The Critical Dimension Problem: No Compromise Feature Selection

Divya Suryakumar, Andrew H. Sung
Department of Computer Science and Engineering
New Mexico Institute of Mining and Technology
Socorro, New Mexico 87801, USA
{divya, sung}@cs.nmt.edu

Qingzhong Liu
Department of Computer Science
Sam Houston State University
Huntsville, Texas 77341, USA
liu@shsu.edu

Abstract— **The important feature selection problem has been studied extensively and a variety of algorithms has been proposed for data analysis and mining tasks in diverse applications. As the era of "big data" arrives, the development of effective techniques for identifying important features or attributes in very large datasets will be highly valuable in dealing with many of the challenges that come with it. This paper describes work in progress regarding a related general problem: for a given dataset, is there a "Critical Dimension" or minimum number of features that are necessary for achieving good results? In other words, for a dataset with many features, how many are truly relevant and important to be included in, say machine learning and/or data mining tasks to ensure that acceptable performance is achieved? Moreover, if a Critical Dimension indeed exists, how to identify the features that need to be included? The problem is first analyzed formally and shown to be intractable. An ad hoc method is then designed for obtaining approximate solution; next experiments are performed on a selection of datasets of varying sizes to demonstrate that for many datasets there indeed exist a Critical Dimension. The significance of the existence or lack thereof in datasets is explained.**

Keywords-machine learning; ranking; feature reduction; Critical Dimension; large datasets.

I. INTRODUCTION

One of the challenges of "big data" is how to reduce the size of data without losing information contained therein. In that regard, effective feature ranking and selection algorithms [1] can guide us in significantly reducing the size of the dataset by eliminating features that are insignificant, irrelevant, or useless. In some bio- or medical- informatics datasets, for example, the number of features can reach tens of thousands. This is partly due to the reason that many datasets constructed today for intended data mining purposes, without prior knowledge about what is to be specifically explored or derived from the data, likely have included measurable attributes that are actually insignificant or irrelevant, and inevitably resulting in large numbers of useless attributes (or features) that can be deleted to greatly reduce the size of datasets without any negative consequences in data analytics or data mining [6].

We investigate in this paper the general question: Given a dataset with n features, is there a Critical Dimension, or the smallest number of features that are necessary for a particular data mining application to ensure a minimal performance requirement? The term performance in this context means the overall accuracy of the training model. That is, any machine learning, statistical analysis, or data mining, etc. tasks performed on the dataset must include at least a number of features no less than the Critical Dimension – or it would not be possible to obtain acceptable results. This is a useful question to investigate since feature selection methods generally provide no guidance on the number of features to retain for a particular task; moreover, for many complex problems to which big data brings hope of breakthrough there is very little or no prior knowledge which may be otherwise relied upon in determining this number.

The question is analyzed in the next section and shown to be intractable. In Section 3, an ad hoc method is proposed as a first attempt to approximately solve the problem. In Section 4, experimental results on selected datasets are presented to demonstrate the existence of the Critical Dimension for most of them. Section 5 provides conclusions and discussions.

II. CRITICAL DIMENSION

The intuitive concept of the Critical Dimension of a dataset with n features is that there may exist, with respect to a specific "machine" M and a fixed performance threshold T, a unique number $\mu \leq n$ such that the performance of M exceeds T when a suitable set of μ features is selected and used (and the rest $n - \mu$ features discarded); further, the performance of M is always below T when any feature set with less than μ features is used. Thus, μ is the critical number of features that are necessary to ensure that the performance of M meets the given threshold.

A. Formal Definition of Critical Dimension

Formally, for dataset D_n with n features, machine M (a learning machine, an algorithm, etc.) and performance threshold T (the accuracy of training, etc.), we call μ the T-Critical Dimension of (D_n, M) if the following two conditions hold:

1. There exists a μ-dimensional projection of D_n, which lets M to achieve a performance of at least T, i.e., **($\exists D_{\mu}$**

 ISBN: 978-1-61208-329-2

$\propto D_n$) [$P_M(D_*) \geq T$], where $P_M(D_*)$ denotes the performance of M on input dataset D_*.

2. For all $j < \mu$, a j-dimensional projection of D_n fails to let M achieve performance of at least T, i.e., ("$D_j \propto D_n$) [$j < \mu \Rightarrow P_M(D_j) < T$]

To determine whether a Critical Dimension exists for a D_n and M combination is a very difficult problem. Specifically, the problem of deciding, given D_n, T, k ($k \leq n$), and a fixed M, whether k is the T-Critical Dimension of (D_n, M) actually belongs to the class $D^P = \{L_1 \cap L_2 \mid L_1 \in NP, L_2 \in coNP\}$ (C.H. Papadimitriou et al, 1982), where it is assumed that the fixed machine M runs in polynomial time in n, the dimension of the dataset. In fact, it is shown in the next subsection that the problem is D^P-hard [4].

Since NP and coNP are subclasses of D^P (note that D^P is not the same as NP ∩ coNP), the D^P-hardness of the Critical Dimension problem indicates that it is both NP-hard and coNP-hard, and likely to be intractable.

B. Proof That CDP is D^P-Hard

The Critical Dimension Problem (CDP) is stated formally as follows: **Given D_n, T, k ($k \leq n$), and a fixed P_M (the performance of M). Is k the T-Critical Dimension of (D_n, M)?** The problem to decide if k is the T-Critical Dimension of the given dataset belongs to the class D^P under the assumption that, given any $D_i \propto D_n$, whether $P_M(D_i) \geq T$ can be decided in polynomial time of n, i.e., the machine M can be trained and tested with D_i in polynomial time. Otherwise, the problem belongs to some larger complexity class, e.g., Δ^P_2. Note here that (**NP ∪ coNP**) $\subseteq$ $D^P \subseteq \Delta^P_2$.

To prove that the CDP is a D^P-hard problem, we take a known D^P-complete problem and transform it into the CDP. Let us consider the maximal independent set problem as an example. In graph theory, a Maximal Independent Set (MIS) is an independent set that is not a subset of any other independent set. That is, it is a set S such that every edge of the graph has at least one endpoint not in S and every vertex not in S has at least one neighbor in S. A MIS is also a dominating set in the graph, and every dominating set that is independent must be maximal independent, so it is also called independent dominating sets. A graph may have many MIS's of widely varying sizes; a largest maximal independent set is called a MIS.

EXACT-MIS problem – Given a graph with n nodes, and $k \leq n$, decide if there is a maximal independent set of size exactly k in the graph is a problem is D^P-complete as proved by Papadimitriou and Yannakakis, 1982. Now we will transform this D^P-complete to an instance of CDP. To construct the instance of the *CDP*, let: dataset D_n be a representation of the given graph with n nodes (e.g., D_n can be made to contain n data points, with n features, representing the adjacency matrix of the graph), T be the value 'True' from the binary range {T, F}, $\mu = k$ be the value in the given problem and M be an algorithm that decides if the dataset represents a maximal independent set of size μ, if yes P_M = True otherwise P_M = False, then a given instance of the D^P-complete problem is transformed into an instance of the *CDP*. Three examples are shown below and explained. If the threshold is T (True) from the binary range {T,F}, then the problem is considered to be the NP-complete EXACT-MIS problem, and F (False) if it is not a NP-complete EXACT-MIS problem. Figure 1 is a graph with 5 nodes, containing an EXACT-MIS of size 3.

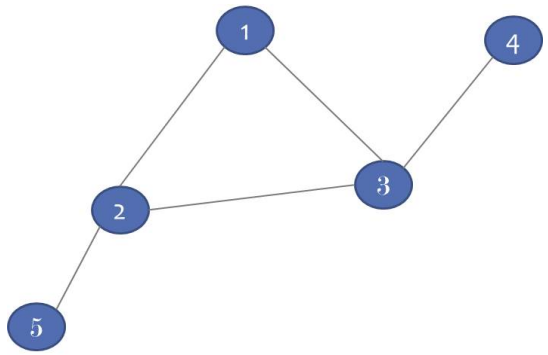

Figure 1. A Graph with 5 nodes showing exactly one MIS with 3 nodes {1,4,5}

D_5=

1	1	1	0	0
1	1	1	0	1
1	1	1	1	0
0	0	1	1	0
0	1	0	0	1

Figure 2. The adjecency matrix of graph with 5 nodes

Example 1: k=3
Threshold T = 'True' from the binary range {T, F}. μ = 3 exist; i.e., an EXACT-MIS of size 3 exists in D_5 and is as highlighted in the adjacency table shown in Figure 2. So, M is an algorithm that decides if the input dataset represents a maximal independent set of size μ, or M "verifies" that some $D\mu$ corresponds to a maximal independent set; i.e., $P_M(D_*)$ = 'True' if $D\mu$ allows M to construct a maximal independent set of G of size μ, where $D_* \propto D_n$ and D_n represents the adjacency matrix of G. Since the solution to the EXACT-MIS problem is True, solution to an instance of the CPD transformed from this is YES.

Example 2: k=4
Threshold T = 'True' from the binary range {T, F}. μ = 4 exists but is not an EXACT-MIS. From D_5 table it can be seen

that there does not exists any independent sets of size 4, so no EXACT-MIS of size 4 exists. Let M be an algorithm that decides if the input dataset represents a graph containing a maximal independent set of size μ, or M "verifies" that some $D\mu$ corresponds to a maximal independent set; i.e., $P_M(D_n)$ = 'True' if $D\mu$ allows M to construct a maximal independent set of G of size μ, where $D_n \propto D_n$ and D_n represents the adjacency matrix of G. In this example since no independent set of size 4 exists the solution to the instance of constructed *CDP* is NO, so $P_M(D_4)$ = 'False' for all D_4.

Example 3: k=2
Threshold T = 'True' from the binary range {T, F}. μ = 2 exists but is not an EXACT-MIS. Again, from D_5 table it can be seen that there exist independent sets of size 2 but they not EXACT-MIS. So, algorithm M decides $D\mu$ \corresponds to a maximal independent set if $D\mu$ allows M to construct a maximal independent set of G of size μ, where $D_n \propto D_n$ and D_n represents the adjacency matrix of G. In this example since no independent set of size 2 exists the solution to the EXACT-MIS is 'False' so a solution to an instance of constructed *CDP* is NO, $P_M(D_2)$ = 'False' for all D_2.

The D^P-hardness of the Critical Dimension problem indicates that it is both NP-hard and coNP-hard; therefore, it's most likely to be intractable, that is, unless P = NP.

III. METHOD TO FIND THE CRITICAL DIMENSION

We can see from the CDP Problem analyzed above that even deciding if a given number is a Critical Dimension is intractable, to find that number is certainly even more difficult. So, a heuristic method is proposed in the following. The heuristic method represents a feasible and practical approach in attempting to find the Critical Dimension of a given dataset and a given performance threshold with respect to a fixed machine. Though the heuristic method actually corresponds to a different definition of the Critical Dimension, it serves to validate the concept that μ exists for datasets, though maybe not for all of them; and we will see that for most of the datasets with which experiments were conducted a Critical Dimension indeed exists. Finally, the μ determined by this heuristic method is hopefully close to the theoretical Critical Dimension as defined in the formal definition.

In the heuristic method, the Critical Dimension of a dataset is defined as that number (of features) where the performance of a specific learning machine would begin to drop below the performance threshold significantly, and would not rise again when smaller number of features is used. To make the method feasible, the features are initially sorted in descending order of significance (according to some feature ranking algorithm) and the feature set is reduced by deleting the least significant feature after each iteration of the experiment when performance of the machine is observed. For cross validation purposes, therefore, multiple experiments can be conducted when attempting to determine the Critical Dimension of a dataset: the same machine is used in conjunction with different feature ranking algorithms; also, the same feature ranking algorithm is used in conjunction with different machines; then we analyze if the different experiments resulted in similar numbers for the Critical Dimension.

A. Heuristic Method to find the Critical Dimension

The term Critical Dimension of a dataset has been described as the minimum number of features required for a learning machine to perform prediction or classification with high accuracy. Empirically, the 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.

In other words, 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 and which can be determined experimentally. Specifically, let A = {a_1, a_2, …, a_n} be the feature set where a_1, a_2, …, 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, …, a_m} where $m \leq n$. For a learning machine M and a feature ranking method R, we call μ ($\mu \leq n$) is the Critical Dimension of [D_n, M, T], if the following conditions satisfy; If T is a given performance threshold that is considered acceptable, when M uses feature set Aμ the performance of M is $\geq T$, and whenever M uses less than μ features its performance drops below T_μ. As an illustration, the Hypothyroid disease dataset was classified using SMO (Sequential Minimal Optimization) classifier. This dataset was ranked using Chi-squared ranking algorithm. Figure 3 shows the Critical Dimension and was found to be 18; and it can be observed that this point satisfies the heuristic methods definition of a Critical Dimension.

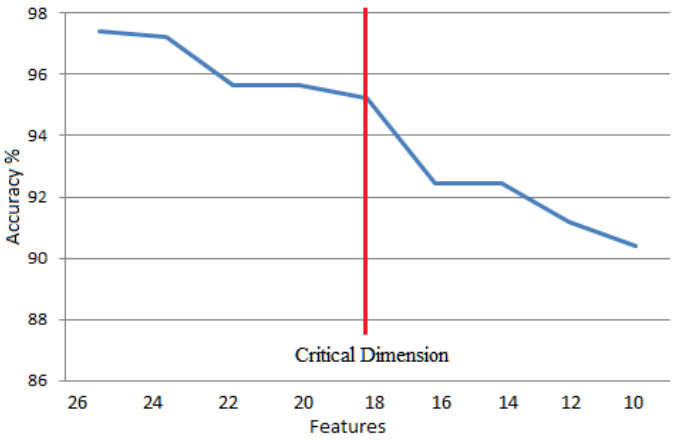

Figure 3. The Critical Dimension of Hypothyroid disease dataset

IV. FINDING THE CRITICAL DIMENSION USING FEATURE RANKING METHODS

To find approximate solutions to the Critical Dimension problem, a heuristic method based on feature ranking algorithms is applied. In this method, the performance

threshold *T* will not be specified beforehand but will be defined during the iterative process where a learning machine classifier's performance is observed as the number of features is decreased. The Figure 4 below shows the method to find the Critical Dimension.

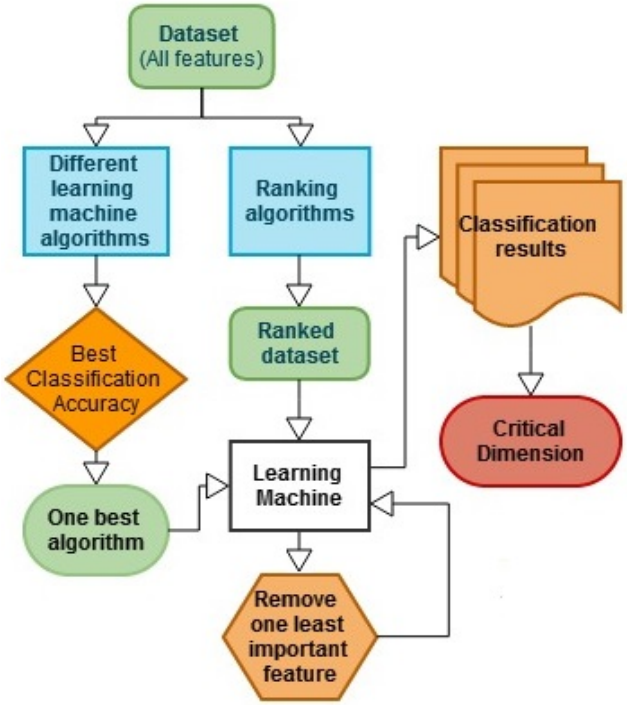

Figure 4. Method to find the Critical Dimension

A. Choosing the best classification algorithm

The dataset is first classified by building a model based on six different algorithms, namely, Bayes net, function, rule based, meta, lazy and decision tree learning machine algorithm [12][8][9]. The machine or model with the best prediction accuracy is chosen as the classifier to find the Critical Dimension for that dataset. Table 1 shows the accuracy results of the learning machine model built based on one of the six algorithms discussed above. Figure 5 shows the method in which the best classifier is choosen.

TABLE I. CHOOSING THE CLASSIFICATION ALGORITHM FOR THROMBIN DATASET

Accuracy %	Method		
	Algorithm	***Learning Machine***	***Best***
42.82	Bayes	Naive bayes	**C4.5 with 69.33%**
18	Functions	SMO	
63.59	Lazy	kStar	
66.18	Meta	Ada Boost	
68.39	Rule	Decision table	
69.33	Tree	C4.5	

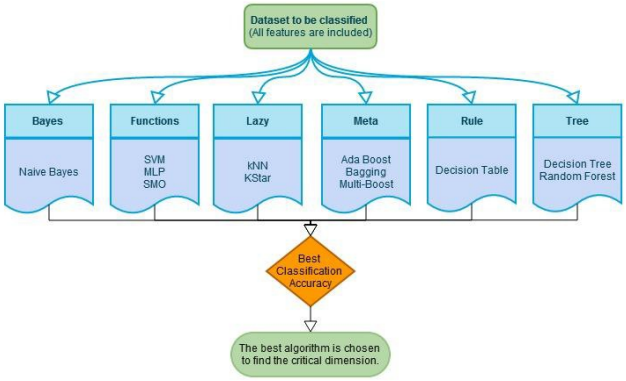

Figure 5. Choosing the best classification model

B. Ranking algorithm

The Chi-square ranking method [3] evaluates the worth of an attribute by computing the value of the chi-squared statistic 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.

V. EXPERIMENTAL RESULTS

There are three large datasets used in this experiment. The datasets are explained and the results are discussed. The dataset for the experiments are divided into 60% for training and 40% for testing. The model is retrained by changing the parameters to decrease the error rate. Six different models are built and retrained to get the best accuracy. The model that gives the best training accuracy is used to find the Critical Dimension.

A. Amazon 10,000 dataset results

The Amazon commerce review dataset [2] is a writeprint dataset. Internet users share attractive information with openness and anonymity to the online community to freely express their opinions. People with vested interests may take the opportunity to post biased information in anonymous ways, significantly harming the purpose of the open review. Therefore, authorship identification of online texts such as verifying the authorship of emails and messages on the cyber community, plagiarism detection and personal blogs is becoming important. Similar to biological fingerprint, the unique writing-style hidden in texts is vividly described as writeprint. Online writeprint identification is the task of predicting the most likely authorship of anonymous texts by using stylistic information in language. This can be seen as a single-label multiclass text categorization problem where the candidate authors represent different classes. The key task of writeprint identification is to extract fine-grained features from texts for quantifying the style of an author. Character n-grams have been proved to be very effective for capturing

complicated stylistic information hidden in the texts. For example, the most frequent character 4-grams of an experimental text indicate lexical (|_the|, |_to_|, |that|), word-class (|_was|, |ing_|), and punctuation usage (|,_wh|, |,_s|). This dataset are derived from the customers' reviews in Amazon Commerce Website for authorship identification. This dataset was originally created to examine the robustness of classification algorithms. Studies conducted the identification experiments for fifty authors in the online context reviews. These are the most active users (represented by a unique ID and username) who frequently posted reviews in the Amazon newsgroups. The number of reviews collected for each author is 30. This is a classification dataset and contains 50 authors x 30 reviews each = 1500 instances. There are 10,000 attributes and they include authors' linguistic style, such as usage of digit, punctuation, words and sentences' length and usage frequency of words and so on. This is a multiclass classification problem with 50 classes. The dataset contains numerical values for all features.

The classification results and the Critical Dimension are shown below. It can be seen from Figure 6 that the Amazon 10,000 dataset shows a Critical Dimension at feature size 2486. The graph below shows the results of the Amazon dataset and the plot shows the Critical Dimension.

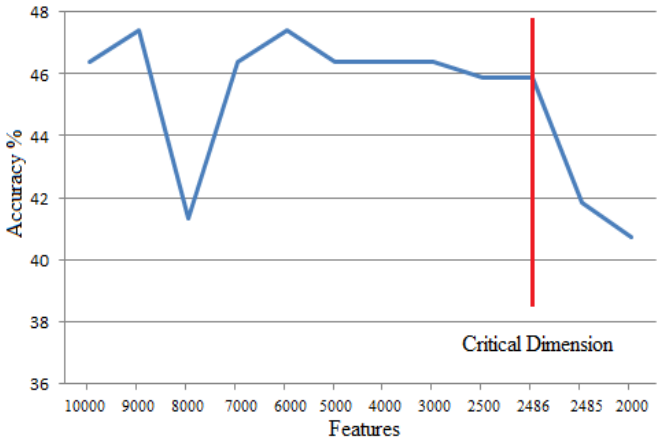

Figure 6. Critical Dimension of the Amazon 10,000 dataset

B. *Amazon ad. or non ad. dataset results*

The Amazon commerce reviews Internet advertisement dataset represents a set of possible advertisements on Internet web pages. The features encode the geometry of the image (if available) as well as phrases occurring in the URL, the image's URL and alt text, the anchor text, and words occurring near the anchor text. The task is to predict whether an image is an advertisement ("ad") or not advertisement ("nonad"). The dataset contains 459 advertisements and 2820 non ad. images, hence a total of 3279 instances. The attributes in this dataset contains 3 continuous and others binary; one or more of the three continuous features are missing in 28% of the instances. There are 1558 features in the Internet advertisement dataset.

The classification results and the Critical Dimension of the ad. and non ad. dataset is shown below. It can be seen from Figure 7 that the Ad dataset shows a Critical Dimension at feature size 383.

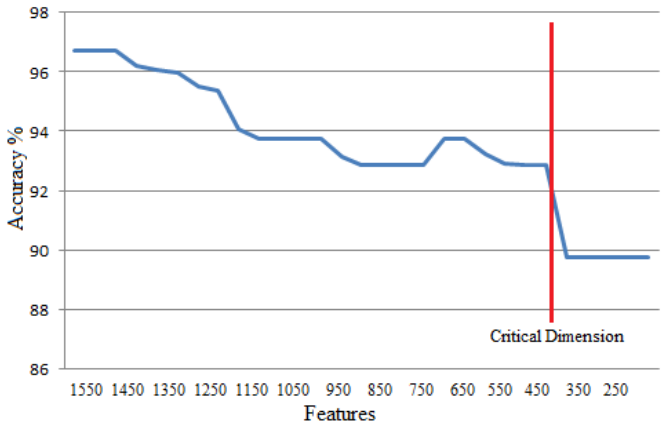

Figure 7. Critical Dimension of the Amazon ad. or non-ad. dataset

C. *Thrombin dataset results*

The present training data set consists of 1909 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting. The chemical structures of these compounds are not necessary for our analysis and are not included. Of these compounds, 42 are active (bind well) and the others are inactive. Each compound is described by a single feature vector comprised of a class value (A for active, I for inactive) and 139,351 binary features, which describe three-dimensional properties of the molecule. The definitions of the individual bits are not included - we don't know what each individual bit means, only that they are generated in an internally consistent manner for all 1909 compounds. Biological activity in general and receptor binding affinity in particular, correlate with various structural and physical properties of small organic molecules. The task is to determine which of these properties are critical in this case and to learn to accurately predict the class value. Drugs are typically small organic molecules that achieve their desired activity by binding to a target site on a receptor. The first step in the discovery of a new drug is usually to identify and isolate the receptor to which it should bind, followed by testing many small molecules for their ability to bind to the target site. This leaves researchers with the task of determining what separates the active (binding) compounds from the inactive (non-binding) ones. Such a determination can then be used in the design of new compounds that not only bind, but also have all the other properties required for a drug (solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc.).

The classification results and the Critical Dimension of the thrombin dataset are shown below. It can be seen that the thrombin dataset shows a Critical Dimension at feature size 8486. Figure 8 below shows the results of the thrombin dataset and the graph plot shows the Critical Dimension.

 ISBN: 978-1-61208-329-2

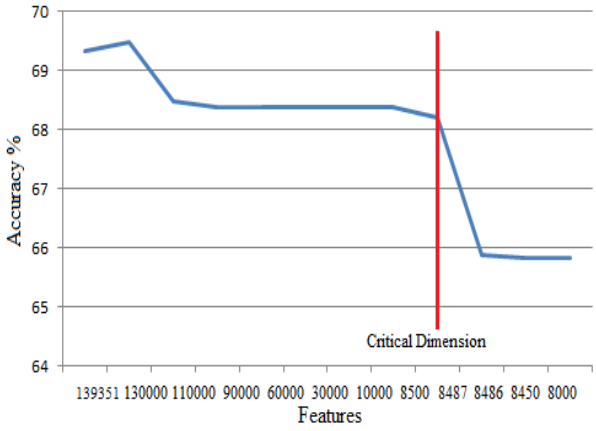

Figure 8. Critical Dimension of the thrombin dataset

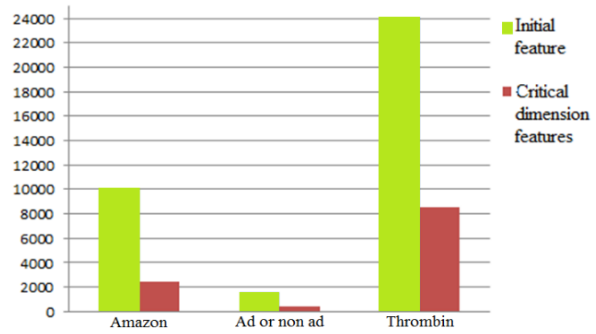

Figure 9. Reduction in the feature size of three large datasets at the Critical Dimension

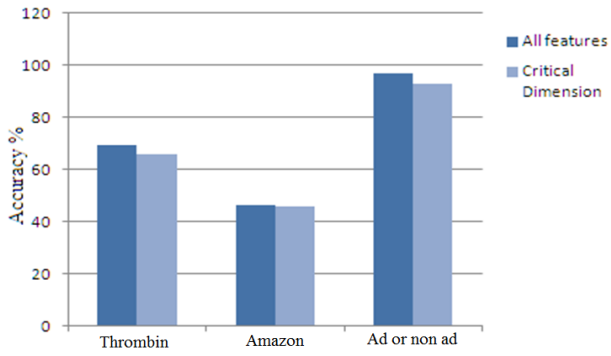

Figure 10. Prediciton accuracy of three large datasets at the Critical Dimension and intial condition (all features included)

Three large featured sized datasets namely the Amazon 10,000, the ad. or non ad. and the thrombin datasets were studied in this experiment. All three datasets shows an obvious existence of Critical Dimension. Figure 9 shows that the feature size has largely decreased at Critical Dimension and the performance is of each of these datasets are maintained 'high'. Figure 10 shows difference in accuracies at initial condition and at Critical Dimension. The initial Amazon 10,000 dataset contains 10,000 features and the accuracy with which the bagging classifier predicted the 50 classes was 46.39%. However, at the Critical Dimension the number of features was reduced to 2486 features and the accuracy to predict 50 classes using the bagging classifier was 45.88%. Similarly, for the binary class classification ad. or non ad. dataset, the initial number of features was 1559 and at Critical Dimension was 383 features. The random forest classifier the accuracy of classifying the initial dataset into the two classes was 96.71% and at Critical Dimension was 92.74%. The largest dataset namely the thrombin dataset contained 139351 features initially and using a bagging classifier, the classification accuracy to predict the two classes was 69.33%. The Critical Dimension for this dataset was then found and the number of feature at this Critical Dimension was found to be 8487 which is an enormous decrease in the feature size. At this Critical Dimension the classification accuracy was 65.87% using the bagging classifier. The results of this paper show us that a Critical Dimension is not only found in smaller datasets but also in much larger datasets. Results of 16 different datasets that were studied earlier are shown below [5]. The chart in Figure 11 shows the accuracies of all datasets. It can be seen that the accuracies at initial condition and at Critical Dimension are not very different; infact for some datasets like the Parkinson's disease and some text mining datasets the performance of the model to correctly classify has increased at Critical Dimension when compared to the performance accuracy measured at the initial condition. While the performance is maintained 'high', the feature size has decreased largely. This is shown in Figure 12.

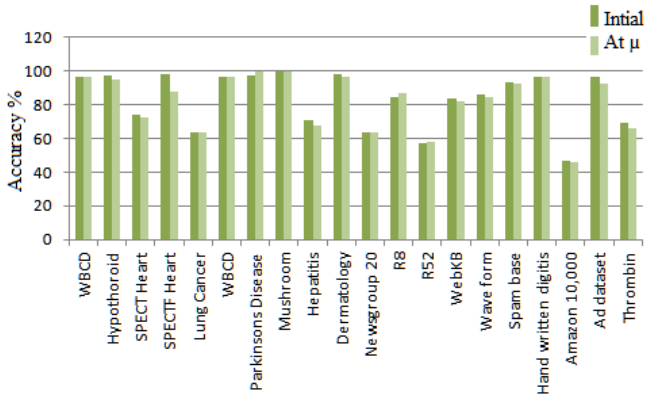

Figure 11. Accuracies of All Datasets at Initial Condition and at Critical Dimension

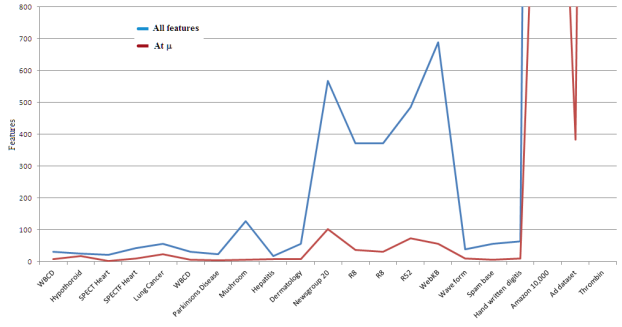

Figure 12. Reduction in the feature size of all datasets at the Critical Dimension.

VI. CONCLUSIONS AND FUTURE WORK

The concept of a Critical Dimension of datasets and a heuristic method for finding it are introduced. Firstly, we ahve shown that finding the Critical Dimension is an intractable problem, and therefore, justifies the use of heuristic methods for finding the Critical Dimension. We also presented the results showing that the heuristic method succeeded in finding the Critical Dimension of some large datasets.

Even though different feature ranking methods are used in the heuristic method, it is emphasized that this paper is not about feature tanking or selection–rather, it is about finding the Critical Dimension. The feature ranking algorithms were merely employed in the heuristic method as a means to help determine if a Critical Dimension exists for a given dataset.

However, many interesting questions are raised that deserve further investigation:

- The heuristic method may find a Critical Dimension for a given dataset and a given machine. In addition, the method identifies the features to be included. But how good the solution is (relative to the formal definition of Critical Dimension) is really unclear, since the method relies on a selected feature ranking algorithm which may well have overlooked the effect of combinations of features–though this seems inevitable for all general feature ranking algorithms that do not take into account prior knowledge about the features and/or the specific problem or application underlying the datasets.
- Using different ranking algorithms and different machines and apply the same heuristic method may lead to very different CD values, what does that mean?
- What does the existence of a CD mean for a dataset? Does it mean that the quality of data is low–since insignificant and/or useless features are included? Or does it mean that the amount of data is in fact insufficient–once the dataset is expanded with more data, might the CD disappear? Both seem to be possibilities.

More fundamentally, how do we count features? What is a feature? In many problems, features are developed and computed from the collected simple measurements (e.g., the TF-IDF feature in text classification). But this seems a different issue regarding prior knowledge. The authors are pursuing, as the next steps of this work, to develop and experiment with more sophisticated (than the linear) heuristic methods for finding the features that constitute the Critical Dimension, and apply the methods to larger datasets.

ACKNOWLEDGMENT

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.

REFERENCES

[1] A. Blum and P. Langley, “Selection of relevant features and examples in machine learning”, Artificial Intelligence, vol. 97, pp. 1-2, 1997.

[2] A. Frank and A. Asuncion, UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, http://archive.ics.uci.edu/ml, [retrived: May, 2010].

[3] A. M. A. Mesleh, “Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System”, Journal of Computer Science, ISSN 1549-3636, pp. 430-435, 2007.

[4] C. H. Papadimitriou and M. Yannakakis, “The complexity of facets (and some facets of complexity)”, Proceedings of the fourteenth annual ACM Symposium on Theory of Computing, pp. 255-260, 1982.

[5] D. Suryakumar, A. H. Sung, and Q. Liu, “Determine the Critical Dimension in data mining (experiments with bioinformatics datasets)”, In Intelligent Systems Design and Applications, 2011 11th International Conference, pp. 48-486, 2011.

[6] H. Almuallim and T. G. Dietterich, “Learning with many irrelevant features”, Ninth National Conference on Artificial Intelligence, MIT Press, pp. 547-552, 1991.

[7] H. Buhrman and J. M. Hitchcock, “NP-hard sets are exponentially dense unless coNP ⊆ NP/poly,” IEEE Conference on Computational Complexity, IEEE Computer Society Press, pp. 600-601, 2008.

[8] I. Guyon, A. Elisseeff, “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research, pp. 1157-1182, 2003.

[9] J. G. Dy and C. E. Brodley, “Feature Subset Selection and Order Identification for Unsupervised Learning”, Proceedings of the Seventeenth International Conference on Machine Learning, pp. 247-254, 2001.

[10] J. R. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann Machine Learning series, 1993.

[11] L. Breiman, “Random forests”, Journal Machine Learning, Vol 45(1) pp. 5-32, 2001.

[12] W. Buntine, “Theory refinement on Bayesian networks”, Proceedings of the Seventh, Annual Conference on Uncertainty in AI, pp. 52–60 1991.

[13] X. Geng, T. Liu, T. Qin, and H. Li, “Feature selection for ranking”, In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 407-414, 2007.

 ISBN: 978-1-61208-329-2