On Software Defect Prediction Using Machine Learning

Jinsheng Ren¹, Ke Qin¹, Ying Ma²†, Guanghun Luo¹
¹University of Electronic Science and Technology of China, China, 611731
²Xiamen University of Technology, China, 361024

Abstract

The goal of this paper is to catalog the software defect prediction using machine learning. Over the last few years, the field of software defect prediction has been extensively studied because of its crucial position in the area of software reliability maintenance, software cost estimation and software quality assurance. An insurmountable problem associated with software defect prediction is the class imbalance problem. Recently, the Asymmetric Partial Least Squares Classifier (APLSC) reported by Qu et al [1] and the Kernel Principle Component Analysis (KPCA) reported in [2] have been shown to be able to tackle the class imbalance problem. Although both the APLSC and KPCA are quite novel, they have their own disadvantages: First, the APLSC is a bilinear classifier, in which the dimension is mapped to a bilinear subspace. Second, the APLSC suffers from high overlapping, especially when the data sets are nonlinear separate. The KPCA regression model does not consider the correlation between principal components and the class attribution. In this paper, we propose two kernel based classifiers, called the Asymmetric Kernel Partial Least Squares Classifier (AKPLSC) and Asymmetric Kernel Principal Component Analysis Classifier (AKPCAC) to solve the class imbalance problem. The kernel function we use is the Gaussian kernel function. Experiments are conducted on the NASA and SOFTLAB data sets. Theoretical analysis and experimental results (using F-measure, Friedman test and Tukey test) confirm the correctness and effectiveness of our methods. As far as we know, the results presented in this paper are novel.

Keywords: Software Defect Prediction, Asymmetric Learning, Machine Learning, Kernel Method, Friedman test, Tukey test.

1 Introduction

Software defect prediction is an essential part of software quality analysis and has been extensively studied in the domain of software-reliability engineering [3, 4, 5, 6, 7]. However, as pointed out by Menzies et al. [4] and Khoshgoftaar et al. [6], the performance of defect predictors can be greatly degraded by class imbalance problem of the real-world data sets. Here the “class imbalanced” means the majority of defects in a software system are located in a small percentage of the program modules. Current approaches to solve the class imbalance problem can be roughly categorized into two ways: in a data-level way or algorithm-level way, as reported in [6]. The literature [6] shows that the algorithm-level method AdaBoost almost always outperforms even the best data-level methods in software defect prediction. AdaBoost is a typical adaptive algorithm which has received great attention since Freund and Schapire’s proposal [8]. Adaboost attempts to reduce the bias generated by majority class data, by updating the weights of instances dynamically according

*Corresponding author: qinke@uestc.edu.cn
†Corresponding author: maying@xmut.edu.cn
to the errors in previous learning. Some other studies improved dimension reduction methods for the class imbalanced problem by means of Partial Least Squares (PLS) [9], Linear Discriminant Analysis (LDA) [10] and Principle Component Analysis (PCA) [11, 12]. Although PLS was not inherently designed for problems of classification and discrimination, it is widely used in many areas need class proclamation. The authors of [9] reported that rarely will PLS be followed by an actual discriminant analysis on the scores and rarely is the classification rule given a formal interpretation. Still this method often produces nice separation. Based on the previous work, recently, Qu et al investigated the effect of PLS in unbalanced pattern classification. It is reported that beyond dimension reduction, PLS is proved to be superior to generate favorable features for classification. Thereafter, they proposed an asymmetric partial least squares (APLS) classifier to deal with the class imbalance problem. They illustrated that APLS outperforms other algorithms because it can extract favorable features for unbalanced classification. As for the PCA, it is an effective linear transformation, which maps high-dimensional data to a lower dimensional space. Based on the PCA, the authors of [2] proposed Kernel Principal Component Analysis (KPCA) which can perform nonlinear mapping \( \Phi(x) \) to transform an input vector to a higher dimensional feature space, where kernel function \( \Phi(x) \) is introduced to reduce computation for mapping the data nonlinearly into a feature space. Then linear PCA is used in this feature space.

While both the APLS and KPCA are of great value, they have their own disadvantages. For example, the APLS classifier is a bilinear classifier, in which the dimension is mapped to a bilinear subspace, which is, to some degree, obscure and not easy to implement. The KPCA regression model does not consider the correlation between principal components and the class attribution. PCA dimension reduction is affected inevitably by asymmetric distribution. In this paper, we propose two kernel based learning methods to solve the class imbalance problem, called Asymmetric Kernel Partial Least Squares Classification (AKPLSC) and Asymmetric Kernel Principal Component Analysis Classification (AKPCAC) respectively. The former is able to nonlinearly extract the favorable features and retrieve the loss caused by class imbalance problem, while the latter is more adaptive to imbalance data sets.

It is not out of place to explain the relationship between this paper and our previous papers [13] and [14]: The AKPLSC and AKPCAC were firstly proposed in [13] and [14] respectively. However, recently we found some errors when we proceeded to our work. And it's very because of such errors, the AKPCAC and AKPLSC proposed in [13] and [14] show superiority only on part of the data sets. We carefully rectified the source code and then tested the AKPCAC and AKPLSC again on the whole data sets by means of statistical tools, such as Friedman test and Tukey test. The outcomes show that our classifiers indeed outperform the others, namely, APLSC, KPCA, AdaBoost and SMOTE. We carefully examine the theory and experimental results and then form this paper in more detail.

2 State of the art

In software defect prediction, \( L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell)\} \subset X \times Y \) denotes the labeled example set with size \( \ell \) and \( U = \{x_{\ell+1}, x_{\ell+2}, \ldots, x_{\ell+u}\} \subset X \) denotes the unlabeled example set with size \( u \). For labeled examples, \( Y = \{+1, -1\} \), the defective modules are labeled ‘+1’, the non-defective modules are labeled ‘-1’. Software defect data sets are highly imbalanced, i.e. the examples of the minority class (defective modules) are heavily under-represented in comparison to the examples of majority class (non-defective modules). Thereby, lots of algorithms are proposed to cope with this problem, as will be seen below.
2.1 Software Defect Predictor Related with Partial Least Squares

Linear Partial Least Squares (PLS) [9] is an effective linear transformation, which performs the regression on the subset of extracted latent variables. Kernel PLS [15] first performs nonlinear mapping \( \Phi : \{x_i\}_{i=1}^n \in \mathbb{R}^N \to \Phi(x) \in \mathbb{F} \) to project an input vector to a higher dimensional feature space, in which the linear PLS is used.

Given the center \( M_i \), the radius of the class region \( r_i \), and the parameter of overlapping \( \eta_i \), the relationship of the two classes can be expressed as \( M_i = M_{+1} - \eta(r_{i+1} - r_{i-1}) \). The parameter \( \eta_i \) indicates the level of overlapping between the region of the two classes (The smaller value of \( \eta \), the higher overlapping).

APLSC can be expressed as \( \tilde{Y} = \text{sign} \left( \sum_{i=1}^k m_i t_i - b \right) \), which is derived from the regression model of the linear PLS, \( \tilde{y} = \sum_{i=1}^k m_i t_i \), where \( k \) is the number of the latent variables, \( t_i \) is the \( i \)th score vector of testing data, \( m_i \) indicates the direction of \( i \)th score, and the bias \( b \) is equal to \( m_1(M_{+1} - r_{i+1}) \).

APLSC suffers from the high overlapping, especially when the data sets are nonlinear separable [1]. A suggestion of solving such overlapping problem is by using a kernel method. Kernel PLS [15] corresponds to solving the eigenvalue equation as follows:

\[
\Phi \Phi^T \Psi \Psi^T \tau = \lambda \tau
\]

where \( \Phi \) and \( \Psi \) denote the matrix of mapped X-space data \( \Phi(x) \) and the matrix of mapped Y-space data \( \Psi(y) \) in the feature space \( \mathbb{F} \), respectively. The nonlinear feature selection methods can reduce the overlapping level of the two classes, but the class imbalance problem makes them fail to distinguish the minority class [1]. In order to retrieve the loss caused by class imbalance problem, we want to get the bias \( \tilde{b} \) of the kernel PLS Classification, KPLSC [15].

Different from the APLSC, the kernel PLS regression is \( \tilde{y} = \sum_{i=1}^\ell \alpha_i \kappa(x_i, x) \), where \( \ell \) is the size of labeled example set, \( \kappa(x_i, x) \) is a kernel function, and \( \alpha_i \) is dual regression coefficient. Consequently, we may combine the APLSC and kernel PLS so that we get the Asymmetric kernel PLS, as will be seen in section 3.1.

2.2 Kernel Principal Component Analysis Classifier for Software Defect Prediction

Principal Component Analysis (PCA) [12] is an effective linear transformation, which maps high-dimensional data to a lower dimensional space. Kernel Principal Component Analysis (KPCA) [2] first performs nonlinear mapping \( \Phi(x) \) to transform an input vector to a higher dimensional feature space. And then linear PCA is used in this feature space.

For both the algorithms demonstrated in [2, 12], the input data are centralized in the original space and the transformed high-dimensional space, i.e. \( \sum_{i=1}^\ell x_i = 0 \) and \( \sum_{i=1}^\ell \Phi(x_i) = 0 \), where \( \ell \) is the number of the labeled data, and \( x_i \) is the \( i \)th instance of the data set. In the proceeding of PCA, the correlation matrix \( C = \frac{1}{\ell} \sum_{i=1}^\ell x_i x_i' \) should be diagonalized, while in KPCA, the correlation matrix \( C^\Phi = \frac{1}{\ell} \sum_{i=1}^\ell \Phi(x_i) \Phi(x_i)' \) should be diagonalized. It is equal to solving the eigenvalue problem \( \lambda V = C^\Phi V \), where \( \lambda \) is an eigenvalue, \( V \) is a matrix of eigenvectors in KPCA. It can also be written as \( n\lambda \alpha = K \alpha \), where \( K = nC^\Phi \) is the kernel matrix.

The kernel principal component regression algorithm has been proposed by R. Rosipal et al. [2]. The standard regression model in the transformed feature space can be written as

\[
f(x) = \sum_{k=1}^p w_k \beta(x)_k + b
\]
Table 1: Data sets

<table>
<thead>
<tr>
<th>project</th>
<th>modules</th>
<th>attributes</th>
<th>size[loc]</th>
<th>%defective</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ar3</td>
<td>63</td>
<td>30</td>
<td>5,624</td>
<td>11.11</td>
<td>Embedded Controller</td>
</tr>
<tr>
<td>ar4</td>
<td>107</td>
<td>30</td>
<td>9,196</td>
<td>18.69</td>
<td>Embedded Controller</td>
</tr>
<tr>
<td>ar5</td>
<td>36</td>
<td>30</td>
<td>2,732</td>
<td>22.22</td>
<td>Embedded Controller</td>
</tr>
<tr>
<td>cm1</td>
<td>327</td>
<td>38</td>
<td>14,763</td>
<td>12.84</td>
<td>Spacecraft instrument</td>
</tr>
<tr>
<td>kc1</td>
<td>2109</td>
<td>21</td>
<td>42,965</td>
<td>15.46</td>
<td>Storage management</td>
</tr>
<tr>
<td>kc2</td>
<td>521</td>
<td>22</td>
<td>19,259</td>
<td>20.54</td>
<td>Storage management</td>
</tr>
<tr>
<td>kc3</td>
<td>194</td>
<td>40</td>
<td>7,749</td>
<td>18.56</td>
<td>Storage management</td>
</tr>
<tr>
<td>sw1</td>
<td>253</td>
<td>38</td>
<td>8341</td>
<td>10.67</td>
<td>A zero gravity experiment</td>
</tr>
<tr>
<td>pc1</td>
<td>705</td>
<td>38</td>
<td>25,924</td>
<td>8.65</td>
<td>Flight software</td>
</tr>
</tbody>
</table>

Table 2: Metrics used in our experiment

<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loc</td>
<td>5</td>
<td>Halstead’s count of blank lines; McCabe’s line count of code;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Halstead’s line count;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Halstead’s count of lines of comments; line count of code and comment</td>
</tr>
<tr>
<td>McCabe</td>
<td>3</td>
<td>cyclomatic complexity; essential complexity; design complexity</td>
</tr>
<tr>
<td>Halstead</td>
<td>12</td>
<td>unique operators; unique operands; total operators; total operands;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>total operators and operands; volume; program length; difficulty;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>intelligence; effort; volume on minimal implementation; time estimator</td>
</tr>
<tr>
<td>Branch Count</td>
<td>1</td>
<td>branch count</td>
</tr>
<tr>
<td>Others</td>
<td>18</td>
<td>global data complexity; cyclomatic density; decision count; decision</td>
</tr>
<tr>
<td></td>
<td></td>
<td>density; global data density; essential density; design density; loc</td>
</tr>
<tr>
<td></td>
<td></td>
<td>executable; parameter count; percent comments; normalized cyclomatic</td>
</tr>
<tr>
<td></td>
<td></td>
<td>complexity; modified condition count; multiple condition count; node</td>
</tr>
<tr>
<td></td>
<td></td>
<td>count; maintenance severity; condition count; global data complexity;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>call pairs; edge count</td>
</tr>
</tbody>
</table>

where $p$ is the number of components, $w_k$ is the $k$th primal regression coefficient, and $b$ is the regression bias. $\beta(x)_k = V_k \Phi(x)$, $V_k$ is the $k$th eigenvector of $V$. $V$ and $\Lambda$ are the eigenvectors and eigenvalues of the correlation matrix respectively.

2.3 Data set

There are many data sets for machine learning tests, such as the UCI [17] and the Promisedata repository [18], which is a data collecting from real world software engineering projects. The choice that which data set should be used depends on the area of the machine learning where it will be applied. In this paper, the experimental data sets come from NASA and SOFTLAB, which can be obtained from PROMISE [18], as shown in Table 1 and 2. These software modules are developed in different languages, at different sites by different teams, as shown in Table 1. The SOFTLAB data sets (ar3, ar4, ar5), are drawn from three controller systems for a washing machine, a dishwasher, and a refrigerator respectively. They are all written in C. The rests are from NASA projects. They are all written in C++, except for kc3, which is written in JAVA. All the metrics are computed according to [18].

1Since the contributors maintain these data set continuously, the metrics listed in Table 1 may vary at different time. What we are using in this paper are the latest ones updated in June 2012. They are different from the data set we used in our previous papers[13, 14].
3 Design the Asymmetric Classifiers Based on Kernel Method

3.1 The Asymmetric Kernel Partial Least Squares Classifier (AKPLSC)

As we illustrated in Section 2.1, APLSC can be expressed as \( \hat{Y} = \text{sign}\left( \sum_{i=1}^{k} m_i t_i - b \right) \) and the kernel PLS regression is \( \hat{y} = \sum_{i=1}^{\ell} \alpha_i \kappa(x_i, x) \), thus the AKPLSC can be well characterized as:

\[
\hat{Y} = \text{sign}\left( \sum_{i=1}^{\ell} \alpha_i \kappa(x_i, x) - \hat{b} \right)
\]  

(3)

where \( \alpha_i \) is dual regression coefficient, which can be obtained from kernel PLS, as shown in Algorithm 1. \( \hat{b} \) is the bias of the classifier.

Since kernel PLS put most of the information on the first dimension, the bias in the AKPLSC can be computed similarly as [1]:

\[
\hat{b} = c_1 \times (M_{+1} - r_{+1}\eta) = c_1 \times \frac{M_{+1}r_{-1} + M_{-1}r_{+1}}{r_{-1} + r_{+1}}
\]  

(4)

where \( c_1 \) indicates the direction of the first score \( \tau_1 \), the centers \( (M_{+1}, M_{-1}) \) and radiiuses \( (r_{+1}, r_{-1}) \) are computed based on \( \tau_1 \), which can be obtained from Eq. (1). Then we move the origin to the center of mass by employing data centering, as reported in [15]:

\[
K = K - \frac{1}{\ell} J J' K - \frac{1}{\ell} K J J' + \frac{1}{\ell^2} (J' K J) J J',
\]  

(5)

where \( J \) is the a vector with all elements equal to 1. After data centering, the AKPLSC can be described as Algorithm 1.

3.2 The Asymmetric Kernel Principal Component Analysis Classifier (AKPCAC)

The KPCA regression model does not consider the correlation between principal components and the class attribution. PCA dimension reduction is inevitably affected by asymmetric distribution [1]. We analyze the effect of class imbalance on KPCA. Considering the class imbalance problem, we propose an Asymmetric Kernel Principal Component Analysis Classification (AKPCAC), which retrieves the loss caused by this effect.

Suppose \( S^b_w = \sum_{i=1}^{2} n_i (u_i - \bar{u})(u_i - \bar{u})' \) denotes the between-class scatter matrix, \( S^\phi_w = \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\Phi(x_j^i) - \bar{u}_i)(\Phi(x_j^i) - \bar{u}_i)' \) the within-class scatter matrix, where \( \bar{u}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \Phi(x_j^i) \) is class-conditioned mean vector, \( \bar{u} \) is mean vector of total instances, \( \Phi(x_j^i) \) is the \( j \)th instances in the \( i \)th class, and \( n_i \) is the number of instances of the \( i \)th class. The total non-centralized scatter matrix in the form of kernel matrix is

\[
K = \sum_{i=1}^{2} \sum_{j=1}^{\ell} \left( \Phi(x_j^i) - \bar{u}_i \right) \left( \Phi(x_j^i) - \bar{u}_i \right)'
\]

\[
= \sum_{i=1}^{2} \sum_{j=1}^{\ell} \left( \Phi(x_j^i) - \bar{u}_i + \bar{u}_i - \bar{u} \right) \left( \Phi(x_j^i) - \bar{u}_i + \bar{u}_i - \bar{u} \right)'
\]

\[
= S^\phi_w + S^b_w + \sum_{i=1}^{2} \sum_{j=1}^{\ell} \left( \Phi(x_j^i) - \bar{u}_i \right) \left( \bar{u}_i - \bar{u} \right) + \sum_{i=1}^{2} \sum_{j=1}^{\ell} \left( \bar{u}_i - \bar{u} \right) \left( \Phi(x_j^i) - \bar{u}_i \right)'
\]  

(6)
Algorithm 1 AKPLSC

Input: Labeled and unlabeled data sets, \( L \) and \( U \); number of components, \( k \).
Output: Asymmetric Kernel Partial Least Squares Classifier, \( H \).
Method:
1. \( K_{ij} = \kappa(x_i, x_j), i, j = 1, \ldots, \ell, x_i, x_j \in L \);
2. \( K = K_y Y = Y \), where \( K \) is the kernel matrix, \( Y \) is the label vector.
3. for \( j = 1, \ldots, k \) do
4. \( \beta_j = \beta_j / ||\beta_j|| \), where \( \beta_j \) is a projection direction.
5. repeat
6. \( \beta_j = \hat{Y} \hat{Y}' K_{j\beta_j} \)
7. \( \beta_j = \beta_j / ||\beta_j|| \)
8. until convergence
9. \( \tau_j = K_{j\beta_j} \), where \( \tau_j \) is the score
10. \( c_j = \hat{Y} \tau_j / ||\tau_j||^2 \), where \( c_j \) is the direction of the score
11. \( \hat{Y} = \hat{Y} - \tau_j c_j \), where \( \hat{Y} \) is the deflation of \( Y \)
12. \( K_{j+1} = (I - \tau_j \tau_j' / ||\tau_j||^2) K_j (I - \tau_j \tau_j' / ||\tau_j||^2) \)
13. end for
14. \( B = [\beta_1, \ldots, \beta_k], T = [\tau_1, \ldots, \tau_k] \)
15. \( \alpha = B(T' KB)^{-1}T' Y \), where \( \alpha \) is the vector of dual regression coefficients
16. Calculate \( b \) according to Eq. (4);
17. \( H(x) = \text{sign} \left( \sum_{i=1}^{\ell} \alpha_i \kappa(x_i, x) - \hat{b} \right), x \in U \);
18. return \( H \);
End Algorithm AKPLSC.

The third term of Eq. (6) can be rewritten as
\[
\sum_{i=1}^{2} \sum_{j=1}^{\ell} \left( \Phi(x_j^l) - \tilde{u}_i \right) \left( \tilde{u}_i - \bar{u} \right)' = \sum_{i=1}^{2} \left( \sum_{j=1}^{\ell} \left( \Phi(x_j^l) - \tilde{u}_i \right) \right) \left( \tilde{u}_i - \bar{u} \right)' = \sum_{i=1}^{2} \left( \sum_{j=1}^{n_i} \left( \Phi(x_j^l) - n_i \tilde{u}_i \right) \right) \left( \tilde{u}_i - \bar{u} \right)' \tag{7}
\]

Note that \( n_i \tilde{u}_i = \sum_{j=1}^{n_i} \Phi(x_j^l) \). Then the third term and fourth term of Eq. (6) are equal to zero. Thus, we have the relation \( K = S_\Phi + S_{w_\Phi} = S_\Phi + P \Sigma_P + N \Sigma_N \), where \( P \) is the number of positive instances, \( N \) is the number of negative instances, \( \Sigma_P \) is the positive covariance matrices, and \( \Sigma_N \) is the negative covariance matrices. Since class distribution has a great impact on \( S_{w_\Phi} \), the class imbalance also impacts the diagonalization problem of KPCA.

In order to combat the class imbalance problem, we propose the AKPCAC, based on kernel method. It considers the correlation between principal components and the class distribution. The imbalance ratio can be denoted as \( \frac{\sum_{i=1}^{\ell} I(y_i, +1)}{\sum_{i=1}^{\ell} I(y_i, -1)} = \frac{p}{N} \), which is the probabilities of the positive instances to the negative instances of training data, where \( I(\cdot) \) is an indicator function: \( I(x, y) = 1 \) if \( x = y \), zero otherwise. We assume that future test examples are drawn from the same distribution,
so the imbalance ratio of the training data is the same as that of the test data. Then, we have

\[
\frac{\sum_{i=1}^{\ell} (\hat{y}_i - \hat{b}) I(y_i, +1)}{\sum_{i=1}^{\ell} (\hat{y}_i - \hat{b}) I(y_i, -1)} = \frac{\sum_{i=1}^{\ell} I(y_i, +1)}{\sum_{i=1}^{\ell} I(y_i, -1)} = \frac{P}{N}
\]

where \( \hat{b} \) is the bias of the classifier, \( \hat{y}_i \) is the regression result of \( x_i \). \( \hat{y}_i \) can be computed by regression model Eq. (2). Note that the regression is conducted on the \( p \) principal components. Solving this one variable equation, we get

\[
\hat{b} = \frac{N \left( \sum_{i=1}^{\ell} (\hat{y}_i I(y_i, +1)) \right) - P \left( \sum_{i=1}^{\ell} (\hat{y}_i I(y_i, -1)) \right)}{N^2 - P^2}
\]

(9)

Based on principal components, Eq. (9) describes the detail deviation of the classifier. This deviation may be caused by class imbalance, noise et al. In order to retrieve the harmful effect, we compensate this deviation. By transforming the regression model Eq. (2), the classifier model can be written as

\[
H(x) = \text{sign} \left( \sum_{k=1}^{p} w_k \beta(x)_k + \hat{b} \right)
\]

\[
= \text{sign} \left( \sum_{k=1}^{p} w_k \sum_{i=1}^{\ell} \alpha_i^k \kappa(x_i, x) + \hat{b} \right)
\]

\[
= \text{sign} \left( \sum_{i=1}^{\ell} c_i \kappa(x_i, x) + \hat{b} \right)
\]

(10)

where \( \{ c_i = \sum_{k=1}^{p} w_k \alpha_i^k \}, i = 1, 2, \ldots, \ell \).

AKPCAC is summarized in Algorithm 2. Since the AKPCAC was firstly studied for reducing the effect of class imbalance for classification, it inherently has the advantage of kernel method, which can conduct quite general dimensional feature space mappings. In this paper, again, we have illustrated how the unreliable dimensions based on KPCA can be removed, thereafter, the imbalance problem based on the PCA has also been solved.

4 Experimental Result

The experiments are conducted under the data set from NASA and SOFTLAB. The Gaussian kernel function \( K(x, y) = \exp(-||x - y||^2) \) is adopted for the performance investigation for both AKPLSC and AKPCAC. The efficiency are evaluated by F-measure and Friedman test, as will be explained presently.

4.1 Validation Method Using F-measure

F-measure method is widely used for assessing a test’s accuracy. It considers both the precision \( P \) and the recall \( R \) to compute the score. \( P \) is defined as the number of correct results divided by the number of all returned results. \( R \) is the number of correct results divided by the number of results
Algorithm 2 AKPCAC

Input: The set of Labeled samples, \( L = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell) \} \);
The set of unlabeled samples, \( U = \{ x_{\ell+1}, x_{\ell+2}, \ldots, x_{\ell+u} \} \);
Output: Kernel Principal Component Classifier, \( H \);
Method:
1: \( K_{ij} = \kappa(x_i, x_j), i, j = 1, \ldots, \ell \);
2: \( K = K - \frac{1}{\ell}JJ^T K - \frac{1}{\ell}KJ^T J + \frac{1}{\ell^2} (J^T K) J J^T \), where \( J \) is a vector with all elements equal to 1.
3: \([V, \Lambda] = \text{eig}(K)\);
4: \( \alpha = \sum_{j=1}^{p} \frac{1}{\lambda_j} (V_j^T Y^s) V_j \); % \( Y^s = \{ y_1, y_2, \ldots, y_\ell \} \) is the label vector
5: Calculate \( \hat{b}, H(x) \) according to Eq.(9), Eq.(10);
6: return \( H \);
End Algorithm APPCC.

that should have been returned. For the clarity of this paper, we give a short explanation of the F-measure as below. Obviously, there are four possible outcomes of a predictor:

1. TP: True positives are modules classified correctly as defective modules.
2. FP: False positives refer to non-defective modules incorrectly labeled as defective.
3. TN: True negatives correspond to correctly classified non-defective modules.
4. FN: False negatives are defective modules incorrectly classified as non-defective.

Thereby, the precision is defined as: \( P = TP/(TP + FP) \) and the recall is \( R = TP/(TP + FN) \). The general formula of the F-measure is:

\[
F_\beta = \frac{(1 + \beta^2)PR}{\beta^2 P + R}
\]  

(11)

where \( \beta \) is a positive real number. According to the definition of \( P \) and \( R \), Eq.(11) can be rewritten as:

\[
F_\beta = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2 FN + FP}
\]  

(12)

Generally, there are 3 commonly used F-measures: \( F_1 \) (which is a balance of \( P \) and \( R \)), \( F_{0.5} \) (which puts more emphasis on \( P \) than \( R \)), \( F_2 \) (which weighs \( R \) higher than \( P \)). In this paper, \( F_1 = 2PR/(P + R) \) is used to evaluate the efficiency of different classifiers. The F-measure can be interpreted as a weighted average of the precision and recall. It reaches its best value at 1 and worst score at 0.

We compare the F-measure values of different predictors including AKPLSC, AKPCAC, APLSC [1], KPCAC [2], AdaBoost [6], and SMOTE [19]. The results are listed in Table 4. For each data set, we perform a 10 \times 5-fold cross validation.

From the table we may see clearly that the AKPLSC and the AKPCAC are superior than the other 4 classifiers, which validates our contributions of this paper.

4.2 Validation Method Using Friedman Test and Tukey Test

The Friedman test is a non-parametric statistical test developed by the Milton Friedman [24, 25]. It is used to detect differences in algorithms/classifiers across multiple test attempts. The procedure involves ranking each block (or row) together, then considering the values of ranks by columns. In this section, we present a multiple AUC value comparison among the six classifiers using Friedman test.
Table 3: Statistical $F$-measure (mean $\pm$ std) values of six classifiers on all data sets. The value of the winner method of each row is emphasized in bold.

<table>
<thead>
<tr>
<th></th>
<th>APLSC</th>
<th>KPCAC</th>
<th>AdaBoost</th>
<th>SMOTE</th>
<th>AKPCAC</th>
<th>AKPLSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ar3</td>
<td>0.698 (0.062)</td>
<td>0.594 (0.115)</td>
<td>0.357 (0.026)</td>
<td>0.415 (0.063)</td>
<td>0.569 (0.052)</td>
<td>0.500 (0.009)</td>
</tr>
<tr>
<td>ar4</td>
<td>0.421 (0.023)</td>
<td>0.422 (0.058)</td>
<td>0.412 (0.013)</td>
<td>0.457 (0.000)</td>
<td>0.474 (0.024)</td>
<td>0.421 (0.001)</td>
</tr>
<tr>
<td>ar5</td>
<td>0.326 (0.058)</td>
<td>0.562 (0.065)</td>
<td>0.435 (0.012)</td>
<td>0.555 (0.001)</td>
<td>0.625 (0.025)</td>
<td>0.553 (0.022)</td>
</tr>
<tr>
<td>cm1</td>
<td>0.254 (0.033)</td>
<td>0.069 (0.031)</td>
<td>0.234 (0.007)</td>
<td>0.237 (0.000)</td>
<td>0.069 (0.031)</td>
<td>0.235 (0.002)</td>
</tr>
<tr>
<td>kc1</td>
<td>0.808 (0.005)</td>
<td>0.309 (0.006)</td>
<td>0.325 (0.005)</td>
<td>0.428 (0.000)</td>
<td>0.462 (0.009)</td>
<td>0.433 (0.023)</td>
</tr>
<tr>
<td>kc2</td>
<td>0.431 (0.021)</td>
<td>0.433 (0.014)</td>
<td>0.534 (0.002)</td>
<td>0.520 (0.000)</td>
<td>0.723 (0.011)</td>
<td>0.480 (0.036)</td>
</tr>
<tr>
<td>kc3</td>
<td>0.382 (0.036)</td>
<td>0.231 (0.025)</td>
<td>0.386 (0.002)</td>
<td>0.340 (0.000)</td>
<td>0.234 (0.015)</td>
<td>0.412 (0.000)</td>
</tr>
<tr>
<td>mv1</td>
<td>0.285 (0.011)</td>
<td>0.276 (0.060)</td>
<td>0.224 (0.002)</td>
<td>0.290 (0.000)</td>
<td>0.371 (0.037)</td>
<td>0.268 (0.025)</td>
</tr>
<tr>
<td>pc1</td>
<td>0.215 (0.019)</td>
<td>0.212 (0.025)</td>
<td>0.354 (0.001)</td>
<td>0.387 (0.000)</td>
<td>0.254 (0.015)</td>
<td>0.398 (0.001)</td>
</tr>
</tbody>
</table>

Table 4: Comparison of AUC between six classifiers. The ranks in the parentheses are used in computation of the Friedman test.

<table>
<thead>
<tr>
<th></th>
<th>APLSC</th>
<th>KPCAC</th>
<th>AdaBoost</th>
<th>SMOTE</th>
<th>AKPCAC</th>
<th>AKPLSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ar3</td>
<td>0.626 (4)</td>
<td>0.588 (5)</td>
<td>0.581 (6)</td>
<td>0.590 (4)</td>
<td>0.650 (1)</td>
<td>0.682 (2)</td>
</tr>
<tr>
<td>ar4</td>
<td>0.563 (5)</td>
<td>0.600 (3)</td>
<td>0.555 (6)</td>
<td>0.610 (2)</td>
<td>0.671 (1)</td>
<td>0.579 (4)</td>
</tr>
<tr>
<td>ar5</td>
<td>0.626 (5)</td>
<td>0.710 (2)</td>
<td>0.614 (6)</td>
<td>0.651 (3)</td>
<td>0.722 (1)</td>
<td>0.641 (4)</td>
</tr>
<tr>
<td>cm1</td>
<td>0.611 (4)</td>
<td>0.724 (1)</td>
<td>0.589 (5)</td>
<td>0.550 (6)</td>
<td>0.681 (2)</td>
<td>0.650 (3)</td>
</tr>
<tr>
<td>kc1</td>
<td>0.682 (4)</td>
<td>0.592 (6)</td>
<td>0.627 (5)</td>
<td>0.700 (3)</td>
<td>0.800 (1)</td>
<td>0.768 (2)</td>
</tr>
<tr>
<td>kc2</td>
<td>0.591 (6)</td>
<td>0.601 (5)</td>
<td>0.796 (1)</td>
<td>0.635 (3)</td>
<td>0.732 (2)</td>
<td>0.610 (4)</td>
</tr>
<tr>
<td>kc3</td>
<td>0.598 (5)</td>
<td>0.569 (6)</td>
<td>0.698 (2)</td>
<td>0.612 (4)</td>
<td>0.658 (3)</td>
<td>0.713 (1)</td>
</tr>
<tr>
<td>mv1</td>
<td>0.387 (5)</td>
<td>0.602 (4)</td>
<td>0.534 (6)</td>
<td>0.654 (2)</td>
<td>0.725 (1)</td>
<td>0.639 (3)</td>
</tr>
<tr>
<td>pc1</td>
<td>0.602 (6)</td>
<td>0.718 (5)</td>
<td>0.769 (3)</td>
<td>0.733 (4)</td>
<td>0.841 (2)</td>
<td>0.882 (1)</td>
</tr>
</tbody>
</table>

Sum of ranks $T_1 = 43$ $T_2 = 37$ $T_3 = 40$ $T_4 = 31$ $T_5 = 14$ $T_6 = 24$

At first, we make two hypotheses:

$H_0$: the six classifiers have equal classification probability;

$H_1$: at least two of them have different probability distribution.

In order to determine which hypothesis should be rejected, we compute the statistic:

$$F_r = \frac{12}{bk(k+1)} \sum_{i=1}^{k} R_i^2 - 3b(k+1)$$

where $b$ is the number of blocks (or rows), $k$ is the number of classifiers and $R_i$ is the summation of ranks of each column. The range of rejection for null-hypothesis is $F_r > \chi^2_\alpha$. In our experiment, the degree of freedom is $k-1 = 5$ and we set $\alpha = 0.05$, thus $F_r = 18.9683 > 11.0705$, which implies $H_0$ should be rejected.

Friedman test just tells us that at least two of the classifiers have different performance, but it doesn’t give any implication which one performs best. In this case, a post hoc test should be proceeded. Actually, there are many post hoc test such as LSD (Fisher’s least significant difference), SNK (Student-Newman-Keuls), Bonferroni-Dunn test, Tukey test and Nemenyi test, which is very similar to the Tukey test for ANOVA. In this paper, the Tukey test [26] is applied.

Tukey test is a single-step multiple comparison procedure and statistical test. It is used in conjunction with an ANOVA to find means that are significantly different from each other. It compares all possible pairs of means, and is based on a studentized range distribution.

Tukey test involves two basic assumptions:

1. The observations being tested are independent.

9
(2). There is equal within-group variance across the groups associated with each mean in the test.

Obviously, our case satisfies the two requirements.

The steps of the Tukey multiple comparison with equal sample size can be summarized as follows:

**Algorithm 3 Tukey Multiple Comparison: Equal Sample Size.**

**Input:** $\alpha$, $p$, $v$, $s$, $MSE$, $n_t$ and samples. The meaning of these parameters is: $\alpha$ is an error rate. $p$ is the number of means, $s = \sqrt{MSE}$, $v$ is the degree of freedom related with $MSE$. $n_t$ is the number of observations of each sample.

**Output:** The minimum significant difference $\omega$ and a deduction;

**Method:**

1. Choose a proper error rate $\alpha$;
2. Calculate the statistic $\omega$ according to
   \[
   \omega = q_\alpha(p, v) \frac{s}{\sqrt{n_t}}
   \]
   where $q_\alpha(p, v)$ is the critical value of Studentized range statistic, which can be found from any statistics textbooks.
3. Compute and rank all the $p$ means
4. Draw a deduction based on the ranks in the confidence level $(1 - \alpha)$.

**End Algorithm Tukey Multiple Comparison.**

In this paper, we set $\alpha = 0.05$. Since we compare 6 classifiers over 9 data sets, then $n = 54$, $p = 6$, $v = n - p = 48$ and $n_t = 9$. $q_\alpha(p, v) \approx 4.2$, which can be found from the Studentized range statistic table. Now the only problem to find the value of $\omega$ is to determine $s$ and $MSE$. This can be calculated accordingly:

\[
SY = \sum_{i=1}^{n} y_i, \quad MY = \frac{SY}{n}, \quad YS = \sum_{i=1}^{n} y_i^2
\]

\[
CM = \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2 = \frac{SY^2}{n}, \quad SS = \sum_{i=1}^{n} y_i^2 - CM = YS - CM, \quad SST = \sum_{i=1}^{p} \frac{T_i^2}{n_i} - CM
\]

\[
SSE = SS - SST, \quad MSE = \frac{SSE}{n - p}
\]

where $y_i$ is the corresponding AUC value in table 4 and $T_i$ is the AUC summation of each column. Now we have the results $MSE = 0.0051$, $s = 0.0713$ and $\omega = 0.0998$. The means comparison is listed in table 5. From this table we can see clearly:

1. The difference $T_6 - T_1 = 0.1058$ is greater than the critical value $\omega$, which hints the AKPCAC is significantly better than the APLSC.
2. But compared to the rests except the APLSC, the two newly proposed methods have no significant difference.
3. Nevertheless, the AKPCAC and AKPLSC have the largest and second largest means, which implies the both indeed outperform the rests, although insignificantly.
4. Compared to the AKPLSC, the AKPCAC is slightly more powerful, which supports our claim that the AKPCAC is more adaptive to dimensional feature space mappings over imbalanced data sets.
5. The deduction is made in the confidence level $(1 - 0.05)$. 
Table 5: Tukey test result for the six classifiers.

<table>
<thead>
<tr>
<th>rank:</th>
<th>APLSC</th>
<th>KPCAC</th>
<th>SMOTE</th>
<th>AdaBoost</th>
<th>AKPLSC</th>
<th>AKPCAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean: $T_i$</td>
<td>0.6196</td>
<td>0.6338</td>
<td>0.6394</td>
<td>0.6402</td>
<td>0.6848</td>
<td>0.7254</td>
</tr>
<tr>
<td>difference: $T_{i+1} - T_i$</td>
<td>-</td>
<td>0.0142</td>
<td>0.0057</td>
<td>0.0008</td>
<td>0.0446</td>
<td>0.0407</td>
</tr>
</tbody>
</table>

5 Conclusion

In this paper, we introduce kernel based asymmetric learning for software defect prediction. To eliminate the negative effect of class imbalance problem, we propose two algorithms called the Asymmetric Kernel Partial Least Squares Classifier and the Asymmetric Kernel Principal Component Analysis Classifier. The former one is derived from the regression model of linear PLS while the latter is from kernel PCA method. The AKPLSC can extract feature information in a nonlinear way and retrieve the loss caused by class imbalance. The AKPCAC is more adaptive to dimensional feature space mappings over imbalanced data sets and has a better performance. F-measure, Friedman test and a post hoc test using Tukey method are used to verify the performance of our algorithms. Experimental results on NASA and SOFTLAB data sets validate their effectiveness.

References


Jinsheng Ren was born in Sichuan, China, in 1981. He received the Master degree from the University of Electronic Science & Technology of China (UESTC) in 2008. He is currently a Ph.D. candidate at the UESTC. His research interests include artificial intelligence and machine learning.

Ke Qin was born in Sichuan, China, in 1980. He received the M.E. and Ph.D. degrees from the UESTC in 2006 and 2010 respectively. He was also a visiting scholar at Carleton University, Ottawa, Canada in 2008. He is currently an associate professor at the UESTC. His research interests include chaos, chaotic pattern recognition and neural networks.

Guangchun Luo received the Ph.D. degrees in computer science from the UESTC in 2004. He is currently a full professor of school of Graduate at the UESTC. His research interests include computer networks and software engineering.

Ying Ma was born in Hu’an, China, in 1982. He received the Ph.D. degree from the University of Electronic Science & Technology of China (UESTC) in 2012. He is currently a lecturer at the Xiamen University of Technology. His research interests include software engineering and machine learning.