Chapter 7
Kernel Machines for Imbalanced Data Problem in Biomedical Applications

Peng Li, Kap Luk Chan, Sheng Fu, and Shankar M. Krishnan

Abstract Kernel machines such as the support vector machines (SVMs) have been reported to perform well in many applications. However, the performance of a binary SVM can be adversely affected by an imbalanced set of training samples, known as the imbalanced data problem. One-class SVMs, as a recognition-based approach, can be used to train and recognize the majority class and such kernel machines have already been developed. In this chapter, we review and study the effects of imbalanced datasets on the performance of both one-class SVMs and binary SVMs. We show that a hybrid kernel machine comprising one-class SVMs and binary SVMs in a multi-classifier system alleviates the imbalanced data problem. We also report the deployment of such hybrid kernel machines in two biomedical applications where the imbalanced data problem exists.

The research presented in this chapter was carried out when all authors were with the Nanyang Technological University, Singapore.

P. Li
University of Bristol, Bristol, UK
e-mail: lipeng@ieee.org

K.L. Chan (✉)
Nanyang Technological University, Singapore
e-mail: eklchan@ntu.edu.sg

S. Fu
KK Hospital, Singapore
e-mail: fu.sheng@kkh.com.sg

S.M. Krishnan
Wentworth Institute of Technology, Boston, USA
e-mail: krishnans@wit.edu

Y. Ma and G. Guo (eds.), *Support Vector Machines Applications*,
DOI 10.1007/978-3-319-02300-7_7,

7.1 Introduction

Kernel machines are algorithms in which kernels are employed to conceptually map data from an input space into a higher-dimensional feature space where the data can be processed using linear methods. The mapping is usually nonlinear and is implemented implicitly through the kernel trick. Many kernel methods have been developed by the machine learning community, such as support vector machines (SVMs) [51], kernel-based principal component analysis (KPCA) [36], kernel-based linear discriminant analysis (KLDA) [35], kernel-based independent component analysis (KICA) [2] and kernel-based nearest neighbour classifier [38].

SVM, a most widely used kernel machine, was originally developed for two-class classification. Based on the principle of structural risk minimization, discriminative binary SVMs, referred to as Binary Support Vector Classifier (*BSVC*) in this chapter, have been reported to perform well in many real applications [9, 12, 37]. However, SVM also suffers from some fundamental problems in statistical pattern recognition, such as the imbalanced data problem [19], in which the size of the training data from one class is significantly larger than that of the other class in a two-class classification task. Such a problem is frequently encountered in many biomedical applications where data from both positive and negative diagnosis categories are not available equally. For example, the data kept by a hospital can be mostly on positive diagnoses where data for negative diagnoses are not all kept. Another scenario can be in screening or patient monitoring where most cases are diagnosed as negative and only a small number of cases are diagnosed as positive. This means the collected data for the two categories are highly imbalanced and they will impact on the performance of binary classifiers, such as the *BSVCs*.

One possible solution to the imbalanced data problem is to use "recognition"-based approach instead of the conventional discriminative two-class classification approach [18]. "Recognition"-based approach is based on a one-class classification model in which only the data from one class (usually the class with more training samples, known as the majority class) are used to train a classifier [47] as opposed to using data from both classes in traditional two-class classifier training. This can prevent the adverse influence due to using a less representative smaller dataset of the minority class and hence avoiding the problem of imbalanced datasets. Two examples of such one-class classification kernel machines are the one-class Support Vector Classifier called the (*νSVC*) [43] and the Support Vector Data Description (*SVDD*) [48]. These one-class SVMs (*OSVC*) are trained using the data from the majority class only. However, the performance of one-class classifiers is reported to be seldom superior to the traditional two-class classifiers in real applications [41]. One reason might be that the data distribution of majority and minority classes is not suitable to be modeled as a one-class classification problem. Another reason may be due to the fact that only the data from one-class are used in one-class classifier training and no information about the other class is used. Hence, the one-class classifiers are to "recognize" the trained class rather than discriminating two classes.

To complement the strengths of these two types of kernel machines, this chapter shows how the hybrid kernel machines, in which the one-class SVMs and binary SVMs work in tandem as a multi-classifier system, handle the imbalanced data problem. We also report the use of such kernel machines in a couple of biomedical applications, namely, abnormal heart beat annotation from ECG waveform and tumor region detection from colonoscopic images.

In the following sections, we first introduce the one-class and binary SVMs and investigate how their training can be affected by the imbalanced data problem. We then present the hybrid kernel machines and show that the classifiers' performance can be improved. After that, we include two biomedical applications and show how the hybrid kernel machines can be used in these applications.

7.2 One-Class and Binary SVMs

In this section, the fundamentals of both discriminative two-class *SVM*s and recognition-based one-class *SVM*s are introduced. Their classification performance on a particular type of imbalanced data problem is investigated in Sect. 7.5 using an artificial dataset.

7.2.1 Discriminative Support Vector Machines for Binary Classification

SVM, a method based on the principles of statistical learning theory [52], can be applied to classification, regression and concept learning. The SVM is originally developed for two-class classification (or binary classification) task and it has been extended for multiple-classification [44]. For the sake of completeness, we briefly introduce the binary SVM in this subsection.

In two-class classification, an SVM classifier is trained using a training set of labeled samples in which the two classes are labelled as $+1$ and -1, respectively,

$$X = \{x_i \in R^d | i = 1, 2, \cdots, N\} \tag{7.1}$$

where N is the number of samples in the training set. Each sample x_i is represented by a feature vector of d dimensions and labelled as $y_i \in \{+1, -1\}$. The classifier can be represented by a function $f(x) : x \rightarrow y$. The label y can be obtained for each pattern x by the classifier. It is assumed that the training and test data are drawn from the same distribution $P(x, y)$. The optimal function f can be found by minimizing the expected risk

$$R(f) = \int r(f(x), y) dP(x, y) \tag{7.2}$$

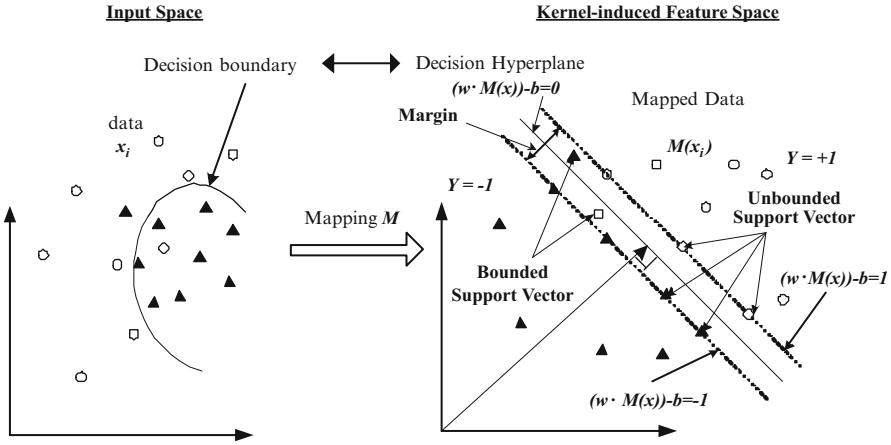

Fig. 7.1 Kernel mapping and optimal separating hyperplane of SVM

where r is a loss function. Conveniently, $r(f(x),y) = |f(x) - y|$ can be defined as 0 for correct classification and 1 for incorrect classification, known as $0/1$ loss.

In practice, the underlying probability distribution $P(x,y)$ is usually unknown. Therefore, the risk R cannot be minimized directly. However, the risk can be approximated by minimizing the empirical risk

$$R_{em}(f) = \frac{1}{N}\sum_{i=1}^{N} r(f(x_i),y_i). \tag{7.3}$$

The empirical risk R_{em} converges to the expected risk R when the number of training samples tends to infinity ($N \to \infty$). However, overfitting may occur when the number of training samples is small [3]. Instead, the expected risk can be estimated while avoiding overfitting using the Vapnik Chervonenkis (VC) theory and the structural risk minimization (SRM) principle [51].

In practice, the bound on the expected risk is often difficult to compute. Fortunately, the decision functions in SVMs are restricted to hyperplanes whose VC-dimension can be bounded in terms of another quantity, called the "margin" [51].

Given that the two-class training set X with N samples are not linearly separable, the data are mapped to another feature space using a mapping M by which the mapped data $M(x)$ can be separated by an optimal separating hyperplane expressed as

$$f(x) = (w \cdot M(x)) - b \tag{7.4}$$

in which w is a weight vector, b is a bias item. $(.)$ is an inner product. Such a mapping is illustrated in Fig. 7.1.

The "margin" is defined as the minimal distance of a sample to the decision hyperplane $f(x)$. w and b can be scaled so that the closest point to the hyperplane satisfies $|w \cdot M(x) - b| = 1$. Then the margin can be calculated using two samples from opposite classes $M(x_1)$ and $M(x_2)$ which have $w \cdot M(x_1) - b = 1$ and $w \cdot M(x_2) - b = -1$, respectively, and thus,

$$\frac{w}{\|w\|} \cdot (M(x_1) - M(x_2)) = \frac{2}{\|w\|} \tag{7.5}$$

The optimization problem then becomes:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{7.6}$$

subject to constraints

$$y_i((w \cdot M(x_i)) - b) \geq 1, i = 1, 2, \cdots, N \tag{7.7}$$

For noisy data, some slack variables θ_i can be introduced to relax the constraints in (7.7):

$$y_i((w \cdot M(x_i)) - b) \geq 1 - \theta_i, \quad \theta_i \geq 0, \quad i = 1, 2, \cdots, N \tag{7.8}$$

The optimization problem in (7.6) can be reformulated as

$$\min_{w,b,\theta} \left\{ \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \theta_i \right\} \tag{7.9}$$

where $C > 0$ is a regularization parameter to control the trade-off of the empirical error and the capacity terms.

The minimization problem in (7.9) is called primal in optimization theory. Its first item is related to the model complexity and the second item is the empirical risk R_{em}. Therefore, minimizing (7.9) can minimize the expected risk R. This problem can be solved by introducing Lagrange Multipliers $\beta_i \geq 0$ and $\gamma_i \geq 0$, $i = 1, 2, \cdots, N$, and with the constraints in (7.8), this leads to the dual problem:

$$\max_{\beta} \left\{ \sum_{i=1}^{N} \beta_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \beta_i \beta_j (M(x_i) \cdot M(x_j)) \right\} \tag{7.10}$$

with constraints

$$\sum_{i=1}^{N} y_i \beta_i = 0 \tag{7.11}$$

and

$$C \geq \beta_i \geq 0, i = 1, 2, \cdots, N \tag{7.12}$$

This is a quadratic programming problem, which can be solved using standard algorithms, such as sequential minimization optimization [39]. Related codes can be found in [1].

In fact, only the inner product is calculated in (7.10) and (7.4). No explicit mapping M is needed. Such an inner product can be replaced using a kernel function

$$K(x_i, x_j) = M(x_i) \cdot M(x_j) \tag{7.13}$$

provided that this kernel $K(x_i, x_j)$ satisfies the Mercer's theorem. Then, equations (7.10) and (7.4) can be reformulated as follows

$$\max_{\beta} \left\{ \sum_{i=1}^{N} \beta_i - \frac{1}{2} \sum_{i,j=1}^{N} y_i y_j \beta_i \beta_j K(x_i, x_j) \right\} \tag{7.14}$$

$$f(x) = \sum_{i=1}^{N} y_i \beta_i K(x_i, x) - b \tag{7.15}$$

Among all possible kernels, the Radial Basis Function (RBF) kernel is a widely used one,

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{\sigma^2}} \tag{7.16}$$

7.2.2 *Recognition-Based One-Class Support Vector Machines*

As mentioned at the beginning of this chapter, recognition-based one-class SVMs can be used to handle the imbalanced data problem by learning from the majority class samples. We give a brief review of this type of SVMs in the following subsection.

7.2.2.1 One-Class Classification

One-class classification is also known as novelty detection, outlier detection and concept learning [47]. The problem formulation in one-class classification is different from conventional two-class classification. In one-class classification, it is assumed that only information of one of the classes, **the target class**, is available, and no information is available from the other class, known as **the outlier class**. The task of one-class classification is to define a boundary around the target class such that it accepts as much of the targets as possible and excludes the outliers as much as possible (Fig. 7.2b).

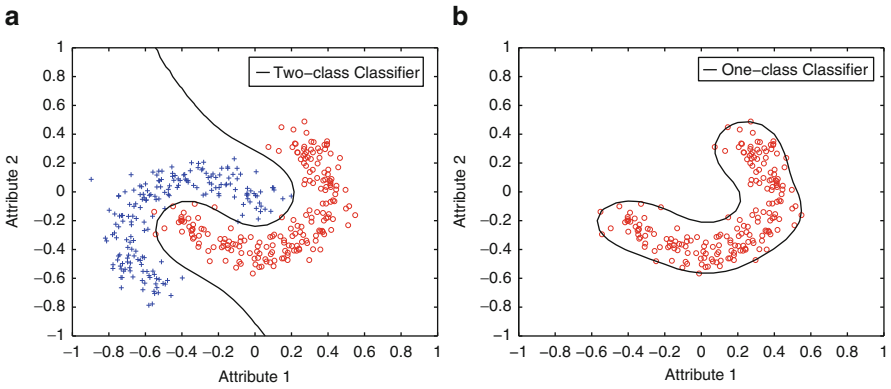

Fig. 7.2 Typical decision boundaries of (**a**) Two-class classifier and (**b**) One-class classifier in a 2-D toy problem

The philosophy behind one-class classification is in agreement with the way that human beings learn a concept. Suppose one expects to teach a child the concept of "car." One only needs to give him or her some examples of cars and it is not necessary to give the examples of non-"car," such as truck, bus or train. This is to say, people can learn a concept using only the examples of the target class. Of course, the information about non-target or outliers is helpful to improve the discrimination between the target and the non-target classes. However, using the examples from only the target class is sufficient to learn the concept of the target and recognize whether a new pattern belongs to the concept of the "target class."

To sum up, as illustrated in Fig. 7.2, the decision boundary of two-class classifier is supported by the samples of both classes and it utilizes the information from both classes while the decision boundary of one-class classifier is formed using only the data from one class. The two-class classifier is trained for "discrimination" purpose but the one-class classifier is trained to "recognize" the target samples rather than for "discrimination" purpose. Therefore, classification performance of one-class classifiers is usually worse than two-class classifiers when the data from both classes are available [41].

In one-class classifiers, a threshold is usually set so that the decision boundary of the classifier can enclose the target samples as much as possible. This is usually difficult when no information is available from the other class. One way is to reject some target to form a tighter boundary. The threshold can be determined based on the errors of classifying the target class only [47].

One-class classification has been used in many fields. Hojjatoleslami et al. employed a RBF network for density estimation in the detection of micro-calcifications in mammograms [14]. Manevitz and Yousef used One-Class Support Vector Machine for document classification [32]. Tax et al employed Support Vector Data Description in pump failure detection [48] and image retrieval [49]. A survey of the one-class classifiers can be found in [33, 34].

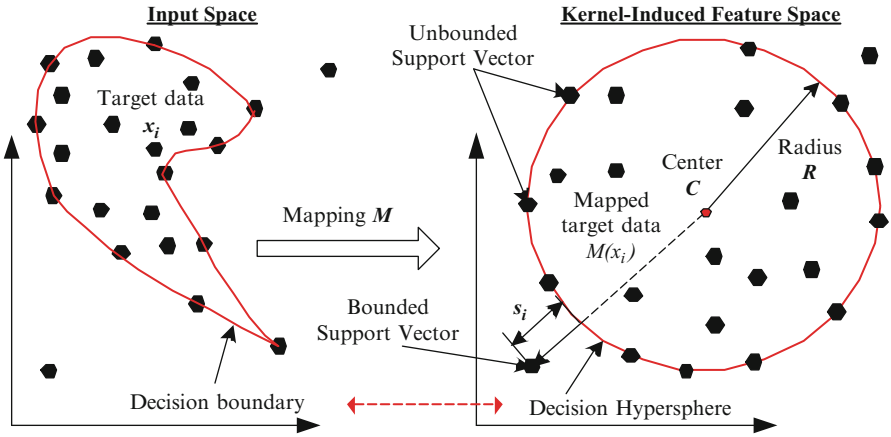

Fig. 7.3 Illustration of the kernel mapping of SVDD

7.2.2.2 Support Vector Data Description

The formulation of SVDD is as follows. Given a set of target data X with N samples, a nonlinear mapping M is sought to map X into some high dimensional kernel-induced feature space in which a hypersphere is sought to enclose the mapped target data $M(X)$ with smallest radius R centered at c. Figure 7.3 illustrates the nonlinear kernel mapping. The problem becomes

$$\min_{\{R,c,S\}} \left\{ R^2 + \frac{1}{\nu N} \sum_{i=1}^{N} S_i \right\} \tag{7.17}$$

subject to

$$\| M(x_i) - c \|^2 \leq R^2 + S_i, \quad i = 1, 2, \cdots, N \tag{7.18}$$

where S_i $(S_i \geq 0)$ are some slack variables to allow soft boundaries, i.e. some target data are allowed to lie outside of the hypersphere so as to control the trade-off between two types of errors. $\nu \in (0, 1]$ is a regularization parameter used to control the trade-off between the size of the hypersphere and the errors. In fact, it is the upper bound of the fraction of target data located outside the hypersphere.

The above problem can be solved by constructing a Lagrangian. Introducing constraints (7.18) to cost function (7.17), we have the following dual problem:

$$\max_{\beta} \left\{ \sum_{i=1}^{N} \beta_i (M(x_i) \cdot M(x_i)) - \sum_{i,j=1}^{N} \beta_i \beta_j (M(x_i) \cdot M(x_j)) \right\} \tag{7.19}$$

with constraints

$$\sum_{i=1}^{N} \beta_i = 1 \tag{7.20}$$

$$0 \leq \beta_i \leq \frac{1}{\nu N}, \quad i = 1, 2, \cdots, N \tag{7.21}$$

Define function $K(\cdot,\cdot)$ as

$$K(x_i, x_j) = M(x_i) \cdot M(x_j) \tag{7.22}$$

Then, Eq. (7.19) becomes

$$\min_{\beta} \left\{ \sum_{i,j=1}^{N} \beta_i \beta_j K(x_i, x_j) - \sum_{i=1}^{N} \beta_i K(x_i, x_i) \right\} \tag{7.23}$$

with the same constraints as (7.19). The cost function of the dual problem (7.23) is convex and quadratic in terms of the unknown parameters β_i. This problem can be solved by quadratic programming for which some standard algorithms such as sequential minimization optimization can be employed [43, 47].

Through quadratic programming, the Lagrangian (7.23) is optimized with respect to β. The center of the hypersphere c and multiplier γ_i can be calculated using the optimal solution β. Because $\| M(x_i) - c \|^2 = R^2$ holds for all the unbounded support vectors (USVs), the radius R can be calculated by choosing any of the USVs x_s

$$R = \left[K(x_s, x_s) + \sum_{i,j=1}^{N} \beta_i \beta_j K(x_i, x_j) - 2 \sum_{i=1}^{N} \beta_i K(x_i, x_s) \right]^{-2} \tag{7.24}$$

Given a new pattern z, the decision function is

$$\begin{aligned} f(z) &= R^2 - \| M(z) - c \|^2 = R^2 - K(z,z) \\ &\quad - \sum_{i,j=1}^{N} \beta_i \beta_j K(x_i, x_j) + 2 \sum_{i=1}^{N} \beta_i K(z, x_i) \end{aligned} \tag{7.25}$$

If the value of the decision function is greater than zero, the new sample lies inside the hypersphere and hence is classified as a target. Otherwise, it is classified as an outlier.

Similar to binary SVMs, kernel function can be used in SVDD. Although nonlinear mapping has been used to improve the effectiveness of the hyperspherical description, neither does the explicit nonlinear mapping $M(.)$ appear in the dual problem of SVDD (7.23), nor in the decision function (7.25). They are expressed completely in terms of $K(x_i, x_j)$, which is the advantage of kernel method. In fact,

since the problem is stated completely in terms of the inner products of the vectors, the inner products of the patterns can be replaced by a kernel function (7.22), provided that this kernel $K(x_i, x_j)$ satisfies the Mercer's theorem [47].

The Gaussian RBF kernel in Eq. (7.16) provides a very flexible description, which has been proven in [48]. Because it only depends on $x_i - x_j$, $K(x,x)$ is constant 1. Therefore, Eq. (7.23) becomes

$$\min_{\beta} \sum_{i,j=1}^{N} \beta_i \beta_j K(x_i, x_j) \tag{7.26}$$

subject to the same constraints as (7.23). The decision function (7.25) can be reformulated as follows using (7.24),

$$f_d(z) = \sum_{i=1}^{N} \beta_i [K(x_i, z) - k(x_i, x_s)] = \sum_{i=1}^{N} \beta_i K(x_i, z) - b \tag{7.27}$$

where the bias term b is

$$b = \sum_{i=1}^{N} \beta_i K(x_i, x_s) = \sum_{i=1}^{N} \beta_i e^{-\frac{\|x_i - x_s\|^2}{\sigma^2}} \tag{7.28}$$

Here, the SVDD decision function behaves as a template-matching detector in the mapped feature space. Since $\beta_i \neq 0$ holds only for those USVs and bounded support vectors (BSVs), these patterns form a known template. Given a new pattern, it is compared with only the USVs and BSVs in the mapped feature space. A pattern similar to all of the USVs and BSVs tends to have a large negative value in (7.27) and it is more likely to be an outlier. A pattern different from all of the USVs and the BSVs tends to have a large positive value in (7.27) and it is more likely to be a target.

7.2.2.3 ν-Support Vector Classifier

Another way of estimating the support of a data distribution in the kernel feature space is the *νSVC*. The kernel mapping is different from that of SVDD. The target data are mapped into a higher-dimensional space called feature space $M(x)$ in which the dot product can be computed using some kernel function. The mapped target data are away from the origin as shown in Fig. 7.4, which can be found by solving the following problem

$$\min_{w, S_i, b} \frac{\| w \|^2}{2} + \frac{1}{\nu N} \sum_{i=1}^{N} S_i - b \tag{7.29}$$

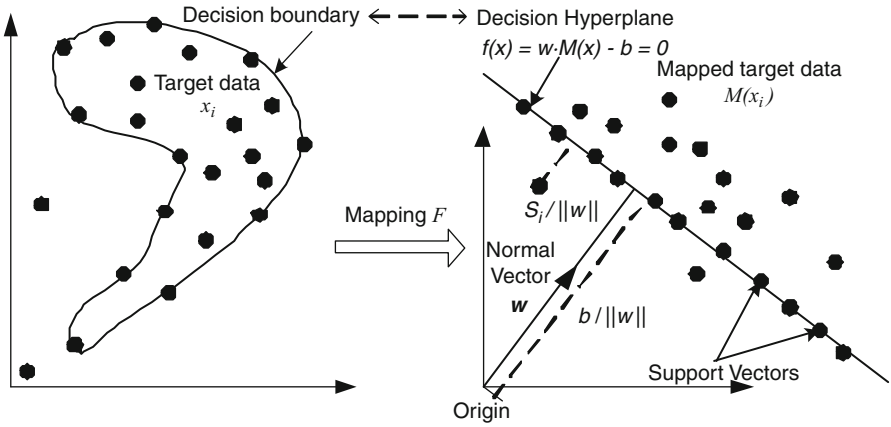

Fig. 7.4 Illustration of the kernel mapping of *vSVC*

subject to

$$w \cdot M(x_i) - b + S_i \geq 0, \quad \gamma_i \geq 0, \quad i = 1, 2, \cdots N \tag{7.30}$$

where S_i are slack variables. $\nu \in (0, 1]$ is a regularization parameter to control the effect of outliers and allows for target samples falling outside the decision boundary. The decision function corresponding to the hyperplane is

$$f(x) = w \cdot M(x) - b \tag{7.31}$$

By similar analysis as in SVDD, this problem can be solved as a quadratic programming problem which is exactly the same as the dual problem (7.26) in SVDD when the Gaussian kernel is used. Hence, *vSVC* and *SVDD* are equivalent to each other [43].

7.3 The Imbalanced Data Problem

The imbalanced data problem has received considerable attention in recent years in the machine learning community. This is the problem when the size of the training set from one class is significantly larger than that of the other class in a two-class classification setting. An example is illustrated in Fig. 7.5. This problem is often encountered in real applications such as in medical screening for abnormalities, image retrieval, and oil spill in satellite images [5, 23]. In applications such as medical screening for abnormalities, the data of the normal class can be easily obtained. On the other hand, the data of the abnormal class are more difficult to be collected than the normal ones. Therefore, the data from the abnormal class (usually

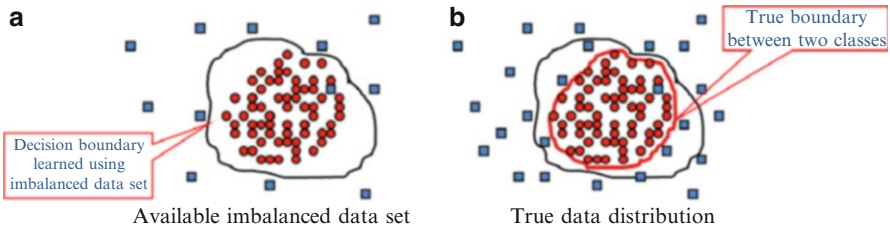

Fig. 7.5 Example of the imbalanced data problem. (**a**) is the decision boundary learned using the available (imbalanced) dataset which is different from the true boundary between the two classes in (**b**)

the minority class) cannot represent its true distribution well compared to the other class (usually the majority class, i.e. the normal class). It can be assumed that the minority class is positive and the majority class is negative in this chapter and it is formulated as a simple binary classification problem. Hence, it is logical to use discriminative binary classifiers such as binary SVMs. However, such classifiers are designed to minimize the overall misclassification rate on the training set, their classification performance degrades if they are trained with a highly imbalanced dataset.

Some attempts have been reported to deal with the imbalanced data problem, which can be classified into three approaches [17, 19] presented in the following subsections.

7.3.1 Resampling

The first approach is resampling the training dataset to make it balanced, such as in [10, 24]. Resampling is probably the most extensively studied approach, which consists of two main techniques:

1. *Undersampling:* The data from majority class are down-sampled so that the size of the majority class matches the size of the minority class. The sampling can be either done randomly [19] or based on some rules [24]. But the problem is that some of the information may be lost if down-sampling is not done properly.
2. *Oversampling:* The data from minority class are over-sampled so that the size of minority class matches the size of the majority class. Similar to random under-sampling, random oversampling has been shown to be effective in improving the classification [19]. There are also some attempts to improve the performance of oversampling. For example, Chawla et al developed a Synthetic Minority Oversampling Technique (SMOTE) by generating artificial data (the nearest neighbors of the original minority data) [4].

It is unclear which of these two is more effective in solving the imbalanced data problem [8, 10]. Therefore, some attempts have also been made to combine these two approaches [4, 10].

7.3.2 Using Different Costs to Two Classes

The second approach is to compensate for the class imbalance by altering the costs of the minority and majority classes in the training of classifiers. For example, Karakoulas et al proposed an algorithm called ThetaBoost, which is a boosting algorithm with unequal loss functions [20]. Some attempts have also been made to compensate the class imbalance by using different costs to the two classes in the training of SVMs [53]. Raskutti et al used different penalizing factors for two classes and resampling for SVM in [41]. Wu et al proposed the class-boundary alignment algorithm to deal with imbalanced data problem in SVM [56].

7.3.3 Recognition-Based Approach

The third approach is to use recognition-based instead of discrimination-based learning strategy by leaving one of the two classes totally unused (usually the minority class). The recognition-based method resembles that of a density estimation without finding the true density explicitly. This is an extreme case where only the data from one class are used to construct the learning model. For example, Japkowicz proposed to use an autoencoder to solve the imbalanced data problem [18]. This method works well when the majority class can be well modelled by a novelty detector such as an autoencoder. However, a recognition-based method is usually outperformed by a discrimination-based one due to the exclusion of the information from the minority class in training the model [41].

7.4 Hybrid Kernel Machine Ensemble

It has been discussed in the previous sections that discriminative two-class SVMs have problems in dealing with imbalanced datasets and the recognition-based one-class SVM cannot always do better than two-class SVMs. Basically, a two-class classifier *BSVC* benefits from the information from two classes while suffering from inadequate representation of the minority class. But, a one-class classifier *OSVC* benefits from more precise representation of the majority class but is not highly discriminative. There is a need to develop a classifier which is in-between the one-class classifier and the two-class classifier. Such a classifier can be named as ***one and half (1.5) classifier***. By exploiting the different properties of the two types of kernel

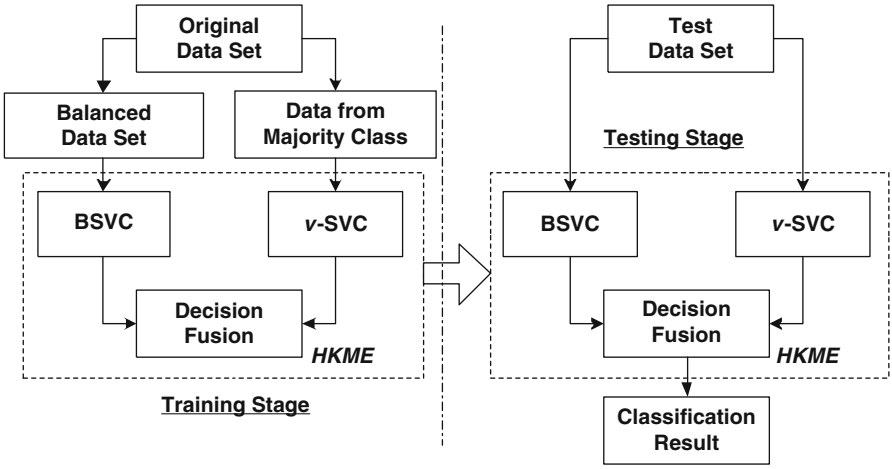

Fig. 7.6 Illustration of hybrid kernel machine ensemble (*HKME*) framework [26]

machines, an ensemble can be constructed by combining these two types of kernel machines. Such an ensemble is called Hybrid Kernel Machine Ensemble (*HKME*). *HKME* is designed to benefit from both discriminative *BSVC* and recognition-based *OSVC* and such an ensemble is expected to perform better in some applications such as in imbalanced datasets or in other cases where there is a need to combine the information from these two types of kernel machines. Most of the material presented in the following subsections has been published in [26].

7.4.1 Hybrid Kernel Machine Ensemble Framework

The hybrid kernel machine ensemble (*HKME*) framework is illustrated in Fig. 7.6. A *HKME* consists of two different types of *SVM*s, i.e. a discriminative *BSVC* and a non-discriminative recognition-based *vSVC* (or *SVDD*). Hence, the *HKME* is expected to benefit from the strength of both *BSVC* and *vSVC*.

HKME is designed for problems where *BSVC* does not perform well or costly to construct while *vSVC* shows good performance. For example, there is a type of imbalanced data problem in which the majority class is compactly clustered and the minority class is scattered in the input space. One example is in heart patient monitoring using ECG. The ECG signal morphologies from normal activities (normal class) are similar and the data from this class can be easily collected (majority class), while those from abnormal activities (abnormal class) may exhibit various morphologies and are more difficult to collect (minority class). A discriminative model, such as a *BSVC*, can be trained by manually balancing the data or compensating the imbalance using different costs to the two classes. Thus, the discriminative model uses the information from both majority class and minority class. However,

its performance can still be poor due to the poorly represented minority class. A recognition-based one-class *SVM* may do better than the discriminative *BSVC* in this situation by modeling the well-represented majority class only. Since the majority class satisfies the assumption of one-class classification where the majority class is well represented and compactly clustered, it avoids the problem faced by the binary *SVM* due to the inadequate representation of the minority class. However, as a descriptive model, such a recognition-based model is not highly discriminative because the information from the minority class is left totally unused. Hence, there is a need to incorporate the information from the minority class to the recognition-based model or exploit the well-represented majority class further in the discriminative model. Exploiting the complementary nature of these two different types of models, a combination of them is expected to perform better than using either of them separately for the classification of this type of imbalanced dataset. Hence, constructing a *HKME* by integrating these two types of kernel machines in an ensemble is presented here to address this type of imbalanced data problem.

In this framework, a *νSVC* can be trained using only the data of majority class, so it can avoid the problem of poor representation of the minority data. On the other hand, a *BSVC* can be trained using balanced dataset using oversampling or undersampling, so it benefits from the information from both classes. The outputs of the two *SVM*s can be integrated using some fusion rules. Since the *νSVC* and *BSVC* are trained using different datasets, the training sets of such two kernel machines can be considered diverse. Furthermore, the different nature of the two *SVM*s can further help to increase the diversity. Therefore, the ensemble of such two kernel machines is expected to improve the classification compared to using either of the two types of *SVM*s.

7.4.2 *Binary* SVM *Training*

Performance of the classifiers is closely related to the parameters used by the classifiers. There are two hyper-parameters to be tuned in *BSVC* when using the Gaussian RBF kernel, the width parameter σ of the RBF kernel and the regularization parameter C which is used to control the trade-off of errors. The hyper-parameters of *BSVC* can be optimized using cross validation on the training set. The use of cross validation is able to avoid over-fitting [3]. The values of the hyper-parameters are chosen so that the errors of both classes on the validation set are minimized.

Another problem in *BSVC* is its training using imbalanced datasets. It has been shown that balanced dataset generally leads to results which are no worse than or superior to those of using natural class distribution, although it does not always produce the optimal results [54]. Since *BSVC* suffers from the imbalanced data problem, the original dataset can be balanced first using oversampling, undersampling or *SMOTE* algorithms aforementioned. The trained *BSVC* using the balanced dataset can then be integrated with the one-class *SVM* to form the *HKME*.

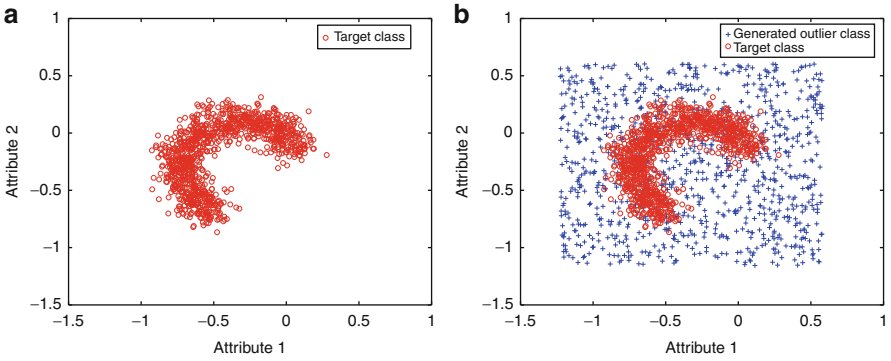

Fig. 7.7 (**a**) Original target dataset and (**b**) Generated artificial outliers around the target class in a toy problem in 2-D space

7.4.3 *One-Class* SVM *Training*

The hyper-parameters of *νSVC* or *SVDD* using Gaussian RBF kernel are the same as those of the BSVCs, i.e. the width parameter of the RBF kernel σ and the regularization parameter ν are used to control the trade-off of errors. The parameters of two-class classifiers can be optimized using cross validation on the training set. However, the information about the outlier class is assumed to be unavailable for one-class classifiers, hence the hyper-parameters can only be estimated using the data from target class or be chosen heuristically. This problem can be solved by generating artificial outliers [50]. Given a set of target samples, some outlier samples are generated randomly with the assumption that the outliers are uniformly distributed around the target class. The union of targets and generated outliers is used as a validation set to optimize the hyper-parameters of one-class *SVM*. A toy dataset and generated artificial outliers are illustrated in Fig. 7.7.

As for the imbalanced data problem in question, there are still some outlier samples, i.e., data from the minority class. The hyper-parameters may be tuned to minimize the training error on the whole training set which consists of both majority and minority classes. But, this might be undesirable if the minority class is not well represented by the sampled data. This problem will be discussed further in the experimental section.

7.4.4 *Fusion Rules for Integration of Hybrid* SVM*s*

Integrating two *SVM*s in a hybrid is posed as a decision level fusion problem. It is nontrivial to properly combine the two sources of information from these two types of *SVM*s.

Many ensemble learning methods have been developed. In this subsection, several ensemble methods are reviewed to understand how the imbalanced data problem can be handled by them and these include Decision Template (*DET*), Stacking, Average (*AVG*), Maximum (*MAX*), Minimum (*MIN*), Product (*PROD*) [22, 25].

Let $C_i(x) = \{C_{i1}(x), C_{i2}(x), \cdots, C_{ik}(x)\}$ be a set of individual classifiers, called an ensemble, each of which gets an input feature vector $\mathbf{x} = [x_1, x_2, \cdots, x_d]^T$ and assigns it to a class label y_i from $Y = \{-1, +1\}$, the goal of the ensemble is to find a class label L_{ens} for $\mathbf{x}$ based on the outputs of k classifiers $C_1(x), C_2(x), \cdots, C_k(x)$ corresponding to labels $L_1(x), L_2(x), \cdots, L_k(x)$. $C_i(x)$ is often an estimate of the posterior probability $P(y_i|\mathbf{x})$.

- *Decision template:* The decision template DET_j for class $y_j \in \{-1, +1\}$ is the average of the outputs of individual classifiers with respect to the training set for class y_j [25]. The ensemble *DET* assigns the input x with the label given by the individual classifier whose Euclidean distance to the decision template DET_j is the smallest.
- *Stacking (Stacked generalization):* Taking the output of individual classifiers $C_i(x)$ as input to an upper layer classifier and the final decision is determined by the upper layer classifier [55].

$$L_{ens}(x) = F(C_1(x), C_2(x), \cdots, C_k(x)) \tag{7.32}$$

  The upper layer classifiers used here include linear discriminant classifiers (*LDC*s) and quadratic discriminant classifiers (*QDC*s) assuming normally distributed classes. Because the covariance matrices for the classes are near singular, *QDC*s may fail when trying to estimate and invert the covariance matrices [25].
- *Average:*

$$L_{ens}(x) = \arg\max_j \left(\sum_{i=1}^{k} \frac{C_{ji}(x)}{k} \right) \tag{7.33}$$

  where $j \in \{-1, +1\}$. The *AVG* rule calculates the average of the outputs of the k individual classifier and assigns the input x to the class with the largest posterior probability.
- *Maximum:*

$$L_{ens}(x) = \arg\max_j \left(\max_i C_{ji}(x) \right) \tag{7.34}$$

  where $j \in \{-1, +1\}$. The *MAX* rule takes the maximum value of the outputs from the k individual classifier for each class and assigns the input x to the class with the largest posterior probability.

- *Minimum:*

$$L_{ens}(x) = \arg\max_j \left(\min_i C_{ji}(x) \right) \tag{7.35}$$

  where $j \in \{-1, +1\}$. The *MIN* rule takes the minimum value of the outputs from the k individual classifier for each class and assigns the input x to the class with the largest posterior probability.
- *Product:*

$$L_{ens}(x) = \arg\max_j \left(\prod_i C_{ji}(x) \right) \tag{7.36}$$

  where $j \in \{-1, +1\}$. The *PROD* rule calculates the product value of the outputs from the k individual classifier for each class and assigns the input x to the class with the largest posterior probability.

The problem here is to fuse the outputs of two classifiers. The generally used majority voting is not suitable here. Furthermore, it can be proved that Maximum, Minimum, Averaging, Product rules are equivalent to each other when they are used to combine two classifiers with posterior probability outputs for a two-class classification task. It has been proved that Maximum and Minimum are equivalent when combining multiple classifiers for two-class classification in [46]. Due to the equivalence of *MAX*, *MIN*, *AVG*, and *PROD* rules for the two-class problem using two classifiers with posterior probability as outputs, only *AVG* is investigated in the following subsection.

7.4.5 Estimating the Posterior Probability for Outputs of *SVMs*

The outputs of *SVM*s are not posterior probabilities and are in different ranges, and hence are not comparable directly. Thus, their outputs have to be normalized for use in this hybrid. It is observed that the outputs of *SVM*s show similar forms. One can estimate the posterior probabilities $P_i(y_j|\mathbf{x})$ of the i-th *SVM* using a sigmoid function by minimizing the negative log likelihood of the training data [40]

$$P_i(y_j|\mathbf{x}) = \frac{1}{1 + e^{p_i f_i(x) + q_i}} \tag{7.37}$$

where p_i is a coefficient to control the shape of sigmoid function and q_i is a coefficient to control the shift along the horizontal axis ($f_i(x)$). Thus, the ensembles can be constructed using these estimated posterior probabilities.

When estimating the posterior probability of *BSVC*, the training set of the *BSVC* has to be balanced. Otherwise, it may lead to biased fitting of a sigmoid to outputs of nonlinear *SVM*s [40, 52]. The balancing of the training set can be done using oversampling such as *SMOTE* [4].

To our best knowledge, this is the first attempt to estimating the posterior probability of *νSVC* or *SVDD*. Since there is only the target data for training a one-class *SVM*, a set of artificial data can be generated whose sample size is the same as that of targets [50]. The union of target data and artificial data can be used to estimate the posterior probability of output from one-class *SVC*.

The posterior probability generated here is only an estimation of the true posterior probability. Bias is unavoidable. This may create some problems to the fusion rules such as *MAX* or *MIN*.

7.5 Experimental Results on Artificial Dataset

The performance of the proposed *HKME* is evaluated on an artificial dataset. The evaluation measure is as follows.

7.5.1 Evaluation Measure

The decision table of two-class classification outcome for calculating the evaluating criteria used in this study is illustrated in Table 7.1.

Four classification outcomes are considered, i.e. true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Let A_+ and A_- denote the classification accuracy rates for positive class and negative class, respectively.

$$A_+ = \frac{TP}{TP+FN} \tag{7.38}$$

$$A_- = \frac{TN}{TN+FP} \tag{7.39}$$

The most commonly used measure is the Average Classification Rate (*ACR*) which is the fraction of all correctly classified samples among all the samples, regardless of the classes:

$$ACR = \frac{TP+TN}{TP+TN+FP+FN} \tag{7.40}$$

Table 7.1 Decision table of two-class classification outcome for calculating the evaluating measure

Ground truth	Classification outcome	
	Positive	Negative
Positive	True positive	False negative
Negative	False positive	Truth negative

In imbalanced datasets, the negative (majority) class dominates. The generally used *ACR* is not valid for evaluating the performance of the classifiers in such an imbalanced dataset. For example, if a classifier classifies all the data as negative samples, it has $A_- = 100\,\%$ and $A_+ = 0\,\%$, but the ACR is still high. Hence, another measure called the Balanced Classification Rate (*BCR*) is used in this study. *BCR* is the algebraic mean of A_+ and A_-:

$$BCR = \frac{A_+ + A_-}{2}. \tag{7.41}$$

This measure has been used for evaluating the performance of classifiers on imbalanced datasets [11, 45]. This measure is more suitable for evaluating the performance of the classifiers here than the generally used *ACR* for the following reason. Only when both A_+ and A_- have large value *BCR* can have a large value. Therefore, the use of *BCR* can give a balanced assessment of the classifiers for the imbalanced datasets as the *BCR* favors both lower false positives and false negatives.

7.5.2 Artificial Dataset

To investigate the effects the imbalanced data have on BSVC and OSVC and HKME, experiments were conducted using a checkerboard dataset similar to the one in [56]. The checkerboard data are shown in Fig. 7.8. The negative samples (majority class) occupy two diagonal squares of the checkerboard in the center and the positive samples (minority) surrounds the negative samples. The data are uniformly distributed and it is in agreement to the assumption that the data of the majority class is compactly clustered and the data of the minority class is scattered in the input space.

7.5.2.1 Influence of Class Imbalance to Discriminative *BSVC*s

In order to show the influence of imbalanced dataset on the performance of discriminative *BSVC*s, the following experiments were conducted.

In the first experiment, the size of negative training data in the 2×2 checkerboard data was fixed at 128, the size of the positive training data was reduced from 128 to 4, with increasing imbalance ratio (majority to minority) from 1:1 to 32:1. The test data consists of 1,000 positive samples and 1,000 negative samples. *BSVC*s with RBF kernel were trained using these data. The hyper-parameters of the *BSVC*s were optimized using threefold cross validation on the training set. The experiment was repeated ten times and the average value and standard deviation of the *BCR*s achieved by *BSVC*s are plotted in Fig. 7.9.

It can be observed that *BSVC* performs well when the training dataset is balanced which is expected. But its performance deteriorates gradually as the imbalance ratio

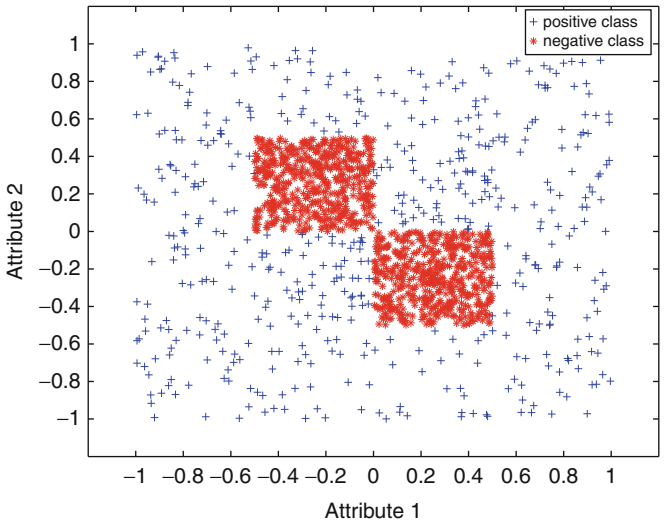

Fig. 7.8 2×2 checkerboard dataset

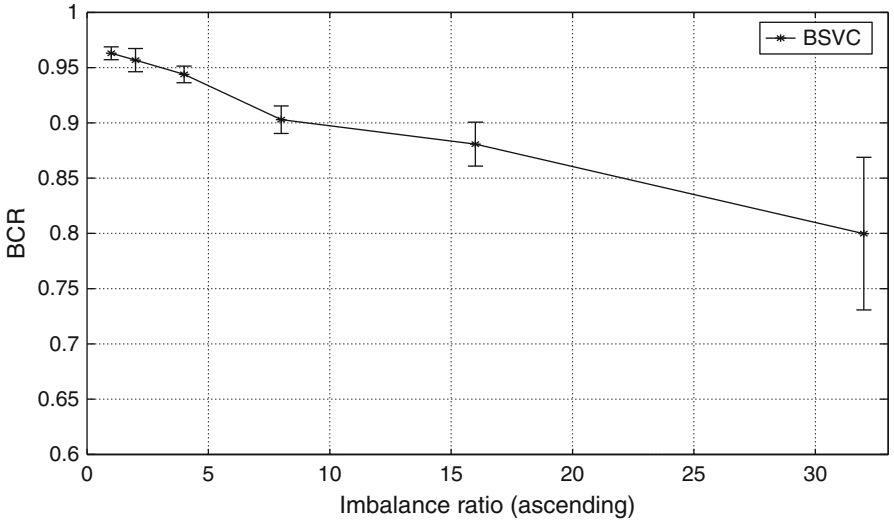

Fig. 7.9 The influence of class imbalance on the performance of *BSVC* using 2×2 checkerboard dataset (with negative samples as majority class)

increases. It indicates that discriminative *BSVC* suffers from the class imbalance. When the number of minority samples is very small, the data from this class cannot represent its true distribution well. This can be observed from the larger variation in the performance of the *BSVC* when the imbalance ratio is large.

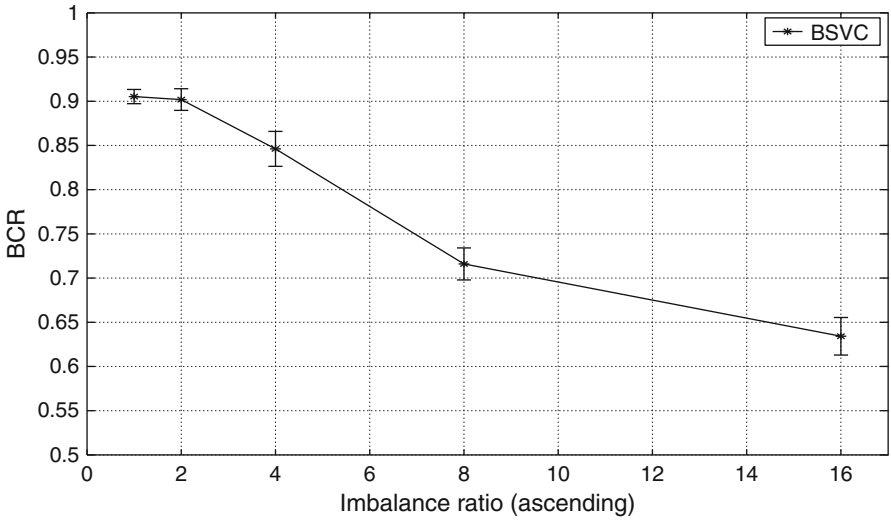

Fig. 7.10 The influence of class imbalance on the performance of *BSVC* using 2×2 checkerboard dataset (with positive samples as majority class)

It is unclear whether *BSVC* also suffers from the imbalanced data problem when the two classes are adequately represented while the samples from two classes are imbalanced. Therefore, the size of negative training data was still fixed at 128 in the second experiment, while the size of the positive training data was increased from 128 to 2,048, with corresponding imbalance ratio (minority to majority) increased from 1:1 to 1:16. Other settings are the same as the first experiment. The experimental result is illustrated in Fig. 7.10.

It can be observed that the result is similar to that in the first experiment. It shows that the discriminative *BSVC* also suffer from the class imbalance when the data are more precisely represented while the two classes are highly imbalanced. In summary, it has been shown that the discriminative *BSVC* suffer from the class imbalance problem. Hence, some measures have to be taken to alleviate this problem.

7.5.2.2 The Performance of Recognition-Based *OSVM*s

It has been mentioned that one approach to address the class imbalance is to use recognition-based model instead of discriminative model by training a one-class classifier using the data from the majority class only. However, one-class classifiers seldom outperform two-class classifiers when the data from two class are available. One reason is that the one-class classifier is designed for describing the majority class rather than for discrimination purpose, leaving the information from another class totally unused. Another reason may be that the concept to be

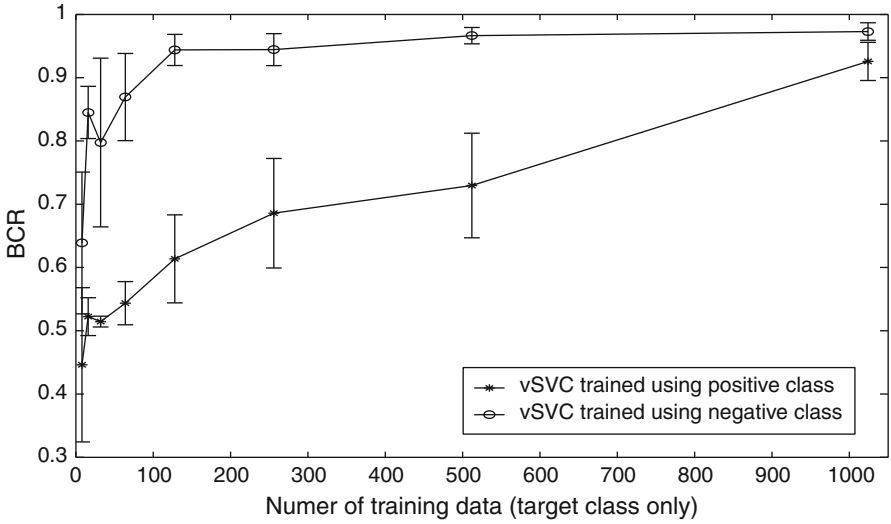

Fig. 7.11 The performance of *vSVC* in terms of *BCR* using 2×2 checkerboard dataset, with negative samples or positive samples as target class for training, respectively

learned is not suitable for description by the one-class classifiers. For example, in patient monitoring, the concept of "normal" is suitable for description by a one-class classifier while the concept of "abnormal" is not. This is because the "normal" class is usually compactly clustered in the input space, while the "abnormal" class is usually scattered. Furthermore, there is no clear boundary between "normal" and "abnormal" class. If a one-class classifier is to enclose the scattered "abnormal" data, it will also include some "normal" data. It may be better to construct a one-class classifier to enclose the compactly clustered "normal" data. The negative class of the checkerboard data in Fig. 7.8 seems more suitable to be described by a one-class classifier than the positive class since it is compactly clustered. This can be ascertained in the following experiment.

In the experiment, the negative samples and positive samples in the checkerboard dataset were taken as target class, respectively, to train a *vSVC*. The number of training data was varied from 8 to 1,024. The test data consists of 1,000 positive samples and 1,000 negative samples. The hyper-parameters of the *vSVC* were optimized using artificially generated dataset described in Sect. 7.4.3. The experiment was repeated ten times and the average value and standard deviation of the *BCR*s achieved by *vSVC*s are reported in Fig. 7.11.

It can be observed that the *vSVC* trained using compactly clustered negative samples outperforms that of using scattered positive samples. This supports the earlier claim that compactly clustered negative class is more suitable for training one-class classifiers than scattered positive class. Furthermore, the performance of *vSVC* is directly related to number of training samples. It seems that $128 \sim 256$ negative samples have been quite good to train a *vSVC* in this dataset, whose

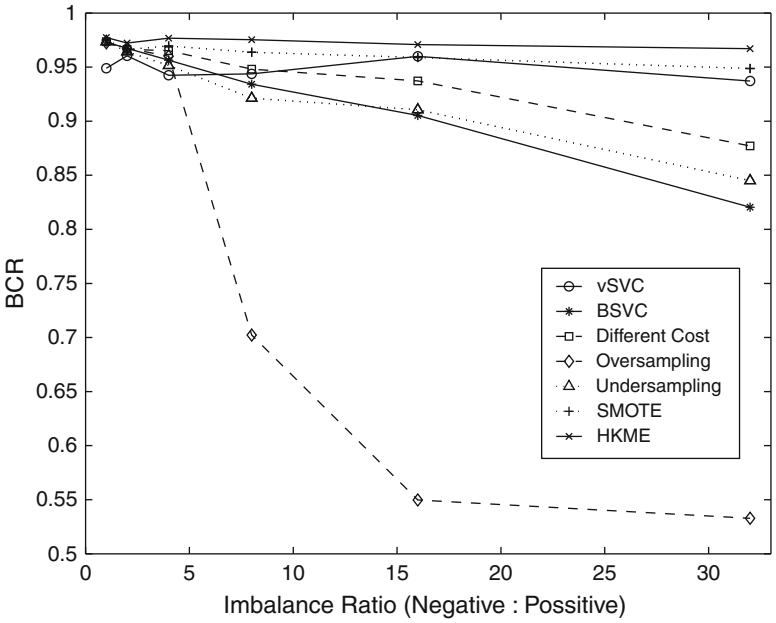

Fig. 7.12 The comparison of different schemes using 2×2 checkerboard dataset with different imbalance ratio [26]

performance is even better than the *vSVC* trained using 1,024 positive samples. It has been pointed out that more data is needed in one-class classification than a two-class classification [47]. So the number of training samples of *vSVC* should be large enough to have a good description of the target class. In imbalanced datasets, a compactly clustered majority class is more suitable for the one-class classifiers to learn.

7.5.2.3 The Performance of HKME

The proposed *HKME* is compared to other commonly used methods to deal with class imbalance using the artificial dataset, including oversampling, down-sampling, *SMOTE* and *BSVC* using different costs to the two classes. The number of negative samples was fixed at 256, the number of positive samples was decreased so that the imbalance ratio (negative to positive) is increased from 1:1 to 32:1. When the imbalance ratio increases to 32:1, the number of positive samples is only eight. The positive samples are too sparse to represent the true distribution. It is thus meaningless to decrease the number of positive samples further. The test data consists of 1,000 positive samples and 1,000 negative samples. The experiment was repeated ten times and the average value of the *BCR*s achieved by different schemes are plotted in Fig. 7.12. The comparison includes the following:

- *Oversampling:* The positive class was randomly oversampled (duplication) so that the training set is balanced.
- *Undersampling:* The negative class was randomly under-sampled so that the training set is balanced.
- *SMOTE:* A balanced dataset was created by adding some artificially generated data in-between the three nearest neighbors of each data point in the original dataset.
- *Using different costs to the two classes:* The *BSVC* was trained using different costs to the two classes. The primal problem in (7.9) becomes

$$\min_{w,b,\theta}\{\frac{1}{2}\|w\|^2 + +C_+\sum_{i=1}^{N_+}\theta_{i+} + C_-\sum_{i=1}^{N_-}\theta_{i-}\} \tag{7.42}$$

where N_+ and N_- are the numbers of positive and negative samples, respectively. ($N_+ < N-$) and θ_{i+} and θ_{i-} are the errors of positive and negative samples, respectively. The regularization parameter becomes:

$$C_+ = \frac{C}{2N_+}, \quad y_i = +1 \tag{7.43}$$

and

$$C_- = \frac{C}{2N_-}, \quad y_i = -1 \tag{7.44}$$

Hence the error of minority negative class is penalized more than for the majority positive class in order to compensate for the class imbalance.

The parameters of all the *BSVC*s are optimized using threefold cross validation. The parameters of the *νSVC* are optimized using artificially generated outlier data aforementioned. The *BCR* achieved by *HKME* using *AVG*, *DET*, *LDC*, and *QDC* fusion rules are shown in Fig. 7.13.

It can be observed from Fig. 7.12 that discriminative *BSVC* (trained using original dataset) perform well when the imbalance ratio is not very high, but its performance deteriorates with the increasing imbalance ratio. *HKME* using *AVG* rule performs the best among all the approaches. The *BSVC* trained using different costs to the two classes perform quite well compared to the *BSVC* trained using the same cost to the two classes. Undersampling performs better than original *BSVC*, but is outperformed by using different costs. SMOTE performs reasonably well. It is better than both original *BSVC* and *νSVC*. Oversampling performs the worst among all the approaches.

The good performance of *HKME* may come from the fact that it benefits from the strength of both of its individual classifiers in the ensemble, the discriminative *BSVC* and recognition-based *νSVC*. This can be explained using their decision boundaries as illustrated in Fig. 7.14. *νSVC* performs well due to its ability to model compactly

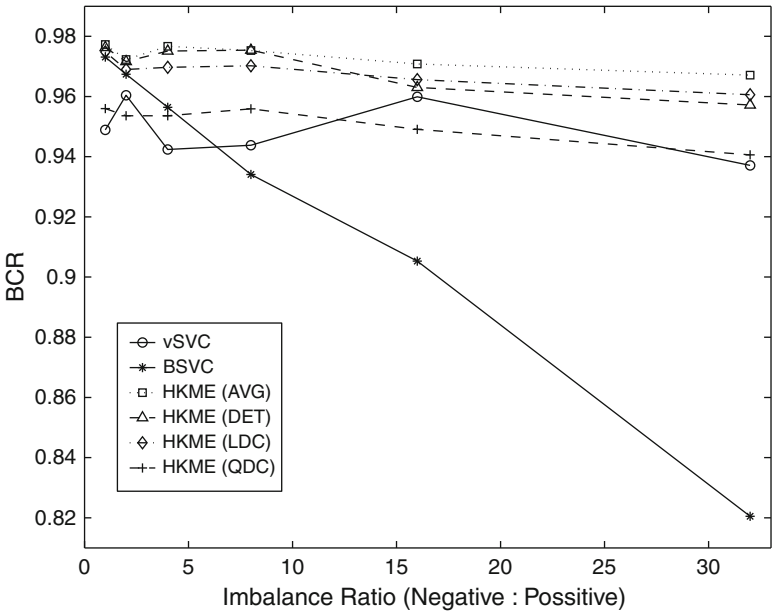

Fig. 7.13 The comparison of different fusion rules for *HKME* using 2×2 checkerboard dataset with different imbalance ratio

clustered target class. But it has to reject some target samples to form a tighter boundary as mentioned in Sect. 7.2.2, so it tends to push the decision boundary towards the majority (negative) class. However, discriminative *BSVC* tends to push the decision boundary toward the minority positive class. The ensemble of these two *SVM* tends to compensate these two different trends and strike a balanced compromise. As shown in the figure, the decision boundary of *HKME* is located in-between two classifiers, which is closer to the ideal decision boundary (two squares in the checkerboard).

The *HKME* using four different fusion rules are compared in Fig. 7.13. *AVG*, *DET*, and *LDC* performs well. But *QDC* does not perform well in some cases. This is because the covariance matrices for the classes are nearly singular in these cases, *QDC*s failed when trying to estimate and invert the covariance matrices [25]. However, it still performs quite well when it is properly trained.

The performance of other methods in Fig. 7.12 may also be explained using their decision boundaries on a checkerboard dataset with 256 negative samples and 16 positive samples, as shown in Fig. 7.15. The *BSVC* tends to push the decision boundary toward the minority class as aforementioned. Using different costs to the two classes, the decision boundary tends to be closer to the majority negative class as shown in Fig. 7.15a. So this approach performs better than *BSVC* trained using original dataset. The artificially generated positive data using *SMOTE* seems

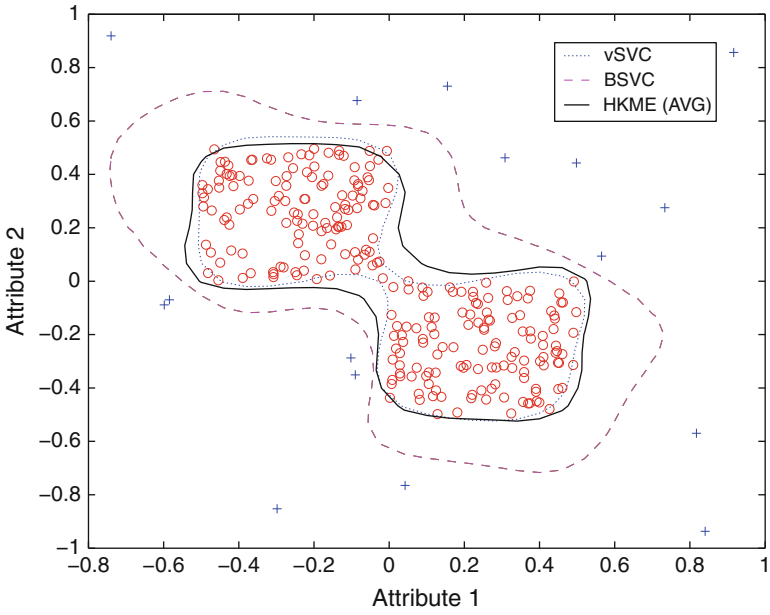

Fig. 7.14 The decision boundaries of *vSVC*, *BSVC*, and *HKME* using 2×2 checkerboard dataset (256 negative samples and 16 positive samples) [26]

closer to its original distribution in Fig. 7.15b, which makes the *BSVC* trained using *SMOTE* performs much better than the others. But it must satisfy the assumption that the samples between the nearest neighbors of a sample are from the same class. Due to the duplication of the minority samples in oversampling, the *BSVC* overfits the minority positive class, which is clearly illustrated in Fig. 7.15c. This leads to poor performance of oversampling shown in Fig. 7.12, especially when the imbalance ratio is high. Therefore, random oversampling the minority data is not suitable in *BSVC* training for imbalanced datasets. Undersampling the majority class seems to produce better decision boundary than that using oversampling. But the shape of the decision boundary is quite different from the ideal one as shown in Fig. 7.15d. This may be because that some useful information is lost when some samples from the majority class are removed from the training set. This detrimental effect is especially obvious when the size of the minority class is very small.

To sum up, *HKME* performs well in the checkerboard dataset. SMOTE and using different costs to the two classes seem quite efficient for this dataset. Random undersampling is better than the *BSVC* trained using original imbalanced dataset. Random oversampling is not suitable for *BSVC* when the imbalance ratio is high.

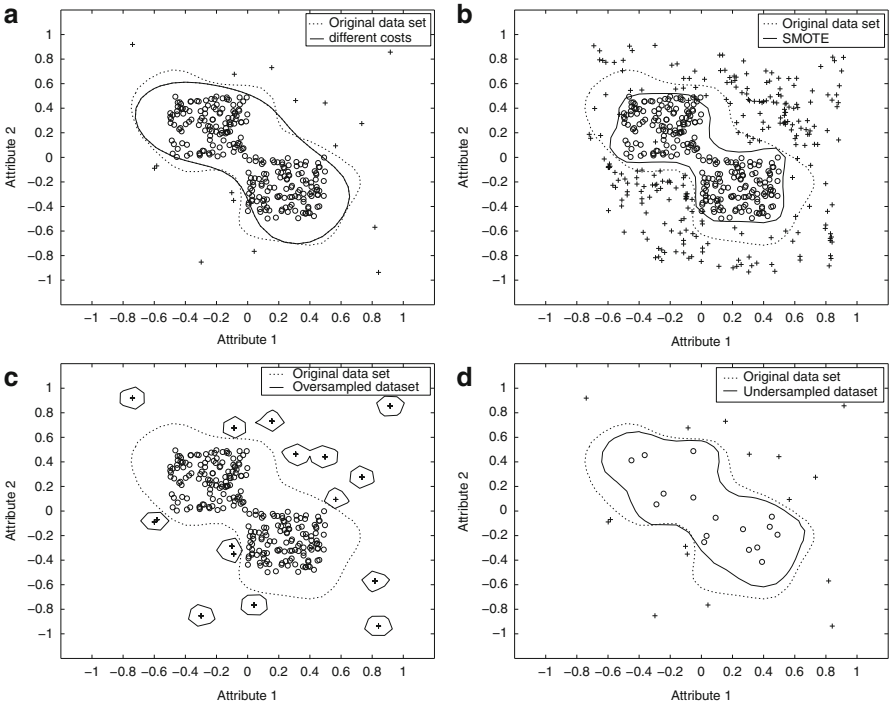

Fig. 7.15 The decision boundary of *BSVC* training using original dataset and those of *BSVC*s trained using (**a**) Different costs to two classes, (**b**) SMOTE, (**c**) Over-sampled dataset, and (**d**) Under-sampled dataset

7.6 Application of HKME in ECG Annotation and Colonoscopic Image Analysis

Since *HKME* is seen to perform well on the artificial dataset, it is deployed in two biomedical applications in which the problem of class imbalance exists and sharing similar properties in the distribution of the class samples to the artificial dataset.

7.6.1 Abnormal ECG Beat Annotation for Long-Term Monitoring of Heart Patients

7.6.1.1 Abnormal ECG Beat Annotation

ECG is a recording of the heart's electrical activity obtained from electrodes attached on the body surface of a patient [57]. Different segments of the ECG signal characterize different cardiac activities. A typical normal ECG beat is illustrated in Fig. 7.16.

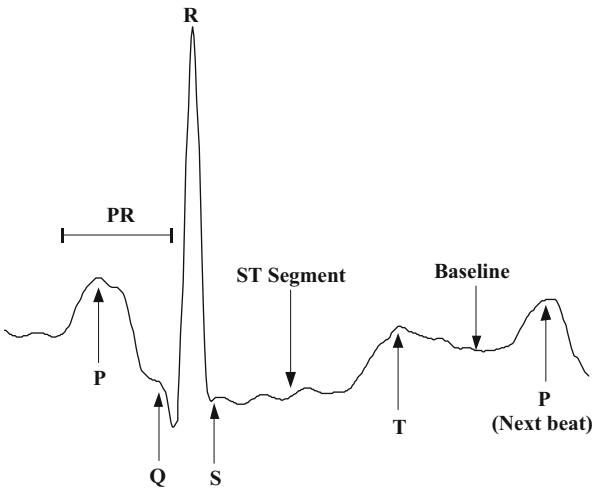

Fig. 7.16 A typical normal ECG beat

The analysis of heart beat cycles in ECG signal is very important for long-term monitoring and diagnosis of the patients' heart conditions in an intensive care unit or at patients' homes through a telemedicine network. However, it is very costly for the physicians to analyze the ECG recordings beat by beat since the ECG recordings may last for hours. Therefore, it is significant to develop a computer-assisted technique to examine and annotate the ECG recordings automatically, so to facilitate review by medical experts. This computer annotation will assist physicians to select only the informative (abnormal) beats for further analysis.

7.6.1.2 Generalization and Imbalanced Data Problem in ECG Beat Annotation

Generalization Problem

A fundamental assumption in the field of pattern recognition is that the underlying distribution of the training samples is the same as that of the test samples. However, such assumption may not hold in practical application. The abnormal ECG beat annotation problem is one of the examples. Figure 7.17 illustrates the distribution of the first two principal components of the original $181 - dimensional(D)$ feature vector of ECG beats obtained by using Karhunen–Loeve transform (PCA) from 4 recordings of MIT/BIH arrhythmia database [13], preserving 69 % of the total variance, where the circles indicate normal ECG beats and the cross signs are abnormal ones. Although some discriminative information may be lost using PCA, it can be observed that the distributions of "normal" ECG beats are different among patients.

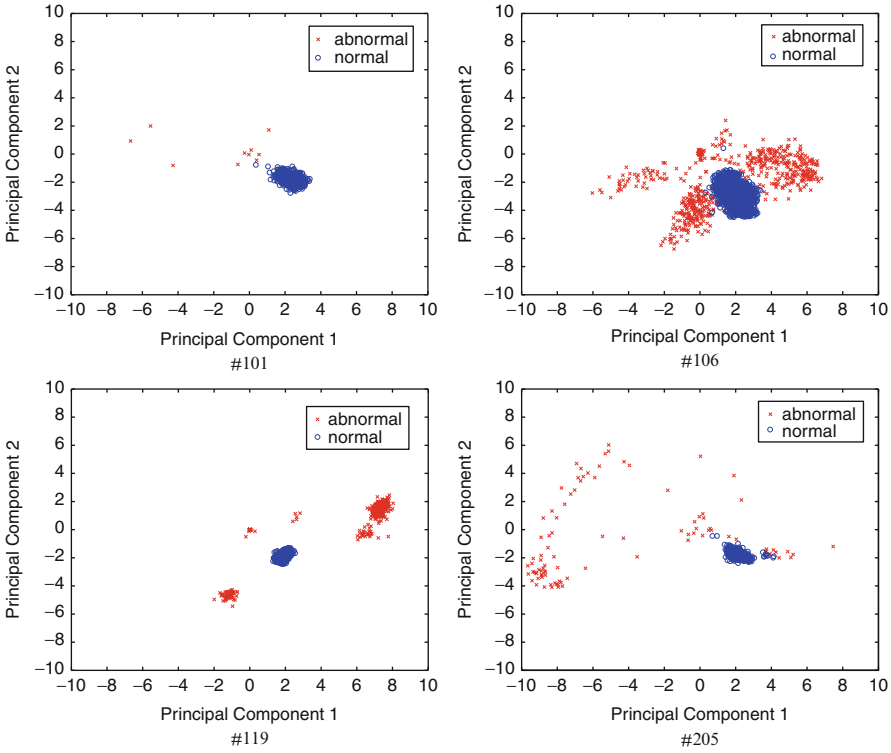

Fig. 7.17 Scatterplot of ECG data of four patients in MIT/BIH arrhythmia database [13] showing the first two principal components of PCA projection

Figure 7.18 illustrates the distribution of the ECG data from 44 recordings of MIT/BIH arrhythmia database using same PCA projection. Although an ECG detector can be finely trained using the ECG beats from a large database which consists of the ECG beats from different patients, it may perform poorly in annotating the ECG beats of other patients who are not in the database. This is the problem of poor generalization.

The solution to such generalization problem lies in the incorporation of local information of a specific patient to the ECG annotator. Since the distribution of the training samples is not the same as that of the test samples, some information about the true distribution of samples from each patient has to be added to train the ECG annotator properly.

In long-term monitoring of patients suffering from cardiovascular diseases, the normal ECG beats usually dominate the ECG recordings such as in patients suffering from or suspected to suffer from asymptomatic heart failure, congestive heart failure, cardiac dysfunction, and cardiac arrhythmias, etc., i.e. the number of abnormal ECG beats is far less than that of the normal ones. It may take a long time

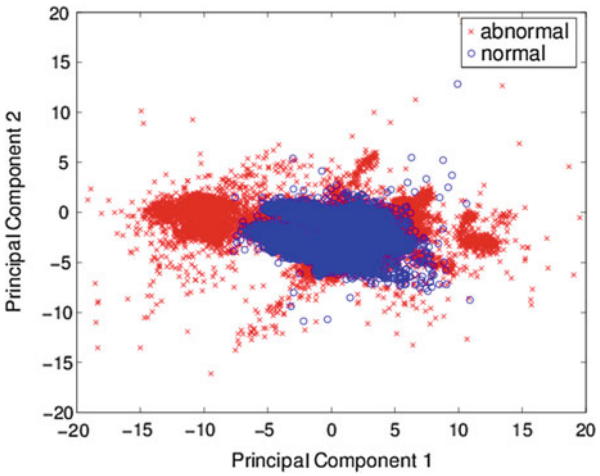

Fig. 7.18 Scatterplot of ECG data of all the ECG data from 44 patients in MIT/BIH arrhythmia database showing the first two principal components of PCA projection

to collect sufficient and balanced normal and abnormal ECG data to construct a good classifier; otherwise, the classifier may suffer from the imbalanced data problem [19].

7.6.1.3 HKME for ECG Annotation

One-Class Classification-Based Approach

A straight way to solve the generalization and imbalanced data problem aforementioned is recognition-based approach. A one-class classifier, *νSVC* can be trained using only about 5 min of normal ECG beats from a patient to adapt to the specific reference value of the patient. The trained model can then be used to determine whether the other ECG beats from the same patient belong to the "normal beats." Hence the abnormal ECG beats can be annotated automatically for further analysis. Such model has been proposed in [28].

As there is an innate difference between the normal range of each patient and that of a group of patient, there is a need to incorporate the local information of each patient to improve the generalization of the ECG beat annotator. Since the distribution of training data of the *νSVC* model is more similar to the ECG beats of the same patient than those of the data from a large group of patients, the *νSVC* model trained properly using about 5 min of the "normal" ECG data from a patient is expected to perform better than the classifiers trained using the data from a large group of patients.

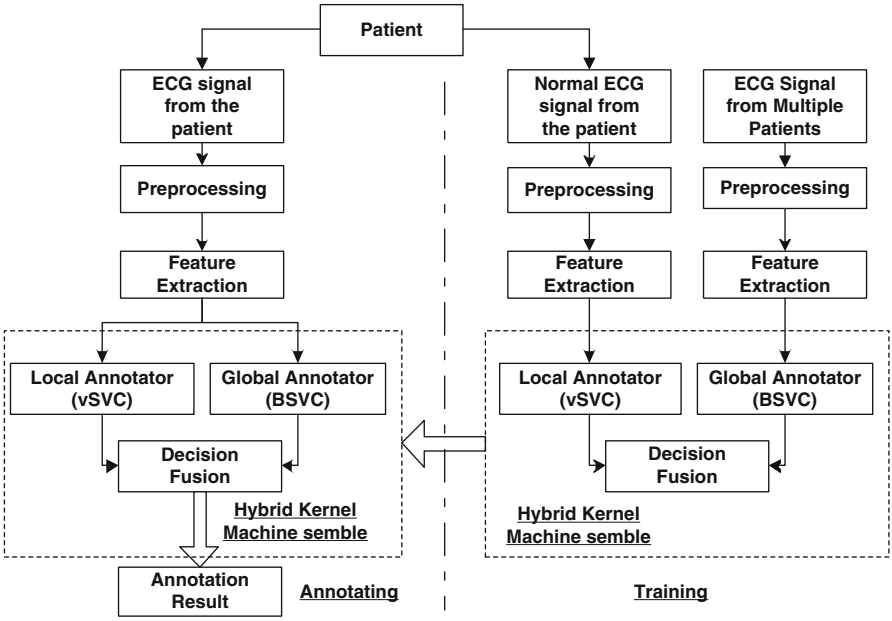

Fig. 7.19 Flowchart of the proposed framework for abnormal ECG beat annotation

HKME-Based Approach

Although the one-class classification-based approach proposed in the previous section performs well in the abnormal ECG beat annotation, it may be possible to be further improved. When a physician examines the ECG recordings of a patient, the physician considers not only the specific reference from each patient to be examined but also the standard reference from the patient-group due to the innate difference between the normal range of each patient and that of a group of patient aforementioned. That is to say, the diagnosis made by the physician is based on the information from both the patient-group and each specific patient. Motivated by this, a *HKME*-based approach is proposed for the ECG beat annotation problem for long-term monitoring heart patients. The following is based on authors' earlier publication in [27].

Figure 7.19 illustrates the flowchart of the proposed Hybrid Kernel Machine Ensemble (*HKME*)-based ECG beat annotator. This *HKME* consists of two base classifiers, one is a binary *SVM* trained using the ECG data from a large group of patients, the other is a one-class classification model, *νSVC* trained using only about 5 min of normal ECG beats from each patient to be monitored. The final decision is determined by a fusion rule. The recognition-based *νSVC* has been described in the previous section. It represents the specific reference value of the patient. The discriminative binary *SVM* is incorporated the global information of a large group of people and thus it can be regarded as the reference values based on the general

patient population. Due to different information learned by these two *SVM*s, they usually perform differently in classifying the ECG beats in the long-term ECG recording of the patient. Furthermore, *νSVC* is a non-discriminative recognition-based model and *BSVC* is a discriminative model. Due to the complementary nature of such two types of *SVM*s, integration of the two types of kernel machines using an ensemble is expected to perform better than using either of them separately.

In this study, the raw amplitude of the time domain ECG signals after noise suppression and baseline shift removal was investigated as feature vectors to represent the ECG beats. After the R-peak is detected, the ECG signal in a window of 500 ms is taken as an ECG beat. The lengths of the signal before and after the R-peak in each beat are 167 and 333 ms, respectively, such that the window covers most of the characterization of the ECG beat (for an ECG signal sampled at 360 Hz, 180 samples around each R-peak are taken in a window, with 59 samples before the R-peak and the other 120 samples behind the R-peak). The amplitude of sampled signal in each window is then taken to form a feature vector of 180-dimensions. It has been shown that R–R interval (the interval between two consecutive R-peaks) is useful in recognition of some abnormal ECG beats [6, 16]. Therefore, it is also included in this study by appending it to the 180-dimensional(D) feature vector. The length of the feature vector to represent the ECG beat is then 181.

Normalization

There are some variations in the amplitude ranges of ECG signals among the human beings. Hence a normalization procedure to the ECG feature vectors is necessary; otherwise, the ECG beats may not be comparable. The feature vectors are divided by the mean value of R-peaks in the training data of each patient, such that the maximum amplitude in each ECG beat window is around 1. The normalized ECG feature vectors are then used for the annotation process using the trained HKME models.

7.6.1.4 Experiments

Experimental Setting

The proposed *HKME*-based patient-adaptable ECG beat annotator was evaluated on MIT/BIH arrhythmia database [13]. The experimental setting is as follows.

1. 22 recordings are selected as local training sets and test sets, in which the number of abnormal beats is significantly less than that of the normal beats, to be in agreement to the scenario in long-term monitoring of patients suffering from cardiovascular diseases. These recordings include # 100, 105, 106, 108, 114, 119, 121, 200, 203, 205, 208, 209, 210, 213, 215, 221, 222, 223, 228, 230, 233, and 234. Each of the 22 recordings is split into two sets.

Table 7.2 Results (average ± standard deviation) of abnormal ECG beat annotation (in percentage)

Classifiers	BCR	SEN	SPE	ACR
BSVC	80.3 ± 16.9	81.3 ± 24.5	79.3 ± 29.6	80.2 ± 26.1
νSVC	83.6 ± 14.7	87.5 ± 22.5	79.7 ± 21.3	81.7 ± 18.5
MAX	86.2 ± 16.4	87.0 ± 21.9	85,4 ± 20.6	85.8 ± 19.5
LDC	86.5 ± 16.2	82.6 ± 27.3	90.5 ± 16.1	90.1 ± 14.1
QDC	83.1 ± 17.3	74.6 ± 35.1	91.6 ± 12.2	90.3 ± 10.3
DET	87.5 ± 15.0	83.3 ± 26.6	91.6 ± 13.7	91.2 ± 11.5

- The first 200 normal ECG beats in each of the 22 recordings (about 3 min) are used as the local training set to construct the *νSVC*s.
- The first 350 normal ECG beats in each of the 22 recordings (about 5 min) are used as the training set to train the ensembles.
- The second $1/2$ of each of the 22 recordings (about 15 min or 1,000 beats) is used as test set to evaluate the performance of the ECG annotators.

2. 10,000 ECG beats (with half normal beats and half abnormal beats) from 22 recordings are used as global training set (DB_G) to train some classical binary classifiers for comparison with the proposed *HKME*-based patient-adaptable ECG beat annotator. These recordings are # 101, 103, 109, 111, 112, 113, 115, 116, 117, 118, 122, 123, 124, 201, 202, 207, 212, 214, 220, 222, 231, and 232.

Results and Discussion

The annotation results of using the proposed *HKME* with different fusion rules, the global binary *SVM* and the local *νSVC* are given in Table 7.2.

The reported results are averaged over 22 test ECG recordings as shown in Fig. 7.20. Test sets #1 – #22 correspond to recording 100, 114, 119, 121, 200, 203, 205, 208, 209, 210, 213, 215, 221, 222, 223, 228, 230, 233, 234, 105, 106, and 108 in MIT/BIH arrhythmia database, respectively.

- *Overall Performance and Generalization*

  It is observed from Table 7.2 that all the *HKME*s except using QDC rule outperforms both the global *BSVC* trained using the ECG data from a large patient-group and the local *νSVC* trained using some normal ECG data from each patient. The generalization of both the global *RSVC* and the local *νSVC* can be improved when the information from these two sources are integrated properly by the *HKME*.

  The best BCR achieved by *HKME* is using DET rule, whose BCR is 7.2 and 3.9 % higher than the global *RSVC* and the local *νSVC*, respectively. DET-based *HKME* outperforms both *SVM*s in more than 80 % of the test sets as reported in Table 7.3. The second best BCR achieved by *HKME* is using LDC-based stacking rule, which outperforms both *SVM*s in about 72 % of the test sets.

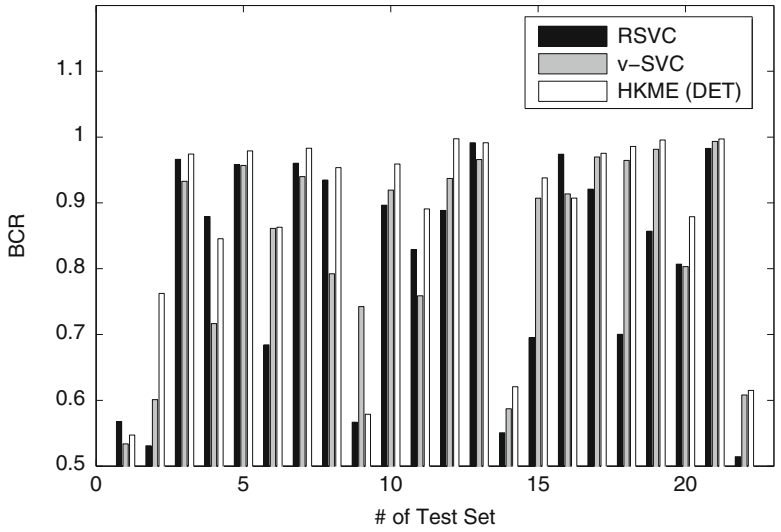

Fig. 7.20 Comparison of the annotation results of the global *BSVC*, local *νSVC* and *HKME* (with DET fusion rule) in each recording of the test sets in terms of BCR

Table 7.3 Number and percentage of test sets in which *HKME* outperforms both global *BSVC* and local *νSVC* among all 22 test sets

Fusion rule	*MAX*	*LDC*	*QDC*	*DT*
In number	14	16	11	18
In percentage	63.6	72.7	50.0	81.8

The performance improvement in terms of BCR using MAX rule is observed in about 64 % of the test sets. Its average of BCR is 5.9 and 2.6 % greater than the global *BSVC* and the local *νSVC*, respectively. The only exception is QDC rule. This may be resulted by the fact that the covariance matrices for the classes are near singular sometimes, these QDC classifiers may fail when trying to estimate and invert the covariance matrices [25]. It is expected that its performance may be better if it is properly trained.

It can be observed that the performance of trained *HKME*s such as DET and LDC is better than that of the non-trained *HKME*, the MAX rule. It shows that the proper training of the fusion rules is helpful to the improvement of the ensemble over the base classifiers. These findings are similar to those in previous section.

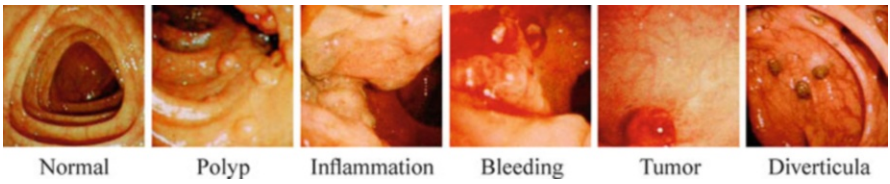

Fig. 7.21 A normal colonoscopic image and five colonoscopic images with different types of abnormalities

7.6.2 Application to Colonoscopic Image Analysis

7.6.2.1 Colonoscopic Image Analysis

In this section, an application of the HKME method to the detection of abnormal region in colonoscopic images is presented. This work has been published in [29]. Colonoscopy is a minimal invasive procedure of screening the colon and rectum using a colonoscope. The procedure is used to look for signs of cancer in the colon and rectum and diagnose the causes of unexplained changes in the bowel such as inflamed tissue, abnormal growths, ulcers, and bleeding. Analyzing colonoscopic images for clinical diagnosis of abnormalities relies on the experience and expertise of the medical experts, which need years of training to acquire. It is thus significant to develop a computer-assisted technique to help the screening process of these potentially lethal diseases by the health-care provider.

Previous research on colonoscopic image analysis focused on the classification between normal tissues and tumors. However, few work has been done to discriminate normal tissues from different kinds of abnormalities including tumors in colonoscopic images, which is more significant for screening purpose. In fact, many categories of abnormalities can be seen in colonoscopic images, such as polyps, tumors, inflammation, bleeding, ulceration, and diverticula (Fig. 7.21) and their image content shows large variations. The abnormal regions usually do not occupy the whole image and vary in color, size, and shape, which add more difficulties to the discrimination of the normal regions from the abnormal ones in colonoscopic images. In addition, this leads to an good example of imbalanced data problem.

The patch-based approach seems to be a good representation model for image segmentation in which the full image is cropped into a set of image patches and these patches can be classified into different categories corresponding to different types of segments. This approach has been used extensively in many applications, such as face detection [15], object detection [7], and image segmentation [42].

In this section, a *HKME*-based approach using multi-size patches is presented for detecting abnormal regions in colonoscopic images. Multiple sizes of patches provide multiple level visual cues of the image regions, which can help produce better perceptually agreeable segmentation. Represented as multi-size patches, the abnormal region detection in colonoscopic images turns into a binary classification

problem to discriminate the patches from normal regions (normal class) and those from abnormal ones (abnormal class). Each pixel in a given image can be categorized as normal or abnormal using a trained patch-based classifier. Using multiple sizes of patches, multi-labels can be given to a pixel, the final label of the pixel can be obtained using the ensemble of these multiple classifiers based on different patch sizes.

The performance of the ensemble depends on the individual classifiers used. A set of individual classifiers have to be trained for the binary classification problem to discriminate the normal patches from those abnormal ones. This problem can be solved using a discriminative model, such as *BSVC*s . Such a *BSVC*-based abnormal region detection approach in colonoscopic images using multiple-size patches was published in [30] (A preliminary work by the authors).

Rather than a typical binary classification problem, the abnormal region detection in colonoscopic images can also be treated as a one-class classification problem. A lot of patterns from abnormal regions in colonoscopic images for each categories of abnormalities have to be collected for training a reliable classifier, which means the concept "abnormal" is not easy to learn. On the other hand, the normal patterns show smaller variations than those of the abnormal ones and are much easier to be obtained. This means the concept "normal" can be easier to learn. Therefore, the concept "normal" can be learned using a one-class classifier, such as *νSVC* or *SVDD*. Such a one-class classification-based abnormal region detection approach in colonoscopic images was published in [31] (Another preliminary work by the authors). Trained using only the data from one class, *νSVC*s try to find a decision boundary around the training data—called targets, which is different from the decision boundary of *BSVC* trained using the data from both normal and abnormal classes. As explained in the previous section, *νSVC* tries to represent of target samples rather than for discrimination purpose. On the one hand, multi-size patches produce multi-level cues of image content, which in turn produce a diverse feature set. On the other hand, the combination of the two different types of kernel machines *νSVC* and *BSVC* can produce more diversity to the ensemble, which may further improve the abnormal detection in colonoscopic images. Experimental results show that the multi-size patch-based hybrid kernel machine ensemble method is superior to that of using single patch size only for the abnormal region detection in colonoscopic images and can produce more perceptual agreeable image segmentation.

7.6.2.2 HKME for Detecting Abnormal Regions in Colonoscopic Image Analysis

Figure 7.22 illustrates the flowchart of the proposed *HKME*-based approach using multi-size patches for abnormal region detection in colonoscopic images. The detail is as follows.

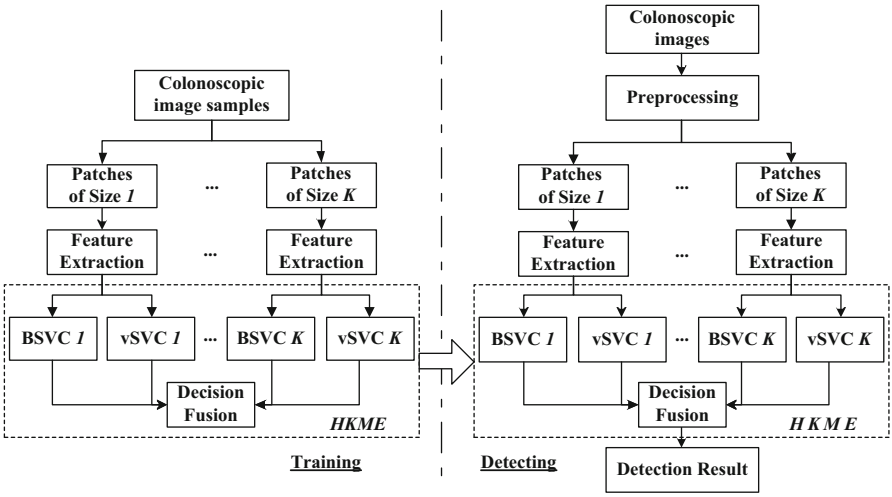

Fig. 7.22 The flowchart of *HKME*-based approach for abnormal region detection in colonoscopic images

Image Region Representation Using Multiple-Size Patches

As illustrated in Fig. 7.21, the abnormal regions in colonoscopic images come from different categories and they vary in location, shape, color, and size. The representation of these regions has to be considered carefully. It is similar to object detection in which the abnormal regions are the objects to be detected.

Patch-based approach turns the abnormal region segmentation into a binary classification problem [21]. As illustrated in Fig. 7.23, each colonoscopic image can be cropped into a set of overlapping image patches and these image patches can be categorized as abnormal region class or normal region class by a classifier. The abnormal regions can thus be segmented from the normal ones.

An open problem in patch-based approach is what patch-size to choose. Compared to large ones, small-size patches (the extreme of a small-size patch is single pixel) can represent the image regions more precisely, but it contains less information of the image content than large-size patches (the extreme of large-size patch is the full image) and usually lead to larger classification error. The large-size patches contain more information of the object, but small abnormal regions in these patches may be missed. This is why it is very difficult to determine the appropriate patch-size to use.

In this section, a multi-size patch-based representation is presented in which multi-size patches are used simultaneously to represent the image regions in colonoscopic images. Using patches of multiple sizes aims at overcoming the scale problem, i.e, an abnormal region may appear at different sizes in different images. Multi-size patches provide multiple-level representation of the image contents. At least some among all the patch sizes can better characterize the object. Hence, the

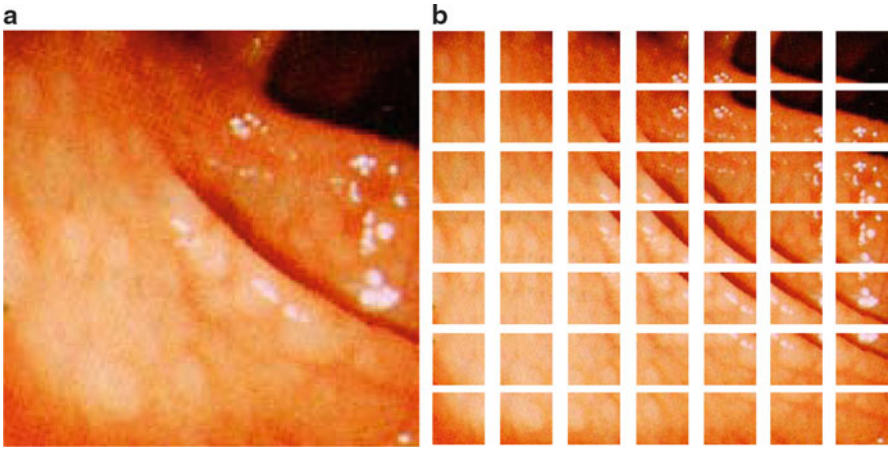

Fig. 7.23 Patch-based image region representation. (**a**) Original colonoscopic image, (**b**) Overlapping image patches

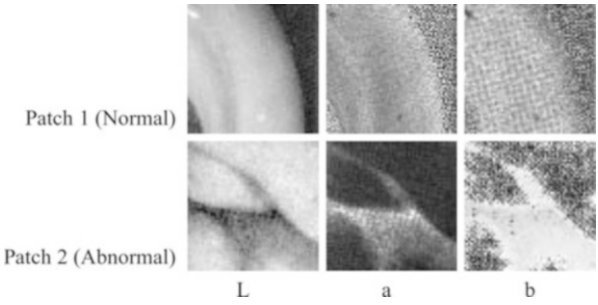

Fig. 7.24 Two examples of patches from colonoscopic images. The component L, a and b of the image patches are illustrated listed in a row

integration of the detection result based on multi-size patches is expected to detect the abnormal regions more precisely than those based on single-size patches only. This is the novelty of this method.

The colonoscopic images in RGB color space are transformed into three bands in *CIELab* color space through which the color and luminance component can be processed and analyzed individually. The image is scanned across and cropped into a set of fixed sizes of patches, respectively. The patches are overlapped by 50 % to ensure that no abnormal region is missed. Here, 3 sizes of the image patches are investigated for abnormal region detection in the colonoscopic images, namely, 48×48, 32×32 and 16×16 (pixels). Two samples of the image patches are illustrated in Fig. 7.24. Feature can be then extracted from these image patches for classification.

Both color and texture features are extracted.

- *Color features:* The color features are Two-dimensional(D) histograms of the components *a* and *b* in *CIELab* color space. The number of bins of the histogram is 8 for 2-D histograms.

- *Textural features:* Two-level Discrete Wavelet Transform (DWT)-based statistical features and 1-D histograms of luminance (the number of bins of the 1-D histogram is 16) are employed as textual features. The image patches are processed using two-level DWT. The mean and standard deviation of the absolute value of the approximate and detailed coefficients from the two-level DWT decomposition of the image patches in the three channels of the *CIELab* color space are calculated as the textural features.

Altogether 128 features are extracted, giving rise to a feature vector of 128-D. Then the feature vectors from patches can be used to form the dataset for classification. A set X of N feature vector x_i, $X = \{x_i \in R^{128} | i = 1, 2, \cdots, N\}$ for N patches, are labelled as $y_i \in \{+1, -1\}$ to indicate whether it is a normal patch or a patch containing abnormalities.

Learning *SVM*s for Image Patch Classification

BSVC Learning

Using overlapped image patches, each pixel in the patch can be classified as normal or abnormal by an *SVM* classifier corresponding to the patch size. Thus each pixel in the original image can have at least one label. If a pixel is classified differently by overlapped patches, the label of the patch that has the largest absolute decision value (confidence) is chosen as the label of that pixel.

νSVC Learning

The classification between normal and abnormal patches can also be solved by a one-class classifier, such as *νSVC*. A *νSVC* can be trained using the data from normal image patches for each patch size. Using overlapped image patches, each pixel in the patch can be classified as normal or abnormal by the trained *νSVC* classifier corresponding to the patch size. Thus each pixel in the original image can have at least one label. If a pixel is classified differently by overlapped patches, the label of the patch that has the largest confidence is chosen as the label of that pixel.

7.6.2.3 Decision Fusion Using *HKME*

Since there are many kinds of abnormalities in colonoscopic images showing large variation, many patterns from abnormal regions in colonoscopic images have to be collected for training a reliable classifier and it is difficult to collect. This leads to an imbalanced data problem. One class—"normal" has many training samples and is easier to model, while the other class—"abnormal" is difficult to model because it has more diverse distributions than the normal class. Therefore, *νSVC* is very suitable for this problem. As a recognition-based model, *νSVC* tries to describe the target data rather than for discrimination purpose, it can handle the problem of

missing information. However, *νSVC* is often inferior to *BSVC* for discrimination purpose. There is a need to combine these two types of kernel machines for this problem.

A set of 2-*SVC*s can be constructed for the classification, while *ν*-*SVC*s can be used to provide further decision information. The classification results of the two kernel machines can be aggregated using an ensemble. The different natures of the two types of *SVM*s adds more diversity to the ensemble, which may further improve the performance of the ensemble.

7.6.2.4 Experimental Results and Discussions

Data Preparation

The proposed approaches were evaluated using a database which consists of 58 clinically obtained colonoscopic images. There are 12 normal images and 46 images with abnormal regions. The abnormal regions mostly occupy only some parts of the whole image and the abnormalities include polyps, tumors, inflammation, bleeding, ulceration, and diverticula, etc. The images are RGB images with the resolution of 256×256 pixels. The pixels in the original images were manually labeled to provide the ground truths. The detection results were compared with the ground truth and evaluated.

In the experiment, the numbers of collected image patches for training of 48×48, 32×32, and 16×16 (pixels) patches are 2,002, 2,090, and 2,126, respectively. The pixels in the original image are manually labeled as the ground truth for comparison. The patches containing mostly abnormal region were labeled as a positive sample, otherwise, a negative one. A leave-one-out experiment was performed to evaluate the performance of the proposed method for abnormal region detection in colonoscopic images. In each round, one of the colonoscopic images was selected for testing and the patches from other 57 images were used for training. The experiment was repeated 58 times, the detected results were compared to the ground truth image and the average value of the total 58 results was taken as the final result.

Evaluation Measure

The evaluation criteria are specificity (SPE), sensitivity (SEN), and Balanced classification rate (*BCR*). Where SPE is the fraction of normal regions detected among all the normal regions, SEN is the abnormal regions detected among all the abnormal regions and BCR is the weighted average of SPE and SEN.

$$BCR = \lambda SPE + (1 - \lambda)SEN \tag{7.45}$$

Table 7.4 Results of abnormal region detection using single patch sizes

Patch size	Classifier	*BCR*	SPE	SEN
48×48	2-SVC	0.744	0.675	0.813
48×48	ν-SVC	0.539	0.991	0.088
32×32	2-SVC	0.738	0.675	0.802
32×32	ν-SVC	0.546	0.998	0.094
16×16	2-SVC	0.745	0.668	0.822
16×16	ν-SVC	0.538	0.946	0.094

where $\lambda \in [0,1]$ can be tuned to favor *SPE* or *SEN*. Smaller λ favors more on *SEN*, which means that the error on the abnormal class is punished more seriously. On the contrary, larger λ favors more on *SPE*, which means that the error on the normal class is taken more seriously. At the extreme case, only *SEN* or *SPE* will be considered when λ is 0 or 1, respectively. $\lambda = 0.5$ is used here, so that *SPE* and *SEN* are treated as equally important. Other values might be selected with respect to the requirement of the medical experts.

νSVC vs *BSVC* Using Single Patch Size

In Table 7.4, it is observed that *BSVC*s outperform *νSVC* in all the cases which agrees with the postulate that discriminative models are superior to that of recognition-based models. *BSVC*s achieved *BCR* around 74 %, while *νSVC*s achieved only 55 %. The *νSVC*s have a very high SPE, but almost completely fail for SEN. This may be resulted that the training set size used for *νSVC* was too small and it also suffered from the curse of dimensionality. Compared to *νSVC*s, *BSVC* have higher SEN while much less SPE, which may be good for adding more diversity to the ensembles. The best *BCR* is 74.5 % which was achieved using patches of size 16×16.

Multi-Size Patch Ensemble of *νSVC*s or *BSVC*s

Table 7.5 illustrates the detection results of three patch size ensembles using *νSVC*s or *BSVC*s separately. Obviously, the best ensembles outperform that of the best *SVM*s using single patch size, which supports the claim that multi-size patch-based *SVM* ensemble can achieve better abnormal region detection in colonoscopic images. Due to the poor performance of individual *νSVC*s, the improvement of their ensemble is limited although there are still some.

HKME Using Single-Size Patches

Table 7.6 shows the detection results of the ensemble of a *νSVC* and a *BSVC* based on single-size patches. Only *DET* and LDC achieved *BCR* comparable to the best

Table 7.5 Detection results (in terms of *BCR*) of abnormal region detection using different patch sizes and ensemble schemes

Ensemble	MAX	AVG	PROD	MV	*DET*	LDC	QDC
BSVC	0.737	0.751	0.745	0.751	0.753	0.763	0.765
νSVC	0.532	0.551	0.535	0.551	0.538	0.533	0.536

The ensembles are constructed using same types of *SVM*s, *BSVC* or *νSVC*

Table 7.6 Detection results (in terms of *BCR*) of abnormal region detection using same patch sizes by *HKME*

Patch size	MAX	AVG	PROD	MV	*DET*	LDC	QDC
48×48	0.551	0.551	0.551	–	0.744	0.746	0.540
32×32	0.556	0.556	0.556	–	0.738	0.741	0.548
16×16	0.563	0.563	0.563	–	0.745	0.745	0.736

single classifier and the performance of other ensembles did not outperform the best single classifier. This may be due to the fact that the *νSVC* and *BSVC* are trained using the same features, which limit the performance of this scheme.

HKME Using Multi-Size Patches

Table 7.7 illustrates the detection results of the *HKME* ensemble of *BSVC*s using all three patch sizes plus 1 to 3 *νSVC*(s) trained using 1 to 3 patch size(s). Most of the ensembles show improvement over the best single *SVM* based on single-size patches. The performance of LDC and AVG outperforms others. Figure 7.25 illustrates the result of the ensemble of *BSVC*s using all 3 patch sizes and a *νSVC*(s) trained using patches with size of 48×48. Obviously, the detection results by the *HKME* ensemble is closer to the ground truth compared to those using single-size patches.

The results of detection of abnormal region using learned *HKME* for four colonoscopic images are illustrated in Fig. 7.25.

7.7 Conclusion

In this chapter, we briefly reviewed the one-class SVM and two-class SVM. Their principles of classification are discussed and the strengths and weaknesses for dealing with imbalanced datasets are illustrated with the checkerboard dataset. The imbalanced data problem is also discussed and the various ways of handling such a problem are also presented. The chapter shows that the one-class SVM and two-class SVM can be integrated into an ensemble classifier to form what we call the Hybrid Kernel Machine Ensemble—*HKME*. This ensemble classifier has been evaluated with artificial dataset. The evaluation results show the benefit

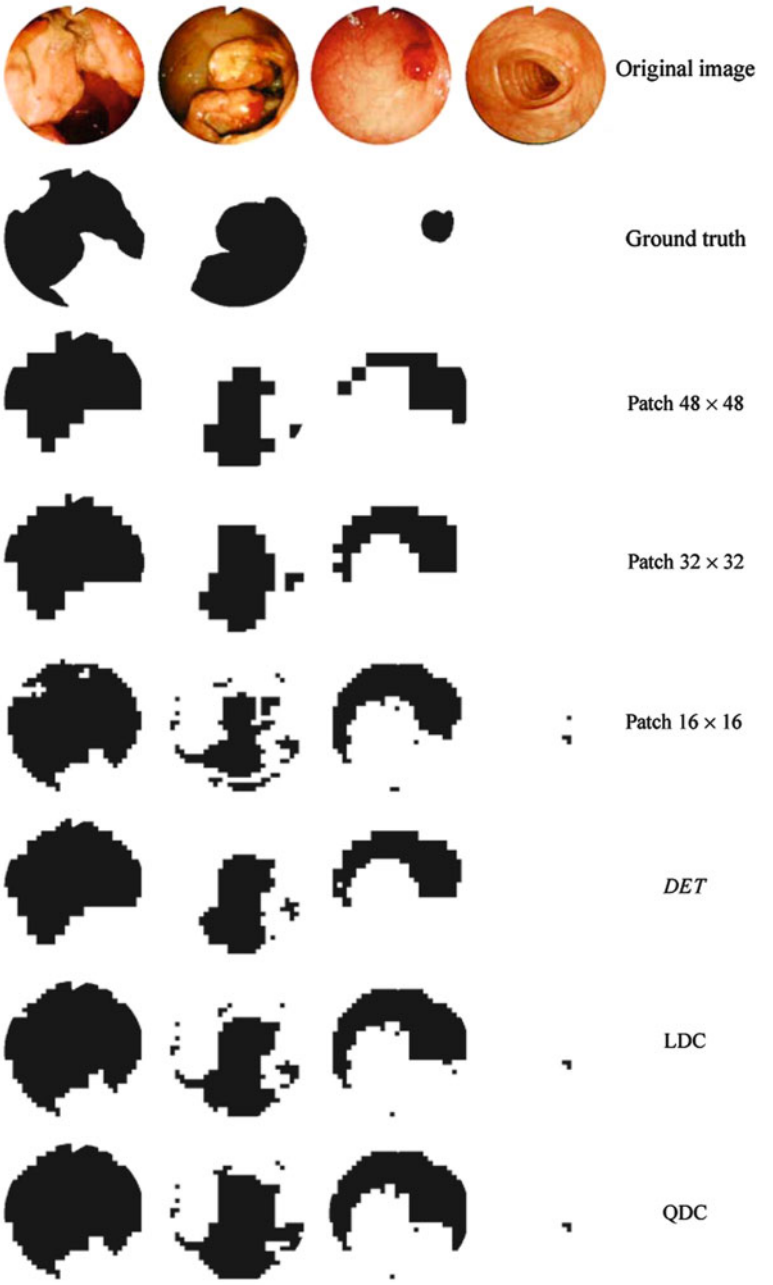

Fig. 7.25 Detection results of four colonoscopic images. The regions in *white* are normal regions detected and the regions in *black* are abnormal ones

Table 7.7 Detection results (in terms of *BCR*) of abnormal region detection using different patch sizes and *HKME*

Ensemble	MAX	AVG	PROD	MV	*DET*	LDC	QDC
A+1	0.705	**0.761**	**0.768**	0.751	0.753	**0.765**	**0.766**
A+2	0.541	**0.761**	0.659	0.751	0.753	**0.764**	0.667
A+3	0.588	**0.754**	0.730	**0.751**	**0.753**	**0.756**	0.730
A+1+2	0.539	**0.769**	0.598	**0.765**	0.753	**0.765**	0.542
A+1+3	0.587	**0.765**	0.656	**0.765**	0.753	**0.763**	0.743
A+2+3	0.538	**0.762**	0.572	**0.765**	0.753	**0.763**	0.559
ALL	0.537	0.565	0.548	0.704	0.751	**0.764**	0.549

Row $A+*$ are the results of ensembles using all 3-size patches learned by *BSVC*s plus 1-size or 2-size patches learned by *νSVC*(s) (1 for 48×48, 2 for 32×32, 3 for 16×16). Row *ALL* are the ensemble results using all 3-size patches and both *BSVC* and *νSVC*

of using such an ensemble to handle an imbalanced dataset. It has been shown that the *HKME* can achieve better performance than using either one-class SVM or two-class SVM alone. Discussions are given on the possible reasons for its better performance. Since such imbalanced data problem exists in many biomedical applications, and encouraged by the good performance of *HKME*, it is deployed in two biomedical applications, namely, abnormal ECG beats annotation and abnormal region detection in colonoscopic images. Experimental results further confirm the superiority of using *HKME*.

References

1. The Kernel-Machine.org http://www.kernel-machines.org/
2. Bach, F., Jordan, M.: Kernel independent component analysis. J. Mach. Learn. Res. **3**(1), 1–48 (2003)
3. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon, Oxford (1995)
4. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: synthetic minority over-sampling technique. Artif. Intel. Res. **16**, 321–357 (2002)
5. Chawla, N.V., Japkowicz, N., Kotcz, A.: Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations **6**(1), 1–6 (2004)
6. de Chazal, P., O'Dwyer, M., Reilly, R.: Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE T. Bio-Med. Eng. **51**(7), 1196–1206 (2004)
7. Deselaers, T., Keysers, D., Ney, H.: Discriminative training for object recognition using image patches. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 2, pp. 157–162 (2005)

8. Drummond, C., Holte, R.C.: C4.5, class imbalance, and cost sensitivity: why undersampling beats over-sampling. In: Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Data Sets II, vol. 11, Washington, DC (2003)
9. El-Naqa, I., Yang, Y., Wernick, M.N., Galatsanos, N.P., Nishikawa, R.M.: A support vector machine approach for detection of microcalcifications. IEEE T. Med. Imaging **21**(12), 1552–1563 (2002)
10. Estabrooks, A., Jo, T., Japkowicz, N.: A multiple resampling method for learning from imbalances data sets. Comput. Intell. **20**(1), 18–36 (2004)
11. Gal-Or, M., May, J.H., Spangler, W.E.: Assessing the predictive accuracy of diversity measures with domain-dependent asymmetric misclassification costs. Inform. Fusion J. (Special issue on Diversity in Multiple Classifier Systems) **6**(1), 37–48 (2005)
12. Gokturk, S.B., Tomasi, C., Acar, B., Beaulieu, C.F., Paik, D., Jeffrey, B.J., Yee, J., Napel, S.: A statistical 3D pattern processing method for computer aided detection of polyps in CT colonography. IEEE T. Med. Imaging **20**(12), 1251–1260 (2001)
13. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), e215–e220 (2000)
14. Hojjatoleslami, A., Sardo, L., Kittler, J.: An RBF based classifier for detection of microcalcifications in mammograms with outlier rejection capability. In: International Conference on Neural Networks, vol. 3, pp. 1379–1384 (1997)
15. Hsu, R.L., Abdel-Mottaleb, M., Jain, A.K.: Face detection in color images. In: Proceedings of 2001 International Conference on Image Processing, vol. 1, pp. 1046–1049 (2001)
16. Hu, Y.H., Palreddy, S., Tompkins, W.J.: A patient-adaptable ECG beat classifier using a mixture of experts approach. IEEE T. Bio-Med. Eng. **44**(9), 891–900 (1997)
17. Japkowicz, N.: The class imbalance problem: significance and strategies. In: Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000), vol. 1, pp. 111–117 (2000)
18. Japkowicz, N., Myers, C., Gluck, M.: A novelty detection approach to classification. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, pp. 518–523. Morgan Kaufmann, San Francisco, CA (1995)
19. Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**(5), 429–450 (2002)
20. Karakoulas, G.J., Shawe-Taylor, J.: Optimizing classifiers for imbalanced training sets. In: Proceedings of the 1998 conference on Advances in Neural Information Processing Systems II, pp. 253–259 (1999)
21. Karkanis, S.A., Iakovidis, D.K., Maroulis, D.E., Karras, D.A., Tzivras, M.D.: Computer aided tumor detection in endoscopic video using color wavelet features. IEEE T. Inf. Technol. B. **7**(3), 141–152 (2003)
22. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE T. Pattern Anal. **20**(3), 226–239 (1998)
23. Kubat, M., Holte, R., Matwin, S.: Detection of oil-spills in radar images of sea surface. Mach. Learn. **30**, 195–215 (1998)
24. Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: one-sided selection. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179–186. Morgan Kaufmann, Nashville, Tennessee (1997)
25. Kuncheva, L.I., Bezdek, J., Duin, R.: Decision templates for multiple classifier fusion: an experimental comparison. Pattern Recogn. **34**(2), 299–314 (2001)
26. Li, P., Chan, K.L., Fang, W.: Hybrid kernel machine ensemble for imbalanced data sets. In: 18th International Conference on Pattern Recognition (ICPR'06), vol. 1, pp. 1108–1111 (2006)
27. Li, P., Chan, K.L., Fu, S., Krishnan, S.M.: An abnormal ECG beat detector approach for long-term monitoring of heart patients based on hybrid kernel machine ensemble. In: International Workshop on Multiple Classifier Systems (MCS 2005), Lecture Notes in Computer Science, vol. 3541, pp. 346–355. Springer (2005)

28. Li, P., Chan, K.L., Fu, S., Krishnan, S.M.: Neural networks in healthcare: potential and challenges. In: A Concept Learning-Based Patient-Adaptable Abnormal ECG Beat Detector for Long-Term Monitoring of Heart Patients, pp. 105–128. Idea Group Publishing, Hershey, PA (2006)
29. Li, P., Chan, K.L., Krishnan, S.M.: Learning a multi-size patch-based hybrid kernel machine ensemble for abnormal region detection in colonoscopic images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 670–675 (2005)
30. Li, P., Chan, K.L., Krishnan, S.M., Gao, Y.: Detecting abnormal regions in colonoscopic images by patch-based classifier ensemble. In: 17th International Conference on Pattern Recognition (ICPR), vol. 3, pp. 774–777. Cambridge, UK (2004)
31. Li, P., Krishnan, S.M., Chan, K.L., Gao, Y.: Abnormal region detection in colonoscopic images using novelty detection technique. In: Proceedings of 7th International Workshop on Advanced Imaging Technology (IWAIT'2004). Singapore pp. 139–154, MIT, Cambridge, USA (2004)
32. Manevitz, L.M., Yousef, M.: One-class SVMs for document classification. J. Mach. Learn. **2**, pp. 139–154. MIT, Cambridge, USA (2001)
33. Markou, M., Singh, S.: Novelty detection: a review-part 1: statistical approaches. Signal Process. **83**(12), 2481–2497 (2003)
34. Markou, M., Singh, S.: Novelty detection: a review-part 2: neural network based approaches. Signal Process. **83**(12), 2499–2521 (2003)
35. Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller, K.R.: Fisher discriminant analysis with kernels. In: Hu, Y.H., Larsen, J., Wilson, E., Douglas, S. (eds.) Neural Networks for Signal Processing IX, pp. 41–48. IEEE (1999)
36. Mika, S., Schölkopf, B., Smola, A., Müller, K.R., Scholz, M., Rätsch, G.: Kernel pca and de-noising in feature spaces. In: Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II, pp. 536–542. MIT, Cambridge, MA (1999)
37. Osowski, S., Hoai, L., Markiewicz, T.: Support vector machine-based expert system for reliable heartbeat recognition. IEEE T. Bio-Med. Eng. **51**(4), 582–589 (2004)
38. Peng, J., Heisterkamp, D., Dai, H.: Adaptive quasiconformal kernel nearest neighbor classification. IEEE T. Pattern Anal. **26**(5), 656–661 (2004)
39. Platt, J.C.: Fast training of support vector machines using sequential minimal optimization. In: Advances in Kernel Methods: Support Vector Learning, pp. 185–208. MIT, Cambridge, MA (1999)
40. Platt, J.C.: Probabilities for SV Machines. In: Smola, A.J., Bartlett, P.J., Scholkopf, B., Schuurmans, D. (eds.) Advances in Large Margin Classifiers, pp. 61–74. MIT, Cambridge, MA (2000)
41. Raskutti, B., Kowalczyk, A.: Extreme re-balancing for SVMs: a case study. SIGKDD Explorations **6**(1), 60–69 (2004)
42. Rothganger, F., Lazebnik, S., Schmid, C., Ponce, J.: Segmenting, modeling, and matching video clips containing multiple moving objects. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 914–921 (2004)
43. Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7), 1443–1471 (2001)
44. Schölkopf, B., Smola, A.J.: Learning with Kernels Support Vector Machines, Regularization, Optimization, and Beyond. MIT, Cambridge, MA (2002)
45. Shin, H., Cho, S.: How to deal with large dataset, class imbalance and binary output in SVM based response model. In: Proceedings of the Korean Data Mining Conference, pp. 93–107 (2003)
46. Shipp, C.A., Kuncheva, L.: Relationships between combination methods and measures of diversity in combining classifiers. Inform. Fusion **3**(2), 135–148 (2002)
47. Tax, D.: One-class classification:concept-learning in the absence of counter-examples. Asci dissertation series, Delft University of Technology (2001)
48. Tax, D., Duin, R.: Support vector data description. Pattern Recogn. Lett. **20**(11–13), 1191–1199 (1999)

49. Tax, D., Duin, R.: Image database retrieval with support vector data description. In: Proceedings of the Sixth Annual Conference of the Advanced School for Computing and Imaging, ASCI Delft (2000)
50. Tax, D., Duin, R.: Uniform object generation for optimizing one-class classifiers. J. Mach. Learn. Res. **2**, 155–173 (2002)
51. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
52. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
53. Veropoulos, K., Cristianini, N., Campbell, C.: Controlling the sensitivity of support vector machines. In: Proceedings of the International Joint Conference on Artificial Intelligence, (IJCAI99), Stockholm, Sweden (1999)
54. Weiss, G., Provost, F.: The effect of class distribution on classifier learning: an empirical study. Tech. Report ML-TR-44, Department of Computer Science, Rutgers University, August 2001
55. Wolpert, D.H.: Stacked generalization. Neural Networks **5**, 241–259 (1992)
56. Wu, G., Chang, E.Y.: Class-boundary alignment for imbalanced dataset learning. In: The Twentieth ICML Workshop on Learning from Imbalanced Datasets, pp. 49–56. Washington, DC (2003)
57. Yanowitz, F.G.: The Alan E. Lindsay ECG learning center in cyberspace, http://medlib.med.utah.edu/kw/ecg/ (2003)