BEST: International Journal of Humanities, Arts,
Medicine and Sciences (BEST: IJHAMS)
ISSN (P): 2348-0521, ISSN (E): 2454-4728
Vol. 4, Issue 10, Oct 2016, 75-86

CREDIT SCORING IN IRAN WITH CSVM AND LOGIT MODEL

MOHAMMADREZA MOHAMMADI[1], HAMID ASAYESH[2] & MOHAMMAD JAVAD TAHERITIKORDI[3]

[1]Imam Khomeini Naval University of Noshahr, Iran

[2]Lecturer Faculty of Humanities, University of Ayatollah Alozma Boroujerdi (RA), Iran

[3]Expert Supreme Audit Court, Iran

ABSTRACT

Features and specifications such as The status of the client's existing checking account, The duration of the credit period in months, The client's credit history, The purpose for the credit, The credit amount requested, The client's savings account/bonds balance and ...., And the different methods used to determine good customers from the bad accounts, but the large volume of outstanding bank loans. This paper investigates the practice of credit scoring and introduces the use of the clustered support vector machine (CSVM) for credit scorecard development in Iran. Accordingly, with a sample of 3000, this study shows that the CSVM can achieve comparable levels of classification performance while remaining relatively cheap computationally. Classifications by model CSVM factors affecting lack of timely repayment of credit facilities include: The client's housing arrangements (i.e. own their home, rent, or live for free), the applicant's monthly income, the duration of the credit period in months, How to refund, the credit amount requested.

KEYWORDS : Credit Risk, Credit Scoring, Clustered Support Vector Machine

INTRODUCTION

In recent years, credit risk assessment has attracted significant attention from managers at financial institutions around the world. This increased interest has been in no small part caused by the weaknesses of existing risk management techniques that have been revealed by the recent financial crisis and the growing demand for consumer credit. Addressing these concerns, over past decades credit scoring has become increasingly important as financial institutions move away from the traditional manual approaches to this more advanced method, which entails the building of complex statistical models. Many of the statistical methods used to build credit score cards are based on traditional classification techniques such as logistic regression or discriminated analysis. Thus, many specially tailored optimization algorithms have been proposed. The first class of such algorithms tries to solve the entire optimization problem by solving a series of small problems. The basic techniques include chunking and decomposition, which were discussed by Boser et al. The success of algorithms depends on an appropriate criterion for the active set selection and an efficient strategy to cache the matrix.

This paper investigates the suitability for credit scoring of a recently developed support vector machine based algorithm that has been proposed by Gu and Han (2013). Their clustered support vector machine has been shown to offer comparable performance to kernel based approaches while remaining cheap in terms of computational time.

CREDIT SCORING

Credit scoring can be viewed as a method of measuring the risk attached to a potential customer, by analyzing their data to determine the likelihood that the prospective borrower will default on a loan. According to Eisenbeis (1978), Hand and Jacka (1998), and Hand, Sohn, and Kim (2005) credit scoring can also be described as the statistical technique employed to convert data into rules that can be used to guide credit granting decisions.

As a result, it represents a critical process in a firm's credit management toolkit. Durand (1941) posited that the procedure includes collecting, analyzing and classifying different credit elements and variables in order to make credit granting decisions. He noted that to classify a firm's customers, the objective of the credit evaluation process, is to reduce current and expected risk of a customer being ‘‘bad’’ for credit. Thus credit scoring is an important technology for banks and other financial institutions as they seek to minimize risk.

RELATED WORKS

Over the years, the demand for consumer credit has increased exponentially. According to Steenackers and Goovaerts (1989), this increase in the demand for credit can be attributable to the increased levels of consumption and the reliance on credit to support this activity. In the United States, this rising level of consumerism followed the introduction of the first modern credit card in 1950s, so that by the 1980s over 55% of American households owned a credit card. Crook, Edelman, and Thomas (2007) posited that by this time, in the US, the total amount of outstanding consumer credit was over $700 billion. Comparatively, at the end of June 2013 this figure had risen to a staggering $2800 billion, a 400% increase.

Henley (1994) noted that the increasing demand for consumer credit has led to the development of many practical the scoring models, which have adopted a wide range of statistical and nonlinear methods.

Orgler (1970) was one of the first researchers to use linear regression for credit scoring. He used regression analysis to model commercial loans and to evaluate outstanding consumer loans. Conducting this work he realized that this approach was somewhat limited but could be used to review loans.

Similarly, Mays (2001) posited that a number of various techniques have been used to build credit scoring applications by credit analyst, researchers, and software developers. These techniques have included; discriminated analysis, linear regression, logistic regression, decision trees, neural networks, support vector machines, k-means, etc.

In recent times, the use of more complex non-linear techniques, such as neural networks, and support vector machines, to build credit scoring applications has seen significant increases in the reported accuracy and performance on benchmarking datasets. Irwin, Warwick, and Hunt (1995) and Paliwal and Kumar (2009) both provide evidence that advanced statistical techniques yield superior performance when compared to traditional statistical techniques, such as discriminated analysis, probity analysis and logistic regression. Masters (1995) also provided evidence that the use of sophisticated techniques, such as neural networks, was essential because they had the capability to more accurately model credit scoring data that exhibits interactions and curvature. However, as pointed out by Hand (2006) the increased performance of these more advanced techniques

Could be illusionary and if real, diminished due to shifts in the class distribution over time. The following sub-sections present a brief discussion concerning some of the classical and advanced statistical models used for credit risk assessment.

As has been noted in the preceding paragraphs a wide range of statistical and more.

MEASURES & METHOD OF RESEARCH

Modern classification technique applied to credit scoring is the support vector machine. First developed by Cortes and Vapnik (1995) for classification problems, the support vector machine attempts to find the optimal separating hyper plane among the classes by maximizing the margin between them. Here, points lying on the boundaries are referred to as support vectors, while the middle of the margin is called the optimal separating hyper plane. It is this margin maximization property of the support vector machine that is said to lead to its state-of-the-art classification performance. However, like the artificial neutral network this technique is often criticized as being a ‘‘black box.‘‘

Furthermore, it remains computationally expensive relative to traditional statistical methods. Support Vector Machines (SVM) technique is one of the most powerful classification techniques that was successfully applied to many real world problems. Support Vector Machines are based on the idea of mapping data points to a high dimensional feature space where a separating hyper plane can be found. This mapping can be carried on by applying the kernel trick which implicitly transforms the input space into another high dimensional feature space. The hyper-plane is computed by maximizing the distance of the closest patterns, i.e., margin maximization, avoiding the problem of over fitting. The separating hyper plane, i.e., maximal margin, is found using a quadric programming routine which is computationally very expensive. Furthermore, this routine depends on the data set size, taking impractical time when dealing with huge data sets. Proposed a fast training algorithm based on quick identification of support vectors, however, their algorithm appears to have some difficulties in dealing with linearly separable training data set. In addition, Keerthi et al proposed an algorithm based on observations about the geometrical properties of support vector machines.

Understanding the Support vector machines (SVM) are new generation of machine learning techniques and have shown strong generalization capability for many data mining tasks. SVM can handle the nonlinear classification by implicitly mapping Input samples from the input feature space into another high dimensional feature space with the nonlinear kernel function. Many applications, such as Data Mining and Bio-Informatics, require the processing of huge data sets. The training time of SVM is a serious obstacle for this kind of data sets. , it would take years to train SVM on a data set of size one million records. Many proposals have been submitted to enhance SVM to increase its training performance, either by random selection or by approximation of the marginal classifier. However, they are still not feasible with large data sets where even multiple scans of entire data set are too expensive to perform, or they end up losing the benefits of using an SVM by over-simplification.

Therefore, SVM may be more effective to reveal the nonlinear sequence to structure relationship than K-means clustering does. The superior performance for non-linear classification inspires us to explore the relationship between the protein sequence and its structure with SVM.

Since SVM is not favorable for a large dataset, modeling of one SVM over the whole sample space containing almost half million data samples is impractical. Furthermore, each subspace of the whole sample space corresponds to different local 3D structures in our application. This paper proposes a new approach for enhancing the training process of SVM when dealing with large data sets. It is based on the combination of SVM and clustering analysis. The idea is as follows: SVM computes the maximal margin separating data points; hence, only those patterns closest to the margin can affect the computations of that margin, while other points can be discarded without affecting the final result. Those points lying close to the margin are called support vectors. We try to approximate these points by applying clustering analysis.

Traditionally, clustering algorithms can be classified into two main types, namely, hierarchical clustering and partitioning clustering. Partitioning, also called flat clustering, directly seeks a partition of the data which optimizes a predefined numerical measure. In partitioning clustering, the number of clusters is predefined, and determining the optimal number of clusters may involve more computational cost than clustering itself. Furthermore, a priori knowledge may be necessary for initialization and the avoidance of local minima. As a result, construction of one SVM for the whole sample space cannot take advantage of the strong generalization power of SVM efficiently. The disadvantage of building one SVM over the whole sample space motivates us to consider the theory of granular computing. Using the divide-and-conquer principle, granular computing is able to divide a complex data-mining problem into a series of smaller and computational simpler problems. To combine the theory of granule computing and principles of the statistical learning algorithms, we propose a new computational model called Clustering Support Vector Machines (CSVMs) in our work. In this new computational model, one SVM is built for each information granule defined by sequence clusters created by the clustering algorithm. CSVMs are modeled to learn the nonlinear relationship between protein sequences and their structures in each cluster. SVM is not favorable for large amount of data samples. This paper formulated a supervised clustering method SVM cluster based on an SVM framework for learning structured outputs. The algorithm accepts a series of “training clusters,” a series of sets of items and clustering’s over that set. The method learns a similarity measure between item pairs to cluster future sets of items in the same fashion as the training clusters the learning algorithm’s correctness depends on an ability to iteratively find and introduce the most violated constraint. Since finding the most violated constraint is intractable for clustering, we use existing clustering methods to help find an approximation. However, CSVMs can be easily parallelized to speed up the modeling process. After gaining the knowledge about the sequence to structure relationship, CSVMs are used to predict distance matrices, torsion angles and secondary structures for backbone a-carbon atoms of protein sequence segments.

Since K-means clustering is computationally efficient for large data sets with both numeric and categorical attributes, K-means clustering is chosen as the granulation method in our study. With the K-means clustering algorithm, data samples with similar characteristics can be grouped together. As a result, the whole sample space is partitioned into subspaces intelligently and the complex data mining work is mapped into a series of computationally tractable simpler tasks. Different number of initial clusters were tried and based on these results, clusters were chosen.

In many problems of machine learning, the data points are distributed non-linearly. In this case, linear classifiers such as linear support vector machine and linear regression are not able to classify the data points correctly, because they fail to consider the underlying structures of complex data (e.g., clusters and manifolds). One way to solve this problem is to train a nonlinear classifier such as kernel support vector machine, which implicitly maps the input data into a high dimensional (or even infinite dimensional) feature space and learns a hyper plane in the new feature space to separate the mapped data. Unfortunately, although the training of linear SVM has received substantial advance in the past years, the best known time complexity of training a kernel SVM is still quadratic to the number of examples.

Therefore, it is necessary to develop some classifiers which are able to handle the nonlinear data, while having considerably lower time complexity than nonlinear classifiers. In our study, we are interested in large margin classifiers, because they are based on max-margin principle, which is theoretically sound. They have also achieved great success in a wide range of applications.

In this paper, we propose a novel large margin classifier namely Clustered Support Vector Machine (CSVM). In particular, we first divide the data into several clusters by K-means, and in each cluster, we train a linear support vector machine. To avoid over-fitting of each local SVM, we add a global regularization, which requires the weight vector of linear SVM in each cluster aligning with a global reference weight vector. The resulting CSVM can be efficiently solved by stochastic gradient descent, or dual coordinate descent.

Although SVMs are stable in high-dimensional space. But the choice of optimal subset of features with additional features and waste removal classifier performance in terms of accuracy, speed and cost improve. In recent years, support vector machines as one of the most efficient classifier have been stable. One of the factors affecting the performance of support vector machines, parameters is. Choosing the optimal subset classified as input another important step in the optimization of support vector machines to classify trans-spectral images.

Based on the literature, characteristics that affect customer credit behavior include: Features and specifications such as: The status of the client's existing checking account, The duration of the credit period in months, The client's credit history, The purpose for the credit, The credit amount requested, The client's savings account/bonds balance, The client's present employment status, The client's personal (marital) status and sex, Whether the client is a debtor or guarantor of credit granted, by another institution, The number of years spent at present residence, The type of property possessed by client, The client's age in years, Whether the client has other installment plans, The client's housing arrangements (i.e. own their home, rent, or live for free), The number of existing credits the client has at the bank, The client's job, The number of people for whom the client is liable to provide , maintenance for Whether the client has a telephone, Whether the client is a foreign worker, The number of months at current address, the applicant's marital status, The number of dependents, The age of first dependent, The age of second dependent, The age of third dependent, The applicant's employment status, The number of years employed with current employer, The loan amount, The loan purpose, The loan type, The applicant's monthly income, The applicants monthly expenditure of the credit file is received from clients used to determine good customers from the bad accounts.

Consistent with Harris (2013) the term evaluation metric will be used when referring to the metric used during the training phase, and the term performance metric used to refer to the measure used to report models performance at the reporting phase.

The Area under the Receiver Operating Characteristic (ROC) curve (AUC) is designated as the primary model evaluation metric and performance metric in this study. The AUC measure is highlighted as in below where, S1, represents the sum of the ranks of the creditworthy clients.

Sensitivity= True Positive/(True Positive + False negative)

Specificity= True Negative/(False Positive + True Negative)

AUC=(◌́s1 _ Sensitivity)*(◌́ Sensitivity +1)*0.5/(Sensitivity +Specificity)

A number of other performance metrics are also used to report the performances of the classifiers developed in this paper. For example, Test accuracy, as in is also reported as it measures how accurately the credit applicants on a withheld test dataset are classified.

Test accuracy = (True Positive/(True Positive + False Positive))+(True Negative/(False Negative +True Negative))

An arguably more meaning full measure of classifier performance is the balanced accuracy (BAC). This measure avoids the misleading affects on accuracy caused by imbalanced datasets by showing the arithmetic mean of sensitivity and specificity. Since skewed datasets are a common occurrence with real world credit scoring datasets this measure may be more relevant.

MODEL CSVM

In this section CSVM approach and the use of MATLAB and variables listed above, to classify the customer's credit status will be discussed. For this purpose because when drawing classification, data should be two-dimensional drawing, significant criteria in selecting pairs of variables in the log it regression or homogeneity of the variables to be considered on the basis of logic. Then, for each pair of first (class) variable considered in the MATLAB code as it is created Classifieds. And then classified based on the output software offers the graph. In the final evaluation for each class CSVM models with test data collection with analysis of variance and sensitivity criteria diagnostic criteria and benchmarks to measure overall accuracy has to be well fitted to each class.

- Classifieds how to ask about the monthly income CSVM (The applicant's monthly income, How to refund)

Software code, this class is as follows:

```
clc
clear all;
Din=xlsread('D3In.xls');
load DOUT
X =Din(,[2,6]);
y = DOUT;
%------------------------------------------------------------------------
%SVMModel = CSVM(X,y);%returns an SVM classifier trained using the predictors
%in the matrix X and class labels in the vector Y for one- or two-class classification.
SVMModel = fitCSVM(X,y);
classOrder = SVMModel.ClassNames;
sv = SVMModel.SupportVectors;
figure;
gscatter(X(,1),X(,2),y);
hold on
plot(sv(,1),sv(,2),'ko','MarkerSize',10);
legend('badhesab','khoshhesab','Support Vector');
```

hold off

% Compute misclassification rate using stratified 10-fold cross validation

% cp = cvpartition(y,'k',10); % Stratified cross-validation

% classf = @(XTRAIN, ytrain,XTEST)(classify(XTEST,XTRAIN,...

% ytrain));

%cvMCR = crossval('mcr',X,y,'predfun',classf,'partition',cp)

xlabel('How to pay');

ylabel('Average monthly income');

title('Clustering based on the average amount of monthly income and How to pay');

Output classification for class CSVM (The applicant's monthly income, How to refund) as follows

Chart 1 class diagrams CSVM (The applicant's monthly income, How to refund)

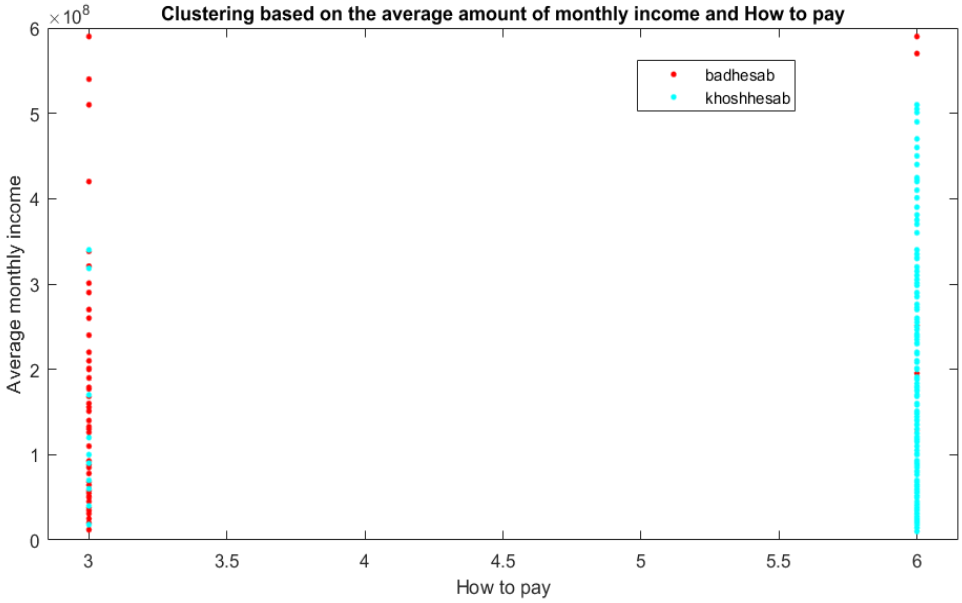

Chart 1

Table 1: Assessment Exam Review Classes CSVM (The Applicant's Monthly Income, How to Refund)

Clustering Made	TPB	TPG	SP	ST	TA	BAC
The applicant's monthly income, How to refund	404	1785	0.99	0.95	1.94	0.97

As can be observed classification accuracy and high detection and classification modeling is very good.

Classifieds housing situation and about the amount of monthly income CSVM (The client's housing arrangements (i.e. own their home, rent, or live for free), The applicant's monthly income)

Software code, this class is as follows:

clc

```
clear all;
Din=xlsread('D3In.xls');
load DOUT
X =Din(,[6,7]);
y = DOUT;
%-------------------------------------------------------------------------
%SVMModel = CSVM(X,y);%returns an SVM classifier trained using the predictors
%in the matrix X and class labels in the vector Y for one- or two-class classification.
SVMModel = fitCSVM(X,y);
classOrder = SVMModel.ClassNames;
sv = SVMModel.SupportVectors;
figure;
gscatter(X(,1),X(,2),y);
hold on
plot(sv(,1),sv(,2),'ko','MarkerSize',10);
legend('badhesab','khoshhesab','Support Vector');
hold off
% Compute misclassification rate using stratified 10-fold cross validation
% cp = cvpartition(y,'k',10); % Stratified cross-validation
% classf = @(XTRAIN, ytrain,XTEST)(classify(XTEST,XTRAIN,...
% ytrain));
%cvMCR = crossval('mcr',X,y,'predfun',classf,'partition',cp)
xlabel('Average monthly income');
ylabel('Housing');
title('Clustering based on the average amount of monthly income and housing situation');
```

Output classification for class CSVM (The client’s housing arrangements (i.e. own their home, rent, or live for free), The applicant’s monthly income) as follows:

Chart 2 class diagrams CSVM (The client’s housing arrangements (i.e. own their home, rent, or live for free), The applicant’s monthly income)

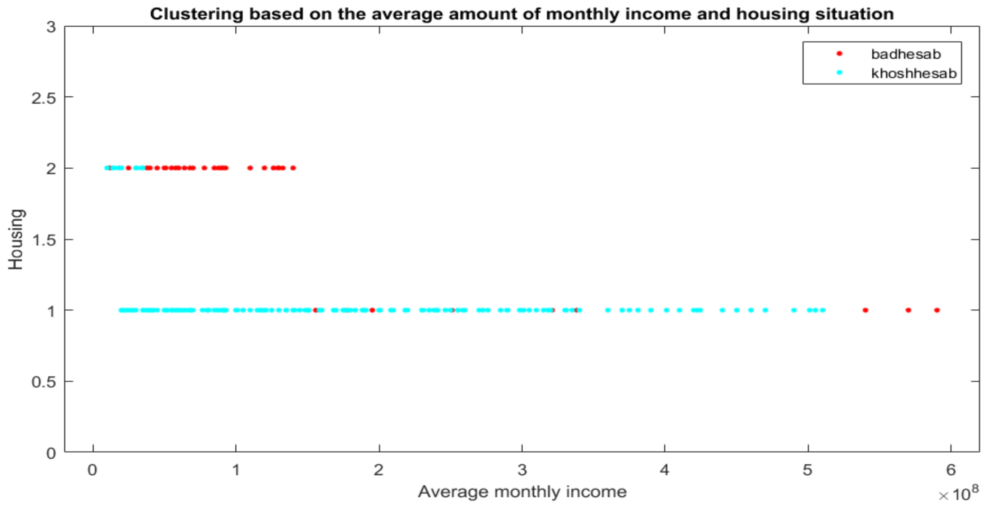

Chart 2

Table 2 assessment exam review classes CSVM (The client's housing arrangements (i.e. own their home, rent, or live for free), The applicant's monthly income)

Table 2

Clustering Made	TPB	TPG	SP	ST	TA	BAC
The client's housing arrangements (i.e. own their home, rent, or live for free), The applicant's monthly income	392	17.84	0.99	0.92	1.91	0.95

As can be seen very high classification accuracy and detection and classification modeling is very good.

As can be observed classification accuracy and high detection and classification modeling is very good.

CONCLUSIONS

Due to the outflow of MATLAB and evaluation tests for each class CSVM model that includes measures analysis of variance and sensitivity (the proportion of bad customer accounts that were correctly classified) diagnostic criteria (the accuracy of the model in the correct identification good customers implies), and the overall accuracy (which reflects the ability of the model to classify all customers, whether good or bad account and the account is) be concluded that CSVM could very well be the probability of repayment of bank predict the actual customer, according estimated based on classifications by model CSVM factors affecting lack of timely repayment of credit facilities include:

- The client's housing arrangements (i.e. own their home, rent, or live for free)
- The applicant's monthly income
- The duration of the credit period in months
- How to refund
- The credit amount requested

REFERENCES

1. Abdou, H. A., & J. Pointon, (2009). Credit scoring and decision making in Egyptian public sector banks.

International Journal of Managerial Finance, 5(4), 391–406.

2. Abdou, H. A., & J. Pointon, (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. Intelligent Systems in Accounting, Finance and Management, 18(2–3), 59–88.
3. Akmal, M &S., Muhammad. (2008). Technical Efficiency of the Banking Sector in Pakistan, SBP Research Bulletin, Vol .4, No. 1, Nov.
4. Bellotti, T., & J.Crook, (2009). Support vector machines for credit scoring and discovery of significant features. Expert Systems with Applications, 36(2), 3302–3308.
5. BGFRS (2013). Consumer Credit Release Board of Governors of the Federal Reserve
6. Crook, J. N., Edelman, D. B., & L.C., Thomas, (2007). Recent developments in consumer credit risk assessment. European Journal of Operational Research, 183(3), 1447–1465.
7. De Morsier, F.; Tuia, D.; Borgeaud, M.; Gass, V.; & J.P., Thiran, (2015), Cluster validity measure and merging system for hierarchical clustering considering outliers. Pattern Recognit. 48, 1478–1489.
8. Francisco, L.; Paulo H; Ferreira-Silva,Carlos A. & R. Diniz.,(2012). On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data, Expert Systems with Applications. 39 , 8071–8078.
9. Ganapathiraju, A., Hamaker, J. and J. Picone, (2004) ‘Applications of support vector machines to speech recognition’, IEEE Transactions on Signal Processing, Vol. 52, No. 8, pp.2348–2355.
10. Gersbach, H., & A., Lipponer, (2003), Firm defaults and the correlation effect, European Financial Management, 9(3)
11. Gu, Q., & J. Han, (2013). Clustered support vector machines. In Paper presented at the proceedings of the sixteenth international conference on artificial intelligence and statistics.
12. Guohai L., Dawei Z., Haixia X. & M. Congli.(2010), Model optimization of SVM for a fermentation soft sensor, Expert Systems with Applications, Vol. 37, pp. 2708–2713
13. Hand, D. J. (2006). Classifier technology and the illusion of progress. Statistical science, 21(1), 1–14.
14. Hastie, T.; Tibshirani, R. & J.Friedman, (2001),The Elements of Statistical Learning; Data Mining, Inference and Prediction; Springer Verlag: New York, NY, USA.
15. Hong W.C., (2011) , Traffic flow forecasting by seasonal SVR with chaotic simulated annealing algorithm, Neuro computing, Vol. 74, 2096-2107
16. Hsu, C.W., Chang, C.C. and C.J. Lin, (2003) A Practical Guide to Support Vector Classification, Working Paper
17. Huang, C.L., Chen, M.C., & C.J. Wang. (2007), Credit scoring with a data mining approach based on support vector machines. Expert Systems with Applications, 847-856.
18. Lee, G.; & C. Scott, (2007), The one class support vector machine solution path. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Honolulu, HI, USA, 15–20; Volume 2, pp.

521–524.

19. Minqiang P.; Dehuai Z. and X. Gang.(2010), Temperature prediction of hydrogen producing reactor using SVM regression with PSO, Journal of computers, Vol 5, No.3
20. Paliwal, M., & U. A. Kumar, (2009). Neural networks and statistical techniques: A review of applications. Expert Systems with Applications, 36(1), 2–17.
21. Pattern Recognit. 2013, 46, 305–316.
22. Qi, Z.; Tian, Y.; & Y. Shi, (2012), Laplacian twin support vector machine for semi-supervised classification. Neural Netw. 35, 46–53.
23. Qi, Z.; Tian, Y.; & Y.Shi, (2013), Structural twin support vector machine for classification. Knowl. Based, Syst, 43, 74–81.
24. Qi, Z.; Tian, Y.; & Y.Shi, (2015), Successive overrelaxation for laplacian supports vector machine. IEEE Trans. Neural Netw. Learn. Syst. 26, 667–684.
25. Qi, Z.; Tian, Y.; Shi, & Y. Twin, (2012), support vector machine with Universum data. Neural Netw, 36, 112–119.
26. Tian, Y.; Qi, Z.; Ju, X.; Shi, Y.& X. Liu, (2014), Nonparallel support vector machines for pattern classification. IEEE Trans. Cybern, 44, 1067–1079.
27. Vert, R.; & J.Vert, (2006), Consistency and convergence rates of one-class SVMs and related algorithms.J. Mach. Learn. Res, 7, 817–854.
28. Wang, Z., Yan, S.C., & C.S. Zhang, (2011). Active learning with adaptive regularization. Pattern Recognition, 44(10–11), 2375–2383.
29. Yang, X.; Peng, G., Cai, Z., & K. Zeng, (2013), Occluded and low resolution face detection with hierarchical deformable model. J. Converge, 4, 11–14.
30. Yin J. (2011). LogP prediction for blocked tripe tides with amino acids descriptors (HMLP) by multiple linear regression and support vector regression", Procedia Environmental Sciences 8, pp. 173–178
31. Yoon, S.H.; & J. Min, (2013), an intelligent automatic early detection system of forest fire smoke signatures using Gaussian mixture model. J. Inf. Process. Syst, 9, 621–632.
32. Yu, L. (2008). Bio-inspired credit risk analysis: Computational intelligence with support vector machines.
33. Zhou, L.; Lai, K. K.; & L. Yu. (2010), Least squares support vector machines ensemble models for credit scoring", Expert System with Applications, 127-133.