Parameter Tuning: Exposing the Gap between Data Curation and Effective Data Analytics

Catherine Blake and Henry A. Gabb

ILLINOIS
GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

The Challenge and Promise of "Big Data"

Finding patterns in large datasets can deliver better decision making support to all aspects of our lives. However, the complexity of the data and the underlying analytics makes reproducibility and curation of the analysis difficult. This can lead to confusing results and redundant experimentation. This is the gap between the real and potential value of data analytics.

The case study presented here illustrates the vastness of a typical parameter space and the impact this has on models generated from data. It also illustrates the importance of parameter details to better understand the relationships between data, computational models, and subsequent model accuracy.

Materials and Methods

Our goal is to automatically identify results in scientific articles. The problem was framed as a classification task, where the classifier was trained to distinguish a result from a non-result sentence. Seventeen full-text articles from PubMed were annotated, giving 2556 total sentences with 965 reporting a result. A random sample of 10% of these sentences were used for model evaluation. The other 90% were used to train the classifiers.

The chi-squared (CHI) feature selection method was used to find high-information terms in the vocabulary. The vocabulary consists of all terms from the corpus reduced to their base forms with stopwords removed.

We tested four classifiers as implemented in the Oracle Data Miner 11g (ODM, release 1): support vector machines (SVM), decision trees (DT), general linear model (GLM), and naïve Bayes (NB).

Hundreds of modeling experiments were conducted to explore the parameter space.

Conclusions

The optimal settings for a given modeling problem are data dependent so optimal parameter settings cannot be known *a priori*.

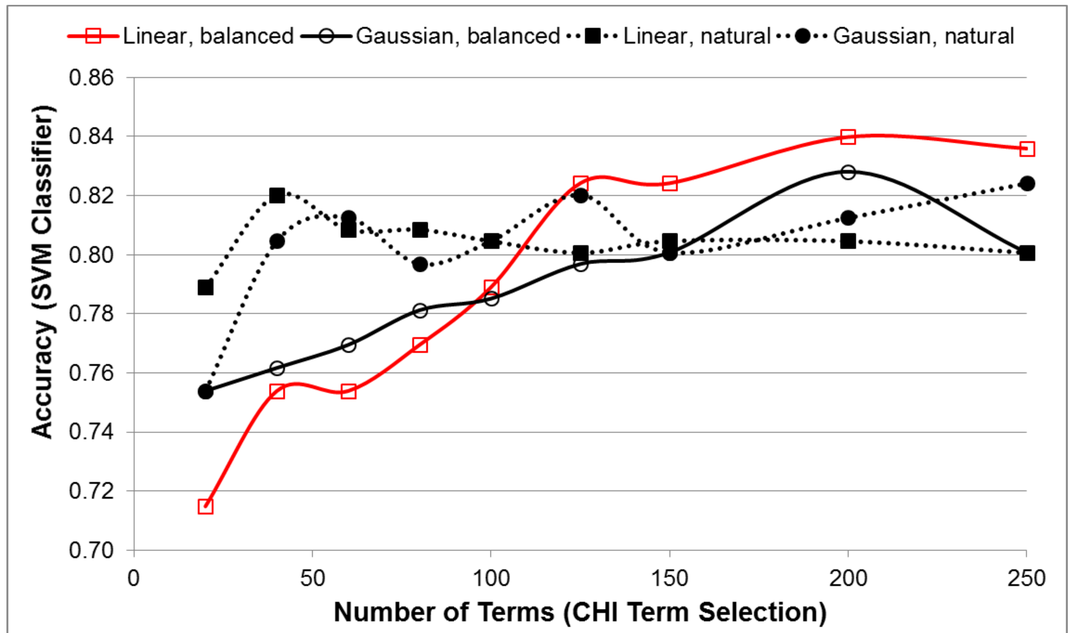
An exhaustive search of the parameter space is difficult. The parameter space is so large that no single researcher can explore all possible settings. Hence, better curation of the dataset and its resulting models is needed.

Acknowledgments

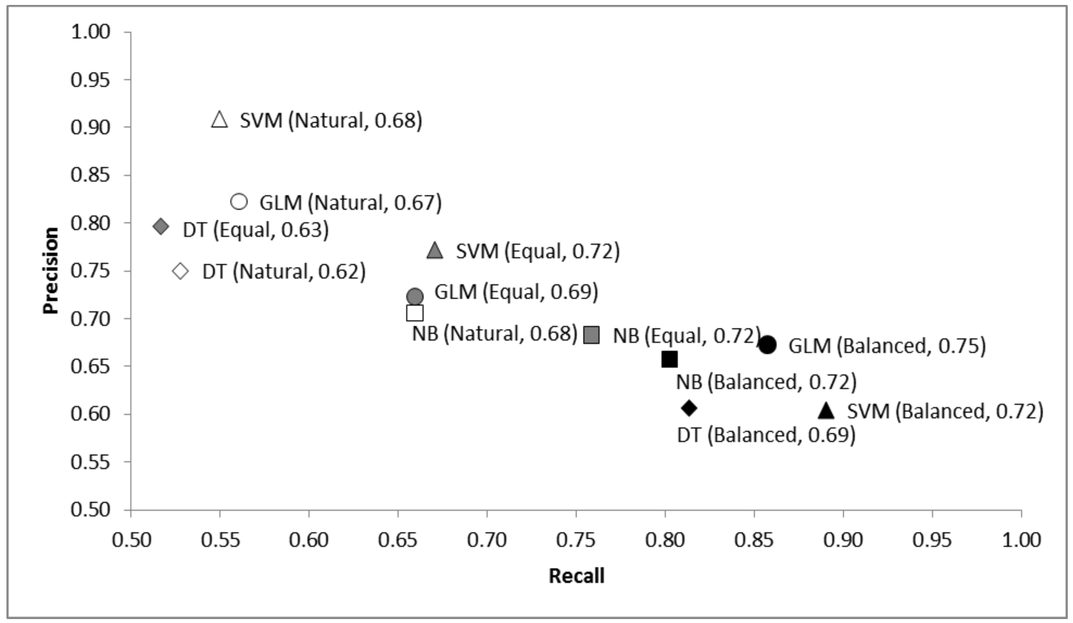
This poster is made possible in part by a grant from the U.S. Institute of Museum and Library Services (IMLS), Laura Bush 21st Century Librarian Program Grant Number RE-05-12-0054-12, Developing a Model for Sociotechnical Data Analytics (SODA) Education.

INSTITUTE of Museum and Library SERVICES

Parameter Interactions Further Increase Complexity



This plot illustrates the interaction between parameters, class weights, and the number of features on the SVM classifier. The red line indicates the ODM default settings. Notice that accuracy varies widely for the default model while other models are less sensitive to the number of features.

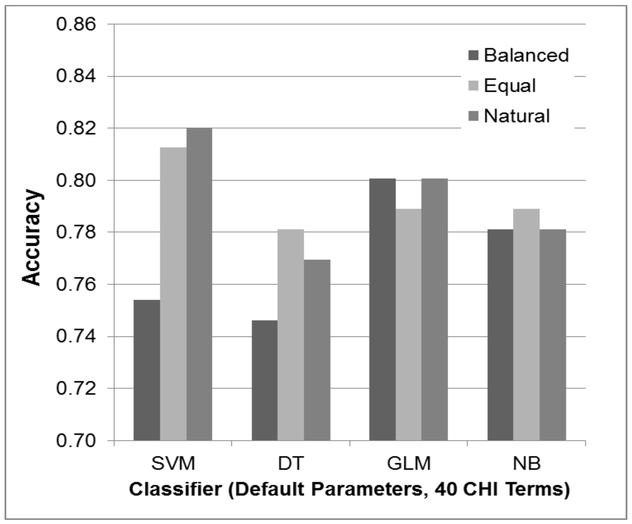


The precision vs. recall data show that this interaction exists for all four ODM classifiers. The data points are from models built with the default classifier settings, 40 CHI terms, and varying class weights. F1 scores are shown after the class weight setting.

Parameter Space of Typical Data Mining Experiments

Parameter	Classifier	Default
Generate Row Diagnostics	GLM	Off
Confidence Level	GLM	0.95
Reference Class Name	GLM	System
Missing Value Treatment	GLM	Mean
Specify Row Weights	GLM	Off
Enable Ridge Regression	GLM	On
Ridge Value	GLM	System
Singleton Threshold	NB	0
Pairwise Threshold	NB	0
Kernel Function	SVM	Linear
Tolerance Value	SVM	0.001
Complexity Factor	SVM	System
Active Learning	SVM	On
Homogeneity Metric	DT	Gini
Maximum Depth	DT	7
Minimum Records in a Node	DT	10
Minimum Percent of Records in a Node	DT	0.05
Minimum Records for a Split	DT	20
Minimum Percent of Records for a Split	DT	0.1

This table lists the many tuning parameters of the classifiers in the Oracle Data Miner 11g (ODM release 1). Some of these parameters are continuous, which leads to an enormous search space. Reproducibility is compromised unless all modeling parameters are reported, but this is rarely the case in the literature. Likewise, software versions must also be reported precisely. Data mining techniques are evolving so the parameters and default settings often change between versions.



In addition to the classifier-specific settings, data-specific class weights can significantly effect model accuracy. In the ODM, the default class weight varies by classifier but this parameter is hidden among the "advanced" settings.