A Machine Learning Approach: SVM for Image Classification in CBIR

Sonali Jain
Department of Computer Science, MITM, RGPV, Indore

ABSTRACT
Image classification is technique of finding out most prominent features of image to the query image and with the help of these features classifying image. To get perfect and fast result is the main task of Content based image retrieval (CBIR). Many researchers used different methods to classify the large collection of data. In this paper, we use machine learning technique such as support vector machine. Extract all the relevance and irrelevance features of image like colour, texture, shape, size, mean, standard deviation, histogram value and entropy but exact selection of feature is very difficult task. SVM is used to find out the optimal result. It also evaluates the generalization ability under the limited training samples. It gives faster result as compared to other. In this paper we explain the need of support vector machine and mathematically describes the proposed system i.e. optimal separating hyperplanes and its other cases and at last shows the different kernel function for mapping purpose upon all the data samples.

Keywords: Image classification, CBIR, Machine learning, SVM, Optimal separating hyper planes, Kernel function.

1. INTRODUCTION
SVM was first proposed by Vapnik and is gaining popularity in field of machine learning due to many attractive features and to show practical performance [2]. It gives higher better performance in classification of image than other data classification algorithm. It is mainly used in real world problem like voice recognition, tone recognition, text categories, image classification, object detection, handwritten digital recognition, and data classification [2]. Image classification is the process of grouping of similar types of image into a single unit i.e. called cluster of image. So the classification is a very exciting task to find exact result. To improve the result of classification, extract the related feature of image, because of this we also get good accuracy [8]. In previous image retrieval system have some problem in general applications. First, mostly users want to complete their search in a single iteration generally on the web. Second, it is very time consuming and show lots of negative result with different variety. Third, some iteration introduces many noisy examples in the resulting image [1]. So with the help of CBIR system we get similar image to the query image from the large collection of database and SVM gives optimal result. CBIR mainly used to overcome the problem occur in keyword annotation method i.e. traditional image classification method. The main focus of CBIR is to extract all the features of image based on their visual content to the query image [12].

2. FEATURE EXTRACTION
It is very important step for image classification in CBIR. In this, all the relevance or irrelevance features of image are extracted and on the basis of this classification of image performed. Basically feature extraction is a process of mapping image from image space to feature space. Feature space is a type of input space where similarity measures with the help of kernel function.

In digital image, basically there are many feature like colour, shape, text, size and dimension etc. which are mainly used for feature extraction but extracting those feature which are more relevance to our work in difficult task. The output given by this step is in the form of vector [5] [10] [13].

Basically image features can be divided into 2 parts-

1. Visual Features
2. Semantic Features

Feature which are extracted by human vision are called visual feature.

They are further divided into-

1. General Features
2. Domain specific
General features are those which can be used for searching like colour, shape, texture and feature which are used for particular domain and have knowledge about them [13]. For example, we are searching for face of girl which belongs to human category, so here domain is human. Another one is we are searching for elephant which belong to animal category. These features are domain specific [5].

Some features are semantic which are very difficult to extract. Semantic features are those which have same meaningful information about image. In this category mean value, RGB value, Histogram value, Standard deviation and entropy are belong. These features are not easy to find [5]. So to analyse the set of feature with the help of input data is called feature extraction [8]. Below diagram shows the process of image classification and feature extraction [10].

![Image Classification & Feature Extraction Process](image)

**Figure1** Image Classification & Feature Extraction Process

### 3. SVM BASED FEATURE EXTRACTION

#### 3.1 Why we use SVM?

Previously neural network was used for supervised and unsupervised learning. This gives good result for such type of learning. MLP uses feed forward and recurrent network. Here simple NN shows for simple input-output for learning and another multilayer perception (MLP) shows multiple input and output for universal approximation of continuous nonlinear function [1].

![Simple Neural Network](image)

**Figure2** Simple Neural Network

But there are some facts which come in front of us

1. Local minima
2. How many neurons might be needed for a task?
3. If NN solution come together this may not give unique result [1].

SVM does not hypothesing number of neurons in the middle layer or defining the centre of Gaussian function in RBF. SVM uses an optimal linear separating hyperplanes to separate two set of data in feature space. Optimal hyperplanes is produced by maximum margin between the two set [5].

If we plot the data about X and Y axis and to classify it, we see that there are many hyperplanes which can classify it. But to choose which one is the best or correct solution is very difficult task. For removing this type of problem SVM used [1].

![Image of hyperplanes](image)

**Figure 2** shows various hyper planes which can classify this data but which one is the best? This was the challenges faced.

There are many linear classifiers or hyperplanes that separate the data, but only to choose those which one have maximum margin. Reason behind this if we use a hyperplanes for classification it might be closer to one data set compared to other. So we use the concept of maximum margin classifier or hyperplanes. This one is an apparent solution [1].

Another question can arise in our mind, why we use maximum margin? The main reason to use maximum margin is that

a. If we have done a bit error at boundary location this will gives least chance of misclassification.

b. Other advantages would be avoiding local minima and gives better classification result [1].

So the main focus of SVM, separating data or support vectors with the help of hyperplanes and extend this to non-linear boundaries using kernel trick and choosing those hyperplanes which one have maximum distance. With the help of kernel trick nonlinear boundaries in feature space are formed [1].

### 3.2 Overview of SVM

In machine learning there are two types of methods supervised and unsupervised. Supervised learning based on learn by result and unsupervised based on learn by example. Supervised learning takes input as a set of training data. Support vector machine is a supervised learning technique that analyse data and identify pattern used for classification. It takes a set of input, read it and for each input desired output form [14]. This type of method is known as classification. There are two types of output-If output is discrete then classification performed and if output is continuous than regression performed.

For constructing maximum separating hyperplanes SVM maps input vector to a higher dimension feature space. Feature space refers to an input space which is reserved for measuring similarity with the help of kernel function. It is high dimension space where linear separation becomes very easier than input space [5]. In this, raw data is transformed into a fixed length sample vectors. There are two terms which are used in feature space i.e. called feature values and feature vectors. The characteristics of image is called feature values and these feature values presented the machine in a vectors is known as feature vectors.

Kernel function used in the kernel method performing some operation such as classification, clustering upon different categories of data like text document, sequence, vectors, set of points, image and graphs etc. it maps the input data into a higher dimension feature space because in this data could be easily separated or better structured [5].

There are some points in the feature space which are separated by some distance is called support vectors. It is the point between origin and that point and define the position of the separator. The distance from the decision surface to the closest data point determines the margin the classifier.
The main feature of SVM is to construct a hyperplane or a set of hyperplanes with the help of support vectors in a higher dimension space [15]. These are mainly used for classification. It separate the space into two half space. A ‘good separation’ is achieved by hyperplanes that has the largest distance to the nearest data points. Here good separation means larger the separation between two hyperplanes gives lower generalization error. That’s by it is called maximum margin classifier [2], [5], [6], [8]. Two parallel hyperplanes are constructed on each side of the hyperplanes that separate the data. If geometric gap between hyperplanes high than classification error is low [2], [5].

3.3 Mathematically describe the SVM

There are several cases in this approach which are given below [5], [6].

3.3.1 Optimal separating hyperplanes

A brief description of SVM algorithm is given as follows. Consider a given set of points in the form of training data which separate two classes of pattern based on given training set [3]:

\[ S = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots (x_n, y_n)\} \]

Where \( x_i \) is a p-dimension real vectors, \( y_i = \{-1, +1\} \) and n is a number of sample.

According to Vapnik’s formula

\[ y_i = <w, x_i> + b \]  \[ (1) \]

\[ = \sum \alpha_i y_i x_i + b \]

Where \( w \) is a p-dim vector and b is constant or scalar. By adding a scalar value b it increases the margin between hyperplanes and in the absence of b hyperplanes is forced to pass through the origin. So in SVM we always use parallel hyperplanes which maintain distance between them. Parallel hyperplanes can be described by equation [5]:

\[ w \cdot x_i + b = 1 \quad \text{for} \quad y_i = 1 \]

\[ w \cdot x_i + b = -1 \quad \text{for} \quad y_i = -1 \]
In the case of linearly separable, we can select these hyperplanes so that there are no points between them and then make an effort to maximize their distance [5]. As we know in feature space there are number of hyperplanes but choose the one for which the distance to the closest point is maximal is called optimal separating hyperplanes (OSH) [7]. Since the distance to the closest point is 1/||w||. After subtracting the two distances we get the summed distance from separating hyperplanes to the nearest points [1].

Maximum Margin \( M = \frac{2}{||w||} \)

The quantity \( \frac{2}{||w||} \) is called the margin and it is used for measure the generalization ability: the larger the margin, the better the find generalization error [7].

To finding optimal separating hyperplanes with maximum margin should minimize \( ||w||^2 \). This can be achieved with the help of Lagrangian multipliers. This optimization problem can be converted into dual form called quadratic Equation problem (QP).

Quadratic Equation problem (QP)-

It is the process of minimizing \( ||q|| \) but it makes the optimization process difficult [8].

To Minimize (in \( w, b \)) [2] [4] [8]

\[
1/2 ||x||^2
\]

Subject to (for \( i = 1, 2 \ldots \ldots n \))

\[
Y_i (w.x - b) \geq 1 \]

\[
(2)
\]

With the help of this we can get maximum separation between hyperplanes and then the Lagrangian is as follows-

\[
L (w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \alpha_i (y_i (w^T x_i + b) - 1) \]

\[
(3)
\]

Where \( \alpha_i \) is a Langrangian multipliers. If we denote by \( \alpha = (\alpha_1, \alpha_N) \) and \( N \) non negative Lagrangian multipliers associated with constraints, our optimization problem amount maximizing [4][6], [8].

\[
W (\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j
\]

\[
(4)
\]

Then partially differentiate eq. (3) with respect to \( b \) saddle point \( (w_0, b_0, \alpha_0) \) [2]

\[
\frac{\delta L}{\delta b} = 0
\]

i.e. \( w_0 = \sum \alpha_0 y_i x_i \)

\[
(5)
\]

And \( \frac{\delta L}{\delta \alpha_i} = 0 \)

\[
\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{for all } i
\]

\[
(6)
\]

The Langrangian has to be minimized with respect to \( \alpha_i \).

Substituting (5) and (6) in (3) and get primal form into dual form [2].

\[
L_\alpha (\alpha) = \sum \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j
\]

\[
(7)
\]

If we find optimal hyperplanes dual Lagrangian \( L_\alpha \) has to be maximized with respect to \( \alpha_i \). Here dual Lagrangian is expressed in terms of training data and depend on scalar product of input patterns \( (X_i^T X_j) \) [2].

And the decision function become [9]

\[
f (x) = \text{sgn} \left( \sum_{i=1}^{N} \alpha_i y_i x_i \cdot x + b \right)
\]

3.3.2 Linearly non-separable case

When the data is not linearly separable, we introduce a new variable i.e. called slack variable[7]. It is denoted by \( \xi \). If we take variables \( \xi_1, \xi_2, \ldots \ldots, \xi_N \) with \( \xi_i \geq 0 \) such that

\[
Y_i (w.x + b) + \xi_i = 1 - \xi_i, \text{ i=1, \ldots, \ldots, N}
\]

The main reason to use the slack variable is to allow misclassified points, which have their corresponding \( \xi > 1 \).
The solution of the following problem [3]:
Minimize: \( \frac{1}{2} w \cdot w + C \sum_{i=1}^{N} \xi_i \)

The first term is minimized to control the learning capacity and second term is to control the number of misclassified points. The parameter C is chosen by the user. If the value of C is high then error rate is to be high [7].

### 3.3.3 Nonlinear support vector machines

In the nonlinear case, we mapped the data into other space with the help of some nonlinear mapping function. This mapping function is called kernel function and space is called Euclidean distance [4]. If we replace \( x \) by its mapping in the feature space \( \phi(x) \)

\[ x = (x_1, x_2, \ldots, x_n) \rightarrow \phi(x) = (\phi(x_1), \ldots, \phi(x_n)) \]

Then the mapping function is, Eq. (4) becomes

\[ w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \]

In the form of dot product, so we use kernel function \( K(x, y) = \phi(x), \phi(y) \). With the help of Mercer’s theorem, we define kernel function \( K(x, y) \). Mercer’s theorem states that there exist a mapping function \( \phi \) such that \( K(x, y) = \phi(x), \phi(y) \) [2], [7].

If a kernel function \( K \) satisfying Mercer’s condition, the training algorithm consists of minimizing

\[ w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \]

And the decision function can be written as [9]:

\[ f(x) = \text{sgn} (\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b) \]

### 4. KERNEL SELECTION OF SVM

In this mapping is done with the help of training vector. Here training vectors are mapped into higher dimensional space by function \( \phi \). Then we get linear separating hyperplanes with the maximum gap. There are many kernel functions in SVM which are used for mapping [2], [4], [6], [11].

Local kernel - In this category only those data are comes that are close or some proximity with each other means which have some distance functions. Examples of local kernel are:

- a) RBF kernel: \( K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \) \( \gamma > 0 \)
- b) KMOD: \( K(x_i, x_j) = \exp(1/1+\|x_i - x_j\|^2) - 1 \)
- c) Inverse Multiquadric: \( K(x_i, x_j) = 1/\sqrt{\|x_i - x_j\|^2 + 1} \)

Global kernel - Data which are far from each other are called global function and all the functions based on dot product. Examples of global kernel are [3]

- a) Linear kernel: \( K(x_i, x_j) = x_i^T x_j \)
- b) Polynomial kernel: \( K(x_i, x_j) = (\gamma x_i^T x_j + r)^d \) \( \gamma > 0 \)
- c) Sigmoid kernel: \( K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \)

Here \( \gamma, r \) and \( d \) are kernel parameters.
5. CONCLUSION

The main focus of proposed work was to show the comparative study of neural network and support vector machine. Here we use SVM as a classifier for the classification of image and apply this classification process to all the features of image which are extract from feature extraction step. It is mainly used to find maximum margin hyperplanes in a high dimensional feature space. Here we also explain all the conditions such as optimal separating hyper planes, linearly non-separable case and nonlinear support vector machines in which SVM work. A kernel based learning method used for the mapping purpose. So with the help of support vector machine we get much better performance than the other traditional method and get optimal result.

Reference


[7] Olivier Chapelle, Patrick Haffner and Vladimir Vapnik, “SVMs for Histogram-Based Image Classification”.


