24–27, 2016.
[27] A. K. Sharma and S. Sahni, "A comparative study of classification algorithms for spam email data analysis," International Journal on Computer Science and Engineering, vol. 3, no. 5, pp. 1890–1895, 2011.
[28] M. Kumari and S. Godara, "Comparative study of data mining classification methods in cardiovascular disease prediction 1," 2011.
[29] S. Kasif, S. Salzberg, D. Waltz, J. Rachlin, and D. W. Aha, "A probabilistic framework for memory-based reasoning," Artificial Intelligence, vol. 104, no. 1, pp. 287–311, 1998.
[30] D.-C. Li, W.-K. Lin, L.-S. Lin, C.-C. Chen, and W.-T. Huang, "The attribute-trendsimilarity method to improve learning performance for small datasets," International Journal of Production Research, vol. 55, no. 7, pp. 1898–1913, 2017.
[31] G. Coqueret, "Approximate norta simulations for virtual sample generation," Expert Systems with Applications, vol. 73, pp. 69–81, 2017.
[32] R. Andonie and L. Sasu, "Fuzzy artmap with input relevances," IEEE Trans. Neural Networks, vol. 17, no. 4, pp. 929–941, 2006.
[33] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.
[34] F. Eibe, M. Hall, I. Witten, and J. Pal, "The weka workbench," Online appendix for data mining: practical machine learning tools and techniques.: Fourth Morgan Kaufmann, 2016.
[35] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10–18, 2009.
[36] C. L. Blake and C. J. Merz, "Uci repository of machine learning databases [http://www. ics. uci. edu/˜ mlearn/mlrepository. html]. irvine, ca: University of california," Department of Information and Computer Science, vol. 55, 1998.
[37] S. Garc´ıa, J. Luengo, and F. Herrera, "Tutorial on practical tips of the most influential data preprocessing algorithms in data mining," Knowledge-Based Systems, vol. 98, pp. 1–29, 2016.
[38] R. Kohavi et al., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Ijcai, vol. 14, no. 2. Stanford, CA, 1995, pp. 1137–1145.
[39] Y.-J. Lee, Y.-R. Yeh, and H.-K. Pao, "Introduction to support vector machines and their applications in bankruptcy prognosis," in Handbook of computational finance. Springer, 2012, pp. 731–761.
[40] S. R. Gunn et al., "Support vector machines for classification and regression," ISIS technical report, vol. 14, pp. 85–86, 1998.