International Journal of Grid and Distributed Computing
Vol. 9, No. 6 (2016), pp.159-168
http://dx.doi.org/10.14257/ijgdc.2016.9.6.16

A New Ensemble Model based Support Vector Machine for Credit Assessing

Jianrong Yao[1], Cheng Lian[2]

[1,2] Information College, Zhejiang University of Finance and Economics
[1] y6310@163.com, [2] ppoi1477@163.com

Abstract

With the rapid growth of internet finance, the credit assessing is becoming more and more important. An effective classification model will help financial institutions gain more profits and reduce the loss of bad debts. In this paper, we propose a new Support Vector Machine (SVM) based ensemble model (SVM-BRS) to address the issue of credit analysis. The model combines random subspace strategy and boosting strategy, which encourages diversity. SVM is considered as a state-of-art model to solve classification problem. Therefore, the proposed model has the potential to generate more accuracy classification. Accordingly, this study compares the ANN, LR, SVM, Bagging SVM, Boosting SVM techniques and experience shows that the new SVM based ensemble model can be used as an alternative method for credit assessing.

Keywords*: credit analysis; ensemble learning; support vector machine; boosting*

1. Introduction

Internet finance is becoming a hot topic for its huge value to economy. Credit assessment is an important part for granting loan. A good credit risk assessment method can help financial institutions to grant loans to creditable applicants, thus increase profits; it can also deny credit for the non-creditable applicants, so decrease losses [1]. There are two main ways applied into this field. One is statistical learning method, another is intelligent method.

Since Fisher created linear regression to solve classification problem[2], many statistical methods were proposed to deal with two-class problem, such as logistic regression model (LR) [3],discriminant analysis[4][5]. However, statistical approaches required certain data distributions, which don’t correspond to reality. Artificial intelligence (AI) methods address the above shortcomings and outperform statistical methods [6].The popular AI methods include the support vector machine (SVM) [7] [8], artificial neural network (ANN) [9] [10] [11], and genetic programming (GA) [12].

Recent researches reported that classification accuracy can be improved by ensemble method, which combine a set of weak models into a strong model [14] [15].Bagging, random subspace, boosting are three popular ensemble strategies, which have been widely applied in machine learning. Bauer and Kohavi report that boosting performs better than bagging, when data does have little noise [16]. Boosting may create ensembles that are less accurate than a single classifier and bagging is almost always more accurate than a single classifier, which is sometimes much less accurate than boosting [17].AdaBoost is the most widely used boosting method. Li, Wang, Sung proposed AdaBoostSVM to solve classification problems [18]. Compared with Decision Trees and Neural Networks classifiers using AdaBoost approaches, AdaBoostSVM performs better. Zhou, *et al*. investigated the performance of least squares SVM ensemble models for credit scoring [13]. Experiment show that ensemble models can provide excellent solutions for credit risk assessment. Wang and Ma proposed RSB-SVM, which is based on bagging and random subspace and SVM as

ISSN: 2005-4262 IJGDC

base learner [7].They use linear and polynomial kernel function. The results show that RSB-SVM is a promising solution to analysis credit risk.

Therefore, SVM based ensemble strategies (SVM-RSB), which is based on bagging and random subspace ensemble approaches, is proposed for credit risk assessment. Two conditions affect ensemble model performance: accuracy and diversity. For accuracy, we adopt SVM as base model, which is a state of art model in practice and theory. For diversity, random subspace creates diversity through selection of features. Boosting changes training data weighting and combines a set of weak classifiers.

The organization of this paper is as follows. Section 2 gives an overview of ANN, bagging, and random subspace. In section 3, this article describes the proposed SVM-RSB approach for credit risk analysis. Section 4 compares SVM-RSB, ANN, SVM, and LR over German credit data. In section 5, this paper draw a conclusion.

2. Background

2.1 Bagging

Breiman was the first man of introducing the bagging method [14]. Bagging approach generates different training data by replacing original data. Suppose we have T sets of training sets. Through repeated sampling, we get a new training sets with the same size of original sets. Repeated above step N times, there are N different training sets derived from original sets. The general idea comes from the bootstrap. Then, different classifiers are trained in each training set. Classifiers are combined into a final model by majority voting.

It can be able to produce enough training sets, especially for data shortage. The method is used to reduce the base classifiers variance, which can avoid overfitting and boost the performance of model.

2.2 Random Subspace

Random subspace is a widely used ensemble strategy, which is first proposed by Ho [19]. The main idea is very similar to bagging in the sense that we bootstrap samples, however, the process is taken place in feature space. Suppose that sample dimension is p. We random draw m variables from the p variables, constructing a new training set. By repeating n times, we can construct many different classifiers. By simple majority voting, we get a final model.

2.3 Boosting

AdaBoost, proposed by Freund and Schapire [20], is the widely applied boost algorithms. The key process of boost is that it combines a set of weak classifiers into a strong classifiers. A weak classifier is the one who accuracy is a litter better than randomly guessing.

Through modifying the weight of instances, we get a sequence of weak classifiers. Instances that incorrectly classified in the current iteration will have a higher weights in the next iteration, otherwise the observation weights are decreased. In the final, the predictions of all weak classifiers will be aggregated. The more accurate classifiers will get high weights in the final model, whereas get low weights.

2.4 Support Vector Machine

The support vector machine, which was developed by Cortes and Vapnik [21], is a kind of modern classification method applied to credit risk assessment. Kernels are used to enlarge feature space by mapping original input space into a high-dimensional feature space. Then, it seeks to find an optimal hyperplane which separates two different data well. Figure 1 explain the main idea of SVM.

Now suppose that we have a set of training observations $\mathrm{D} = \{(x_{1,}y_1), \dots, (x_{m,}y_m)\}$,where $x_i \epsilon R^n, y_i \epsilon \{-1, +1\}$,represents two different class. Using training data, the main idea of SVM is to find a perform classifier by solving the following convex quadratic optimization problem.

$\min \frac{1}{2} w^T w + C \sum_{i=1}^{N} \varepsilon_i$

Subject to $y_i(w^T x_i + b) \geq 1 - \varepsilon_i \ (\varepsilon_i \geq 0, i = 1, \dots, N)$

In the equation, w is norm vector of the optimal hyperplane, ε_i is called error variable that x_i is allowed to stand the wrong side of classify, C stands for the nonnegative tuning parameter on the training error and is used to trade off the classifier complexity and training error. $W^T x_i = K(x, x')$, which is called kernel function. Other popular kernel functions are as follows:

Polynomial:$\mathrm{K}(\mathrm{x}, \mathrm{x}') = (1 + < \mathrm{x}, \mathrm{x}' >)^d$

Radial basis:$\mathrm{K}(\mathrm{x}, \mathrm{x}') = \exp(-\gamma ||\mathrm{x} - \mathrm{x}'||^2)$

3. A New SVM based Ensemble Model for Credit Analysis

Statistical methods have been applied into building models to analysis risk for years. However, it depended on some hypotheses between input and output, which can't be satisfied in the real world. AI doesn't need any priori training data information, so AIs are proposed to construct risk assessment model. Considering that accuracy is a key issue for business profit, we always seek better model.

In the world of academy, a lot of researches show that ensemble approach can improve the classification accuracy. In the industry, ensemble approaches also show their powerful energy. The Netflix prize was won by combining 107 different machine learning models.

Accuracy and diversity are two conditions for guaranteeing that ensembles model perform better than stand-alone model [22]. In the previous section, it is shown that SVM is one of the most accuracy for credit risk assessing. It satisfies the first condition. For the second condition, we choose random subspace and boosting as ensemble strategies. Random subspace construct different training data through choosing input features. However, boosting generates diversity by combing a sequence of classifiers. Marina compare three combining techniques, *i.e.*, bagging, random subspace and boosting, with linear discriminant analysis carried out for several data sets [23]. Though ensemble methods often perform better than a single model, different ensemble methods have different performance depended on data sets. Wang and Ma proposed RSB-SVM, which is based on bagging and random subspace and uses SVM as a base learner [7]. They find that RSB-SVM is the best methods among SVM, DT, ANN, Boosting SVM, Bagging SVM, and Random Subspace SVM. However, it doesn't test boosting and random subspace. Wang and Ma propose that credit risk model based on integrating boosting and random subspace [24]. The paper indicates that RS-Boosting is a better way to analysis credit risk. However, it only uses DT as base learners.

Based on the above literature, we propose a new RFB-SVM method. It uses SVM as a base classifier and blends two popular ensemble strategy, which are random subspace and boosting. Random subspace is an attribute partitioning methods and Boosting is an instance partitioning method. Through them, we get enough diversity. It has a distinct advantage than random subspace and boosting individually. First, random subspace divides the training sets into subsets by resampling feature. Then, SVM was trained in sub data. After that, instances were reweighted by boosting. The whole framework of SVM-BRS is showed in the Figure 2.

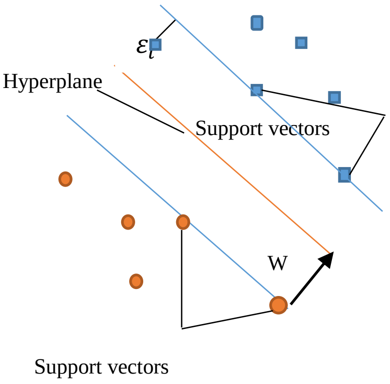

Figure 1. Support Vector Machine

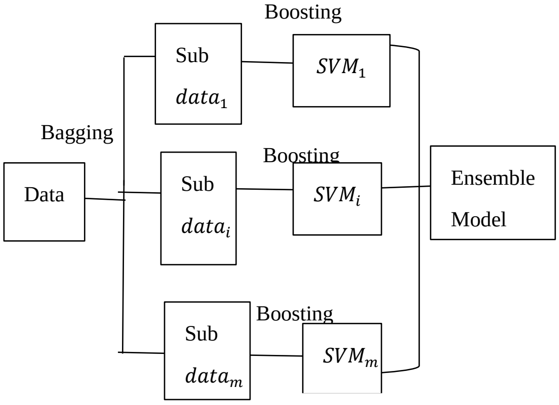

Figure 2. Framework of RSB-SVM

3.1. Partitioning Original Data

In order to perturbing training sets, random subspace is used for creating samples. The original data D has n instances, $e.g. x_i = (x_1, x_2, \dots, x_n)$. Each instances has p features, $e.g. x_{ij} = (x_{i1}, x_{i2}, \dots, x_{ip})$. Instances are represented by i and j represents features. The data set can be image expressed in the below table. Suppose that we extract r(r<p) features in every iteration and get new subsets $\tilde{x} = (x_{i1}, x_{i2}, \dots, x_{ir})$.

The explicit algorithm is shown in the Figure 4.

```
Input original data
For n=1 to N
    Select r(r<p) features from original data and construct new subset $\tilde{x} = (x_{i1}, x_{i2}, \dots, x_{ir})$.
    Fit a base algorithm $T_n$ to the new subset
end.
Output ensemble model $T = sign \sum_{n=1}^{N} T_n(x)$.
```

Figure 3.The Random Subspace Algorithm

3.2. Creating diversity support vector machine

Hansen and Salamon [22] revealed that a necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diversity. Therefore, how to generate diversity is important factor. For SVM model, training data and parameter setting are two way to create diversity. In the previous step, we get diversity of training data. The sector proposes diversity of parameter setting by grid research. Hsu,Chang,&Lin reveal that grid research can be used to choose parameters of SVM[25]. This is an intuitive way. It tries pairs of parameters and select some of them according to standard.

In linear kernel SVM, parameter C control the margin. Larger C lead to wide margin and small C lead to large margin. We randomly choose a set of C and get corresponding receiver operating characteristic (ROC). Then, we select some parameter C according to ROC.

The true positive (TP) is the number of companies that are actually good for which we predict good. The true negative (TN) is the number of companies that are actually bad for which we predict bad. The false positive (FP) is the number of companies that are actually bad for which we predict good. The false negative (FN) is the number of companies that are actually good for which we predict bad.

$$\text{Sensitivity} = \frac{TN}{TN+FP}$$

$$\text{Specificity=} \quad \frac{TP}{TP+FN}$$

Receiver operating characteristic (ROC) explain the relationship between specificity and sensitivity,where sensitivity in the Y axis and (1-specificity) in the X axis. The area under the ROC curve(AUC) is the figure area consisting of the curve and X axis. It ranges from 0.5 to 1, with larger values representing higher model performance[26]. AUC is a traditional measure to evaluate the model.

The range of parameter C value is 0.01 to 100. Fitting with subsets, we receive diversity of SVM models corresponding AUC. Then, we select m models which have large AUC value. The grid research approach is shown in the Figure 6.

```
Input training data
For parameters C=a to b    % a and b is constant
    Fit a base algorithm $T_c$ to the data
    Calculate corresponding AUC
End
Order base algorithm $T_c$ according to AUC
Select m models with large AUC value
Output m models $T_m$
```

Figure 3.The Gird Research Algorithm

There are two parameters in the radial kernel SVM. Applying to the same way, we get a set of SVM models with radial kernel.

3.3. Creating Boosting SVM

In the last decade years, boosting is widely applied to classification problem. Thought it often combines weak classifiers, it also has positive influence in strong classifiers. Error rate is representatives by ε, which is defined as the number of misclassifications dividing summary of instances. Algorithm of boost is described in the Figure 4.

```
Input training data
$T_m(x)$ is the base classifier, m=1,2,…,M
Initialize the weight distribution of training data $w_i = \frac{1}{n}, i = 1,2,\dots,n$
For m from 1 to M
   Compute the misclassification error rate $\varepsilon_m = \mathrm{P}(T_m(x_i) \neq y_i)$
   Calculate coefficients of $T_m(x)$
        $\alpha_m = \frac{1}{2} log \frac{1-\varepsilon_m}{\varepsilon_m}$
   Normalization constant is $Z_m = \sum_{i=1}^{N} w_i \exp(-\alpha_m y_i T_m(x_i))$
   Update the weight $w_i = \frac{w_i}{Z_m} \exp\left(-\alpha_m y_i T_m(x_i)\right)$
End
Output $\mathrm{T(x)} = \mathrm{sign}(\sum_{m=1}^{M} \alpha_m T_m(x))$
```

Figure 4.The AdaBoosting Algorithm

3.4. Integrating Diversity Classifiers into an Ensemble Output

Depended the previous work, we collect a number of single models. Next, we adopt appropriate ensemble strategy to aggregate these appropriate members. Majority voting is widely used ensemble strategy for its easy implementation. The whole algorithm is descripted in the Figure 8. In the following table, a,b,c and d is constant.

```
Input original data
For n=1 to N
    Select r(r<p) features from original data
    Construct new subset $\tilde{x} = (x_{i1}, x_{i2}, \dots, x_{ir})$      % random subspace
    For parameter C= a to b        % liner kernel
        Or pairs of parameter $(C,\gamma)$=(a, b) to (c,d) % radial basis function kernel
        Select m models according to its AUC          % grid research
    End
    Accept boosting algorithm $T_n(\mathrm{x}) = \mathrm{sign}(\sum_{m=1}^{M} \alpha_m T_m(x))$    % boosting
End
Output ensemble classifier $\mathrm{T(x)} = sign(\sum_{n=1}^{N} T_n(x))$    % majority voting
```

Figure 5.The SVM-BRS Algorithm

4. Experimental Analysis

4.1. Data Set

In order to evaluate the performance of SVM based ensemble strategy, we use credit dataset from the website kaggle. The Credit Dataset includes 150000 instances. The datasets consists of 11 attributes. We determine whether to accept or reject loan application by using classification model. The particular explanations of these attributes are showed in Table I.

Table I. Variables of Credit Data

Variable Name	Description	Type
SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	Percentage
Age	Age of borrower in years	Integer
NumberOfTime30-59 DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	Integer
DebtRatio	Monthly debt payments, alimony, living costs divided by monthy gross income	Percentage
MonthlyIncome	Monthly income	Real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or	Integer

	mortgage) and Lines of credit (*e.g.* credit cards)	
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	Integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	Integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	Integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children *etc*.)	Integer

4.2. Evaluation Criteria

Typically, total accuracy, Type I error and Type II error are three evaluation criteria in the field of binary classification .Table 2 is a confusion matrix, which presents evaluation criterion. The true positive (TP) is the number of companies that are actually good for which we predict good. The true negative (TN) is the number of companies that are actually bad for which we predict bad. The false positive (FP) is the number of companies that are actually bad for which we predict good. The false negative (FN) is the number of companies that are actually good for which we predict bad.

Accuracy= $\frac{TP+TN}{TP+FP+FN+TN}$ Type II error rate= $\frac{FP}{TN+FP}$

Type I error rate= $\frac{FN}{TP+FN}$

Table II. Confusion Matr

	Predicted positive(non-risk)	Predicted negative(risk)
Actual positive(non-risk)	True Positive(TP)	False Negative(FN)
Actual negative(risk)	False Positive (FP)	True Negative(TN)

Normally, confusion matrix is used to estimate the accuracy. A confusion matrix is a matrix that shows the relationships between true class and predicted class. The below figure shows a confusion matrix.

Additionally, 5-fold cross validation is presented in order to reduce the variance of the results. In detail, the training data randomly split 5 equal size folds. Specially, the 4 data sets will be used as the training sets while the remaining sets will be reserved as the testing data. The procedure is repeated 5 times so as to each fold can be considered as testing data. The final result are means of 5-fold cross validation.

4.3. Experimental Results

We do experiment in the data mining tool R. With R, an open source software, we can extract all the information using a set of machine learning algorithms. In this paper, we compare SVM based method with LR, ANN. Traditionally, we set r=$\sqrt{p}$ in random subspace, while other parameters are set according to sector III. We choose radial basis

The average accurate, Type I and Type II error are showed in the table III.

Table III. Results of Different Model

Methods	Accuracy	Type I error	Type II error
LR	76.1%	11.71%	52.33%
ANN	70.5%	21%	49%
SVM	76%	11%	53.67%
Bagging SVM	76.2%	11.32%	56%
Random Subspace SVM	75.1%	9.14%	61.67%
Boosting SVM	76.2%	11%	53.7%
SVM-BRS	77.4%	12%	47.3%

From the table, SVM-BRS has the highest average accuracy of 77.4%. Study shows that SVM based model has a perfect performance, which is around 76%.Therefore, SVM is ideally chosen as a base learner. The new SVM-RSB significantly boosts performance.

Though RSB-SVM has the highest accuracy, it increases Type I error. Meanwhile, the new model reduces Type II error. Type I error is the number of instances that are actually good for which we predict bad, while Type II error is the number of instances that are actually bad for which we predict good. As everyone know, bad loan get great loss, which could enough offset profits of good grant. From the above discussion, we can draw a conclusion that RSB-SVM is an alternative approach for credit analysis.

5. Conclusions

Credit risk assessment is important for institutions who will decide whether to accept or reject loan application. In this paper, we propose a new SVM-RSB method based on random subspace, boosting and support vector machine classifier to address the issue of credit risk analysis. In the empirical analysis part, German credit dataset is used to check the model performance. Experiments show that SVM-RSB gets the best performance among 7 methods, i.e., LR, ANN, SVM, Random Subspace SVM, Bagging SVM, and Boosting SVM.

The paper prove that ensemble model performs better than a single model. However, the study only tests German credit data and consider SVM as a base learner. Machine learning is criticized for black box, which we only know the result and hardly explain the result. For further study, it can use other base learner except SVM and explore more credit data. On the other hand, more interpretative model should be explored in the future.

References

[1] G. Wang, J. Ma, L. Huang, & K. Xu, “Two credit scoring models based on dual strategy ensemble tree”, Knowledge-Based Systems, vol. 26, **(2012)**, pp. 61-68.

[2] R. A. Fisher, “The use of multiple measurements in taxonomic problems. Annals of eugenics”, vol. 7, no. 2, **(1936)**, pp. 179-188**.**

[3] V. S. Desai, J. N.Crook, & Overstreet, G. A, “A comparison of neural networks and linear scoring models in the credit union environment”, European Journal of Operational Research, vol. 95, no. 1, **(1996)**, pp. 24-37.

[4] E. I. Altman, “Financial ratios, discriminant analysis and the prediction of corporate bankruptcy”, The journal of finance, vol. 23, no. 4, **(1968)**, pp. 589-609.

[5] T. K Sung, N. Chang, & G. Lee, “Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction”, Journal of Management Information Systems, **(1999)**, pp. 63-85.

[6] S. Lee, & W. S. Choi, A multi-industry bankruptcy prediction model using back-propagation neural

network and multivariate discriminant analysis. Expert Systems with Applications, vol. 40, no. 8, **(2013)**, pp. 2941-2946.

[7] G., Wang, & J. Ma, “A hybrid ensemble approach for enterprise credit risk assessment based on Support Vector Machine”, Expert Systems with Applications, vol. 39, no. 5, **(2012)**, pp. 5325-5331.

[8] T. Harris, “Credit scoring using the clustered support vector machine”, Expert Systems with Applications, vol. 42, no. 2, **(2015)**, pp. 741-750.

[9] C. F. Tsai, & C. Hung, “Modeling credit scoring using neural network ensembles”, Kybernetes, vol. 43, no. 7, **(2014)**, pp. 1114-1123.

[10] L. Wang, S. & Lai, K. K, “Credit risk assessment with a multistage neural network ensemble learning approach”, Expert Systems with Applications, vol. 34, no. 2, **(2008)** , pp. 1434-1444.

[11] D. West, “Neural network credit scoring models”, Computers & Operations Research, vol. 27, no. 11, **(2000)**, pp. 1131-1152.

[12] C. S. Ong, J. J. Huang, & G. H. Tzeng, “Building credit scoring models using genetic programming”, Expert Systems with Applications, vol. 29, no. 1, **(2005)** pp. 41-47.

[13] L. Zhou, K. Lai, K., & Yu, L, “Least squares support vector machines ensemble models for credit scoring”, Expert Systems with Applications, vol. 37, no. 1, **(2010)**, pp. 127-133.

[14] L. Breiman, “Bagging predictors”, Machine learning, vol. 24, no. 2, **(1996)**, pp. 123-140.

[15] L. Breiman, “Prediction games and arcing algorithms”, Neural computation, vol. 11, no. 7, **(1999)**, pp. 1493-1517.

[16] E. Bauer, & Kohavi, R. “An empirical comparison of voting classification algorithms: Bagging, boosting, and variants”, Machine learning, vol. 36, no. 1-2, **(1999)**, pp. 105-139.

[17] D. Opitz, & R. Maclin, “Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research”, **(1999),** pp. 169-198.

[18] X. Li, L. Wang, & E. Sung, “AdaBoost with SVM-based component classifiers”, Engineering Applications of Artificial Intelligence, vol. 21, no. 5, **(2008)**, pp. 785-795.

[19] T. K. Ho, “The random subspace method for constructing decision forests”, Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 20, no. 8, **(1998)**, pp. 832-844.

[20] Y. Freund, & R. E. Schapire, “Experiments with a new boosting algorithm”, In ICML, vol. 96, **(1996)**, pp. 148-156.

[21] C. Cortes, & V. Vapnik, “Support-vector networks. Machine learning” , vol. 20, no. 3, **(1995)**, pp. 273-297.

[22] L. K Hansen, & P. Salamon, “Neural network ensembles. IEEE Transactions on Pattern Analysis & Machine Intelligence”, 10, **(1990)**, pp. 993-1001.

[23] M. Skurichina, & Duin, R. P, “Bagging boosting and the random subspace method for linear classifiers”, Pattern Analysis & Applications , vol. 5, no. 2, pp. 121-135 **(2002)**.

[24] G. Wang, & J. Ma, “Study of corporate credit risk prediction based on integrating boosting and random subspace”, Expert Systems with Applications, vol. 38, no. 11, **(2011)**, pp. 13871-13878.

[25] C. W Hsu, C. C. Chang, & C. J Lin, “A practical guide to support vector classification”, **(2003).**

[26] S. Wang, D Li, N. Petrick, B. Sahiner, M. G., Linguraru, & R. M Summers, “Optimizing area under the ROC curve using semi-supervised learning”, Pattern recognition, vol. 48, no. 1, **(2015)**, pp. 276-287.