Crop Coverage data Classification using Support Vector Machine

By Tarun Rao, N. Rajasekhar & N C Naveen  
Dayananda Sagar College of Engineering

Abstract- A statistical tool which can be used in various applications ranging from medical science to agricultural science is support vector machines. The proposed methodology used is support vector machine and it is used to classify a raster map. The dataset used herein is of Gujarat state agriculture map. The proposed approach is used to classify raster map into groups based on crop coverage of various crops. One group represents rice crop coverage and the other millets crop coverage and yet another that of cotton crop coverage. Various statistical parameters are used to measure the efficacy of the proposed methodology employed.

Keywords: mining, SVM, supervised classification.

GJCST-C Classification: C.1.2
Crop Coverage data Classification using Support Vector Machine

Taran Rao\textsuperscript{a}, N. Rajasekhar\textsuperscript{b} & N C Naveen\textsuperscript{c}

Abstract: A statistical tool which can be used in various applications ranging from medical science to agricultural science is Support vector machine (SVM). The proposed methodology used is SVM and it is used to classify a raster map. The dataset used herein is of Gujarat state agriculture map. The proposed approach is used to classify the raster map into groups based on crop coverage of various crops. One group represents rice crop coverage and the other millets crop coverage and yet another that of cotton crop coverage. Various statistical parameters are used to measure the efficacy of the proposed methodology employed. Keywords: Mining, SVM, supervised classification.

I. INTRODUCTION

Crop mapping is widely used in agriculture and remote sensing science. Crop mapping using classification methodologies serves various applications in agricultural science like gauging water and soil demand etc. For such applications information on the spatial distribution of classification error is of particular interest. Recent progress in Information Technology systems, lead to dynamic communication among people of every profession. Information technology systems have changed the way people meet and communicate. There is an increasing tendency of professionals and experts in the agriculture sector to communicate best practices in the field of agriculture via the medium of internet. Farmers who use the medium of internet get benefited from the various forums used therein to communicate advanced crop yield technologies. Crop mapping can also facilitate the farmers in planning their crop management in advance and they do not see internet and modern technologies has a hurdle.

Data is everywhere, abundant, continuous, increasing and heterogeneous. Extracting meaningful information from that data is useful but very difficult: rich data but poor information is a common phenomenon in the world. Data mining (DM) refers to extracting or mining useful knowledge from large amounts of data. One of the various phases of data mining is classification.

Classification is the process in which available data items are categorized into two or more categories depending on the various criterions. Methodologies in which the class label is known a priori is called supervised classification and those in which class labels are not known a priori are called unsupervised classification or clustering. Supervised classification can be further categorized as parametric and non-parametric categories. Based on whether or not the approach is based on probability distribution or density functions.

A well-known statistical method that can be used to solve optimization problems is Support Vector machines (SVM). The proposed methodology used here is SVM. The data items can be represented as feature vectors in a hyperplane and a line passing through the hyperplane can be used to categorize the data items into two different categories. The line can be considered a naive form of SVM. The advantage of SVM as a classification method is that in has feature extraction method inbuilt in its architecture. SVM is better compared to other existing classification methodologies like Naïve bayes, Artificial neural network, decision tree based classification etc. depending on previous research.

SVM which is inherently linear in nature. However, by using kernel function it can be extended to non-linear space as well. In either of the approaches SVM takes a lot of time to classify the data items. SVM approach is used to solve a multi-class classification problem in this research work. Its finds a suitable line which is far off from all equidistant points in the hyperplane.

SVM has numerous applications as inland analysis, species mapping, medicine, error identification, text and speech analysis, signal analysis etc. SVM is used in this research to classify raster TIFF datasets. Subsequent section explains about Literature Survey on SVM. Later Proposed methodology is explained followed by result analysis. The final section deals with conclusion followed by references.

II. LITERATURE SURVEY

a) Introduction to SVM

SVM is a promising methodology which is used in various applications. It solves both two class and multi-class problems. Problems in which input data items need to be categorized into two categories are called two class problems and the ones in which data items need to be categorized into multiple classes are called multi-class problems. The multi-class classification problem can be solved using divide and...
conquer approach. In this approach the problem can be divided into many two-class problems and in the future the results can be merged to arrive at the final solution to the problem.

b) SVM has two major features

Margin maximization: The classification boundary functions of SVMs maximize the margins, which leads to maximizing generalization performance [18].

SVM can be categorized as linear and non-linear SVM as in Fig 1. In linear SVM the hyper plane categorized under two different class labels by a line passing through the hyperplane[18][19][20].

![Fig.1: Linear SVM](image)

The line representing the SVM can be denoted by (1)[21]:

\[
m\mathbf{\theta}i + c = \pm 1
\]

Data items can be represented by (2)[22]:

\[
f(x) = \text{sign}(mc + b)
\]

Where \(\text{sign()}\) represents sign function, \(m\) denotes slope and \(\theta\) happens to be the angle. Sign function is denoted by:

\[
\text{sign}(c) = \begin{cases} 
1 & \text{if } c > 0 \\
0 & \text{if } c = 0 \\
-1 & \text{if } c < 0
\end{cases}
\]

Numerous lines might be able to split the planeas different categories but the one that maximizes the distance between itself and the data items in the two categories is known as the support vector as denoted in Figure 2.

![Fig. 2 : Distance amid data items in a feature space](image)

The above distance can be denoted as:

\[
M = \frac{(\mathbf{\theta}i - \mathbf{\theta}j).m}{|m|} = \frac{2}{|m|}
\]

\[
h(m) = \frac{1}{2}m^t m
\]

subject to \(y(m\theta + b) \geq 1, \forall i\)

The solution can be denoted with the help of a Lagrange multiplier \(\alpha\) as:

\[
m = \sum \alpha_i y_i \theta_i
\]

\[
b = y_k - m^t x_k \text{ for any } x_k \text{ such that Lagrange multiplier } \alpha_k \neq 0
\]

Classifier representation[16]:

\[
f(\theta) = \sum \alpha_i y_i \theta_i x + b
\]

Systematic nonlinear classification via kernel tricks: SVMs effectively handle non-linear classifications using kernel tricks.

To improve the efficiency of the solution the input data item space can be mapped to a higher dimensional feature space denoted by [18]:

\[
K(\theta_i, \theta_j) = f(\theta_i)^t f(\theta_j)
\]

The above representation is also known as a kernel function and can be denoted as [23]:

Linear Kernel function = \(\theta_i^t \theta_j\)

Polynomial kernel function = \((1 + \theta_i^t \theta_j)^p\)

Gaussian kernel = \(\exp\left(-\frac{|\theta_i - \theta_j|^2}{\sigma^2}\right)\)

Sigmoid kernel = \(\tanh(\omega_0 \theta_i^t \theta_j + \omega_1)\)

(9)

c) Multi-class SVM

Multi class SVM can be categorized as one-versus-all, one-versus-one, and k-class SVM’s[18].

i. One-versus-all support vector machines

In this approach SVM classifiers are constructed which separate one class from remaining patterns[18].

ii. One-versus-one support vector machines

In this approach k different SVM classifiers are constructed for every pair of classes [18].

iii. k-Class support vector machines

In this approach K binary classifiers are built concurrently [18].

III. PROPOSED METHODOLOGY

a) Datasets used

A TIFF data set is used in this research and SVM is used to classify the said data set[24]. The TIFF data set is a Gujarat map which has crop coverage data across the state for rice, cotton and millet.

b) Proposed Approach

The TIFF dataset is initially pre-processed. [25]. Later Region Of Interest (ROI) is created from the image. In the next stage training set samples are selected from the ROI. Each of these training set samples correspond to a particular crop coverage in Gujarat map data set used. Three crop coverages are used for performing the said classification. They are rice, cotton and millet crop coverage’s. After the training data sample are collected the SVM classification methodology is used as explained[26]:
Begin

Step 1: Extract features from the data sets

Step 2: Select feature vectors and form the input data set

Step 2: Start dividing the input data set into two sets of data corresponding to two different categories

Step 3: If a data item is not assigned any of the regions mentioned then add it to set of support vectors V

Step 4: end loop

End

Finally the built model is validated against the test data set. Herein the test data set under consideration is the crop coverage area that is not covered as part of the selected training data set sample. One of the key steps involved in the classification process is feature extraction as mentioned below:

Energy (E): It facilitates in computing homogeneity in the data set and is denoted by:

$$E = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} (p(i,j))^2$$  \hspace{1cm} (9)

Contrast (C): Contrast helps identify local data set variation and is denoted as:

$$C = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} (i - j)^2 p(i,j)$$  \hspace{1cm} (10)

Inverse difference moment (IDM): Local texture alterations can be located using:

$$IDM = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} \frac{1}{1+(i-j)^2} p(i,j)$$  \hspace{1cm} (11)

Entropy (S): The data set complexity can be computed by:

$$S = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} p(i,j) \log(p(i,j))$$  \hspace{1cm} (12)

Where μ_k and mxn are the mean and size of the block B_k

Spatial Frequency (SF): Frequency changes in the data set can be computed using:

$$SF = (RF)^2 + (CF)^2$$

Where

$$RF = \left( \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=2}^{n} [I(i,j) - I(i,j-1)]^2 \right)^{1/2}$$

and

$$CF = \left[ \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=2}^{n} [I(i,j) - I(i-1,j)]^2 \right]^{1/2}$$

Variance (V): Level of focus in a data set can be computed using:

$$V = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} (I(i,j) - \mu)^2$$  \hspace{1cm} (16)

Where μ is the mean value of the block image and mxn is the image size

Energy of Gradient (EOG): Measure of focus can also be computed using:

$$EOG = \sum_{i=1}^{m-1} \sum_{j=1}^{n-1} (f_i^2 + f_j^2)$$  \hspace{1cm} (17)

Where, f_i = f(i+1, j) - f(i, j)

f_j = f(i, j+1) - f(i, j)

IV. Result Analysis

a) Environment Setting

Agricultural map of Gujarat was used as a dataset to perform the said classification. A region of interest (ROI) was extracted from the map that acted as a training data and it was validated against the complete data segment pertaining to a particular crop in the map. Environment in which the research was undertaken is shown in Table 1[27].

Table 1: Environment Setting

<table>
<thead>
<tr>
<th>Item</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>Intel CPU @2 GHz processor</td>
</tr>
<tr>
<td>Memory/OS</td>
<td>4GB/Win 7</td>
</tr>
<tr>
<td>Applications</td>
<td>Monteverdi</td>
</tr>
</tbody>
</table>

b) Result Analysis

The ratio of correctly classified and uncorrectly classified data items can be represented using confusion matrix view as mentioned in Table 2. It helps measure the efficacy of the performed classification. Classification results is given in Figure 4.

Table 2: Confusion Matrix

<table>
<thead>
<tr>
<th>Classification result</th>
<th>No Event</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Event</td>
<td>True Negative (TN)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td>Event</td>
<td>False Negative (FN)</td>
<td>True Positive (TP)</td>
</tr>
</tbody>
</table>

(a)  

Global Journal of Computer Science and Technology (C) Volume XVI Issue III Version I Year 2016
Fig. 3: (a) ROI from the TIFF data set. (b) Classified image with various crop coverage in the state of Gujarat displayed in various colors (Rice-Brown, Millets-Violet, Cotton-Brown). (c) Edge Feature extracted image of the crop data set.

Accuracy and kappa statistics are used to measure the efficacy of the classification methodology used. These parameters are denoted by equations (18) and (19)[28][29][30]:

\[
\text{Accuracy} = \frac{TP + TN}{(TP + FN + FP + TN)} \times 100
\]  

(18)

\[
\text{Kappa statistics} = \text{Sensitivity} + \text{Specificity} - 1
\]  

(19)

Confusion matrix in research is mentioned in Table 3.

Table 3: Confusion Matrix

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Reference</th>
<th>Rice</th>
<th>Millets</th>
<th>Cotton</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rice</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Millets</td>
<td>0</td>
<td>16</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Cotton</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

V. Conclusion

SVM classification methodology is used to classify the Gujarat map TIFF data set. Accuracy and kappa statistics parameters are used to measure the efficacy of the said method and the values obtained for the said evaluation parameters prove beyond doubt that the method used classifies the data set with better accuracy.

References Références Referencias


27. David J. Rogers, Jonathan E. Suk, Jan C. Semenza, Using global maps to predict the risk of dengue in Europe, ActaTropica, Volume 129, January 2014, Pages 1-14, ISSN 0001-706X.