Critical Dimension in Data Mining

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 - **Data mining is an increasingly important means of knowledge acquisition for many applications in diverse fields such as biology, medicine, management, engineering, etc. When tackling a large-scale problem that involves a multitude of potentially relevant factors but lacking a precise formulation or mathematical characterization to allow formal approaches to solution, the available data collected for the application can often be mined to extract knowledge about the problem. Feature ranking and selection, thereby, are immediate issues to consider when one prepares to perform data mining, and the literature contains numerous theoretical and empirical methods of feature selection for a variety of problems. This work in progress paper concerns the related question of critical dimension, i.e., for a specific data mining task, does there exist a minimum number (of features) which is required for a specific learning machine to achieve satisfactory performance? As a first step in addressing this question, a simple ad-hoc method is employed for experiment and it is shown that the phenomenon of critical dimension indeed exists for several of the datasets studied. The implications are that each of these datasets contains irrelevant features or input attributes, which can be eliminated to achieve higher accuracy in model building using learning machines.**

Keywords-feature selection; critical dimension; machine learning.

I. INTRODUCTION

Data mining is aimed at extracting useful information or knowledge from datasets; to achieve this goal, feature selection is often necessary to eliminate lesser or insignificant features in order to reduce the size of the dataset and to facilitate model building (e.g., using learning machines) for knowledge extraction. Many methods have been proposed for feature selection [1]. The interesting fact about extracted features are that sometimes not all extracted features are individually useful; however, correlation of features itself an intriguing question.

We may use learning machines to find feature correlation or to discover important or relevant features. Some theoretically optimal criteria could become practically intractable [2]. The ultimate, guaranteed optimal feature selection method requires exhaustive analysis of all possible subsets of features; this is infeasible for datasets with a large number of features; so, the next best goal is to find a satisfactory set of subsets. Feature selection is usually done in two different ways, namely subset selection or entropy-based selection and feature ranking. Feature ranking uses ranking algorithms which scores all features using certain metrics and ranks them accordingly [3]. A subset selection method uses an algorithm to find a best possible subset in arbitrary time. Here, the term best possible subset refers to the best subset found among satisfactory set of subsets [4].

II. FEATURE RANKING

The main objective of feature selection is to improve the prediction performance or accuracy, to provide faster and cost-effective predictors and understand the correlation among data [5]. For our experiments, we use both feature selection and subset selection.

A supervised 'Chi-squared Ranking Filter' [6] and a supervised 'Support Vector Machines (SVM) feature evaluator' [7] method are used for ranking features. A 'Ranker' search method ranks attributes according to their relevance and individual evaluations. Using Ranker we can set the threshold to reduce the attribute set to consider or also specify the set of attributes to ignore; hence it is comfortable for our experiments to eliminate some unwanted features. The Chi-squared Ranking Filter evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. It is a statistical test to find the independence of two events for goodness of fit of an observed distribution to a theoretical one whose value is in zero to infinity range and cannot be negative. SVM feature evaluator evaluates the worth of an attribute by using an SVM classifier. Attributes are ranked by the square of the weight assigned by the SVM feature evaluator. Attribute selection for multiclass problems is handled by ranking attributes for each class separately using a one to all method and then dealing from the top of each pile to give a final ranking.

To find the best feature subset, we use supervised CFS Subset Evaluator method and a greedy stepwise search algorithm. The algorithm 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. Subsets of features that are highly correlated with the class while having low inter-correlation are preferred [8][9].

The two feature selection methods discussed above are the most widely used methods but there could always be that one subset which is the best feature subset or the correlation among a certain low ranked features could increase the

 ISBN: 978-1-61208-181-6

performance. Hence, in this paper, we show results of a method called critical dimension which can provide us the minimum number of features that are required for a learning machine to perform accurately.

III. CRITICAL DIMENSION

The *critical dimension* of a dataset is the minimum number of features required for a learning machine to perform prediction or classification with high accuracy. As such, it is an informal concept and empirical methods are called for to determine the critical dimension. 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 $m \leq n$. For a learning machine *M* and a feature ranking method *R*, we call μ ($\mu \leq n$) the critical dimension of [*M*, *R*], if whenever *M* uses feature set A_k with $k \geq \mu$ 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.

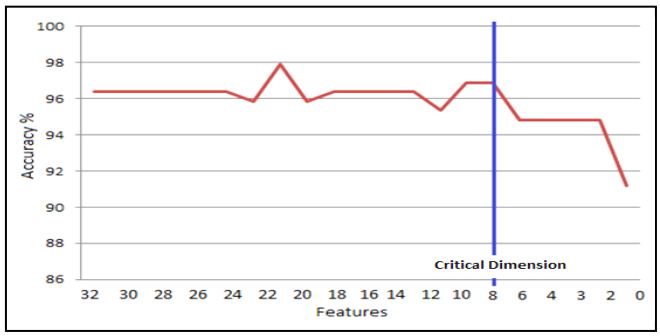

Figure 1. Showing the critical dimension at feature size 8

The graph in Figure 1 shows that there exists a μ at 8 features in the Wisconsin breast cancer dataset [10] dataset.

AdaBoost was used to classify this dataset. The graph is plotted with the number of features on the x-axis and prediction accuracy on the y-axis. From the graph we can see that the performance decreases if we choose lesser features than μ and the performance never rises above the measure at μ.

The first step in find μ is to rank all features using ranking algorithms. In this experiment we used *Chi Squared Attribute Evaluator* as the attribute evaluator and Ranker as the search method for feature ranking and a SVM subset evaluator for subset feature selection. Once the datasets are ranked the prediction accuracy is calculated. In the following iterations prediction accuracy is calculated by removing one least important feature each time till and beyond the critical dimension point. The results are studied and the point at which the performance curve as shown in Figure 1 drops drastically and never rises above that point is defined as the unique μ for that dataset.

Utilizing results from experiments carried out earlier, we can say that μ exist in most datasets and that this μ is a unique number pertaining to that dataset for that particular or specific learning machine classifier and ranking combination. Results using similar classifiers by other experimenters are in the UCI database.

The table below shows the results of experiments performed previously on six different datasets from the UCI repository [11] which, either has an obvious critical dimension (O), or no obvious critical dimension (N/O), as shown in the last column of the table. The classifiers used for classifying the datasets are also shown. The initial condition is when all ranked features or the best subset features are analyzed. For some of datasets, all features are feature ranked and then a learning machine classifier is used to find the accuracy and for others the best feature subset is found and classification accuracy is found using a learning machine classifier. Experiments were performed to find μ by removing one least important feature at the beginning of iteration and calculating the performance accuracy at the end of each iteration. In the table below, the accuracy at μ and accuracy during the first iteration are shown. The classifiers used for each dataset are also tabulated. For the Wisconsin breast cancer dataset (WBDC) two different classifiers were used to experiment. It can also be seen that the critical dimension is unique to that (dataset, machine learning algorithm and ranking) combination.

TABLE I. RESULTS OF BIO-MEDICAL DATASETS

SN	Name	Initial condition		At critical dimension		Classifier	Type
		# of features	*Accuracy %*	*# of features*	*Accuracy %*		
1	WBCD	31	96.3731	8	96.8912	Ada Boost	O
2	Hypothoroid	25	97.3953	18	95.2483	SMO	O
3	SPECT Heart	22	74.1573	3	72.6592	Attribute selected	O
4	SPECTF Heart	44	98.001	10	87.9121	Bagging	O
5	Lung Cancer	56	63.6364	24	63.6364	Multi Boost	O
6	WBCD	31	96.3731	6	96.8549	Multilayer Perceptron	N/O
7	Parkinsons Disease	23	96.9697	5	100	Ada Boost	N/O

a. The dataset used for this experiment are from the UCI repository.

 ISBN: 978-1-61208-181-6

Critical dimension is an innovative and cost effective method to reduce the problems involved in feature selection as it is almost always impossible to find the best possible feature subset possible. The main idea is finding the minimum set of features necessary for the successful development of learning machine classifiers for a given dataset. The results from the table demonstrate that this is indeed the case for the several bioinformatics dataset studied.

We can see from the results presented that there exists a unique critical dimension in some datasets which, when found can reduce the feature dimension, without compensating in performance. The accuracy of performance with all the features and at the critical point for all dataset in Table 1 shows that there is not much difference in the performance.

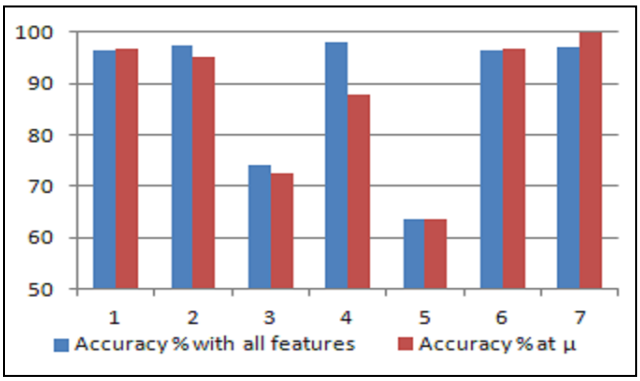

Figure 2. x-axis representing the datasets as numbered in Table 1 and the y-axis are the accuracies

A mushroom dataset was created by us with 127 features. There are two types of genes recorded in this dataset, *Lentinus Fr. and Marasmius Fr.* and 40 samples of each. The 127 features are the mushrooms habitat, details of the macroscopic pileus, macroscopic gills, macroscopic stipe, microscopic context, microscopic basidispore, microscopic basidia, microscopic cystidia, microscopic trama and microscopic Pileipellis. The dataset contains subgenus of both *Lentinus Fr. and Marasmius Fr. Lentinus Fr. and Marasmius Fr.* are the broader classification. The dataset contains some missing values or gaps.

This paper concludes the results of a new method to identify mushroom gene using machine learning methods. Different types of mushrooms are used as an extract to cure certain cancers and hence it is highly important to classify them [12]. In this experiment, we are trying to identify the species into the broader class classification. For example, the *Lentinus Fr.* has subgenus type such as *Lentinus cladopus, Lentinus squarrosulus, Lentinus cyathiformis* etc. which are classified as *Lentinus Fr.* Similarly the subgenus of *Marasmius Fr.* are grouped into type *Marasmius Fr.* Machine learning methods were used to identify the types. This is a binary class classification.

The dataset contains a total of 127 features and 80 samples. The datasets for the experimentation was divided into testing and training sets. The split is 66% for training and the rest for testing. The performance measure was the prediction accuracy of the test set. The mushroom dataset was classified using different classifiers, namely Rule based classifier ZeroR, classifier, SMO, AdaBoost and ADTree. We can see that the rule-based classifier accuracy was poor and SMO and ADTree showed 100% accurate results.

TABLE II. RESULTS OF DIFFERENT LEARNING MACHINE CLASSIFIER

Method	Accuracy%
ZeroR	40.7407
AdaBoost	96.2963
SMO	100
ADTree	100

A ranking algorithm was then used to rank the dataset. The ranking method used was CfsSubsetEval and the selection was made using greedy stepwise algorithm. The output was a feature set of 20 features. The feature numbers of the best feature subset was {3,7,14,22,31,36,58,68,72, 73,74,75,78,84,93,113,121,122,125,127}. Using this best feature subset and SMO classifier the results obtained are shown below.

TABLE III. SMO RESULTS OF BEST FEATURE SUBSET

Method	Accuracy%	Confusion matrix
SMO (using best feature subset)	100	a b 0 16 | a = Lentinus Fr. 11 0 | b = Marasmius Fr.

Now, using Ranker algorithm and ChiSquareAttributeEval method, all 127 features were ranked, for example, the 84th feature was ranked the highest or most important feature and 67th feature was ranked as the least important feature. We then use only the top twenty features ranked by the Ranker and run our learning machine classifier. The dataset was split into training (66%) and testing dataset (34%). The second line shows the output of SMO classifier using the top 20 features.

TABLE IV. RESULTS OF MUSHROOM DATASET

Feature	TP rate	FP Rate	F Measure	ROC Area	Mean abs error	Relative abs error	Accuracy%
31	0.96	0.04	0.964	0.994	0.045	9.758	96.373
30	0.96	0.04	0.964	0.994	0.045	9.758	96.373
11	0.96	0.04	0.964	0.99	0.055	11.903	96.373
10	0.95	0.55	0.953	0.988	0.058	12.416	95.336
9	0.97	0.03	0.969	0.993	0.0562	11.974	96.891
8	**0.97**	**0.03**	**0.969**	**0.993**	**0.0562**	**11.974**	**96.891**
7	0.95	0.07	0.948	0.993	0.0591	12.591	94.818
6	0.95	0.07	0.948	0.993	0.0591	12.591	94.818
5	0.95	0.07	0.948	0.993	0.0607	12.941	94.818
4	0.96	0.03	0.964	0.992	0.0581	12.384	94.818

The critical dimension was found for the mushroom dataset. We can see from the table above that a critical dimension exists and is 7 features. We can see that when the experiment was run using top 6 features the accuracy drops

 ISBN: 978-1-61208-181-6

to 96.2963%. Hence, a critical number 7 can be assigned to this mushroom dataset using SMO.

The above experiments show that a 100% accurate perditions result was obtained by means of a SMO classifier using the best feature subset and also using the same number of features as in the best feature subset, i.e., top twenty features. The dataset was analyzed to find the critical dimension and a feature set containing the top 7 features namely, Microscopic Context type (homoiomerous or heteromerous), presence of Annulus or partial veil, Macroscopic Stipe Color, Microscopic Trama breath, Microscopic Cystidia Cheilocystidia shape, Macroscopic Pileus Shape, and Macroscopic Stipe consistency was found to be the critical dimension of the mushroom dataset. The results of this paper are a breakthrough in mushroom identification for broader genus identification.

The graph showing the critical dimension of the mushroom dataset is shown below. From the plot and the table, we observe that this dataset possesses an obvious critical dimension.

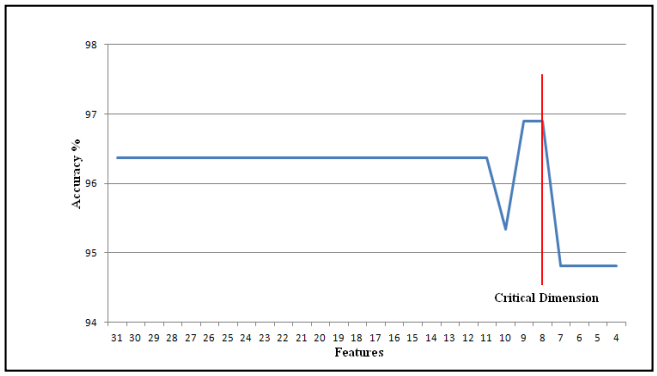

Figure3. Mushroom dataset showing μ = 8

CONCLUSION AND FUTURE WORK

As we continue to explore the concept of critical dimension and seek to develop a more formal framework, we are also trying to study and verify the ramifications of this phenomenon. Clearly, a dataset that exhibits an obvious critical dimension indicates that it contains irrelevant features which can be eliminated, or that the dataset itself is not large enough or sufficiently representative of the problem's whole input space to allow the construction of accurate models using learning-machine-based approaches (i.e., the inclusion of more data points may make the critical dimension disappear). Experiments are also being carried out to study critical dimensions in relation to different learning machines and feature ranking methods, since it appears that the critical dimension of a dataset is dependent on both the adopted learning machine and the adopted feature ranking/selection method for mining the data. It is believed that this research complements the research on feature ranking and selection in several aspects by addressing the question of how many features are essential in building, e.g., a learning machine classifier that delivers acceptable performance. Also, the existence of a critical dimension for a dataset indicates a measure of poor data quality and points to the opportunity of dimension reduction by eliminating useless or irrelevant features.

We are creating a much larger dataset for the mushroom study to perform experiments on multiclass classification and to see if the results are as expected or as good as the binary classification. New dataset will be tested using the top 7 features given by experiments performed in this dataset. Sub genus identification and classification using data mining is the next step after multiclass classification experiments are carried out.

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) and the mushroom dataset provided by CAS in Botany department, University of Madras, Guindy campus, Chennai, India are gratefully acknowledged.

REFERENCES

[1] Dy, J. G. and Brodley, C. E., Interactive visualization and feature selection for unsupervised data, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 360–364, 2000

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

[3] Guyon, I. and Elisseeff, A., An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3, pp. 1157-1182, 2003

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

[5] Hong, Z.Q. and Yang, J.Y., 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

[6] Mesleh, A. M. A., Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System, Journal of Computer Science, pp. 430-435, 2007

[7] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V., Gene selection for cancer classification using support vector machines, Machine Learning, pp. 389-422, 2001

[8] Hall, M. A., Correlation-based Feature Subset Selection for Machine Learning, Hamilton, New Zealand, 1998

[9] Geng, X., Liu, T. Y., Qin, T., and Li, H., Feature Selection for Ranking, SIGIR, 2007

[10] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets <retrieved: March, 2010>

[11] Wolberg, W.H., Street, W.N., and Mangasarian, O.L., Machine learning techniques to diagnose breast cancer from fineneedle aspirates, Cancer Letters, Vol. 77, pp. 163-171, 1994

[12] Borchers, A.T., Stern, J.S., Hackman, R.M., Keen, C.L., and Gershwin, M.E., Mushrooms, tumors, and immunity, Exp Biol Med, Vol. 221, pp. 281-293, 1999

 ISBN: 978-1-61208-181-6