Influence of Machine Learning vs. Ranking Algorithm on the Critical Dimension

Divya Suryakumar, Andrew H. Sung, and Qingzhong Liu

***Abstract*—The critical dimension is the minimum number of features required for a learning machine to perform with "high" accuracy, which for a specific dataset is dependent upon the learning machine and the ranking algorithm. Discovering the critical dimension, if one exists for a dataset, can help to reduce the feature size while maintaining the learning machine's performance. It is important to understand the influence of learning machines and ranking algorithms on critical dimension to reduce the feature size effectively. In this paper we experiment with three ranking algorithms and three learning machines on several datasets to study their combined effect on the critical dimension. Results show the ranking algorithm has greater influence on the critical dimension than the learning machine.**

***Index Terms*—Critical dimension, feature selection, machine learning, ranking algorithm.**

I. Introduction

In datasets containing a large number of features, it is difficult to mine useful information. The complexity of analysis increases as the dataset is projected at a higher dimensional plane [1], [2]. For these reason, we usually reduce the features using feature reduction methods. The implication is that, datasets contain irrelevant features or attributes, which when eliminated can help achieve higher accuracy. Feature ranking and elimination, and subset selection are two ways in which feature reduction is traditionally performed. Feature ranking algorithms rank individual features using some metrics. Each feature is given a score based on factors such as correlation among some or all features. The features with a high score are ranked higher and those which do not meet an adequate score are eliminated. In subset selection method, random subsets are created from original feature set and the subset with the highest correlation coefficient among itself is considered as the best feature subset. Most feature subset algorithm select subsets based on greedy algorithm and some use an exhaustive search method with stop time defined. The main objective of feature selection is to improve the prediction performance, to provide faster and cost-effective predictor and better understand the correlation among data. The interesting fact about extracted features are that sometimes not all extracted features are individually useful; however, correlation of features itself is an intriguing question.

Manuscript received September 20, 2012; revised November 18, 2012. Support for this work received from ICASA (Institute for Complex Additive Systems Analysis) of New Mexico Tech and the National Institute of Justice, U.S. Department of Justice (Award No. 2010-DN-BX-K223) is gratefully acknowledged.

D. Suryakumar and A. H. Sung are with the Department of Computer Science and Engineering, New Mexico Tech, Socorro, NM 87801, USA (e-mail: divyasuryakumar@gmail.com)

Q. Liu is with the Department of Computer Science and Engineering, Sam Houston State University, Huntsville, Texas 77341, USA (e-mail: liu@shsu.edu)

II. Feature Ranking Algorithms

There are many ranking methods available and are used for different kinds or purposes [3]. The ranking algorithms used in our experiments are discussed below

A. Chi-squared Ranking

This evaluates the worth of an attribute by computing the value of the chi-squared statistic [4],[5] with respect to the class. There are several ways in which a chi-squared statistics is used; one such is using a contingency table. To rank features, we look at the chi-square distribution table against its degree of freedom value to find the corresponding probability level α; search method ranker, ranks these based on higher probability.

B. SVM Attribute Evaluator

This method, evaluates the worth of an attribute by using an SVM classifier [6]. Attributes are ranked by the square of the weight assigned by the SVM. Attribute selection for multiclass problems is handled by ranking attributes for each class separately using a one vs. all method and then "dealing" from the top of each pile to give a final ranking. This also uses a ranker search method to rank.

C. Cfs Attribute Evaluator

The Cfs evaluator evaluates the worthiness of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them [7]. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred. This uses a greedy step wise search method to rank.

III. Machine Learning Algorithms

There are several machine learning algorithms [8]. Three of the commonly used algorithms are used in this experiment.

A. Multilayer Perceptron Classifier

MLP is a feed forward Neural Network that uses back propagation algorithm. The MPL contains hidden layers and there can be any number of hidden layers. The weights change in the hidden layer and it is a black box approach. The first hidden layer of the helps the draw simple boundaries and the inner layers are used to draw complex boundaries. Hence, an MLP is very powerful algorithm if the right parameters

were chosen.

B. Naive Bayes Classifier

The NB Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, NB can often outperform more sophisticated classification methods. NB does not make use of only the prior information available, but also uses the likelihood of the instance.

C. Random Forest Classifier

Random forest (RF) is a learning machine and a trademark of Leo Breiman and Adele Cutler, and uses an ensemble of decision trees to make the final prediction [9]. A number of trees are grown and each represents one of the classes. If there are N cases in the training set then N sample cases are created at random with replacement, from the original dataset. This set is used as the training set for growing the trees. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing. Each tree is grown to the largest extent possible. There is no pruning. The error rate depends on correlation between any two trees in the forest and strength of each individual tree in the forest. After each tree is built, all of the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees.

IV. Critical Dimension

The term critical dimension of a dataset can be described as the minimum number of features required for a learning machine to perform prediction or classification with high accuracy [10]. It is an informal concept and empirical methods are used to determine the critical dimension in many datasets. Thus critical dimension of a dataset can be defined as that number (of features) where the performance of a specific learning machine would begin to drop significantly, and would not rise again when smaller number of features is used. Specifically, it is postulated that for a dataset there possibly exists a critical dimension μ which is a unique number for a specific machine learning and feature ranking combination. More clearly, let A = $\{a_1, a_2, \ldots, a_n\}$ be the feature set where $a_1, a_2, \ldots, a_n$ are listed in order of decreasing importance as determined by some feature ranking algorithm. Let $A_m \subseteq$ A contains the m most important features, i.e., $A_m = \{a_1, a_2, \ldots, a_m\}$ where m ≤ n. For a learning machine *M* and a feature ranking method *R*, we call μ (μ ≤ n) the critical dimension of [*M*, *R*], if whenever *M* uses feature set A_k with k ≥ μ the performance of *M* is ≥ T, where T represents a performance threshold deemed satisfactory; and whenever *M* uses less than μ features its performance drops below T; further, *M*'s performance from μ to μ−1 features decreases significantly. An example, the hypothyroid dataset was classified using SMO classifier. This dataset was ranked using Chi-squared ranking algorithm. The critical dimension was found to be 18. The table below shows the classification accuracies and other measurement for different feature sizes.

TABLE I: Results of Hypothyroid using SMO Classifier

Feature	TP rate	FP Rate	F Measure	ROC Area	Kappa statistic	Accuracy %
25	0.97	0.53	0.97	0.722	0.588	97.4
24	0.97	0.55	0.967	0.711	0.558	97.21
22	0.95	0.96	0.935	0.5	0	95.63
20	0.95	0.96	0.935	0.5	0	95.63
18	0.95	0.95	0.929	0.5	0	95.25
16	0.92	0.92	0.912	0.5	0	92.44
14	0.92	0.92	0.912	0.5	0	92.44
12	0.91	0.91	0.896	0.5	0	91.17
10	0.90	0.90	0.882	0.5	0	90.38
25	0.97	0.53	0.97	0.722	0.588	97.4

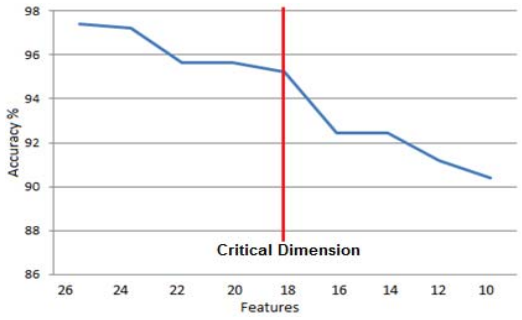

Fig. 1. Showing critical dimension at feature size 18.

TABLE II: Critical Dimension and Accuracies of Datasets

Dataset	R	Accuracy %								
		MLP			NB			RF		
		μ	At μ	All	μ	At μ	All	μ	At μ	All
WBCD	Chi	8	93.26	94.3	8	94.74	98.25	8	92.62	95.18
	SV	6	91.67	94.3	8	96.49	98.25	6	92.62	95.18
	Cfs	8	89.47	94.3	8	95.18	98.25	8	93.86	95.18
Hypothyroid	Chi	1 8	96.68	97.31	1 8	97	97.31	1 7	97.23	98.34
	SV	1	95.26	97.31	1	95.26	97.31	1	94.86	98.34
	M	6	26	31	6	26	31	6	86	34
	Cfs	1 8	95.26	97.31	1 6	95.26	97.31	1 6	95.86	98.34
Hand written	Chi	1 0	88.54	85.29	1 0	83.54	92.15	1 0	89.69	96.46
	SV	1	91.62	85.29	1	84.23	92.15	1	92.62	96.46
	M	1	62	29	1	23	15	1	69	46
	Cfs	1 6	96.77	85.29	1 6	89.69	92.15	1 5	92.85	96.46
Dermatology	Chi	2 2	95.16	98.38	2 2	95.97	99.19	2 4	96.77	98.39
	SV	1	92.75	98.38	1	94.35	99.19	2	97.58	98.39
	M	7	75	38	7	35	19	1	58	39
	Cfs	9	92.74	98.38	9	97.58	99.19	9	92.75	98.39

V. Method

The critical dimension is found for the four different datasets [11], [12]. A comparative study is carried out to find the influence of M, R on μ. Each of the four dataset was ranked by three different ranking algorithm and three different learning machines were used to classify them. The critical dimension is found as mentioned above. A total of 36 experiments were performed using four different datasets. Two of these are binary and two others are multiclass

classification datasets.

VI. Results

The table below shows the results of 36 different experiments for four different datasets. The critical dimension and the performance accuracies at μ and including all features are shown.

VII. Observations

The three machine learning algorithm (M1, M2 and M3) are studied. Let M1 be Multilayer Perceptron algorithm, M2 be Naïve Bayes and M3 be Random Forest algorithm. The following table shows the influence of M on μ.

TABLE III: Influence of M1 on Critical Dimension

Dataset	R	μ	Acc % At μ	Acc % All
WBCD	Chi	8	93.26	94.3
	SVM	6	91.67	94.3
	Cfs	8	89.47	94.3
Hypothyroid	Chi	18	96.68	97.31
	SVM	16	95.26	97.31
	Cfs	18	95.26	97.31
Handwritten	Chi	10	88.54	85.29
	SVM	11	91.62	85.29
	Cfs	16	96.77	85.29
Dermatology	Chi	22	95.16	98.38
	SVM	17	92.75	98.38
	Cfs	9	92.74	98.38

TABLE IV: Influence of M2 on Critical Dimension

Dataset	R	μ	Acc % At μ	Acc % All
WBCD	Chi	8	94.74	98.25
	SVM	8	96.49	98.25
	Cfs	8	95.18	98.25
Hypothyroid	Chi	18	97	97.31
	SVM	16	95.26	97.31
	Cfs	16	95.26	97.31
Handwritten	Chi	10	83.54	92.15
	SVM	11	84.23	92.15
	Cfs	16	89.69	92.15
Dermatology	Chi	22	95.97	99.19
	SVM	17	94.35	99.19
	Cfs	9	97.58	99.19

TABLE V: Influence of M3 on Critical Dimension

Dataset	R	μ	Acc % At μ	Acc % All
WBCD	Chi	8	92.62	95.18
	SVM	6	92.62	95.18
	Cfs	8	93.86	95.18
Hypothyroid	Chi	17	97.23	98.34
	SVM	16	94.86	98.34
	Cfs	16	95.28	98.34
Handwritten	Chi	10	89.69	96.46
	SVM	11	92.69	96.46
	Cfs	15	92.85	96.46
Dermatology	Chi	24	96.77	98.39
	SVM	21	97.58	98.39
	Cfs	9	92.75	98.39

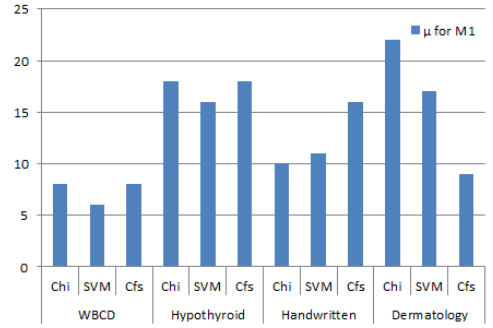

Fig. 2. Critical dimension using MLP and 3 ranking algorithm.

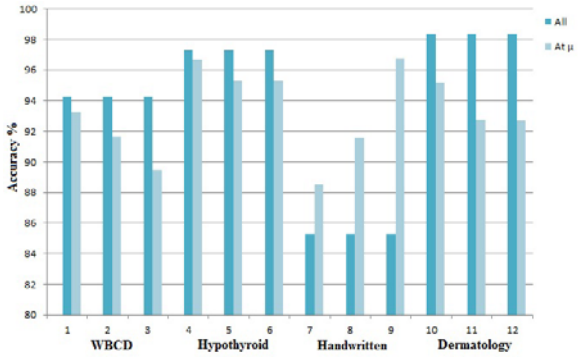

Fig. 3. Accuracies of all datasets using MLP.

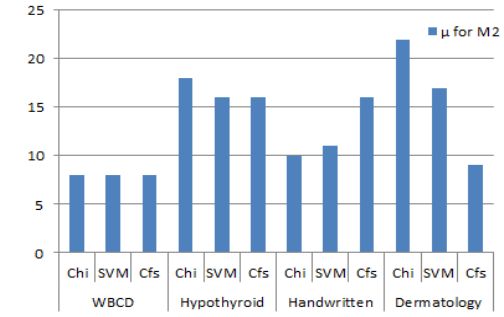

Fig. 4. Critical dimension using M2 and 3 ranking algorithm.

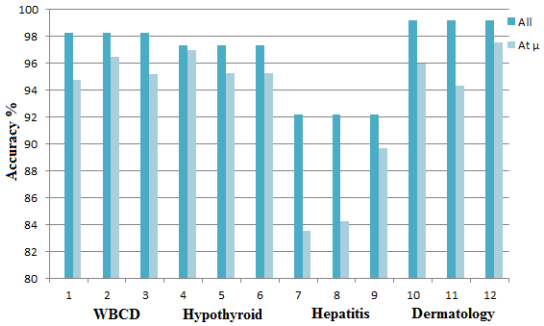

Fig. 5. Accuracies of all datasets using M2.

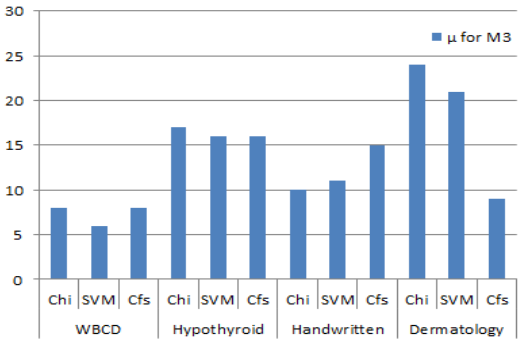

Fig. 6. Critical dimension using M3 and 3 ranking algorithm.

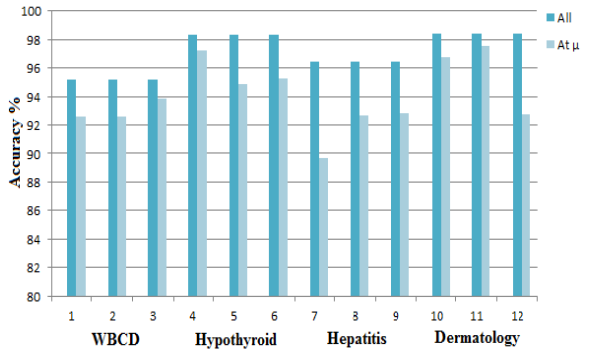

Fig. 7. Accuracies of all datasets using M3.

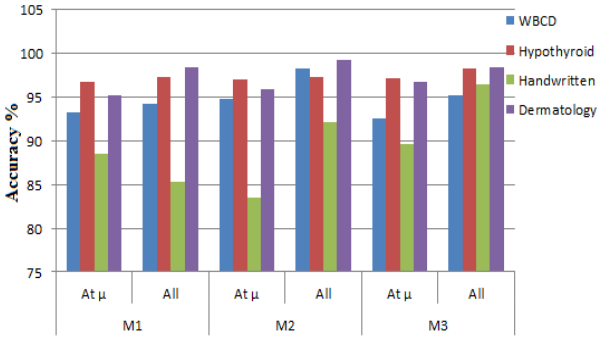

Fig. 8. Accuracy of datasets using R1 algorithm.

TABLE VI: CRITICAL DIMENSION AND ACCURACY USING R1

Dataset	Accuracy %								
	M1			M2			M3		
	μ	At μ	All	μ	At μ	All	μ	At μ	All
WBCD	8	93.26	94.3	8	94.74	98.25	8	92.62	95.18
Hypothyroid	18	96.68	97.31	18	97	97.31	17	97.23	98.34
Handwritten	10	88.54	85.29	10	83.54	92.15	10	89.69	96.46
Dermatology	22	95.12	98.38	22	95.97	99.19	24	96.77	98.39

TABLE VII: CRITICAL DIMENSION AND ACCURACY USING R2

Dataset	Accuracy %								
	M1			M2			M3		
	μ	At μ	All	μ	At μ	All	μ	At μ	All
WBCD	6	91.67	94.3	8	96.49	98.25	6	92.62	95.18
Hypothyroid	16	95.26	97.31	16	95.26	97.31	16	94.88	98.34
Handwritten	11	91.62	85.29	11	84.23	92.15	11	92.69	96.46
Dermatology	17	92.75	98.38	17	94.35	99.19	21	97.58	98.39

TABLE VIII: CRITICAL DIMENSION AND ACCURACY USING R3

Dataset	Accuracy %								
	M1			M2			M3		
	μ	At μ	All	μ	At μ	All	μ	At μ	All
WBCD	8	89.47	94.3	8	95.18	98.25	8	93.86	95.18
Hypothyroid	18	95.26	97.31	16	95.26	97.31	16	95.26	98.34
Handwritten	16	96.77	85.29	16	89.69	92.15	15	92.85	96.46
Dermatology	9	92.74	98.38	9	97.58	99.19	9	92.75	98.39

We can see from Fig. 2, Fig. 4 and Fig. 6 that the bar graph shows fluctuations, which means there is a difference in critical dimension. Fig.4 shows for WBCD dataset using M2 algorithm the critical dimension is the same, but not for other datasets. Fig.3, Fig. 5 and Fig. 7 show that there exists a critical dimension for all four datasets studied using 3 different learning machine algorithms.

The three ranking algorithms studied are Chi, SVM and Cfs algorithms. To study the influence of R on μ, we look at the following tables. Let R1, R2 and R3 are Chi, SVM and Cfs be the ranking algorithms respectively.

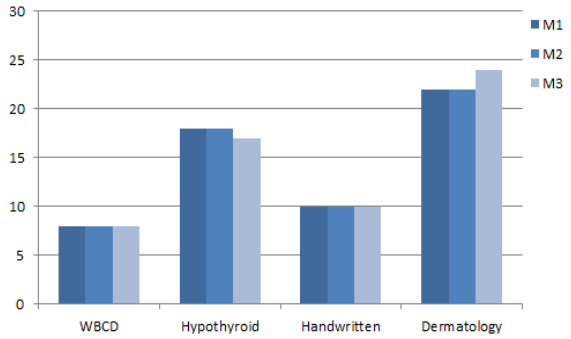

Fig. 9. μ for datasets using R1 algorithm.

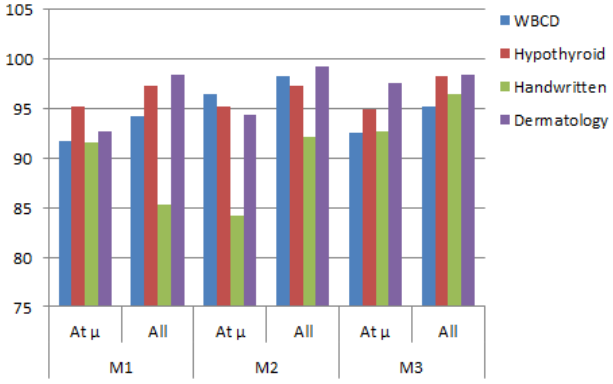

Fig. 10. Accuracy of datasets using R2 algorithm.

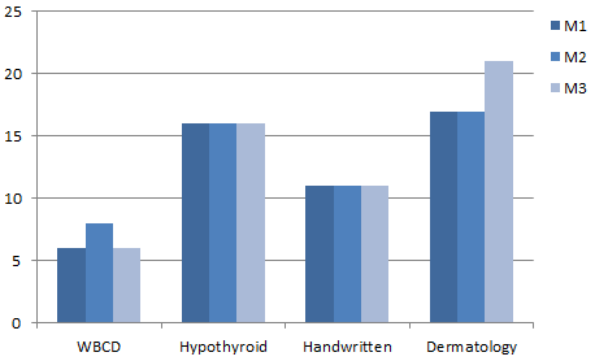

Fig. 11. μ for datasets using R2 algorithm.

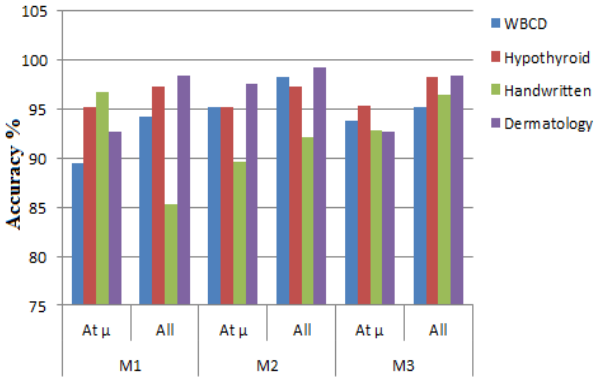

Fig. 12. Accuracy of datasets using R3 algorithm.

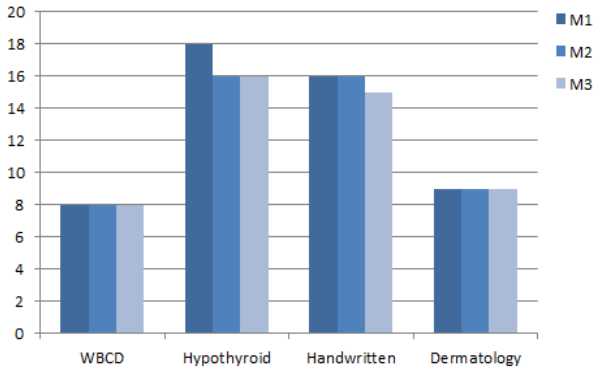

Fig. 13. μ for datasets using R3 algorithm.

Fig.8, fig.10 and fig.12 show us that there exists a critical dimension in the datasets studied for different learning machines and using three ranking algorithms. Fig.9 shows the results using R1 algorithm. We can see that for the WBCD and the handwritten datasets, the critical dimension is the same but for hypothyroid and dermatology dataset, one value is different. Fig.11 shows the results using R2 algorithm. We can see that for the hypothyroid and the handwritten dataset have the same critical dimension for all three M's. However, the WBCD and dermatology datasets show a little difference in critical dimension for one M. Similarly, fig.13 shows the results using R3 learning algorithm and we can see that WBCD and dermatology dataset show the same critical dimension for all three M's. Hypothyroid and handwritten datasets show a difference in critical dimension for one of the M.

Though there is a difference in the critical dimension of some of the datasets using the same R and different M's, when compared to Fig.2, Fig.4, and Fig.6 in which the difference in the critical dimensions are very different from each, we can conclude that R has a greater influence on μ than M on μ.

VIII. Conclusion

Three machine learning algorithm, multilayer perceptron, naive bayes and random forest and three ranking algorithms namely chi-squared feature ranking, support vector machine ranking and correlation based feature ranking methods were studied in 36 different combinations to find the influence of M and R on μ. It can be seen that though both M and R had an influence in determining μ, the influence of the ranking algorithm played a major role. It can be seen that by keeping the same learning machine algorithm, M1 and three different ranking algorithms R1, R2 and R3 changed the μ number of features. While keeping the same ranking algorithm R1 and three different machine learning algorithms M1, M2 and M3, the critical dimension did not very much. This indicates that μ highly varies as the ranking algorithm is changed. This gives us awareness that to find a low critical dimension number, an analytical search of different ranking algorithms with the same learning machine can be performed.

References

[1] L. Parson, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: A review," *SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, 2004.

[2] Z. Q. Hong and J. Y. Yang, "Optimal discriminant plane for a small number of samples and design method of classifier on the plane," *Pattern Recognition*, vol. 24, pp. 317-324, 1991.

[3] I. Guyon and A. Elisseeff, "An Introduction to variable and feature selection," *Journal of Machine Learning Research*, pp. 1157-1182, 2003.

[4] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles," *Lecture Notes in Computer Science*, pp. 106-115, 2006.

[5] X. Geng, T. Y. Liu, T. Qin, and H. Li, "Feature selection for ranking," SIGIR, ACM 1-58113-000-0/00/0004, 2007.

[6] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with SVMs," *Journal of Information Retrieval*, 2010.

[7] M. A. Hall and L. A. Smith, "Feature selection for machine learning through a correlation based approach to the Wrapper," FLAIRS Conference, 1999.

[8] K. J. Cios and L. A. Kurgan, "Machine learning algorithms inspired by the Work of Ryszard Spencer Michalski," *Advances in Machine Learning I, Studies in Computational Intelligence*, vol. 262, pp. 49-74, 2010.

[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[10] D. Suryakumar, A. H. Sung, and Q. Liu, "Critical Dimension in Data Mining," *IARIA Conference, Intelligent Systems Design and Applications*, pp. 481-486, ISBN: 978-1-61208-181-6, 2012.

[11] A. Frank and A. Asuncion, "Hypothyroid dataset," UCI Machine Learning Repository Irvine, CA: University of California, School of Information and Computer Science, 2010.

[12] UCI Repository. [Online]. Available: http://archive.ics.uci.edu/ml, 2010.

[13] D. P. Foster and R. A. Stine, "Variable selection in data mining: Building a predictive model for bankruptcy," *Journal of the American Statistical Association*, vol. 99, pp. 303-313, 2004.

Divya Suryakumar is a Ph.D. candidate from Computer Science Department, New Mexico Tech (NMT), USA. The author was born in India and did her undergraduate degree in Electronics and Communication and Engineering, Madras University, Chennai, India. Divya did her Masters in Computer Science with Information Technology option in NMT, Socorro, NM, USA in 2007 and is currently doing her research in data mining and bio-informatics. She has worked in NMT as a teaching assistant, research assistant and an instructor form 2005. She worked in Biomoda Inc., Albuquerque, NM, USA as a summer intern in 2008 and in Bioinformatics laboratory, Anna University as a summer intern in 2011. She has worked as a research assistant in the artificial intelligence laboratory in the Indian Institute of Technology (Madras), Chennai, India. Her research interests are in data mining, soft computing, artificial intelligence, medical-informatics and bio-informatics. Ms. Suryakumar is a member of IEEE. Her paper on "Determining the critical dimension" won the best paper award in IARIA, eknow 2012 conference held in Valencia, Spain.