Spring February 14, 2018

Automated Classification to Improve the Efficiency of Weeding Library Collections

Kiri Lou Wagstaff
Geoffrey Liu, *San Jose State University*

The Journal of Academic Librarianship 44 (2018) 238–247

Contents lists available at ScienceDirect

The Journal of Academic Librarianship

journal homepage: www.elsevier.com/locate/jacalib

Automated Classification to Improve the Efficiency of Weeding Library Collections

Kiri L. Wagstaff*, Geoffrey Z. Liu

School of Information, San Jose State University, One Washington Square, San Jose, CA 95192-0029, United States

A B S T R A C T

Previous studies have shown that weeding a library collection benefits patrons and increases circulation rates. However, the time required to review the collection and make weeding decisions presents a formidable obstacle. This study empirically evaluated methods for automatically classifying weeding candidates. A data set containing 80,346 items from a large-scale weeding project running from 2011 to 2014 at Wesleyan University was used to train six machine learning classifiers to predict a weeding decision of either 'Keep' or 'Weed' for each candidate. The study 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 variables found to be most relevant were: librarian and faculty votes for retention, item age, and the presence of copies in other libraries.

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 A key component of this process is weeding, the selective removal of items that are outdated, physically worn, no longer relevant to patron interests and needs, and/or available in electronic form. Librarians generally agree that weeding benefits not only the library by reducing the number of items that have to be maintained, but also the user population by making desired items easier to find (Dilevko & Gottlieb, 2003). There is also a general belief that pruning the collection to remove unwanted items can increase library circulation rates, although experimental studies assessing the effect of weeding on circulation have yielded mixed results (Moore, 1982; Roy, 1987; Slote, 1997). Weeding 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 these benefits, weeding is often low on the priority list for busy librarians. Only 24% of libraries weed continuously, and 39% weed at regular intervals (Dilevko & Gottlieb, 2003).

Multiple factors contribute to librarians' reluctance to weed their collections. Weeding provides no immediate observable benefit. As noted by Dilevko and Gottlieb (2003), weeding can impose a psychological strain on those tasked to implement it, and many librarians find it stressful to make the decision to discard an item. One of the largest obstacles is that weeding is extremely time-consuming. Making a decision about a single title can take several minutes (Zuber, 2012), and the amount of reviewing that can be done is limited by the number of people who can devote time to the task. Large-scale weeding projects can require reviewing tens of thousands of titles, and the weeding project can take years to complete. For example, Concordia University reviewed 25,000 books per year for two years, weeded a total of 12,172 items before deciding that this level of review "could not be maintained" and consequently reduced 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 from a collection of 286,000 items over two years (Snyder, 2014). Wesleyan University weeded 46,000 of approximately 90,000 candidates over three years, from 2011 to 2014. They began by identifying 90,000 weeding candidates that were reviewed individually by the librarians and then reviewed again by interested faculty members (Tully, 2011). 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 (P. Tully, personal communication, October 16, 2014).

One possible way to remove the time obstacle and reduce librarians' psychological stress is automation. Existing methods (e.g., Lugg, 2012; McHale, Egger-Sider, Fluk, & Ovadia, 2017) enable librarians to specify

* Corresponding author. Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, United States.
E-mail addresses: wkiri@wkiri.com (K.L. Wagstaff), geoffrey.liu@sjsu.edu (G.Z. Liu).

https://doi.org/10.1016/j.acalib.2018.02.001
Received 30 October 2017; Received in revised form 31 January 2018; Accepted 6 February 2018
Available online 14 February 2018
0099-1333/

a set of weeding criteria and apply them to a library's circulation records to generate the initial list of weeding candidates automatically. Each candidate is manually reviewed and marked 'Keep' or 'Weed'. 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. An ideal candidate list would be one that contains only items that the librarian would agree to weed. There is potential for significant time savings if an automated method could be employed to filter and refine the list of weedable candidates.

This paper reports on an experimental study that was designed to assess the potential of improving weeding efficiency by using a data mining approach. Specifically, using existing records from the Wesleyan University Library's weeding project, a set of automated classifiers based on different machine learning algorithms were trained to predict the librarians' weeding decisions. The study found statistically significant agreement between the automated classifiers' predictions and the librarians' weeding decisions.

Prior work on weeding library collections

There are two primary approaches to weeding a collection: inclusive and exclusive. The inclusive approach considers each item in turn to decide whether it should be weeded or kept (Soma & Sjoberg, 2010; Tully, 2011), as exemplified by the widely employed Continuous Review, Evaluation, and Weeding (CREW) method (Larson, 2012). In contrast, the exclusive approach first identifies the "core collection" for the library and then weeds items that fall outside of this subset (Trueswell, 1966). In either approach, the weeding process ultimately comes down to deciding if a given item ought to be removed or not based on criteria typically formulated in terms of some conditional factors and instilled in the library's collection development/weeding policy.

Factors in weeding decisions

According to Dilevko and Gottlieb (2003), 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 A. The following sections examine each of the key factors in weeding decisions identified from the literature.

Circulation records

Many weeding efforts are motivated by the empirical observation that a large fraction of the library collection never circulates (Kent et al., 1979; Silverstein & Shieber, 1996; 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. 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).

Slote (1997) surveyed the literature on weeding and found that past use of items 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 determining objectively 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). However, Goldstein (1981) studied eleven libraries and found that none of them took shelf-time into consideration when making weeding decisions, although they did employ use statistics (e.g., number of checkouts) as a weeding criterion. Others have argued that demand (number of checkouts per year) may be more informative than shelf-time (Snyder, 2014).

Circulation records alone may be insufficient for making informed weeding decisions. Some studies found that in-house use mirrors that of circulation, while others found that they could be quite different. Selth, Koller, and Briscoe (1992) found that 11% of the books in their library had in-house use but zero circulation. Weeding based only on circulation records could potentially remove these items despite their evident popularity and utility for visiting patrons (Slote, 1997). The process of weeding requires the examination of these and other additional factors, which increases the time required to evaluate weeding candidates accurately.

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 negative factors that relate to the quality of an item's content and summarizes them with the acronym "MUSTIE" (Larson, 2012, p. 57). These negative factors are: 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).

Additional factors

Several other factors 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 or available in digital form (Metz & Gray, 2005). 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 participating in 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.

Faculty input

Weeding is not always viewed favorably by library patrons, and involving them in the process is helpful. 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 or institutional prestige (Dubicki, 2008). In a psycholinguistic analysis of relevant literature, Agee (2017) identified several negative emotions expressed in faculty responses to weeding projects, which include anger, sadness, and anxiety, in decreasing order of occurrence. Public library patrons may disapprove of discarding items purchased with tax dollars. To overcome opposition to weeding, librarians often devote time to educating and involving patrons. For example, Wesleyan University librarians attended several faculty meetings and set up a website for interested faculty to 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 and remove the flag of any book they wanted to keep (Snyder, 2014). Librarians at Virginia Tech worked to head off criticism by publicizing weeding criteria in advance and inviting interested faculty members to review weeding decisions until they were comfortable with the results (Metz & Gray, 2005).

Automation attempts

To reduce the time required for weeding, systems have been developed to aid in compiling the initial list of weeding candidates (Lugg, 2012). These systems require that the librarian specify a set of weeding rules that define which items shall be considered to be weeding candidates. A commercial service such as Sustainable Collection Services (SCS) (Lugg, 2012) applies these rules to the collection and iterates, sometimes many times, with the librarians until the list of candidates appears satisfactory in terms of its summary statistics (e.g., number of candidates identified). McHale et al. (2017) developed an interactive spreadsheet that linked candidates to external resources such as WorldCat, Amazon.com, and Wikipedia to streamline the manual review process.

The initial list of candidates generated with these approaches may still be overwhelmingly long, containing tens or hundreds of thousands of items that require review. A possible approach to help prune or prioritize the list is to employ machine learning; that is, to construct a computational model of human weeding decisions that can be applied automatically to new items. Silverstein and Shieber (1996) experimented with a similar idea by employing 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 was the number of months since the item's last checkout, akin to Slote's (1997) shelf-time. When only a few items were chosen for off-site storage, incorporating knowledge about the LC classification 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 82%, compared to storage selection 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. However, to the best of our knowledge, a machine learning approach to weeding has not yet been investigated, and it is not known which machine learning model provides the best performance or which variables are most relevant. Silverstein and Shieber's (1996) result suggests that methods employing multiple variables are likely to yield the best performance.

Machine learning classifier methods

This study examined five major types of machine learning classifiers that can be employed to predict weeding decisions. Different types of classifiers employ different model representations, and they possess different strengths and weaknesses, as explained below.

Nearest-neighbor classifier

The simplest approach to classifying an item is to identify the most similar examples previously classified and use them to predict the class of a new item. A nearest-neighbor classifier accumulates a database of classified examples and classifies a new item by selecting the most common label in the k most similar examples (Cover & Hart, 1967). 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 label the new item. 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 most similar examples. For a large data set, this classifier can be very slow.

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). The "naïve" assumption that the variables are not correlated does not always hold, 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 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.

Decision tree

A decision tree creates a series of hierarchically organized tests such that a new item can be classified by applying the tests in order (Quinlan, 1986). This is the model that was employed by Silverstein and Shieber (1996) to predict future demand for books. Parameters to specify include the criterion used to select the best test at each node, the maximum tree depth, and the maximum number of variables to consider for each split. The strengths of the decision tree approach 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 because they are often calculated from the fraction of examples that reach a given leaf node, which may be a very small sample.

Random forest

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 decision tree is trained on a data set that was sampled randomly, with replacement, from the full data set. The collective decisions made by a random forest are more reliable than those made by a single decision tree (the *ensemble effect*). 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 reliable posterior confidence value. Its main weakness is that because it is composed of many individual models, its interpretability is diminished compared to a single decision tree.

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). The support vectors for the weeding problem will consist of those items from the data set that are most difficult to decide whether to weed or keep. SVMs require that all inputs be numeric. Parameters to be specified include the type of kernel (item similarity) function K (linear or Gaussian) and a regularization parameter C to help avoid over-fitting to the training data. 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. SVMs are generally treated as black boxes that generate predictions without any explanation or justification.

Definition of research problem

This research was motivated by the desire to identify automated methods to improve the efficiency of weeding. This section first describes the role of machine learning in weeding projects and then defines the research questions that drove the study.

Machine learning to assist weeding decisions

Weeding decisions are complex and involve both objective and subjective factors (Slote, 1997). While shelf-time can be a strong predictor of 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. In addition, 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.

Making weeding decisions can be viewed as a binary classification problem ('Weed' or 'Keep'). Machine learning provides the ability to train a custom model for any given library 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 libraries' weeding criteria. Machine learning models also naturally accommodate the incorporation of multiple variables into the trained model. This study sought to evaluate whether such classifiers could produce sufficiently reliable and accurate predictions of weeding decisions, by addressing these research questions:

R1. Can machine learning classifier methods predict librarians' weeding decisions with sufficient accuracy?

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 (1981). 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. A prioritized list enables the librarian to review first the items most likely to be weeded.

R2. Which factors are most relevant for making the best predictions of librarians' 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 ones reported by librarians (Dilevko & Gottlieb, 2003). Information about an item's physical condition and the quality of its content is 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. Identification of the most relevant factors is necessary to ensure that future weeding projects collect the right information for each candidate.

Research hypothesis

The major hypothesis that was tested in this experimental study is stated in both null and alternative forms 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.

Research design and methods

This study employed a somewhat unusual experimental design. Conceptually, the experiment can be visualized as a single-group multi-treatment design with repeated measures. The collection of weeding candidates served as a sample of experimental subjects, the weeding decisions made by librarians or classifiers as treatments, and the binary status of weeding candidates ('Weed' or 'Keep') as the dependent variable. In fact, the treatment of predicting with machine learning classifiers consists of multiple levels itself, each corresponding to a different type of machine learning classifier: nearest-neighbor, naïve Bayes, decision tree, random forest, or support vector machine (with linear and Gaussian kernels). Instead of testing for significant differences between treatments in a dependent variable, the statistical analyses tested for significant agreement between the librarians' and classifier-generated weeding decisions.

Data set of weeding decisions

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 Wesleyan data set, provided by the Wesleyan Library for the study in March 2015, 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): (1) publication date before 1990, (2) acquisition date before 2003, (3) no checkouts since 2003, (4) ≤ 2 checkouts since 1996, (5) held by > 30 other U.S. libraries, and (6) 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, an item had to satisfy all of the specified criteria.

Data set variables

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

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

hathicopy and *hathipub* were converted to a representation in which True = 1 and False = 0.

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, the shelf-time variable (ideally to be derived from the last circulation date) was excluded 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 categories of weeding criteria 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 were not recorded in the data set, they were not available for use by the machine classifiers. Data on in-house use of the items were also unavailable.

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

Data set pre-processing

An initial assessment of data quality and the distribution analysis of values for each variable identified some inconsistencies and errors in the data set. The following steps were taken to correct and reduce the data set: (1) One item (*Germany's Stepchildren*, by Solomon Liptzin) had an invalid publication year of "5704", and 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, but the difference in criteria 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 as well. After these adjustments, the data set contained 80,346 items.

Data set characteristics

The data set contained 48,445 (60.3%) items that were marked 'Keep' and 31,901 (39.7%) that were marked 'Weed.' The minimum, mean, and maximum values for the three variables with more than two distinct values are listed in Table 2.

Fig. 1 shows the distribution of values observed for these three variables. Separate distributions are plotted for items marked 'Keep' and 'Weed'. Fig. 1(a) shows that the distribution of ages for items marked 'Keep' is shifted slightly lower (younger/newer) than for items marked 'Weed.' Fig. 1(b) shows that the distribution of values for the number of U.S. libraries holding the item is shifted slightly higher (more holdings) for items marked 'Weed.' Fig. 1(c) shows that items with any faculty votes at all are much more likely to be kept than withdrawn.

Table 2
General distribution of data set by three major variables.

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

Table 3 summarizes the distribution of 'Keep' and 'Weed' decisions for variables that took on only two possible values. The *checkouts* variable was dominated by items that had never been checked out, and the probability of an item with 0 checkouts being withdrawn was much higher. 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 libraries and those held by three peer libraries, with the latter less likely to be withdrawn. As shown in the variables *hathicopy* and *hathipub*, items in the Hathi Trust were a bit more likely to be held in copyright (57%) than to be in the public domain (13%), and those held in copyright were more likely to be withdrawn. Finally, 5.5% of the items were marked to keep by librarians' votes (with value 1 for the *librariankeep* variable), of which only 31 items (0.7%) were withdrawn despite the librarian 'Keep' vote; these were either lost or duplicate items. This suggests a very strong correlation between librarian votes and final decisions. Likewise, the items that received at least one faculty vote of 'Keep' ($>$ 0 for the *facultykeep* variable) were very likely to be kept, and only 1.7% of these items were eventually withdrawn.

Classifier training and evaluation

All machine learning classifiers need to be trained to develop an internal model. During the training process, each classifier analyzes a set of training data in which each item has been labeled with a human classification decision (the dependent variable). Through training, the classifier's internal model captures relationships between the independent (feature) variables and the dependent (prediction) variable.

The data set was divided randomly into two equal halves: D_t for training and D_e for evaluation. Each machine learning classifier was trained on D_t with known labels and then used to generate blind predictions for D_e. All of the variables were normalized to achieve a mean value of 0 and a standard deviation of 1. The shifting/scaling coefficients were determined from D_t and then applied to D_e.

The parameters that were used to train each machine learning classifier are summarized in Table 4. 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. The parameter values that resulted in the highest held-out performance (accuracy) were selected. For some classifiers (linear and Gaussian SVMs), all specified parameter values were tested; these classifiers are marked "All" in Table 4. For others (marked "50R" or "100R"), only 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 selected parameter values.

Performance measures and significance testing

To compare the predictions generated by a classifier against the decisions made by a librarian statistically for hypothesis testing, two statistical measures of agreement, Yule's Q and the φ coefficient, were employed. There is no single best measure of agreement that is widely agreed upon for all possible types of data. However, Yule's Q and φ are two widely used measures that factor in the amount of agreement that would be expected by random chance, without an assumption of Gaussianity.

Yule's Q (Yule, 1900) is based on the odds ratio of two outcomes (agreement and disagreement) to enable the determination of whether

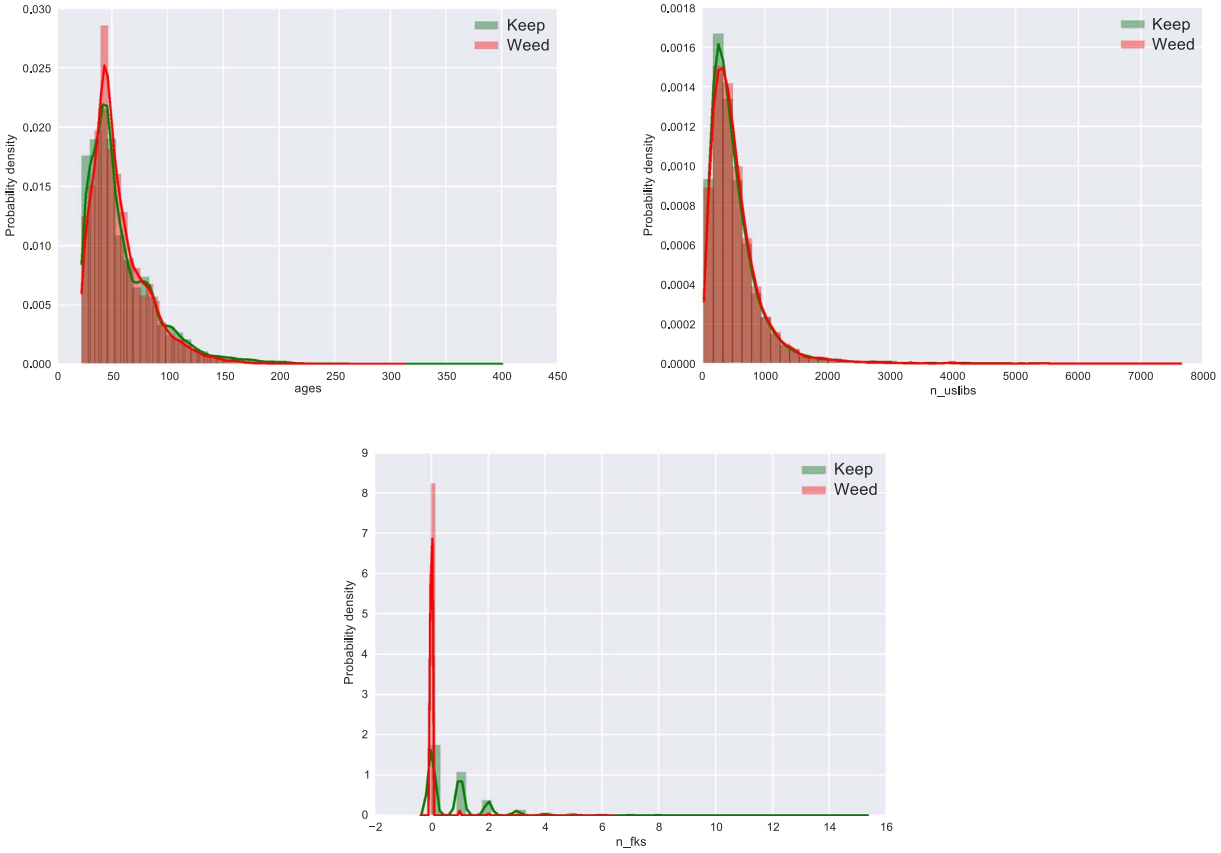

Fig. 1. Distribution of items marked 'Keep' vs. 'Weed' for three variables: (a) *age*, (b) *uslibs*, (c) *facultykeep*.

Table 3
Variable-specific characteristics of data set.

Variable	Value	Frequency	Percentage	Weed	Keep	*P*(weed)
checkouts	0	61,993	77.16	25,252	36,741	0.41
	1	18,353	22.84	6649	11,704	0.36
peerlibs	2	47,738	59.41	19,600	28,138	0.41
	3	32,608	40.58	12,301	20,307	0.38
hathicopy	False	34,660	43.14	12,531	22,129	0.36
	True	45,686	56.86	19,370	26,316	0.42
hathipub	False	69,997	87.09	27,932	42,045	0.40
	True	10,369	12.91	3969	6400	0.38
librariankeep	0	75,926	94.50	31,870	44,056	0.42
	1	4420	5.50	31	4389	0.01
facultykeep	0	56,451	70.26	31,503	24,948	0.56
	> 0	23,895	29.74	398	23,497	0.02

Table 4
Classifier parameters and candidate values evaluated for optimization.

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

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'. Let the value a be the number of times that the librarian and the classifier both voted to weed a particular item, and d be the number of times they both voted to keep an item; these are the agreements. Let b be the number of items that the librarian voted to keep and the classifier voted to weed, and let c be the number of items that the librarian voted to weed and the classifier voted to keep; these are the disagreements. Yule's Q is defined as $Q = \frac{ad - bc}{ad + bc}$; the φ coefficient is calculated as $\varphi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$. Both values can be assessed for statistical significance by a χ^2 test with one degree of freedom, which was done at the significance level of $p = 0.001$.

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' likely are worse mistakes than incorrect predictions of 'Keep', but Yule's Q and the φ coefficient treat both equally. To gain further insight into these types of errors, each method was also assessed in terms of recall (R), precision (P), and accuracy (A), which are defined respectively as: $R = \frac{a}{a+c}$; $P = \frac{a}{a+b}$; and $A = \frac{a+d}{a+b+c+d}$. The statistical significance of the recall, precision, and accuracy scores was assessed with a univariate χ^2 analysis by testing the observed values against those expected of a random process. The expected values of recall, precision, and accuracy are 0.5, 0.4, and 0.5 respectively (see Appendix B for derivation and further details about the χ^2 testing).

Table 5
Parameter values selected for machine learning classifiers using cross-validation.

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

Implementation

The experiment was implemented in Python, using a combination of new code and the freely available scikit-learn library for implementing the machine learning classifiers.

Parameter setting

After performing three-fold cross-validation on the training set D_t, the parameter values were selected for each classifier as shown in Table 5. 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. 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 variables (of the eight available), and the maximum tree depth was also five (with no pruning). In contrast, the random forest was composed of 100 trees, each of which was only allowed to use three variables 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 classifier. In contrast, the Gaussian SVM selected a C value of 10, which means that its more complex modeling yielded a better fit to the data. The Gaussian parameter (γ) was set to 10. This parameter is an intuitive measure of how far the influence of a given example reaches in the 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.

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 variable, estimated from the training data (given in Table 3). The support vector machine models do not lend themselves well to interpretation, and they are generally treated as black boxes that generate good predictions but do not provide insights into the data being classified. Consequently, it is only necessary to describe the learned decision trees here.

The top three layers of the trained decision tree are shown in Fig. 2. The most likely outcome after three layers of testing (W = 'Weed' and K = 'Keep'), with its associated probability, is shown in the bottom row of the diagram; the complete (five-layer) classifier conducts two more layers of tests before classifying a given item. The first variable 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 variable 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 variable assessment. The *librariankeep* and *facultykeep* variables had the strongest discriminatory power between the two classes, while the other variables provided less separation.

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

Experimental results

Table 6 presents the results of performance testing for two baseline approaches (keep all items, weed all items) and the six machine learning classifiers. The best values are marked in boldface in each column. If a column has multiple values marked in boldface, these marked values are not significantly different from each other, as determined by z-score tests with $p = 0.01$.

For all the machine learning classifiers, the accuracy values were found to be statistically significantly better than a random process ($\chi^2 \geq 7827.0$, $p = 0.001$). They all achieved approximately the same level of accuracy (~72%), well above the best baseline performance of 60.0%. However, a z-score test with $p = 0.05$ found that there is no significant difference in accuracy between classifiers. Their recall values were all significantly better than random ($\chi^2 \geq 21{,}889.3$, $p = 0.001$), and so were their precision scores ($\chi^2 \geq 6286.1$, $p = 0.001$). Finally, their φ and Yule's Q values of agreement were all statistically significant (χ^2 as listed in Table 6, $p = 0.001$) as well, despite the difference in their range of values. Both of the baseline approaches did significantly worse than a random process.

Although all classifiers reached about the same level of accuracy, they did not all make the same type of errors. Recall varied noticeably across classifiers, while precision 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 60%. That is, there are many items in the data set that should be kept but are difficult to distinguish, given the variables 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 predict 'Keep' mistakenly for an item that was actually withdrawn, but its 'Weed' predictions were most reliable. This classifier had significantly higher precision than the naive Bayes ($z = 3.18$, $p = 0.01$) and linear SVM classifiers ($z = 3.18$, $p = 0.01$), and to a

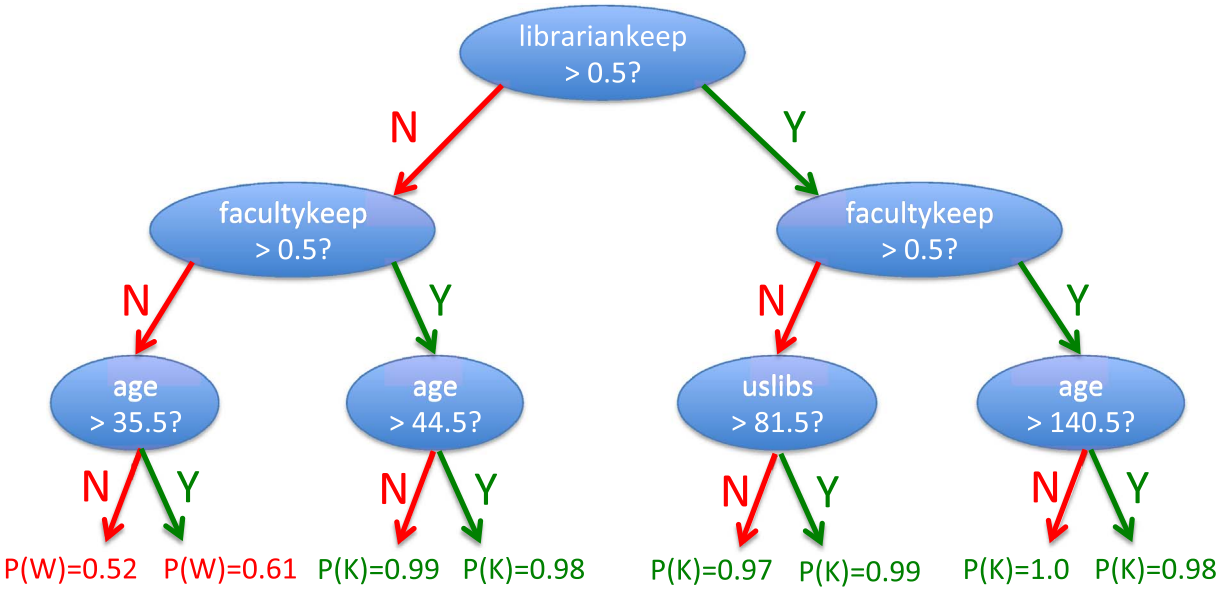

Fig. 2. Top three layers of the learned decision tree model.

Table 6
Performance statistics of predicting weeding decisions by different classifiers. The best result for each column is marked in bold. Multiple values are in bold if they are not statistically significantly distinguishable.

Method	Accuracy	Recall	Precision	φ		Yule's Q	
				value	χ^2	value	χ^2
Baseline (keep all)	0.600	0.000	0.000	N/A	N/A	N/A	N/A
Baseline (weed all)	0.400	1.000	0.400	N/A	N/A	N/A	N/A
K-nearest-neighbor	**0.721**	0.869	**0.605**	0.486	9965.8	0.832	217.0
Naïve Bayes	**0.724**	0.980	0.594	**0.552**	12,142.1	0.968	538.3
Decision tree	**0.725**	0.967	0.596	0.545	10,921.3	0.949	278.3
Random forest	**0.724**	0.919	**0.601**	0.516	12,066.1	0.887	446.5
SVM (linear)	**0.725**	**0.986**	0.594	**0.557**	12,329.4	**0.978**	645.2
SVM (Gaussian)	**0.725**	0.954	0.598	0.537	11,653.0	0.930	361.7

lesser degree than the decision tree ($z = 2.45$, $p = 0.05$) and Gaussian SVM ($z = 2.00$, $p = 0.05$). In contrast, the linear SVM had the highest recall (98.6%) but the lowest precision (59.4%), and the differences are statistically significant in both cases (for recall, $z \geq 6.69$, $p = 0.01$; for precision, all but naïve Bayes and the decision tree, $z \geq 2.20$, $p = 0.05$). The linear SVM therefore was more likely to predict 'Weed' mistakenly for an item that was actually kept.

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 6 cause us to reject the null hypothesis. Although the two measures are not directly comparable in terms of their values, they ranked the methods identically. The naive Bayes and linear SVM classifiers had the highest φ values (0.552 and 0.557), and the difference was not statistically significant 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, it was found that both measures correlated well with recall ($r = 0.918$, 0.999 for φ and Yule's Q respectively) but not with precision ($r = 0.196$ for φ, and 0.003 for Yule's Q). Thus, classifiers with high recall values tended to have higher agreement values as well, regardless of their precision.

Findings and discussion

The experimental results indicate that the alternative hypothesis should be accepted, i.e., there is statistically significant agreement between human decisions and automated classifier predictions. 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 librarians in making weeding decisions in a given project. For the purpose of refining the initial list of weeding candidates, precision is more important than recall, since mistakenly discarding an item has more impact than mistakenly keeping it. It is important that the prediction of weeding any item 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 helps select the most appropriate model. This consideration favors 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.

Classifier precision may be improved by including more variables (e.g., shelf-time and physical condition, in-house use) in the data set. On the other hand, analysis of the learned models revealed that librarian and faculty votes for retention emerged as the most relevant variables, while item age and presence in other U.S. libraries were also important.

One must interpret the agreement measurements carefully. To our knowledge, quantitative assessment of 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, it is unknown what to consider as the best achievable agreement in reality; probably not $\varphi = 1.0$. Assessing agreement between librarians would help interpret the classifier performance in context.

Conclusion

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 the 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, sometimes as a continual ongoing process (Larson, 2012), methods that can further reduce the amount of human effort required are vital.

This study produced the first empirical evidence of agreement between human weeding decisions and predictions by machine learning classifiers. The learned models will not replace human processing, but they can instead provide an initial assessment of the list of candidates, which allows librarians to focus their time and attention on those items most likely to be weeded. 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 weeding decisions made at one library could be directly applied to the collection at another library. Even within a given library, McAllister and Scherlen (2017) suggest that different models may be needed for individual disciplines or collections due to different patterns in book use. It is recommended that each library label a relevant portion of their weeding candidates, or compile past weeding decisions, to provide a custom 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 resulting list of candidates by specifying a minimum confidence and generate a new list of only those with a weeding prediction confidence greater than this threshold. For greater flexibility, the entire candidate list may be sorted 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.

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 needed to train a model sufficiently so that it can attain a certain level of accuracy or agreement. In this study, half of the data set was used to train each model, which amounted to > 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. Another promising area of investigation is the use of active learning (Cohn, Ghahramani, & Jordan, 1996). With active learning, the machine learning method starts with a few labeled examples and then actively suggests which items the librarian should label first to provide the most informative labeled examples. This strategy has been shown to reduce dramatically the number of labeled items required. Second, one could assess whether deep neural networks could provide a more accurate model than the six classifiers examined in this study. Third, this study included only information about age, circulation, and other library holdings. As discussed earlier, there are several other potentially useful variables that could not be evaluated in the present study. These include information about the item's shelf-time, physical condition, quality of content, library in-house use, etc. A similar empirical study with a data set including those variables would yield more definitive findings and determine if improvement in classifier performance (if any) warrants the cost of compiling additional data.

Ultimately, the 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.

Acknowledgments

We wish to thank Diane Klare, Lori Stethers, and Patricia Tully for providing access to the Wesleyan University data set and generously sharing their time and expertise. We also thank the anonymous reviewers for their feedback and recommendations.

Appendix A. 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 > 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
- > 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 created a list of ~90,000 candidates using these criteria (Tully, 2011):

- Fewer than two checkouts since 1996
- Published before 1990
- Acquired before 2003

- > 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.

Appendix B. Expected values for performance measures and χ^2 testing

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 (Table 3), 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:

$$E[R] = \frac{E[a]}{E[a+c]} = \frac{N \bullet P_{rand}(W) \bullet P_{label}(W)}{N \bullet P_{label}(W)} = \frac{N \bullet 0.5 \bullet 0.4}{N \bullet 0.4} = 0.5$$

Likewise, the expected value of precision is:

$$E[P] = \frac{E[a]}{E[a+b]} = \frac{N \bullet P_{rand}(W) \bullet P_{label}(W)}{N \bullet P_{rand}(W)} = \frac{N \bullet 0.5 \bullet 0.4}{N \bullet 0.5} = 0.4$$

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, $O(o_i)$:

$$X^2 = N \bullet \left(P(o_1) \left(\frac{O(o_1) - P(o_1)}{P(o_1)} \right)^2 + P(o_2) \left(\frac{O(o_2) - P(o_2)}{P(o_2)} \right)^2 \right)$$

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 $\overline{V}$ be the average of $V_1$and V_2.

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

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

References

Agee, A. (2017). Faculty response to deselection in academic libraries: A psycholinguistic analysis. *Collection Management, 42*(2), 59–75.

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, and 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., ... 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.

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

McAllister, A. D., & Scherlen, A. (2017). Weeding with wisdom: Tuning deselection of print monographs in book-reliant disciplines. *Collection Management, 42*(2), 76–91.

McHale, C., Egger-Sider, F., Fluk, L., & Ovadia, S. (2017). Weeding without walking: A mediated approach to list-based deselection. *Collection Management, 42*(2), 91–108.

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 dissertationUniversity 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* (4th 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.

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. *Proceedings of the American Society for Engineering Education Annual Conference (pp. 6139–6152). Austin, TX.*