

Learning with Imbalanced Datasets using Fuzzy ARTMAP-based Neural Network Models

Shing Chiang Tan

Faculty of Information Science and Technology
Multimedia University, Malaysia
sctan@mmu.edu.my

Zuwarie Ibrahim

Centre of Artificial Intelligence and Robotics (CAIRO)
University of Technology Malaysia, Malaysia
zuwairie@fke.utm.my

Lee Wen Jau

ATTD Automation (APAC) Pathfinding Intel Technology
Sdn. Bhd. Kulim, Malaysia
wen.jau.lee@intel.com

Junzo Watada

Graduate School of Information, Production and Systems
Waseda University, Japan
watada@waseda.jp

Marzuki Khalid

Centre of Artificial Intelligence and Robotics (CAIRO)
University of Technology Malaysia, Malaysia
marzuki@utm.my

Lim Chun Chew

ATTD Automation (APAC) Pathfinding Intel Technology
Sdn. Bhd. Kulim, Malaysia
chun.chew.lim@intel.com

Abstract—One of the main difficulties in real-world data classification and analysis tasks is that the data distribution can be imbalanced. In this paper, a variant of the supervised learning neural network from the Adaptive Resonance Theory (ART) family, i.e., Fuzzy ARTMAP (FAM) which is equipped with a conflict-resolving facility, is proposed to classify an imbalanced dataset that represents a real problem in the semiconductor industry. The FAM model is combined with the Dynamic Decay Adjustment (DDA) algorithm to form a hybrid FAMDDA network. The classification results of FAM and FAMDDA are presented, compared, and analyzed using several classification metrics. The outcomes positively indicate the effectiveness of the proposed FAMDDA network in undertaking classification problems with imbalanced datasets.

Keywords—Data classification; imbalanced data; Adaptive Resonance Theory Neural Networks; supervised learning

I. INTRODUCTION

Many problems in modern engineering applications can be undertaken as a classification task. Over the years, various machine learning techniques have been investigated and applied to process and analyze real databases from industries. In particular, data classification is an active research field in machine learning. A classification model can be constructed by capturing associations between the input and the output variables of a dataset through a supervised learning procedure. The effectiveness of the classification model can be affected by an imbalanced dataset in which its data samples from one class exist in a large quantity (i.e., the majority class) whereas data samples with other classes are available in small quantities (i.e., the minority classes). Imbalanced datasets are commonly observed in real applications. Many real-world datasets exhibit such characteristics, for examples, in medical prognosis and diagnosis [1], vision recognition [2], credit card fraud detection [3], and bioinformatics [4]. In these applications, the

occurrence of minority/rare cases is occasionally given more attention than that of the majority.

Learning with an imbalanced dataset is never an easy task. One of the main challenges is how a classification model, whereby its knowledge base is dominated by information from the majority class, is able to learn information of the minority classes accurately. The reason is that a classification model may tend to neglect information acquired from data samples with the minority classes owing to overwhelming information of the majority class. The situation becomes worse if the imbalanced dataset is large in size and has an arbitrary high complexity. From the literature, data mining and machine learning techniques such as decision trees, k -nearest neighbor and naïve Bayesian [5] have been employed as classification models for handling imbalanced datasets. Nevertheless, they are unable to yield satisfactory results [5]. Modifications made on these techniques (e.g., [6, 7]) are able to improve the classification performances in dealing with imbalanced datasets. In this regard, two common approaches have been adopted to construct a more effective classification model for tackling imbalanced datasets, i.e., modifications are at either the data level [6] or the algorithm level [7]. At the data level, re-sampling methods, which include methods for under sampling the instances from the majority class, over sampling the instances from the minority classes, and the use of these two methods, have been introduced. The re-sampling methods are applied to determine a good mix of samples belonging to different classes from the imbalanced dataset for achieving good classification performances. On the other hand, at the algorithm level, the learning algorithm of the classification model is modified so as to learn the characteristics of samples from the minority classes more effectively.

In this research, a supervised Adaptive Resonance Theory (ART) neural network, namely Fuzzy ARTMAP (FAM) [8], is proposed as a base classifier to handle imbalanced datasets. FAM, an incremental learning model, is able to overcome the

stability-plasticity dilemma [9]. The FAM network is *plastic* enough to absorb new information from new samples and, at the same time, is *stable* enough to retain previously learned information from being corrupted by newly learned information. An interesting feature of FAM is that it integrates fuzzy set theory [10] and the stability-plasticity characteristic of ART into a common framework, and this allows the network to encode both analog and binary input patterns (i.e., information) in the form of rectangular prototypes. The focus of this research is on evaluating the effectiveness of a variant of FAM in dealing with a large imbalanced dataset collected from a manufacturing industry. Original FAM does not have the conflict resolving capability for handling overlaps among prototypes of different classes. In this case, a Dynamic Decay Adjustment (DDA) algorithm [11] can be integrated within the supervised learning mechanism of FAM, and the resulting network is known as FAMDDA. As far as imbalanced datasets are concerned, such additional facility is helpful for FAM to adapt and distinguish information of the minority classes from those of the majority class more accurately; hence, the network is able to provide a better solution to problem under scrutiny.

The organization of this paper is as follows. Section II explains the classification models (i.e., FAM and FAMDDA) used for learning imbalanced datasets. Section III describes the indicators for assessing the performance of a classification model with imbalanced datasets. In Section IV, a case study with a large imbalanced dataset from a semiconductor manufacturing plant is carried out. The results and the associated discussion are presented. Concluding remarks are given in Section V.

II. FUZZY ARTMAP-BASED NEURAL NETWORKS

A. Fuzzy ARTMAP (FAM)

FAM [8] is an incremental learning neural network from the ART family. The network is capable of self-organizing and self-stabilizing information gained from training samples during supervised learning. The structure of FAM is depicted in Fig. 1. The network is constructed by a pair of fuzzy ART modules (i.e., ART_a and ART_b) that are interconnected through a map field, F^{ab} . Each ART module consists of three layers of nodes, i.e., $F_0^a(F_0^b)$ is the normalization layer wherein an M -dimensional input vector, $\mathbf{a}$, is complement-coded [8] to a $2M$ -dimensional vector $\mathbf{A}$; $F_1^a(F_1^b)$ is the input layer wherein the nodes receive the complement-coded input vectors; $F_2^a(F_2^b)$ is the recognition layer containing recognition nodes that encode prototypes of input patterns. Note that new nodes can be created in $F_2^a(F_2^b)$ when necessary to include new information.

The learning procedure of FAM is as follows. During supervised learning, an input pattern is presented to ART_a whereas its corresponding target output is presented to ART_b . At ART_a , the input pattern is complement-coded as pattern $\mathbf{A}$ and then propagated to F_2^a through F_1^a . The activation of each node j in F_2^a is measured by a choice function [8]

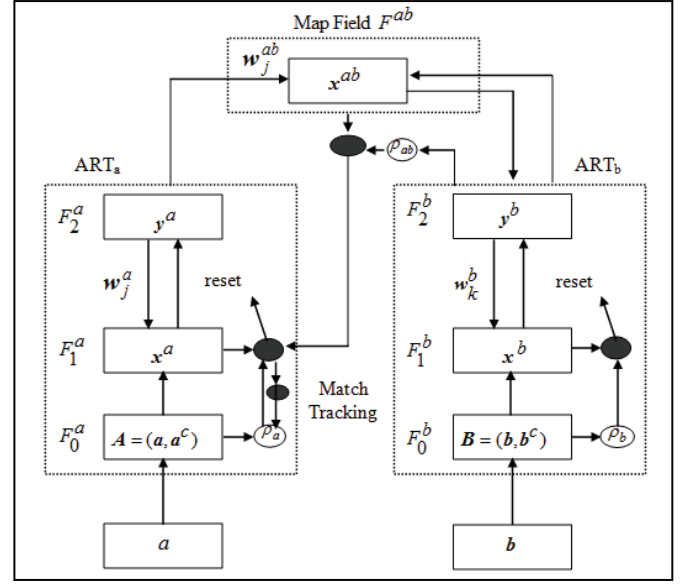


Figure 1. The network structure of Fuzzy ARTMAP.

$$T_j = \frac{|\mathbf{A} \wedge \mathbf{w}_j^a|}{\chi_a + |\mathbf{w}_j^a|} \quad (1)$$

where χ_a is the choice parameter and is set close to zero [8]; $\mathbf{w}_j^a$ the weight of node j . Based on the *winner-take-all* competition scheme, the node with the largest activation is selected as the winning node, and is denoted as node J . FAM imposes a vigilance test that measures the similarity between the winning prototype patterns, $\mathbf{w}_J^a$, and $\mathbf{A}$ against a threshold (vigilance parameter, ρ_a), as follows [8]

$$\frac{|\mathbf{A} \wedge \mathbf{w}_J^a|}{|\mathbf{A}|} \geq \rho_a \quad (2)$$

If the winning node fails in the vigilance test, a new search cycle for another winning node is exercised. The process of searching for a winning node is continued until the selected node is able to satisfy the vigilance test. If no such node exists, a new node is created in F_2^a to encode the input pattern. The same pattern-matching cycle takes place in ART_b using the target vector to find a winning node that encodes the target class.

Upon the identification of the winning nodes in F_2^a and F_2^b , a prediction is sent from F_2^a to F_2^b via F^{ab} . A map-field vigilance test is conducted to ascertain the prediction, i.e., [8]

$$\frac{|\mathbf{y}^b \wedge \mathbf{w}_J^{ab}|}{|\mathbf{y}^b|} \geq \rho_{ab} \quad (3)$$

where $\mathbf{y}^b$ denotes the output vector of F_2^b ; $\mathbf{w}_J^{ab}$ denotes the weight vector from F_2^a to F^{ab} ; and ρ_{ab} is the map-field

vigilance parameter. If the map field vigilance test fails, this indicates that the winning node in F_2^a has made a wrong prediction of the target class in F_2^b ; hence, a matching-tracking process [8] ensues. Parameter ρ_a , which is initially set to a user-defined baseline vigilance parameter $\bar{\rho}_a$, is now increased to

$$\rho_a = \frac{|A \wedge w_J^a|}{|A|} + \delta \quad (4)$$

where δ is a small positive value. By increasing ρ_a , the current winning node in F_2^a will fail in the ART_a vigilance test. A new search cycle in ART_a is initiated again with a new level of ρ_a . This process is repeated until a correct prediction is made between the winning nodes in F_2^a and F_2^b . Then, FAM enters a learning phase, whereby the weight vector of the winning node in F_2^a is updated as [8]

$$w_J^{a(new)} = \beta_a (A \wedge w_J^{a(old)}) + (1 - \beta_a) w_J^{a(old)} \quad (5)$$

where β_a is the learning rate of ART_a . The same learning process as in ART_a also takes place in ART_b . Note that all equations in ART_a are applicable to ART_b , but with superscript or subscript a replaced by b .

B. Fuzzy ARTMAP with the Dynamic Decay Adjustment (FAMDDA) Algorithm

The architecture of FAMDDA is the same as that of FAM. However, FAMDDA is equipped with an additional conflict-resolving facility contributed by the DDA algorithm. In this regard, the FAM learning procedure, specifically in its ART_a module, is modified for resolving overlaps (conflicts) among prototypes of different classes.

In order to include the idea of infinite/finite volume as in [11] within the FAM framework, each dimension d of prototype p in the recognition layer (i.e., F_2^a) of the ART_a module is given a status of $S_{pd}=0$ (i.e., infinite dimension) or $S_{pd}=1$ (i.e., finite dimension). All dimensions of a newly committed F_2^a prototype are initialized as 0. When the prototype is involved in width shrinking, the status of the selected dimension is changed to 1. In addition, a new set of reference vector, w_J^r , is introduced to each prototype in F_2^a . Each reference vector is initialized as a zero vector. When learning takes place, besides the weight vector w_J^a [8], the reference vector of the J -th F_2^a winning node is also updated according to a recursive center estimation procedure [12], as follows.

$$(w_J^r)^{new} = (w_J^r)^{old} + \frac{1}{N_J} (A - (w_J^r)^{old}) \quad (6)$$

where N_J is the number of input patterns of the J -th node, $N_J = N_J + 1$.

If x , which represents the weights of the M -dimension (or $2M$ -dimension with complement coding) winning prototype falls in the region formed by the prototype of different classes (i.e., λ), a conflict is said to occur. As such, a width shrinking procedure is executed to avoid the conflict. The width of the conflicting prototypical region is shrunk. The shrinking procedure is applied successively between the winning prototype and other conflicting prototypes. In general, three cases of width shrinking are considered [11]. First, if the existing finite dimensions of the conflicting prototype q can be shrunk without falling below a pre-set $\epsilon_{\min}$, the one with the smallest loss in volume ($\gamma_{best,k}$) is chosen, as follows [11].

$$\gamma_{best,k} = \min \{ |w_{qk}^r - x_k| : \forall I \leq i \leq n, i \neq k : \left(\frac{\lambda_{qk} - |w_{qk}^r - x_k|}{\lambda_{qk}} \leq \frac{\lambda_{qi} - |w_{qi}^r - x_i|}{\lambda_{qi}} \right) \wedge (\lambda_{qk} \geq \epsilon_{k,min}) : S_{qi}, S_{qk} = 1 \} \quad (7)$$

where $\lambda_{qj} = |w_{qj} - w_{qj}^r|$ and w_{qj} represents the weight of the prototype q at dimension j . Equation (7) indicates that shrinking of an existing finite dimension is preferred for not losing the “infinite volume” of other infinite dimensions. Such idea also complies with the learning paradigm of FAM, which supports the stability property; i.e., it helps protect previously learned knowledge of FAMDDA from being washed away. However, if the first option is not satisfied, the following options are considered, i.e., either one of the remaining infinite dimensions is shrunk ($\gamma_{\max,l}$),

$$\gamma_{\max,l} = \max \{ |w_{ql}^r - x_l| : S_{ql} = 0 \} \quad (8)$$

or, the width of the finite dimension of the prototype q is shrunk

$$\gamma_{\min,m} = \min \{ |w_{qm}^r - x_{qm}| : \forall I \leq i \leq n, i \neq m : \left(\frac{\lambda_{qm} - |w_{qm}^r - x_{qm}|}{\lambda_{qm}} \leq \frac{\lambda_{qi} - |w_{qi}^r - x_{qi}|}{\lambda_{qi}} \right) : S_{qi}, S_{qm} = 1 \} \quad (9)$$

Equation (8) is selected if $\gamma_{\max,l} > \gamma_{\min,m}$. Upon the selection of γ_z from (7)–(9), region λ_{qk} is reduced by adjusting w_{qz} , as follows.

$$w_{qz}^{new} = \begin{cases} w_{qz}^{old} + \gamma_z & \text{if } w_{qz}^{old} < w_{qz}^r \\ w_{qz}^{old} - \gamma_z & \text{if } w_{qz}^{old} > w_{qz}^r \end{cases} \quad (10)$$

If $w_{qz} > 1$ or $w_{qz} < 0$, it is rounded as $w_{qz} = 1$ or $w_{qz} = 0$, respectively. For clarity, FAMDDA undergoes sequential learning which synchronizes the operation of the DDA algorithm with the learning process of FAM in a fast manner. In this situation, the DDA algorithm is not possible to find an optimum solution which can eliminate all conflicts within reasonable time in this sequential learning paradigm. Instead,

the proposed FAMDDA network attempts to reduce and/or avoid overlapping among prototypes of conflicting classes during its training session. The conflict-resolved prototypes and their associations obtained in the training phase are used, during the test phase, to recall a prediction when an unseen pattern is presented to ART_a .

In general, the training procedure of FAMDDA can be summarized as follows.

1. An M -dimensional input pattern $\mathbf{a} \in [0,1]^M$ in F_0^a is complement-coded to a $2M$ -dimensional vector $\mathbf{A}$.
2. The input vector $\mathbf{A}$ is propagated from F_1^a to F_2^a through the weight vector, $\mathbf{w}^a$. The response of each node is computed according to the match function as in (1).
3. The prototype of node J is propagated back from F_2^a to F_1^a to participate in the vigilance test as in (2).
4. If the vigilance test fails, a new search cycle is initiated by re-visiting step 3. (The above cycle runs simultaneously in ART_b).
5. A prediction is sent from F_2^a (i.e., $\mathbf{w}_J^a$) and from F_2^b (i.e., $\mathbf{y}^b$), respectively, to F^{ab} , where the map-field vigilance test as in (3) is performed.
6. If the map-field vigilance test fails, a matching-tracking procedure as in (4) is triggered. Similar to FAM, the matching-tracking procedure only affects the ART_a module of FAMDDA, and step 3 is re-visited.
7. The weight vectors $\mathbf{w}_J^a$ and $\mathbf{w}^r$ are updated according to (5)–(6). The weight vector $\mathbf{w}_J^b$ of the winning node in ART_b is updated using (5) by replacing superscript or subscript a with b .
8. The F_2^a nodes with different class from that of the winning node J are identified. Width shrinking of the existing conflicting nodes is initiated by considering one of the three cases as in (7)–(9).

The training session is completed when all training patterns have been presented to the network. Otherwise, the training session is continued by presenting the next pattern to the network, and step 1 is re-visited.

III. LEARNING WITH IMBALANCED DATASETS

The class distributions of an imbalanced dataset are uneven whereby most of the data samples belong to the majority class and very few data samples belong to the minority classes. Learning from such an imbalanced dataset is a challenging task especially when the database is huge and complex. For a binary classification problem, many performance indicators have been introduced, and they are based on a confusion matrix as shown in Table I.

A popular metric in pattern classification is the overall accuracy, which is the *correctly classified rate* (CCR) as shown in (11). Indeed, the CCR is suitable for assessing the performance of a classification model with balanced datasets. A tendency of a typical classification model when dealing with imbalanced datasets is that the model can be biased to classify

TABLE I. CONFUSION MATRIX

	Predicted Positive (Label=0)	Predicted Negative (Label=1)
Actual Positive (Label=0)	True Positive (TP)	False Negative (FN)
Actual Negative (Label=1)	False Positive (FP)	True Negative (TN)

all data samples as belonging to the majority class. In this regard, the model is likely to classify data samples from the majority class well (i.e., very high TP) but unlikely to classify data samples from the minority class efficiently (i.e., very low TN). Under this circumstance, a high CCR value can be contributed by a very high value of TP and a very low value of TN. As such the CCR is not an appropriate measure for classification performance when dealing with imbalanced datasets.

$$CCR = \frac{TP + TN}{TP + FP + FN + TN} \quad (11)$$

From the literature, geometric mean, G-mean, is a more feasible performance indicator for imbalanced data classification. It is a nonlinear measurement, and is independent of the distribution of data samples among different classes [13]. G-mean evaluates the performance of a classification model in a fair way by considering the results of both the majority and minority classes. G-mean is expressed as follows:

$$G - \text{mean} = \sqrt{TPR \times TNR} \quad (12)$$

where

$$TPR = \frac{TP}{TP + FN}, \quad TNR = \frac{TN}{TN + FP} \quad (13)$$

TPR and TNR are the rates of TP and TN, respectively. TPR and TNP, respectively, represent the rates of the majority and minority classes being predicted correctly.

IV. CLASSIFICATION WITH A LARGE IMBALANCED DATASET

In the manufacturing industry, finished goods are normally produced in large quantities. The volume of production is a key performance index to indicate processing capacity and quality. In semiconductor as well as in other manufacturing industries, production growth from the mass production of one finished good is a target to achieve because it could bring a reduction in cost and also an increase in revenue. If the volume of production could be estimated by analyzing data from an upstream process with an intelligent system, then it could assist the production unit to identify impending yield or quality issues related to the final product. In this case, the undesirable variations from the upstream process could be identified to make improvements, and this helps to assure high product quality and efficient supply line management. Indeed, as pointed out in [14], classification is a useful way to help understand complex manufacturing processes.

A. Data Background

In this research, a case study involving the use of an encrypted dataset collected from a semiconductor company. The input and output features of the dataset are shown in Table II. The dataset comprises three input attributes where one of them is a categorical attribute whereas the other two are numerical attributes. The categorical attribute has a record of

215 values, ranging from 1A to 215A. Both numerical attributes have discrete values; the first has a range from 174610 to 299370 whereas the second from 174650 to 301770. The output class a binary number.

The dataset consists of 169910 samples, and they were divided into training and test sets. Table III summarizes the distributions of the data samples in the training and test sets. As can be seen from Table III, the data distribution in training and test sets was highly skewed. This is because in both training and test sets, the number of samples with label 0 is more than those with label 1. Indeed, the dataset exhibits a property of data imbalance, i.e., relative imbalance [15]. Thus, it was a challenging task to predict the samples from the minority class accurately.

B. Experimental Setups

Data normalization is required for use with FAM so as to mitigate the effect of category proliferation [8] during its learning process. In this case, the value of an attribute is normalized between 0 and 1 using

$$(x - \min(x)) / (\max(x) - \min(x)) \quad (14)$$

where $\min(x)$ and $\max(x)$ are respectively the lowest and the highest values of the attribute, x , in the dataset.

Data re-sampling techniques are useful for assisting a classification model in learning imbalanced datasets [15]. As such, an under sampling technique [6] was applied to select randomly a portion of training samples from the majority class in this experiment. The FAM-based networks were trained with the selected samples from the majority class together with all samples from the minority class. In this case, the numbers of samples selected from the majority class were respectively 10%, 30%, 50% and 100% of number of the samples from minority class (i.e., four ratios, namely 1:10, 3:10, 5:10 and 10:10).

The parameter settings of FAM-based networks are as follows. Original FAM used a fast training approach with $\beta = 1$; baseline vigilance of ART_a , $\bar{\rho}_a = 0.0$; choice parameter $\chi \approx 0$. As for FAMDDA, it was trained with the same parameter settings as original FAM in addition to a minimum width value, $\varepsilon_{\min} = 0.1$. In order to ascertain the stability of the classification performance, the experiment was repeated 10 times. The average results were computed using the bootstrap method [16], with the 95% confidence intervals estimated. The bootstrap method was employed because it did not rely on the assumption that the samples must be drawn from a normal distribution. In order to compare the average results between FAM and FAMDDA, a bootstrap hypothesis testing was conducted. In this case, the null hypothesis assumes the average results of FAM and FAMDDA were the same whereas the alternative hypothesis stated that the average result of FAM is lower than that of FAMDDA. The significance level α is set at 0.05. If the probability value (p -value) is lower than α , then the null hypothesis is rejected.

C. Results and Discussion

Fig. 2 compares graphically the G-mean results of the trained networks on the test set. Note that the networks were trained with samples from majority and minority classes in

TABLE II. AN OVERVIEW OF AN ENCRYPTED SEMICONDUCTOR DATASET

Dataset	Attribute 1 (categorical)	Attribute 2 (numerical)	Attribute 3 (numerical)	Output (numerical)
Training Set	1A to 215A	174610–261970	174650–278140	0/1
Test Set	1A to 215A	180800–299370	176670–301770	0/1
Interpolate Rate of Test Data (%)	100	44.66	93.06	100
Extrapolate Rate of Test Data (%)	0	55.34	6.94	0

TABLE III. CLASS DISTRIBUTIONS OF THE SAMPLES IN TRAINING AND TEST SETS

Dataset	Samples of the Majority Class	Samples of the Minority Class	Total Samples	Minority Rate (Minority/Total)
Training Set	44924	7469	52393	0.166
Test Set	108622	8895	117517	0.076

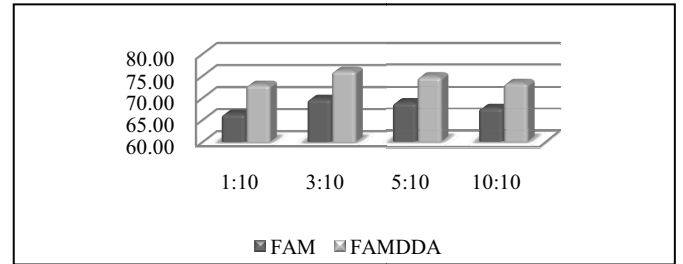


Figure 2. The G-mean results of the trained networks on the test set.

four different ratios. By observing Fig. 2, the FAM and FAMDDA, which were trained with a combination of majority and minority samples with a ratio 3:10, could give the highest G-mean results. Such classification results of the trained networks on the test set are presented in more details in Table IV. Even though 55% of the test samples recorded by attribute 2 (and a small portion of the test samples recorded by attribute 3) are extrapolated, both FAM and FAMDDA are able to classify the samples with good performances that exceed the extrapolation rates. In particular, the performances of FAM in terms of CCR, TPR, and G-mean are significantly lower than those from FAMDDA statistically (at the 95% confidence level). The performances of FAM and FAMDDA in terms of TNR are statistically similar. However, the latter is able to yield a slightly higher TNR by predicting more test samples from the minority class correctly. Overall, FAMDDA is able to outperform FAM in dealing with the imbalanced dataset from this problem.

In addition, despite the presence of a substantially large number of unseen samples that are different from those in the training set, the learned FAMDDA model is able to classify the unseen samples with very similar TPR (i.e., 77.57%) and TNR (i.e., 74.96%). This observation indicates that FAMDDA is able to predict equally well samples from the majority and minority classes, whereby it does not show a bias to the majority class that contains more samples.

TABLE IV. PERFORMANCE COMPARISON BETWEEN FAM AND FAMDDA

Metric (%)	FAM			FAMDDA			<i>p</i> -value
	Average	Highest	Lowest	Average	Highest	Lowest	
CCR	66.04	72.62	60.85	77.36	84.02	70.74	0.000
TPR	65.31	72.42	59.53	77.57	84.92	70.06	0.000
TNR	74.70	78.25	70.13	74.96	79.07	71.05	0.425
G-Mean	69.78	73.71	66.47	76.14	78.73	74.43	0.000

TABLE V. PERFORMANCE COMPARISON WITH OTHER CLASSIFICATION METHODS

Model	CRR	TPR	TNR	G-mean
KNN	0.91	0.95	0.43	0.64
SVM-RBF	0.75	0.74	0.83	0.79
FAM (best)	0.7262	0.7242	0.7503	0.7371
FAMDDA (best)	0.8402	0.8492	0.7300	0.7873

The classification performances of FAM and FAMDDA are also compared with those results of other methods, which include K-nearest neighborhood (KNN) algorithm and support vector machines with radial basis function kernels (SVM-RBF) [17]. In this regard, the best individual classification performances of FAM and FAMDDA, which are based on the highest G-mean, are listed in Table V for comparison. KNN represents an instance-based learning algorithm whereas SVM-RBF is a kernel-based model making use of selected instances (i.e., support vectors, for which the margins between their locations and the concept boundary (hyperplane) have been maximally optimized) for performing pattern classification. Both FAM and FAMDDA could give more *balanced* results in terms of TPR and TNR as compared to KNN and thus, they could contribute to higher G-mean than the latter. The classification results of FAM are not better than SVM-RBF in terms of G-mean and CRR. On the other hand, FAMDDA could give G-mean that is compatible with that of SVM-RBF and the former has a slight advantage in performing with a higher CRR. This finding indicates the potential of FAMDDA as a useful model for learning and classifying imbalanced dataset from the semiconductor industry.

V. SUMMARY

In this paper, a variant of FAM, which has been equipped with a conflict-resolving facility, i.e., fuzzy ARTMAP with Dynamic Decay Adjustment algorithm (FAMDDA), has been proposed for undertaking learning with imbalanced datasets. It is used together with a data re-sampling technique for learning and classifying a large imbalanced dataset collected from a semiconductor company. The results and analysis showed that the proposed FAMDDA network can identify the samples from both the minority and majority classes more accurately than FAM. A benchmark study was also conducted to check on the classification performances of FAM and FAMDDA with other methods. The finding indicated that FAMDDA could serve as one of the favorable classification models in handling the imbalanced dataset problem from the semiconductor industry.

As for future work, additional experimentation should be conducted to evaluate the applicability of the proposed FAMDDA model in dealing with imbalanced datasets from different application domains. This allows validation of FAMDDA as a useful classification model for imbalanced datasets from different application domains. Besides, a long-

term target in this research is to improve the classification performances, especially in terms of TPR and TNR. Hence, additional research effort will be devoted to devise more effective learning algorithms based on the existing framework for achieving this target.

ACKNOWLEDGMENT

The first author was a fellowship recipient of the Matsumae International Foundation (MIF) in 2010 for a six-month research visit to the Watada Lab. He is grateful to the MIF and the Watada Lab for providing financial supports and warm hospitality during his research stay in Japan.

REFERENCES

- [1] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Toward breast cancer survivability prediction models through improving training space," *Expert Systems with Applications*, vol. 36, pp. 12200–12209, 2009.
- [2] M. Merler, Y. Rong, and J.R. Smith, "Imbalanced rankboost for efficiently ranking large-scale image/video collections," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2607–2614, 2009.
- [3] P. Juszczak, N.M. Adams, D.J. Hand, C. Whitrow, and D.J. Weston, "Off-the-peg and bespoke classifiers for fraud detection," *Computational Statistics and Data Analysis*, vol. 52, pp. 4521–4532, 2008.
- [4] H. Wang and H. Zheng, "A bayesian approach to improving decision making in support vector machine and its application in bioinformatics," *Fifth International Conference on Natural Computation ICNC '09*, vol. 2, pp. 89–93, 2009.
- [5] Y.-Y. Nguwi and S.-Y. Cho, "An unsupervised self-organizing learning with support vector ranking for imbalanced datasets," *Expert Systems with Applications*, vol. 37, pp. 8303–8312, 2010.
- [6] N. Japkowicz and S. Stephen, "The class imbalance problem: a systematic study," *Intelligent Data Analysis*, vol. 6, pp. 429–449, 2002.
- [7] X. Hong, S. Chen and C.J. Harris, "A kernel-based two-class classifier for imbalanced data sets," *IEEE Transactions on Neural Networks*, vol. 18, pp. 28–41, 2007.
- [8] G.A. Carpenter, S. Grossberg, N. Markuzon, J. Reynolds, and D. Rosen, "Fuzzy ARTMAP: a neural network architecture for incremental learning of analog multidimensional maps," *IEEE Transactions on Neural Networks*, vol. 3, pp. 698–713, 1992.
- [9] G.A. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer Vision Graphics and Image Processing*, vol. 37, pp. 54–115, 1987.
- [10] L. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.
- [11] K.-P. Huber and M.R. Berthold, "Building precise classifiers with automatic rule extraction," *Proceedings of the IEEE International Conference on Neural Networks*, vol. 3, pp. 1263–1268, 1995.
- [12] C.P. Lim and R.F. Harrison, "An incremental adaptive network for on-line supervised learning and probability estimation," *Neural Networks*, vol. 10, pp. 925–939, 1997.
- [13] M. Kubat, R.C. Holte and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," vol. 30, pp. 195–215, 1998.
- [14] A. R. Chaturvedi, "Acquiring implicit knowledge in a complex domain," *Expert Systems with Applications*, vol. 6, pp. 23–35, 1993.
- [15] H. He and E.A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 1263–1284, 2009.
- [16] B. Efron, "Bootstrap methods: another look at the jackknife," *The Annals of Statistics*, vol. 7, pp. 1–26, 1979.
- [17] L. Ding, J. Watada, C.C. Lim, Z. Ibrahim, W.J. Lee, M. Khalid, "Building a svm-rbf-based classifier for forecasting large imbalanced dataset problem," *Proc Int'l Symposium on Intelligent Informatics ISI2010*, 2010.