Novel RVM Approach to Structuring and Classifying Epidemic Outbreak Data

Sunaina Sharma
Student, Department of I.T
UIET, Panjab University
Chandigarh, India

Veenu Mangat
Assistant Professor, Department of I.T.
UIET, Panjab University
Chandigarh, India

ABSTRACT
Classifying this indefinite big data, is computationally intensive as a large amount of data is related with an existential probability of undefined or undetermined values of raw data. Classifying is hindered by a large amount of data from various sources. RVM, a Bayesian formulation of the linear model both for classification and regression, has lately involved a lot of interest in the research community. The paper aims at learning kernelized RVM classifier to evaluate Ebola virus outbreak, using generalization error, intra class separability, missing probability Pi is compared to SVM.RVM relevance impact with other epidemic diseases of Ebola Virus is also compared.

Keywords
classification, relevance vector machine, support vector machine, Naive Bayes, neural network, generalization error, intra class separarbility, missing probability , Predictive value imputation , distributed based imputation

1. INTRODUCTION
Assigning an object to a certain class based on its similarity to previous examples of other objects. Can be done with reference to original data or based on a model of that data. It consists of instance, concept, target concept, hypothesis, sample, candidate, testing.

The Instance is an input or value or set of values or whatever used to describe input. The concept is the kind of function. It maps inputs to an output. Take some input that is instance and mapping it to an output, is a concept. It is an idea that describe a set of things. The target concept is the thing that is we trying to find, the actual answer or a particular thing that we are finding. The hypothesis class all possible classification functions that could be defined. The sample is the training set. Candidate it the concept that we think might be the target concept. The testing set check whether the candidate concept is right is wrong.

1.1 Types of Classification
1.1.1 Neural Network
Artificial neural networks are parallel computational models (unlike our computers, which have a single processor to collect and display information). These networks are commonly made up of multiple simple processors which are able to act in parallel alongside one another to model changing systems. This parallel computing process also enables faster processing and computation of solutions. Neural networks follow a dynamic computational structure and do not abide by a simple process to derive the desired output. The basis for these networks originated from the biological neuron and neural structures - every neuron takes in multiple unique inputs and produces one output. Similarly, in neural networks, different inputs are processed and modified by a weight, or a sort of equation that changes the original value. The network then combines these different weighted inputs with reference to a certain threshold and activation function and gives out the final value. Neural networks are a new concept whose potential we have just scratched the surface of. They may be used for a variety of different concepts and ideas and error correction during the testing phase. By properly minimizing the error, these multi-layered systems may be able to one day learn and conceptualize ideas alone, without human correction.

1.1.2 Support Vector Machine
In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis [1]. The Support Vector Machine classifier is widely used in bioinformatics (and other disciplines) due to its highly accurate; able to calculate and process the high dimensional data such as gene expression in modelling diverse sources of data .SVMs belong to the general category of kernel methods and a kernel method is an algorithm that depends on the data only through dot-products. This is the case; the dot product can be replaced by a kernel function which computes a dot product in some possibly high dimensional feature space. It has two advantages: First; the ability to generate nonlinear decision boundaries using methods designed for linear classifiers and second; the use of kernel functions allows the user to apply a classifier to data that have no obvious fixed-dimensional vector space representation [2]. SVM is being used successfully in many real-world problems as a classifier like gait recognition, text (and hypertext) categorization, image classification, bioinformatics (Protein and Cancer classification)

1.1.3 Relevance Vector Machine
The relevance vector machine (RVM) uses an approach named as sparse Bayesian modeling approach which is used to obtain parsimonious solutions for regression and classification. It is basically a machine learning technique which enables sparse classification by linear weights of fixed small size functions from a large number of potential candidates. It has a similar functional form to that of SVM i.e. support vector machine but to add to it, RVM has probabilistic classification property. Comparatively to SVM, the Bayesian approach of RVM avoids set of free parameters that usually requires cross validation post optimization. Characteristically, our predictions are based on upon some function y(x) defined over the input space, and `learning' is the process of inferring (perhaps the parameters of) this function. A flexible and popular set of candidates for y(x) is that of the form [3]

$$y(\mathbf{x};\mathbf{w}) = \sum_{i=1}^{M} w_i \phi_i(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) \quad (1)$$

where the output is a linearly-weighted sum of M, generally nonlinear and fixed, basis functions Φ(x)=(Φ 1(x); Φ 2 (x);…;-M (x))^{T}. Exploration of functions of the type (1) is facilitated since the adjustable parameters (or `weights') w = (w1; w2;:::;wM )T appear linearly, and to estimate `good' values for those parameters [3]. Thus, it's a Bayesian approach to SVM, that is, it operates over distributions and outputs a PDF of scores instead of a point estimate. RVMs and techniques such as Platt Scaling are useful for domains where knowing the uncertainty can allow to do interesting things.

1.1.4 Naive Bayes

Naive Bayes assumes that all the features are conditionally independent of each other. This, therefore, permits us to use the Bayesian rule for probability. Usually this independence assumption works well for most cases if even in actuality they are not really independent. Naive Bayes and ANNs have different performance characteristics with respect to the amount of training data they receive. The Naive Bayes classifier has been shown to perform surprisingly well with very small amounts of training data that most other classifiers, and especially ANNs, would find significantly insufficient. As a result, if you find yourself with a small amount of training data Naive Bayes would be a good bet. However, the two classifiers also behave differently on the other end of the spectrum, when provided with large amounts of training data. As it is fed increasing quantities of training data, the performance of the Naive Bayes classifier plateaus above a certain threshold. Its simplicity prevents it from benefiting incrementally from training data past a certain point. Naive Bayes Classifier is based on the fundamental ('Naive') assumption of independence between every pair of features, i.e. all input variables are stochastically independent of each other. If that assumption is not true (there exist correlations between input variables) then it can impact the accuracy of the Naive Bayes classifier.

2. LITERATURE REVIEW

1. Fatemeh Sheikholesalmi and Georgios B. Giannakis [4] The paper proposed an algorithm for the joint online subspace learning and kernel-based SVM classification approach. The algorithm is of sequential high dimensional data vector applications streaming big data. The algorithm handles the missing entries and leverages the low dimensional data matrix, for the representation of incomplete vectors .Thus, classifiers are developed in the lesser dimension of the projection coefficients. The numerical result and performance analysis depicts the significant effectiveness of the proposed algorithm.

2. Geng Fan, Dengwu Ma, Xiaoyan Qu, Xiaofeng Lv [5] (2012) The paper shows for good performance of RVM is achieved by correct selection of kernel function and its parameters. A multi-scale RVM classification approach is developed to overcome the limitation of single kernel based RVM, which is based on intelligent optimization. Quantum- behaved particle swarm optimization (QPSO) algorithm is used for kernel parameters for better robust ability and multiple Gaussian kernels combined by linear weighting is used ,which shows better results in context to classification accuracy than a particular single kernel. Currently, the Gaussian kernel is the special kernel. However, a particular kernel possibly will not remain always appropriate for complex classification. The learning approach of multiple kernel functions have multi-scale expression ability. The paper depicts that to process complex classification problems with uniform model form and comprehensive scale selections, usage of multi kernel method is benefitted.

3. Carson Kai-Sang Leung, Fan Jiang [6] (2014) depicted in their paper an experiment to diminish the additional calculations in mining. To mine indefinite Big data for frequent patterns with an adequate user-listed anti-monotonic constraints, a data science solution that uses MapReduce is proposed. To mine interesting patterns many experiments are preformed and those show the efficiency of data science solution used to mine patterns from uncertain big data. Many mining algorithms focus on association and its analysis from precise databases. But currently the databases are uncertain and is more worsen when we move into the era of big data. To avoid waste of time and space in computing all the patterns a solution called BigAnt is proposed ,that allow users to express their interest in terms of anti-monotonic(AM) constrain.

4. Maytal Saar-Tsechansky,Foster Provost , Rich Caruana [7] (2007) the Paper compares several different methods of PVI and DBI imputations using reduced models. It shows hybrid approached to estimate the accuracy and to evaluate the balance between the imputation methods. Paper discussed the brief on missing values and alternative methods to them. Imputation outperform are conducted in context to decision tress. It has various evidence of generalizability with logistic regression

5. Jianfeng Fen [8] (1998) The paper discussed an open problem of learning machine for finding an exact form for the generalization error. With the help of theory of statics, the paper shows the exact form of generalization error of simple perceptron. A perceptron rule is easily implementation and its implementation is shown by simple additions and subtractions and with famous convergence theorem. It shows the possibility of rigorous generalization error analysis.

6. Junhui Wang and Xiaotong Shen [9] (2006) This paper gave a brief description on generalization error, it shows the methodology to estimate generalization error. The technique to optimize the measure is proposed. The outperforms of CV of both fixed and random designs is demonstrated via simulation also. Tuning and combining applications are discussed.

7. S. Ravikumar, A. Shanmugam [10] (2014) proposed a method for detection of white blood cell based on relevance vector machine. The methodology effectively works for WBC detection. The computational time of images is observed reduced. For increasing the efficiency the various feature vectors are used such as length, area, perimeter etc. The paper shows the increase in efficiency by 91% compare to other methods. A proposed method gives best result in medical field.

8. Stephen Kaisler,Frank Armour,J. Alberto Espinosa,William Money [11] The paper discussed about the increase in amount of data in this era. The authors have thrown light on the characteristics of Big Data and the issue and challenged faced in examining this large amount of data. Data management, its storage and cost is

the challenge majorly forward seen. The main concern is to identify the issues in technology, so as to overcome the storage and management obstructions.

3. PARAMETER USED

The missing probability is the parameter, which defines the probability of unclassified data after the classification process [20].

3.1 Missing Values

Missing values refers to the useful attribute features that may be missing either at induction time or at the prediction time. In predictive modeling applications, it is required to distinguish between two perspectives that are values unknown in training data or in the test case. Imputation is the most important and most basic approach to deal important value or feature missing for a particular instance. It estimates the missing value from the present data. There are two key types of imputation methodologies are Predictive value imputation (PVI), distributed based imputation (DBI). PVI the value to be used by model is estimated by predictive value imputation. DBI prediction will be based on this prediction and its conditional distribution of the missing value is estimated. Most of time, missing values occur completely at random.

The missing value probability for each data set and each treatment is the probability obtained by the difference between the classification accurateness for the treatment and (as a baseline) the accurateness attained if all features had been recognized both for training and for testing (the "complete" setting). The missing value probability (improvement) is given by 100·

$$(A_{CT} - A_{CK})/ A_{CK} \qquad (2)$$

Where A_{CK} is the prediction accuracy found in the complete setting, and A_{CT} denotes the accuracy obtained when a test instance includes missing values and a treatment T is applied [2].

3.2 Generalization Error

The generalized or generalization error for a machine learning scheme is defined as the capability of the machine learning method to learn and generalize the unseen or unstructured data. The distance is measured between the error on training dataset and test dataset and then is computed as the averaging factor for the each iteration during the whole learning process.

The generalization error has been tested containing the dataset of multiple entries of disease impacts, where each row defines one entry. The RVM kernel is training with different numbers of training and testing dataset sizes, with the standard deviation 0.1 versus π. The linear classifier (RVM) has been used for the classification purposes to reduce the generalization error.

3.3 Intra -Class Separability

Consider a data set consisting of N paired data values ($x_{n,1}$, $x_{n,2}$), for n = 1, ..., N. The intra-class correlation are originally proposed by Ronald Fisher is [13]:

$$r = \frac{1}{Ns^2} \sum_{n=1}^{N} (x_{n,1} - \bar{x})(x_{n,2} - \bar{x}) \qquad (3)$$

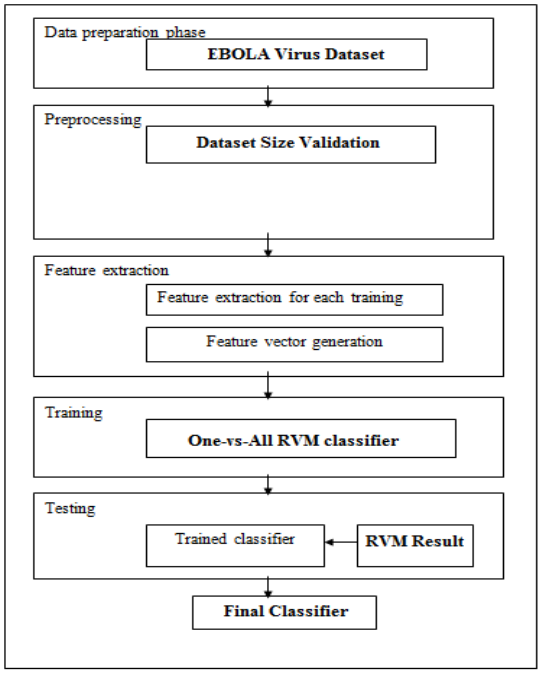
Where,

$$\bar{x} = \frac{1}{2N} \sum_{n=1}^{N} (x_{n,1} + x_{n,2}) \qquad (4)$$

$$s^2 = \frac{1}{2N} \left\{ \sum_{n=1}^{N} (x_{n,1} - \bar{x})^2 + \sum_{n=1}^{N} (x_{n,2} - \bar{x})^2 \right\} \qquad (5)$$

Later versions of this statistic used the degrees of freedom 2N −1 in the denominator for calculating s^2 and N −1 in the denominator for calculating r, so that s^2 becomes unbiased, and r becomes unbiased if s is known [14].

4. IMPLEMENTATION SCENARIO



5. RESULTS

5.1 Dataset Description

A dataset of Ebola virus death toll is taken. The dataset has many problems which lend themselves to useful lessons that can be applied to Big Data as well [15].

- WHO (World Health Organization), www.who.int/csr/disease/ebola/situation-reports/en/ (Data after Aug 29, 2014).
- CDC www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/index.html (before Aug 29, 2014).

Dataset Type	Number of Rows/Entries	Size on Disk
Dataset1-Cumulative Death cases	5706	3387898 bytes
Dataset2-Ebola Death rate cases	102691	60973698 bytes
Ebola Dataset3-death rate cases	1026901	609732498 bytes

5.2 RESULT ANALYSIS

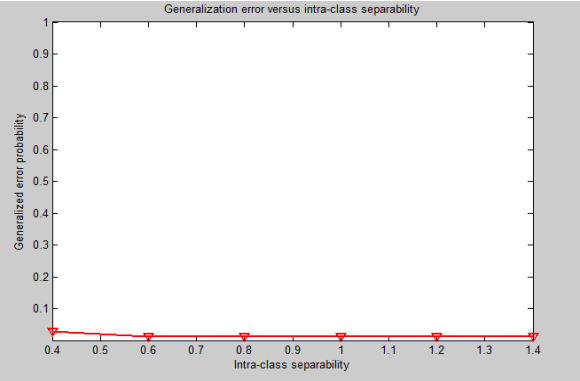


Figure 1 Generalization error versus intra-class separability of proposed model

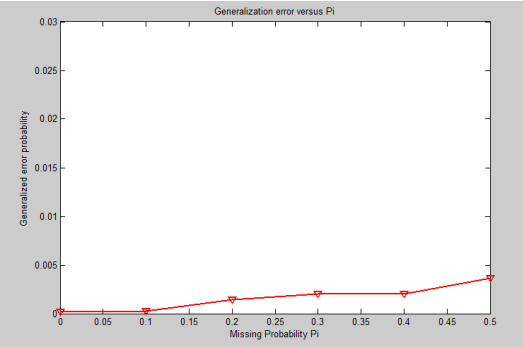


Figure 2 Generalization error versus Missing Probability Pi of proposed model

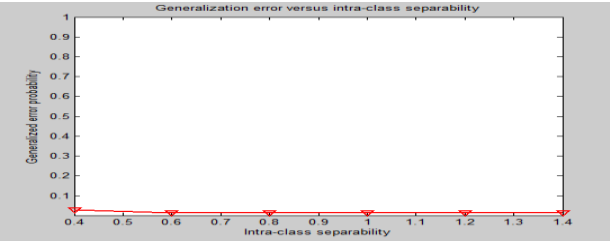


Figure 3 Generalization error and intra- class separability of proposed Model using RVM with large dataset size

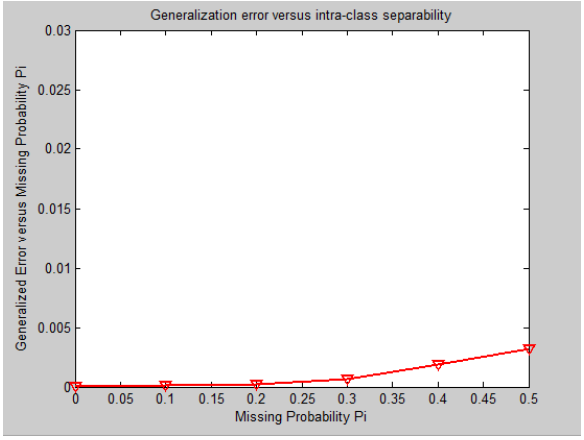


Figure 4 Generalization error and Missing probability Pi of proposed Model using RVM with large dataset size

Figure 3 and 4 shows generalisation error against Intra class separability and missing probability pi, respectively of larger dataset size. RVM shows lower generalization error probability against the missing probability (Pi) or intra-class separability**.**

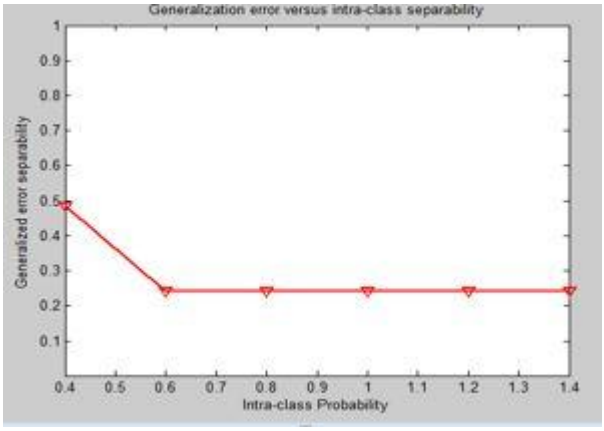


Figure 5 Generalization error versus intra-class separability of existing model using SVM

The generalization error has been compared against the intra-class separability under the results and discussion model. The intra-class separability is the ability of the classifier to correctly classify the given data into the various classes. The intra-class separability is inversely proportional to the generalization error. The higher intra-class separability and lower generalization errors depict the better performance of the classification model**.**

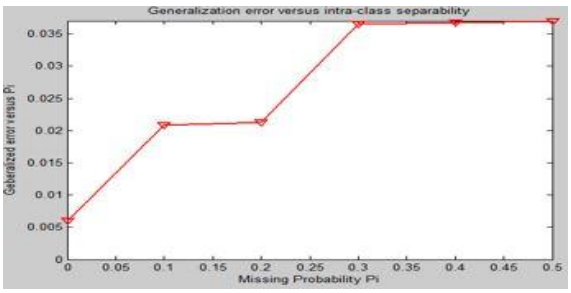


Figure 6 Generalization error versus missing probability Pi of existing model using SVM

The generalization error has been also compared against the missing probability or Pi. With the rise is the missing probability, the generalization error is also rising, which shows the directly proportional relationship between the missing probability (Pi) and Generalization error. The missing probability is the parameter, which defines the probability of unclassified data after the classification process.

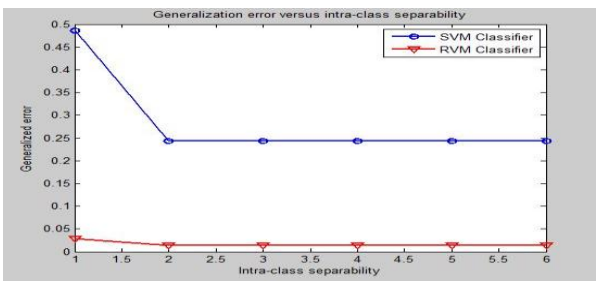


Figure 7 Comparison between RVM and SVM classifier showing Generalization error versus Intra class separability

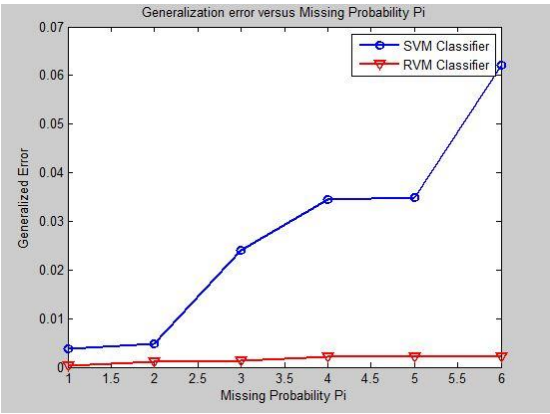


Figure 8 Comparison between RVM and SVM classifier showing Generalization error versus Missing probability Pi

Table 1 Evaluating Generalisation Error of SVM and RVM using Intra class separability

Intra-Class Separability	Generalization Error of SVM	Generalization Error of SVM
0.4	0.485714	0.028571
0.6	0.242857	0.014286
0.8	0.242857	0.014286
1	0.242857	0.014286
1.2	0.242857	0.014286
1.4	0.242857	0.014286

Intra class separability range lies between 0-10. In our record, minimum interval value is 0.4 and maximum is 1.4 for intra - class separability and accordingly Generalization error is estimated which lies between 0-1 is shown in above table.

Table 2 Evaluating generalisation error of SVM and RVM using Missing probability Pi

Missing probability Pi	Generalization Error of SVM	Generalization Error of SVM
0	0.00653	7.14E-05
0.1	0.014963	0.000104
0.2	0.025005	0.000218
0.3	0.028349	0.000621
0.4	0.028906	0.001848
0.5	0.051785	0.003169

Missing Probability range lies between 0-10. In our record, minimum interval value is 0 and maximum is 0.5 for intra class separability and accordingly Generalization error is estimated which lies between 0-1 is shown in above table.

6. CONCLUSION

The proposed model results have been obtained on the basis of various performance parameters. The result evaluation has been performed against the existing model, where the support vector machine (SVM) has been used for the purpose of data classification. The proposed model has been developed using the relevance vector machine (RVM) model. The major performance parameter of generalization error has been evaluated against the missing probability and intra-class separability. The proposed model with RVM has been performed atleast two times (2x) better than the existing system with SVM classifier i.e. It has been found that lower generalization error probability against the missing probability (Pi) or intra-class separability is seen. Hence, the experimental results of the proposed model have been proved the proposed model as the efficient model than the existing model.

7. FUTURE WORK

In the future, the proposed model can be enhanced using the amalgamation of effective feature descriptor with some effective classifier. The proposed model performance can be evaluated and compared against any other similar model on the appropriate dataset.

8. REFERENCES

[1] P. Jenifer Martina , P. Nagarajan , P. Karthikeyan," Hand Gesture Recognition Based Real-time Command System",International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology IJCSMC, Vol. 2, Issue. 4, April 2013, pg.295 – 299.

[2] Rapanjot Kaur, Gagangeet Singh Aujla,"Review on: Enhanced Offline Signature Recognition Using NeuralNetwork and SVM", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3648-3652

[3] Michael E. Tipping,"Sparse Bayesian Learning and the Relevance Vector Machine ",Journal of Machine Learning Research ,1 (2001) 211–244 Submitted 5/00; Published 6/01

[4] Fatemeh Sheikholesalmi and Georgios B. Giannakis," Online Subspace Learning and Nonlinear Classification of Big Data with Misses",Dept. of ECE and Digital Tech. Center, Univ. of Minnesota.

[5] Geng Fan, Dengwu Ma, Xiaoyan Qu, Xiaofeng Lv, " Multi-scale Relevance Vector Machine Classification Based on Intelligent Optimization", 2012 International Conference on Systems and Informatics (ICSAI 2012)

[6] Carson Kai-Sang Leung,Fan Jiang," A Data Science Solution for Mining Interesting Patterns from Uncertain Big Data", IEEE Fourth International Conference on Big Data and Cloud Computing, 2014,3-5 Dec. 2014 , pp.235 - 242 ,IEEE,10.1109/BDCloud.2014.136.

[7] Maytal Saar-Tsechansky ,Foster Provost , Rich Caruana," Handling Missing Values when Applying Classification Models",Journal of Machine Learning Research 8 (2007), pp. 1625-1657 Submitted 7/05; Revised 5/06; Published 7/07.

[8] Jianfeng Feng, "Generalization errors of the simple perceptron", J. Phys. A: Math. Gen. 31 (1998) 4037–4048. Printed in the UK PII: S0305-4470(98)86524-4

[9] Junhui Wang and Xiaotong Shen,"Estimation Of Generalization error :Random and fixed Inputs", Statistica Sinica 16(2006), pp.569-588.

[10] S. Ravikumar , A. Shanmugam,"WBC Image Segmentation and Classification Using RVM" , Applied Mathematical Sciences, Vol. 8, 2014, no. 45, 2227 - 2237

HIKARI Ltd, www.m-hikari.com http://dx.doi.org/10.12988/ams.2014.43191

[11] Stephen Kaisler Frank Armour J. Alberto Espinosa William Money, "Big Data: Issues and Challenges Moving Forward", 46th Hawaii International Conference on System Sciences (HICSS 2013) Subscribe Jan. 7, 2013 to Jan. 10, 2013 ISBN: 978-1-4673-5933-7 pp: 995-1004

[12] Maytal Saar-Tsechansky, "Handling Missing Values when Applying Classification Models",Journal of Machine Learning Research 8 (2007) 1625-1657 Submitted 7/05; Revised 5/06; Published 7/07

[13] https://en.wikipedia.org/wiki/Intraclass_correlation_coefficient

[14] Lu, Yi Qing, "Research on E-Government Model Based on Big Data", Advanced Materials Research, 2014

[15] http://www.kdnuggets.com/2014/10/ebola-analytics-data-science-lessons.html?