A hybrid Random Forest based Support Vector Machine Classification supplemented by boosting

Tarun Rao[1]

[1] Acharya Nagarjuna University

Received: 7 December 2013 Accepted: 5 January 2014 Published: 15 January 2014

Abstract

This paper presents an approach to classify remote sensed data using a hybrid classifier.Random forest, Support Vector machines and boosting methods are used to build the said hybrid classifier. The central idea is to subdivide the input data set into smaller subsets and classify individual subsets. The individual subset classification is done using support vector machines classifier. Boosting is used at each subset to evaluate the learning by using a weight factor for every data item in the data set. The weight factor is updated based on classification accuracy. Later the final outcome for the complete data set is computed by implementing a majority voting mechanism to the individual subset classification outcomes.

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

1 Introduction

any organizations maintain huge data repositories which store data collected from various sources in different formats. The said data repositories are also known as data warehouses. One of the prominent sources of data is remote sensed data collected via satellites or geographical information systems software's [1].

The data thus collected can be of use in various applications including and not restricted to land use [2] [3], species distribution modeling [4] [5] [6] [7], mineral resource identification [8], traffic analysis [10], network analysis [9] and environmental monitoring systems [11] [12]. Data mining is used to extract information from the said data repositories. The information thus mined can help various stakeholders in an organization in taking strategic decisions. Data can be mined from the data repositories using various methodologies like anomaly detection, supervised classification, clustering, association rule learning, regression, characterization and summarization and sequential pattern mining. In this paper we shall be applying a hybrid classification technique to classify plant seed remote sensed data.

A lot of research has been undertaken to classify plant functional groups, fish species, bird species etc... [7][13] [14].The classification of various species shall help in conserving the ecosystem by facilitating ins predicting of endangered species distribution [15]. It can also help in identifying various resources like minerals, water resources and economically useful trees. Various technologies in this regard have been developed. Machine learning methods, image processing algorithms, geographical information systems tools etc..have added to the development of numerous systems that can contribute to the study of spatial data and can mine relevant information which can be of use in various applications. The systems developed can help constructing classification models that in turn facilitate in weather forecasting, crop yield classification, mineral resource identification, soil composition analysis and also locating water bodies near to the agricultural land.

Classification is the process wherein a class label is assigned to unlabeled data vectors. It can be categorized into supervised and un-supervised classification which is also known as clustering. In supervised classification learning is done with the help of supervisor ie. learning through example. In this method the set of possible class labels is known apriori to the end user. Supervised classification can be subdivided into non-parametric and parametric classification. Parametric classifier method is dependent on the probability distribution of each class. ??on without supervisor ie. learning from observations. In this method set of possible classes is not known to the end user. After classification one can try to assign a name to that class. Examples of un-supervised classification methods are Adaptive resonance theory(ART) 1, ART 2,ART 3, Iterative Self-Organizing Data Analysis Method, K-Means, Bootstrapping Local, Fuzzy C-Means, and Genetic Algorithm [17]. In this paper we shall discuss about a hybrid classification method. The said hybrid method will make use of support vector machine(SVM) classification, random forest and boosting methods. Later its performance is evaluated against traditional individual random forest classifiers and support vector machines.

A powerful statistical tool used to perform supervised classification is Support Vector machines. Herein the data vectors are represented in a feature space. Later a geometric hyperplane 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. It helps in solving equally two class and multi class classification problem. 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 have 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 [18] [19].

In Random forest(RF) classification method many classifiers are generated from smaller subsets of the input data and later their individual results are aggregated based on a voting mechanism to generate the desired output of the input data set. This ensemble learning strategy has recently become very popular. Before RF, Boosting and Bagging were the only two ensemble learning methods used. RF can be applied for supervised classification, unsupervised learning and regression. RF has been extensively applied in various areas including modern drug discovery, network intrusion detection, land cover analysis, credit rating analysis, remote sensing and gene microarrays data analysis etc... ??20][21].

Other popular ensemble classification methods are bagging and boosting. Herein the complex data set is divided into smaller feature subsets. An ensemble of classifiers is formed with the classifiers being used to classify data items in each feature subset. The said feature subsets are regrouped together iteratively depending on penalty factor also known as the weight factor applied based on the degree of misclassification in the feature subsets. The class label of data items in the complete data set is computed by aggregating the individual classification outcomes at each feature subset [22] [23].

A hybrid method is being proposed in this paper which makes use of ensemble learning from RF classification and boosting algorithm and SVM classification method. The processed seed plant data is divided randomly into feature subsets. SVM classification method is used to derive the output at each feature subset. Boosting learning method is applied so as to boost the classification adeptness at every feature subset. Later majority voting mechanism is applied to arrive at the final classification result of the original complete data set.

Our next section describes Background Knowledge about Random Forest classifier, SVM and Boosting. In section 3 proposed methodology has 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.

2 II.

Background Knowledge a) Overview of SVM Classifier Support vector machine (SVM) is a statistical tool used in various data mining methodologies like classification and regression analysis. The data can be present either in the form of a multi class or two class problem. In this paper we shall be dealing with a two class problem wherein the seed plant data sets need to be categorized under two class labels one having data sets belonging to North America and the other having data sets belonging to South America. It has been applied in various areas like species distribution, locating mineral prospective areas etc..It has become popular for solving problems in regression and classification, consists of statistical learning theory based heuristic algorithms. The advantage with SVM is that the classification model can be built using minimal number of attributes which is not the case with most other classification methods [24]. In this paper we shall be proposing a hybrid classification methodology to classify seed plant data which would lead to improving the efficiency and accuracy of the traditional classification approach.

The seed plant data sets used in the paper have data sets with known class labels. A classification model is constructed using the data sets which can be authenticated against a test data set and can later be used to predict class labels of unlabeled data sets. Since class labels of data sets are known apriori this approach is categorized as supervised classification. In unsupervised classification method also known as clustering the class label details is not known in advance. Each data vector in the data set used for classification comprises of unique attributes which is used to build the classification model [25] [19]. The SVM model can be SVM is represented by a separating hyper plane f (x) that geometrically bisects the data space thus dividing it into two diverse regions thus resulting in classification of the input data space into two categories.

Figure ?? : The Hyperplane The function f(x) denotes the hyperplane that separates the two regions and facilitates in classification of the data set. The two regions geometrically created by the hyperplane correspond to the two categories of data under two class labels. A data point x n belongs to either of the region depending on the value of f(x n). If f(x n) > 0 it belongs to one region and if f(x n) < 0 it belongs to another region. There are many such hyperplanes which can split the data into two regions. But SVM ensures that it selects the hyperplane that is at a maximum distance from the nearest data points in the two regions. There are only few hyperplanes that shall satisfy this criterion. By ensuring this condition SVM provides accurate classification results [27].

SVM's can be represented mathematically as well. Assume that the input data consists of n data vectors where each data vector is represented by x i ? R n , where i (=1, 2, ?.., n). Let the class label that needs to be assigned to the data vectors to implement supervised classification be denoted by y i , which is +1 for one category of data vectors and -1 for the other category of data vectors. The data set can be geometrically separated by a hyperplane. Since the hyperplane is represented by a line it can also be mathematically represented by [8][3] [28]:mx i + b >= +1 mx i + b <= -1**(1)**

The hyperplane can also be represented mathematically by [31][32] [33]:f(x)= sgn(mx+ b) = sgn((? ? n i=1 i y i x i). x + b) (2)

where sgn() is known as a sign function, which is mathematically represented by the following equation:sgn(x)=? 1 if x > 0 0 if x = 0 ?1 if x < 0 (3)

The data vectors are said to be optimally divided by the hyperplane if the distance amid the adjoining data vectors in the two different regions from the given hyperplane is maximum.

This concept can be illustrated geometrically as in Figure 2, where the distance between the adjoining data points close to the hyperplane and the hyperplane is displayed [29][30] [28].

This hyperplane which has maximum distance d from adjoining points is computed to implement the said classification. This SVM can be represented as a primal formulation given by the equation [8][5] [31]:h(m)= 1 2 ||m|| 2 + Training error (5) subject to y i (mx i + b) >=1,?i

The idea is to increase the margin and reduce the training error. The data sample records in the training data set belong to input set. Each of the data vectors have precise attributes based on which the classification model is built. These set of attributes are said to form a feature space. The kernel function bridges the gap between the feature space and the input space and enables to carry out classification on input space rather than complicated feature space. [29].

In this paper we have used Gaussian radial basis functions (RBF). SVM's make use of the radial basis kernel function to be able to work at the simpler input space level. The RBF kernel used is represented mathematically by [3][29]: can be solved using various methods. One method is to move the data vectors to a different space thereby making the problem linear. The other method is to split the multi class problem into numerous two class problems and later with a voting mechanism combine the solutions of individual two class problems to get the solution of the original multi class problem. [8].K(x1,x2)=exp(|x 1 ?x 2 | 2 2? 2)**(6)**

The steps followed while using SVM in classifying data are mentioned in the below algorithm [16]:

- **??**——————————————— **??**——————————————— **??**———————————————————— In RF classification method the input data set is first subdivided into two subsets, one containing two thirds of the data points and the other containing the remaining one third. Classification tree models are constructed using the subset comprising of two thirds of data points The subset which contains one third data of data points which are not used at any given point of time to construct classification trees and are used for validation are called out of bag(OOB) data samples of the trees. There is no truncation applied at every classification tree. Hence every classification tree used in RF classification method is maximal in nature. Later RF classification method follows a majority voting process wherein classification output of every classification tree casts a vote to decide the final outcome of the ensemble classifier ie.. assigning a class label to a data item x [21]. The set of features are used to create a classification tree model at every randomly chosen subset [37]. This set of features shall remain constant throughout the growing of random forest.

In RF, the test set is used to authenticate the classification results and also used for predicting the class labels for unlabeled data after the classification model is built. It also helps in cross validation of results among different classification results provided by various classification trees in the ensemble. To perform the said cross validation the out of bag(OOB) samples are used.. The individual classification tree outcomes are aggregated with a majority vote and the cumulative result of the whole ensemble shall be more accurate and prone to lesser classification error than individual classification tree results [26].

Every classification tree in the random forest ensemble is formed using the randomly selected two thirds of input variables, hence there is little connection between different trees in the forest. One can also restrict the number of variables that split a parent node in a classification tree resulting in the reduction of connection between classification trees. The Random forest classification method works better even for larger data sets. This is not the case with other ensemble methods [1] [2]. In this paper we shall be using the both boosting and random forest ensemble classification methods along with support vector machines to give a more accurate classification output. This hybrid method shall be more robust to noise as compared to individual classification method.

RF classification method works with both discreet and continuous variables which is not the case with other statistical classification modeling methods. Furthermore, there is no limit on the total number of classification trees that are generated in the ensemble process and the total number of variable or data samples(generally two thirds are used) in every random subset used to build the classification trees [36].

RF rates variables based on the classification accuracy of the said variable relative to other variables in the data set. This rank is also known as importance index. It reflects the relative importance of every variable in the process of classification. The importance index of a variable is calculated by averaging the importance of the variable across classification trees generated in the ensemble. The more the value of this importance index, the greater is a variables importance for classification. Another parameter obtained by dividing the variable's importance index by standard error is called z-score. Both importance index as well as z-score play a significant role in ensuring the efficiency of the classification process [25][36][39] [38].

The importance of a variable can also be assessed by using two parameters, Gini Index decrease and OOB error estimation. Herein relative importance of variables are calculated which is beneficial in studies wherein the numbers of attributes are very high and thus leading to relative importance gaining prominence [40].

3 Global Journal of Computer Science and Technology

Volume XIV Issue I Version I46 (D D D D) Year C 2014 ? ? ? k(C i ,X) |X| ? . ? k(C j ,X) |X| ? j?i (7)

where k(C i ,X) |X| is the is the probability that a selected case belongsto class C i .

RF method provides precise results with respect to variation and bias [39].. The performance of the RF classification method is better compared to other classifiers like support vector machines, Neural Networks and discriminant analysis. In this paper a hybrid classification method coalescing the advantages of both Random forest and Support vector machines in addition to boosting is used. The RF algorithm is becoming gradually popular with applications like forest classification, credit rate analysis, remote sensing image analysis, intrusion detection etc.

Yet another parameter that can contribute in assessing the classification is proximity measure of two samples. The proximity measure is the number of classification trees in which two data samples end up in the same node. This parameter when divided by the number of classification trees generated can facilitate in detecting outliers in the data sets. This computation requires large amount of memory space, depending on the total number of sample records and classification trees in the ensemble [1]. The pseudo code for Random Forest algorithm is mentioned below [42]:

- **??**——————————————-Random Forest Algorithm: ——————————————Input: D: training sample a: number of input instance to be used to generate classification tree T: total number of classification trees in random forest OT: Classification Output from each tree T 1) OT is empty 2) for i=1 to T 3) Db = Form random sample subsets after selecting 2/3rd instances randomly from D /* For every tree this sample would be randomly selected*/ 4) Cb = Build classification trees using random subsets Db 5) Validate the classifier Cb using remaining 1/3rd instances //Refer Step 3. 6) OT=store classification outputs of classification trees 7) next i 8) Apply voting mechanism to derive output ORT of the Random forest(ensemble of classification trees) 9) return ORT **??**——————————————c) Overview of Boosting Ensemble learning is a process wherein a data set is divided into subsets. Individual learners are then used to classify and build the model for each of these subsets. Later the individual learning models are combined so as to determine the final classification model of the complete data set. As the complex large data set is divided into smaller random subsets and classification model is applied on these smaller subsets the said process of ensemble learning results in improving classification efficiency and gives more accurate results. Numerous classification methodologies like bagging, boosting etc...can also be used in learning by constructing an ensemble [43][44] [45].

In this research paper boosting method has been used to create the said ensemble. It works by rewarding successful classifiers and by applying penalties to unsuccessful classifiers. In the past it has been used in various applications like machine translation [46], intrusion detection [47], forest tree regression, natural language processing, unknown word recognition [48] etc.

Boosting is applied to varied types of classification problems. It is an iterative process wherein the training data set is regrouped together into subsets and various classifiers are used to classify data samples in the subsets. The data samples which were difficult to classify by a classifier also known as a weak learner at one stage are classified using new classifiers that get added to the ensemble at a later stage [49][50] [51]. In this way at each stage a new classifier gets augmented to the ensemble. The difficulty in classifying a data item Xi at stage k is represented by a weight factor Wk(i). The regrouping of training sets at each step of learning is done depending on the weight factor Wk(i) [22]. The value of the weight factor is proportional to the misclassification of the data. This way of forming regrouped data samples at every stage depending on the weight factor is called re-sampling version of boosting. Yet another way of implementing boosting is by reweighting wherein weight factor is assigned iteratively to every data item in the data set and the complete data set is used at every subsequent iteration by modifying the weights at every stage [48] [52].

The most popular boosting algorithm called Adaboost [23]. Adaboost stands for Adaptive Boosting. It adapts or updates weights of the data items based on misclassification of training samples due to weak learners and regroups the data subsets depending on the new weights. The steps of Adaboost algorithm is mentioned below:

- end for ———————————————————In the next section the proposed hybrid methodology is discussed in detail.——————————————————- Adaboost Algorithm ——————————————————————

4 III.

5 Proposed Methodology

In this paper we shall construct a hybrid classification model which shall facilitate in predicting the class label of seed plant data from test data sets. The methodology recommended has been denoted as a schematic diagram as mentioned in Fig 3 and the detailed explanation of the steps followed has been given in the following subsections. The data sets are randomly divided into n different random subsets each subset comprising of two third of the whole data set. Classification methods are applied to each of these random subsets. The remaining one third data sets at each subsets is used as a test set. At each random subset the following attributes were used so as to implement the classification method discussed in the next subsection: id, continent, specificEpithet and churn. Now churn is a variable that is set to yes if the seed plant data belongs to North America or if it belongs to South America it is set to no.

6 d) Selection of an appropriate classification method

In this paper seed plant data sets are classified using a hybrid classification method which makes use of Random forest, SVM classifier and boosting ensemble learning method. In the hybrid methodology the input data set is randomly subdivided into subsets. Each data item in each of the subset has a weight factor associated with it. The data items in the subsets are classified by SVM classifier. If a misclassification has occurred then the weight factor of the data items is increased otherwise it is reduced. The data subsets are rearranged and again SVM classifier is used to perform classification at each subset. The weights are again updated depending on whether it is a proper classification or a misclassification. These steps are iteratively repeated till all the weights get updated to a very low value. The output of the input data set is computed by applying voting mechanism to all the random subsets classification outputs [34]. The algorithm for the proposed hybrid methodology is givenin the sample code herein:

——————————————————Algorithm 1 Hybrid classification using RF and SVM supplemented by boosting - **??**—————————————————— The obtained classification output at each random subset is validated by using the hybrid classifier model to test against the complete data set.

In this paper 10 random feature subsets were used and at every subset SVM classifier was used to perform the said classification. Voting mechanism was then applied to derive the final classification output. In this paper a total of 180 support vectors were used.

IV.

7 Performance Analysis a) Environment Setting

The study area included is from North and South America. It includes data pertaining to localities wherein seed plant species are present.

A total of 599 data set records from North American region and a total of 401data set records from South American region are analyzed in order to execute the proposed method. Sample records used in this paper are shown in Table **??** It is observed that the most conventionally utilized evaluation metrics in classification are accuracy, specificity, positive predictive value and negative predictive value. The formulae for accuracy, specificity, prevalence and negative predictive value are provided by equations (**??**), (**??**), (**??**0) and (**??**1 The confusion matrix or error matrix view for SVM Classifier is given in Table V and for RF Classifier in Table **??**I. Performance Measures using evaluation metrics are specified in Fig 5 which are calculated using equations (**??**), (**??**), (**??**0)and (11).

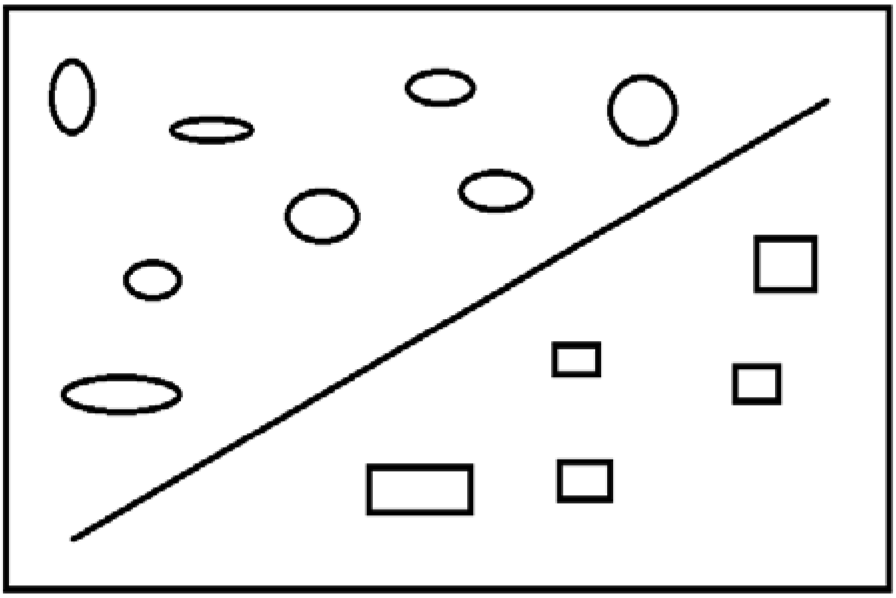
8 Conclusion

In this paper hybrid classifier based on random forest, SVM and boosting methods is used to classify seed plant data. The hybrid classification results are compared with the results attained by implementing classification using traditional SVM and RF classifiers. The research has established that the hybrid approach of classification is more efficient as compared to traditional SVM and RF classifiers since it gives higher values of accuracy, specificity, positive predictive value and negative predictive value.

The reason for better results in the case of hybrid classification methodology used in this paper is since it makes use of the advantages of each of the individual traditional SVM, RF classifications methods. Furthermore, the classification results are supplemented using boosting ensemble classification method. In the future the proposed method can be used so as to classify vector, raster remote sensed data that can be collected via satellites and various geographical information systems.

9 VI.

[1] [2]



2

Figure 1: Figure 2 :

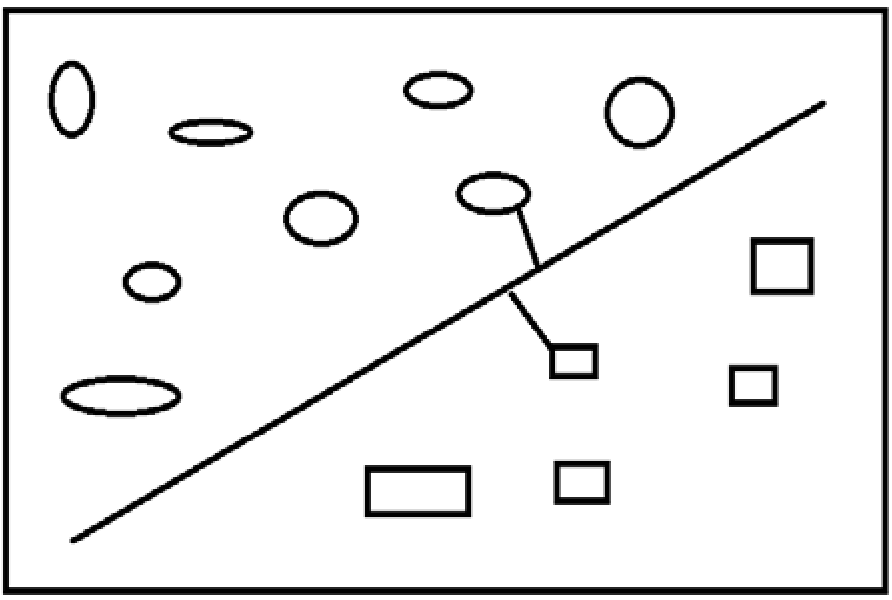
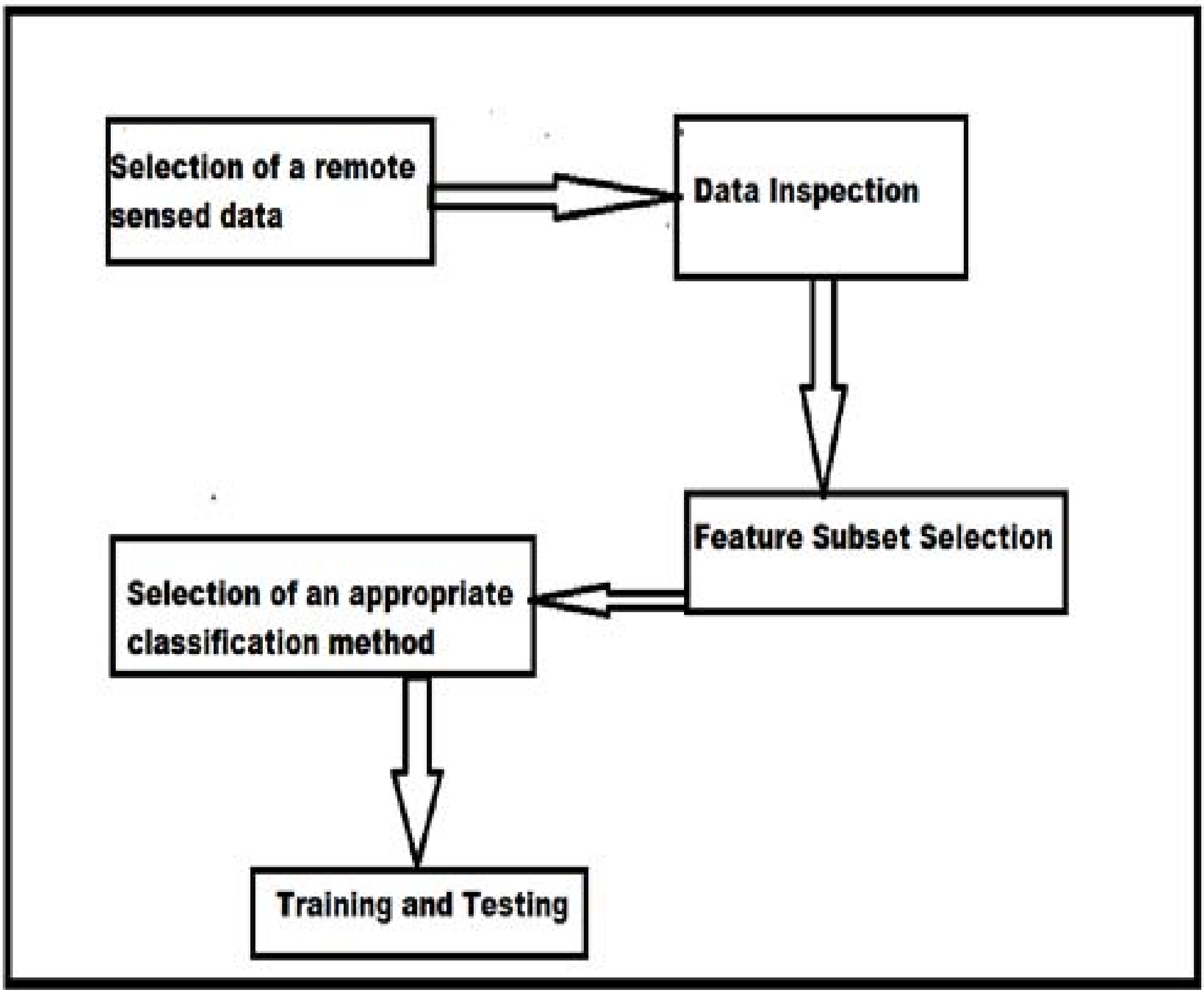


Figure 2:



2

Figure 3: - Algorithm 2

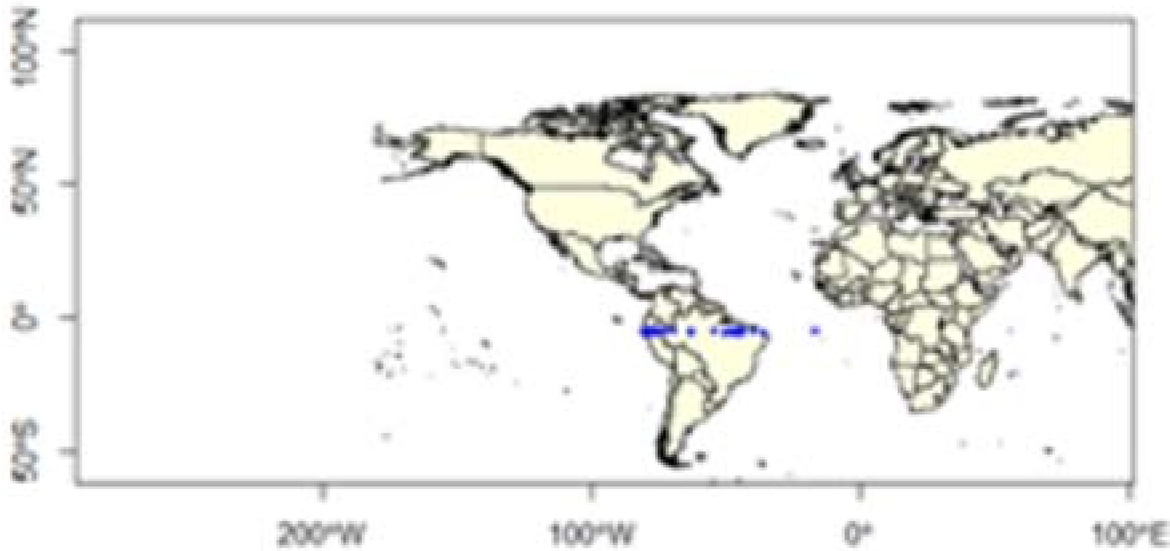


Figure 4:

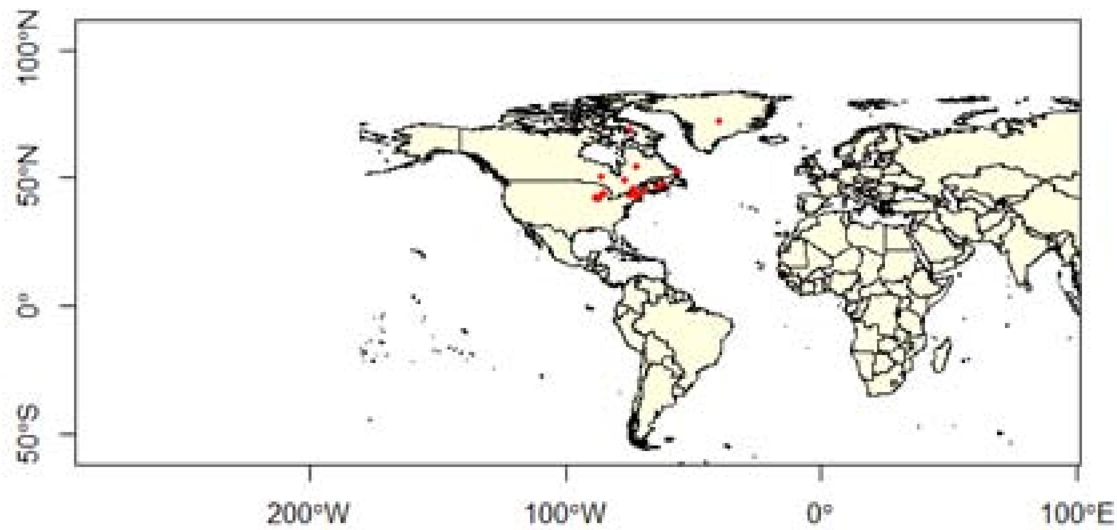


Figure 5:

3

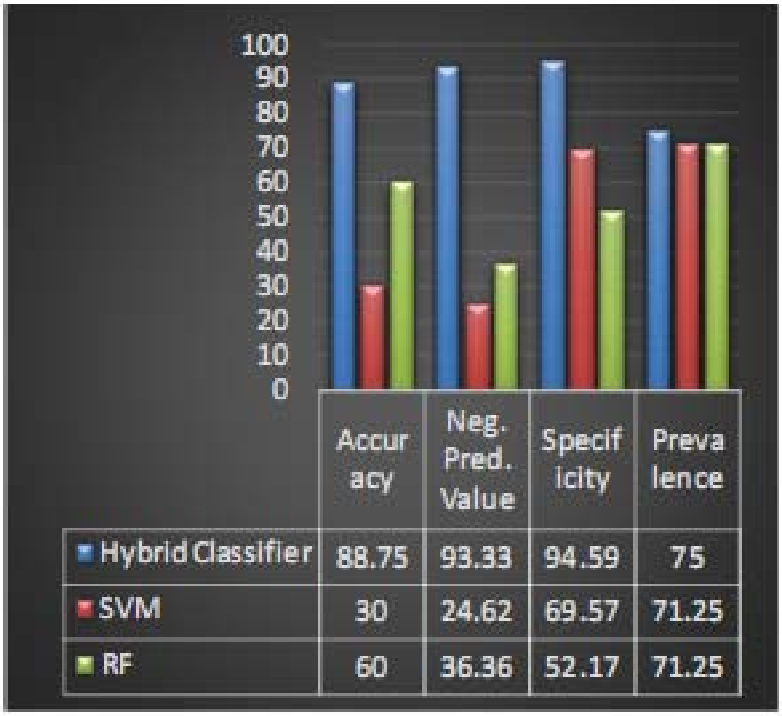


Figure 6: Figure 3 :

1

id	higherGeography	continent	family	scientificName		decimalLatit ude	decimalLongi tude	genus	specificEpi thet
2759 86	North America, GREENLAND	North America	Lycoperdac eae	Calvatiaarctica		72	-40	Calvati a	arctica
3333 01	North America,	North America	Ericaceae	Empetrumeamesii Fernald &Wiegand		52	-56	Empetr um	eamesii
2717 58	North America,	North America	Ranuncula ceae	Thalictrum Greene	terrae-novae	52	-56	Thalictr um	terrae-novae

[Note: A]

Figure 7: Table 1 :

2

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

[Note: b) Result AnalysisClassification of the spatial data sets can be represented as a confusion or error matrix view as shown in]

Figure 8: Table 2 :

III

Figure 9: Table III .

3

Real group	Classification result	
	North America	South America

[Note: North America True Negative(TN) False Positive(FP) South America False Negative(FN) True Positive(TP)]

Figure 10: Table 3 :

5

Prediction	Reference South America North America
South America 8	7
North America 49	16

Figure 11: Table 5 :

6

Prediction	Reference South America North America
South America 36	11
North America 21	12

Figure 12: Table 6 :

.1 Acknowledgment

We direct our frank appreciativeness to the Field Museum of Natural History(Botany)-Seed Plant Collection (accessed through GBIF data portal, http://data.gbif.org/datasets/resource/14346,2013-06-03) for providing us with different seed plant data sets. We also thank ANU university for providing all the support in the work conducted.

[Rumpf et al.] , Till Rumpf , Christoph Römer , Martin Weis , Markus Sökefeld , Roland Gerhards , Lutz Plümer . http://dx.doi.org/10.1 (Sequential Pages89-96, ISSN0168-1699)

[Dec] , Dec . 2012doi10.1109/NUICONE.2012.6493213.

[(2012)] , 10.5815/ijitcs.2012.07.06. July 2012. (Published Online)

[Yeh et al. (2012)] *A hybrid KMV model, random forests and rough set theory approach for credit rating, Knowledge-Based Systems*, Ching-Hiang Yeh , Chih-Yu Fengyilin , Hsu . 10.1016/j.knosys.2012.04.004. http://dx.doi.org/10.1016/j.knosys.2012.04.004 September 2012. 33. (Pages 166-172, ISSN0950-7051)

[Elbasiony et al. (2013)] 'A hybrid network intrusion detection framework based on random forests and weighted k-means'. Reda M Elbasiony , Elsayed A Sallam , Tarek E Eltobely , Mahmoud M Fahmy . 10.1016/j.asej.2013.01003. http://dx.doi.org/10.1016/j.asej.2013.01003 *A in Shams Engineering Journal* 2090-4479. Available online 7 March 2013.

[Kumar (2010)] *A hybrid SVM based decision tree, Pattern Recognition*, M Kumar , M . 10.1016/j.patcog.2010.06.010. http://dx.doi.org/10.1016/j.patcog.2010.06.010 December 2010. 43 p. .

[Zhang and Zhang ()] 'A local boosting algorithm for solving classification problems'. Chun-Xia Zhang , Jiang-She Zhang . 10.1016/j.csda.2007.06.015. http://dx.doi.org/10.1016/j.csda.2007.06.015 *Computational Statistics & Data Analysis* 0167-9473. 10 January 2008. 1928-1941. 52 (4) .

[Cho (2013)] 'A support vector machine approach to CMOS-based radar signal processing for vehicle classification and speed estimation'. Hsun-Jung Cho , Ming-Tetseng . 10.1016/j.mcm.2012.11.003. http://dx.doi.org/10.1016/j.mcm.2012.11.003 *Mathematical and Computer Modelling* July 2013. 58 (2) . (Pages 438-448,ISSN0895-7177)

[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. .

[Gray et al. (2013)] 'Alexander Hammers, Daniel Rueckert, for the Alzheimer's Disease Neuroimaging Initiative, Random forest-based similarity measures for multimodal classification of Alzheimer's disease, NeuroImage'. Katherine R Gray , Paul Aljabar , Rolf A Heckemann . 10.1016/j.neuroimage.2012.09.065. http://dx.doi.org/10.1016/j.neuroimage.2012.09.065 *Pages 167-175, ISSN10538119*, 15 January 2013. 65.

[Rodriguez-Galiano et al. (2012)] 'An assessment of the effectiveness of a random forest classifier for land-cover classification'. V F Rodriguez-Galiano , B Ghimire , J Rogan , M Chica-Olmo , J P Rigol-Sanchez . 10.1016/j.isprsjprs.2011.11.002. http://dx.doi.org/10.1016/j.isprsjprs.2011.11.002 *ISPRS Journal of Photogrammetry and Remote Sensing* January 2012. 67. (Pages 93-104, ISSN0924-2716)

[Zhang et al. (2008)] 'An efficient modified boosting method for solving classification problems'. Chun-Xia Zhang , Jiang-She Zhang , Gai-Ying Zhang . 10.1016/j.cam.2007.03.003. http://dx.doi.org/10.1016/j.cam.2007.03.003 *Journal of Computational and Applied Mathematics* 0377- 0427. 1 May 2008. 214 (2) p. .

[Kuncheva et al. (2002)] *An experimental study on diversity for bagging and boosting with linear classifiers, Information Fusion*, L I Kuncheva , M Skurichina , R P W Duin . 10.1016/S1566-2535(02. http://dx.doi.org/10.1016/S1566-2535(02 December 2002. 3 p. .

[Yugal Kumar and Sahoo ()] 'Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA'. G Yugal Kumar , Sahoo . *I.J. Information Technology and Computer Science* 2012. 7 p. .

[Xiao et al. (2013)] 'Bagging and Boosting statistical machine translation systems'. Tong Xiao , Jingbo Zhu , Tongran Liu . 10.1016/j.artint.2012.11.005. http://dx.doi.org/10.1016/j.artint.2012.11.005 *Artificial Intelligence* February 2013. 195. (Pages 496-527, ISSN0004-3702)

[Zitouni et al. (2003)] 'Boosting and combination of classifiers for natural language call routing systems'. Imed Zitouni , Hong-Kwang Jeff Kuo , Chin-Hui Lee . 10.1016/0167-6393(03. http://dx.doi.org/10.1016/0167-6393(03 *Speech Comm unication* November 2003. 41 (4) p. . (Pages 647-661, ISSN0167-6393)

[Tanha et al. (2013)] 'Boosting for Multiclass Semi-Supervised Learning'. Jafar Tanha , Maartenvan Someren , Hamideh Afsarmanesh . 10.1016/j.patrec.2013.10.008. http://dx.doi.org/10.1016/j.patrec.2013.10.008 *Pattern Recognition Letters* 21 October 2013. p. .

[Techo et al. ()] 'Boosting-based ensemble learning with penalty profiles for automatic Thai unknown word recognition'. Jakkrit Techo , Cholwich Nattee , Thanaruk Theeramunkong . 10.1016/j.camwa.2011.11.062. http://dx.doi.org/10.1016/j.camwa.2011.11.062 *Computers &Mathematics with Applications* March 2012, Pages 1117-1134, ISSN08981221. 63 (6) .

[Idris et al. ()] 'Churn prediction in telecom using Random Forest and PSO based data balancing in combination with various feature selection strategies'. Adnan Idris , Muhammad Rizwan , Asifullah Khan . 10.1016/j.compeleceng.2012.09.001. http://dx.doi.org/10.1016/j.compeleceng.2012.09.001 *Pages 1808-1819, ISSN0045-7906*, Novem ber 2012. 38.

[Naidoo et al. (2012)] 'Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment'. L Naidoo , M A Cho , R Mathieu , G Asner . 10.1016/j.isprsjprs.2012.03.005. http://dx.doi.org/10.1016/j.isprsjprs.2012.03.005 *ISPRS Journal of Photogrammetry and Remote Sensing* April 2012. 69 p. .

[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'. Miao Liu , Mingjun Wang , Duo Junwang , Li . 10.1016/j.snb.2012.11.071. http://dx.doi.org/10.1016/j.snb.2012.11.071 *Pages 970-980, ISSN0925-4005*, February 2013. 177.

[Liu et al. ()] '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* 0925-4005. February2013. 177 p. .

[Shao and 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, Volume70*, Yang Shao , Ross S Lunetta . 10.1016/j.isprsjprs.2012.04.001. http://dx.doi.org/10.1016/j.isprsjprs.2012.04.001 June2012, Pages78-87,ISSN0924-2716.

[Chau et al.] *Convex and concave hulls for classification with support vector machine, Neuro computing*, Asdrúbal López Chau , Xiaoou Li , Wen Yu . 122 p. 25.

[Ruano-Ordás et al. ()] 'Effective scheduling strategies for boosting performance on rule-based spam filtering frameworks'. D Ruano-Ordás , J Fdez-Glez , F Fdez-Riverola , J R Méndez . 10.1016/j.jss.2013.07.036. http://dx.doi.org/10.1016/j.jss.2013.07.036 *Journal of Systems and Software* 0164-1212. Volume86,Issue12, December2013,Pages 3151-3161.

[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

[Zintzaras and Kowald] 'Forest classification trees and forest support vector machines algorithms: Demonstration using microarray data'. Elias Zintzaras , Axel Kowald . 10.1016/j.compbiomed.2010.03.006. http://dx.doi.org/10.1016/j.compbiomed.2010.03.006 *Computers in Biology and Medicine* (Pages 519-524, ISSN0010-4825)

[Yeh et al. (2013)] 'Going-concern prediction using hybrid random forests and rough set approach'. Ching-Chiang Yeh , Der-Jang Chi , Yi-Rong Lin . 10.1016/j.ins.2013.07.011. http://dx.doi.org/10.1016/j.ins.2013.07.011 *Information Sciences* 0020-0255. August 2013.

[Aguilera et al. (2010)] 'Hybrid Bayesian network classifiers: Application to species distribution models'. P A Aguilera , A Fernández , F Reche , R Rumí . 10.1016/j.envsoft.2010.04.016. http://dx.doi.org/10.1016/j.envsoft.2010.04.016 *Environmental Modelling & Software, Volume25, Issue12*, December 2010. p. .

[Björnwaske et al. (2012)] 'imageRF -A user-oriented implementation for remote sensing image analysis with Random Forests'. Sebastian Björnwaske , Carsten Vander Linden , Benjamin Oldenburg , Jakimow . 10.1016/j.envsoft.2012.01.014. http://dx.doi.org/10.1016/j.envsoft.2012.01.014 *Environmental Modeling & Software*, (Andreas Rabe, Patrick Hostert) July 2012, Pages192-193. 35.

[Löw et al. (2013)] *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 November 2013. p. .

[Borra and Di Ciaccio (Volume38, Issue4, 28February2002. Pages 407-420,ISSN01679473)] 'Improving nonparametric regression methods by bagging and boosting'. Simone Borra , Agostino Di Ciaccio . 10.1016/S0167-9473(01. http://dx.doi.org/10.1016/S0167-9473(01 *Computational Statistics &Data Analysis* Volume38, Issue4, 28February2002. Pages 407-420,ISSN01679473. p. .

[Rodríguez-Galiano et al. ()] 'Incorporating Spatial Variability Measures in Land-cover Classification using Random Forest'. V F Rodríguez-Galiano , F Abarca-Hernández , B Ghimire , M Chica-Olmo , P M Akinson , C Jeganathan . 10.1016/j.proenv.2011.02.009. http://dx.doi.org/10.1016/j.proenv.2011.02.009 *Procedia Environmental Sciences* 1878- 0296. Volume3,2011. p. .

[Özyer et al. (2007)] 'Intrusion detection by integrating boosting genetic fuzzy classifier and data mining criteria for rule prescreening'. Tansel Özyer , Reda Alhajj , Ken Barker . 10.1016/j.jnca.n2005.06.002. http://dx.doi.org/10.1016/j.jnca.n2005.06.002 *Journal of Network and Computer Applications* January 2007. 30 (1) . (Pages 99-113, ISSN1084-8045)

[Rajasekhar et al.] 'Magnetic resonance brain images classification using linear kernel based Support Vector Machine'. N Rajasekhar , S J Babu , T V Rajinikanth . *2012Nirma University International Conference on*, 5 p. 68. (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 *ISPRS Journal of Photogrammetry and Remote Sensing* November 2006. 61 (2) . (Pages 125-133,ISSN0924-2716)

[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, Volume40, Issue6*, 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, Pages1925-1934.

[Topouzelis and Psyllos (2012)] 'Oil spill feature selection and classification using decision tree forest on SAR image data'. Konstantinos Topouzelis , Apostolos Psyllos . 10.1016/j.isprsjprs.2012.01.005. http://dx.doi.org/10.1016/j.isprsjprs.2012.01.005 *ISPRS Journal of Photogrammetry and Remote Sensing* 0924-2716. March 2012. 68 p. .

[Pages 198-209,I SSN 0925-2312 (2013)] 10.1016/j.neucom.2013.05.040. http://dx.doi.org/10.1016/j.neucom.2013.05.040 *Pages 198-209,I SSN 0925-2312*, December 2013.

[Shataeea et al. ()] 'Plot-level Forest Volume Estimation Using Airborne Laser Scanner and TM Data, Comparison of Boosting and Random Forest Tree Regression Algorithms'. Shaban Shataeea , Holger Weinaker , Manoucher Babanejad . 10.1016/j.proenv.2011.07.013. http://dx.doi.org10.1016/j.proenv.2011.07.013 *Procedia Environmental Sciences* 1878- 0296. 2011. 7 p. .

[Pino-Mejías et al. (2010)] 'Predicting the potential habitat of oaks with data mining models andtheR 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

[Song Feng (2012)] 'QBoost: Predicting quantiles with boosting for regression and binary classification'. Zheng Song Feng . 10.1016/j.eswa.2011.06.060. http://dx.doi.org/10.1016/j.eswa.2011.06.060 *Expert Systems with Applications* 0957- 4174. 1 February 2012. 39 (2) p. .

[Rahim et al. ()] N , Abdul Rahim , M P Paulraj , A H Adom . 10.1016/j.proeng.2013.02.054. http://dx.doi.org/10.1016/j.proeng.2013.02.054 *Adaptive Boosting with SVM Classifier for Moving Vehicle Classification, Procedia Engineering*, 2013. 53 p. .

[Özçift (2011)] 'Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis'. Akin Özçift . 10.1016/j.compbiomed.2011.03.001. http://dx.doi.org/10.1016/j.compbiomed.2011.03.001 *Computers in Biology and Medicine* May 2011. 41 (5) . (Pages 265-271 ,ISSN0010-4825)

[Pall Oskar Gislason et al. (2006)] 'Random Forests for land cover classification'. Jon Atli Pall Oskar Gislason , Johannes R Benediktsson , Sveinsson . 10.1016/j.patrec.2005.08.011. http://dx.doi.org/10.1016/j.patrec.2005.08.011 *Pattern Recognition Letters* 0167-8655. March 2006. 27 (4) p. .

[Xiao et al. (2013)] 'Sleep stages classification based on heart rate variability and random forest'. Meng Xiao , Hong Yan , Jinzhong Song , Yuzhou Yang , Xianglin Yang . 10.1016/j.bspc.2013.06.001. http://dx.doi.org/10.1016/j.bspc.2013.06.001 *Biomedical Signal Processing and Control* November 2013. 8. (Pages 624-633, ISSN1746-8094)

[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. 49 p. .

[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. p. .

[Abedi et al. ()] '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* 0098-004. September2012. 46 p. .

[Bosch et al. (2013)] 'Support vector machine under uncertainty: An application for hydroacoustic classification of fish-schools in Chile'. Paul Bosch , Julio López , Héctor Ramírez , Hugo Robotham . 10.1016/j.eswa.2013.01.006. http://dx.doi.org/10.1016/j.eswa.2013.01.006 *Expert Systems with Applications* August 2013. 40 (10) . (Pages 4029-4034,ISSN0957-4174)

[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)

[Meyer (2012)] *Support Vector Machines The Interface to lib svm in package e1071*, David Meyer . September, 2012. Austria. Technische University of Wien

[Yang et al. (2013)] 'The one-against-all partition based binary tree support vector machine algorithms for multi-class classification'. Xiaowei Yang , Qiaozhen Yu , Lifang He , Tengjiao Guo . 10.1016/ j.neucom.2012.12.048. *Neurocomputing* 0925-2312. 3 August 2013. 113 (1-7) .

[Martínez-Muñoz and Suárez (2007)] 'Using boosting to prune bagging ensembles'. Gonzalo Martínez-Muñoz , Alberto Suárez . 10.1016/j.patrec.2006.06.018. http://dx.doi.org/10.1016/j.patrec.2006.06.018 *Pattern Recognition Letters* 1 January 2007. 28. (Pages 156-165, ISSN0167-8655)

[Genuer et al. (2010)] 'Variable selection using r random forests'. Robin Genuer , Jean-Michel Poggi , Christine Tuleau-Malot . 10.1016/j.patrec.2010.03.014. http://dx.doi.org/10.1016/j.patrec.2010.03.014 *Pattern Recognition Letters* 15 October 2010. 31 (14) . (Pages 2225-2236, ISSN0167-8655)

[Hu et al. (2012)] *XiuliSi, Fish species classification by color, texture and multi-class support vector machine using computer vision, Computers and Electronics in Agriculture*, Jing Hu , Daoliang Li , Qingling Duan , Yueqi Han , Guifen Chen . 10.1016/j.compag.2012.07.008. http://dx.doi.org/10.1016/j.compag.2012.07.008 October 2012. 88. (Pages 133-140,ISSN0168-1699)