Kernel Based Asymmetric Learning for Software Defect Prediction

Ying MA†1, Guangchun LOU†, Nonmembers, and Hao CHEN†, Member

SUMMARY A kernel based asymmetric learning method is developed for software defect prediction. This method improves the performance of the predictor on class imbalanced data, since it is based on kernel principal component analysis. An experiment validates its effectiveness.

key words: defect prediction, class imbalance, kernel principal component analysis, machine learning.

1. Introduction

Software defect prediction is to predict the defect-prone modules for the next release of software or cross project software. Real world data mining applications, including software defect prediction domain, must address the issue of learning from imbalanced data sets. As pointed out by Khoshgoftaar et al. [1] and Menzies et al. [2], the majority of defects in a software system are located in a small percentage of the program modules, software defect data sets are highly class imbalanced.

Existing approaches to solving class imbalanced problem mainly include data sampling methods and adaptive algorithm methods. The results reported in [3] show that AdaBoost almost always outperforms even the best data sampling techniques in software defect prediction. AdaBoost is a typical adaptive algorithm, and has received a good deal of attention since being introduced by Freund and Schapire [4]. It attempts to reduce the bias generated by majority class data, by updating the weights of instances dynamically according to the errors in previous learning.

Besides these methods, some studies improved dimension reduction methods for the class imbalanced problem [5], [6], [7]. Most recently, Qu et al. [7] proposed an asymmetric partial least squares classifier (APLSC) to tackle the class imbalance problem. They suggested that APLSC outperform other existing algorithms, because it can extract favorable features for unbalanced classification. However it is a bilinear classifier, in which the dimension is mapped to a bilinear subspace. In this paper, we develop a kernel based asymmetric learning method, called Asymmetric Kernel Principal Component Classification (AKPCC), which is more adaptive to general situations.

2. Kernel Based Asymmetric Learning

2.1 Background

Principal Component Analysis (PCA) [8] is an effective linear transformation, which maps high-dimensional data to a lower dimensional space. Kernel Principal Component Analysis (KPCA) [9] first performs nonlinear mapping \( \Phi(x)^T \) to transform an input vector to a higher dimensional feature space. And then linear PCA is used in this feature space.

In both algorithms, 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. PCA diagonalizes the correlation matrix, \( C = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i x_i^T \), while KPCA diagonalizes the correlation matrix, \( C^0 = \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(x_i) \Phi(x_i)^T \). It is equal to solving the eigenvalue problem \( \lambda V = C^0 V \), where \( \lambda \) is an eigenvalue, \( V \) is a matrix of eigenvectors in KPCA. It can also be written as \( n \alpha = K \alpha \), where \( K = n C^0 \) is the kernel matrix.

The kernel principal component regression algorithm has been proposed by R. Rosipal et al. [9]. The standard regression model in the transformed feature space can be written as

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

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.

2.2 Asymmetric Kernel Principal Component Classification (AKPCC)

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 [7]. We analyze the effect of class imbalance on KPCA. Considering the class imbalance problem, \( \Phi(x)^T \) is introduced to reduce computation, for mapping the data nonlinearly into a feature space.

†The authors are with University of electronic science and technology of China, Chengdu, China.
a) E-mail: may@uestc.edu.cn
DOI: 10.1587/transinf.E93.D.1

Copyright © 2010 The Institute of Electronics, Information and Communication Engineers
we propose an Asymmetric Kernel Principal Component Classification (AKPCC), which retrieves the loss caused by this effect.

In defect predictor, \( L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell)\} \subset X \times Y \) denotes the labeled example set with size \( \ell \) and \( U = \{x_{\ell+1}, x_{\ell+2}, \ldots, x_{n_u}\} \subset X \) 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’.

Suppose \( S^b_\alpha = \sum_{i=1}^{n_u} n_t (\tilde{u} - \bar{u})(\tilde{u} - \bar{u}) \) denotes the between-class scatter matrix, \( S^w_\alpha = \sum_{i=1}^{n_u} n_t (\Phi(x'_i) - \tilde{u})(\Phi(x'_i) - \tilde{u}) \) the within-class scatter matrix, where \( \tilde{u}_t = \frac{1}{n_u} \sum_{i=1}^{n_u} \Phi(x'_i) \) is class-conditional mean vector, \( \bar{u} \) is mean vector of total instances, \( \Phi(x'_i) \) is the \( j \)th instances in the \( i \)th class, and \( n_t \) is the number of instances of the \( i \)th class. The total non-centralized scatter matrix in the form of kernel matrix is

\[
K = \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (\Phi(x'_i) - \bar{u})(\Phi(x'_j) - \bar{u})' = \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (\Phi(x'_i) - \bar{u} + \bar{u} - \bar{u})(\Phi(x'_j) - \bar{u} + \bar{u} - \bar{u})' = S^w_\alpha + S^b_\alpha + \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (\Phi(x'_i) - \bar{u})(\Phi(x'_j) - \bar{u})' + \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (\bar{u} - \bar{u})(\Phi(x'_i) - \bar{u})' \tag{2}
\]

The third term of Eq. (2) can be rewritten as

\[
\sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (\Phi(x'_i) - \bar{u})(\Phi(x'_j) - \bar{u})' = \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} (\Phi(x'_i) - \bar{u}')(\bar{u} - \bar{u})' = \sum_{i=1}^{\ell} \sum_{n=1}^{n_u} (\Phi(x'_i) - \tilde{u})(\tilde{u} - \bar{u})' \tag{3}
\]

Note that \( n\tilde{u} = \sum_{i=1}^{n_u} \Phi(x'_i) \). Then the third term and fourth term of Eq. (2) are equal to zero. Thus, we have the relation \( K = S^w_\alpha + S^b_\alpha = S^w_\alpha + P \Sigma^p_N + N \Sigma^b_\alpha \), where \( P \) is the number of positive instances, \( N \) is the number of negative instances, \( \Sigma^p_N \) is the positive covariance matrices, and \( \Sigma^b_\alpha \) is the negative covariance matrices. Since class distribution has a great impact on \( S^w_\alpha \), the class imbalance also impacts the diagonalization problem of PCA.

In order to combat the class imbalance problem, we propose the AKPCC, based on kernel method. It considers the correlation between principal components and the class distribution. The imbalance ratio can be denoted as \( \frac{\sum_{i=1}^{\ell} K(y_i, +1)}{\sum_{i=1}^{\ell} K(y_i, -1)} = \frac{P}{N} \), which is the probabilities of the positive instances to the negative instances of training data, where \( I() \) is an indicator function\(^{117}\). 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

\[
\frac{\sum_{i=1}^{\ell} (\hat{y}_i - \hat{b})I(y_i, +1)}{\sum_{i=1}^{\ell} I(y_i, +1)} = \frac{\sum_{i=1}^{\ell} I(y_i, +1)}{\sum_{i=1}^{\ell} I(y_i, -1)} = \frac{P}{N} \tag{4}
\]

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. (1). Note that the regression is conducted on the \( p \) principal components. Solving this one variable equation, we get

\[
\hat{b} = \frac{N\left( \sum_{i=1}^{\ell} (\hat{y}_i I(y_i, +1)) - P \left( \sum_{i=1}^{\ell} (\hat{y}_i I(y_i, -1)) \right) \right)}{N^2 - P^2} \tag{5}
\]

Based on principal components, Eq. (5) 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. (1), the classifier model can be written as

\[
H(x) = \text{sign}(\sum_{k=1}^{p} w_k \beta(x_k) + \hat{b}) = \text{sign}(\sum_{k=1}^{p} w_k \sum_{i=1}^{\ell} a_k \Phi(x_i, x) + \hat{b}) = \text{sign}(\sum_{i=1}^{\ell} c_i \Phi(x_i, x) + \hat{b}) \tag{6}
\]

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

AKPCC is summarized as in Algorithm 1. AKPCC originates from the need to combat the class imbalance problem in the classification. It inherits the advantage of kernel method, which can conduct quite general dimensional feature space mappings. This algorithm removes the unreliable dimensions based on KPCA, and solves the imbalance problem based on the principal components.

**Algorithm 1 AKPCC**

**Require:**

The set of Labeled samples, \( L = \{(x_1, y_1), (x_2, y_2), \ldots, (x_\ell, y_\ell)\} \);

The set of unlabeled samples, \( U = \{x_{\ell+1}, x_{\ell+2}, \ldots, x_{n_u}\} \);

**Ensure:**

Kernel Principal Component Classifier, \( H \);

1: \( K_{ij} = \kappa(x_i, x_j), i, j = 1, \ldots, \ell \);
2: \( K = K - \frac{1}{\ell} J K J - \frac{1}{\ell} J K J + J K J J ; \% J \) is the all 1s vector
3: \( [V, A] = \exp(K) \);
4: \( a = \sum_{j=1}^{\ell} \frac{1}{\ell}(V')^j Vj; \% Y' = (y_1, y_2, \ldots, y_\ell) \) is the label vector
5: Calculate \( \hat{b}, H(x) \) according to Eq. (5), Eq. (6);
6: return \( H \);

### 3. Experimental Result

The experimental data sets come from NASA and SOFTESTLAB, which can also be obtained from PROMISE [10], as
shown in Table 2. The SOFTLAB data sets (ar3, ar4, ar5), are drawn form three controller systems for a washing machine, a dishwasher, and a refrigerator respectively. They are all written in C code. Remaining data sets are from NASA projects, which are developed at different sites by different teams. They are all written in C++ code, except for kc3, which is written in JAVA.

In order to investigate the performance of AKPCC (Gaussian kernel $K(x,y) = \exp(-||x-y||^2)$ is used here), we compare it with APLSC [7], KPCC [9], AdaBoost [3], and SMOTE [7, 11]. For each data set, we perform a 10 × 5-fold cross validation. We use $F$-measure = $2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$ and GMeans = (true positive rate $\times$ true negative rate) $\pm$ performance metrics which are commonly used in the field of software defect prediction. And then, the Mann-Whitney U test$^{1111}$ with a 95% confidence interval is conducted.

The results for the five methods are shown in Table 1. We can see that AKPCC outperforms other algorithms in the aspect of $F$-measure. In the aspect of GMeans, AKPCC is no better or even worse than other algorithms. But, AKPCC outperforms KPCC in the aspect of the two metrics. It means that, asymmetric learning method of kernel form is expected to set up a model adapted to more general situations.

4. Conclusion

In this paper, we introduce a kernel based asymmetric learning method for software defect prediction. It can combat the class imbalance problem in the software defect data. Experimental results on real data sets validate its effectiveness.

Acknowledgments

We thank R. Rosipal for providing us with the kernel PCA code, and LeVis Group for APLSC code. We also thank the anonymous reviewers for their great helpful comments.

References


---

$^{1111}$ Mann-Whitney U test is a non-parametric statistical hypothesis test to compare two independent groups of sampled data, which is without an assumption of a normal distribution [12].

| Table 1 | Statistical $F$-measure(mean ± std) and GMeans(mean ± std) values of five classifiers on all data sets. The line w/t means that the algorithm at the corresponding AKPCC wins in w data sets, ties in t data sets, and loses in f data sets, compared with the algorithm at the corresponding column. |
|---|---|---|---|---|---|
| project | APLSC | KPCC | AKPCC | AdaBoost | SMOTE |
| ar3 | 0.505 (0.070) | 0.407 (0.115) | 0.570 (0.087) | 0.377 (0.037) | 0.455 (0.002) |
| ar4 | 0.431 (0.026) | 0.454 (0.058) | 0.477 (0.037) | 0.443 (0.002) | 0.457 (0.000) |
| ar5 | 0.537 (0.061) | 0.570 (0.085) | 0.613 (0.037) | 0.464 (0.014) | 0.575 (0.001) |
| cm1 | 0.265 (0.043) | 0.074 (0.033) | 0.091 (0.022) | 0.244 (0.003) | 0.217 (0.000) |
| kc1 | 0.408 (0.005) | 0.309 (0.006) | 0.450 (0.006) | 0.339 (0.005) | 0.407 (0.000) |
| kc2 | 0.421 (0.011) | 0.465 (0.016) | 0.550 (0.026) | 0.530 (0.001) | 0.517 (0.000) |
| kc3 | 0.377 (0.037) | 0.232 (0.032) | 0.255 (0.072) | 0.328 (0.001) | 0.336 (0.000) |
| mw1 | 0.290 (0.010) | 0.266 (0.083) | 0.375 (0.049) | 0.220 (0.002) | 0.283 (0.000) |
| pc1 | 0.205 (0.021) | 0.233 (0.028) | 0.268 (0.026) | 0.347 (0.001) | 0.393 (0.000) |
| $w/t$ | 3/0/2 | 6/1/0 | -- | 0/0/3 | 1/0/3 |

| Table 2 | Data sets |
|---|---|---|---|---|
| project | modules | attributes | size(loc) | %defective |
| ar3 | 63 | 29 | 5,624 | 12.7 |
| ar4 | 107 | 29 | 9,196 | 18.69 |
| ar5 | 36 | 29 | 2,732 | 12.69 |
| cm1 | 498 | 21 | 14,763 | 9.83 |
| kc1 | 522 | 21 | 42,965 | 15.46 |
| kc2 | 522 | 21 | 19,259 | 20.49 |
| kc3 | 522 | 21 | 19,259 | 20.49 |
| mw1 | 403 | 38 | 8341 | 7.69 |
| pc1 | 1109 | 21 | 21,592 | 6.94 |

1111 Mann-Whitney U test is a non-parametric statistical hypothesis test to compare two independent groups of sampled data, which is without an assumption of a normal distribution [12].


H.B. Mann and D.R. Whitney, “On a test of whether one of two random variables is stochastically larger than the other,” The annals of mathematical statistics, vol.18, no.1, pp.50-60, 1947.