*MEDINFO 2013*
*C.U. Lehmann et al. (Eds.)*


*doi:10.3233/978-1-61499-289-9-739*

# Unsupervised Medical Image Classification by Combining Case-Based Classifiers

**Thien Anh Dinh[a], Tomi Silander[a], Bolan Su[a], Tianxia Gong[c], Boon Chuan Pang[b], C. C. Tchoyoson Lim[b], Cheng Kiang Lee[b], Chew Lim Tan[a], Tze-Yun Leong[a]**

[a] *School of Computing, National University of Singapore, Singapore*
[b] *National Neuroscience Institute, Singapore*
[c] *Bioinformatics Institute, Singapore*

**Abstract**

*We introduce an automated pathology classification system for medical volumetric brain image slices. Existing work often relies on handcrafted features extracted from automatic image segmentation. This is not only a challenging and time-consuming process, but it may also limit the adaptability and robustness of the system. We propose a novel approach to combine sparse Gabor-feature based classifiers in an ensemble classification framework. The unsupervised nature of this non-parametric technique can significantly reduce the time and effort for system calibration. In particular, classification of medical images in this framework does not rely on segmentation, nor semantic-based or annotation-based feature selection. Our experiments show very promising results in classifying computer tomography image slices into pathological classes for traumatic brain injury patients.*

***Keywords:***

Medical images, image processing, traumatic brain injury, classification.

## Introduction

Large numbers of medical images are being generated every day. These images contain valuable information that might be useful for medical diagnosis, treatment, research, and education. Automatic image annotation attempts to extract symbolic knowledge from images to facilitate text-based retrieval of the relevant images pertaining specific abnormalities or diseases. Image-based computer aided diagnosis (CAD) systems can thus make use of such images and serve as a second source of opinion for clinicians on abnormality detection and pathological classification. The general task of automated knowledge extraction from medical image databases, however, remains a significant technological challenge.

In this paper we study automated knowledge extraction based on the images of traumatic brain injury (TBI). This is one of the major causes of death and disability in the world. Computed tomography (CT) images of the brain are widely used for the clinical diagnosis of TBI. Automatic classification of TBI brain images could help training radiologists assess clinical prognosis and pose content-based queries to image repositories. Recently, many image classification methods have been proposed to address this problem [1-3]. These methods usually consist of a feature selection phase in which the areas of interest are first detected from the image.Discriminative features of these segments are then extracted and selected for building a classification rule that is capable of labeling any image (or image segment) with one of the predefined labels.

Most current research has focused on inducing the classification rule from a finite sample of manually labeled training data. However, it is generally acknowledged that the quality of the segmented images and the relevance of the extracted features are critical factors in building a high quality classifier. Hence, constructing the classifier usually requires considerable amount of manual work. Automated segmentation is considered one of the most challenging tasks in medical imaging, and manually devising highly sensitive and accurate segmentation requires a lot of time and effort. The skills of a radiologist appear to be tacit knowledge, which makes it hard to explicate specific sets of useful features in analyzing different medical images. On the other hand, automatic feature selection is also non-trivial. To start with, it is hard to even establish a space of relevant feature candidates. Furthermore, the set of relevant features may well vary from case to case depending on context and the particular image, so that no single set of selected features is optimal for all images.

We propose a brain image classification architecture that does not rely on segmentation, nor semantic- or annotation-based feature selection. Consequently, the knowledge in the image databases can be extracted automatically and efficiently. The proposed system is case-based (or non-parametric). Thus, classification is performed directly on previously seen data without an intermediate modeling phase. Such methods are not new. However, recent advances in inducing sparsity have provided principled ways to regularize these methods. Recently, they have been successfully used for various classification problems such as robust face recognition [4].

Naturally, the images still have to be represented in some meaningful manner. For this purpose we deploy Gabor-filters that extract localized, low level features from the image. These "what and where"-features are known to resemble the primitive features extracted by the human visual cortex. However, with such a low level feature space, selection of a small set of significant features appears counterintuitive. We therefore borrow the idea of ensemble learning and form a collection of weak classifiers that specialize on different random subsets of the features. The final classification is then a product of many parallel classifiers operating on separate but possibly overlapping feature subspaces in a case-based manner.

Automated classification of TBI image slices has been studied before. Cosic and Longaric proposed a rule-based approach for classifying intracerebral hemorrhage (ICH) on CT brain images [1]. The rules were manually generated based on hematoma region information. Liao et al. [2] used a decision

tree-based approach to classify CT brain images into three hematoma types: extradural hematoma (EDH), subdural hematoma (SDH) and ICH. Gong et al. [3] proposed an unsupervised approach to label images with keywords extracted from their corresponding text reports, and used this "weakly" labeled training data to construct the TBI classification system. All these classifiers require pre-segmented image regions; thus the quality of image processing is the critical component. Furthermore, additional adjustments of the segmentation method are often needed when dealing with different datasets. This limits the practicality of the whole system. Manually crafting features is also a time-consuming process. Recently, Liu et al. [5] proposed an automatic Alzheimer's disease (AD) classification system that makes use of sparse classification based on spatially normalized tissue density maps. These maps still require segmentation of the gray matter, white matter and cerebrospinal volumes.

In our previous work [6], we have studied the classification of full stacks of volumetric TBI images using methods that require selecting handcrafted features and segmented regions. In this work, we concentrate on classifying single image slices. We propose to eliminate the segmentation process out of the system flow and embrace the unsupervised and autonomous Gabor feature extraction methods that aim at a more robust image classification system. By doing so, we enhance the practicality of the system by reducing the requirement of human expertise in selecting features and segmentation. To the best of our knowledge, this is the first attempt for TBI classification under these assumptions.

## Materials and Methods

In this section, we introduce a Gabor feature subspace ensemble classification framework, using sparse representation-based classifiers. To demonstrate its performance, the CT images commonly used for detection of TBI are evaluated . However, the theory makes no assumption about the specific neuroimaging modality.

### System Architecture

The workflow of the system is shown in Figure 1. First, the skull regions in TBI brain images are removed. The image is then normalized in terms of pose, direction and intensity. The images are further de-noised by reducing their resolution to 128 × 128 grayscale pixels. The image resolution is reduced for computational efficiency. Visually, the reduced resolution images are not very different from the original images.

Gabor features of the low resolution images are then extracted. The features are sampled randomly many times to form different feature subspaces that serve as input features for different "weak" classifiers. Finally, the results of all weak classifiers are combined to generate the final classification.

### Gabor Feature Extraction

Gabor filters have been widely used for image/texture recognition and detection [7]. The Gabor function extracts edge-like features in different scales and orientations at different locations of the image. Gabor features closely resembles the features extracted by the human visual system [8]. In particular, 2D Gabor filters have been used for texture segmentation, analysis, and recognition [9, 10].

In our work, we deploy 2D spatial Gabor filter [7] as a feature extraction mechanism. The filter $g: \Re^2 \to C$ is defined as a product of a Gaussian elliptical envelope with major/minor axis of γ/η and a complex sinusoidal plane wave that can be parameterized by an angle θ and central frequency $f_0$ by first rotating its plain coordinates *(x,y)* via

$$x' = x\cos\theta + y\sin\theta,$$
$$y' = -y\sin\theta + y\cos\theta,$$

and then expressing the filter as

$$g_{f_0,\theta}(x,y) = \frac{f_0^2}{\pi\gamma\eta} e^{-\left(\frac{f_0^2}{\gamma^2}x'^2 + \frac{f_0^2}{\eta^2}y'^2\right)} e^{i2\pi f_0 x'}. \quad (1)$$

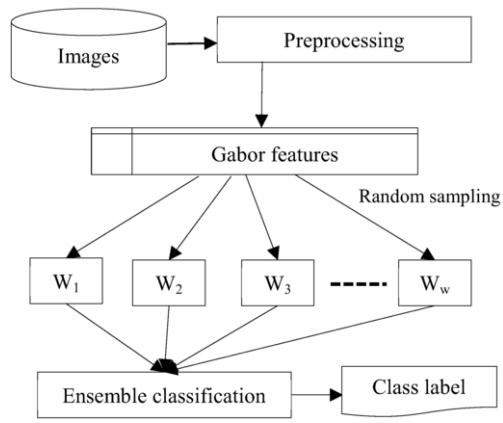



*Figure 1 - System architecture: The Gabor features of preprocessed images are randomly sampled to form input spaces for weak SRC classifiers $W_1,\ldots,W_W$, which are then combined to get the final class label.*

The Gabor feature *G* attached to the location $(x_0,y_0)$ of a grayscale image *img* is now defined by the magnitude of the convolution

$$G_{f_0,\theta}(x_0,y_0) = \left\| \sum_{(x,y)=(0,0)}^{(127,127)} g_{f_0,\theta}(x-x_0, y-y_0) img(x,y) \right\|. \quad (2)$$

We use the Matlab package [11] that can automatically adjust the sharpness parameters γ and η to create a Gabor filter bank with 5 frequencies and 8 orientations. Thus, each of the 16,384 locations in the image has 40 different features, resulting in a feature vector with 655,360 features. Performing classification directly on such long vectors is computationally demanding and carries the danger of overfitting.

One possible solution would be to select just a small number of discriminative features for constructing a classifier. However, as discussed in the introduction, pathological image classification is a skill that requires holistic tacit knowledge. Therefore, we opt for designing an ensemble of weak classifiers that together would accomplish the task.

In order to construct one weak classifier, we randomly select 4,000 Gabor features to form a feature subspace. We do this

sampling several times, and each random sampling defines a feature subspace for one weak classifier. In the experiments, we study the effect of the ensemble size by varying the number of weak classifiers from 5 to 100.

### Sparse Representation-Based Classifier

Recently, a sparse representation-based classifier (SRC) was proposed by Wright et al. [12]. SRC has achieved high performance and high robustness to noise upon classifying face images. Instead of using the sparsity to identify a relevant model or relevant features which can be used for classifying all test samples, the SRC aims at reconstructing the test image as a linear combination of a small number of training images. Classification is then done by evaluating how the images, belonging to the different classes, contribute to the reconstruction of the test image. Using the original training images for reconstruction, however, allows "lazy" classification without forcing us to induce a classification rule for all possible test images before they are encountered.

In our work, we propose to use this non-parametric sparse representation to construct the individual weak classifiers. More formally, suppose we have N training samples which belong to K classes. Each class $X^k$ consists of $N_i$ training vectors of length M, i.e., $X^k = [x_1^k, \dots, x_{N_k}^k] \in \Re^{M \times N_k}$, and $N = \sum_{k=1}^{K} N_k$. We can join the training data matrices to a one big block matrix $X = [X^1, \dots, X^K] \in \Re^{M \times N}$. The pseudo-code for this classification algorithm can be found in Algorithm 1.

**Algorithm** Sparse representation-based classifier (SRC)

**Input:**

- A block matrix $X = [X^1, \dots, X^K] \in \Re^{M \times N}$ of training samples.
- A test sample as a column vector $y \in \Re^M$.

**Program:**

1. Scale each column of *X* and the test sample *y* to have unit *L2*-norm.
2. Find a sparse coefficient vector π of length *N* by solving the *L1-norm* minimization problem:

   $$\pi = argmin_\alpha \|\alpha\|_1 subject\ to\ \|X\alpha - y\|_2 \le \varepsilon$$

3. For *k=1,…,K*:

   Evaluate the class specific reconstruction error $r_k(y) = \|X^k \pi^k - y\|_2$ using vector $\pi^k$ in which elements of $\pi$ that do not correspond to training vectors in class *k* have been set to zero.
4. **return** the residuals $r_k(y)$ for all classes *k*.

*Algorithm 1- The algorithm of sparse representation-based classifier*

The optimization problem in step 3 can be solved efficiently by L1-regularized sparse coding methods [13, 14].

### Ensemble of Weak Classifiers

Combining multiple weak classifiers often yields more accurate and robust classification than using individual classifiers [15]. The method of combining the results of individual classifiers isimportant. Many classifiers simply output the class label. In this case, it is common to decide the class label by majority voting. This simple method cannot account for the different confidences of weak classifiers.

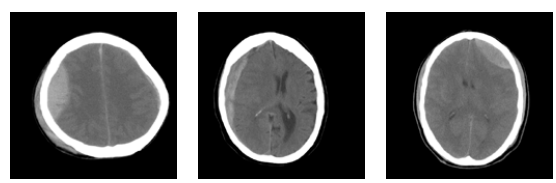
When using SRC, we can take the class specific residuals as confidence measurers. The better a test sample is, approximated by the sparse reconstruction of training samples belonging to a certain class, the smaller the corresponding residual. Instead of performing majority voting on class labels, we thus compute the average residuals of all weak classifiers for each class. The test sample *y* is assigned to the class with minimum average residual. Suppose that we have W weak classifiers. Let $r_k^w(y)$ be the residual for class *k* given by the classifier *w*. The average residual for a given class *k* will be calculated as $E_k(y) = \frac{1}{W}\sum_w^W r_k^w(y)$. The final label of a test *y* is defined as $Label(y) = argmin_k E_k(y)$.

### Experimental Materials

Data used for evaluation of our proposed method are taken from the database of the Neuroradiology Department in a tertiary referral hospital specializing in neurological diseases in Singapore. We obtained the Institutional Review Board approval for this anonymised dataset for retrospective data mining.

TBI brain damages are classified into several types mainly based on location, shape or size of hemorrhages (or hematoma). The commonly used types are subdural hematoma (SDH), extradural hematoma (EDH), intracerebral hemorrhage (ICH), subarachnoid hemorrhage (SAH) and intraventricular hematoma (IVH). External mechanical forces such as traffic accidents, violence and falls are usually the main reasons behind these injuries. Radiologists, trained specialists and senior residents usually categorize TBI by scrutinizing the CT images. The categorization is based on the doctor's previous experience and knowledge. It may be inaccurate due to the individual doctor's limited experience.

Our data set (Figure 2) consists of images featuring three types of TBI: EDH (24 patients), ICH (21 patients) and SDH (58 patients). Each case is in the form of a volumetric stack consisting of 18-30 slices. There are a total of 531 slices exhibiting SDH, 165 slices with EDH and 151 slices with ICH.



(a) (b) (c)

*Figure 2 - Images (a) and (c) are examples of extradural hematoma (EDH), while image (b) features subdural hematoma (SDH)*

Each image is 512 × 512 pixels of 8-bit grayscale. The dataset is manually labeled with TBI subtype and verified with findings from medical text reports. In this work, we only focus on detection of abnormal slices, thus slices without abnormalities are removed from the dataset. The number of slices without any abnormalities is different from case to case. It could vary from 4 to 15 slices per patient. We assume that each remaining slice only exhibits a single type of hematoma.

### Experimental Setup

We used stratified ten-fold, cross-validation to evaluate the performance of our system. We ran validation 50 times with different random folding and measured the average precisions and recalls. Although the system works at the slice level, the

training and testing dataset should be separated at the case level to avoid slices from the same case being both in training and test data. Since there are almost three times more SDH cases than EDH and ICH cases, our dataset is skewed. It is known that imbalance in training data may considerably reduce the accuracy of a classification system [16]. Although we could equalize the number of samples for each class, we chose to follow the real class distribution for the sake of more relevant evaluation.

The main components of our classifier architecture are the sparse case-based SRC method that abolishes the need for segmentation; and the ensemble of weak classifiers that abolish the need for semantic- or annotation-based feature selection. To evaluate these design choices, we compared its performance to other classifiers that differ in these aspects. The baseline classifier is the standard support vector machine (SVM) classifier, one of the most popular and successful classification techniques in machine learning [17]. Like SRC, the SVM is non-parametric and does not need feature selection for regularization. Therefore, it could potentially yield similar benefits as our method. To study the aspect of feature selection, we have also constructed an SVM+FS classifier that first selects the original Gabor-based features and then uses the SVM algorithm. Features are ranked according to their variance in the dataset: the top 1,000 features with largest variance serve as an input space for the SVM. To study the role of combining weak classifiers, we constructed an Ensemble+SVM method. Using an ensemble architecture similar to our proposed approach, instead of employing SRC as a weak classifier, we use the standard SVM. All the SVM classifiers use weighted SVM to deal with imbalanced data. We have also included a plain SRC method that does not use weak classifiers but, like the plain SVM, directly uses all the 655,360 features.

## Results

In general, our Ensemble+SRC method yields promising results in classifying abnormal TBI slices (Table 1). The performance in the majority class SDH is better than in EDH and ICH. This might be explained by a limited number of training samples in these two classes compared to SDH. The proposed framework improves precision and recall compared to other classifiers. The regularization of the SRC appears to suit this domain better than the maximum margin principle used by SVM. Even without feature selection, the SRC generally outperforms the SVM. Combining SVM with feature selection or using it as a weak classifier does not appear to bring about the desired results. SVM+FS and Ensemble+SVM do not perform very well in EDH and ICH classes. We conclude that combining SRC and ensemble learning is a good classifier architecture for this domain.

*Table 1- Average precision and recall for different methods. The standard deviations over several foldings listed in parenthesis.*

| | SDH | | EDH | | ICH | |
|---|---|---|---|---|---|---|
| | Prec. | Recall | Prec. | Recall | Prec. | Recall |
| SVM | 76% (±1%) | 56% (±1%) | 42% (±4%) | 63% (±4%) | 43% (±2%) | 61% (±3%) |
| SVM + FS | 76% (±2%) | 57% (±1%) | 57% (±2%) | 42% (±3%) | 64% (±2%) | 42% (±3%) |
| Ens. + SVM | 76% (±2%) | 56% (±2%) | 43% (±2%) | 64% (±2%) | 43% (±2%) | 60% (±3%) |
| **SRC** | 77% (±1%) | 71% (±2%) | 45% (±2%) | **65%** (±3%) | 52% (±3%) | 44% (±4%) |
| **Ens. + SRC** | **84%** (±2%) | **80%** (±2%) | **64%** (±3%) | 60% (±4%) | **71%** (±4%) | **65%** (±1%) |

There are two parameters that might have important influence on the overall performance of our system. They are the number of weak classifiers in the ensemble, and the number of features sampled from the original Gabor features. We will next investigate the effects of these parameters on classification results.

Table 2 illustrates the performance of the system with different numbers of weak classifiers. All cases are given 1,000 randomly selected features. The number of classifiers appears to affect both the precision and recall. In particular, recruiting more classifiers helps in classifying the smaller EDH and ICH subtypes.

*Table 2 - Average precision and recall of classifiers when varying the ensemble size and fixing the number of features at 1,000*

| | SDH | | EDH | | ICH | |
|---|---|---|---|---|---|---|
| | Prec. | Recall | Prec. | Recall | Prec. | Recall |
| Ens.+ SRC 5 classifiers | 84% (±3%) | 60% (±8%) | 63% (±7%) | 56% (±6%) | 51% (±7%) | 65% (±7%) |
| Ens.+ SRC 50 classifiers | 82% (±2%) | 66% (±5%) | 63% (±6%) | **57%** (±4%) | **58%** (±5%) | **66%** (±5%) |
| Ens. + SRC 100 classifiers | **83%** (±1%) | **65%** (±3%) | **65%** (±6%) | 55% (±3%) | 57% (±4%) | 61% (±6%) |

The number of sampled features appears to have a reasonably small effect. When using 50 weak classifiers (Table 3), increasing the number of sampled features from 1,000 to 2,000 does not appear to improve the results. When conducting the experiments, we observed that when the number of features are increased, the ensemble size should be increased as well to avoid overfitting.

*Table 3 - Average precisions and recalls of classifiers when varying number of features and fixing the number of classifiers at 50*

| | SDH | | EDH | | ICH | |
|---|---|---|---|---|---|---|
| | Prec. | Recall | Prec. | Recall | Prec. | Recall |
| Ens.+ SRC 500 features | 80% (±3%) | 62% (±3%) | 60% (±6%) | **60%** (±5%) | 53% (±3%) | 65% (±4%) |
| Ens.+ SRC 1000 features | 82% (±2%) | **66%** (±5%) | **63%** (±6%) | 57% (±4%) | **58%** (±5%) | **66%** (±5%) |
| Ens. + SRC 2000 features | **83%** (±2%) | 64% (±4%) | 63% (±6%) | 55% (±6%) | 57% (±4%) | 61% (±8%) |

To further assess our proposed framework, we compared our results with one of the most recent methods in classifying TBI subtype. Gong et al. [3] proposed to use SVM classification on selected features extracted from segmentation. The comparison is possible since our dataset originates from the same source. However, only one particular folding is available to us for the comparison. Table 4 illustrates the results of the two methods. Although Gong et al.'s method yields better results, it takes more effort to manually craft the features to be extracted. In addition, this technique depends on the reliability of the segmentation technique involved. We believe that our method has more potential in terms of adaptability and automation.

*Table 4- Comparison of the proposed framework to Gong et al.'s [3]*

| | SDH | | EDH | | ICH | |
|---|---|---|---|---|---|---|
| | Prec. | Recall | Prec. | Recall | Prec. | Recall |
| Ens.+SRC | 84% | 80% | 64% | 60% | 71% | 65% |
| Gong et al [3] | 80.6% | 87.9% | 82.9% | 79.1% | 83.3% | 78.9% |

## Conclusion and Future Work

We have introduced an ensemble classification framework with sparse Gabor-feature based classifiers. The system does not require segmentation nor supervised feature selection. This reduces the need for manual work in extracting useful information from medical images. Our experiments show that in the domain of classifying TBI images, we achieve reasonable results as compared to segmentation-dependent techniques that rely on manually selected handcrafted features. The proposed approach does not make any modality dependent assumptions. Testing on other modalities is a natural extension for future work.

The proposed method is non-parametric, thus when more training data is used, the classification will take a longer time. Parallel computation or compression of the training data could help to improve this situation. The dilemma of imbalanced data has also not been addressed adequately in our current system. Weights for each class are still manually assigned to avoid the dominance of large classes. We feel that these two issues are related, and devising a solution while still maintaining the case-based distributed representation, is one of the main foci for future studies.

Currently, the proposed system classifies single image slices. In future, we would like to investigate the problem of classifying regions of interest, possibly utilizing the slice level classification. In addition, unsupervised features extracted from text report could be used to enhance the classifier. Finally, integration with existing medical image retrieval system is needed to assess the practical value of the system.

**Acknowledgments**

This research was partially supported by Academic Research Grants no. T1251RES1211 and MOE2010-T2-2-111 and a research scholarship from the Ministry of Education, Singapore. Part of this work was done when the first author was supported by a Singapore Millennium Fellowship.

## References

[1] Ćosić D, Lončarić S. Rule-based labeling of CT head image. In: Artificial intelligence in medicine. Berlin / Heidelberg: Springer; 1997. p. 453-6.

[2] Liao CC, Xiao F, Wong JM, Chiang IJ. A knowledge discovery approach to diagnosing intracranial hematomas on brain CT: recognition, measurement and classification. Medical Biometrics. Berlin / Heidelberg: Springer; 2007. p. 73-82.

[3] Gong T, Li S, Wang J, Tan CL, Pang BC, Lim CCT, Lee CK, Tian Q, Zhang Z. Automatic labeling and classification of brain CT images. 18th IEEE Int Conf on Img Proc (ICIP). 2011; 2011:1581-1584.

[4] Huang K, Aviyente S. Sparse representation for signal classification. Adv in Neur Inf Proc Syst. 2007; 19:609.

[5] Liu M, Zhang D, Shen D. Ensemble sparse classification of Alzheimer's disease. Neuroimage. 2012; 60(2):1106-1116.

[6] Dinh TA, Silander T, Lim CC, Leong TY. An automated pathological class level annotation system for volumetric brain images. AMIA Ann Symp Proc. 2012; 2012:1201-1210.

[7] Jain A, Healey G. A multiscale representation including opponent color features for texture recognition. IEEE Trans on Img Proc.1998; 7(1):124-8.

[8] Clausi DA, Jernigan M. Designing Gabor filters for optimal texture separability. Patt Recog. 2000; 33(11):1835-49.

[9] Bovik AC, Clark M, Geisler WS. Multichannel texture analysis using localized spatial filters. IEEE Trans on Patt Anal and Mach Intel. 1990; 12(1):55-73.

[10] Jain AK, Farrokhnia F. Unsupervised texture segmentation using Gabor filters. Patt Recog. 1991; 24(12):1167-86.

[11] Ilonen J, Joni-Kristian, Kämäräinen, Kälviäinen H. Efficient computation of Gabor features. Research Report. Lappeenranta University of Technology, Department of Information Technology. 2005.

[12] Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. IEEE Trans on Patt Anal and Mach Intel. 2009; 31(2):210-27.

[13] Boyd S, Vandenberghe L. Convex optimization. Cambridge University Press; 2004.

[14] Kim SJ, Koh K, Lustig M, Boyd S, Gorinevsky D. An interior-point method for large-scale L1-regularized least squares. IEEE J Sel Top in Sig Proc. 2007; 1(4):606-17.

[15] Dietterich TG. Ensemble methods in machine learning. In: Multiple Classifier Systems. Berlin / Heidelberg: Springer; 2000. p.1-15 .

[16] He H, Garcia EA. Learning from imbalanced data. IEEE Trans on Knowl and Data Eng. 2009; 21(9):1263-84.

[17] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer; 2009.

**Address for correspondence**

Thien Anh Dinh (thienanh@comp.nus.edu.sg)
School of Computing, National University of Singapore
13 Computing Drive, Singapore 117417.