Identifying Relevant Data for a Biological Database: Handcrafted Rules Versus Machine Learning

Aditya Kumar Sehgal, Sanmay Das, Keith Noto, Milton H. Saier, Jr. and Charles Elkan

Abstract— **With well over one thousand specialized biological databases in use today, the task of automatically identifying novel, relevant data for such databases is increasingly important. In this paper, we describe practical machine learning approaches for identifying MEDLINE documents and Swiss-Prot/TrEMBL protein records, for incorporation into a specialized biological database of transport proteins named TCDB. We show that both learning approaches outperform rules created by hand by a human expert. As one of the first case studies involving two different approaches to updating a deployed database, both the methods compared and the results will be of interest to curators of many specialized databases.**

Index Terms— **Bioinformatics (genome or protein) databases; Clustering, classification, and association rules; text mining; biomedical text classification; data mining.**

I. INTRODUCTION

Rapid advances in sequencing technology have resulted in the availability of large amounts of protein, RNA, and DNA data. The need to organize this data has prompted the development of large-scale, general-purpose databases such as Swiss-Prot, Protein Data Bank (PDB) and NCBI GenBank. At the same time, many specialized databases have been developed that contain information about biomolecules that are related functionally, structurally, and/or phylogenetically. The *Nucleic Acids Research* journal lists 1,170 specialized databases as of the January 2009 update [13]. Examples of such databases include the Secreted Protein Database [3], the Nuclear Protein Database [7], and FlyBase [15]. Keeping these databases up-to-date is a significant task. Currently, they are primarily curated manually, so a significant amount of human time and labor goes into adding new information to them.

New information typically comes either from the primary biological literature, or from existing broad biomolecular databases. Browsing the literature or broad biological databases manually for information has become infeasible as the volume of data has grown. For example, MEDLINE, a database of approximately 17 million published articles related to the life sciences, adds approximately 50,000 new articles each month, while Swiss-Prot, a general protein database (http://ca.expasy.org/sprot), added over 100,000 new proteins in the first half of 2008. Therefore, developing automatic methods to identify new relevant information has received considerable attention in recent years [10], [22].

Most research on automated curation has been in the context of evaluation challenges. For instance, the first task of the 2002 KDD Cup data mining contest was to automatically identify articles that should be added to FlyBase [36]. Typically in these challenges, a document dataset is released and researchers have the opportunity to work on it and submit their relevance judgments, which are compared with expert judgments to rate the researchers' systems. While the development of these challenges has been a big step towards rigorous and systematic evaluation of biomedical information retrieval systems, such systems are still not widely deployed in practice. This is because practical deployment involves an additional number of steps that are as important as the development of the actual system that learns to judge relevance. These include (i) the development of the training corpus (including human decisions on the true labels for items), (ii) appropriately assessing the performance of the final classifier when labels are not available, and (iii) optimizing the workload of experts so they do not end up doing more work than it takes to do manual curation. These parts of the pipeline are taken care of by the organizers in evaluation challenges. In real applications, these considerations often dominate, and curators may choose to use hand-coded rules rather than more flexible machine learning or information retrieval systems. Such questions explain the recent call for an increased focus on user evaluation from the biomedical informatics community [16].

This paper addresses these problems in the context of a specialized database, the Transport Classification Database (TCDB) [29]. Our eventual goal is to automate the entire process of adding new information to TCDB. The first task is to identify appropriate sources of new information, keeping in mind the issues described above. In order to understand whether machine learning approaches can bring real benefits to the curators of TCDB, we develop and evaluate machine learning methods without asking too much of human experts, and compare these with expert-designed rules. We find that the machine learning approaches have substantial benefits even without imposing an unreasonable workload on experts. In particular, they perform significantly better in terms of both precision and recall. To the best of our knowledge, this paper is the first to compare rule-based and machine learning approaches in the context of a real application. This evaluation will be important for practitioners choosing an approach on which to focus their efforts.

II. BACKGROUND

TCDB is a database that provides free access through the web (http://www.tcdb.org) to comprehensive information on transport proteins. A transport protein is a protein that imports or exports molecules through the membrane of a cell or organelle. Currently, TCDB contains information about over 6,000 distinct transport proteins organized into more than 600 families, and compiled

from over 5,000 published papers. Data are added to TCDB continually as new functional information is published and new transport systems are identified. A human expert reads through published literature and identifies papers that describe transport-related proteins of interest.[1]

If a newly identified protein is not homologous to a protein that is already included in TCDB, or if it has a novel function, then information relating to the protein's amino acid sequence, the organism in which it is found, its function, and relationship to other proteins is imported into TCDB.

The first step towards automating the updating of TCDB is to automate the selection of potential sources of new data. In addition to the primary literature, we also consider existing structured general protein databases, in particular Swiss-Prot/TrEMBL. We compare the accuracy of hand-crafted expert rules with that of learned classifiers in distinguishing papers or database records that are relevant to transmembrane transport from the rest.

While some tasks are best evaluated using a task-specific utility measure (for instance, the document classification task in the 2005 TREC Genomics Track [17]), we use standard precision and recall to evaluate our techniques for the sake of generality. Precision and recall can be difficult to measure in the absence of a large set of hand-labeled data. We describe our method for doing so in detail in Section IV-B.

III. METHODS

A. Rule-Based Approaches

The rule-based approaches that we evaluate are classification schemes designed by a domain expert. In this approach, records from other protein databases are treated as sources of potentially relevant information. We use the Swiss-Prot and TrEMBL databases that are maintained by the Swiss Institute for Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). Swiss-Prot is a manually annotated database containing 230,150 protein records.[2] A record consists of an official name for a protein, its sequence, synonyms, gene name, taxonomy, published references, references to other protein databases, as well as other information. Some records also contain information extracted by hand from associated publications. This information can include functional descriptions and subcellular locations.

Importantly, records in Swiss-Prot contain keywords assigned by human experts from a controlled vocabulary. The vocabulary of keywords contained 865 entries in 2006, and still does in 2009; the current list of keywords is available at http://www.expasy.org/cgi-bin/keywlist.pl. The keywords assigned to a protein provide a high level description of it. Example keywords include *transport*, *transmembrane*, and *ion transport*.

Human experts also assign Gene Ontology (GO) [14] terms to records. The Gene Ontology is a separate controlled vocabulary composed of three different ontologies, namely *Biological Process*, *Molecular Function*, and *Cellular Component*. It provides terms to describe gene and gene product attributes, for example GO:*ammonium transport* and GO:*angiogenesis*.

There is a known mapping between Swiss-Prot keywords and GO terms [2]. However, this mapping is many-to-many and partial: about 25% of keywords have no mapping to GO terms.

TrEMBL is a supplement to Swiss-Prot that contains records for proteins that are not yet included in Swiss-Prot.[3] These records contain information similar to Swiss-Prot records. However, keywords and GO terms are automatically assigned to TrEMBL records and are hence less reliable on average than those in Swiss-Prot records [23]. Additionally, Swiss-Prot is cross-referenced with over 50 other databases, providing quick access to other types of relevant information.

The goal is to distinguish automatically between relevant (in our case, to TCDB) and irrelevant Swiss-Prot and TrEMBL records. A record is relevant if and only if it describes a transmembrane transporter protein. We evaluate two handcrafted rule-based methods for this task. Both methods use the metadata information (keywords and GO terms) previously described. Of the 230,150 records in the August 22, 2006 Swiss-Prot release, 198,613 have been assigned keywords and 28,958 have been assigned GO terms. Only 2,579 records (under 1.2%) have neither keywords nor GO terms. In the August 22, 2006 version of TrEMBL, 948,399 records have keywords assigned and 1,795,208 have GO terms assigned; 338,328 (about 11%) have neither. Most records in Swiss-Prot have keywords but this is not the case for TrEMBL, which is to be expected as only the former is manually annotated. Interestingly, a higher percentage of TrEMBL records have GO terms.

TCDB contains 3,226 unique proteins as of August 4, 2006, with a unique accession number for each one that points to a record in an external database. 3,205 accession numbers point to Swiss-Prot or TrEMBL records, while 21 point to other databases including NCBI, GenBank, and PDB. We ignore these 21 proteins in this paper; given their small number, this choice has no significant impact on our results. Of the 3,205 linked records in Swiss-Prot or TrEMBL, 6 have been deleted due to incorrect data. Henceforth, when we refer to proteins in TCDB, we mean the 3,199 proteins that have valid Swiss-Prot or TrEMBL records.

Of the 3,199 proteins, 2,088 have Swiss-Prot records and 1,111 have TrEMBL records. 2,823 have keywords assigned and 1,385 have GO terms assigned; 129 (about 4%) have neither. We use two different rule-based methods to identify relevant records. Both involve matching keywords or terms associated with a record to keywords or terms indicative of transport. We consider only exact matches for keywords and terms, since these come from restricted, controlled vocabularies.

The term *transport* (GO:0006810) is the ancestor of all transport-related terms in the *Biological Process* ontology. It is defined as "The directed movement of substances (such as macromolecules, small molecules, ions) into, out of, within or between cells." Similarly, the term *transporter activity* (GO:0005215) is the ancestor of all transport-related terms in the *Molecular Function* ontology. It is defined as "Enables the directed movement of substances (such as macromolecules, small molecules, ions) into, out of, within or between cells". No directly comparable ancestor term exists in the *Cellular Component* ontology. While this ontology contains the term *membrane*, which encompasses proteins associated with the cell membrane, it may be assigned to non transmembrane transporter proteins, such as Guanine nucleotide-

---

[1] The human expert for TCDB is Professor Milton H. Saier Jr., one of the authors of this paper and the architect of the Transporter Classification System.

[2] The May 5, 2009 version of Swiss-Prot contains 466,739 protein records. Our experiments were done with the August 22, 2006 version which contains 230,150 records.

[3] The August 22, 2006 version of TrEMBL contains 3,081,935 protein records.

binding protein G(t) subunit alpha-1, which is a transmembrane protein that functions as a signal transducer rather than a transporter.

Rule 1: This rule says that a record is relevant if and only if any GO term in the record is *transport* or *transporter activity*, or either of these terms is an ancestor in the GO hierarchy of any GO term in the record. ∎

For the second rule, the human expert has classified the keywords assigned to a Swiss-Prot or TrEMBL record into three sets listed in Table I. Keywords in set 1 are the best indicators of transmembrane transporter activity, so they are deemed sufficient to predict relevance. Keywords in sets 2 and 3 also indicate relevance but are not sufficient by themselves. To predict relevance, they must be augmented by additional keywords, *transport* in the case of set 2 and *transmembrane* in the case of set 3. Note that the term *transport* by itself is too broad to accurately identify proteins involved in transmembrane transport. Such a query will result in many false positives, such as 'PDCD6-interacting protein' and 'Autophagy-related protein 12' which are bulk transport proteins and not transmembrane transporters.

Rule 2: This rule says that a record is relevant if and only if at least one of the following conditions is true:

- The record has any keyword in set 1.
- The record has the keyword *transport* and any keyword in set 2.
- The record has the keyword *transmembrane* and any keyword in set 3.
- The record has any keyword corresponding to a GO term that is or has an ancestor in the GO ontology that is the term *transport* or *transporter activity*. ∎

Note that Rule 1 only looks at the GO terms in a record, while Rule 2 only looks at the keywords in a record. Therefore, neither rule is a special case of the other one. Rule 2 uses the correspondence between keywords and GO terms [2], but Rule 1 does not.

Rule-based methods for annotating proteins based on information in Swiss-Prot records have been described in the prior literature (*e.g.* Kretschmann *et al.* [23]). That research is similar in some ways to ours, but there are major differences. The goal of Kretschmann *et al.* is to assign keywords to proteins in the TrEMBL database using rules learned from particular information in Swiss-Prot records. They utilize machine learning algorithms to learn rules from the basic data (taxonomic and sequence information) present in Swiss-Prot records. Each rule outputs a single keyword based on the basic data of the protein to be annotated (in TrEMBL). In contrast, our goal is to create rules that utilize a protein record's annotation data to discriminate between transmembrane transport related proteins and other proteins. Unlike Kretschmann *et al.*, our rules have been manually designed by a domain expert. Also, while the overall research goal of Kretschmann *et al.* was to learn rules, our research goal is to compare documents and database records as possible sources of data for transmembrane transport proteins.

B. Learning to Classify MEDLINE Documents

For this approach, we consider published articles in the primary literature as a source of potentially relevant information to be added to TCDB. A document is relevant if it contains information on proteins related to transmembrane transport and irrelevant otherwise.

TABLE I
THREE SETS OF KEYWORDS USED IN RULE 2.

Keyword set 1
Ion transport, sugar transport, amino-acid transport, symport, sodium transport, ionic channel, hydrogen, ion transport, potassium transport, porin, antiport, phosphotransferase system, voltage-gated channel, peptide transport, calcium channel, tonB box, potassium channel, zinc transport, sodium channel, copper transport, neurotransmitter transport, chloride channel, phosphate transport, bacteriocin transport, signal recognition particle, sulfate transport, gap junction, cobalt transport, ammonia transport, polysaccharide transport, nickel transport, sodium/potassium transport, phosphonate transport.
Keyword set 2 (In combination with "transport")
Membrane, transmembrane, antimicrobial, antibiotic, antibiotic resistance, heme, molybdenum, nickel, antibiotic, cytochrome c-type biogenesis, decarboxylase, cadmium resistance, cadmium, hormone, arsenical resistance, plant toxin, neurotoxin, anion exchange, chromate resistance, bacteriochlorophyll, tungsten, mercury, enterotoxin, vitamin A, thyroid hormone, selenium, mercuric resistance, hemagglutinin, folate-binding, bacteriocin immunity.
Keyword set 3 (In combination with "transmembrane")
Protein transport, electron transport, iron transport, calcium transport, lipid transport.

We restrict ourselves to the documents in MEDLINE, which currently contains data on over 17 million documents, and furthermore we consider only the subset of 557,458 that contain the term "protein" in the title or abstract and come from one of 108 journals selected by a human expert.[4] We consider this subset of documents to be our universe. The rest of MEDLINE is excluded from consideration. This greatly limits the scope our search for relevant documents, at a risk of omission considered low by the human expert.

Training a classifier normally requires a set of example documents that are known to be positive and a set of documents that are known to be negative. We use 3,168 MEDLINE documents referenced by TCDB as the positive training set. Unfortunately, no curated set of negative example documents is available; a similar difficulty occurs with numerous other text classification tasks also. Learning from positive and unlabeled examples is an important research question that has received a great deal of attention (*e.g.* [8], [9], [12], [25], [37]). Methods for learning with unlabeled examples include using a biased classifier [24], one-class SVMs [32], [33], and transductive SVMs [21].

We use a random sample of documents from the universe as the negative training set. Even though the random sample must include some positive examples (we estimate about 5%), this procedure is justified. We have shown in previous research that classifiers trained using unlabeled documents as negative examples can rank documents as accurately as classifiers trained with actual negative examples. For a thorough explanation and evaluation of this approach, with comparisons to alternative methods, see [11].

Specifically, we randomly select 6,336 documents (twice the number of positive documents) that are not in TCDB, from the

---

[4]There were 557,458 such documents on July 6, 2006, when these experiments were carried out.

universe of 557,458 documents to serve as the negative training set. Choosing twice the number of positive documents is a design decision that balances the competing needs of getting a comprehensive sample of the population and avoiding problems known to be associated with trying to learn a classifier on highly skewed training data (classifiers can be tuned after training to deal with differently balanced test sets by tweaking parameters, but having too many negatives in the training set can lead to a loss in discriminative ability), as well as the practical aspects of downloading many MEDLINE records.

Distinguishing between relevant and irrelevant documents is a text categorization task, which is a well-known research problem in computer science [34]. In addition to domain-independent methods, specialized methods for classifying biomedical documents have been proposed that incorporate information from outside the documents, such as from the Unified Medical Language System [18] or from the Web [4]. Other methods use special information within documents such as figure captions [28] and image data [35]. Although all these methods could be useful here, we begin by evaluating two standard classification methods that are applied to documents represented as "bags of words," *i.e.* vectors of features based on individual word counts. The vector components we use are words from the document's title and abstract, and terms from the medical subject headings (MeSH) and chemical abstracts service (CAS) vocabularies that have been assigned to the document. We divide each document's title and abstract into individual words and we group multi-word MeSH and CAS terms into phrases. We stem individual words and we filter out common English words. Following [19] and many others, we apply a version of the tfidf weighting procedure described by [31]. Specifically, the feature value f_{ij} for a word i in document j is

$$f_{ij} = \log(1 + \frac{n_{ij}}{D_j}) \times \log(\frac{C}{C_i})$$

where n_{ij} is the number of occurrences of word i in document j, D_j is the number of words in document j, C is the total number of documents, and C_i is the number of documents that contain word i. Other document representations and term weighting techniques yield results similar to those below, so they are not described in this paper.

For the classifier, we use a support vector machine (SVM). Specifically, we use the SVM-Light package of SVMs [20]. We use the SMART IR system [30] to index documents and create document vectors from the weighted index terms to provide as input to SVM-Light.[5]

C. Learning to Classify Database Records

The third approach that we evaluate is training a classifier on Swiss-Prot records. These experiments are published and described in detail in [6], so we only summarize the method here. As mentioned above, each Swiss-Prot record consists of a protein name, titles of related publications, human-annotated comments, curator-assigned keywords, and GO terms, among other things.

We consider each record to be a document, and employ a bag-of-words representation, similar to the one described above for MEDLINE documents. We use a maximum-entropy classifier, implemented as part of the MALLET toolkit [27].[6]

For the experiments classifying Swiss-Prot records, the positive training set consists of 2,453 Swiss-Prot records mentioned in TCDB as of September 23, 2006. The negative training set consists of 4,906 Swiss-Prot records (twice the number of positive examples, following the methodology above) selected randomly from the set of Swiss-Prot records for proteins not present in TCDB. Strictly speaking, these examples are unlabeled, although we know most are negative. However, as described in [6], our methods involve labeling false positives iteratively, so the negative training set is highly accurate and contains very few remaining actual positive examples, if any.

IV. EXPERIMENTS

A. Relevance

Eventually, we wish to develop methods that update TCDB automatically in a way that satisfies the biologists who currently do the task manually. This paper focuses on the first step of the updating process, which is to identify sources describing proteins and the novel information about them that should be included in the database. Ideally, our methods would find all proteins that should be included in TCDB, without finding any proteins that a human expert would deem unfit for inclusion. A protein must at least be involved in transmembrane transport to be included in this database. However, there are many reasons why a transport protein is not included in TCDB, for instance because it is homologous and performs the same function as a protein already in TCDB, it is not sufficiently well characterized functionally, or the information about it is not published in a well-established journal.

For the purposes of evaluation, we adopt a simple definition of relevance: A document or database record is relevant if contains information about any protein related to transmembrane transport. However, we also consider the notion that we call "particularly interesting" proteins in our discussion in Section V. These are transport-related proteins that are judged by the human expert to be the most valuable additions to TCDB.

B. Measuring Precision and Recall

To evaluate the success of the approaches we consider, we begin with the standard notions of precision and recall from the information retrieval literature.

Measuring precision and recall properly when gold standard data is absent or costly to obtain is a non-trivial problem that is starting to receive significant attention in the literature [1]. This question has also been asked in the context of biomedical information retrieval tasks. In particular, [5] considers a combination of manual labeling by experts and bootstrapping "weak labels" from existing datasets for a relational learning problem, but they do not consider the recall problem in the detail we do here.

Our task uses the relevance judgments of a human expert as the gold standard. For each example retrieved by any method, the human expert indicates whether or not it is genuinely about a

---

[5]We also evaluated the naïve Bayes classifier, using the Rainbow toolkit [26]. Initial experiments show that SVMs consistently have a lower false positive rate. Because it has higher precision, the SVM classifier is more likely to satisfy the experts who maintain TCDB, whose immediate need is to reduce their workload by narrowing the number of papers they must read, even at the potential cost of missing some relevant papers. Thus we report results based on the SVM classifier.

[6]MALLET also has an implementation of naïve Bayes but the discriminative MaxENT model proved to have superior precision in initial experiments. See [6] for details.
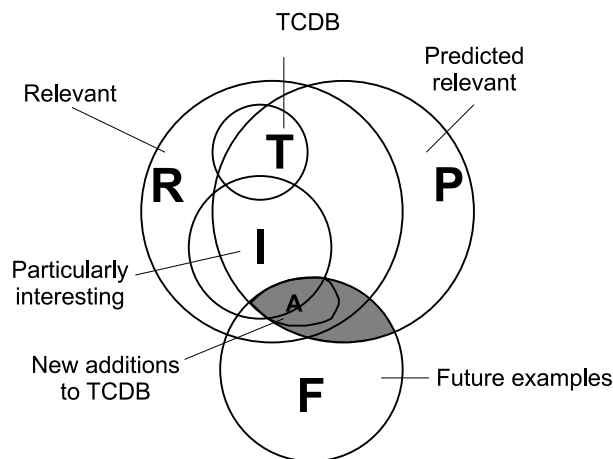
Fig. 1. A Venn diagram showing the classes of examples (*i.e.* MEDLINE documents or Swiss-Prot/TrEMBL protein records) in this project. Examples in R are relevant (in our case, to TCDB). Examples in P are predicted to be relevant by a classifier, *i.e.* $P \cap R$ represents the set of examples discovered by the classifier, $P \setminus R$ represents false positives, and $R \setminus P$ represents false negatives. T is the set of examples that are already in a specialized database (TCDB). I represents the set of examples that are particularly interesting. These are relevant, novel and scientifically important examples that are especially valuable, as discussed in Section V. F represents a set of future examples. In our experiments, these are sampled randomly. The shaded area, $F \cap P$, is the set of examples that the human expert must review and classify. Some of the relevant examples will be added to the database, represented by the set A. A equals $F \cap P \cap I$ plus part of $F \cap P \cap R$. Not all of $F \cap P \cap R$ will necessarily be added to TCDB because we do not add new examples to TCDB if they correspond to proteins that are homologous to ones already in the database.

protein related to transmembrane transport, regardless of whether the protein is homologous to a protein already in TCDB, or might otherwise be unsuitable for inclusion in the database.

Consider the Venn diagram in Figure 1. Here, the set R refers to the set of relevant documents. the set P refers to the set of documents that are predicted to be relevant by a classifier. Precision, $|P \cap R|/|P|$, and recall, $|P \cap R|/|R|$, are difficult to measure because most MEDLINE documents and Swiss-Prot records are not labeled as relevant or irrelevant.

To measure precision and recall, we perform two separate experiments. For measuring precision we obtain a sample of unlabeled examples, F in Figure 1. After the classifier makes its predictions, the human expert labels the examples in this set that are predicted to be positive ($F \cap P$, the shaded area in Figure 1). Precision is then computed as $|F \cap P \cap R|/|F \cap P|$.

Unfortunately, we do not have a similar way to measure recall without labeling the entire set F, which would require prohibitively large effort from a human expert. Therefore, to measure recall we use cross-validation. TCDB itself provides the set of known relevant examples, since examples are guaranteed to be relevant if they actually are included in TCDB. Specifically, we divide the entire training set into ten equal parts, each maintaining the original positive/negative ratio. Ten times, we train a classifier on nine parts and test on the remaining part. Recall is then measured with test set examples only, as $|T \cap P|/|T|$, where T is the set of examples (documents or protein records) in TCDB, and P is the set of examples predicted to be relevant by the classifier.

TABLE II
OBSERVED PRECISION BASED ON CLASSIFYING RANDOMLY SELECTED TEST EXAMPLES.

Method	Test Set Size	Predictions	Correct	Precision
Rule 1	1,000	101	68	0.673
Rule 2	1,000	77	60	0.779
Learned Record Classifier	4,906	278	192	0.691
Document Classifier	1,000	62	51	0.823

TABLE III
OBSERVED RECALL BASED ON CLASSIFYING EXAMPLES IN TCDB (HENCE CERTAINLY RELEVANT). NUMBERS VARY BECAUSE EXAMPLES ARE EITHER SWISS-PROT RECORDS OR MEDLINE DOCUMENTS, AND EXPERIMENTS WERE CARRIED OUT AT DIFFERENT TIMES.

Method	Examples in TCDB	Correct	Recall
Rule 1	2,088	1,534	0.735
Rule 2	2,088	1,492	0.715
Learned Record Classifier	2,453	2,193	0.894
Document Classifier	3,168	2,715	0.857

C. Results

1) Handcrafted Rule-Based Classifiers: To measure precision, we select random examples from the same distribution as our training sets, not including those examples that are in TCDB (training and evaluation sets are always made to be mutually exclusive). In the case of our rule-based methods, we randomly select 1,000 Swiss-Prot records. We use only Swiss-Prot records because they are more likely to be well-characterized and therefore better suited to the goals of updating TCDB. For each Swiss-Prot record predicted to be relevant, a human expert determines whether or not it describes a protein related to transmembrane transport. 101 of the 1,000 Swiss-Prot records matched Rule 1 (GO terms *transport* or *transporter activity*). The human expert considers 68 of these to be actually transport-related (a precision of 0.673). 77 of the 1,000 records matched Rule 2 (having a combination of keywords). The human expert considers 60 of these to be transport-related (a precision of 0.779).

Both rules fail to recognize at least 25% of records known to be relevant. Specifically, Rule 1 matches 1,534 of the 2,088 Swiss-Prot records present in TCDB (a recall of 0.735) while Rule 2 matches 1,492 (a recall of 0.715) of these records. Rule 1 matches a further 678 of 1,111 TrEMBL records present in TCDB (recall=0.610) and Rule 2 matches 649 (0.584). As mentioned in Section III-A, 129 of the 3,199 total records (about 4%) contain neither keywords nor GO terms, hence cannot be recognized.

Rule 2 fails to retrieve many records with a keyword in set 3 because the records have not been assigned the keyword *transmembrane*. This reveals the incompleteness of protein annotations. Consider the keyword *electron transport*, which is in keyword set 3. Electrons can be transported within the cell as well as across cell membranes, so to remove ambiguity it is necessary for the keyword *transmembrane* to be assigned also, if it is appropriate. However, many records labeled *electron transport* that should be labeled *transmembrane* also are not so labeled. The same is true for other keywords such as *protein transport* and *lipid transport*.

TABLE IV
COMPUTED F1 VALUES FOR ALTERNATIVE CLASSIFIERS.

Method	F1
Rule 1	0.703
Rule 2	0.745
Learned Record Classifier	0.779
Document Classifier	0.839

2) Learned Protein Database Record Classifier: In the case of our learned protein record classifier, we use the set of 4,906 randomly-selected Swiss-Prot records described in [6]. These records are used as training data, so we use cross-validation to evaluate them as test set results. 278 of the 4,906 records were predicted to be relevant. After the human expert labeled all of them, 192 turned out to be relevant, which is a precision of 0.691. Cross-validation experiments predict 2,193 of the 2,453 Swiss-Prot records in TCDB to be positive, for a recall of 0.894.

When the human expert examines the positive predictions and labels them as relevant or irrelevant, the set of labeled examples that can be used for training increases in size. Although here, we aim only to measure the accuracy of a classifier on held-aside data representing future instances, it is worth noting that by iteratively collecting labels and retraining the classifier, we can improve its accuracy by a significant amount [6].

3) Learned Document Classifier: In the case of the MEDLINE document classifier, we randomly choose 1,000 documents from the universe of MEDLINE documents described previously, *i.e.* that contain the word "protein" and come from the set of 108 journals recommended by the human expert. Of the 1,000 documents, the classifier predicted 62 to be relevant. The human expert reviewed all of these, and 51 were judged to be relevant, which is a precision of 0.823. Using cross-validation to classify documents in TCDB, 2,715 of the 3,168 MEDLINE documents included in TCDB are predicted to be positive for a recall of 0.857.

All measures of precision and recall are summarized in Tables II and III, respectively. In order to evaluate precision and recall on the same types of records, results in Table III are based only on Swiss-Prot (not TrEMBL) records. No method dominates the others in terms of both precision and recall.

One way of ranking the methods is by their F1 scores (Table IV). F1, a commonly used statistic, is the harmonic mean of precision and recall ($\frac{2 \times P \times R}{P+R}$). Since we estimate precision and recall using different sets of examples, the usual tradeoff between precision and recall may not be reflected in the F1 scores we report. Therefore, they may not be directly comparable to F1 scores reported in other papers. However, they are meaningful for comparing the methods described in this paper.

V. DISCUSSION

Precision and recall results are important indicators, but they are not the only way to gauge success. Future users will be concerned with whether or not the methods can identify especially novel and interesting documents or records. During the manual evaluation of the SVM classifier, the expert identified 11 documents (18% of 62 retrieved documents, and 22% of the 51 identified as relevant by the expert) as particularly interesting (*I* in Figure 1). Similarly, using Rule 2, the expert identified 12 Swiss-Prot or TrEMBL records (16% of the 77 retrieved records,

and 20% of the 60 identified as actually relevant) as particularly interesting. These examples are interesting because the expert was not previously aware of the proteins mentioned in these documents or records, or because these proteins belong to a new class of proteins not yet included in TCDB. All of these examples are more valuable to TCDB than the other relevant examples. In fact, a new class of toxin proteins (Transport Classification 8.B) that target transport proteins was added to TCDB as a direct result of the particularly interesting articles about them that were discovered during this research. This ability to detect truly novel relevant proteins is a success for our methods.

An important advantage of the handcrafted methods is that the rules used to query the database are easily comprehensible to humans, while at the same time providing high precision in manual evaluations. However, our results indicate that the machine learning methods are able to predict relevance with even higher precision.

The classifier trained on MEDLINE documents has high precision also, but it is typically more difficult to import information from documents into a database than it is to import information from another database, since in the former case the information needs to be extracted from the text of a paper. However, the use of MEDLINE documents has the advantage that new information is made available sooner. Papers are often available months before the corresponding protein data makes it into Swiss-Prot or TrEMBL. It is worth noting also that there can be papers that are about more than one protein, thus making some papers potentially more valuable than others. This should be taken into account when choosing a method for updating a specialized database.

In our experiments to measure precision, the human expert reviewed and labeled all of the classifiers' predictions. However, a further advantage of machine learning approaches over fixed rules like the ones evaluated here is that many classifier-learning algorithms (including all the ones mentioned and used here) can be used to score and rank predictions in terms of their likelihood of relevance. This means that the curators of a specialized database are able to tune the recall/precision tradeoff, for instance by reviewing predictions only until the false positive rate gets too high.

VI. CONCLUSIONS AND CURRENT STATUS

This paper shows the value of machine learning methods for updating specialized databases. Many research groups have focused on developing good algorithms for similar tasks, but there has been a paucity of evidence that machine learning approaches outperform simpler rule-based systems in real case studies. By performing such a comparison in the context of TCDB, this paper demonstrates both the utility of the machine learning approach and principled methods for performing comparisons between different approaches.

We have deployed our classifiers to create working systems. These systems are always online in the sense that training and deployment data are constantly being updated with new training labels from TCDB and new examples from MEDLINE or Swiss-Prot. The document classification system and protein record classification system use up-to-date article references in TCDB and protein records in TCDB, respectively, as positive training data. They use new MEDLINE articles and new Swiss-Prot records, respectively, as unlabeled training data. The workflow
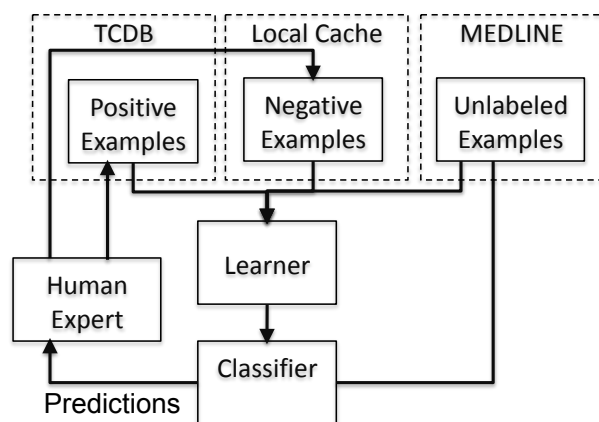
Fig. 2. TCDB update process workflow. Documents of all labels are used to train the classifier. Positive examples are obtained from TCDB, negative examples are articles previously rejected by the human expert. Unlabeled articles are ranked according to the likelihood of relevance by the classifier and then deployed to the human expert, who labels the documents and incorporates the relevant ones into TCDB.

for the MEDLINE document classification process is shown in Figure 2.

After either classifier is run, each example (article or protein record) is assigned a score proportional to the likelihood of relevance to TCDB. The examples with the top-ranking scores are delivered to the human expert for examination. It takes either system approximately 2.5 hours to download data, retrain a classifier from scratch, and rank all examples. We run the system overnight so that up-to-date predictions are available on demand. The system is run on a 2GHz Power Mac computer with 2 GB RAM and running Apple OS X.

In the past few years, the automated systems have identified thousands of new proteins that have been added to TCDB. The last year has been the fastest period of growth in TCDB's history, and nearly all of the articles and proteins that have been added to TCDB were found by the automated systems.

ACKNOWLEDGMENTS

REFERENCES

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proc. ACM SIGIR*, pages 541–548, 2006.

[2] E. Camon et al. The Gene Ontology Annotation (GOA) project: Implementation of GO in Swiss-Prot, TrEMBL, and InterPro. *Genome Res.*, 13(4):662–672, 2003.

[3] Y. Chen et al. SPD–A web-based secreted protein database. *Nucleic Acids Res.*, 33(Database Issue):D169–D173, 2005.

[4] F. M. Couto, B. Martins, and M. J. Silva. Classifying biological articles using web resources. In *Proc. ACM Symp. Appl. Computing*, pages 111–115, 2004.

[5] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proc. 7th Intl. Conf. on Intelligent Systems for Molecular Biol.*, 1999.

[6] S. Das, M. H. Saier Jr., and C. Elkan. Finding transport proteins in a general protein database. In *Proceedings of the Eleventh European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 54–66, 2007.

[7] G. Dellaire, R. Farrall, and W. A. Bickmore. The Nuclear Protein Database (NPD): Sub-nuclear localisation and functional annotation of the nuclear proteome. *Nucleic Acids Res.*, 31:328–330, 2003.

[8] F. Denis. PAC learning from positive statistical queries. In *Proceedings of the 9th International Conference on Algorithmic Learning Theory (ALT'98), Otzenhausen, Germany*, volume 1501 of *Lecture Notes in Computer Science*, pages 112–126. Springer, 1998.

[9] F. Denis, R. Gilleron, and M. Tommasi. Text classification from positive and unlabeled examples. In *Proceedings of the Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2002)*, pages 1927–1934, 2002.

[10] I. Donaldson et al. PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4(1), 2003.

[11] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pages 213–220, 2008.

[12] G. P. C. Fung, J. X. Yu, H. Lu, and P. S. Yu. Text classification without negative examples revisit (sic). *IEEE Transactions on Knowledge and Data Engineering*, 18(1):6–20, 2006.

[13] M.Y. Galperin and G.R. Cochrane. Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. *Nucleic Acids Research*, 37(Database issue):D1–D4, 2009.

[14] The Gene Ontology Consortium. Gene Ontology: Tool for the unification of biology. *Nat. Genet.*, 25:25–29, 2002.

[15] G. Grumbling, V. Strelets, and The FlyBase Consortium. Flybase: Anatomical data, images and queries. *Nucleic Acids Res.*, 34:D484–D488, 2006.

[16] W. Hersh. Evaluation of biomedical text-mining systems: Lessons learned from information retrieval. *Briefings in Bioinformatics*, 6(4):344–356, December 2005.

[17] W. Hersh, A. Cohen, Yang J., R. T. Bhupatiraju, P. Roberts, and M. Hearst. Trec 2005 genomics track overview. In *Proc. TREC*, 2005.

[18] W. Hou, C. Lee, and H. Chen. Classifying biological full-text articles for multi-database curation. In *Proc. Eur. Chap. of ACL*, pages 159–162, April 2006.

[19] M. Iwayama, A. Fujii, N. Kando, and Y. Marukawa. An empirical study on retrieval models for different document genres: patents and newspaper articles. In *Proc. ACM SIGIR*, pages 251–258, 2003.

[20] T. Joachims. Making large-scale support vector machine learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods: Support Vector Machines*, pages 169–184. MIT Press, Cambridge, MA, 1998.

[21] T. Joachims. Transductive inference for text classification using support vector machines. In Ivan Bratko and Saso Dzeroski, editors, *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann, 1999.

[22] M. Krallinger and A. Valencia. Text-mining and information-retrieval services for molecular biology. *Genome Biol.*, 6(7):224–230, 2005.

[23] E. Kretschmann, W. Fleischmann, and R. Apweiler. Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, 17(10):920–926, 2001.

[24] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003)*, pages 179–188, 2003.

[25] H. Liu, G. Xu, M. Torii, Z. Hu, and J. Goll. Learning from positives and unlabeled for bacterial protein-protein interaction document retrieval. *Lecture Notes in Bioinformatics*, page (to appear), 2009.

[26] A. K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. Unpublished, 1996.

[27] A. K. McCallum. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[28] Y. Regev, M. Finkelstein-Landau, R. Feldman, R. Gorodetsky, X. Zheng, S. Levy, R. Charlab, C. Lawrence, R. A. Lippert, Q. Zhang, and H. Shatkay. Rule-based extraction of experimental evidence in the biomedical domain - the KDD Cup (Task 1). *SIGKDD Explorations*, 4(2):90–92, 2002.

[29] M. H. Saier Jr., C. V. Tran, and R. D. Barabote. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res.*, 36(Database Issue):D181–D186, 2006.

[30] G. Salton. *The SMART Retrieval System; Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[31] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inform. Process. Manag.*, 24(5):513–523, 1988.

[32] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

[33] B. Schölkopf, A. Smola, R. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[34] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47, 2002.

[35] H. Shatkay, N. Chen, and D. Blostein. Integrating image data into biomedical text categorization. *Bioinformatics*, 22(14):e446–e453, 2006.

[36] A. S. Yeh, L. Hirschman, and A. A. Morgan. Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup. *Bioinformatics*, 19(Suppl. 1):i331–i339, 2003.

[37] D. Zhang and W. S. Lee. A simple probabilistic approach to learning from positive and unlabeled examples. In *Proceedings of the 5th Annual UK Workshop on Computational Intelligence (UKCI)*, pages 83–87, September 2005.