Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 14 Issue 1 Version 1.0 Year 2014
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Supervised Classification of Remote Sensed data Using Support Vector Machine

By Tarun Rao & T.V. Rajinikanth

Acharya Nagarjuna University, India

Abstract- Support vector machines have been used as a classification method in various domains including and not restricted to species distribution and land cover detection. Support vector machines offer many key advantages like its capacity to handle huge feature spaces and its flexibility in selecting a similarity function. In this paper the support vector machine classification method is applied to remote sensed data. Two different formats of remote sensed data is considered for the same. The first format is a comma separated value format wherein a classification model is developed to predict whether a specific bird species belongs to Darjeeling area or any other region. The second format used is raster format which contains image of Andhra Pradesh state in India.

Keywords: classification, data mining, support vector machine, remote sensed data.

GJCST-C Classification: G.4

Supervised Classification of Remote Sensed Data using Support Vector Machine

Tarun Rao [α] & T.V. Rajinikanth [σ]

Abstract- Support vector machines have been used as a classification method in various domains including and not restricted to species distribution and land cover detection. Support vector machines offer many key advantages like its capacity to handle huge feature spaces and its flexibility in selecting a similarity function. In this paper the support vector machine classification method is applied to remote sensed data. Two different formats of remote sensed data is considered for the same. The first format is a comma separated value format wherein a classification model is developed to predict whether a specific bird species belongs to Darjeeling area or any other region. The second format used is raster format which contains image of Andhra Pradesh state in India. Support vector machine classification method is used herein to classify the raster image into categories. One category represents land and the other water wherein green color is used to represent land and light blue color is used to represent water. Later the classifier is evaluated using kappa statistics and accuracy parameters.

Keywords: classification, data mining, support vector machine, remote sensed data.

I. Introduction

Data mining is the process of extracting useful information from various data repositories wherein data might be present in different formats in heterogeneous environments[1][2]. Various methods like classification, association, clustering, regression, characterization, outlier analysis can be used to mine the necessary information. In this paper we shall be focusing on classification.

Classification is the process wherein a class label is assigned to unlabeled data vectors. Clas-sification can be further categorized as supervised and unsupervised classification. In supervised classify-cation the class labels or categories into which the data sets need to be classified into is known in advance. In unsu-pervised classification the class label is not known in advance[3]. Unsupervised classification is also known as clustering. Supervised classification can be subdivided into non-parametric and parametric classification. Parametric classifier method is dependent on the pro-bability distribution of each class. Non parametric cla-ssifiers are used when the density function is not known[4].

One of the very prominent parametric supervised classification methods is support vector machines(SVM). In this paper SVM are used to perform the said classification. Herein the data vectors are represented in a feature space. Later a hyperplane that geometrically resembles a slope line is constructed in the feature space which divides the space comprising of data vectors into two regions such that the data items get classified under two different class labels corresponding to the two differrent regions[5]. It helps in solving equally two class and multi class classification problem[6][7]. The aim of the said hyper plane is to maximize its distance from the adjoining data points in the two regions. Moreover, SVM's do not have an additional overhead of feature extraction since it is part of its own architecture. Latest research has proved that SVM classifiers provide better classification results when one uses spatial data sets as compared to other classification algorithms like Bayesian method, neural networks and k-nearest neighbors classification methods[8][9].

SVM have been used to classify data in various domains like land cover classification[10], species distribution[11], medical binary classification[9], fault diag-nosis[12], character classification[5], speech recog-nition[13], radar signal processing[14], habitat prediction etc... In this paper SVM is used to classify remote sensed data sets. Two formats of remote sensed data viz. raster format and comma separated value(CSV) file formats have been used for performing the said classification using SVM.

Our next section describes Background Knowledge about SVM classifiers. In section 3 materials and methods viz. data acquired and the proposed methodology have been discussed. Performance analysis is discussed in Section 4. Section 5 concludes this work and later acknowledgement is given to the data source followed by references.

II. Background Knowledge

a) Overview of SVM Classifier

Support vector machine (SVM) is a powerful tool used in solving either two class or multi class classification problem[15][16]. In a two class problem the input data has to be separated into two different categories wherein each category is assigned a unique class label[17]. A multi class classification problem can be solved by dividing it into multiple two class classification problems and later aggregating the individual results to get the final result of the multi class problem.

Author α: Acharya Nagarjuna University, Guntur, India. e-mail: tarun636@gmail.com
Author σ: Sreenidhi Institute of Science and Technology, Hyderabad, India. e-mail: rajinitv@gmail.com

SVM can be categorized into non-linear and linear SVM. Data can be represented in space as shown in Fig 1. Linear SVM can be geometrically represented by a line which divides the data space into two different regions thus resulting in classifying the said data which can be assigned two class labels corresponding to the two regions[18][19][20].

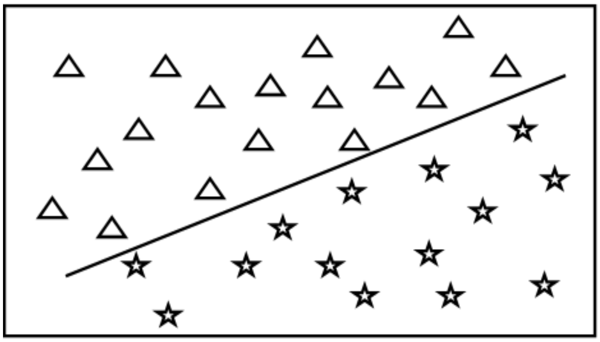


Figure 1 : The Hyperplane

The line mentioned herein is called a hyperplane and can be mathematically represented by equation (1)[21]:

$$\begin{aligned} mx_i + b &>= +1 \\ mx_i + b &<= -1 \end{aligned} \quad (1)$$

The data points can be represented by equation (2) [22]:

$$f(x) = \mathrm{sgn}(mx + b) \quad (2)$$

where sgn() is known as a sign function, which is mathematically represented by the following equation:

$$\mathrm{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \quad (3)$$

There can be many hyperplanes which can divide the data space into two regions but the one that increases the distance amid the bordering data points in the input data space is the result to the two class problem. The adjoining data points close to this hype-rplane are called support vectors. This concept can be illutrated geometrically as in Figure 2.

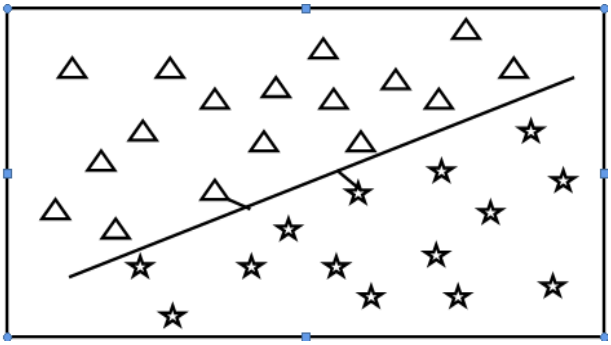


Figure 2 : Distance of the nearest data vectors from the Hyperplane

The margin width can be represented mathematically by the equation:

$$M = \frac{(x^{+} - x^{-}).m}{|m|} = \frac{2}{|m|} \quad (4)$$

This maximization problem viz. maximizing the distance between the hyperplane and the adjoining support vectors can be represented as a Quadratic Optimization Problem as in equation(5)[22][23]:

$$h(m) = \frac{1}{2} m^t m \quad (5)$$

subject to $y_i(mx_i + b) >= 1, \forall i$

The solution for this problem can be provide by a Lagrange multiplier α_i which is associated with every constraint in the main problem. The solution can be represented as:

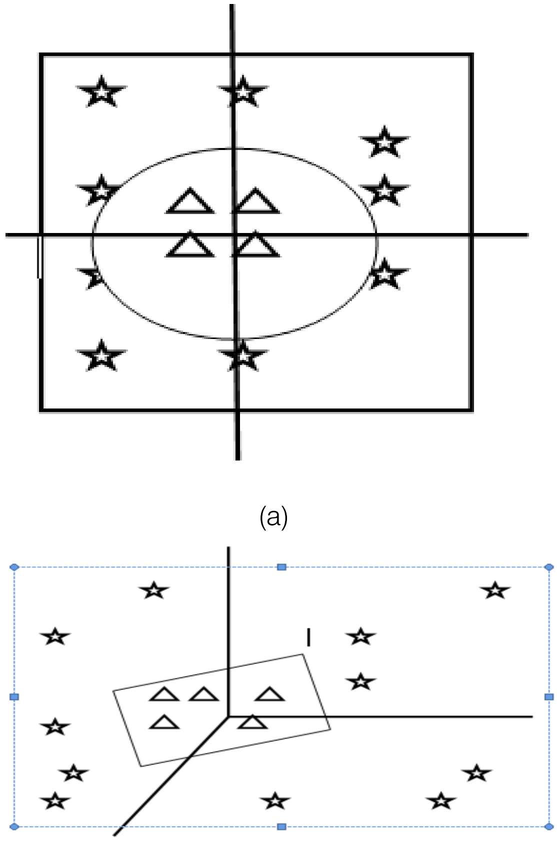
$m = \sum \alpha_i y_i x_i$

$b = y_k - m^t x_k$ for any x_k such that Lagrange multiplier $\alpha_k \# 0$ (6)

The classifier can be denoted as[16]:

$$f(x) = \sum \alpha_i y_i x_i \, x + b \quad (7)$$

In the case of non-linear SVM's the input data space can be generalized onto a higher dimensional feature space as illustrated in Fig 3.



(a)

(b)

Figure 3 : (a) Input space (b) Higher dimensional feature space

If every data point in the input data space is generalized onto a higher dimensional feature space which can be represented as [18]:

$$K(x_i,x_j)=f(x_i)^{t.}f(x_j) \qquad (8)$$

This is also called a kernel function. It is computed using an inner dot product in the feature space. Various kernel functions can be used to do the said mapping as mentioned in the below equations[23]:

Linear Kernel function $= x_i^t x_j$

Polynomial kernel function $= (1 + x_i^t x_j)^p$

Gaussian radial based kernel function $= \exp(-\frac{|xi - xj|^2}{2\sigma^2})$

Sigmoid kernel function$= \tanh(\beta_0 x_i x_j + \beta_1)$ (9)

One of the major advantages of SVM is that feature selection is automatically taken care by it and one need not separately derive features.

III. Materials and Methods

a) *Data Acquisition*

In this paper SVM classification methodology is applied on two different data set formats. The first format of data sets used is a comma separated value(CSV) file which shall have all relevant attributes necessary for the said classification separated by comma. The data sets used in this category is taken from the birds species occurrences of North-east India[24]. The second format of data sets for classification is in raster format[25]. Raster image is a collection of pixels represented in a matrix form. Raster images can be stored in varying formats. The raster format used herein is TIFF format. A map of Andhra Pradesh state in India used.

b) *Proposed Method*

The data under consideration is first prep-rocessed. [26]. In the case of csv datasets comprising of information of birds of North-east India the attributes considered are id, family, genus, specific_ epit-het, latitude, longitude, ver-b-atim _scientific_ name, ve-rba-tim_ family, verbatim_ genus, verbatim_ specific_ ep-ithet and locality. A variable called churn acts as a class label which would categorize the data into two cate-goriesviz onehaving data sets of birds from Darjeeling area and the other having data sets of birds belonging to other north eastern parts in India. Before applying the clas-sif-ication the data sets are cleaned to remove any mis-sing values. In the case of raster data set, a TIFF image is used. The image comprises of a map of Andhra Pradesh, a state in India. Initially a region of interest(ROI) is captured and later supervised SVM classification m-ethodology is applied. Algorithm that explains imple-mentation of SVM is given below [27]:

```
Begin
Step 1: Loop the n data items
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
```

IV. Performance Analysis

a) *Environment Setting*

A total of 695 data set records act as test data set and are used to authenticate the classification results obtained for CSV data sets and in the case of TIFF raster data sets one Region of interest is extracted from the given input image. The proposed method has been implemented under the environment setting as shown in Table 1[28][29].

Table1 : Environment Setting

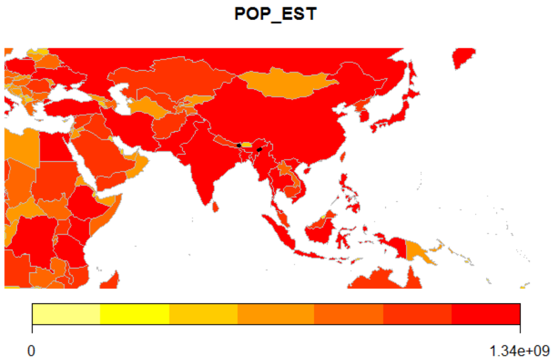
Item	Capacity
CPU	Intel CPU G645 @2.9 GHz processor
Memory	8GB RAM
OS	Windows 7 64-bit
Tools	R, R Studio, Monteverdi tool

b) *Result Analysis*

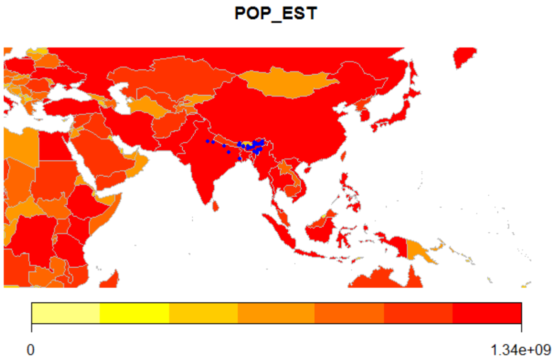
Classification accuracy can be measured using parameters of a confusion or error matrix view depe-nding on whether the event is correctly classified or no event is correctly classified as shown in Table 2[9]. And the classified results for CSV format data sets is demo-nstrated in Figure 4.

Table 2 : Confusion / Error Matrix View

Real group	Classification result	
	No Event	Event
No Event	True Negative(TN)	False Positive(FP)
Event	False Negative(FN)	True Positive(TP)



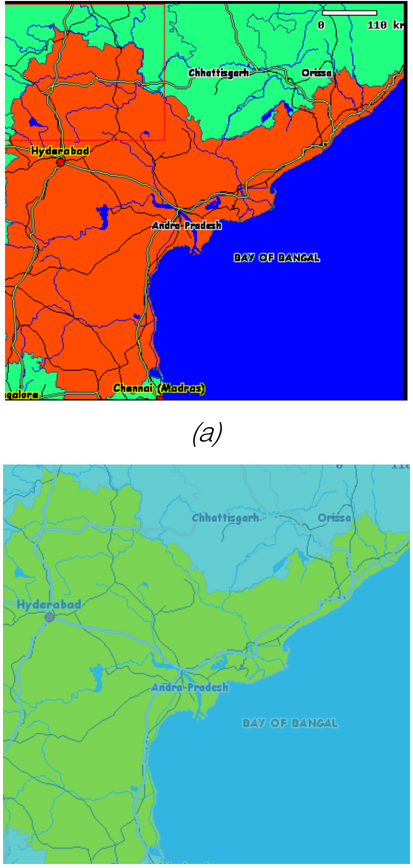
(a)



(b)

Figure 4 : (a) Birds data belonging to Darjeeling area from input dataset in black color(b) Birds data belonging to parts other tan Darjeeling area from input dataset marked in blue color

The region of interest for the raster data set and the classified image is shown in Fig 5.



(a)

(b)

Figure 5 : (a) Region of Interest from the input raster data set. (b) Classified image with Andhra Pradesh land represented with green and water represented with light blue color.

In this paper the parameters used to evaluate the classification is Accuracy and kappa statistics. The formulae for accuracy, specificity, sensitivity and kappa statistics are provided by equations (10), (11), (12) and (13)[30][31][32]:

$$\text{Accuracy} = \frac{TP+TN}{(TP+FN+FP+TN)} \times 100 \tag{10}$$

$$\text{Specificity} = \frac{TN}{(TN+FP)} \times 100 \tag{11}$$

$$\text{Sensitivity} = \frac{TP}{(TP+FN)} x100 \tag{12}$$

$$\text{Kappa statistics} = \text{Sensitivity} + \text{Specificity} - 1 \tag{13}$$

The efficiency of the proposed SVM classifier is evaluated using the said parameters. The confusion or error matrix view for SVM classifier while classifying the CSV data sets is given in Table 3.

Table 3 : Confusion Matrix for CSV data sets

Prediction	Reference	
	Other parts	Darjeeling
Other parts	571	1
Darjeeling	7	116

The confusion matrix or error matrix view for SVM Classifier while classifying raster TIFF data sets is given in Table 4.

Table 4 : Confusion Matrix for rater datasets

Prediction	Reference	
	Land	Water
Land	78	0
Water	0	56

Performance Measures using evaluation metrics are specified in Table 5 which are calculated using equations (10), (11), (12)and (13).

Table 5 : Performance measures for CSV and raster datasets

Data set type	Accuracy	Kappa Statistics
CSV data sets	98.85	95.97
Raster TIFF data sets	100	100

V. Conclusion

In this paper SVM classification method is used to build a classification model for two datasets. The first data set is of CSV format and the second one is a raster TIFF image. Later the classification model is validated against a test data set which is a subset of the input dataset. The performance of SVM is calculated using kappa statistics and accuracy parameters and it is established that for the given data sets SVM classifies the raster image dataset with better accuracy than the CSV dataset. The SVM classification methodology discussed herein can help in environment monitoring, land use, mineral resource identification, classification of remote sensed data into roads and land etc.. in the future.

VI. Acknowledgment

We direct our frank appreciativeness to CSV data which was accessed via GBIF data portal, http://data.gbif.org/ and image data accessed through http://maptell.com for providing us with CSV and raster image data sets. We also thank ANU university for providing all the support in the work conducted.

References Références Referencias

1. E.W.T. Ngai, Yong Hu, Y.H. Wong, Yijun Chen, Xin Sun, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decision Support Systems, Volume 50, Issue 3, February 2011,Pages559569,ISSN01679236,http://dx.doi.org/ 10.1016/ j.dss.2010.08.006.
2. Wilson A. Castillo Rojas, Claudio J. Meneses Villegas, Graphical Representation and Exploratory Visualization for Decision Trees in the KDD Process, Procedia - Social and Behavioral Sciences, Volume 73, 27 February 2013, Pages 136-144, ISSN 1877-0428,http://dx.doi.org/10.1016/j.sbspro.2013.02.033.
3. Borja Calvo, Pedro Larrañaga, José A. Lozano, Learning Bayesian classifiers from positive and unlabeled examples, Pattern Recognition Letters, Volume 28, Issue 16, 1 December 2007, Pages 23752384,ISSN01678655,http://dx.doi.org/10.1016/j. patrec.2007.08. 003.
4. Yugal kumar, G. Sahoo, Analysisof Parametric & Non Parametric Classifiers for Classification Technique using WEKA, I.J. Information Technology and Computer Science, 2012, 7, 43-49 Published Online July 2012 in MECS DOI:10.5815/ijitcs.2012.07.06a
5. Ismail Hmeidi, Bilal Hawashin, Eyas El-Qawasmeh, Performance of KNN and SVM classifiers on full word Arabic articles, Advanced Engineering Informatics, Volume 22, Issue 1, January 2008, Pages106111,ISSN14740346,10.1016/j.aei.2007.12. 001.
6. Jing Hu, Daoliang Li, QinglingDuan, Yueqi Han, Guifen Chen, Xiuli Si, Fish species classification by color, texture and multi-class support vector machine using computer vision, Computers and Electronics in Agriculture, Volume 88, October 2012, Pages133140,ISSN01681699,10.1016/j.compag.201 2.07.008
7. Xiaowei Yang, Qiaozhen Yu, Lifang He, Tengjiao Guo, The one-against-all partition based binary tree support vector machine algorithms for multi-class classification, Neuro computing, Volume 113, 3 August2013,Pages17,ISSN09252312,http://dx.doi.or g/10.1016j.neucom.2012.12.048.
8. Yang Shao, Ross S. Lunetta, Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 70, June2012,Pages7887,ISSN0924716,10.1016/j.isprsj prs.2012.04.001
9. Rajasekhar, N.; Babu, S.J.; Rajinikanth, T.V., "Magnetic resonance brain images classification using linear kernel based Support Vector Machine, "Engineering (NUiCONE), 2012 Nirma University International Conference on , vol., no pp.1,5, 6-8 Dec 2012 doi10.1109/NUICONE.2012.6493213.
10. Yang Shao, Ross S. Lunetta, Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 70, June 2012, Pages 78-87, ISSN 0924-2716, http://dx.doi.org/10.1016/j.isprsjprs.2012.04. 001.
11. Hongji Lin, Han Lin, Weibin Chen, Study on Recognition of Bird Species in Minjiang River Estuary Wetland, Procedia Environmental Sciences, Volume 10, Part C, 2011, Pages 2478-2483, ISSN 1878-0296,http://dx.doi.org/10.1016/j.proenv.2011.09.386.
12. Baoping Tang, Tao Song, Feng Li, Lei Deng, Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine, Renewable Energy, Volume 62, February 2014, Pages 1-9, ISSN 0960-1481, http://dx.doi.org/10.1016/j.renene.2013.06.025.
13. A.D. Dileep, C. Chandra Sekhar, Class-specific GMM based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines, Speech Communication, Volume 57, February 2014, Pages126143,ISSN01676393,http://dx.doi.org/10.10 16/ j.specom.2013.09.010.
14. Hsun-Jung Cho, Ming-TeTseng, A support vector machine approach to CMOS-based radar signal processing for vehicle classification and speed estimation, Mathematical and Computer Modeling, Volume 58, Issues 1–2, July 2013, Pages 438-448 , ISSN08957177,http://dx.doi.org/10.1016/j.mcm.2012 .11 .003
15. Lam Hong Lee, Rajprasad Rajkumar, Lai Hung Lo, Chin Heng Wan, Dino Isa, Oil and gas pipeline failure prediction system using long range ultrasonic transducers and Euclidean-Support Vector Machines classification approach, Expert Systems with Applications, Volume 40,Issue 6,May 2013, Pages 1925-1934,ISSN0957-4174,http:// dx.doi.org/10.1016 /j.eswa.2012.10.006.
16. Elias Zintzaras, Axel Kowald, Forest classification trees and forest support vector machines algorithms: Demonstration using microarray data, Computers in Biology and Medicine, Volume40,Issue5, May2010, Pages519524,ISSN00104825,http://dx.doi.org/10.10 16/j.compbiomed.2010.03.006.
17. Xinjun Peng, Yifei Wang, Dong Xu, Structural twin parametric-margin support vector machine for binary classification, Knowledge-Based Systems, Volume 49,September2013,Pages6372,ISSN09507051,http:/ / dx.doi.org/10.1016/j.knosys.2013.04.013.

18. Maysam Abedi, Gholam-Hossain Norouzi, Abbas Bahroudi, Support vector machine for multi-classification of mineral prospectivity areas, Computers & Geosciences, Volume 46, September 2012,Pages272283,SSN0098004,http://dx.doi.org/10.1016/ j.cageo.2011.12.014.
19. Liu ,Mingjun Wang, Jun Wang, Duo Li, Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar, Sensors and Actuators B: Chemical, Volume 177, February 2013, Pages970980,ISSN09254005,http://dx.doi.org/10.1016/ j.snb.2012.11.071.
20. F. Löw, U. Michel, S. Dech, C. Conrad, Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using Support Vector Machines, ISPRS Journal of Photogrammetry and Remote Sensing, Volume85,November2013,Pages102119,ISSN09242716, http://dx.doi.org/10.1016/j.isprsjprs.2013.08.007.
21. Steve R. Gunn, Support Vector Machines for Classification and Regression, A Technical Report Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, University Of South Hampton,May1998
22. Till Rumpf, Christoph Römer, Martin Weis, Markus Sökefeld, Roland Gerhards, Lutz Plümer, Sequential support vector machine classification for small-grain weed species discrimination with special regard to Cirsiumarvense and Galiumaparine, Computers and Electronics in Agriculture, Volume 80, January 2012, Pages89-96,ISSN01681699,http://dx.doi.org/10.1016/j.compag.2011.10.018.
23. Hassiba Nemmour, Youcef Chibani, Multiple support vector machines for land cover change detection: An application for mapping urban extensions, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 61, Issue 2, November 2006, Pages 125-133,ISSN09242716,http://dx.doi.org/10.1016/j.isprsjprs. 2006.09.004.
24. Sujit Narwade, Mohit Kalra, Rajkumar Jagdish, Divya Varier, Sagar Satpute, Gautam Talukdar (2011), Collaborative work of Environmental Information System (ENVIS) Centre and Important Bird Areas Programmes-Indian Bird Conservation Network (IBA-IBCN) projects of the BNHS, India, Zookeys, Narwade, S., et al (2011): Literature based species occurrencedataofbirdsofNortheastIndia.www.maptell.com/.
25. D.Lu & Q. Weng (2007):A survey of image classification methods and techniques for improving classification performance, International Journal of Remote Sensing, 28:5,823-870,http://dx.doi.org/10.1080/01431160600746456
26. S.N.Jeyanthi, Efficient Classification Algorithms using SVMs for Large Datasets, A Project Report Submitted in partial fulfillment of the requirements for the Degree of Master of Technology in Computational Science, Supercomputer Education and Research Center, IISC, BANGALORE, INDIA, June 2007
27. R Core Team},R:A Language and Environment for Statistical Computing, R Foundation for Statistical Computing,Vienna,Austria,2013,http://www.Rproject.orgwww.orfeotoolbox.org/otb/monteverdi.html
28. Jennifer A. Taylor, Alicia V. Lacovara, Gordon S. Smith, Ravi Pandian, Mark Lehto, Near-miss narratives from the fire service: A Bayesian analysis, Accident Analysis & Prevention, Volume 62, January 2014,Pages119129,ISSN00014575,http://dx.doi.org/10.1016/ j.aap.2013.09.012.
29. David J. Rogers, Jonathan E. Suk, Jan C. Semenza, Using global maps to predict the risk of dengue in Europe, Acta Tropica, Volume 129, January 2014, Pages 1-14, ISSN 0001-706X, http://dx.doi.org/10.1016/j.actatropica.2013.08.008.
30. Rafael Pino-Mejías, María Dolores Cubiles-de-la-Vega, María Anaya-Romero, Antonio Pascual-Acosta, Antonio Jordán-López, Nicolás Bellinfante-Crocci, Predicting the potential habitat of oaks with data mining models and the R system, Environmental Modelling & Software, Volume 25 ,Issue 7, July 2010, Pages 826-836, ISSN 1364-8152,http://dx.doi.org/10.1016/j.envsoft.2010.01.004.

Global Journals Inc. (US) Guidelines Handbook 2014

www.GlobalJournals.org