SUPPORT VECTOR MACHINES, PRESENTED FOR THE PROBLEM OF IDENTIFYING TWO GROUPS OF POINTS ON THE PLANE

Luong Thai Hien[1], Dinh Thi Tam[2]
[1,2] Van Hien University
[1]HienLT@vhu.edu.vn
Received: 17/3/2017; Accepted: 06/6/2017

ABSTRACT

SVM (Support Vector Machine) is a concept in statistics and computer science for a set of supervised learning methods related to each other for classification and regression analysis.

SVM is a binary classification algorithm, Support vector machine (SVM) to build a hyperplane to classify the data set into two separate classes.

A hyperplane is a function similar to the line equation, y = ax + b. In fact, if we need to classify a dataset with only two features, the hyperplane is now a straight line.

In terms of ideas, SVM uses tricks to map the original dataset to more dimensional spaces. Once mapped to a multidimensional space, SVM will review and select the most suitable superlattice to classify that data set.

Keywords: *binary extraction, two-dimensional space, data classification, data clustering, data stratification, identification, SVM (Support Vector Machine).*

TÓM TẮT
Trình bày về Support Vector Machines
cho vấn đề nhận dạng hai nhóm điểm trên mặt phẳng

SVM (Support Vector Machine) là một khái niệm trong thống kê và khoa học máy tính cho một tập hợp các phương pháp học có giám sát liên quan đến nhau để phân loại và phân tích hồi quy.

SVM là một thuật toán phân loại nhị phân, Support vector machine (SVM) xây dựng (learn) một siêu phẳng (hyperplane) để phân lớp (classify) tập dữ liệu thành 2 lớp riêng biệt.

Một siêu phẳng là một hàm tương tự như phương trình đường thẳng, y = ax + b. Trong thực tế, nếu ta cần phân lớp tập dữ liệu chỉ gồm 2 feature, siêu phẳng lúc này chính là một đường thẳng.

Về ý tưởng thì SVM sử dụng thủ thuật để ánh xạ tập dữ liệu ban đầu vào không gian nhiều chiều hơn. Khi đã ánh xạ sang không gian nhiều chiều, SVM sẽ xem xét và chọn ra siêu phẳng phù hợp nhất để phân lớp tập dữ liệu đó.

Từ khóa: *trích rút nhị phân, không gian hai chiều, phân loại dữ liệu, phân cụm dữ liệu, phân lớp dữ liệu, nhận dạng, SVM (Support Vector Machine).*

1. Preamble

Given a training set, expressed in vector space, where each material is a point, this method finds the best Super flat decision that can divide the points in space into two distinct classes. Respectively, class + and class -. The quality of this hyperplane is determined by the distance (called boundary) of the nearest data point of each layer to this plane. Then, the larger the boundary, the better the decision plane, and the more accurate the classification.

The purpose of the SVM method is to find the maximum boundary distance, which is illustrated as follows.

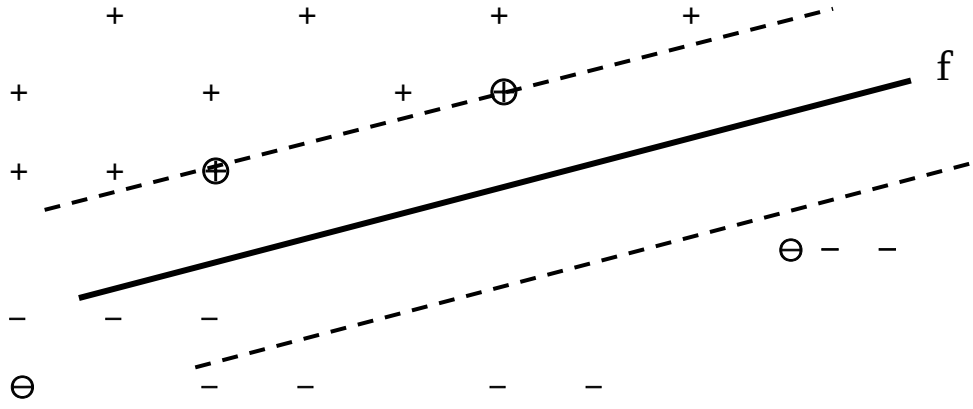

Figure 1: The hyperplane divides the study data into two classes + and - with the largest boundary distance. The closest points (circleed points) are the Support Vector.

2. Research methodology

2.1. Theoretical concept

SVM is actually an optimization problem, the goal of this algorithm is to find a space F and Super flat decision f decide on F such that the smallest classification error. For the set of samples $\{(x_1, y_1), (x_2, y_2),... (x_f, y_f)\}$ with $x_i \in R^n$, belong to the two labels class: $y_i \in \{-1,1\}$ is the corresponding class label of x_i (-1 denotes class I, 1 denotes grade II).

We have, the hyperplane equation contains the vector $\overrightarrow{X_t}$ in space.

$$\overrightarrow{W}.\ \overrightarrow{X_1} + b = 0$$

$$f(\overrightarrow{X_l}) = \text{sign}(\overrightarrow{X_1}.\ \overrightarrow{W} + b) = \begin{cases} +1, & \overrightarrow{X_1}.\ \overrightarrow{W} + b > 0 \\ -1, & \overrightarrow{X_1}.\ \overrightarrow{W} + b < 0 \end{cases}$$

So f($\overrightarrow{X_t}$) Represents the classification of $\overrightarrow{X_t}$ Into two classes as stated. I say $y_i = +1$ if $\overrightarrow{X_t} \in$ Class I and $y_i = -1$ if if $\overrightarrow{X_t} \in$ Class II. Then, To have a super flat f I will have to Solving the following problem:

Find min $\|\overrightarrow{w}\|$ with W Satisfies the following circumstances.

$y_i(\sin(X_i.W + b)) \geq 1$ with với $i \in [1,n]$

The SVM problem can be solved by using the Lagrange operator to transform the equation into an equation. An interesting feature of the SVM is that the decision plane depends only on the Support Vector and that the distance to the decision plane is $1/\|\vec{w}\|$. Even though the other points are deleted, the algorithm still produces the same result as the original. This is the highlight of the SVM method compared to other methods because all the data in the training set is used to optimize the results.

In the case of binary linear separations, classification is performed by the decision function f (x) = sign (<wx> + b), which is obtained by changing the standard vector $\vec{w}$, which is the vector To maximize the functional border. Expanding SVM for multi-layered layering is still being researched. One approach to solve this problem is to build and combine multiple SVM binary classifiers. (For example, during the training with SVM, the class m class problem can be transformed into a differential problem. 2*m class, then in each of the two classes, the deterministic function will be defined for maximal generalizability). In this approach one can refer to two ways: one-to-one, one-to-all.

2.2. Two-layer problem with SVM

The problem is to determine the class function to classify future patterns, ie, with a new data sample xi, it is necessary to determine whether xi is assigned to the +1 or -1 class. To determine the classifier function based on the SVM method, we will proceed to find two parallel hyperplanes such that the distance y between them is the largest possible to separate these two classes into two. The decomposition function corresponds to the hyperplane equation located between the two superimposed flattens.

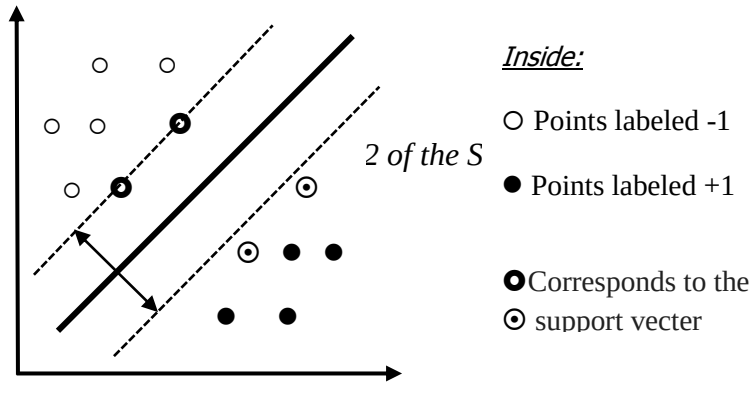

Figure 2: Illustration 2 of the SVM class

Points that lie on two separable hyperbolas are called the Support Vector. These points will determine the decomposition function.

Consider the simplest categorization problem - classify the two subclasses with the training data set consisting of n samples given in the form < $\overrightarrow{x_i}$, $\mathbf{y_i}$ > *i = 1..n.* Inside $\overrightarrow{x_i} \in R^n$ is a vector consisting of m elements containing the value of m attributes or characteristics, and y_i is the classification label that can receive the +1 value (corresponding to the x_i form of the domain of interest) Or -1 (corresponding to the sample x_i not in the field of interest). It is possible to visualize data as points in mucosal Euclidean space and be labeled. SVM is built on the basis of two main ideas.

The first idea: Mapping original data to a new space is called a characteristic space with larger dimensions such that in the new space it is possible to construct a hyperplane that allows the data to be split into two distinct parts, each consisting of Points have the same classification label. The idea of mapping to a feature space is illustrated below.

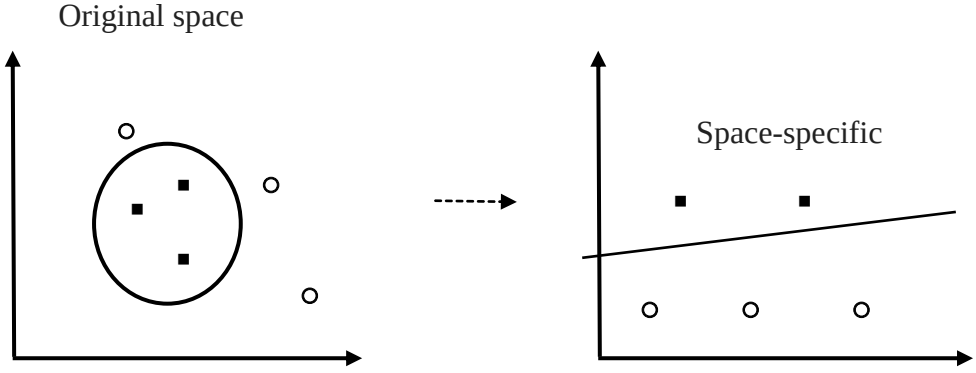

Figure 3: Mapping data from the root space to zero
Characteristic space that allows data to be partitioned by super flat

Second idea: Among such super-flat should choose the largest flat margin with the largest margin. The margin here is the distance from the hyperplane to the nearest points on either side of the hyperplane (each side corresponds to a classification mark). Note that the hyperplane is evenly spaced from the closest points to the different labels. Figure 4 illustrates the hyperplane (solid line) with the maximum margin to data points represented by circles and squares. Original Space Featured Space.

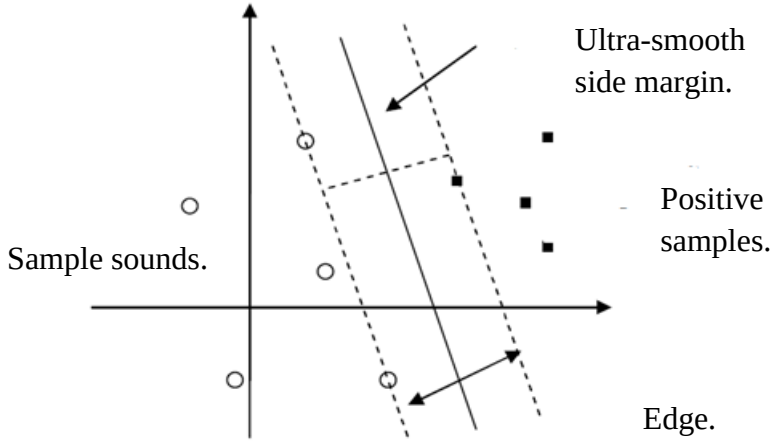

Figure 4: Super flat with maximum margin that allows splitting of squares from circles in the feature space

To avoid direct computation with data in the new space, we use a method called kernel tricks by finding a kernel function K such that:

$$K(\vec{a}, \vec{b}) = <\vec{a}, \vec{b}> \quad (1)$$

Using the Lagrangian multiphysics method and replacing the dot product of the two vectors by the value of the multiplication function by formula (1), the SVM's maximum margin problem is given to the second mathematical planning problem as follows:

Find the coefficient vector $\vec{\alpha} = (\alpha_1, \alpha_2, \dots \alpha_n)$ Allows to minimize the objective function

$$W(\vec{\alpha}) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(\vec{x_i}, \vec{x_j}) + \sum_{i=1}^{n} \alpha_i \quad (2)$$

At the same time satisfy the conditions.

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \quad (3)$$

Và $0 \leq \alpha_i \leq C$ (4)

In ***(2), (3), (4),*** $\vec{x_i}$ and $\mathbf{y_i}$ correspondingly, the data and the classification label of the ith training example, α_i is the coefficient to be determined. In constraint (4), C is the maximum number of data points that are misclassified, ie points located on this side of the hyperplane but labeled on points located on the other side. The use of C allows to correct status of training data Practice has incorrectly labeled examples. After the training is completed, the label value for a new example $\vec{x}$ will be computed by.

$$f(\vec{x}) = sign\left(\sum_{i=1}^{n} y_i \alpha_i K(\vec{x_i}, \vec{x}) + b\right) \quad (5)$$

b is calculated in the following equation.

$$b = y_i - \sum_{j=1}^{n} y_j \alpha_j K(\vec{x_i}, \vec{x_j}) \quad (6)$$

i is a coefficient that satisfies the condition: $0 < \alpha < C$.

3. Simulation

Application of SVM algorithm for handwriting numeral identification.

Algorithm:

Input:

- Blood x;
- Stratification strategy;
- Models have been trained;

Output:

- Label x;

Method

Case Strategy of

Initialization Count[i] = 0; // *i=0,..,N-1*

LoadModel(OVOModel);

for (i=0; i < N-1; i++)

for (j=i+1; j < N; j++)

Count[BinarySVM(x,i,j)]++;

Count[label]=Max(Count[i]);

LoadModel(OVRModel);

label=-1;

for (i=0; i < N; i++)

{

label=BinarySVM(x,i,Rest);

if(label=i) break;

}

EndCase;

Return label;

Inside:

BinarySVM(x,i,j); //Is a function of x in one of two classes i or j

Count[ ] ;// Is a count array to store the class identifier

Demo program:

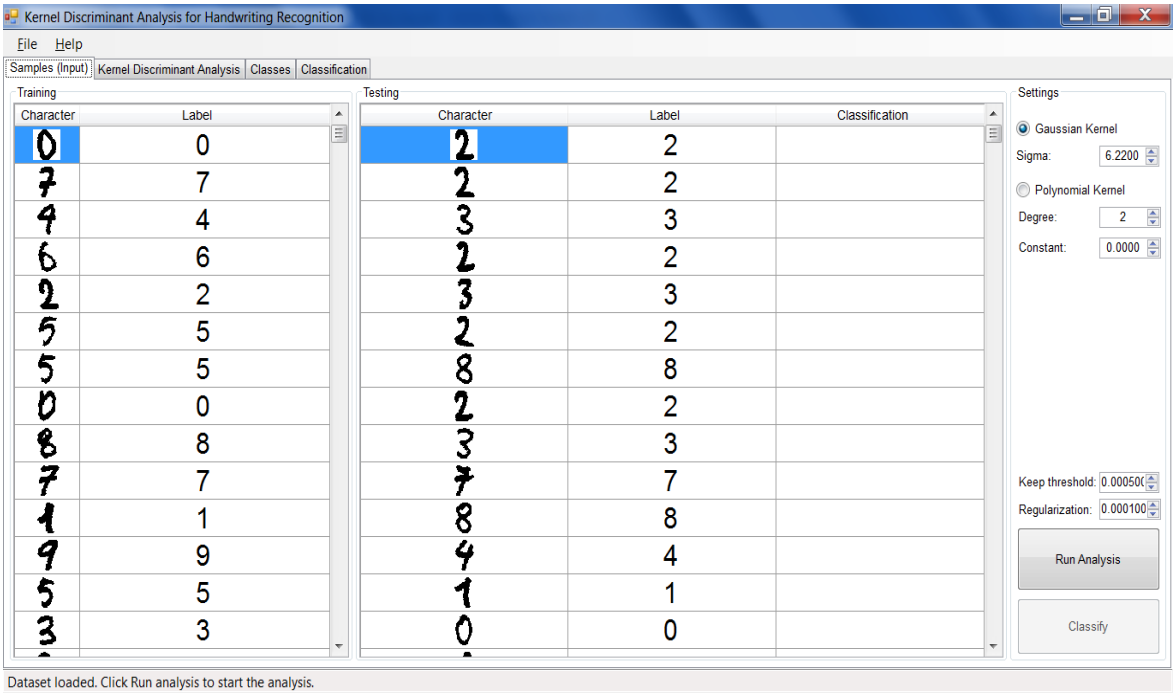

Figure 5: Samples (input)

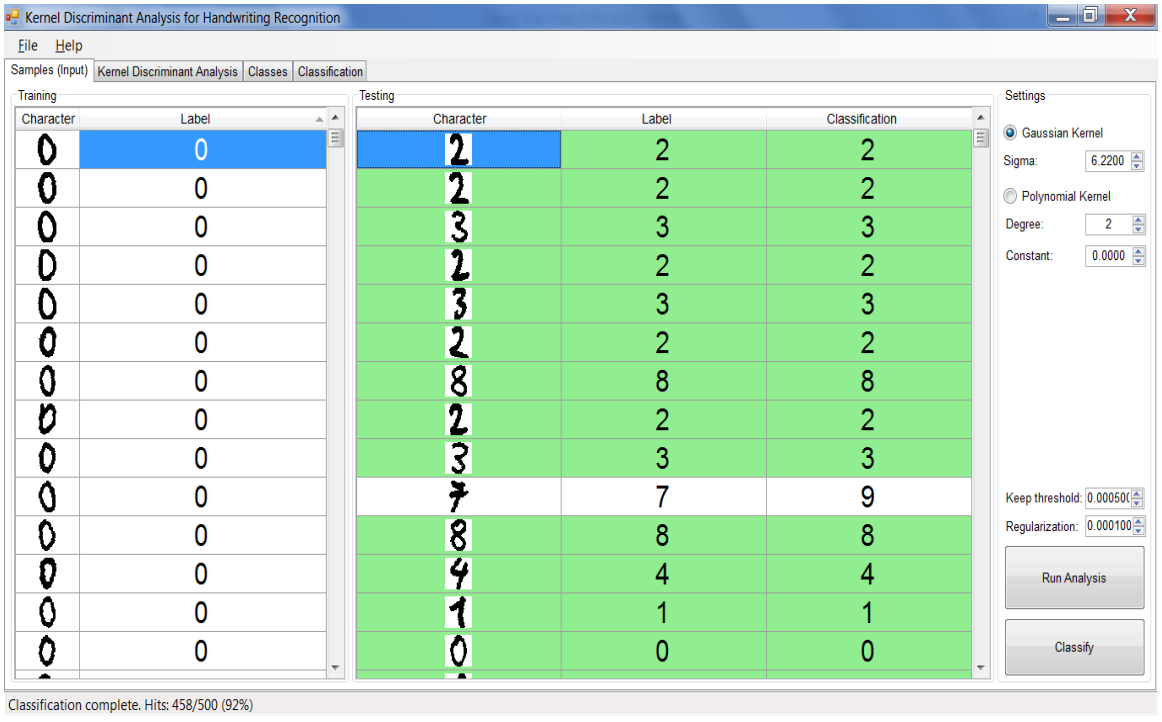

Figure 6: Run Analysis-Classify

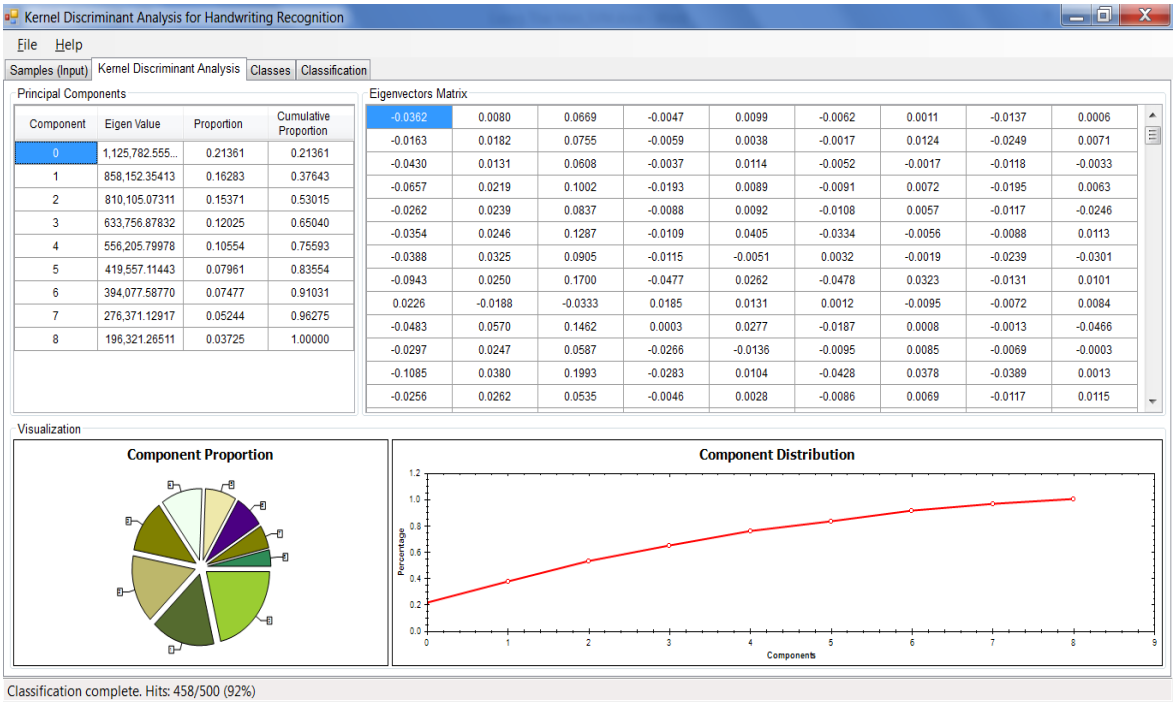

Figure 7: Kenle Descriminant Analysis

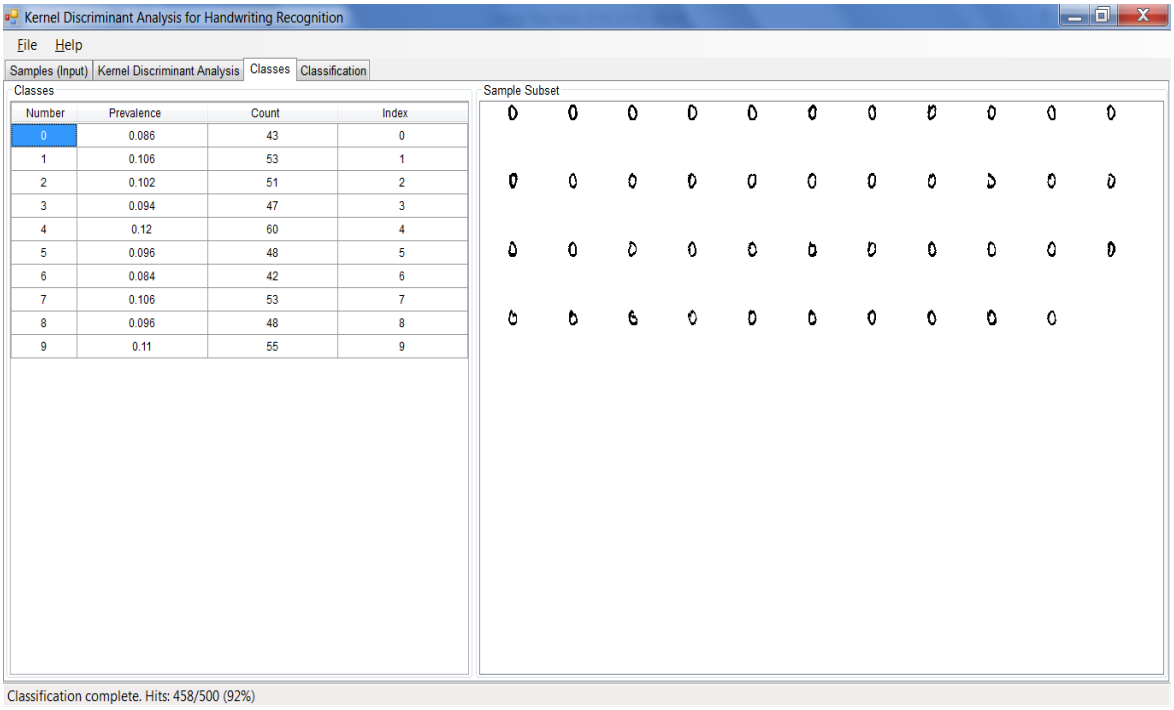

Figure 8: Class

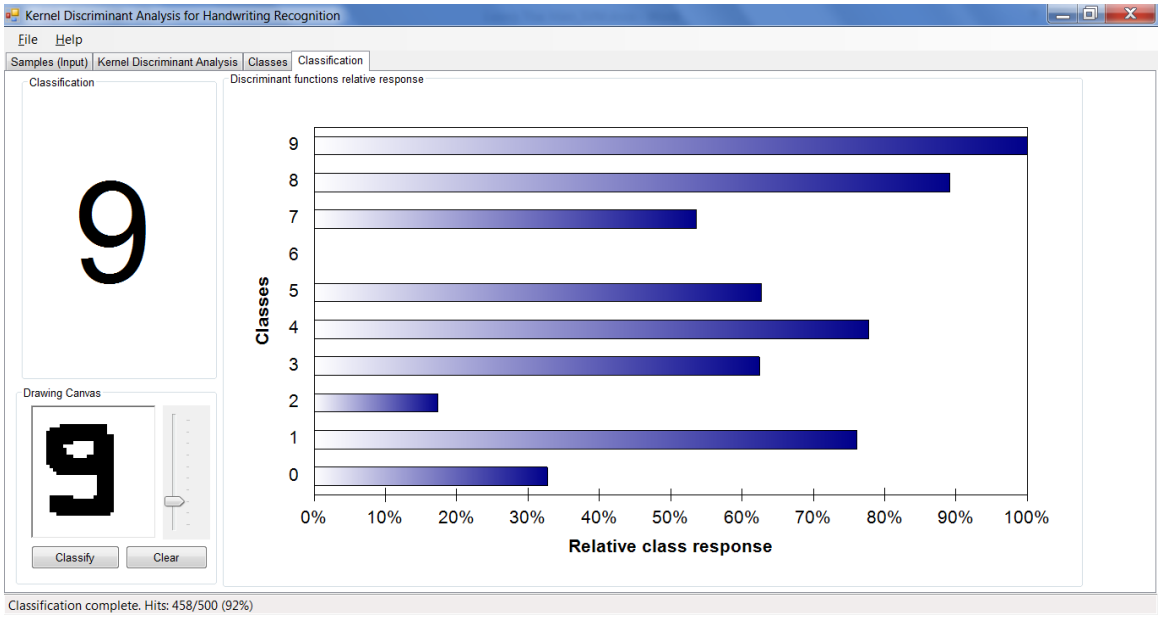

Figure 9: Classification

4. Conclusion

The paper aims to achieve high accuracy of data classification, although all the objectives have been considered, so some issues remain incomplete. However, the article also

achieved some results. Study and present the theoretical basis of the SVM method. This is the most efficient method of classification that has been studied the most recently. Analyze solutions that allow for extensibility and enhancements to improve application performance of SVM.

Binary nature is also a limitation of the SVM, the expansion of the possibility of SVM to solve multi-layer classification problems is not trivial. Neck Many strategies are proposed to extend the SVM to the multi-tier classifier problem. The strengths, weaknesses vary depending on the specific type of data. Until Now, the selection of stratification strategies is usually conducted on a real basis Experience.

REFERENCES

[1] Joachims T., 2009. Text categorization with support vector machines: Learning with many relevant features. Technical Report 23, Universität Dortmund, LS VIII.

[2] Joachims T., 2010. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In International Conference on Machine Learning (ICML).

[3] Kivinen J., Warmuth M., and Auer P., 2011. The perceptron algorithm vs. winnow: Linear vs. logarithmic mistake bounds when few input variables are relevant. In Conference on Computational Learning Theory.

[4] Turk G., O'Brien J.F., 2005. Shape Transformation Using Variational Implicit Functions. Proceedings of ACM SIGGRAPH '05. Los Angeles. California.

[5] Chen D., Bourland H., Thiran J., 2001. Text Identification in Complex Background Using SVM Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2.

[6] Chang Ch., Lin Ch., 2003. LIBSVM: A Library for Support Vector Machines. Department of Computer Science and Information Engineering. National Taiwan University.