



- (51) International Patent Classification:
G06N 3/00 (2006.01)
- (21) International Application Number:
PCT/JP2012/060391
- (22) International Filing Date:
11 April 2012 (11.04.2012)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
13/085,329 12 April 2011 (12.04.2011) US
- (71) Applicant (for all designated States except US): **SHARP KABUSHIKI KAISHA** [JP/JP]; 22-22, Nagaike-cho, Abeno-ku, Osaka-shi, Osaka, 5458522 (JP).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **WANG, Fan. XU, Xinyu. YUAN, Chang. VAN BEEK, Petrus J.L..**

- (74) Agent: **HARAKENZO WORLD PATENT & TRADE-MARK**; Daiwa Minamimorimachi Building, 2-6, Tenjin-bashi 2-chome Kita, Kita-ku, Osaka-shi, Osaka, 5300041 (JP).

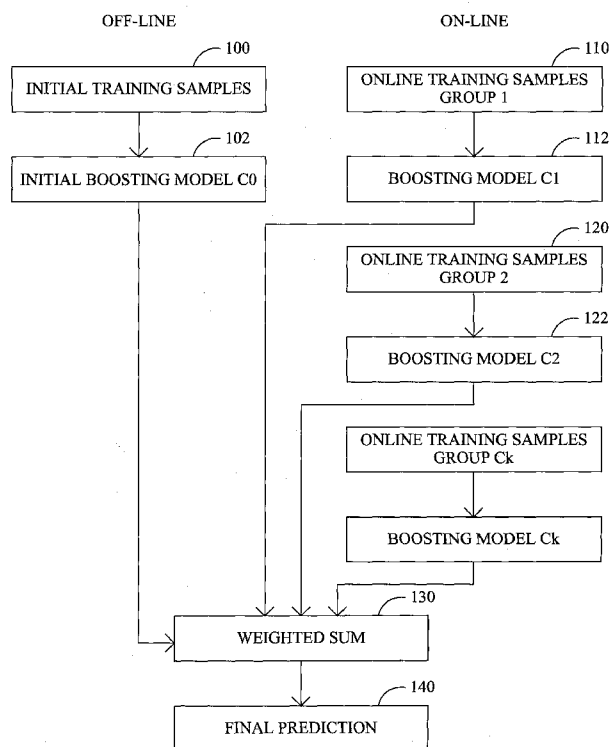
- (81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE,

[Continued on next page]

- (54) Title: SUPERVISED AND SEMI-SUPERVISED ONLINE BOOSTING ALGORITHM IN MACHINE LEARNING FRAMEWORK

FIG. 1



- (57) Abstract: A method for classification of samples comprising providing a trained statistical model based upon a set of initial samples. Receiving a set of first samples and training a first statistical model base upon the first set of samples, where the first statistical model is of the same class as the trained statistical model. Receiving a set of second samples and training a second statistical model base upon the second set of samples, where the second statistical model is of the same class as the trained statistical model. The trained statistical model, the first statistical model, and the second statistical model, being independent of each other and collectively used to classify another sample.



DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, **Published:**

LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE,

SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA,

GN, GQ, GW, ML, MR, NE, SN, TD, TG).

— *with international search report (Art. 21(3))*

DESCRIPTION

TITLE OF INVENTION: SUPERVISED AND SEMI-SUPERVISED
ONLINE BOOSTING ALGORITHM IN MACHINE LEARNING
FRAMEWORK

TECHNICAL FIELD

The present invention relates to online learning of a
classifier in a machine learning framework.

BACKGROUND ART

The present invention relates to online learning of a
classifier in a machine learning framework that includes
supervised and semi-supervised online learning algorithms in
a machine learning framework.

Such classifiers may be trained using training samples,
and then used in a so-called testing or prediction stage, to
classify test samples. For example, such classifiers may be
used in an automated factory inspection application to detect
defects in an image of a product. In this case, a "sample" may
consist of a set of data derived from an image or pixels in a
region of the image, and the task of the classifier is to classify
"sample" as "defect" (positive class) or "non-defect" (negative
class). As another example, such classifiers may be used to
classify defects into different categories.

- 2 -

Machine learning applications usually include two stages: training stage and prediction stage. In the training stage, traditionally all training samples are available as a batch at the beginning; a statistical model can then be trained based on the training samples. In the prediction stage, the statistical model obtained from the training stage is then used to classify new samples into different categories. However, in some machine learning tasks, not all the training samples are available at the initial training stage. More samples will be acquired and may be labeled as time goes on. It is desirable to use this new information to refine and improve the classifier for future usage. In some other applications, the data properties might be changing over time, or even not generated from any distribution. The model trained with the initial samples can only accommodate the initial properties, so it might become useless as new samples arrive over time.

One way of solving this problem would be re-training the model with all samples including the initial samples and the newly obtained samples. However, re-training from scratch will usually be time-consuming, and it's not efficient to perform the re-training frequently in an online application. Therefore, a mechanism is desirable such that the model can be updated by the newly obtained samples in an online fashion during the prediction stage without complete re-

training.

Lots of statistical models can be used as the classifier including Normal Support Vector Machines, Decision Trees, Boosted Decision Trees and Neural Networks. The Boosted
5 Decision Tree (or Boosting Tree) may be the statistical model of the classifier. Thus each of the types of statistical models may be generally referred to as a respective class.

Boosting is based on the use of an ensemble of weak classifiers that is not constrained to specific classifiers. In a
10 boosting algorithm, a weak classifier is trained with respect to the samples and the associated weights. At each iteration, a weak classifier is added to form a final strong classifier. The weak classifiers are typically weighted by their accuracy. After a weak classifier is added, the samples are reweighted: the
15 weights of the misclassified samples will be increased, and the samples that are classified correctly will have decreased weights. Weak classifiers that are subsequently added will be trained based on the re-weighted samples, focusing more on the misclassified samples.

20 The weak classifier may be in the form of a Decision Tree (DT). Decision Tree (DT) is a binary tree (i.e. tree where each non-leaf node has exactly 2 child nodes). The training and prediction of the Decision Tree (DT) is described as follows.

Training Decision Trees. The tree is built recursively,
25 starting from the root node. The whole training data (feature

vectors and the responses) are used to split the root node. In each node the optimum decision rule (i.e. the best "primary" split) is found based on some criteria (gini "purity" criteria is used for classification). Then, if necessary, the surrogate splits are found that resemble at the most the results of the primary split on the training data; all data are divided using the primary and the surrogate splits (just like it is done in the prediction procedure) between the left and the right child node. Then the procedure recursively splits both left and right nodes. At each node the recursive procedure may stop (i.e. stop splitting the node further). When the tree is built, it may be pruned using cross-validation procedure, if need. That is, some branches of the tree that may lead to the model overfitting are cut off. Normally, this procedure is only applied to standalone decision trees, while tree ensembles usually build small enough trees and use their own protection schemes against overfitting.

Predicting with DT: to reach a leaf node, and thus to obtain a response for the input feature vector, the prediction procedure starts with the root node. From each non-leaf node the procedure goes to the left, or to the right based on the value of a certain variable. If the value of the variable is less than the threshold. This pair is called split. Once a leaf node is reached, the value assigned to this node is used as the output of prediction procedure.

- 5 -

The foregoing and other objectives, features, and advantages of the invention may be more readily understood upon consideration of the following detailed description of the invention, taken in conjunction with the accompanying drawings.

SUMMARY OF INVENTION

One embodiment of the present invention discloses a method for classification of samples including:

- (a) providing a trained statistical model based upon a set of initial samples;
- (b) receiving a set of first samples;
- (c) training a first statistical model based upon said first set of samples, said first statistical model being of the same class as said trained statistical model;
- (d) receiving a set of second samples;
- (e) training a second statistical model based upon said second set of samples, said second statistical model being of the same class as said trained statistical model;
- (f) wherein said trained statistical model, said first statistical model, and said second statistical model, are independent of each other and collectively used to classify another sample.

The trained statistical model and the first statistical model are of the same class because they use the same

statistical model type. For example, the trained statistical model and the first statistical model can both be Decision Tree statistical model, or they can both be Support Vector Machine statistical model.

5 Step (a) is a step in which a sample classification apparatus trains an initial statistical model based on a set of initial samples. That is, the initial statistical model is automatically trained with a set of initial samples without user intervention. The initial statistical model can be Decision
10 Tree or Support Vector Machine (or any other machine learning methods).

One embodiment of the present invention discloses a method for classification of samples including:

- 15 (a) providing a trained statistical model based upon a set of initial samples;
- (b) receiving a first sample;
- (c) further training said initial statistical model based upon said first sample;
- 20 (d) wherein said further training of said initial statistical model processes said first sample as if it were already included in said trained statistical model.

One embodiment of the present invention discloses a method for classification of samples including:

- 25 (a) providing a first statistical model based upon a set of samples wherein its evaluation on positive samples has

- 7 -

high positive recall with low precision;

(b) providing a second statistical model based upon a set of samples wherein its evaluation on positive samples has high precision with low positive recall;

5 (c) classifying a sample based upon said first statistical model;

(d) selectively based upon said classifying said sample based upon said first statistical model further classifying said sample based upon said second statistical model.

10 The high positive recall means that the percentage of true positive samples found by the statistical model with respect to all the ground-truth positive samples is high.

One embodiment of the present invention discloses a method for classification of samples including:

15 (a) providing a trained statistical model based upon a set of initial labeled samples;

(b) receiving a set of first non-labeled samples;

(c) training a first statistical model based upon said first set of non-labeled samples, where said first statistical
20 model is of the same class as said trained statistical model, based upon predicting labels for said non-labeled samples and selecting said non-labeled samples based upon a confidence score.

- 8 -

BRIEF DESCRIPTION OF DRAWINGS

FIG. 1 illustrates an off-line and an on-line boosting technique.

FIG. 2 illustrates a cascade prediction for on-line boosting.

FIG. 3 illustrates an initial training model.

FIG. 4 illustrates on-line boosting with additional labeled samples.

FIG. 5 illustrates on-line boosting with additional unlabeled samples.

FIG. 6 illustrates on-line self-training.

FIG. 7 illustrates a more comprehensive boosting technique.

FIG. 8 illustrates semi-supervised prediction.

FIG. 9 illustrates an arrangement of a classification device in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The boosting technique may supplement the training on an image by image basis, or group of images by group of images basis, by expanding the model (statistical model). The statistical model is an algorithm (or a program) for classifying a sample into any of a plurality of predetermined categories.

Suppose with the batch of initial training samples, an ensemble of weak classifiers has already been obtained, then this ensemble can be expanded to a larger one as new training samples arrive. With the initial training samples, a model C_0 (a first statistical model), which is an ensemble of weak classifiers $L_i^0, i=1, \dots, M$, is trained. The prediction of C_0 on a testing sample x may be made by a weighted combination of

the prediction of each weak classifier: $C_0(x) = \sum_{i=1}^M w_i^0 L_i^0(x)$, in which $w_i^0, i=1, \dots, M$ are the weights of each classifier. The label of sample x may be given by $\text{sgn}[C_0(x)]$, which is the sign of the output.

In the online stage, the new training samples are collected and stored until there are a sufficient number of samples to effectively train another new model. The model trained by the i -th group of samples may be denoted as C_i , which is also an ensemble of N_i weak classifiers,

$C_i(x) = \sum_{j=1}^{N_i} w_j^i L_j^i(x)$. The final predicted label C can be made by a

weighted combination of all the models available up until the

current stage as follows $C = \sum_{i=0}^K \{\beta_i C_i(x)\} = \sum_{i=0}^K \left\{ \beta_i \left[\sum_{j=1}^{N_i} w_j^i L_j^i(x) \right] \right\}$, where

β_i is the weight for each model C_i .

Referring to FIG. 1, an exemplary framework for an on-line boosting technique described above is illustrated. An

- 10 -

initial set of samples 100 are trained to determine an initial boosting model (a trained statistical model) C0. As another group of images are obtained as a result of subsequent processing, a first group of additional training samples 110 are obtained to determine a first boosting model (a first statistical model) C1 112 based upon this group of images, and not the initial training samples 100 or initial boosting model C0 102. As a further group of images are obtained as a result of further subsequent processing, a second group of additional training samples 120 are obtained to determine a second boosting model (a second statistical model) C2 122 based upon this further group of images, and not the training samples 100/110 or boosting models C0/C1 102, 112. This process may be continued for additional groups of subsequent images. The result of the available boosting models C0, C1, ..., Ck may be combined using a weighted sum 130 to determine a final prediction 140.

Accordingly, all the models $C_i, i=0, \dots, K$ may be trained with disjointed sets of training samples to utilize the models in a flexible manner. The weights can be determined by the importance of each group of samples, by the order that each group of samples arrives, removal of selected models determined to be of limited value, or any other suitable manner. This technique may be referred to as a voting mechanism, with each model acting as an expert, with the

final results being determined by them together with different weights. The image buffer of the group of images may be cleared when the group is processed to determine a suitable model.

5 The on-line boosting technique may also update the existing model. In many cases, the optimal model that can be achieved for an online application would be the batch model re-trained with all the training samples available up until the current stage. To approach this goal, the online boosting
10 technique may correspond to the batch boosting technique. The model trained with the initial training samples may be updated with the newly added samples, so that the updated model would be approximately the same as the batch model re-trained with all training samples.

15 By way of example, the input technique may be a group of weak classifiers learned so far $L_i^0, i=1, \dots, M$, a new training sample s to arrive (with y_s as it's ground truth label), and an incremental learning technique which utilizes the new training sample to update the existing weak classifier
20 *BaseUpdate*(L_i^0, s). An exemplary set of details may be as follows.

Set the new sample's weight as $\lambda_s = 1$;

For each weak classifier $L_i^0, i=1, \dots, M$:

Generate k according to $Poisson\left(\frac{N\lambda_s}{W_i}\right)$;

Do k times: $L_i^0 = BaseUpdate(L_i^0, s)$;

- 12 -

Set $f_i(x) = \frac{1}{2} \log \frac{p_i(x)}{1-p_i(x)}$, in which $p_i(x) = P(y=1|s, L_i^0)$;

Update the new sample's weight: $\lambda_s \leftarrow \lambda_s \exp(-y_s f_i(s))$;

Update the summation of all samples' weights: $W_i = W_i + \lambda_s$;

The final prediction of a sample x can be made as

$$\text{sgn} \left[\sum_{i=1}^M f_i(x) \right].$$

W_i is the summation of the weights of all samples that have been used until the current stage (excluding the weight of new sample s), and N is the number of all past samples excluding the new sample s . When the new sample is misclassified by a weak classifier L_i^0 , $-y_s f_i(s)$ would be positive, so the weight λ_s associated with this sample is increased when presented to the next weak classifier; otherwise, the weight λ_s will be decreased. The principal idea of the online boosting technique is to process the new sample as if it were already included in the initial training set, i.e., also passing it from the first weak classifier to the last, and modifying the sample's weight at each step before passing to the next classifier. By using such a technique the updating technique does not require the samples to arrive in a group; instead, it can process the samples one by one.

One or more online boosting techniques, such as the expanding models technique and/or the updating the initial model technique may be combined together to achieve

improved performance. In the following, the expanded model is denoted as M_E and the updated model is denoted as M_U .

If the positive samples and negative samples are unbalanced, the two models will focus on different aspects of the classification. Without losing generality, it is assumed that the number of negative samples is much larger than the number of positive samples, and that finding positive samples is more valuable. Then two different kinds of performance evaluation on positive samples may be defined: (1) recall is defined as the percentage of true positive samples found by the model in all the ground-truth positive samples; (2) precision is defined as the percentage of true positive samples found by the model in all the positive samples predicted by the model. In other words, the true defects are classified properly but also include a lot of non-defects classified improperly. Also, in other words, a lot of the true defects are improperly classified as non-defects and only a few true defects are classified as defects.

Experimental results have shown that M_U will provide results with high positive recall but low precision, and M_E will provide results with high precision and low recall. To get both high positive recall and high precision, these two models may be combined together. It is desirable to make a prediction of a test sample using M_E and M_U separately, and take their weighted sum as the final prediction $\text{sgn}(M_E(x) + M_U(x))$.

As the online stage proceeds, M_E will be expanded, so that there will be more and more base weak classifiers in M_E . To make a prediction with M_E , all weak classifiers may be used to make the prediction, which might be time-consuming. Therefore, the use of a cascaded prediction may be used, which saves a lot of computational burden by avoiding using M_E in many unnecessary cases.

The joint technique is further described referring to FIG. 2. First M_U 200 is used to make a prediction 210. If the sample is predicted as negative 220, then the system may return -1 230 as the final output. In this case it is likely that the sample is a defect since M_U tends this way. If the sample is predicted as positive 240, then model M_E 250 is used to make another prediction 260. In this case, the prediction of M_U is not that accurate so it is preferable to also use M_E to obtain further characterization. The system then determines the weighted combination of the two predictions as the final positive output 270 or final negative output 280.

M_U tends to provide results with high recall, which means, most of the positive samples will be included in those samples predicted as positive by M_U . That is to say, if a sample is predicted as negative by M_U , it's most likely that it is a true negative. If it's predicted as positive, it is not sure whether it's true positive or not; then combine it with the prediction result by model M_E to make further confirmation.

- 15 -

The model M_E may be used primarily when a sample is predicted as positive in the first stage, which is a smaller fraction among all samples. Since the number of negative samples is larger than the number of positive samples, M_U will be enough to make the prediction in many cases, and the time required to go through all weak classifiers in M_E will be saved in many cases.

In the aforementioned techniques, the updating of the existing weak classifier $BaseUpdate(L_i^0, s)$ was described. The technique may be specific, if desired, to a particular type of baseline classifier. For example, if the decision tree is trained, and it is desirable to update it with a new sample, the system may first pass the sample from the root to corresponding branch according to the criteria at each internal node, and recalculate the “purity” score for each node. If the purity score of a node is not high enough, it should be re-split based on all previous samples and this new sample. Therefore, some information should be maintained about previous samples so that it’s possible to recalculate the purity score of each node. For variables with discrete value, this statistical information can be obtained by counting the number of samples with each value; however, if the variable is real-valued, to precisely maintain the distribution information, all the previous feature values that have appeared should be stored for future use.

In online learning, when using a decision tree as the base classifier in the boosting model, and the feature of each sample is real-valued, all the feature values that have appeared would normally be stored to update the initial model, which is determined by the characteristics of decision tree. This is highly memory-consuming as the number of samples that must be stored keeps increasing. A more compact representation of the sample features is desirable in a manner that does not result in a significant loss of accuracy to reduce memory usage.

In the initial stage, the system obtains the representative samples of initial training samples by using a suitable technique, such as a K-means clustering technique on positive and negative samples separately. The cluster centers are selected as the representative samples, denoted as $\{\hat{x}_i, i=1, \dots, n\}$, and their weights are taken as the number of samples in the corresponding cluster, denoted as $\{s_i, i=1, \dots, n\}$.

In the online stage, the new training samples are added to the previous representative sample set $\{\hat{x}_i, i=1, \dots, n\}$ with each new training sample assigned a weight of 1. A weighted k-means is performed on this enlarged set as follows:

(1) Randomly select n points as the initial cluster centers $\{center_i, i=1, \dots, n\}$;

(2) Assign each sample x to the nearest cluster center:

if $i = \arg \min_i \|x - center_i\|_2$, then $x \in C_i$.

- 17 -

(3) Re-calculate the cluster center by taking the weighted average of all samples in the cluster:

$$center'_i = \frac{\sum_{x_j \in C_i} s_j x_j}{\sum_{x_j \in C_i} s_j}, \text{ and the weight associated with the center}$$

is $s'_i = \sum_{x_j \in C_i} s_j$.

5 (4) Calculate the cluster error: $error = \sum_{i=1}^n \sum_{x_j \in C_i} s_j \|x_j - center'_i\|_2^2$;

(5) If the error is below some threshold, or doesn't change a lot compared with previous iteration, exit the algorithm, return the up-to-date cluster centers $\{center'_i, i=1, \dots, n\}$ and the associated weights $\{s'_i, i=1, \dots, n\}$ as the final compact representation of all samples;

10

(6) If the error doesn't converge, replace the old centers $\{center_i, i=1, \dots, n\}$ with new centers $\{center'_i, i=1, \dots, n\}$, and back to step 2.

Using this technique, the statistical distribution

15 information contained in the entire sample set is dynamically maintained, while the memory footprint and the computational load is controlled as constant, proportional to the number of centers used in the K-means clustering technique. Although the compact represented is described

20 with respect to updating decision trees, this representation is also useful for other types of classifiers, as long as the

classifier is influenced by the data density.

In supervised learning as previously described, training samples may be "labeled" when used to train a statistical model. The "labeled samples" means samples which are
5 classified into one of plurality of categories in advance by human being or by a preset statistical model. Typically, the labels of samples are provided by human being. Occasionally, an intelligent statistical model can be used to label the samples, however in this case, it is required that the labels
10 provided by the statistical model conform to the understanding of human being, that is the labels must be "meaningful" and "correct" to human being.

This means that the "ground-truth" class label of each sample is available at the time of training. For example, in a
15 defect detection context, a sample may be labeled as either "defect" or "non-defect" in order to "teach" a statistical model/classifier. The labels may be provided by a human operator or expert. In many applications, obtaining the labeled samples for training are expensive, difficult and/or
20 time-consuming. If a substantial number of unlabeled samples are available, these unlabeled samples may also be used to assist in improving the model, which reduces the burden associated with labeling the samples.

For applications where a limited number of labeled
25 samples are available, but a large number of unlabeled

samples are available, it may be desirable to use semi-supervised learning to exploit the useful information embedded in unlabeled samples. This means that samples that are not labeled, e.g. manually by an operator, may be used for training the statistical model of a classifier. This may be beneficial in many applications, since it may be relatively expensive (in terms of time or other resources) to provide labels for the purpose of training statistical classifiers.

In the online stage(s), if the incoming group of samples are labeled, the technique uses the supervised online boosting model described above, i.e. the initial model is updated and a new model is added. If the incoming group of samples are unlabeled, the system may use two models, namely, a newest added model and/or an updated initial model. Both models can predict the labels for these unlabeled samples, and the samples with the most confident labels are selected. Supposing the most confident labels are just the pseudo ground truth labels, each model will provide these samples and their labels to update the other model.

The samples used for training are usually labeled by experts; however, the labeling might be difficult, expensive or time-consuming, while the unlabeled samples are easier to collect. With limited amount of labeled samples, the obtained classifier might be biased or insufficient to distinguish different classes. Given the limited labeled samples, an

effective technique to train a classifier should exploit the information in the large amount of unlabeled samples, which is generally well-known as semi-supervised learning.

A semi-supervised online learning boosting tree framework may make use of the information contained in a large amount of unlabeled data. Their distribution gives some information about the intrinsic connections between data points, but this information is hard to exploit using traditional supervised learning techniques, since the samples do not have class labels associated with them.

A semi-supervised technique, generally referred to as co-training may be utilized. By way of general description there are two different classifiers M_0 and M_1 trained on labeled samples. Firstly M_0 is used to provide (predict) labels for the unlabeled samples. Then, predicted samples with the most confident labels are selected, removed from the unlabeled set, and added to the labeled set associated with the predicted labels. M_1 is re-trained with the expanded labeled set. Then, the role of M_0 and M_1 is changed, and the above steps are repeated. In this stage, M_0 is re-trained based on predictions from the updated M_1 . This co-training procedure may continue, until some convergence criteria is met.

Referring to FIG. 3, with the initial training samples 300, the system has an initial model M_U trained 310. In the subsequent stages, depending on the nature of the received

samples, one or more techniques may be applied.

Referring to FIG. 4, if the incoming group of samples 410, 420 is labeled, M_U 435 may be updated 430 with these samples. In this manner, the initial model is updated based upon the subsequent samples, and therefore reflects the total set of available samples. A new boosting model M_E^i 440 based upon a set of samples is added 450 to M_E 460. In this manner, new boosting models are determined based upon the labeled images each of which is characteristic of the newly received samples, which are then combined with previous newly determined boosting models. In this manner, the system is more suitable to adjust to changes in the characteristics of the samples.

Referring to FIG. 5, if the incoming group of samples is unlabeled 500, and the system has two models M_U and M_E , then a co-training procedure may be carried out between M_U and the most recently added model M_E^j in M_E . In general, the technique predicts 510 the labels for all of the unlabeled samples. Based upon the prediction, the system selects the samples with the most confident labels 520. The model M_U is updated 530 based upon the selected samples resulting in a set of updated models 540. Based upon the models 540 the system may predict the labels 550 for the unlabeled images 560. The prediction 550 may include the labels and a confidence measure (confidence score). The prediction value

of each tree $f_i(x) = \frac{1}{2} \log \frac{p_i(x)}{1-p_i(x)}$ (i denotes a Decision Tree)

recited above is regarded as the confidence score. Then the largest positive prediction value (C_pos) in each tree is found and also the smallest negative prediction value (C_neg) in each tree is found. Next, a confidence ratio as 0.6 is defined, and then the confident samples are selected if its prediction value is larger than 0.6*C_pos or its absolute prediction value is larger than 0.6*abs(C_neg).

The samples with the most confident labels 570 may be selected as likely true labels. The model M_E is updated based upon the selected samples resulting in a set of updated models 590. This process may be repeated for additional unlabeled samples.

Referring to FIG. 6, if the subsequent group of samples is unlabeled and the system only has model M_U , a self-training procedure may be carried out, which is similar to co-learning, with M_U providing the most confident samples to itself.

If the online samples are arriving one by one instead of group by group, the system may use a buffer to store the labeled samples until there are enough labeled samples to train a new model to be added. There is no need to hold another buffer for unlabeled samples, since updating the existing model and the co-training framework can handle the

samples one by one.

FIG. 7 illustrates the resulting technique from combining FIG. 4, FIG. 5, and FIG. 6.

In the prediction of semi-supervised online learning, the
5 initial model M_U and each of M_E^j produce a prediction label. Then these labels are combined by a weighted summation to generate the final predicted label. This prediction procedure is illustrated in FIG. 8.

FIG. 9 illustrates an arrangement of a classification
10 device 1 of the present embodiment. The classification device includes an input section 2, an information processing section 3, an output section 4, and a memory (storage section) 5 (see FIG. 9). The memory 5 stores plurality of types of statistical models (a trained statistical model based upon a
15 set of initial (labeled) samples, a first statistical model (the initial boosting model), and a second statistical model (the second boosting model)).

The information processing section 3 obtains an image of a product to be subjected to classification (a set of initial,
20 first, and second samples), and analyzes the image by use of the plurality of types of statistical models stored in the memory 5, so as to determine whether or not there is a defect in the product in the image. The information processing section 3 outputs a result of the determination via the output
25 section 4.

The information processing section 3 is a CPU or a dedicated arithmetic circuit. The output section 4 is, for example, a display device. The input section 2 is a member which has a function of obtaining the image. The input
5 section 2 is exemplified by a communication device which receives the image and a port which connects a storage medium.

The plurality of types of statistical models may be manually supplied to the classification device 1 (stored in the
10 memory 5) by a user.

The information processing section 3 trains a statistical model in accordance with the method as described earlier, so as to determine the product by use of the statistical model thus trained. The information processing section 3 may
15 determine the product by combining a plurality of statistical models.

The information processing section 3 may receive a set of labeled samples via the input section 2. The labeled samples are samples which are classified into one of a plurality of categories
20 in advance by a preset statistical model.

The following description discusses how to select samples in co-training.

- Predict the labels of unlabeled samples and select the most confident samples:

- 25 -

$$f_m(x) = \frac{1}{2} \log \frac{p_m(x)}{1 - p_m(x)}$$

- Each tree node value is defined as
- The summation of M trees' results:
 - Suppose trees are independent, the posteriori-

probability of the sample being positive is: $p(x) = \prod_{m=1}^M p_m(x)$

$$\sum_{m=1}^M f_m(x) \propto \frac{\prod_{m=1}^M p_m(x)}{\prod_{m=1}^M (1 - p_m(x))} \propto \frac{p(x)}{1 - p(x)}$$

5

- This poster-probability ratio can be regarded as a confidence value.

- Fixed number:
 - 5000 Samples with largest absolute predicted value;

10

- Confidence ratio:
 - Get the largest possible value(positive) _pos suppose in each tree the sample will be classified into the node with larger value, as well as the smallest possible value (negative) C_neg;

15

- Define a confidence ratio as 0.6, the confident samples are selected as:

- a positive sample with its label $C_x > 0.6 * C_{pos}$
- a negative sample with label C_x and $\text{abs}(C_x) > 0.6 * \text{abs}(C_{neg})$

20

- Max difference:
 - Selecting samples from classifier A for classifier B:
 - Classifier A and Classifier B have agreed on the

- 26 -

sign;

Classifier A predicts larger absolute value (more confident) than classifier B.

5 The terms and expressions which have been employed in the foregoing specification are used therein as terms of description and not of limitation, and there is no intention, in the use of such terms and expressions, of excluding
equivalents of the features shown and described or portions thereof, it being recognized that the scope of the invention is
10 defined and limited only by the claims which follow.

CLAIMS

1. A method for classification of samples comprising:

(a) providing a trained statistical model based upon a
5 set of initial samples;

(b) receiving a set of first samples;

(c) training a first statistical model based upon said
first set of samples, said first statistical model being of the
same class as said trained statistical model;

10 (d) receiving a set of second samples;

(e) training a second statistical model based upon said
second set of samples, said second statistical model being of
the same class as said trained statistical model;

(f) wherein said trained statistical model, said first
15 statistical model, and said second statistical model, are
independent of each other and collectively used to classify
another sample.

2. The method of claim 1 wherein said trained
20 statistical model is a boosting model.

3. A method for classification of samples comprising:

(a) providing a trained statistical model based upon a
set of initial samples;

25 (b) receiving a first sample;

- 28 -

(c) further training said initial statistical model based upon said first sample;

(d) wherein said further training of said initial statistical model processes said first sample as if it were already included in said trained statistical model.

4. The method of claim 3 wherein said processing is from a first weak classifier to a last weak classifier, and modifying a sample's weight at each step before passing to the next classifier.

5. A method for classification of samples comprising:

(a) providing a first statistical model based upon a set of samples wherein its evaluation on positive samples has high positive recall with low precision;

(b) providing a second statistical model based upon a set of samples wherein its evaluation on positive samples has high precision with low positive recall;

(c) classifying a sample based upon said first statistical model;

(d) selectively based upon said classifying said sample based upon said first statistical model further classifying said sample based upon said second statistical model.

6. The method of claim 5 wherein said classifying said

- 29 -

sample based upon said first statistical model includes a first output indicating it is likely a defect and a second output indicating it is likely not a defect.

5 7. The method of claim 6 wherein said selectively based upon said second statistic model is based upon said second output.

10 8. The method of claim 1 wherein said training said first statistical model is based upon a compact representation of said first set of samples.

15 9. The method of claim 1 wherein said training said further training said initial statistical model is based upon a compact representation of said first sample.

10. A method for classification of samples comprising:

(a) providing a trained statistical model based upon a set of initial labeled samples;

20 (b) receiving a set of first non-labeled samples;

(c) training a first statistical model based upon said first set of non-labeled samples, where said first statistical model is of the same class as said trained statistical model, based upon predicting labels for said non-labeled samples
25 and selecting said non-labeled samples based upon a

- 30 -

confidence score.

11. The method of claim 10 wherein said first on-labeled samples is more than two samples.

5

12. The method of claim 11 wherein said trained statistical model is a boosting model.

10

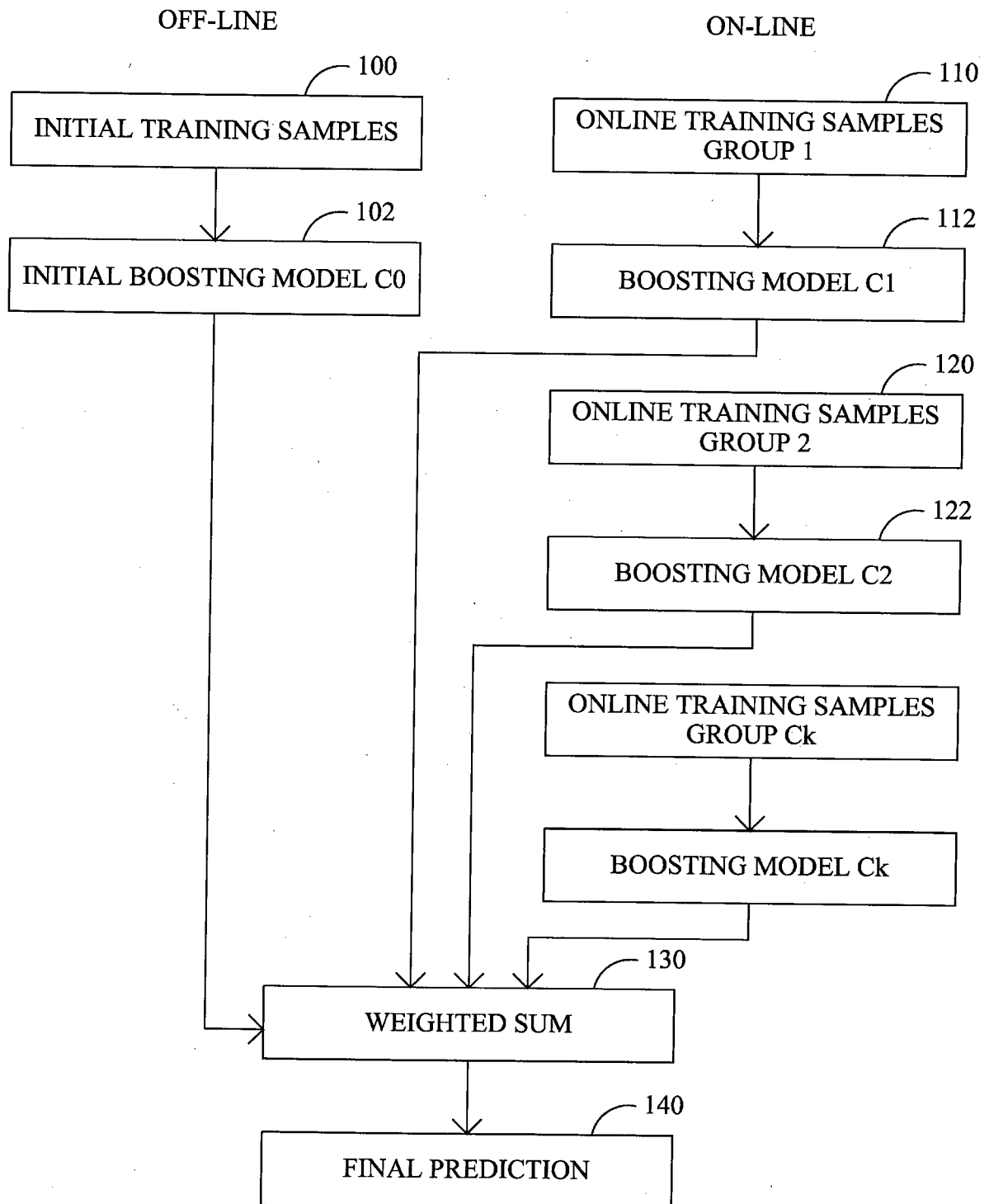
13. The method of claim 3 wherein said trained statistical model is a boosting model.

14. The method of claim 10 wherein said training is based upon at least two different statistical models, each of which are updated.

15

1/8

FIG. 1



2/8

FIG. 2

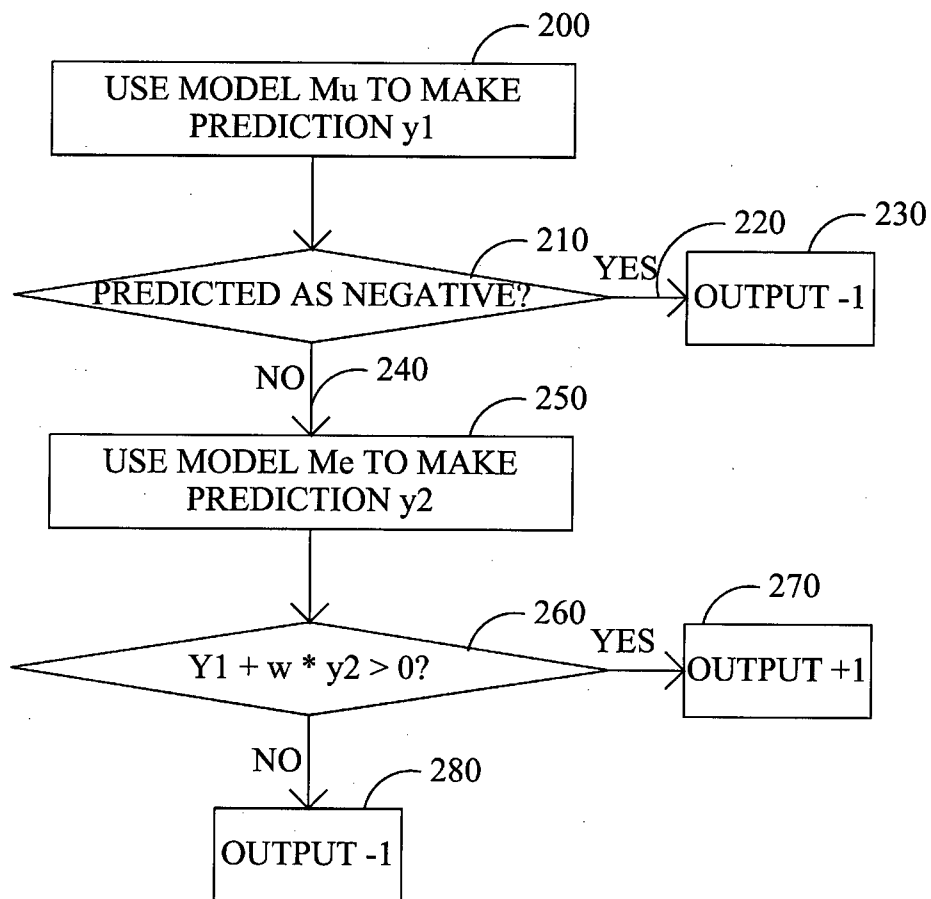
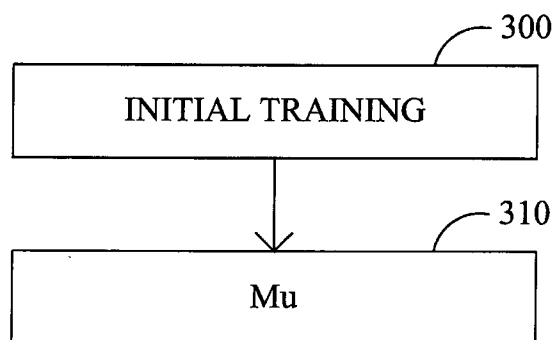
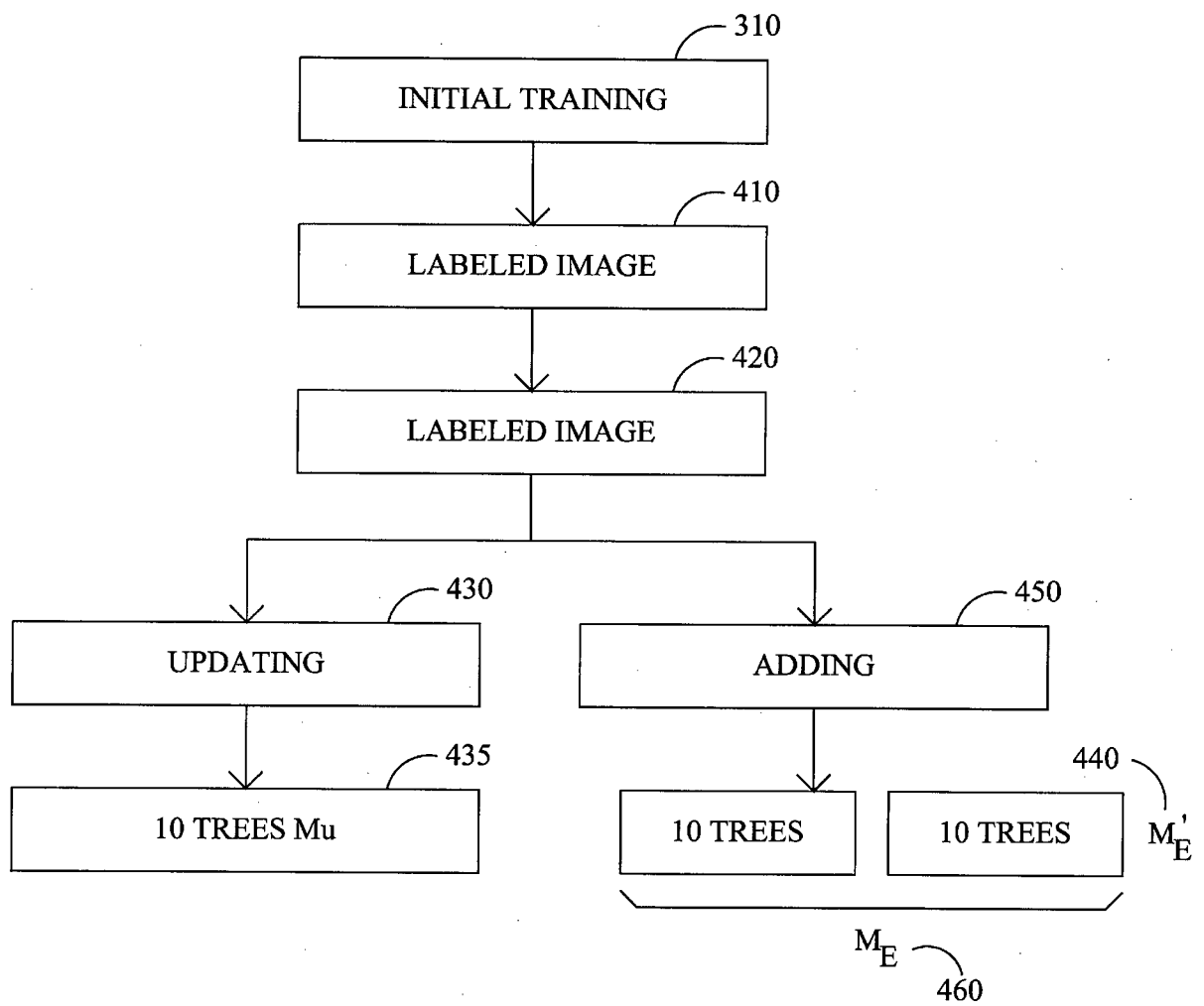


FIG. 3



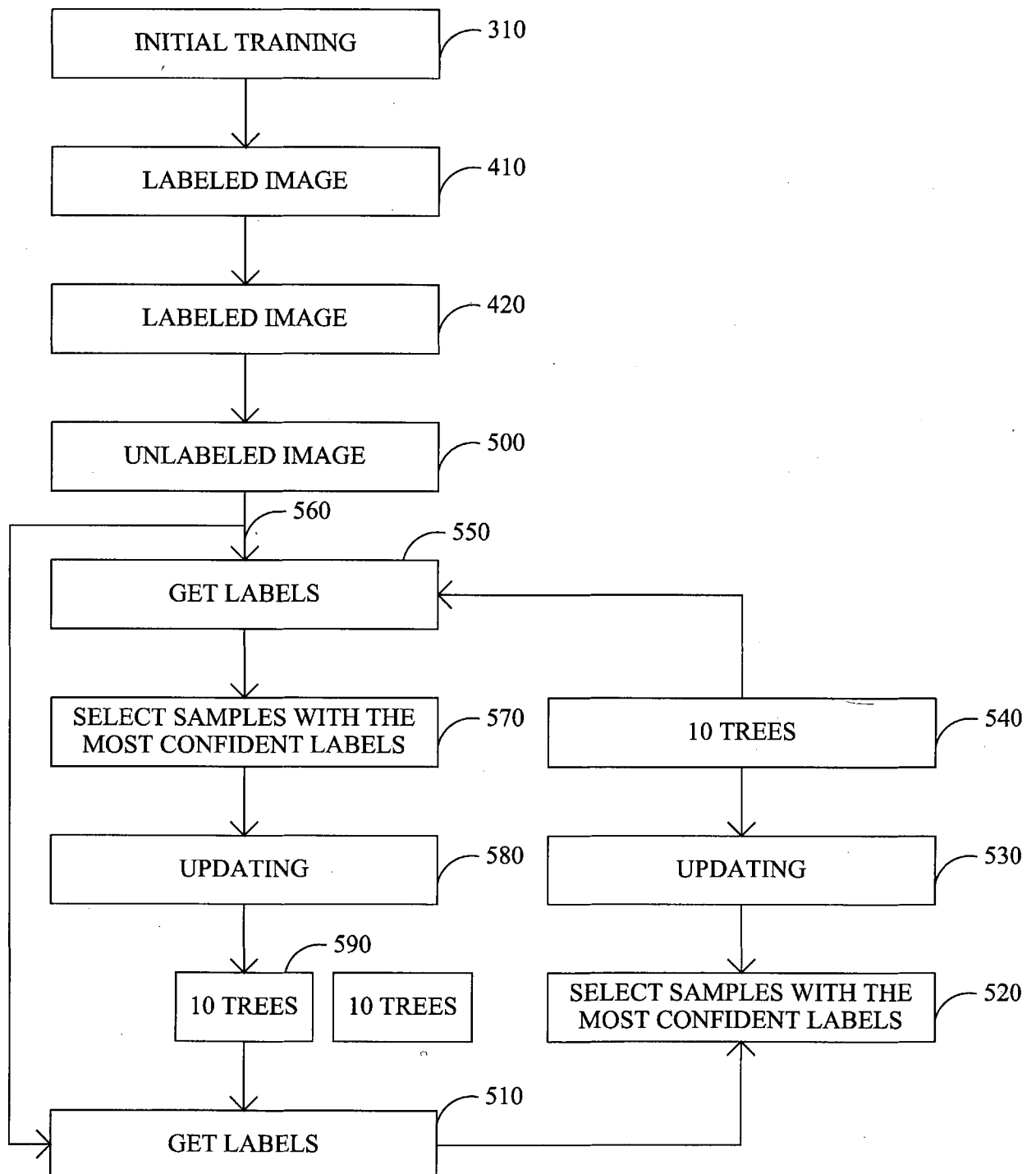
3/8

FIG. 4



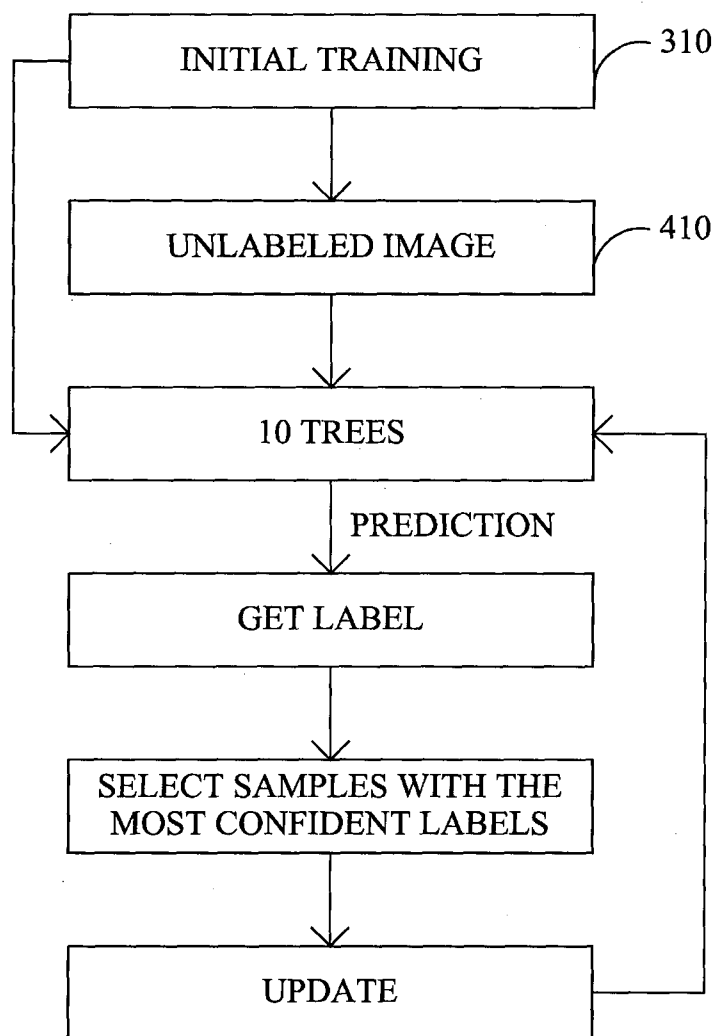
4/8

FIG. 5



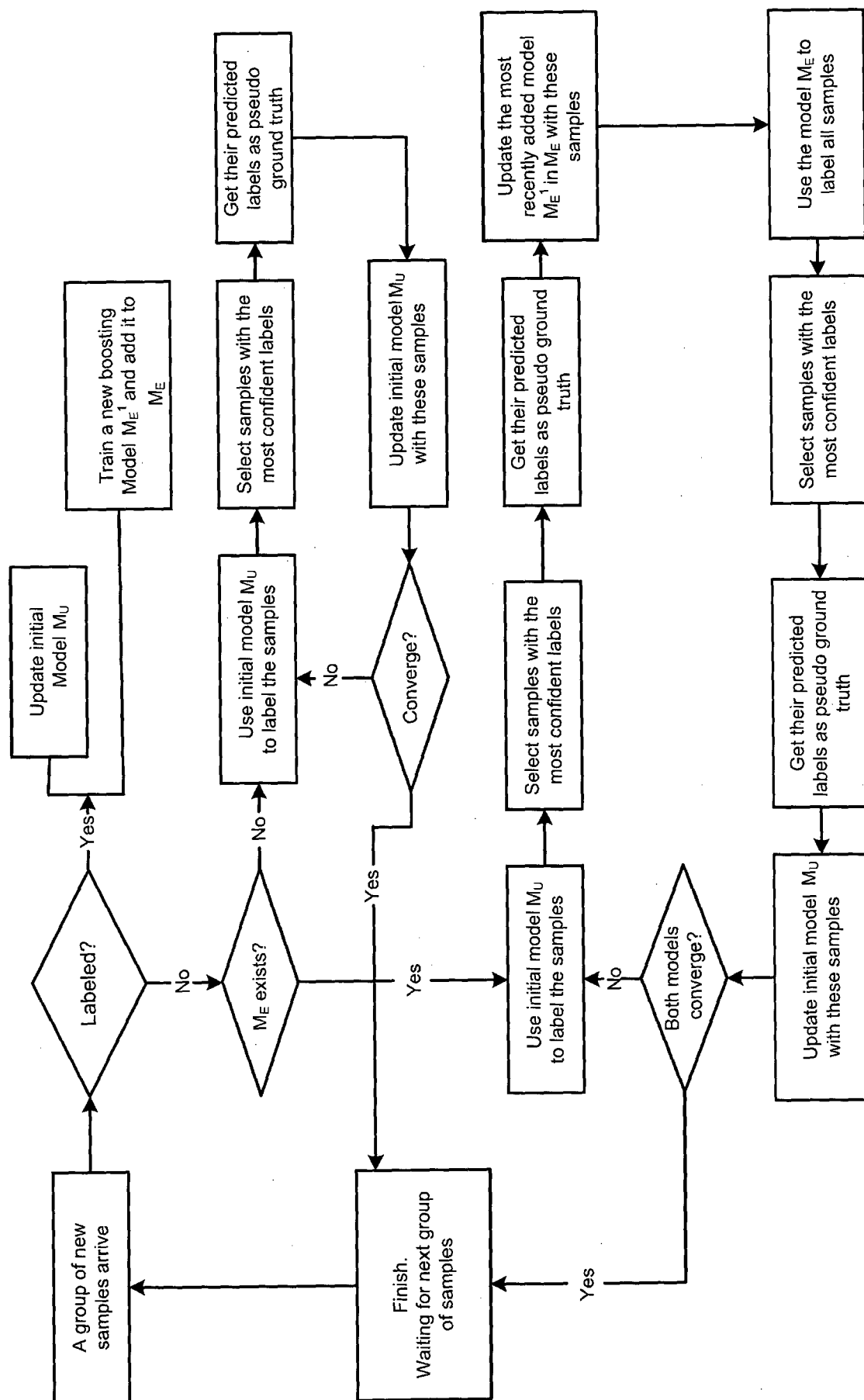
5/8

FIG. 6



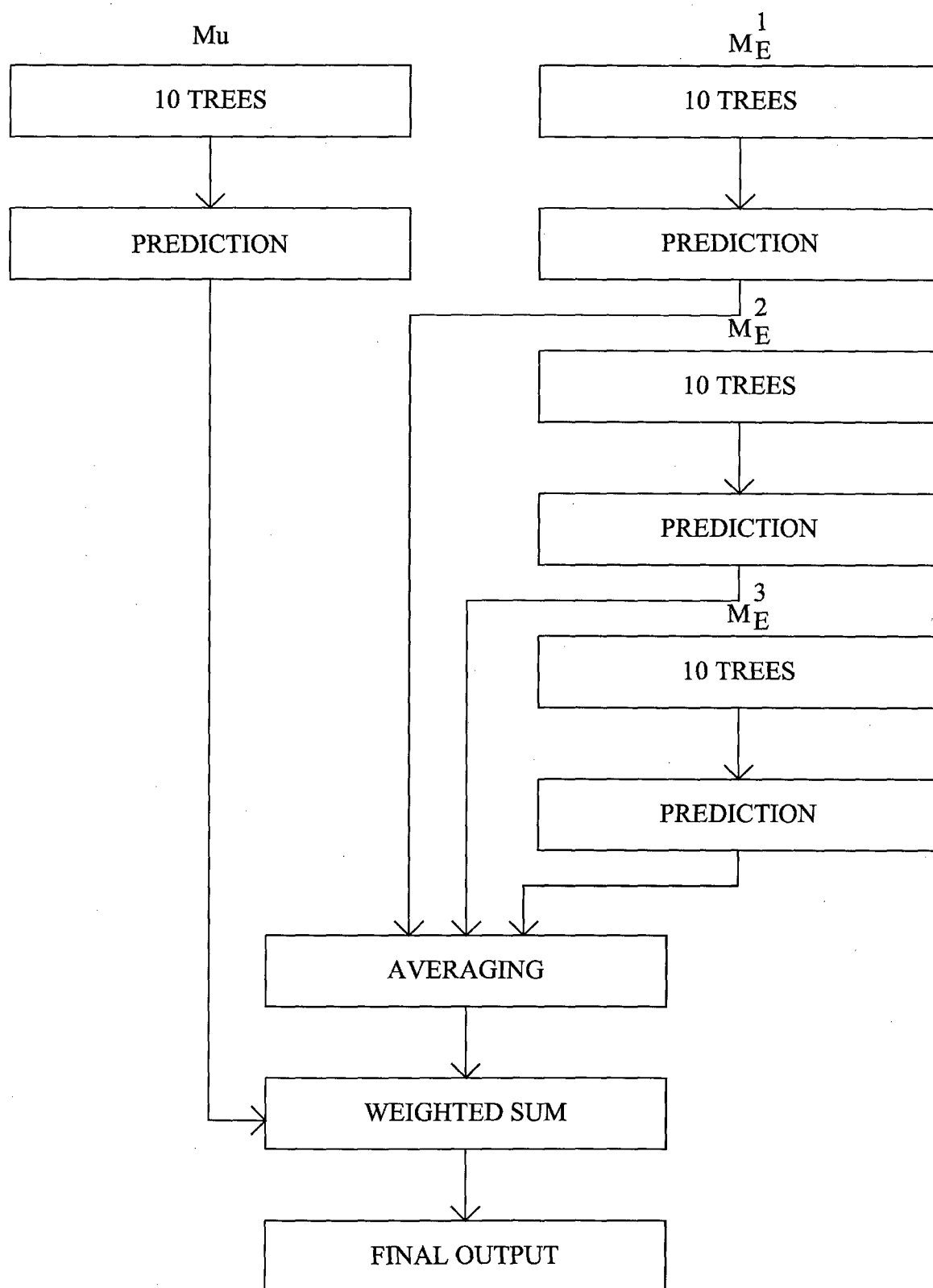
6/8

FIG. 7



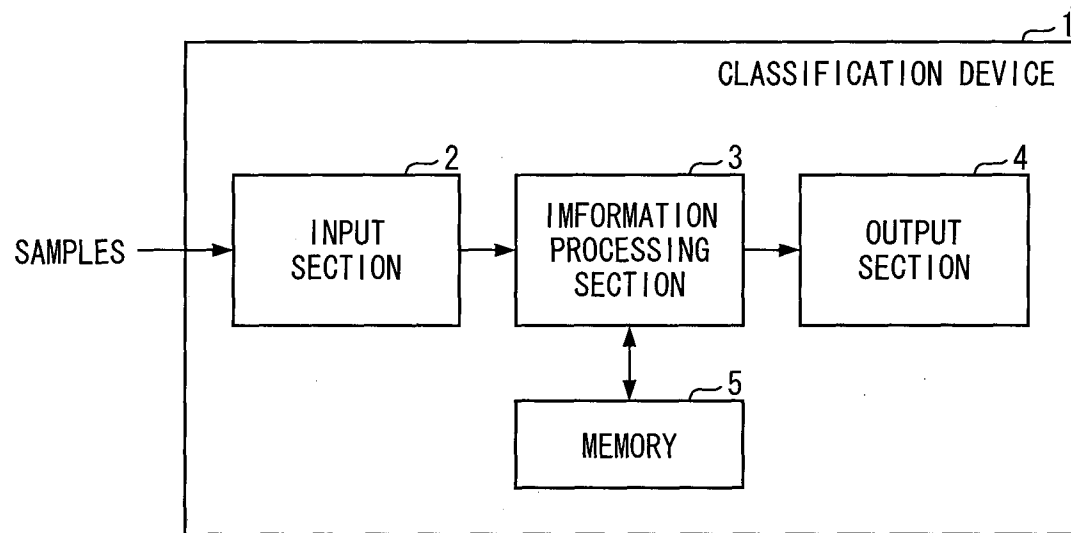
7/8

FIG. 8



8/8

FIG. 9



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/060391

A. CLASSIFICATION OF SUBJECT MATTER		
Int.Cl. G06N3/00 (2006.01) i		
According to International Patent Classification (IPC) or to both national classification and IPC		
B. FIELDS SEARCHED		
Minimum documentation searched (classification system followed by classification symbols)		
Int.Cl. G06N3/00		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Published examined utility model applications of Japan 1922-1996 Published unexamined utility model applications of Japan 1971-2012 Registered utility model specifications of Japan 1996-2012 Published registered utility model applications of Japan 1994-2012		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)		
IEEE Xplore JSTPlus (JDreamII)		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	R. Polikar et al., LEARN++: an incremental learning algorithm for multilayer perceptron networks, Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing 2000, vol.6, 2000, pp.3414 – 3417	1-2, 8
X	Takuya Kidera et al., An Incremental Learning Algorithm of Ensemble Classifier Systems, Proceedings of International Joint Conference on Neural Networks 2006, 2006, p.3423 section III. Paragraphs 1 and 2	1-2, 8
A	NAONORI UEDA, Ensemble Learning, IPSJ Transactions on Computer Vision and Image Media, 2005.10.15, Vol.46, pp.11-20	1-2, 8
☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.		
* Special categories of cited documents: "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search		Date of mailing of the international search report
29.06.2012		10.07.2012
Name and mailing address of the ISA/JP		Authorized officer
Japan Patent Office		HASEGAWA Atsuo
3-4-3, Kasumigaseki, Chiyoda-ku, Tokyo 100-8915, Japan		5B 3465
		Telephone No. +81-3-3581-1101 Ext. 3545

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2012/060391

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Claim 1 lacks novelty over D1 (R. Polikar et al., LEARN++: an incremental learning algorithm for multilayer perceptron networks, Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing 2000, vol.6, 2000, pp.3414 - 3417), and involves no special technical features. Thus, there are 4 inventions in the claims of this application.

- (Invention 1) claims 1-2,8
- (Invention 2) claims 3-4,9,13
- (Invention 3) claims 5-7
- (Invention 4) claims 10-12,14

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-2, 8

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.