Fraud Detection in Transactions of Credit Card- Approach of Classification using Support Vector Machines

Student: Asim Shaik

Under guidance of: Dr Ruoming Jin

05/09/2022

Course: Machine Learning & Deep Learning

ABSTRACT:

This project is used to determine the fraud in the credit card transactions as fraud is increasing along with the development in technology in today's world, So as the increase in the use of credit card. This can be done by different methods like regression, classification etc. Here I have chosen classification approach using Support Vector Machine. The main objective of this approach is to detect frauds with very high accuracy as well as in number.

MOTIVATION:

Credit card payments have grown in recent years. It can be used for both online and offline purchasing. Credit card payments are now both necessary and handy. Due to an increase in fraudulent transactions, a reliable fraud detection model is required.

With the advancement of modern technologies, fraud is on the rise. In today's environment, the volume of information is likewise explosive. Analyzing the cardholder's spending behavior is a promising technique to detect fraud.

Detecting the scam entails locating the suspect. If there is any anomaly in the spending behavior, it is regarded suspicious and is investigated further. For fraud detection, a behavior-based technique based on support vector machines is used.

SVM (Support Vector Machine) is a known study topic that successfully solves classification issues in noisy and complicated environments. Due to its superior generalization performance in a wide range of learning tasks, such as handwritten digit recognition, web page categorization, and face detection, SVM has played a vital role in the field of machine learning. In SVM applications, the problem of overfitting is quite rare.

Statistical learning theory, the theoretical foundation of statistical inference, gave rise to Support Vector Machines. Credit rating analysis, bankruptcy prediction, and time series prediction and classification have all recently used SVM in commercial applications.

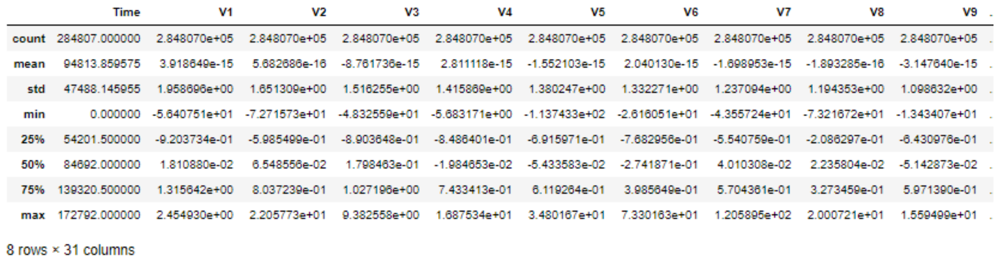
DATASET DESCRIPTION:

The data set that I have used here consists of the transaction details in the form of variables, the amount which has been transacted.

Dataset contains following fields:

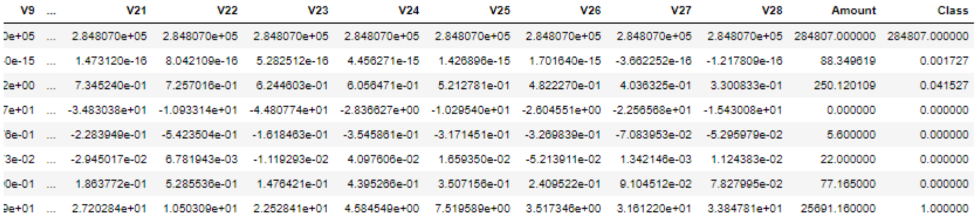
1. Time
2. V1 to V28 transaction details
3. Class

Mean, Standard Deviation has been calculated for this data.
Finally, Quantiles like Minimum, Maximum have been calculated. However, there is no missing and mismatched data in the dataset. I’ve also performed data visualization.



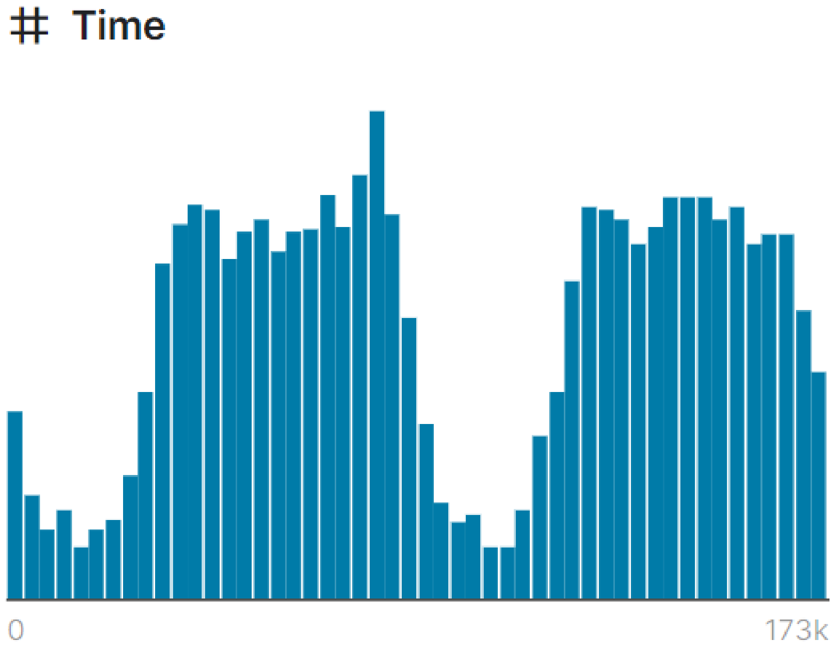
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	.
count	284807.000000	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	.
mean	94813.859575	3.918649e-15	5.682686e-16	-8.761736e-15	2.811118e-15	-1.552103e-15	2.040130e-15	-1.698953e-15	-1.893285e-16	-3.147640e-15	.
std	47488.145955	1.958696e+00	1.651309e+00	1.516255e+00	1.415869e+00	1.380247e+00	1.332271e+00	1.237094e+00	1.194353e+00	1.098632e+00	.
min	0.000000	-5.640751e+01	-7.271573e+01	-4.832559e+01	-5.683171e+00	-1.137433e+02	-2.616051e+01	-4.355724e+01	-7.321672e+01	-1.343407e+01	.
25%	54201.500000	-9.203734e-01	-5.985499e-01	-8.903648e-01	-8.486401e-01	-6.915971e-01	-7.682956e-01	-5.540759e-01	-2.086297e-01	-6.430976e-01	.
50%	84692.000000	1.810880e-02	6.548556e-02	1.798463e-01	-1.984653e-02	-5.433583e-02	-2.741871e-01	4.010308e-02	2.235804e-02	-5.142873e-02	.
75%	139320.500000	1.315642e+00	8.037239e-01	1.027196e+00	7.433413e-01	6.119264e-01	3.985649e-01	5.704361e-01	3.273459e-01	5.971390e-01	.
max	172792.000000	2.454930e+00	2.205773e+01	9.382558e+00	1.687534e+01	3.480167e+01	7.330163e+01	1.205895e+02	2.000721e+01	1.559499e+01	.

8 rows × 31 columns



V9	...	V21	V22	V23	V24	V25	V26	V27	V28	Amount	Class
)e+05	...	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	2.848070e+05	284807.000000	284807.000000
0e-15	...	1.473120e-16	8.042109e-16	5.282512e-16	4.456271e-15	1.426896e-15	1.701640e-15	-3.662252e-16	-1.217809e-16	88.349619	0.001727
2e+00	...	7.345240e-01	7.257016e-01	6.244603e-01	6.056471e-01	5.212781e-01	4.822270e-01	4.036325e-01	3.300833e-01	250.120109	0.041527
7e+01	...	-3.483038e+01	-1.093314e+01	-4.480774e+01	-2.836627e+00	-1.029540e+01	-2.604551e+00	-2.256568e+01	-1.543008e+01	0.000000	0.000000
6e-01	...	-2.283949e-01	-5.423504e-01	-1.618463e-01	-3.545861e-01	-3.171451e-01	-3.269839e-01	-7.083953e-02	-5.295979e-02	5.600000	0.000000
3e-02	...	-2.945017e-02	6.781943e-03	-1.119293e-02	4.097606e-02	1.659350e-02	-5.213911e-02	1.342146e-03	1.124383e-02	22.000000	0.000000
0e-01	...	1.863772e-01	5.285536e-01	1.476421e-01	4.395266e-01	3.507156e-01	2.409522e-01	9.104512e-02	7.827995e-02	77.165000	0.000000
9e+01	...	2.720284e+01	1.050309e+01	2.252841e+01	4.584549e+00	7.519589e+00	3.517346e+00	3.161220e+01	3.384781e+01	25691.160000	1.000000

Here is the Graphical data representation for the Time column in dataset-



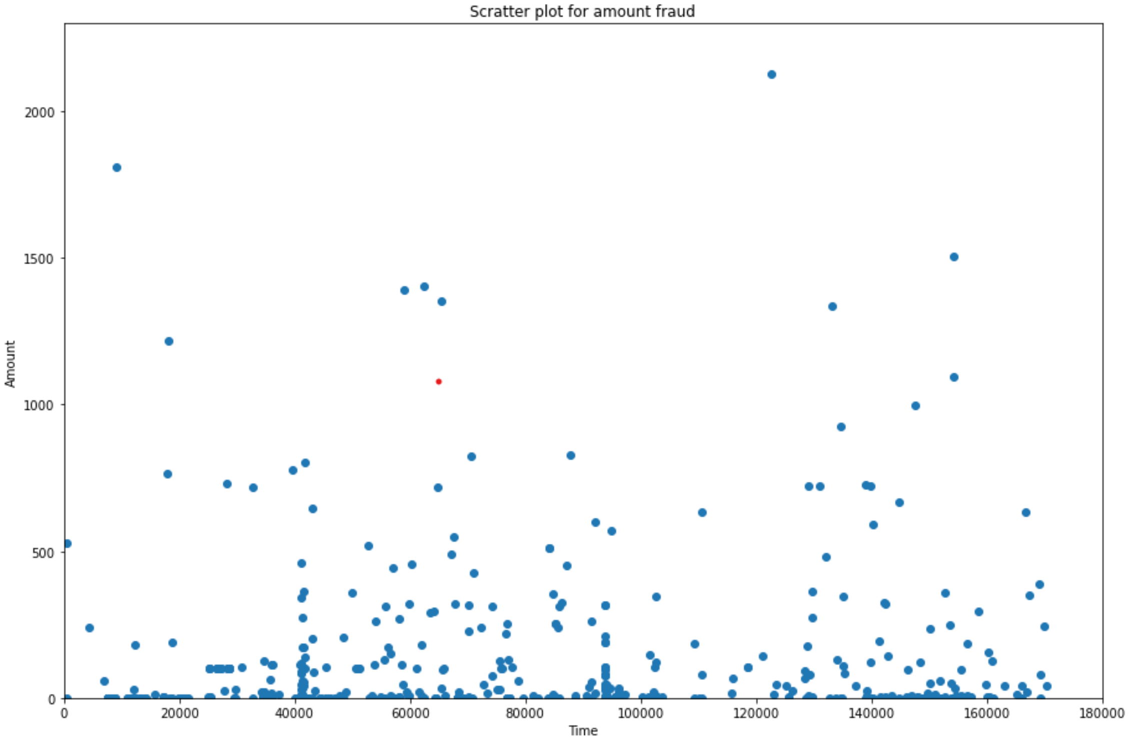
similarly, data graphs for v1 to v28 can be calculated.

MACHINE LEARNING TASKS & EVALUATION CRITERIA:

The main task here is the process of detection of frauds as well as to determine which ones are non-frauds.

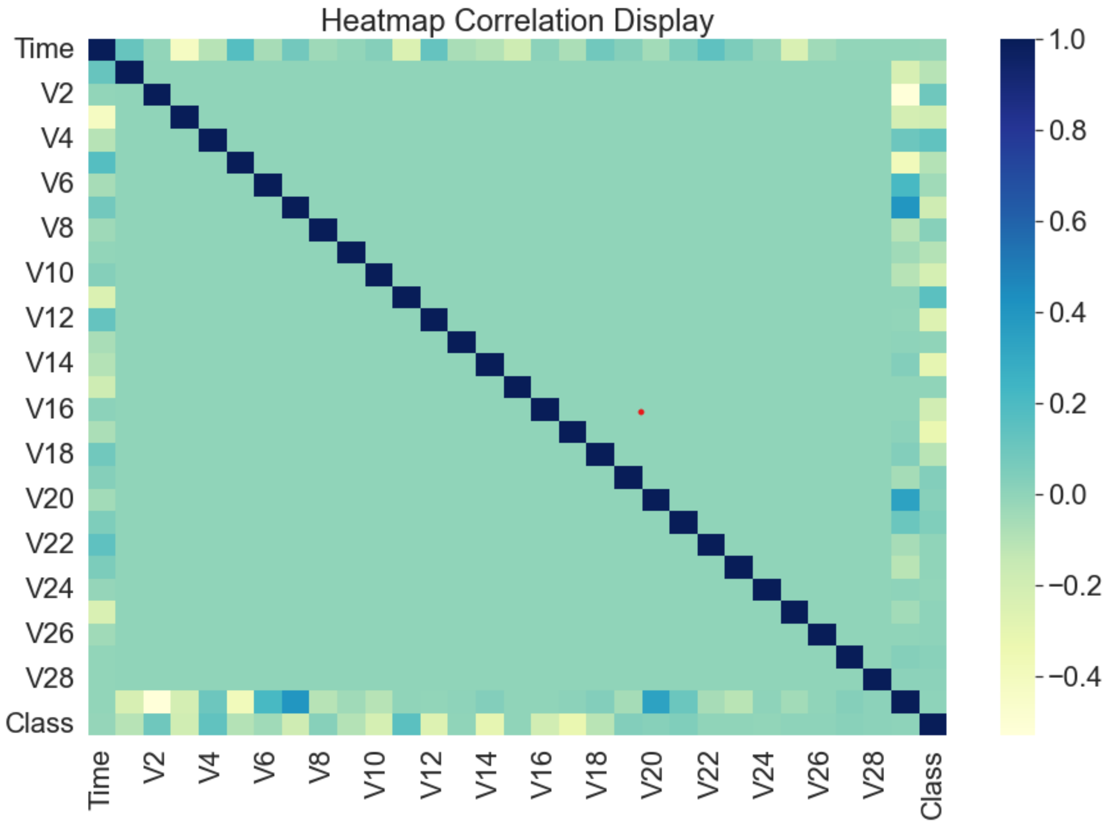
The practice of detecting fraudulent purchase attempts and rejecting them rather than executing the order is the objective of credit card fraud detection.

I have created the Scatter Plot for the amount of frauds so that it gives an idea to see how much fraud is present.



Then I have calculated the data wherein the frauds are more than 1000 amount. Moreover there is very little data that is more than thousand as in general frauds usually are in little amount.

Later I've calculated the accuracy of the classifier, correlation of coefficients. After deriving the heat map, by focusing on the most important dimensions, it is possible to explain the majority of the problem, gaining amount of time while avoiding a significant reduction in accuracy.



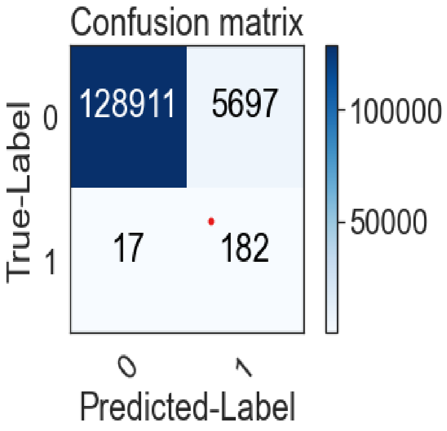
Furthermore, I have created confusion matrix where I have created the True Label and Predicted Label and used SVM model classifier using scikit-learn library.

Next step is to test the model. Hence, I have tested the model by collecting the confusion matrix, then minimize the errors in the svm prediction and then deriving the confusion matrix once again.

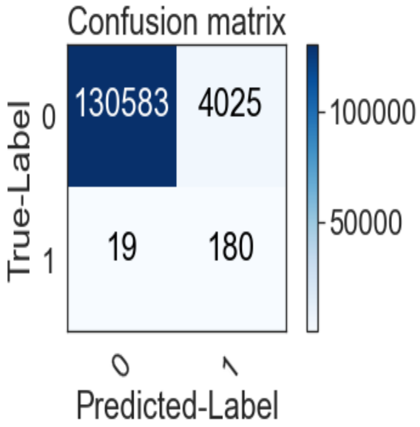
The confusion matrix has errors on the anti-diagonal. However, we can deduce that being incorrect about a fraud transaction is significantly worse than being incorrect about a non-fraud transaction.

As a result, employing accuracy as the sole classification criterion may be regarded unwise. Our criterion will consider precision on the true fraud 4 times more significant than general accuracy for the remainder of this investigation. Despite the fact that the final test result is accurate.

Below is the confusion matrix results-



After testing the model, cm will be-



Indeed, we can choose which class to prioritize during the training phase by adjusting the class weight parameter. The ultimate goal is to lose as little effective fraud as possible.

The class 1 that describes the fraudulent operations will be given precedence over the class 0 in this scenario (non-fraud operation). However, because of the large number of misclassified non-fraud operations, we will prioritize class 0 in this case.

Now the train the data and the again perform the test on it to obtain the final confus ion matrix.

Atlast, the result of probability to detect a fraud I got is `0.8743718592964824`

And Accuracy: `0.9816107472163909`

However, Credit card fraud detection using SVM has the accuracy of 0.99 which similar to the probability 1 across the real time data.

SVM FOR CLASSIFICATION:

Support Vector Machine widely used as classification rather than regression. SVM is a supervised machine learning technique that may be used for both classification and regression. The goal of the SVM algorithm is to find a hyperplane in an N-dimensional space that categorizes data points clearly.

The hyperplane's size is determined by the number of features. If there are only two input characteristics, the hyperplane is just a line. When there are three input features, the hyperplane becomes a two-dimensional plane. When the number of features exceeds three, it becomes more difficult to imagine.

It's a classifier that separates patterns into fraudulent and non-fraudulent. It's great for binary classifications. To construct a learn model, it, like any other artificial intelligence tool, must be trained.

Structural Risk Minimization is the concept where it was derived from.

The decision function of SVM classifier in binary classification is given by:

$$f\ (x) = \mathrm{sgn}(x.w) + b$$

Here, x is the input vector

b is the constant

w is weight

SVM needs either $x.w + b >= 1 – Ei$ or $x.w + b >= -1 + Ei$.

This can be classified by this following equation-

$$yi(w.x + b) >= 1+Ei$$

So, i values are 1,2,3,4…, n.

The optimization problem for w and b will be given by this equation below-

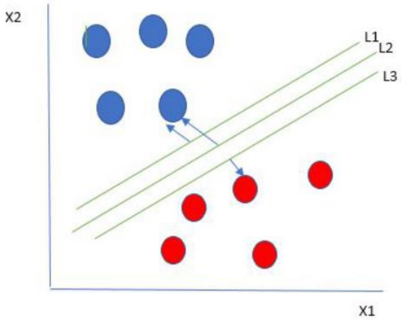
$$(\text{Min } (\tfrac{1}{2})* ||w||\text{\textasciicircum}2) + C \sum Ei$$

Where epsilon values range from i=1 to n.

Decision boundary between 2 classes is derived by this decision function.
We have calculated the criterion used by SVM based on margin maximization of two classes.

Based upon the datapoints, the Best Hyperplane must be calculated which is the largest separation(margin) between two classes.

The cost is 0 if the predicted value and the actual value are of the same sign.



Here x1,x2 are the input feature(variables), Blue and Red circles are nothing but the data points between which the classification is done.

<u>APPLICATIONS:</u>

SVM is used to solve the real time problems like:

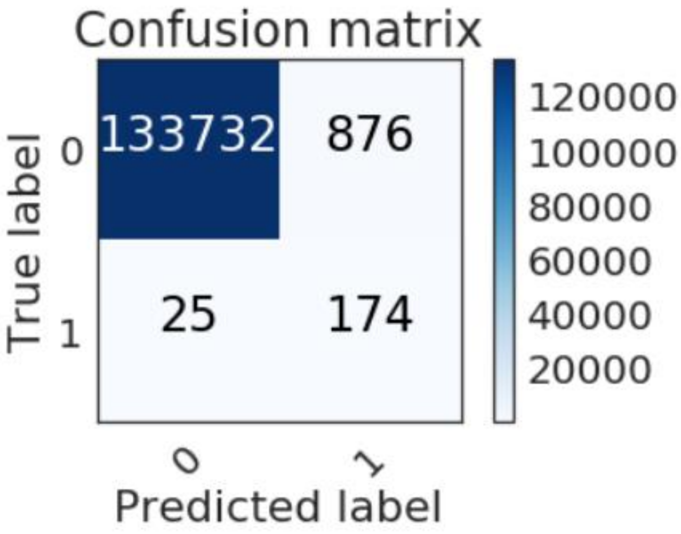
1. In both the standard inductive and transductive situations, SVMs can considerably minimize the need for labeled training instances, making them useful in text and hypertext categorization. Support vector machines are used in several shallow semantic parsing approaches.
2. Image Classification, Face Detection.
3. Neural Networks, Machine Learning.
4. Used in Biological sciences and other sciences.

DRAWBACKS:

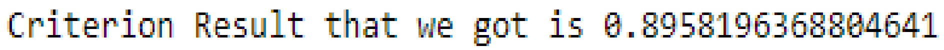
1. Full labelling of input data.
2. Not yet applied on multi-class tasks.
3. Class memberships are not calibrated.

EXPERIMENTAL RESULTS:

Final calculated confusion matrix is shown below-

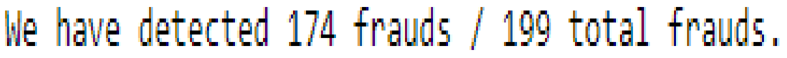


Final Criterion Result that I've computed is-



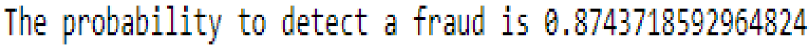
```
Criterion Result that we got is 0.8958196368804641
```

Final total Frauds that I got is-



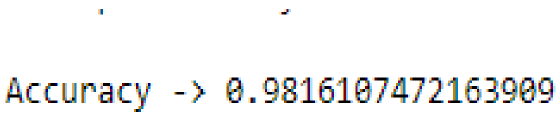
```
We have detected 174 frauds / 199 total frauds.
```

Final Probability of Fraud detection-



```
The probability to detect a fraud is 0.8743718592964824
```

Final accuracy is retrieved as-



```
Accuracy -> 0.9816107472163909
```

FUTURE WORK & CONCLUSION:

In this project, classification is applied using SVM to detect fraud in credit card transactions. This approach gives high accuracy to detect the frauds.

Support Vector Machine has the ability to separate the data in the form of threshold.

It has ability to handle the huge number of transactions with decent results.
In future, the use of SVM with kernel function will be done as a result it works with the low error rate with great accuracy.

REFERENCES:

1. V. Dheepa, R. Dhanapal and D. Remigious, "*A Novel Approach to Credit Card Fraud Detection Model*", *Journal of Computing*, Vol. 2, No. 12, pp. 96, 2010.

2. Tareq Allan and Justin Zhan, "*Towards Fraud Detection Methodologies*", IEEE Proceedings of the Fifth International Conference on Future Technology (Future Tech),

2010.

3. V. Dheepa and R. Dhanapal, “*Analysis of credit card fraud detection systems*”, International Journal of Recent Trends in Engineering, Vol. 2, No. 3, pp. 126-128, 2009.

4. Yashvi Jain, NamrataTiwari, ShripriyaDubey,Sarika Jain, “*A Comparative Analysis of Various Credit Card Fraud Detection Techniques*” IJRTE, Volume-7 Issue-5S2, January 2019.

5. d. l. g. s. chandrahas mishra, “*credit card fraud detection using neural networks*,” international journal of comoputer science, vol. 4, no. 7, July 2017.

6. D. S. G. S.Saranya, “*fraud detection in credit card transaction using bayesian network*,” international research journal of engineering and technology, vol. 4, no. 4, April 2017.

7. T. R. C.Sudha, “*credit card fraud detection in internet using k nearest neighbour algorithm*,” IPASJ international journal of computer science, vol. 5, no. 11, 2017.

8. E. Aji M. Mubarek, “*Multilayer perceptron neural network technique for fraud detection*,” in International Conference on Computer Science and Engineering(UBMK), 2017.

9. V. Dheepa1 and R. Dhanapal, “*Behaviour Based Credit Card Fraud Detection Using Support Vector Machines*”, ICTACT Journal on Soft Computing, vol: 02, issue: 04, July 2012.

10. Xuchen Li, Lei Wang and Eric Sung, “*Ada Boost with SVM–based component classifiers*”, Engineering Applications of Artificial Intelligence, Vol. 21, No. 5, pp. 785-795, 2008.