

Supervised Classification of Remote Sensed data Using Support Vector Machine

Tarun Rao¹ and T.V.Rajinikanth²

¹ Acharya Nagarjuna University,

Received: 8 December 2013 Accepted: 1 January 2014 Published: 15 January 2014

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.

Index terms— classification, data mining, support vector machine, remote sensed data.

1 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. Classification can be further categorized as supervised and unsupervised classification. In supervised classification the class labels or categories into which the data sets need to be classified into is known in advance. In unsupervised 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 probability distribution of each class. Non parametric classifiers 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 different 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 diagnosis [12], character classification [5], speech recognition [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. The line mentioned herein is called a hyperplane and can be mathematically represented by equation (1) [21]:

2 II.

3 Background Knowledge

$$mx + b \geq +1 \text{ or } mx + b \leq -1 \quad (1)$$

The data points can be represented by equation (2): $f(x) = \text{sgn}(mx + b)$ (2)

where $\text{sgn}()$ is known as a sign function, which is mathematically represented by the following equation: $\text{sgn}(x) = 1$ if $x > 0$, 0 if $x = 0$, -1 if $x < 0$ (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 hyperplane are called support vectors. This concept can be illustrated geometrically as in Figure ???. 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 (4) [22][23]:

$$h(m) = \frac{1}{2} m^T m \quad (4)$$

subject to $y_i (mx_i + b) \geq 1, \forall i$ The solution for this problem can be provided by a Lagrange multiplier λ_i which is associated with every constraint in the main problem. The solution can be represented as: $m = \sum \lambda_i x_i$, $b = y_k - m^T x_k$ for any x_k such that Lagrange multiplier $\lambda_k \neq 0$ (6) The classifier can be denoted as [16]: $f(x) = \text{sgn}(\sum \lambda_i x_i^T x + b)$ (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. If every data point in the input data space is generalized onto a higher dimensional feature space which can be represented as [18]: $K(x_i, x_j) = f(x_i)^T \cdot f(x_j)$ (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 = $\exp(-\gamma \|x_i - x_j\|^2)$

Sigmoid kernel function = $\tanh(\gamma x_i^T x_j + c)$ (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.

4 III.

5 Materials and Methods

6 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.

7 b) Proposed Method

The data under consideration is first preprocessed. [26]. In the case of csv datasets comprising of information of birds of North-east India the attributes considered are id, family, genus, specific_epithet, latitude, longitude, verbatim_scientific_name, verbatim_family, verbatim_genus, verbatim_specific_epithet and locality. A variable called churn acts as a class label which would categorize the data into two categories viz one having 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 classification the data sets are cleaned to remove any missing 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 methodology is applied. Algorithm that explains implementation 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

8 End

9 IV. erformance nalysis 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 Classification accuracy can be measured using parameters of a confusion or error matrix view depending 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 demonstrated in Figure 4. κ_0 , κ_1 , κ_2 and κ_3 Kappa statistics= $\text{Sensitivity} + \text{Specificity} - 1$ (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. The confusion matrix or error matrix view for SVM Classifier while classifying raster TIFF data sets is given in Table 4. Performance Measures using evaluation metrics are specified in Table 5 which are calculated using equations κ_0 , κ_1 , κ_2 and (13).

10 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.



Figure 1:

¹© 2014 Global Journals Inc. (US)

²© 2014 Global Journals Inc. (US) Supervised Classification of Remote Sensed data Using Support Vector Machine

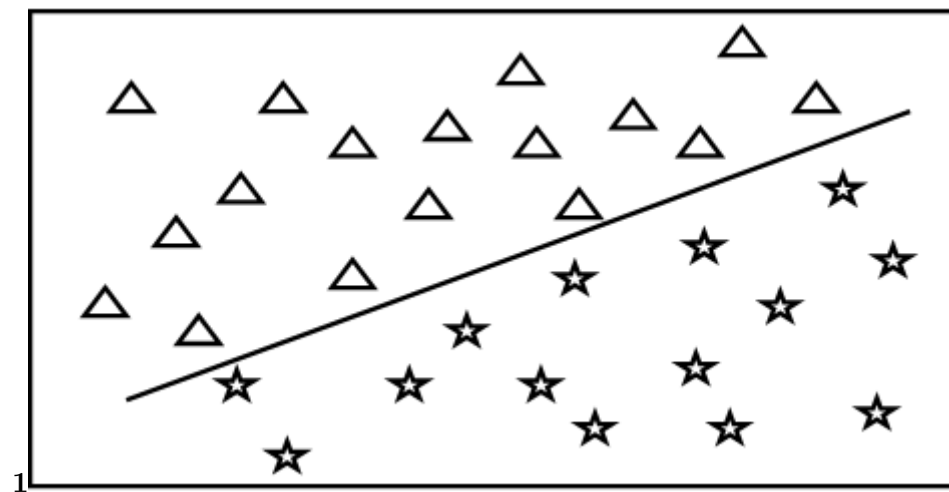


Figure 2: Figure 1 :

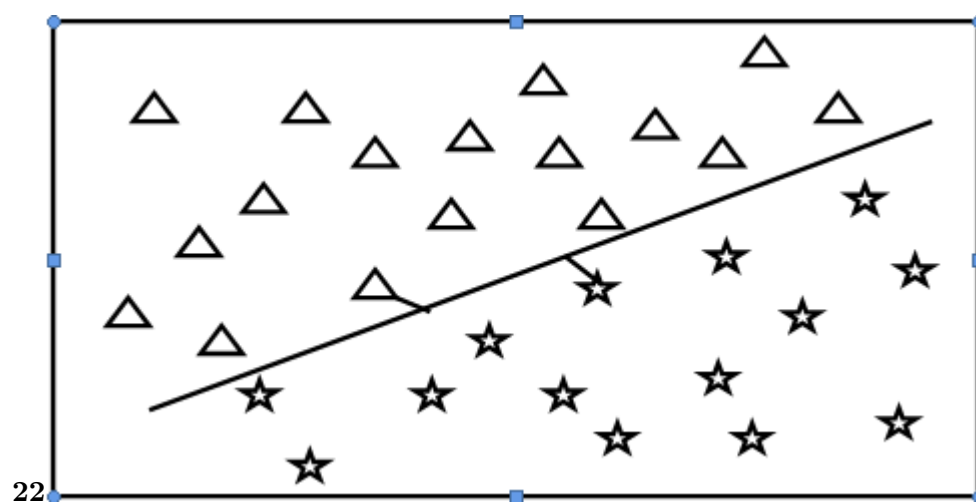


Figure 3: Figure 2 : 2 |??|

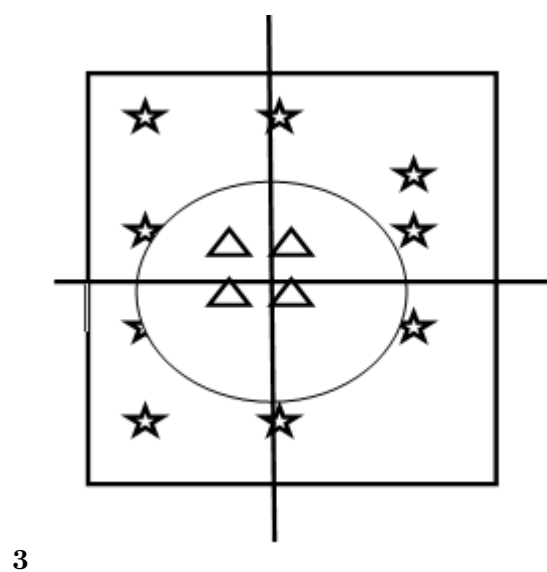
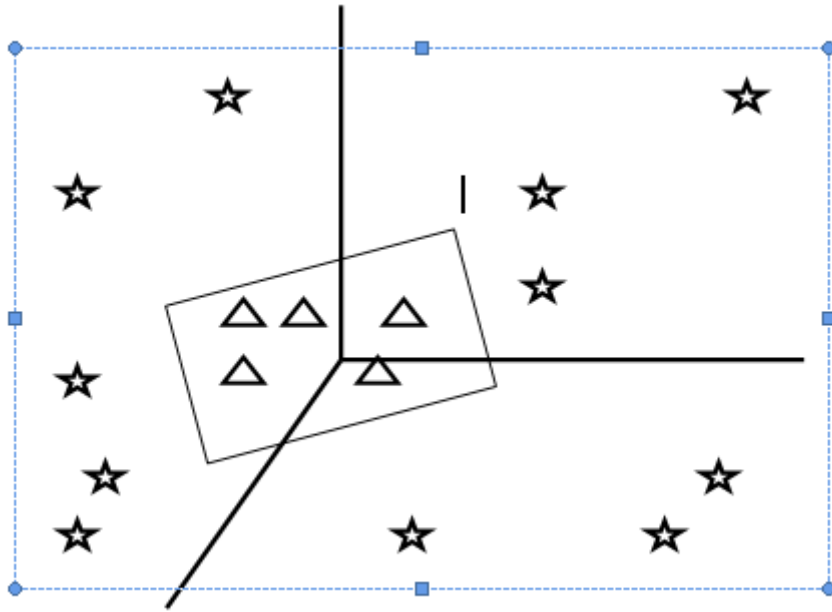
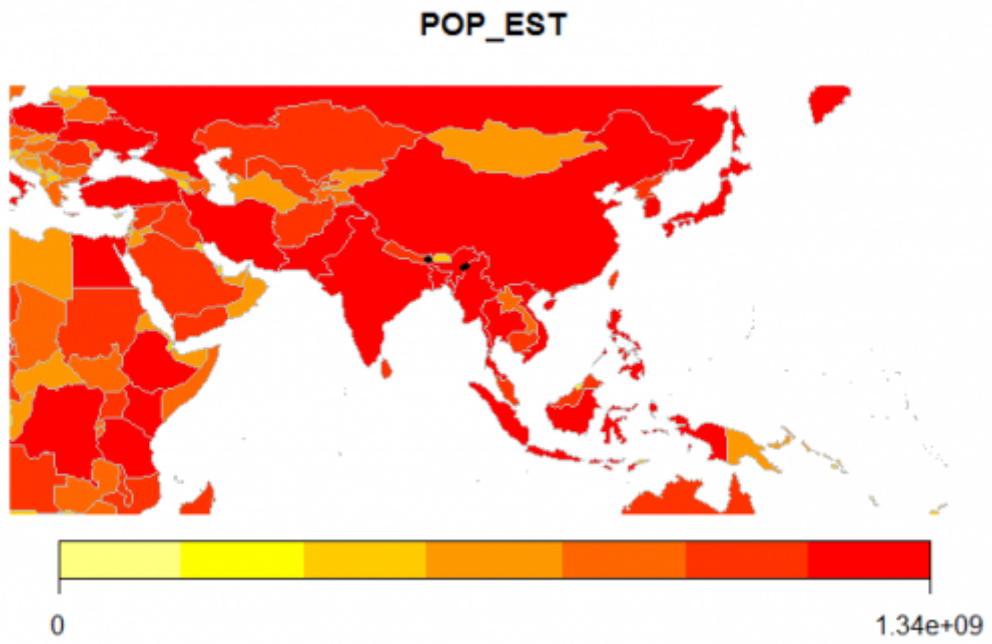


Figure 4: Figure 3 :



4

Figure 5: Figure 4 :



5

Figure 6: Figure 5 :

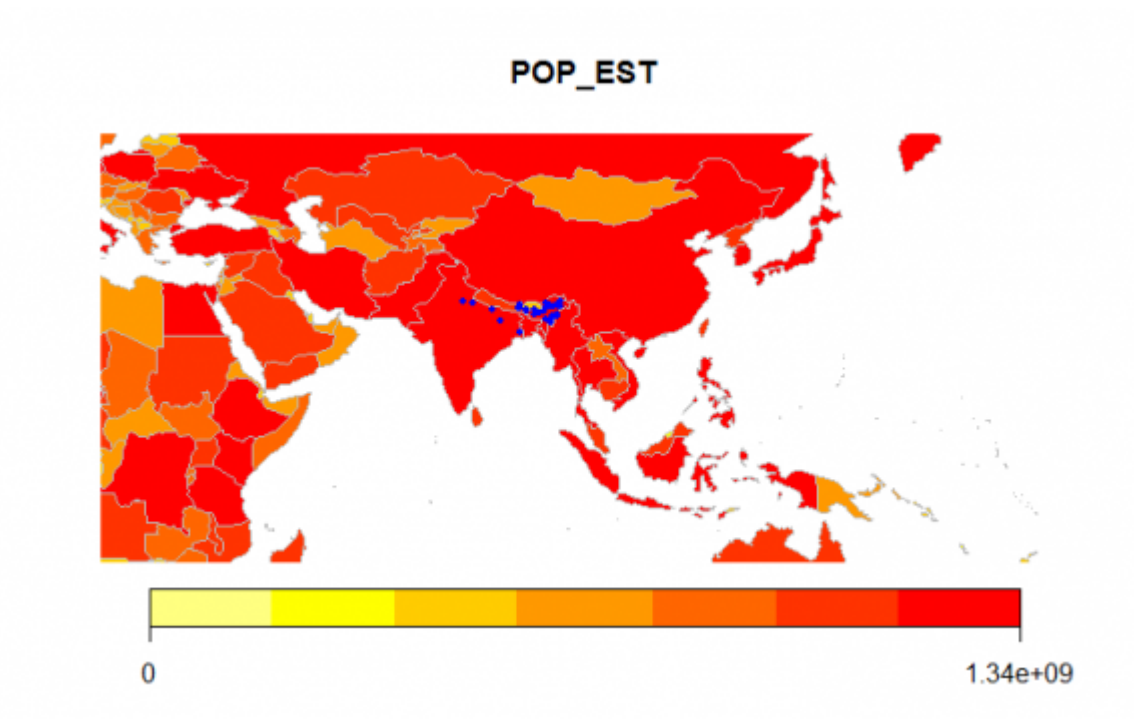


Figure 7:

1

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	

Figure 8: Table 1 [

2

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

Figure 9: Table 2 :

3

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

Figure 10: Table 3 :

4

Prediction	Reference Land	Water
Land	78	0
Water	0	56

Figure 11: Table 4 :

5

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

Figure 12: Table 5 :

.1 Acknowledgment

We direct our frank appreciativeness to CSV data which was accessed via GBIF data portal,

[Cho (2013)] *A support vector machine approach to CMOS-based radar signal processing for vehicle classification and speed estimation*, *Mathematical and Computer Modeling*, Hsun-Jung Cho , Ming-Tetseng . 10.1016/j.mcm.2012.11.003. <http://dx.doi.org/10.1016/j.mcm.2012.11.003> Issues 1-2, July 2013. 58 p. .

[Lu and Weng ()] ‘A survey of image classification methods and techniques for improving classification performance’. D Lu , & Q Weng . 10.1080/01431160600746456. <http://dx.doi.org/10.1080/01431160600746456> *International Journal of Remote Sensing* 2007. 28 (5) p. .

[Yugal Kumar and Sahoo (2012)] ‘Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA’. G Yugal Kumar , Sahoo . 10.5815/ijitcs.2012.07.06a. *I.J. Information Technology and Computer Science* 2012. July 2012. 7 p. . (Published Online)

[Dileep and Sekhar (2014)] ‘Class-specific GMM based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines’. A D Dileep , C Chandra Sekhar . 10.1016/j.specom.2013.09.010. <http://dx.doi.org/10.1016/j.specom.2013.09.010> *Speech Communication* February 2014, Pages126143,ISSN01676393. 57.

[Sujit Narwade et al. ()] *Collaborative work of Environmental Information System (ENVIS) Centre and Important Bird Areas Programmes-Indian Bird Conservation Network (IBA-IBCN) projects of the BNHS*, Mohit Sujit Narwade , Rajkumar Kalra , Divya Jagdish , Sagar Varier , Gautam Satpute , S Talukdar ; Narwade . www.maptell.com/ 2011. 2011. India, Zookeys. (Literature based species occurrence data of birds of Northeast India)

[Liu et al. (2013)] ‘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’. Mingjun Liu , Jun Wang , Duo Wang , Li . 10.1016/j.snb.2012.11.071. <http://dx.doi.org/10.1016/j.snb.2012.11.071> *Sensors and Actuators B: Chemical* February 2013, Pages970980,ISSN09254005. 177.

[Shao and Lunetta (2012)] ‘Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points’. Yang Shao , Ross S Lunetta . 10.1016/j.isprsjprs.2012.04.001. <http://dx.doi.org/10.1016/j.isprsjprs.2012.04.001> *ISPRS Journal of Photogrammetry and Remote Sensing* 0924-2716. June 2012. 70 p. .

[Shao and Lunetta ()] ‘Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points’. Yang Shao , Ross S Lunetta . 10.1016/j.isprsjprs.2012.04.001. *ISPRS Journal of Photogrammetry and Remote Sensing* June2012,Pages7887,ISSN0924716. 70.

[Jeyanthi (2007)] *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*, S N Jeyanthi . June 2007. IISC, BANGALORE, INDIA. Supercomputer Education and Research Center

[Hmeidi and Hawashin (2008)] ‘Eyas El-Qawasmeh, Performance of KNN and SVM classifiers on full word Arabic articles’. Ismail Hmeidi , Bilal Hawashin . 10.1016/j.aei.2007.12.001. *Advanced Engineering Informatics* January 2008, Pages106111,ISSN14740346. 22 (1) .

[Tang et al. (2014)] *Fault diagnosis for a wind turbine transmission system based on manifold learning and Shannon wavelet support vector machine*, *Renewable Energy*, Baoping Tang , Tao Song , Feng Li , Lei Deng . 10.1016/j.renene.2013.06.025. <http://dx.doi.org/10.1016/j.renene.2013.06.025> February 2014. 62.

[Hu et al. (2012)] *Fish species classification by color, texture and multi-class support vector machine using computer vision*, *Computers and Electronics in Agriculture*, Jing Hu , Daoliang Li , Yueqi Qinglingduan , Guifen Han , Xiuli Chen , Si . ISSN01681699,10.1016/j.compag.2012.07.008. October 2012. Pages133140. 88.

[Zintzaras and Kowald ()] ‘Forest classification trees and forest support vector machines algorithms: Demonstration using microarray data’. Elias Zintzaras , Axel Kowald . 10.1016/j.compbio.2010.03.006. <http://dx.doi.org/10.1016/j.compbio.2010.03.006> *Computers in Biology and Medicine* Volume40,Issue5, May2010, Pages519524,ISSN00104825.

[Wilson et al. (2013)] ‘Graphical Representation and Exploratory Visualization for Decision Trees in the KDD Process’. A Wilson , Claudio J Meneses Castillo Rojas , Villegas . 10.1016/j.sbspro.2013.02.033. <http://dx.doi.org/10.1016/j.sbspro.2013.02.033> *Procedia -Social and Behavioral Sciences* 1877- 0428. 27 February 2013. 73 p. .

[Löw et al. ()] *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, F Löw , U Michel , S Dech , C Conrad . 10.1016/j.isprsjprs.2013.08.007. <http://dx.doi.org/10.1016/j.isprsjprs.2013.08.007> November2013,Pages102119,ISSN09242716.

- [Calvo et al. (2007)] ‘Learning Bayesian classifiers from positive and unlabeled examples’. Borja Calvo , Pedro Larrañaga , José A Lozano . 10.1016/j.patrec.2007.08.003. <http://dx.doi.org/10.1016/j.patrec.2007.08.003> Pages 23752384,ISSN01678655, 1 December 2007. 28.
- [Rajasekhar et al. (2012)] ‘Magnetic resonance brain images classification using linear kernel based Support Vector Machine’. N Rajasekhar , S J Babu , T V Rajinikanth . 10.1109/NUICONE.2012.6493213. *Nirma University International Conference on*, 2012. 6-8 Dec 2012. 5. (Engineering (NUiCONE))
- [Nemmour and Chibani (2006)] ‘Multiple support vector machines for land cover change detection: An application for mapping urban extensions’. Hassiba Nemmour , Youcef Chibani . 10.1016/j.isprsjprs.2006.09.004. <http://dx.doi.org/10.1016/j.isprsjprs.2006.09.004> Pages 125-133,ISSN09242716, November 2006. 61.
- [Taylor et al. (2014)] ‘Near-miss narratives from the fire service: A Bayesian analysis’. Jennifer A Taylor , Alicia V Lacovara , Gordon S Smith , Ravi Pandian , Mark Lehto . 10.1016/j.aap.2013.09.012. <http://dx.doi.org/10.1016/j.aap.2013.09.012> *Accident Analysis & Prevention* January 2014,Pages119129,ISSN00014575. 62.
- [Lam Hong Lee et al. ()] *Oil and gas pipeline failure prediction system using long range ultrasonic transducers and Euclidean-Support Vector Machines classification approach, Expert Systems with Applications*, Rajprasad Lam Hong Lee , Lai Hung Rajkumar , Chin Lo , Dino Heng Wan , Isa . 10.1016/j.eswa.2012.10.006. <http://dx.doi.org/10.1016/j.eswa.2012.10.006> May 2013. 1925-1934. 40.
- [Pino-Mejías et al. (2010)] ‘Predicting the potential habitat of oaks with data mining models and the R system’. Rafael Pino-Mejías , María Dolores Cubiles-De-La-Vega , María Anaya-Romero , Antonio Pascual-Acosta , Antonio Jordán-López . 10.1016/j.envsoft.2010.01.004. <http://dx.doi.org/10.1016/j.envsoft.2010.01.004> *Environmental Modelling & Software*, July 2010. 25 p. . Nicolás Bellinfante-Crocci
- [R Core ()] Team} R Core . <http://www.Rproject.orgwww.orfeotoolbox.org/otb/monteverdi.html> *R:A Language and Environment for Statistical Computing, R Foundation for Statistical Computing*, (Vienna,Austria) 2013.
- [Rumpf et al. (2012)] *Sequential support vector machine classification for small-grain weed species discrimination with special regard to Cirsiumarvense and Galiumaparine, Computers and Electronics in Agriculture*, Till Rumpf , Christoph Römer , Martin Weis , Markus Sökefeld , Roland Gerhards , Lutz Plümer . 10.1016/j.compag.2011.10.018. <http://dx.doi.org/10.1016/j.compag.2011.10.018> January 2012, Pages89-96,ISSN01681699. 80.
- [Peng et al. ()] *Structural twin parametric-margin support vector machine for binary classification, Knowledge-Based Systems*, Xinjun Peng , Yifei Wang , Dong Xu . 10.1016/j.knosys.2013.04.013. <http://dx.doi.org/10.1016/j.knosys.2013.04.013> September2013,Pages6372,ISSN09507051. 49.
- [Lin et al. ()] ‘Study on Recognition of Bird Species in Minjiang River Estuary Wetland’. Hongji Lin , Han Lin , Weibin Chen . 10.1016/j.proenv.2011.09.386. <http://dx.doi.org/10.1016/j.proenv.2011.09.386> *Procedia Environmental Sciences* 1878- 0296. 2011. 10 p. .
- [Abedi et al. (2012)] ‘Support vector machine for multiclassification of mineral prospectivity areas’. Maysam Abedi , Gholam-Hossain , Abbas Norouzi , Bahroudi . 10.1016/j.cageo.2011.12.014. <http://dx.doi.org/10.1016/j.cageo.2011.12.014> *Computers & Geosciences* September 2012,Pages272283,SSN0098004. 46.
- [Gunn ()] *Support Vector Machines for Classification and Regression*, Steve R Gunn . May1998. Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, University Of South Hampton (Technical Report)
- [Ngai et al. (2011)] ‘The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature’. E W T Ngai , Yong Hu , Y H Wong , Yijun Chen , Xin Sun . 10.1016/j.dss.2010.08.006. <http://dx.doi.org/10.1016/j.dss.2010.08.006> *Decision Support Systems* February 2011,Pages559569,ISSN01679236. 50 (3) .
- [Yang et al.] *The one-against-all partition based binary tree support vector machine algorithms for multi-class classification, Neuro computing*, Xiaowei Yang , Qiaozhen Yu , Lifang He , Tengjiao Guo . 113 p. 3.
- [Rogers et al. (2014)] ‘Using global maps to predict the risk of dengue in Europe’. David J Rogers , Jonathan E Suk , Jan C Semenza . 10.1016/j.actatropica.2013.08.008. <http://dx.doi.org/10.1016/j.actatropica.2013.08.008> *Acta Tropica* 0001-706X. January 2014. 129 p. .