Spring 2017

Automated Classification to Improve the Efficiency of Weeding Library Collections

Kiri Lou Wagstaff
San Jose State University

AUTOMATED CLASSIFICATION TO IMPROVE THE EFFICIENCY OF WEEDING LIBRARY COLLECTIONS

A Thesis

Presented to

The Faculty of the School of Information

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Library and Information Science

by

Kiri L. Wagstaff

M.S. Geological Sciences

Ph.D. Computer Science

May 2017

The Designated Thesis Committee Approves the Thesis Titled

AUTOMATED CLASSIFICATION TO IMPROVE THE EFFICIENCY OF WEEDING LIBRARY COLLECTIONS

by

Kiri L. Wagstaff
M.S. Geological Sciences
Ph.D. Computer Science

APPROVED FOR THE SCHOOL OF INFORMATION

SAN JOSÉ STATE UNIVERSITY

May 2017

Geoffrey Z. Liu, Ph.D.	School of Information
Virginia M. Tucker, Ph.D.	School of Information
Diane G. Klare, MLIS	Wesleyan University
Patricia A. Tully, MLIS	Ketchikan Public Library

ABSTRACT

AUTOMATED CLASSIFICATION TO IMPROVE THE EFFICIENCY
OF WEEDING LIBRARY COLLECTIONS

by Kiri L. Wagstaff
M.S. Geological Sciences
Ph.D. Computer Science

Studies have shown that library weeding (the selective removal of unused, worn, outdated, or irrelevant items) benefits patrons and increases circulation rates. However, the time required to review the collection and make weeding decisions presents a formidable obstacle. In this study, we empirically evaluated methods for automatically classifying weeding candidates. A data set containing 80,346 items from a large-scale academic library weeding project by Wesleyan University from 2011 to 2014 was used to train six machine learning classifiers to predict "Keep" or "Weed" for each candidate. We found statistically significant agreement ($p = 0.001$) between classifier predictions and librarian judgments for all classifier types. The naive Bayes and linear support vector machine classifiers had the highest recall (fraction of items weeded by librarians that were identified by the algorithm), while the k-nearest-neighbor classifier had the highest precision (fraction of recommended candidates that librarians had chosen to weed). The most relevant variables were found to be librarian and faculty votes for retention, item age, and the presence of copies in other libraries. Future weeding projects could use the same approach to train a model to quickly identify the candidates most likely to be withdrawn.

ACKNOWLEDGMENTS

I am grateful to my advisor, Dr. Geoffrey Liu, for his enthusiasm for this work and his tireless feedback and guidance. I also thank my committee members, Dr. Virginia Tucker, Diane Klare, and Patricia Tully, for their encouragement and support as well as their expertise on weeding methods and the data set from Wesleyan University. This thesis would not have been possible without the expert input of Lori Stethers at Wesleyan who provided vital insights into how the data was collected.

I am grateful for the support of my family and friends throughout this process, with a special thanks to Manuel Martinez-Lavin. I also thank the librarians at the Monrovia Public Library, who initially inspired my interest in Library Science and provided me with several delightful years of volunteer experience in the library setting.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

As library collections grow and patron needs evolve, there is an ongoing need for reviewing and maintaining physical collections. A key component of this process is weeding, the selective removal of items that are outdated, physically worn, no longer relevant to patron interests and needs, and/or available in electronic form. There is widespread agreement among librarians that weeding benefits both the library, by reducing the number of unneeded or unwanted items that are maintained, and the user population, by making desired items easier to find (Dilevko & Gottlieb, 2003). Pruning the collection to remove unwanted items can increase library circulation rates. Weeding also creates space that can be used for new acquisitions or to support other library needs, such as programming, maker spaces, or study areas (Lugg, 2012; Slote, 1997).

Despite the benefits, weeding is often low on the priority list for busy librarians. In contrast to fulfilling an inter-library loan request or helping a patron locate an item, weeding provides no immediate observable benefit. Further, it can impose a psychological strain on those tasked to implement it; many librarians find it stressful to make the decision to discard an item (Dilevko & Gottlieb, 2003).

One of the largest obstacles to weeding is the time it requires. Making a decision about a single title can take several minutes (Zuber, 2012). Large-scale weeding projects can require the review of tens of thousands of titles. For example, Wesleyan University conducted a weeding project from 2011 to 2014 that began by identifying 90,000 weeding candidates for individual review (Tully, 2011). Reviewing was done in two phases, by the librarians and then by interested faculty members. The project involved 17 librarians, two consultant subject specialists, and approximately

20 staff members plus two new employees (a reference librarian and a staff member) who were hired specifically to support the project. The project took three years to complete (P. Tully, personal communication, October 16, 2014).

Other obstacles, such as patron opposition to the act of weeding, can also hinder progress and must be addressed carefully and diplomatically. At Wesleyan University, faculty opposition to weeding was a significant factor, and it was necessary to provide a means for faculty to provide input on which weeding candidates should be retained (Tully, 2012). However, the experience at other university libraries has been different. Monmouth University found that their weeding project was "greeted enthusiastically" by some faculty members due to a general concern about outdated materials (Dubicki, 2008). Each library must assess and respond to the particular patron views and concerns at hand.

One possible solution to the time obstacle is automation, and some automation is already available. Services exist that can apply a set of weeding criteria to a library's circulation records to generate the initial list of weeding candidates (Lugg, 2012). A librarian reviews each candidate and assigns it to either the "Keep" or the "Weed" category. However, for the purposes of weeding, each candidate that is labeled "Keep" on the initial list represents an unproductive expenditure of the librarian's time. The ideal candidate list would be one that contains only items that the librarian would agree to weed. There is a potential for significant time savings if an automated method could be employed to filter and refine the list of weedable candidates.

Motivated by this line of thinking, an experimental study was conducted to assess the potential improvement in weeding efficiency that could be achieved using a data mining approach. The study was performed using existing records from the Wesleyan University weeding project. Automated machine learning classifiers were

trained and evaluated using the weeding decisions that were recorded for each item. Agreement between librarian judgments and weeding decisions predicted by machine learning classifiers was calculated and tested for statistical significance. The study found significant agreement between the predictions made by the classifiers and human judgments. The most relevant variables were found to be librarian and faculty votes for retention, item age, and the presence of copies in other libraries.

This thesis reports on the experimental study of automated classifiers for assistance in library weeding projects. The remaining content is organized as follows. In Chapter 2, the existing literature on weeding and automated methods for collection analysis is reviewed in detail. Chapter 3 defines the research problem and hypothesis. Chapter 4 provides more information about machine learning and the specific classifiers that were used in this study. Chapter 5 describes the methodology for the experimental study of automated weeding recommendations, and Chapter 6 presents and discusses the results of the experiments. Chapter 7 summarizes the findings and conclusions of this study. Appendix A defines key terms.

The results of this study indicate that machine learning classifiers can improve the efficiency of weeding projects by automatically identifying those candidates most likely to be withdrawn. The goal is not to replace the librarian or faculty member with an automated method, but rather to assist with the decision process. Staff involved in a weeding project can use a classifier's recommendations to make the best use of limited review time. We anticipate that the results of this study can serve to aid and improve future weeding projects by reducing the time needed for their completion.

Chapter 2

Literature Review

2.1 Introduction

This chapter reviews the motivation for weeding library collections and describes how such projects are conducted. Next, several weeding project challenges are discussed, including the primary motivator for this study, which is the time that such projects consume. One option for reducing the time required is to employ computer automation to assist in weeding projects. The discussion of current automation methods and tools highlights an opportunity for data mining or machine learning methods to assist in filtering or prioritizing the lists of weeding candidates.

2.2 Motivation for Weeding Projects

Many weeding efforts are motivated by the empirical observation that in many libraries, a large fraction of the collection never circulates. In 1975, Kent, Cohen, Montgomery, Williams, Bulick, Flynn, Sabor, and Mansfield (1979) found that 49% of all titles at the University of Pittsburgh Library had not circulated in the past seven years. Similarly, 40% of the collection in Harvard's Widener library had not circulated in the twenty years prior to 1996, going back as far as computerized records exist (Silverstein & Shieber, 1996). In 2007, Concordia University found that 15% of the library's collection had not circulated in 20 years (Soma & Sjoberg, 2010).

Non-circulating items can be a liability for libraries in that they consume shelf space and resources but do not directly benefit patrons. They also reduce the library's overall circulation rate, which is commonly calculated as the number of annual circulation counts divided by total holdings. High circulation is valued by librarians because it contributes to a feeling that the library is "serving its community well" (Dilevko & Gottlieb, 2003, p. 93). For some libraries, circulation

rates also factor into state decisions about funding support (Roy, 1987). However, it is not the case that all items with low circulation should be automatically discarded. Some items enjoy patron use inside the library despite not being checked out (Selth, Koller, & Briscoe, 1992; Slote, 1997). The process of weeding requires the examination of additional factors, which increases the time required to accurately evaluate weeding candidates.

There is a general belief among librarians that weeding can increase circulation rates (Dilevko & Gottlieb, 2003). However, the few experimental studies that have been done to assess the effect of weeding on circulation produced mixed results. In 1973, Slote weeded 20% of the adult fiction collection at the Harrison Public Library and found that circulation increased by 6.2% after six months and by 21.2% after 20 months, while the rest of the collection experienced no change in circulation (Slote, 1997). Moore (1982) conducted a weeding study with "inconclusive" (p. 41) results about the impact on circulation. Preliminary results suggested that weeding had more impact on circulation for Dewey Decimal classes that had high use but little impact on those with low use.

In contrast, Roy (1987) found that weeding produced no statistically significant improvement in circulation. Working with a team, she weeded 10% of the adult and young adult collections in four libraries and compared circulation over the next eight months to that observed in the same eight-month period in the previous year. Unexpectedly, two libraries that weeded experienced an average decrease in circulation of 1%. Two libraries that used both weeding and book displays experienced an average increase of 19%. However, two control libraries that did not weed during this period experienced an average increase of 13%. Roy found that there was a statistically significant difference between the libraries that only used weeding and those that used weeding and displays, but no significant difference

between either group and the control libraries. However, the small number of libraries involved in the study may have limited the statistical power of the experiment.

Despite these mixed results, librarians continue to employ weeding as part of collection maintenance. Dilevko and Gottlieb (2003) conducted a survey of 294 public librarians in the U.S. and Canada about their weeding practices. They found that 33% of public libraries weeded irregularly (e.g., to make space for new acquisitions), 24% conducted weeding in an ongoing (continuous) fashion, and 39% weeded with a specified frequency (e.g., yearly or monthly). The remaining 4% of respondents skipped this question.

Circulation is not the only motivation for weeding. For example, in academic libraries, weeding projects can be inspired by the cost of the storage space consumed by the collection, a desire to reduce duplication of items on different campuses, a reliance on interlibrary loan to get materials to where they are wanted, and an increased reliance on digital sources over print materials (Metz & Gray, 2005).

2.3 Weeding Approaches

There are two primary approaches to weeding. An *inclusive* approach considers each item in turn to decide whether it should be weeded or kept. For example, the widely employed Continuous Review, Evaluation, and Weeding (CREW) method advocates for the ongoing review of the entire collection, item by item, during which weeding occurs as part of a larger process of collection maintenance (Larson, 2012). Inclusive strategies are also often employed for weeding projects in which the goal is to identify a given number of items for withdrawal, often to make room for new items or because the collection is being moved to a new location (Soma & Sjoberg, 2010; Tully, 2011).

An *exclusive* approach instead first identifies the "core collection" for the library and then weeds items that fall outside of this subset. The core collection is defined as the set of items that collectively satisfy a desired fraction of anticipated patron requests, such as 95% or 99%. Exclusive weeding strategies are exemplified by the Trueswell method (Trueswell, 1966). This approach employs past circulation statistics to identify the library's core collection. Trueswell studied the circulation records of two libraries and found that just 20-25% of the collection could satisfy over 99% of patron requests (Trueswell, 1964). This pattern has been observed in many other libraries as well (Silverstein & Shieber, 1996; Slote, 1997). In general, the best strategy to employ is one that accommodates the particular needs of the library's patrons and community as well as the type and size of the library (Swoger, 2014).

2.4 Factors Employed in Weeding Decisions

The policy used to decide whether a given book should be weeded or retained is specific to the library (and sometimes to the librarian). Goldstein (1981) found that only three of eleven TALON (medical) resource libraries had a written policy for weeding. Four other libraries provided general guidelines and then relied on individual judgment and expertise. Twenty years later, in a broader survey of libraries, Dilevko and Gottlieb (2003) found that 75% of libraries reported having a written weeding policy.

Dilevko and Gottlieb (2003) reported that the criteria most often used by librarians to make weeding decisions were circulation statistics, the physical condition of the item, and the accuracy of its information. This is consistent with the CREW method's advice to weed items with low circulation, poor appearance, or poor content (Larson, 2012) and the methods used by previous weeding projects

such as those of the University of Toledo (Crosetto, Kinner, & Duhon, 2008), Concordia University (Soma & Sjoberg, 2010), and Rollins College (Snyder, 2014). A summary of the criteria used by several weeding projects to identify candidates is given in Appendix B. Next, we examine more closely each of the key factors involved in weeding decisions.

Circulation Records. Slote (1997) surveyed the weeding literature and found that past use of the item consistently emerged as the best single criterion for making weeding decisions. One way to characterize past use is the measure of an item's "shelf-time," i.e., the length of time that has elapsed since the item last circulated. Slote advocated shelf-time as the most reliable criterion for objectively determining which books could be weeded with the least impact on patron needs. In his own 1969 study of five libraries, he found that "past use patterns, as described by shelf-time period, are highly predictive of the future use, and can be used to create meaningful weeding criteria" (p. 63).

In practice, however, not everyone agrees. Goldstein (1981) found that none of the eleven libraries that were studied employed shelf-time in their weeding decision making, although they did employ use statistics (e.g., number of checkouts). Others have argued that demand (number of checkouts per year) may be more informative than the time since last checkout (Snyder, 2014).

Physical Condition. Libraries seek to provide materials that are in a useful state. Items that have been damaged (e.g., food spills, ripped pages, water damage, weakened spines, missing pages) are less valuable to patrons and may even become unusable. As items age, they become more vulnerable to physical decay and damage. Sometimes items can be repaired. If they are deemed unusable, the library must decide whether to simply discard the item or to replace it, based on the value of its content to the user community.

Quality of Content. The CREW manual identifies six factors that relate to the quality of an item's content and summarizes them with the acronym "MUSTIE" (Larson, 2012, p. 57). The negative factors include: Misleading (or factually inaccurate), Ugly (worn), Superseded, Trivial (no longer of literary or scientific merit), Irrelevant (to the user community), or the same information can be easily obtained Elsewhere (e.g., interlibrary loan or electronic format).

Other Factors. There are several other factors that may be used by librarians in making weeding decisions. They may consider whether the item is a duplicate of other items in the same collection and whether it is held by other libraries. They may consult book reviews or canonical bibliographies, assess local relevance, track in-house use of the item, and consider unique features of the book. Soma and Sjoberg (2010) developed a standard checklist to be used by all librarians as part of a collaborative weeding effort. The checklist included circulation and browse statistics as well as an indication of whether the item appeared in *Resources for College Libraries* and how many copies were held by other libraries.

2.5 Weeding Project Challenges

Dilevko and Gottlieb (2003) found that the biggest obstacle to weeding that was reported by public librarian respondents was the amount of time that it consumed. The amount of reviewing that can be done is limited by the number of people who can devote time to the task, which varies by library. Concordia University reviewed 25,000 books per year for two years, dividing the work between five weeding teams, and weeded a total of 12,172 items before deciding that this level of review "could not be maintained" and reducing the review rate by 50% (Soma & Sjoberg, 2010). Monmouth University librarians took two years to review 72,500 items and select 12,800 for removal (Dubicki, 2008). Rollins College weeded 20,000 of their collection

of 286,000 items over two years (Snyder, 2014). Wesleyan University weeded 46,000 of ∼90,000 candidates over three years using 17 librarians and 21 staff members (Tully, 2014).

Reducing the time spent on the weeding process would allow more time to be devoted to communication with library patrons, many of whom have concerns about the practice of weeding. For academic libraries, some faculty members may oppose the entire project and refuse to sanction the removal of any titles. Some are concerned about the loss of the scholarly record; discarding any material raises the chance that some key prior contribution will be forever forgotten. Public library patrons may disapprove of discarding items purchased with tax dollars.

In public and academic libraries, librarian and staff time is often devoted to education and overcoming weeding opposition. For example, Wesleyan University librarians attended several faculty meetings and set up a website where interested faculty could review the candidates and vote on which ones should be retained (Tully, 2012). Olin Library at Rollins College also invited patrons to participate in the weeding process: weeding candidates were flagged but remained on the shelf for two months, during which time faculty members were encouraged to browse relevant call number ranges and remove the flag of any book they wanted to keep (Snyder, 2014). Librarians at Virginia Tech worked to head off criticism by employing advance publicity with clear weeding criteria and inviting interested faculty members to review weeding decisions until they were comfortable with the judgment employed (Metz & Gray, 2005). However, they found that too much project visibility was a problem; they collected discards in a bin for recycling, and every few months a patron or member of the public would notice and object to the project on principle. Eventually they ended up redirecting discards to the university's surplus property sales to reduce visibility and objections.

In contrast, Monmouth University found that "faculty and administrators were, in fact, the easier groups to convince that this [weeding] was a worthwhile endeavor" with "no resistance voiced to the project" (Dubicki, 2008). Methods for addressing these concerns were outside the scope of this thesis, except insofar as we were able to incorporate faculty feedback, when available, into the candidate filtering process.

A final challenge is incomplete data about item usage. Circulation records are a ready source of information about patron interest in an item, but they do not capture item use inside the library. Some studies have found that in-house use mirrors that of circulation, while others found that they can be quite different. Selth et al. (1992) found that 11% of the books in their library had in-house use with zero circulation. Weeding based only on circulation records could potentially remove these items despite their evident popularity and utility for visiting patrons.

2.6 Automation Attempts for Weeding Projects

To reduce the time required for weeding, decision support systems have been developed to aid in the identification of the initial list of weeding candidates (Lugg, 2012). These systems require that the librarian specify a list of weeding rules that define which items shall be considered to be weeding candidates. For example, librarians at Wesleyan University specified that weeding candidates were those items that were published before 1990, acquired by the library before 2003, had fewer than two checkouts since 1996, were held by at least 30 U.S. libraries, and were held by at least two partner libraries (Tully, 2011). Given a list of rules, a commercial service such as Sustainable Collection Services (SCS) (Lugg, 2012) applies the rules to the collection and iterates, sometimes many times, with the librarians until the list appears satisfactory in terms of its summary statistics (e.g., number of candidates identified).

Part of this process could be done in-house by the librarians themselves. The main benefit of partnering with a company such as SCS is that it consolidates data from many libraries in one place, enabling easy application of rules that involve the number of holdings in other libraries (lest the single remaining copy of an item be inadvertently discarded) and digital repositories such as the Hathi Trust. This kind of service therefore enables the potential for a collaborative view of weeding with "coordinated deselection" decisions between libraries (Snyder, 2014, p. 21). Actively coordinating weeding decisions between libraries is beyond the scope of this project, but it could be a powerful future capability.

Little research has been done on methods to further improve the quality of the weeding candidate list, once the general rules have been applied. Silverstein and Shieber (1996) used a machine classifier to predict future demand for individual books. Their goal was to support an off-site storage program and to minimize the number of patron requests for items in storage. They evaluated several strategies for predicting future use. The best single criterion was the number of times the item had been checked out in a ten-year period preceding the prediction period, and the next best criterion was the number of months since the item's last checkout, akin to Slote's shelf-time criterion (Slote, 1997). When only a few items were chosen for off-site storage, incorporating knowledge about the LC class of the item increased prediction performance, but when selecting larger groups it was less reliable and sometimes decreased performance. The best result was obtained using a decision tree classifier, which reduced the number of off-site item requests by a factor of five, compared to a policy based only on previous use statistics. While this classifier was designed to support off-site storage decisions, the same approach could be employed to predict which books may be weeded. Silverstein and Shieber's result suggests that methods that employ multiple variables are likely to yield the best performance.

2.7 Summary

There is a consensus opinion amongst librarians that weeding library collections can yield benefits for libraries and library patrons. It is commonly found that a large fraction of a library's collection never circulates but instead consumes shelf space and imposes an overhead on cataloging and collection maintenance, such as the effort required to relocate or re-organize the collection. While the empirical support for specific outcomes such as increased circulation rates is mixed, librarians continue to employ weeding as a part of collection maintenance, and large weeding projects involving the review of most or all of the collection are often required as part of a move to a different building.

The biggest obstacle to weeding projects is the time required to review and make decisions about each item. One natural solution to reduce manual effort is to employ computer automation. However, to date automation has only been employed to create lists of weeding candidates, by applying criteria articulated by the librarians to filter the collection. No methods have been proposed to improve or expedite the subsequent review and decision process that must be applied to each candidate. This gap in the literature inspired the current study which evaluated the potential for machine learning methods to filter and prioritize the weeding candidate list.

Chapter 3

Definition of Research Problem

The high-level motivation for this research was the desire to identify methods to improve the efficiency of the weeding process. This chapter discusses the reasoning behind the study's goal of assessing whether machine learning methods can improve efficiency and then articulates the research questions that drove the study.

3.1 Machine Learning to Assist Weeding Decisions

Weeding decisions are complex and involve both objective and subjective factors (Slote, 1997). While shelf-time (the time elapsed since the last checkout of an item) is a strong predictor of low or no future demand for an item, and therefore its potential for being withdrawn (Slote, 1997), other factors such as physical condition of the item and the accuracy of its information also come into play (Dilevko & Gottlieb, 2003; Larson, 2012; Soma & Sjoberg, 2010). Ranking candidates based on a single factor would fail to capture the complexity of the decision process.

Weeding decision making also varies between different libraries, both when defining the criteria for identifying weeding candidates and when making final weeding decisions. Individual libraries prepare custom checklists for staff members to apply when weeding items from their collection (Dubicki, 2008; Soma & Sjoberg, 2010). Therefore, any automation that is used to assist in the weeding project must adapt to local library priorities and preferences.

Machine learning provides the ability to train a model using past decisions and then apply it to new items, making predictions that are consistent with past practices (Mitchell, 1997). In this way, a machine learning model can provide the flexibility to accommodate individual library criteria and actual decisions made.

Machine learning models also naturally accommodate the incorporation of multiple factors (or variables) into the trained model.

3.2 Research Questions

Goldstein (1981) suggested that "the incorporation of sophisticated quantitative methods into the already existing subjective framework could produce a remarkably reliable and accurate evaluation tool" (p. 314). In this study, we sought to evaluate whether such methods could indeed produce sufficiently reliable and accurate predictions of weeding decisions.

We asked the following research questions:

R1. Can automated data classification methods accurately predict librarian weeding decisions? Since each library employs custom criteria to guide weeding decisions, there will never be one single equation or model that satisfies the needs of all libraries. Automated methods must adapt to local criteria. For machine learning classifiers, this can be achieved by training a local model using labeled examples of previous weeding decisions made at that library. If the model is sufficiently sophisticated, and the examples are representative of the rest of the candidates to be classified, then the predictions made by the classifier should be consistent with librarian decisions on the same items.

To be of use, it is not necessary for the classifier to have perfect agreement with a librarian on all decisions, since the goal is not to replace the librarian with the classifier, but rather to construct an item evaluation tool, as suggested by Goldstein. The classifier's predictions can be used to prioritize (rank) or shorten (filter) the list of weeding candidates, which are then reviewed and confirmed by a librarian.

R2. Which factors are most relevant for making the best predictions of librarian weeding decisions? According to Slote (1997), the criteria most

commonly used to make weeding decisions include physical appearance, duplicate volumes, poor content, foreign language, item age, and circulation statistics. Physical appearance, content, and circulation correspond to the criteria advocated by the CREW method (Larson, 2012) and the most common criteria self-reported by librarians (Dilevko & Gottlieb, 2003). Information about an item's physical condition and the quality of its content are not likely to be available in a weeding data set, but circulation statistics are recorded by all libraries in some form. Information about item age, its availability in digital form, and its presence in other libraries is also readily available.

3.3 Research Hypothesis

The major hypothesis to be tested in this experimental study is stated as follows.

- H_0: There is no statistically significant agreement between librarian and classifier weeding decisions based on item age, circulation, availability in digital form, and presence in other libraries.
- H_a: There is statistically significant agreement between librarian and classifier weeding decisions based on item age, circulation, availability in digital form, and presence in other libraries.

In the remainder of this thesis, we report on an empirical study designed to test the hypothesis using six different machine learning classifiers trained on data from a large-scale university library weeding project.

Chapter 4

Machine Learning Classification and Evaluation

This study evaluated weeding decision agreement between librarian judgments and several state-of-the-art supervised machine learning classifiers. This chapter provides background and definitions for each of the classifiers that were employed, including nearest-neighbor classifiers, naive Bayes classifiers, decision trees, random forests, and support vector machines. This background is followed by a discussion of the theoretical evaluation strategy. The reader may find it useful to refer to terminology definitions in Appendix A.

A supervised classifier analyzes a set of training data, in which each item has been labeled with a human classification decision (the dependent variable). The independent variables that describe each example are organized into a *feature vector* (list of variable values) for use by the classifier. Through this analysis, the classifier constructs a model that captures the observed relationships between the independent and dependent variables.

All machine learning models make the assumption that items with similar feature vectors have similar labels (dependent variable values). Phrased another way, they assume that there is a correlation between some combination of the independent variables and the single dependent variable. If this is not the case, the model may be unreliable and fail to generalize well to new items. This usually means that the representation employed does not adequately capture important variation within the data set. Poor generalization (test) performance can indicate the need for additional variables to be included in the modeling process.

Each classifier type employs a different model representation, and they possess different strengths and weaknesses.

4.1 Nearest-Neighbor Classifier

The simplest approach to classifying an item is to identify the most similar previously classified examples and use them to predict the class of a new observation. A *nearest-neighbor* classifier does not train an explicit model. Instead, it accumulates a database of classified examples. To classify a new item, the algorithm calculates the distance between the new item's feature vector and the feature vectors for all previously stored examples and selects the k examples with the smallest distances (Cover & Hart, 1967). Of the k labels associated with those examples, the most frequently appearing class is predicted for the new item. To avoid ties, k is usually chosen to be an odd number. Since it relies on a distance metric to identify the most similar items, the nearest-neighbor classifier requires that input variables be numeric.

The strengths of the nearest-neighbor classifier are (1) it is fast to construct, since no explicit model need be trained; (2) it makes no assumption about the distribution of classes in the feature space; and (3) its predictions are easy to explain by displaying the k examples that were used to predict the item's label. Its major weakness is that the time required to classify a new item increases with the size of the training data set, since all examples must be considered to find the k cases that are closest to the new item. For a large data set, this classifier can be very slow.

4.2 Naive Bayes Classifier

The *naive Bayes* classifier uses a probabilistic model of data and labels to predict the most likely label for a new item (Duda & Hart, 1973). It relies on Bayes' rule:

$$P(c|x) = P(c)\frac{P(x|c)}{P(x)} \tag{1}$$

where x is an item and c is a class label. $P(c)$ is the prior probability that an item will belong to class c, before the item is observed. $P(x)$ is the probability of

observing item x which is expressed as the probability of observing an item with the feature vector obtained for item x. The conditional probability $P(x|c)$ is the probability of observing the feature vector for item x given that the item comes from class c.

The calculated value of interest, $P(c|x)$, is the probability that item x belongs to class c. Interestingly, it is not necessary to calculate $P(x)$ to classify a new item. This value is the same for all possible values of c, so if one simply wants to identify the highest-probability class, it is sufficient to calculate the numerator only.

The requisite probabilities in the numerator ($P(c)$ and $P(x|c)$) are derived from empirical statistics on the (labeled) training examples. For $P(x|c)$ to be meaningful, there must be at least one previously observed item with exactly the same feature vector as x. However, we do not expect the training set to include every possible combination of feature values; indeed, if it did, then no learning would be needed, since all possible outcomes would be already memorized. Instead, we decompose

$$P(x|c) = \prod_{i=1}^{d} P(x_i|c) \tag{2}$$

where i ranges from 1 to the number of features (d), and $P(x_i|c)$ is the independent probability of observing feature value x_i for feature i in class c. For Equation 2 to hold, it must be the case that the features are statistically independent (not correlated), given class c. This is referred to as a "naive" assumption, and therefore a classifier built on this equation is called a "naive Bayes" classifier. The naive assumption does not always hold (features are often correlated), but in practice, naive Bayes often still performs well.

The strengths of the naive Bayes classifier are (1) it has a probabilistic foundation, so it naturally provides a posterior probability for each prediction that is made; (2) it can accept numeric or categorical inputs; and (3) there are no

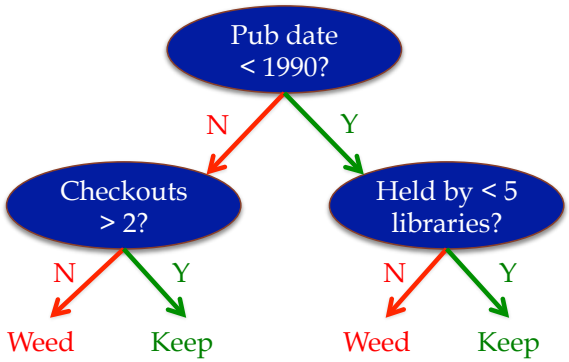


Figure 1: Simple Decision Tree for Classifying Weeding Candidates as "Weed" or "Keep"

parameters to specify. Its primary weakness is that for its predictions to generalize well, the distribution of classes in the training data set must be consistent with the true probabilities of those classes to be observed in new data.

4.3 Decision Tree and Random Forest

Decision trees are another commonly used machine learning method. A decision tree uses a set of labeled items to create a series of tests organized in a tree-shaped hierarchy that allows a new item to be classified (Quinlan, 1986). An illustrative hypothetical decision tree for the weeding task is shown in Figure 1. Starting at the top of the tree, the test at each node is applied to determine which branch to follow until a final prediction is reached. The first test is whether the publication date is earlier than 1990. If not, the next test is whether the number of checkouts in the last ten years is greater than 2. If not, the prediction is "Weed;" otherwise, the prediction is "Keep." If the first test (publication date before 1990) yields a positive result, then the next test is whether the item is held by fewer than five libraries. If not, the prediction is "Weed;" otherwise, the prediction is "Keep." A decision tree is automatically constructed by analyzing a set of previously labeled items and identifying which questions (tests) will correctly classify as many of the items as

possible. Decision trees can accept numeric or categorical inputs. Parameters to specify include the criterion used to select the best feature on which to split the data at each node, the maximum tree depth, and the maximum number of features to consider for each split.

The strengths of a decision tree are (1) it generates an easy-to-understand model that can explain how each prediction was made by simply tracing through the tree, and (2) it can accept numeric or categorical inputs. Its primary weakness is that its posterior probability estimates are usually not very reliable, as they are often calculated from the fraction of examples that reach a given leaf node, which may be a very small sample.

A *random forest* is a collection of decision trees that vote on the classification decision (Breiman, 2001). The forest is "random" in that each individual tree is trained on a data sample that is of the same size but is chosen at random, with replacement, from the full data set. Therefore, by random chance some items will be omitted from a given sample, so each tree develops a slightly different model. In addition, instead of considering all d input variables when constructing a node test, the random forest restricts each node's search to $\sqrt{d}$, thereby introducing additional variation in the tree learning process.

It has been shown that the collective decisions made by a random forest are more reliable than those made by a single decision tree (the *ensemble effect*). In addition, a random forest can generate a real-valued output that characterizes the agreement amongst the ensemble and therefore the confidence in the outcome. Random forests can accept numeric or categorical inputs. Parameters are the same as for a single decision tree, plus the number of trees in the forest.

The strengths of a random forest are (1) it provides more robust decisions due to its ensemble nature, and (2) it can generate a posterior confidence value. Its main

weakness is that because it is composed of many individual models, its interpretability is diminished, as compared to a single decision tree.

4.4 Support Vector Machine

Support vector machines (SVMs) identify a subset of the training data as the "support vectors," which are the items that most strongly constrain a consistent model of output decisions (Cortes & Vapnik, 1995). For example, imagine a binary classification problem with a single independent variable v, in which items with a value for $v < 0$ are assigned to class A, and items with $v \geq 0$ are assigned to class B. The two support vectors that will be chosen are the item from class A with the largest v value and the item from class B with the smallest v value. These two items are the minimal set needed to correctly classify the rest of the items. For data of higher complexity and additional independent variables, more support vectors may be needed to specify the model. The SVM prediction for item x is the weighted sum of the similarity between x and each of the training items x_i, plus a bias offset β:

$$SVM(x) = \sum_i \alpha_i K(x, x_i) + \beta \tag{3}$$

Similarity is defined using a *kernel function*, K, which in its simplest form (linear kernel) is the dot product between the two item vectors. Another common kernel function is the RBF or Gaussian kernel, in which $K(x, x_i) = e^{-\frac{||x-x_i||^2}{2\gamma^2}}$. The training process consists of estimating the weights α_i and bias β. Training items with $\alpha_i > 0$ are considered support vectors. The remaining items have $\alpha_i = 0$.

SVMs require that all inputs be numeric. Parameters to be specified include the type of kernel function K and a regularization parameter C, which specifies an upper limit on the magnitude of each α_i and therefore controls how much influence a single support vector can have. This helps avoid over-fitting to the training data. For RBF kernels, one must also specify a value for γ.

The strengths of SVMs are (1) they are computationally efficient, especially for large data sets, and (2) they have good generalization performance on a wide range of problems. Their major weakness is the lack of interpretability for predictions. If a linear kernel is used, one can calculate an equivalent weight vector w, with one value per feature, that provides insight into which features have more or less influence in the model.

4.5 Evaluation Approach

To test the hypothesis stated in Chapter 3, we required a method to compare the decisions made by a librarian with the predictions generated by a classifier. Examining the fraction of items in which the librarian and classifier agree (accuracy) provided some insight, but it did not allow us to test the hypothesis in a statistical fashion. Instead, we employed two statistical measures of agreement (ϕ and Yule's Q) that permit the assessment of statistical significance by factoring in the amount of agreement that would be expected by random chance. The justification for this choice is provided in the next chapter.

We further investigated the types of errors made by the classifiers using recall and precision measures. This is important to any evaluation of weeding decisions because of asymmetric error costs: the impact of an item that is incorrectly predicted to be weeded is likely to be larger than the impact of an item that is incorrectly predicted to be kept. Statistical measures of agreement do not account for the difference in these two errors, but recall and precision reveal which types of errors are most commonly made. Implementation details for these metrics are provided in the next chapter.

4.6 Summary

In this chapter, we described six machine classification algorithms that each construct a model of prior decisions to enable prediction on future data. The algorithms were nearest-neighbor, naive Bayes, decision tree, random forest, linear SVM, and RBF SVM. We reported the strengths and weaknesses of each method and the parameters that must be specified for each one. We also outlined the evaluation strategy that was employed in the study. The methodology of this study is described in detail in the next chapter.

Chapter 5

Methodology

This chapter describes the design of the empirical study that we conducted to assess whether machine classifiers could generate weeding predictions having sufficiently high agreement with human judgments. Next, the data set, variables, and pre-processing steps are described. We characterize the data set by reporting overall statistics and the correlation of each variable with the weeding decision. Finally, we describe how the classifiers were trained and evaluated and how the research hypothesis was tested using statistical agreement.

5.1 Research Design

This study employed a somewhat unusual experimental design. Rather than testing whether a treatment creates a statistically significant change in a dependent variable, the goal was to test whether librarian and machine-generated weeding decisions about the same set of items were in significant agreement. That is, rather than attempting to predict whether a book should be weeded or kept in an objective fashion, classifiers were designed to construct a model of librarian decisions that could be efficiently applied to a large collection. This is an inclusive approach to weeding (as defined in Section 2.3), and each item was considered independently.

5.2 Research Population and Sampling

The study was structured as a retrospective analysis of data collected by the Wesleyan University Library as part of a large-scale weeding project that took place from 2011 to 2014 under the direction of the Wesleyan University Librarian, Pat Tully (Tully, 2014). The research population (if it can be considered as such) consisted of the items from the library's collection that were identified by Wesleyan librarians and staff as candidates for weeding.

The Wesleyan data set employed in this study was provided by Lori Stethers, systems librarian at the Wesleyan Library, in March 2015. It contained 88,491 weeding candidates. Each item was marked to indicate whether it was withdrawn or kept as a result of the weeding project.

The criteria used to generate the list of weeding candidates were as follows (Tully, 2011):

- Publication date before 1990
- Acquisition date before 2003
- No checkouts since 2003
- ≤ 2 checkouts since 1996
- Held by > 30 other U.S. libraries
- Held by ≥ 2 partner libraries (members of the CTW Consortium, which includes Connecticut College, Trinity College, and Wesleyan University).

To be included in the list of candidates, each item had to satisfy all of the specified criteria.

5.3 Data Collection

The variables available for this study (see Table 1) were limited to what was previously collected as part of the weeding project. The Wesleyan University Library provided information for each item about its publication year (used to calculate *age*), circulation history (*checkouts*), how many other libraries held the same item (*uslibs, peerlibs*), and whether the item was in the Hathi Trust (*hathicopy, hathipub*). Since some of the classifiers in this study require that all inputs be numeric, the study was limited to variables that were naturally numeric or that could be converted into a numeric representation. The variables *hathicopy* and *hathipub* were converted to a representation in which True = 1 and False = 0.

Table 1: *Variables Used to Represent Each Weeding Candidate*

Variable	**Description**	**Type**
age	Number of years between Publication Year and 2012, when the data was collected	integer
checkouts	Number of checkouts since 1996	integer
uslibs	Number of U.S. libraries with a copy of this item, based on OCLC holdings records	integer
peerlibs	Number of peer libraries with a copy of this item, based on OCLC holdings records	integer
hathicopy	Copyrighted digital version exists in the Hathi Trust?	Boolean
hathipub	Public domain digital version exists in the Hathi Trust?	Boolean
facultykeep	Number of Wesleyan faculty votes to keep the item	integer
librariankeep	Number of Wesleyan librarian votes to keep the item	integer
decision	"Keep" or "Weed"	Boolean

Some items had a date of last circulation, but most (77%) did not. While methods exist for inferring missing values in a data set, they are only appropriate if the value exists but is missing (not recorded). For this data set, checkout dates are only available for items checked out since 1996. The remaining items may have been checked out prior to 1996, or they may never have been checked out at all. With no valid observations of checkouts prior to 1996, there is no principled way to infer the possible checkout dates for those items. Since many of the machine learning methods cannot operate on data with missing values, we decided to exclude the shelf-time variable from modeling. It is very possible that higher performance would be achieved if shelf-time information were available for all items.

The Wesleyan data set contained information that covered only two of the six weeding criteria categories identified by Slote (1997). No information was available about each item's physical condition, whether an item was a duplicate of another item, the quality of the item's content, or whether the item was written in a

language not commonly used by patrons of the library. These factors may have been employed by librarians in making their final decisions, but since they are not recorded in the data set, they were not available for use by the machine classifiers. As noted in Chapter 3, data on in-house use of the items was also unavailable.

The Wesleyan study was unusual in its large-scale involvement of university faculty in the weeding project decision process. Faculty members were invited to vote on items that they did not want withdrawn using a web interface (Tully, 2012). The data set contained information about the number of "Keep" votes that each item received from faculty members (*facultykeep*) as well as librarians (*librariankeep*). These variables can potentially capture indirect information about an item's condition and subjective value.

The labeled data set contained an entry for each candidate specifying values for all independent variables (its feature vector) and librarian decision (label variable).

5.4 Data Pre-processing

An initial assessment of data quality and the distribution of values for each feature identified some inconsistencies and errors in the data set. The following steps were applied to correct and reduce the data set:

(1) One item (*Germany's Stepchildren*, by Solomon Liptzin) had an invalid publication year of 5704. This value was replaced by the correct value of 1944.

(2) Items that were marked as part of an "enumeration" (series) were handled separately with different weeding decision criteria during the Wesleyan weeding project. The difference in criteria that were employed could preclude the learning of a consistent model, so these items ($n = 8141$) were excluded from the data set.

(3) Four items had a last circulation date prior to 1996, which was identified as an error. These items were excluded from the data set.

5.5 Data Set Analysis and Characterization

The data set contained 80,346 items, 48,445 (60.3%) of which were marked "Keep" and 31,901 (39.7%) of which were marked "Weed." The minimum, mean, and maximum values for the four variables with more than two distinct values were:

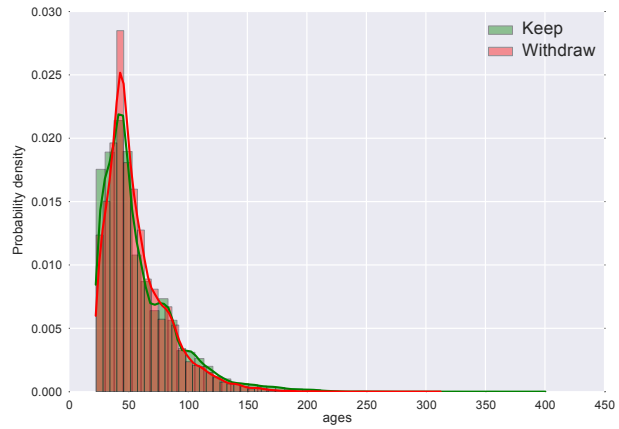
Variable	Units	Minimum	Mean	Maximum
age	years	23	57.59	400
uslibs	libraries	31	544.66	7634
facultykeep	votes	0	0.46	15

Figure 2 shows the distribution of values observed for these three variables. Separate distributions are plotted for items marked "Keep" and "Weed." Figure 2(a) shows that the distribution of ages for items marked "Keep" was shifted slightly lower (younger/newer) than for items marked "Weed." Figure 2(b) shows that the distribution of values for the number of U.S. libraries holding the item was shifted slightly higher (more holdings) for items marked "Weed." Figure 2(c) shows that items with any faculty votes at all are much more likely to be kept than withdrawn.

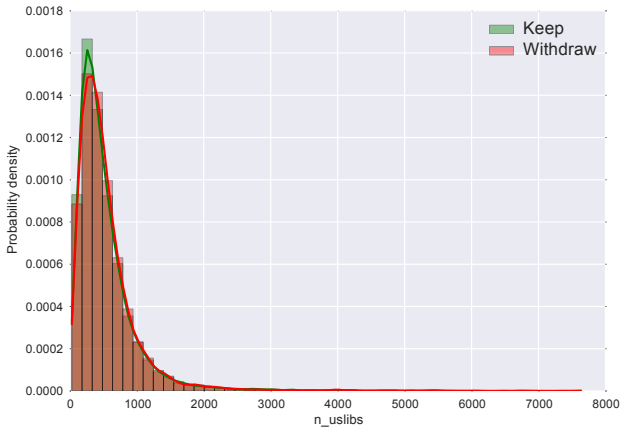
For variables that took on only two possible values, we analyzed the distribution of "Keep" and "Weed" decisions. The *checkouts* variable was dominated by items that had never been checked out: 77% of items in the data set had 0 checkouts. The probability of an item being withdrawn, $P(W)$, given that it had 0 checkouts, was much higher than if it had 1 checkout (0.41 vs. 0.36). The probability that an item was withdrawn, across the whole data set, was 0.40.

checkouts	Number	Fraction	Weed	Keep	$P(W)$
0	61,993	77.16%	25,252	36,741	0.41
1	18,353	22.84%	6,649	11,704	0.36

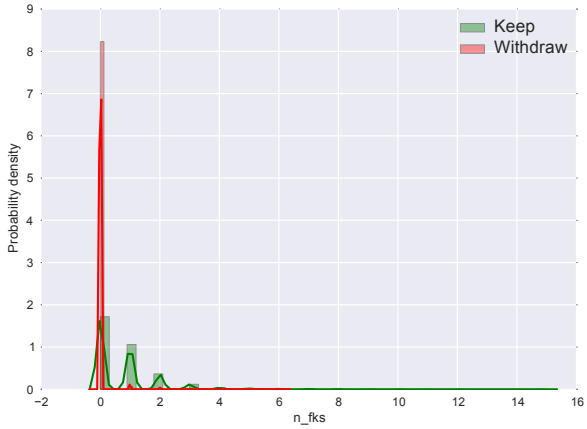
The *peerlibs* variable was less strongly aligned with the weeding decision, but there was still a difference in outcome between items that were held by two peer



(a) *age*



(b) *uslibs*



(c) *facultykeep*

Figure 2: Distribution of Values Observed for Three Variables Compiled Separately for Items Marked "Keep" vs. "Weed"

libraries (59% of the items, with $P(W) = 0.41$) versus those held by three peer libraries (41% of the items, with $P(W) = 0.38$). The latter were less likely to be withdrawn.

peerlibs	Number	Fraction	Weed	Keep	$P(W)$
2	47,738	59.41%	19,600	28,138	0.41
3	32,608	40.58%	12,301	20,307	0.38

Items in the Hathi Trust were more likely to be held in copyright (57%) than to be in the public domain (13%). For those items with a copyrighted version in the Hathi Trust, the probability of being withdrawn was higher (0.42) than the data set average, while those items with a public domain copy had a lower probability of being withdrawn (0.38).

hathicopy	Number	Fraction	Weed	Keep	$P(W)$
False	34,660	43.14%	12,531	22,129	0.36
True	45,686	56.86%	19,370	26,316	0.42

hathipub	Number	Fraction	Weed	Keep	$P(W)$
False	69,997	87.09%	27,932	42,045	0.40
True	10,369	12.91%	3,969	6,400	0.38

Finally, 6% of the items had a value of 1 for the *librariankeep* variable; the rest had a value of 0. There was a very strong relationship between librarian votes and final decisions, as expected: 99% of items with a "Keep" vote were kept (i.e., $P(W) = 0.01$). The 31 items that were withdrawn despite a librarian "Keep" vote were determined to be either lost or duplicates of other items.

librariankeep	Number	Fraction	Weed	Keep	$P(W)$
0	75,926	94.50%	31,870	44,056	0.42
1	4,420	5.50%	31	4,389	0.01

Likewise, the items that received at least one faculty vote ($n = 23,895$) for retention were very likely to be kept. The weeding project overrode the faculty votes for only 398 (1.7%) items.

facultykeep	Number	Fraction	Weed	Keep	$P(W)$
0	56,451	70.26%	31,503	24,948	0.56
>0	23,895	29.74%	398	23,497	0.02

5.6 Classifier Training and Evaluation

We divided the data set into two equal halves: D_t for training and D_e for evaluation. Items were randomly assigned to D_t or D_e. Each machine learning classifier was trained on D_t, with known labels, and then used to generate predictions for D_e, for which the labels were not visible to the classifier.

All of the feature values were normalized to achieve a mean value of 0 and a standard deviation of 1. This is a standard practice that compensates for different ranges in different features, and it tends to improve performance. The shifting/scaling coefficients were determined from D_t, then applied to D_e.

The parameters that were used to train each machine learning classifier are summarized in Table 2. To select parameter values, three-fold cross-validation was conducted on the data in D_t. That is, D_t was further divided randomly into three folds, and a classifier was trained on two of the folds and then evaluated on the third fold, which was not used for training. This process was done three times, so that each held-out fold was evaluated once. This was done for all candidate parameter values, and the values that resulted in the highest held-out performance (accuracy) were selected. For some classifiers (linear and RBF SVMs), all possible parameter values were tested; these classifiers are marked "All" in Table 2. For others (marked "50R" or "100R"), a fixed number (50 or 100) of randomly selected parameter values within the specified range were evaluated, due to computational cost. A final classifier of each type was then trained on all of D_t using the identified parameter values. This classifier was tested on D_e.

Table 2: *Classifier Parameters and Candidate Values Evaluated for Optimization*

Classifier	**Parameter**	**Candidate values**
Nearest neighbor (100R)	Number of neighbors k	$\{1, 2, \ldots, 199, 200\}$
Naive Bayes	None	
Decision tree (100R)	Maximum tree depth Maximum number of features Split criterion	{None, 3, 5} {1, 2, . . . , 7, 8} Gini Index or Entropy
Random forest (50R)	Maximum tree depth Maximum number of features Split criterion Number of trees	{None, 3, 5} {1, 2, . . . , 7, 8} Gini Index or Entropy {10, 20, 50, 100, 500}
SVM (linear kernel) (All)	Regularization parameter C	$10^{\{-10,-9,\ldots,0,1\}}$
SVM (RBF kernel) (All)	Regularization parameter C RBF parameter γ	$10^{\{-10,-9,\ldots,0,1\}}$ $10^{\{-2,-1,0,1,2,3\}}$

5.7 Performance Measures and Hypothesis Testing

To test the study's central hypothesis, the predictions made by each classifier were compared to the librarian decisions for each item in D_e. We employed two statistical measures of agreement: Yule's Q and the ϕ coefficient. There is no single best measure that is widely agreed upon for all possible types of data. Some measures make assumptions about Gaussianity of the underlying trait, and it is not evident that the keep/weed decision would satisfy that assumption. However, Yule's Q and ϕ are two widely used measures that do not make that assumption and therefore are suitable for this task. Yule's Q (Yule, 1900) is based on the odds ratio of two outcomes (agreement and disagreement) to enable the determination of whether the observed agreement is statistically distinguishable from random chance. The ϕ coefficient (Yule, 1912) is an extension of Pearson correlation to dichotomous (binary-valued) data: in this case, the two values are "Weed" and "Keep."

Both measures operate on values calculated as part of a contingency table, which reports the number of occurrences of the possible outcomes for a given

Table 3: *Contingency Table to Tally Agreements and Disagreements Between Librarian Decisions and Classifier Predictions*

	Librarian	
Classifier	Weed	Keep
Weed	a	b
Keep	c	d

prediction (see Table 3). The value a is the number of times that the librarian and the classifier both voted to weed a particular item, and d is the number of times they both voted to keep an item. The values b and c together report the number of times they disagreed, where b is the number of items that the librarian voted to keep and the classifier voted to weed, and c is the number of items that the librarian voted to weed and the classifier voted to keep.

Yule's Q uses the following equation:

$$Q = \frac{ad - bc}{ad + bc}. \tag{4}$$

The ϕ coefficient is calculated as:

$$\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}. \tag{5}$$

Both values can be assessed for statistical significance using a χ^2 test with one degree of freedom. Testing of the null hypothesis was done at the significance level of $p = 0.001$, so we reject the null hypothesis if $p \leq 0.001$ and fail to reject it otherwise.

The amount of agreement between librarian and classifier decisions provides a quantification of the quality of the classifier judgments. However, it does not distinguish between different kinds of disagreements. Incorrect predictions of "Weed" (those items in cell b of the contingency table) likely are worse mistakes than incorrect predictions of "Keep" (cell c). Yule's Q and the ϕ coefficient both

treat these equally. To gain further insight into these types of errors, we also assessed each method in terms of recall, which is sensitive to c, and precision, which is sensitive to b.

We assess recall and precision with respect to the correct identification of items that were labeled "Weed" by humans. Recall is the number of items correctly predicted "Weed" by the classifier (i.e., in agreement with human decisions) divided by the total number of items labeled "Weed" by humans. With respect to the contingency table in Table 3, recall (R) is defined as:

$$R = \frac{a}{a+c}. \tag{6}$$

Precision (P) is the number of items correctly predicted "Weed" (in agreement with human decisions) divided by the total number of items predicted "Weed" by the classifier:

$$P = \frac{a}{a+b}. \tag{7}$$

Finally, accuracy (A) is defined as the fraction of classifier predictions that agree with human decisions, including both "Weed" and "Keep" items:

$$A = \frac{a+d}{a+b+c+d}. \tag{8}$$

We assess the statistical significance of recall, precision, and accuracy scores with a univariate χ^2 analysis by testing the observed values against those expected by a random process. The expected values of recall, precision, and accuracy are determined as follows.

Let N be the total number of items in the data set, $P_{rand}(W)$ be the probability that an item is marked "Weed" by a random process, and $P_{label}(W)$ be the probability that an item is labeled "Weed" by a human. Since there are only two prediction outcomes, "Weed" or "Keep," $P_{rand}(W) = 0.5$. From our analysis of the

data set (Section 5.5), we know that $P_{label}(W) = 0.40$. The expected number of items correctly predicted as "Weed" (a) is $N \times P_{rand}(W) \times P_{label}(W)$. The expected number of items labeled "Weed" by human decision ($a + c$) is $N \times P_{label}(W)$. Then the expected value of recall is:

$$\begin{aligned} E[R] &= \frac{E[a]}{E[a+c]} && (9)\\ &= \frac{N \cdot P_{rand}(W) \cdot P_{label}(W)}{N \cdot P_{label}(W)} && (10)\\ &= \frac{N \cdot 0.5 \cdot 0.4}{N \cdot 0.4} && (11)\\ &= \frac{N \cdot 0.2}{N \cdot 0.4} && (12)\\ &= 0.5 && (13) \end{aligned}$$

Likewise, the expected value of precision is:

$$\begin{aligned} E[P] &= \frac{E[a]}{E[a+b]} && (14)\\ &= \frac{N \cdot P_{rand}(W) \cdot P_{label}(W)}{N \cdot P_{rand}(W)} && (15)\\ &= \frac{N \cdot 0.2}{N \cdot 0.5} && (16)\\ &= 0.4 && (17) \end{aligned}$$

For a random predictor with two outcomes, the expected value of accuracy, $E[A]$, is 0.5.

A non-parametric χ^2 test was used to determine the statistical significance of each ratio value. The random process was modeled as generating a binary variable with two possible outcomes. The predicted probability of each outcome, $P(o_i)$, was compared with its observed probability, $O(o_i)$:

$$\chi^2 = N \cdot \left(P(o_1) \left(\frac{O(o_1) - P(o_1)}{P(o_1)} \right)^2 + P(o_2) \left(\frac{O(o_2) - P(o_2)}{P(o_2)} \right)^2 \right) \qquad (18)$$

The calculated χ^2 value was checked against a standard table of χ^2 distribution values to determine the probability of observing the difference between $P(o_i)$ and $O(o_i)$ by chance, which is the significance level (p-value) for the observed values.

To assess the statistical significance in differences in performance values (recall, precision, or accuracy) between two classifiers (V_1 and V_2), we calculated a z-score based on the ratio of the observed difference to the standard error observed in the combined sample. Since the ratio scores are in the range $[0, 1]$, we first converted them into the range $[0, 100]$ by multiplying each value V_i by 100. Let $\bar{V}$ be the average of V_1 and V_2.

$$z = \frac{V_1 - V_2}{\sqrt{\frac{2}{N}\bar{V}(100 - \bar{V})}} \tag{19}$$

The calculated z-score was checked against a standard table of z distribution values to determine the significance level (p-value).

5.8 Implementation Details and Project Timeline

All experiments and data analyses (including the initial variable analysis reported in Section 5.5) were implemented in Python using a combination of new code and the freely available scikit-learn library for implementing the machine learning classifiers and computing ϕ. Yule's Q was implemented according to Equation 4. Experiments were conducted on a MacBook Air laptop running OS X 10.7.5 with a 1.7 GHz Intel Core i5 processor and 4 GB of RAM. The Python interpreter was version 2.7.5.

The data set was acquired from the Wesleyan University Library in March 2015. We iterated several times with Wesleyan to address some inconsistencies in the data set. Initial experimental results were complete by July 2015 and the process of writing up the results began. In September 2015 we discovered a discrepancy in that some items in the data set had circulated after 2003, yet one of the selection

criteria was that items should not have circulated after 2003. A follow-up discussion with Wesleyan resulted in the determination that these items should be excluded from the analysis.

After re-running the experiments, another obstacle appeared in terms of the statistical measure of agreement that was used (ϕ). It was generating inconsistent results due to the methodology employed (comparing agreement as a function of classifier confidence threshold). The ϕ values were not comparable because the set of data that passed this threshold was different for each classifier. The ϕ statistic requires that the underlying data, or at least its marginal distribution, be fixed for a comparison to be valid (Liu, 1980). After extensive research into other agreement measures and assessment of the assumptions they imposed on the data set, we identified Yule's Q as a better choice for measuring agreement. After more consideration, we abandoned the idea of assessing classifier performance as a function of confidence threshold and instead decided to measure the classifier outputs directly. That meant that ϕ could once again be employed to compare different classifiers, and it simplified the presentation of results greatly.

The final experimentation was completed by December 2015 and the analysis of the results was completed by March 2016.

5.9 Summary

This chapter outlined the methodology of the study. It described the source of the data set as well as its properties and the features that were used to describe each item. An initial assessment of the distribution of values for each feature and their association with the weeding outcome provided initial insights into the contents of the data set. It also described the process of training the classifiers as well as the evaluation of their performance in terms of accuracy, recall, precision,

and agreement, to enable statistical testing of the hypothesis. The chapter concluded with an overview of the research implementation.

Chapter 6

Experimental Results and Discussion

This chapter presents the experimental results, discusses research findings in the context of prior studies, and reflects on the limitations of the study in terms of data set quality, factors excluded from modeling, and other potential pitfalls.

6.1 Results

6.1.1 Parameter selection results. After performing three-fold cross-validation on the training set, D_t, the parameter values were selected for each classifier as shown in Table 4. The number of neighbors used by the nearest-neighbor classifier is quite high (165). This indicates that the items may not be neatly divided into "Keep" and "Weed" groups within the feature space. Instead, they are mixed together, and a large number of neighbors is needed to get a robust vote on the correct prediction. The naive Bayes classifier, as previously noted, does not have any parameters to set.

The decision tree was allowed to use up to five features (of the eight available), and the maximum tree depth was also five (with no pruning). The decision tree's first split was on the *librariankeep* feature.

In contrast, the random forest was composed of 100 trees, each of which were only allowed to use three features and to have a depth of three. Shallower trees tend to generalize better to new data, but they may miss finer nuances. Interestingly, the single decision tree used the Gini Index to determine how to split nodes, while the random forest used the entropy criterion. Either one is acceptable for decision trees.

The linear SVM employed a regularization parameter (C) value of 0.001, which is very small. This signals that the data may not be well modeled by a linear separation, which is consistent with the high k value chosen by the nearest neighbor

Table 4: *Parameter Values That Were Selected for Six Machine Learning Classifiers Using Cross-validation on the Training Set*

Classifier	**Parameter**	**Best value**
Nearest neighbor	Number of neighbors k	165
Naive Bayes	None	
Decision tree	Maximum tree depth	5
	Maximum number of features	5
	Split criterion	Gini Index
Random forest	Maximum tree depth	3
	Maximum number of features	3
	Split criterion	Entropy
	Number of trees	100
SVM (linear kernel)	Regularization parameter C	0.001
SVM (RBF kernel) (All)	Regularization parameter C	10
	RBF parameter γ	10

classifier. In contrast, the RBF SVM selected a C value of 10, which means that its more complex modeling yielded a better fit to the data. The RBF parameter (γ) was set to 10. This parameter is an intuitive measure of how far the influence of a given example reaches, in feature space. A value of 10 is medium-large, indicating a relatively small radius of influence for a given item. One can interpret this to indicate a heterogeneous feature space in which classes may be interspersed rather than cleanly separated.

6.1.2 Learned models. As noted earlier, the nearest neighbor classifier does not learn an explicit model, so there is no model to discuss.

The naive Bayes model consists of the conditional class probabilities for each feature, estimated from the training data (given in Section 5.5).

One of the strengths of the decision tree classifier is that it creates models that are easy to interpret. The top three layers of the trained decision tree are shown in Figure 3. The most likely outcome after three layers of testing, and its associated

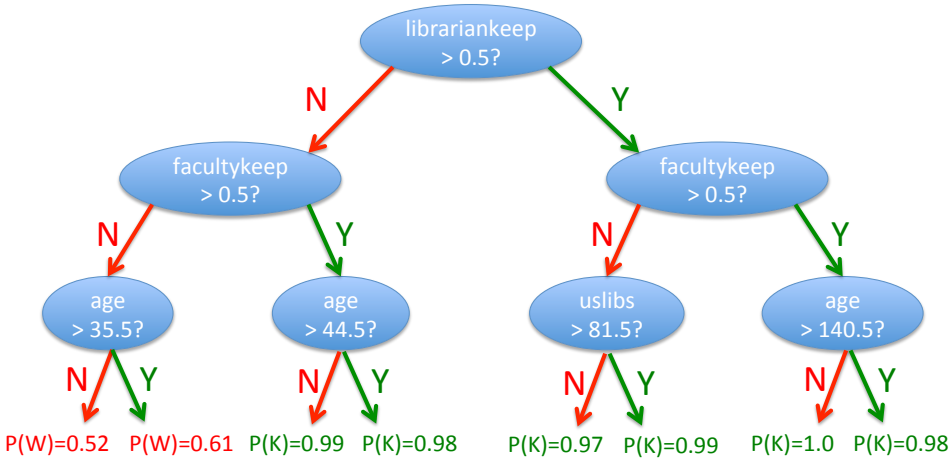


Figure 3: Top Three Layers of the Learned Decision Tree Model

probability, is shown at the bottom of the diagram; the complete classifier conducts two more layers of tests before classifying a given item. The first feature that the classifier tests is *librariankeep*. If there is at least one librarian who voted to keep the item, processing moves to the right sub-tree. There, if at least one faculty member voted to keep the item, the age of the item is tested. For all of the items that follow the first branch to the right, the most likely outcome is "Keep." There are very few exceptions; the probability of "Keep" ranges from 0.97 to 1.0.

If no librarians voted to keep the item, then the first left branch is followed and the classifier likewise checks whether any faculty members voted to keep the item. Any such votes lead to an age check, and the most likely outcome is again "Keep" with probability 0.98 to 0.99. However, items that had neither a faculty vote nor a librarian vote to be kept are most likely to be weeded. This left sub-branch is where most of the complexity and uncertainty exists in the model. The probability of the most likely outcome ("Weed") is higher for older items (those older than 35.5 years), but the probability is still only 0.61, indicating that there are many items in this group that should be kept. For younger items, the probability of being weeded

is not much more than random chance (0.52). The next feature to be checked (not shown in the figure) is *checkouts*, then *uslibs* and *hathicopy*. However, none of these checks served to improve the separation of "Weed" and "Keep" items by much.

The structure of the tree is consistent with the individual feature assessment in Section 5.5. The *librariankeep* and *facultykeep* features had the strongest discriminatory power between the two classes, while the other variables provided less separation. The decision tree structure indicates that even when combining the features in a sequence of tests, accurate weeding decisions on some items remain difficult to achieve.

The random forest used an ensemble of 100 decision trees, which would be tedious to examine individually. However, it also produces a consensus estimate of the importance of each feature based on how often it is employed in individual trees. The most important feature was *facultykeep* (0.77), followed by *librariankeep* (0.19) and *age* (0.01).

The support vector machine models do not lend themselves well to interpretation. They are generally treated as black boxes that generate good predictions but do not provide insights into the data being classified.

6.1.3 Performance results. Table 5 presents the test performance for the baseline approaches (keep all items, weed all items) and the six machine learning methods. The best value(s) for each column is/are in bold; top values that are not significantly different from each other, as determined by z-score tests with $p = 0.01$, are all shown in bold (ties).

All of the accuracy values in Table 5 were found to be statistically significantly better than random performance ($\chi^2, p = 0.001$), except for the "Weed-all" baseline, which was significantly worse than random performance. All of the recall values were significantly better than random ($\chi^2, p = 0.001$), except for the "Keep-all"

Table 5: *Performance Statistics of Predicting Weeding Decisions by Different Classifiers*

Method	Accuracy	Recall	Precision	ϕ	Yule's Q
Baseline (all "Keep")	0.600	0.000	0.000	N/A	N/A
Baseline (all "Weed")	0.400	1.000	0.400	N/A	N/A
K-nearest-neighbor	**0.721**	0.869	**0.605**	0.486	0.832
Naive Bayes	**0.724**	0.980	0.594	**0.552**	0.968
Decision tree	**0.725**	0.967	0.596	0.545	0.949
Random forest	**0.724**	0.919	**0.601**	0.516	0.887
SVM (linear)	**0.725**	**0.986**	0.594	**0.557**	**0.978**
SVM (RBF)	**0.725**	0.954	0.598	0.537	0.930

baseline, which was significantly worse. For precision, all of the classifiers achieved values that were significantly better than random ($\chi^2, p = 0.001$), but both of the baselines were significantly worse. As described in Section 5.7, ϕ and Yule's Q are statistical measures of agreement. In the table, all of the ϕ and Yule's Q values were statistically significant ($\chi^2, p = 0.001$), despite the difference in their range of values.

All of the machine learning classifiers achieved approximately the same level of accuracy (72%). This performance is well above the best baseline performance of 60%, which would be achieved by not withdrawing any items. However, a z-score test with $p = 0.05$ found that there is no significant difference in accuracy between classifiers.

Nevertheless, although all classifiers reached about the same level of accuracy, they did not all make the same type of errors. As noted in Chapter 5, analysis of recall and precision values allows the determination of whether a given classifier is more likely to be wrong when predicting "Keep" or when predicting "Weed." The recall values varied noticeably across classifiers, but the precision values showed little difference between them (albeit statistically significant). This suggests that

some classifiers were better than others at correctly identifying items that should be withdrawn, but all of the classifiers struggled to improve precision beyond about 60%. That is, there is a large number of items in the data set that should be kept but are difficult to distinguish, given the features available, from those that should be withdrawn.

Specifically, the K-nearest-neighbor classifier had the lowest recall (86.9%) but the highest precision (60.5%), which indicated that it was more likely to mistakenly predict "Keep" for an item that was actually withdrawn, but its "Weed" predictions were most reliable. This classifier had significantly higher precision than the naive Bayes, random forest, and linear SVM classifiers (z-score, $p = 0.01$) and to a lesser degree compared to the decision tree and RBF SVM (z-score, $p = 0.05$). In contrast, the linear SVM had the highest recall (98.6%) but the lowest precision (59.4%), which indicated that it was more likely to mistakenly predict "Weed" for an item that was actually kept. It had significantly higher recall than all other classifiers (z-score, $p = 0.01$).

The hypothesis behind this study was that machine learning classifiers could obtain a statistically significant level of agreement with human weeding decisions. The null hypothesis was that there would not be significant agreement. The ϕ and Yule's Q results in Table 5 cause us to reject the null hypothesis ($\chi^2, p = 0.001$). The two measures are not directly comparable in terms of their values, but we found that they ranked the methods identically. The naive Bayes and linear SVM classifiers had the highest ϕ values (0.552 and 0.557; these were not significantly different at the $p = 0.01$ level). The linear SVM had a significantly higher Yule's Q value (0.978), compared to all other classifiers, for the same p-value. The K-nearest-neighbor classifier had the lowest agreement (ϕ of 0.486 and Yule's Q of 0.832).

Like accuracy, ϕ and Yule's Q do not distinguish between different types of errors. In this study, we found that they correlated well with recall (R=0.918 for ϕ, R=0.999 for Yule's Q) but did not correlate well with precision (R=0.196 for ϕ, R=0.003 for Yule's Q). Thus, classifiers with high recall values tended to have higher agreement values as well, regardless of their precision.

6.2 Discussion

The experimental results indicate that we should reject the null hypothesis that there is no statistically significant agreement between human decisions and automated classifier predictions ($\chi^2, p = 0.001$). In fact, there was significant agreement for all six classifiers that were tested, based on the statistical results of two measures of agreement (ϕ and Yule's Q).

In practice, one particular model would be selected and employed to assist in a given weeding project. Which one should be selected? For this particular task (employing a classifier to aid in making weeding decisions), precision is more important than recall, since mistakenly discarding an item has more impact than mistakenly keeping it. It is important that any prediction for an item to be weeded is highly reliable. Problems with this kind of asymmetric "cost" (impact of different types of errors) are common in machine learning applications, and inspecting recall and precision performance aids the user in selecting the most appropriate model.

This would lead us to favor the behavior of the K-nearest-neighbor classifier or the random forest, even though they did not have the highest accuracy or agreement values. However, these outcomes could change if the same experiment were conducted with a different set of items or with data from another library. Fortunately, there is little cost to training and evaluating all models on a new data set to enable a similar assessment and selection of the most appropriate classifier.

An opportunity for improving classifier performance stands out in these results. The inability of all six classifiers to achieve precision above 60% indicates that there is a need for additional descriptive features to capture the information that distinguishes the items with an incorrect "Weed" prediction from those that should truly be withdrawn. These items are the ones in the leftmost branch of the decision tree in Figure 3. Shelf-time, physical condition, or other factors might be the key to correctly handling them. Inspection of specific examples from this group could direct the collection of additional variables.

6.3 Limitations

There are some important limitations of this study. First, we were unable to make use of the shelf-time information for these items. These values were missing for the majority of the items in the data set. We wanted to be able to use the data with all of the classifier candidates, but not all of them can handle missing values. Decision trees are an exception, and they would be able to access this additional source of information, which Slote (1997) claimed was the single best predictor of future item use.

Another option would be to incorporate shelf-time information in a different way. Last checkout dates were available only for checkouts since 1996. Instead of trying to capture the total shelf-time, which is unavailable for 77% of the items, we could create a "shelf-time since 1996" variable that would be set to 16 years for items with no checkout record between 1996 and 2012 when the data was collected.

Second, there are several variables relevant to weeding decisions that are unlikely or impossible to be made available to a machine classifier without further librarian intervention, such as the physical condition of the item, the quality of its content, and its in-house use. This places an (unknown) upper limit on the ability of any

classifier to accurately capture human weeding decisions. However, the classifier will only be used to filter the candidates, not to make a final decision. Its output will still be reviewed by a librarian, who can employ additional judgment based on factors not available to the classifier.

Third, to our knowledge an assessment of quantitative agreement between librarians on weeding decisions has never been attempted empirically. Since libraries differ in their policies and individual librarians may differ in their application of subjective criteria, it is likely that inter-librarian agreement is not perfect. Thus, we do not know what to consider as the best achievable agreement in reality; it is probably not $\phi = 1.0$. Assessing agreement between librarians would help us interpret the classifier performance in context: is there a lot of room for improvement, or is it already as good as a human?

Fourth, the results of this evaluation may not generalize well to new library collections. Evidence suggests that libraries differ enough in their weeding strategies (Swoger, 2014) that the classifier would have to be re-trained to create a custom model for each new library according to the local librarians' weeding philosophy as expressed in their past weeding decisions. Depending on the particular nature of those decisions, the classifier could achieve higher or lower agreement that that observed on the data set employed in this study.

Despite these limitations, the experimental study demonstrated the feasibility of a machine learning approach to constructing a model of human weeding decisions and then applying it to new items in a consistent fashion. Accuracy, recall, precision, and agreement values indicate high performance that could be used to reliably rank weeding candidates for prioritized review by a human librarian and save significant time by directing efforts to the items most likely to require weeding.

6.4 Summary

The results of the study show statistically significant agreement between human and classifier decisions about which items to weed or keep. Agreement was highest for the naive Bayes and linear SVM classifiers. These classifiers had high recall but low precision. The k-nearest-neighbor classifier provided the highest precision but lowest recall. Judgment about the impact of different types of prediction errors is needed to select between the classifiers for the one that best suits a given weeding project.

Analysis of the learned models provided insight into which variables were most relevant for distinguishing between candidates that should be weeded or kept, in this data set. Librarian and faculty votes for retention emerged as the most important features, while item age and presence in other U.S. libraries were also important.

Chapter 7

Conclusions and Future Work

Lack of time is cited as the biggest obstacle to effective weeding projects (Dilevko & Gottlieb, 2003). Current automation to assist weeding efforts is limited to the *a priori* specification of general weeding rules that are used to generate a list of weeding candidates. The time required to review this candidate list can be formidable, and the project may require the efforts of a large number of staff members to complete. Because collection review and weeding is relevant for all libraries at some point, and perhaps even as a continual ongoing process (Larson, 2012), methods that can further reduce the amount of human effort required to accomplish large weeding projects are vital.

This study provides the first empirical analysis of the agreement with human weeding decisions that can be achieved by machine learning classifiers. The machine learning methods provide a data-driven approach to building a model that predicts human decisions. The learned models do not replace humans, but they can provide an initial assessment of the candidate list, which allows librarians to focus their time and attention on those items most likely to be weeded.

7.1 Key Research Findings

All six machine learning classifiers had statistically significant agreement with human judgments. Research question R1 ("Can automated data classification methods accurately reproduce librarian weeding decisions?") was answered by Yule's Q and ϕ values that caused us to reject the null hypothesis at the $p = 0.001$ level.

We also analyzed the models that were learned to address research question R2 ("Which factors are most relevant for obtaining accurate weeding predictions?"). We found that the features representing librarian and faculty votes to keep certain

items were prioritized by the classifiers, followed by the features relating to item age and presence in other libraries. These findings are consistent with the literature on factors relevant to weeding decisions.

7.2 Recommendations for Machine Learning in Weeding Projects

The results of this empirical study suggest that machine learning classifiers can improve the efficiency of weeding projects by pruning or prioritizing the list of weeding candidates prior to their review by a librarian. Because the weeding criteria are defined differently in each library, and because weeding decisions often include an element of subjectivity, it is unlikely that a generic classifier trained on decisions made at one library could be directly applied to the collection at another library. Instead, we recommend that each library label a portion of their weeding candidates to provide a custom, library-specific set of training examples.

Most classifiers, including those used in this study, output a posterior confidence in their predictions as well as the binary outcome. The librarian can filter the list of candidates by specifying a minimum confidence threshold (τ) and generating a list of only those weeding candidates whose prediction confidence is greater than τ. For greater flexibility, we recommend sorting the entire candidate list by posterior confidence values, so that the librarian can start with the candidates most likely to be weeded and work down the list as time permits. Given good agreement between librarian and classifier decisions, the sorted list will direct the librarian quickly to items most likely to be withdrawn.

7.3 Future Research Directions

There are several directions for further research on the best use of machine learning classifiers to assist in weeding projects. First, it would be valuable to determine the minimum number of librarian-labeled examples that are needed to

train a model that can attain a certain level of accuracy or agreement. As noted above, it will be necessary for each library to label a representative set of examples for training a customized model of that library's weeding practices. In this study, we used half of the data set to train each model, which amounted to more than $40,000$ labeled items. It is desirable to do the same evaluation with progressively fewer labeled items to determine whether the same performance could be achieved with less up-front effort.

In practice, for a new weeding project, one could start with a small collection of labeled items and progressively increase the number of items until the desired level of performance is achieved. Another promising area of investigation is the use of active learning (Cohn, Ghahramani, & Jordan, 1996) in which the machine learning method starts with very few labeled examples and then actively chooses which items the librarian should label to provide the most informative labels first. This strategy has been shown to dramatically reduce the number of labeled items required.

Second, this study included only information about age, circulation, and other library holdings. As discussed in Chapter 2, there are many other potentially useful features that could not be evaluated in the present study. These include information about the item's physical condition, quality of content, library in-house use, etc. A similar empirical study with data that included those variables would help to determine their value for predicting weeding decisions. An ablative study, in which each variable in turn is eliminated from the data set and the same evaluation re-run, would help determine which variables contribute relevant information and which do not.

Finally, the vital next step in evaluating this approach is to conduct an experimental study in parallel with an active human weeding project. One method would be to collect data about items that are under consideration for weeding, then

split the list in half. For half of the items, the standard approach of manual review by librarians would be used. For the other half of the items, a classifier would be trained and then used to prioritize the items for subsequent review. The total human time required to identify weeding candidates in each half of the data set would be tracked, and the outcomes would be measured in terms of circulation rate, or the metric of most importance to the weeding project, for each half.

Ultimately, our goal is to facilitate weeding projects and reduce the burden that they currently impose on librarians in terms of time and effort. While most librarians feel that weeding is an important and necessary process, the most common complaint is that it takes too much time (Dilevko & Gottlieb, 2003). It may never be possible to fully automate the weeding process, but the use of automation to provide decision support to busy librarians has the potential to reduce that burden significantly.

References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, *4*, 129–145.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Crosetto, A., Kinner, L., & Duhon, L. (2008). Assessment in a tight time frame: Using readily available data to evaluate your collection. *Collection Management*, *33*(1–2), 29–50.

Dilevko, J., & Gottlieb, L. (2003). Weed to achieve: A fundamental part of the public library mission? *Library Collections, Acquisitions, & Technical Services*, *27*, 73–96.

Dubicki, E. (2008). Weeding: Facing the fears. *Collection Building*, *27*(4), 132–135.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis.* New York: Wiley-Interscience.

Goldstein, C. H. (1981). A study of weeding policies in eleven TALON resource libraries. *Bulletin of the Medical Library Association*, *69*(3), 311–316.

Kent, A., Cohen, J., Montgomery, K. L., Williams, J. G., Bulick, S., Flynn, R. R., Sabor, W. N., & Mansfield, U. (1979). *Use of library materials: The University of Pittsburgh study.* New York: Dekker.

Larson, J. (2012). *CREW: A weeding manual for modern libraries.* Austin, TX: Texas State Library and Archives Commission.

Liu, R. (1980). A note on phi-coefficient comparison. *Research in Higher Education*, *13*(1), 3–8.

Lugg, R. (2012). Data-driven deselection for monographs: A rules-based approach to weeding, storage, and shared print decisions. *Insights*, *25*(2), 198–204.

Metz, P., & Gray, C. (2005). Public relations and library weeding. *The Journal of Academic Librarianship*, *31*(3), 273–279.

Mitchell, T. M. (1997). *Machine learning.* McGraw-Hill.

Moore, C. (1982). Core collection development in a medium-sized public library. *Library Resources & Technical Services*, *26*(1), 37–46.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Roy, L. (1987). *An investigation of the use of weeding and displays as methods to increase the stock turnover rate in small public libraries.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Selth, J., Koller, N., & Briscoe, P. (1992). The use of books within the library. *College & Research Libraries*, *53*(3), 197–205.

Silverstein, C., & Shieber, S. M. (1996). Predicting individual book use for off-site storage using decision trees. *The Library Quarterly*, *66*(3), 296–293.

Slote, S. J. (1997). *Weeding library collections: Library weeding methods* (Fourth ed.). Littleton, CO: Libraries Unlimited.

Snyder, C. E. (2014). Data-driven deselection: Multiple point data using a decision support tool in an academic library. *Collection Management*, *39*(1), 17–31.

Soma, A. K., & Sjoberg, L. M. (2010). More than just low-hanging fruit: A collaborative approach to weeding in academic libraries. *Collection Management*, *36*(1), 17–28.

Swoger, B. (2014). Books are for use: Weeding and deselecting. Scientific American Information Culture Blog. Retrieved from http://blogs.scientificamerican.com/information-culture/2014/01/07/books-are-for-use-weeding-and-deselecting/.

Trueswell, R. W. (1964). *User behavioral patterns and requirements and their effect on the possible applications of data processing and computer techniques in a university library.* Unpublished doctoral dissertation, Northwestern University.

Trueswell, R. W. (1966). Determining the optimal number of volumes for a library's core collection. *Libri*, *16*, 49–60.

Tully, P. (2011). More than you want to know about weeding criteria. Retrieved from http://weeding.blogs.wesleyan.edu/2011/09/26/more-than-you-want-to-know-about-weeding-criteria/.

Tully, P. (2012). Update: April 10, 2012. Retrieved from http://weeding.blogs.wesleyan.edu/2012/04/16/update-april-10-2012/.

Tully, P. (2014). Project wrap-up - July 31, 2014. Retrieved from http://weeding.blogs.wesleyan.edu/2014/07/30/project-wrap-up-july-31-2014/.

Yule, G. U. (1900). On the association of attributes in statistics. *Philosophical Transactions of the Royal Society A*, *75*, 257–319.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, *49*(6), 579–652.

Zuber, P. (2012). Weeding the collection: An analysis of motivations, methods and metrics. In *Proceedings of the American Society for Engineering Education Annual Conference* (pp. 6139–6152). Austin, TX.

Appendix A

Definition of Terms

An *automated classifier* is a computer model that was trained to make predictions in a manner consistent with previously recorded decisions for training examples.

Data mining is the use of statistical methods to extract meaningful characterizations of relationships within a data set.

A *feature vector* is the list of values used by a machine learning classifier to represent a real-world item.

A *label* is the judgment ("Keep" or "Weed") assigned by a librarian to a given item.

Machine learning is the process by which an automated classifier is trained on examples and applied to new data to make predictions.

Weeding is the selective removal (withdrawal) of library items that are outdated, physically worn, no longer relevant to patron interests and needs, and/or available in electronic form.

A *weeding candidate* is an item that has been included in the list of items for review and possible withdrawal.

Appendix B

Weeding Criteria Used in Prior Weeding Projects

Concordia University's Carl B. Ylvisaker Library employed the following variables during a weeding project that reviewed 25,000 items during 2007 and 2008 and removed a total of 12,172 (Soma & Sjoberg, 2010):

- Last circulation date
- Browse count
- Whether the item appears in *Resources for College Libraries*
- Whether there are more than five copies at other U.S. libraries

The article does not specify what thresholds were used to convert these variables into decision criteria.

The Olin Library at Rollins College conducted a weeding project from 2010 to 2012 that removed more than 20,000 items from the collection (Snyder, 2014). The criteria that they used to create the candidate list were:

- Acquired before January 1, 1996
- No in-house use or circulation since January 1, 1996
- More than 100 U.S. libraries hold the item
- Either the University of Florida or Florida State University holds the item
- Not in *Resources for College Libraries* or *Choice Reviews*
- Not about Florida (local interest)

All criteria had to be satisfied for an item to be included in the candidate list.

Wesleyan University creating a list of ∼90,000 candidates using these criteria (Tully, 2011):

- Fewer than two checkouts since 1996
- Published before 1990

- Acquired before 2003
- More than 30 U.S. libraries hold the item
- At least two Wesleyan partner libraries hold the item

Again, all criteria had to be satisfied for an item to be included in the candidate list.