2016

Using Semi-supervised Learning for the Creation of Medical Systematic Review: An exploratory Analysis

Prem Timsina
Dakota State University

Jun Liu
Dakota State University

Omar F. El-Gayar
Dakota State University

Yanyan Shang
Dakota State University

Using Semi-supervised Learning for the Creation of Medical Systematic Review: An exploratory Analysis

Prem Timsina
Dakota State University
ptimsina@pluto.dsu.edu

Jun Liu
Dakota State University
Jun.liu@dsu.edu

Omar El-Gayar
Dakota State University
Omar.el-gayar@dsu.edu

Yanyan Shang
Dakota State University
yshang@pluto.dsu.edu

Abstract

In this research, we explore semi-supervised learning based classifiers to identify articles that can be included when creating medical systematic reviews (SRs). Specifically, we perform comparative study of various semi-supervised learning algorithm, and identify the best technique that is suited for SRs creation. We also aim to identify whether semi-supervised learning technique with few labeled samples produce meaningful work saving for SRs creation. Through an empirical study, we demonstrate that semi-supervised classifiers are viable for selecting articles for systematic reviews and situations when only a few numbers of training samples are available.

1. Introduction

According to Higgins and Green [1], "*a systematic review is a high-level overview of primary research on a particular research question that tries to identify, select, synthesize and appraise all high quality research evidence relevant to that question in order to answer it*". Moreover, Khan et al. [2] notes that "*A review earns the adjective systematic if it is based on a clearly formulated question, identifies relevant studies, appraises their quality and summarizes the evidence by use of explicit methodology*" These systematic reviews (SRs) translate biomedical research into practical guidelines that inform clinicians, researchers, and policymakers for informed decision-making. Each systematic review addresses a clearly formulated problem. An example of systematic review may be "*Screening for Cognitive Impairment in Older Adults: A Systematic Review for the U.S. Preventive Services Task Force*" [3]. This study was aimed to identify the diagnostic accuracy of brief cognitive screening instruments and the benefits and harms of pharmacologic and nonpharmacologic interventions for early cognitive impairment. Developing a medical systematic review is a much more demanding, rigorous, and resource-intensive process. The general workflow of systematic review consists of 1) Perform keyword search to identify potentially relevant articles 2) Perform article triage procedure to identify relevant articles for the topic, and 3) Finally, summarized the articles in the form evidence report via meta-analysis or qualitative analysis technique [2].

The second step is particularly resource intensive. Specifically, articles are triaged in two steps [4]. First, the title and abstract of an article are reviewed to identify if the full text of the article should be examined. This step may involve screening potentially thousands of titles and abstracts. Second, a full text inspection will be conducted of the selected articles based on the titles and abstracts to determine if the articles satisfy the inclusion criteria and should be included in the systematic review. This step entails the screening and review of hundreds to thousands of full-text articles. An initial search by querying databases such as Medline, Cochrane and Embase often returns a large number of articles given a medical topic. For example, Lin et. al [3] retrieved 16,179 articles based on keywords such as "cognitive impairment", "cognitive impairment and older adults" in order to ensure that none of the relevant articles will be missed. Each article was manually inspected by two scientists using highly methodic procedures resulting in only 1,190 articles. Finally, 253 articles were included after full text screening of the 1,190 articles. [4]. Due to the manual workflow of selecting articles for systematic reviews (SRs), developing SRs requires a significant investment in time (1,139 expert hours on average) and

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.151

funds (up to a quarter of a million dollars) from a dedicated and qualified research team [5]

In that regard, machine learning is proposed to automate the article screening for SRs [6-8]. Machine learning has proven helpful in updating existing SRs. Most of the existing research use supervised learning assuming readily available training data and focus on updating reviews. For example, Cohen et al. [9] used 50% training and 50% validation data, Adeva et al. [8] used 90% training and 10% validation data, and other studies have embraced a similar approach. However, supervised machine learning assumes the availability of training data sets that do not necessarily exist when creating SR. A need exist to explore other approaches that are more suited to situations where training data sets are not readily available, e.g., when creating SR.

In that regard, semi-supervised learning approach has received considerable attention due to its potential for reducing the effort of labeling data. Some often used methods include semi-supervised support vector machine, self-training, graph-based algorithm, generative mixture models [10]. Semi-supervised learning falls between supervised and un-supervised learning techniques. This approach holds greater promise if positive class is very rare and labeling through sequential scanning of samples is very costly.

The aim of this research is to perform an exploratory analysis of semi-supervised learning techniques for article selection for medical systematic review creation. More specifically, given the fact that when it comes to creating a new SR, labeled training data (i.e., articles that have been reviewed by human experts to be included in or excluded from a systematic review) is not readily available and is difficult and time-consuming to obtain, we plan to explore semi-supervised learning to overcome this labeling bottleneck and develop data mining models that can classify articles for inclusion or exclusion, thus helping automate SR creation with only a few labeled instances. We perform comparative study of various semi-supervised learning algorithm, and identify best technique that is suited for SRs creation procedure. To our knowledge, the proposed research is one of the first that attempts to address the small-sized training dataset problem that hampers the use of classification algorithms in SR creation.

2. Related Work

There have been some attempts in literature to leverage supervised machine learning to automate SR update procedure [6, 7, 11-13]. There are also other studies, though not in the area of SRs, that demonstrated the possibility of semi-supervised learning in the case of rare training instances. For example, Song et al [14] proposed an approach for Protein-protein interaction (PPI) extraction technique by combining Deterministic Annealing- based semi-supervised learning and an active learning technique to extract protein-protein interaction. Through three experiments with different PPI corpuses, authors showed that PPISpotter is superior to the other techniques incorporated into semi-supervised SVMs such as Random Sampling, Clustering, and Transductive SVMs. In another example, Jin et al. [15] evaluated the self-learning SVM and proved that their method is better than the former algorithm. Using their self-training semi-supervised SVM algorithm, author were able to save much time for labeling the un-labeled data and obtain a better classifier with good performance.

Overall, extant research focuses on applying supervised learning to article selection for a SR, assuming the existence of a large number of labeled training examples. Supervised learning is practical for article selection in SR updates, but less feasible for SR creations that often start with zero or few labeled articles. While there is a number of semi-supervised learning techniques that are proposed in literature; however, we did not find any studies that attempted to thoroughly investigate semi-supervised learning in the context of medical systematic review creation. Also, existing literature in SRs automation indicate that semi-supervised learning system that is carefully designed is possible in principle, and is an interesting area for future research for SR creations [16]. Existing research on semi-supervised learning also demonstrate promise for text classification with few labeled examples. We hence propose to investigate semi-supervised learning to systematic review creation.

3. Research Gap

Our literature review indicates that 1) the generation of a training dataset for article classification is expensive and requires significant human effort 2) it is necessary to identify machine learning techniques that is able to learn with small amount of training dataset 3) it has become essential to perform comparative investigation of semi-supervised learning techniques in the context of systematic review creation This leads us to the following research questions:

1. *What is the most-suited semi-supervised learning technique in the context of systematic review creation?*

We plan to address this question by investigating various semi-supervised based machine learning

approach. Specifically, we will investigate various graph-based algorithms, and semi-supervised support vector learning.

2. *Are semi-supervised learning algorithms always superior than supervised learning algorithms? If not, what is the break-even point of supervised and semi-supervised learning algorithm?*

We plan to compare semi-supervised based algorithm with supervised learning algorithms with different percentage of training dataset, and calculate break-even point of semi-supervised and supervised learning algorithm.

3. *Is semi-supervised learning viable technique for Systematic Review Creation?*

To address this issue, we will compare work saving in semi-supervised learning technique (with few samples) with supervised learning technique with (with complete training set). Here, we will identify if semi-supervised learning based systematic review creation produce meaningful empirical outcome.

4. Methodology

Our analytics approach for this research includes three major components: 1) evaluating the effectiveness of different semi-supervised learning algorithms in systematic review creation, 2) comparing semi-supervised learning vs. supervised learning, and 3) determining if semi-supervised learning is feasible for systematic review creation with empirical evidence. We conduct experiments using four systematic review datasets and compared our approach with others that were proposed in existing research. In following sub-sections, we describe the data source, each component in our approach, and the methods we compare our approach with in detail.

4.1. Data Sets

We used datasets from AHRQ's Evidence-based Practice Center (EPC) at Oregon Health and Science University. Specifically, we selected datasets of four systematic reviews drug topics—ACEInhibitors (ACE), AtypicalAntipsychotics (AT), NSAID, and Estrogens (ESTRO)". The original datasets downloaded from [17] include the PubMed Unique Identifiers (PMID) of all the articles, whether included or excluded from the reviews, and the inclusion and exclusion decisions made by human researchers. Table 1 provides an overview of datasets.

Table 1: Overview of Data Corpus

Dataset	Total number of articles	Number of excluded articles	Number of included articles	Ratio—Included vs. Excluded
ACEInhibitors (ACE)	2546	2362	184	1:13.84
Antihistamines (AT)	1120	751	361	1:2
NSAID	393	305	88	1:3.5
Estrogens (ESTRO)	370	289	81	1:3.6

4.2. Data Pre-processing

We represented each article in our datasets using the bag-of-word model [18] that includes 1-grams (i.e., single words), 2-grams (i.e., two-word phrases) and 3-grams (i.e., phrases including three words). We created a feature vector for each article that includes the words/phrases in the title, abstract, Medline publication type, and Medical Subject Heading of the article. The data pre-processing included 1) removing non-English and non-alphanumeric characters (e.g., characters like !, #, *) , 2) removing English Stop-words (e.g., like a, an, the), and 3) converting all uppercase words into lower case. To create the bag-of-words, we used the term frequency-inverse document frequency (tf-idf) technique [19]. In the tf-idf scheme, we use all words/phrases in the corpus as the features. For each document in the corpus, a count is formed of the number of occurrences of each word/phrase. After suitable normalization, this term frequency count is compared to an inverse document frequency count, which measures the number of occurrences of the word/phrases in the entire corpus (generally on a log scale, and again suitably normalized). The end result is a feature vector for each of the documents in the corpus. The feature vector includes the tf-idf scores of the words/phrases contained in the document (if a word or phrase does not exist in the document, we assign 0 to it).

4.2. Semi-supervised Learning Algorithm

There exist a number of semi-supervised learning algorithms in literatures. Here, we evaluate the effectiveness of three widely used ones. For two algorithms, Label Spreading and Label Propagation, we consider two variations for each.

Label Spreading Algorithm

We investigate label-spreading algorithm with RBF and KNN kernel[20]. The key assumption in label spreading is that geometrically closer data points tend to be similar. There are two general ideas related to label spreading: 1) a example's label propagates to its neighboring examples according to their proximity, and 2) the labeled examples act as sources that push out labels to unlabeled data. Below, we describe the label-spreading algorithm in detail.

- Form the affinity matrix W defined by $W_{ij} = \exp(-\|x_i - x_j\|^2/2\sigma^2)$ if $i \neq j$ and $W_{ii} = 0$.
- Construct the matrix $S = D^{-1/2}WD^{-1/2}$ in which D is a diagonal matrix with its (i, i)-element equal to the sum of the i-th row of W.
- Iterate $F(t+1)=\alpha SF(t)+(1-\alpha)Y$ until convergence,where α is aparameter in (0, 1).
- Let F^* denote the limit of the sequence {F(t)}. Label each point x_i as a label $y = \text{argmax}_{j \leq c} F^*_{ij}$.

This algorithm can be understood intuitively in terms of spreading activation networks [21, 22] from experimental psychology.

Label Propagation Algorithm

We also investigate the label-propagation algorithm with RBF and KNN kernel [23]. The key idea of label propagation is node's labels propagate to neighboring nodes according to their proximity. Meanwhile labels are clamped on the labeled data. The labeled data act like sources that push out labels through unlabeled data [23].

Let $(x_1, y_1)...(x_l, y_l)$ be labeled data, where $Y_L = \{y_1...y_l\} \in \{1...C\}$ are the class labels. We assume the number of classes C is known, and all classes are present in the labeled data. Let $(x_{l+1}, y_{l+1})...(x_{l+u}, y_{l+u})$ be unlabeled data, where $Y_U = \{y_{l+1}...y_{l+u}\}$ are observed; usually, l<<u. Let $X = \{x_1...x_{l+u}\} \in R^D$. The problem is to estimate Y_U from X and Y_L.

Intuitively, we want data points that are close to have similar labels. We create a fully connected graph where the nodes are all data points, both labeled and unlabeled. The edge between any nodes i, j is weighted so that the closer the nodes are in local Euclidean distance dij, the larger the weight w_{ij}. The weight are controlled by a parameter σ.

$$\omega_{ij} = \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) = \exp\left(-\Sigma_{d=1}^{D}(x_i^d - x_j^d)^2 \Big/ \sigma^2\right)$$

The algorithm of label spreading algorithm is as follows:

- All nodes propagate labels for one step: $Y \longleftarrow TY$
- Row-normalize Y to maintain the class probability interpretation.
- Clamp the labeled Data. Repeat from step 2 until Y converges.

Semi-supervised Support Vector Machine (S3VM)

This algorithm start by creating a multidimensional plane with large margin over labeled data. It then aims to fit a plane that separate the data (with large margin) into labeled and unlabeled data. Then, those unlabeled documents that lies in the side of X documents (inclusion trial) are labeled as X (inclusion trial); whereas, other documents are labeled as Y (exclusion trial) [24]. S3VM is applicable wherever SVMs are applicable. In text classification SVMs perform better than other classifier and is expected to outperform other classifiers. The main drawback of S3VM is optimization is currently difficult [25].

Support Vector Machine (SVM)

Existing studies such as [6, 26, 27] have proved the effectiveness of SVM with a linear kernel in text classification in the process of medical systematic reviews. The optimization problem associated with the SVM is shown below.

$$min_{\mathbf{w},b} \frac{\mathbf{w^T w}}{2}$$

$$\text{subject to: } y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 \ (\forall \text{ data points } \mathbf{x_i}).$$

where for each data point (x_i, y_i), y_i is either 1 or −1, indicating the class to which the point belongs. The two hyperplanes $w \cdot x - b = 1$ and $w \cdot x - b = -1$ are called support vectors that separate the data. SVM maximizes the distance (called "margin") between the support vectors.

Soft-margin linear SVM: In our earlier research, we performed comparative investigation of Neural Networks, SVMs, Naïve Bayes, Nearest Neighbor and identified that soft-margin SVM outperforms other algorithms in context of systematic review creation (research published elsewhere, citation after reviewer's comment). Thus, we propose to use the soft-margin Support Vector Machine (SVM) with a linear kernel as a supervised machine-learning algorithm. Soft-margin SVM is an extension of the standard "hard" margin SVM described above.

The "hard-margin" SVM sometimes does not work well since it does not allow data points in the margin. However, data is not often perfectly linearly separable, and it is necessary to allow some data points of one class to appear within the region bounded by the support vectors. Soft-margin SVM provides the flexibility by introducing a slack variable $\epsilon_i \geq 0$, and the optimization problem of soft-margin SVM becomes[28]:

$$min_{\mathbf{w},b,\epsilon} \frac{\mathbf{w}^T\mathbf{w}}{2} + C\sum_i \epsilon_i$$

$$\text{subject to: } y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \epsilon_i \text{ and } \epsilon_i \geq 0 \ (\forall \text{ data points } \mathbf{x_i}).$$

where ϵ_i, the slack variable, represents the degree of error in classification.

5. Evaluation

We evaluated the classification performance using four measures: precision, recall, F1-score and Work Saved a measure proposed in[13]. These measures are defined based on a confusion matrix as shown in Table 2

Table 2: Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Recall refers to the rate of correctly classified positives among all positives and is equal to TP divided by the sum of TP and FN (TP/ (TP+FN)). *Precision* refers to the rate of correctly classified positives among all examples classified as positive and is equal to the ratio of TP to the sum of TP and FP TP/(TP+FP). *F1* means the harmonic mean of recall and precision ((2*recall*precision)/(recall + precision)). WSS defined as percentage of samples that met the initial search criteria that the human reviewers do not have to read because they have been correctly screened by the classifier ((TN + FN)/(TN + FN + TP + FP) – 1+ TP/(TP + FN)).

6. Experimental Design and Results

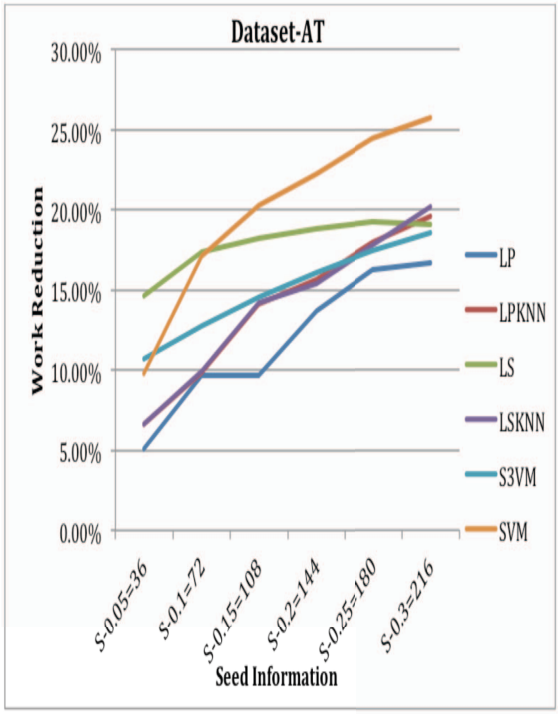
Table 3. Overview of experiments

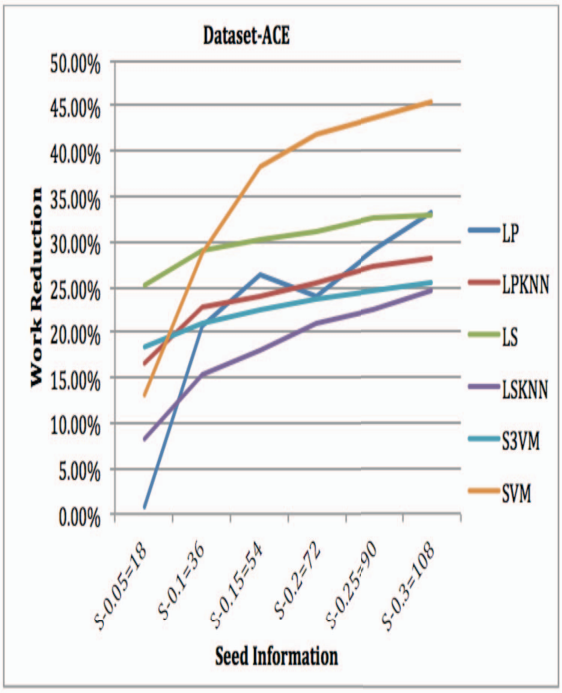
	Description	Algorithms	Goal
Exp. 1	Comparison of semi-supervised algorithms	Label Spreading, Label Propagation, and S3VM	Identify the best semi-supervised learning algorithm.
Exp. 2	Compare semi-supervised learning vs. supervised learning	Label Spreading with RBF kernel, Soft-margin SVM with polynomial kernel	1)Identify a turning point after which supervised learning outperform semi-supervised learning. 2)Evaluate if semi-supervised learning is feasible for systematic review creation.

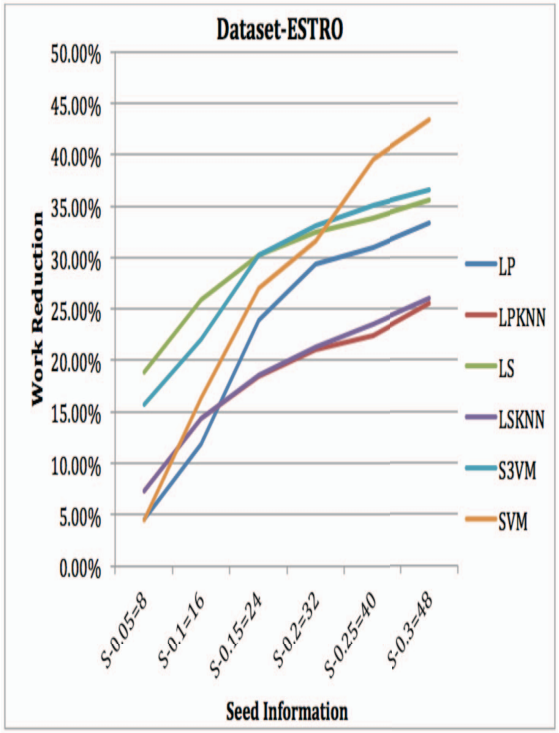
We conducted two experiments to evaluate the effectiveness different semi-supervised learning algorithms vs. a supervised learning algorithm. The datasets we use in the experiments are the four datasets we described in section 4.1. The detail of our experiment design is illustrated in Table 3. In experiment 1, we evaluated the effectiveness of five semi-supervised learning algorithms. Experiment 1 consist of 6 sub-experiments, and in each sub-experiment, we used different numbers of seeds (i.e., initially labeled articles). We used 5% positive examples in sub-experiment 1, 10% positive examples in sub-experiment 2, 15% in sub-experiment 3, 20% in sub-experiment 4, and 25% in sub-experiment 5 and 30% in sub-experiment 6. We also randomly selected the same number of negative examples in each sub-experiment. To ensure the reliability of the results, in each step, we conducted 100 trials. Then, we averaged the results of 100 trials to generate the final results for each sub-experiment. This approach is consistent with an earlier approach used in literature in this kind of research [23]. In experiment 2, we focused on comparing supervised learning technique with semi-supervised learning technique. Our hypothesis is, below a certain number of training samples, semi-supervised learning works better than supervised learning. There is a turning point in terms of number of training samples. After the turning point, supervised learning algorithms perform better than semi-supervised learning. In this experiment, we compared soft-margin SVM (supervised learning algorithm) with the best semi-supervised learning algorithm identified in experiment 1. By comparing the work saving of the semi-supervised learning with a few training examples vs. that of the supervised learning with the whole training dataset, we intend to evaluate the feasibility of using semi-supervised learning for systematic review creation.

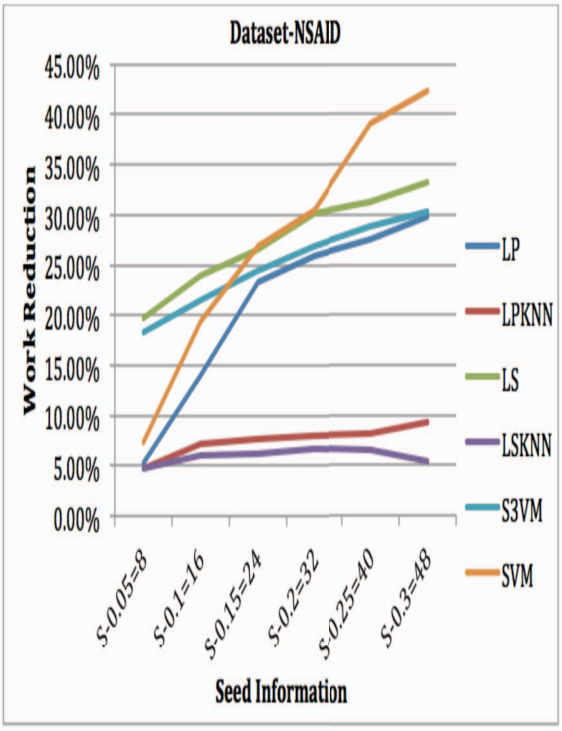
6.1. Comparison of Learning Algorithms

We performed investigation of five semi-supervised learning algorithms (label-spreading with RBF kernel, label-spreading with KNN kernel, label-propogation with RBF kernel, label-propogation with KNN kernel, and S3VM), and compared them with soft-margin as supervised learning algorithm. We tested all algorithms with 5%, 10%, 15%, 20%, 25% and 30% training samples. The results are shown in Figure 1. Each sub-figure in Figure 1 shows the work reduction values we obtained when we applied the different algorithms to a dataset. On the x axis of each sub-figure represents the number and percentages of training examples used. For instance, in the first sub-figures, we used first 5% of training samples, which is equivalent to 36 samples. We then used 10%, 15%,









***LP**= Label Propagation with rbf kernel, **LPKNN**= Label Propagation with knn kernel, **LS**= Label Spreading with rbf kernel, **LSKNN**= Label Spreading with knn kernel, **S3VM**= Semi-supervise Support Vector Machine, **SVM**= Support Vector Machine*
***Horizontal Notation in Figure (like S-0.05:8)=** 5% training set , which equals 8 samples in training set (4 being positive and 4 being negative)*

Figure 1: Comparison of Algorithms

20%, 25% and 30% of training samples in later iterations. As shown in Figure 1, label-spreading with RBF kernel outperforms all other semi-supervised based learning algorithm in all four datasets. S3VM also showed comparatively good results. Figure 1 also shows that the slope of curve increased rapidly first as we increased the number of training samples, but after a certain sample size, there is no crucial improvement in work reduction as we add more training samples. It is obvious that even with a very small size of training data, we can achieve a considerable amount of work saving. For example, in NSAID dataset, with 8 training samples (4 being positive and 4 being negative), we obtained work reduction of 20%. In comparison, given the same dataset, the researchers in [13] used 20,000 samples for training and obtained work saving of 37.1% using supervised learning.

Obviously, semi-supervised learning does not always outperform supervised-based learning algorithm. Figure 1 shows that after 30-40 samples (with 50% of them being positive examples and 50% being negative), the supervised learning algorithm, polynomial SVM started to outperform the label-spreading algorithms in three datasets, ACE, ESTRO and NSAID. In the dataset AT supervised technique outperforms semi-supervised technique after using 72 training examples. It seems that in order to conduct semi-supervised learning, it is necessary to manually identify about 20 to 30 positive articles (articles that will be included in a systematic review) to achieve comparable results with supervised learning.

6.2. Effectiveness of Semi-supervised learning algorithm

Table 4. Effectiveness of semi-supervised learning algorithm (with a small number of samples) as compared with supervised learning (with the whole dataset)

Data set	Algorithm	Train	Validate	No. of Positive Samples-training	No. of Negative Samples-training	TN	FP	FN	TP	Recall	Precision	F1	WSS
ACE	**LS**	**0.10**	**0.9**	**18.00**	**18.00**	**1122.48**	**1220.52**	**27.64**	**137.36**	**0.83**	**0.10**	**0.18**	**29.11%**
ACE	SVM	0.50	0.5	90.00	90.00	1623.74	647.26	19.33	73.67	0.79	0.10	0.18	48.72%
ACE	SVM	0.70	0.5	151.00	151.00	1611.58	623.42	10.88	46.12	0.81	0.07	0.13	51.70%
ACE	SVM	0.90	0.1	162.00	162.00	1623.64	573.36	3.60	15.40	0.81	0.03	0.05	54.48%
AT	**LS**	**0.10**	**0.9**	**36.00**	**36.00**	**313.02**	**407.98**	**59.55**	**267.45**	**0.82**	**0.40**	**0.53**	**17.34%**
AT	SVM	0.50	0.5	180.00	180.00	353.16	223.84	38.01	144.99	0.79	0.39	0.53	30.70%
AT	SVM	0.70	0.5	252.00	252.00	315.84	189.16	21.34	89.66	0.81	0.32	0.46	35.51%
AT	SVM	0.90	0.1	324.00	324.00	282.80	148.20	7.19	29.81	0.81	0.17	0.28	42.53%
ESTRO	**LS**	**0.20**	**0.8**	**16.00**	**16.00**	**136.47**	**135.53**	**6.40**	**57.60**	**0.90**	**0.30**	**0.45**	**32.52%**
ESTRO	SVM	0.50	0.5	40.00	40.00	171.04	76.96	4.15	35.85	0.90	0.32	0.47	50.46%
ESTRO	SVM	0.70	0.5	66.00	66.00	165.48	66.52	2.35	21.65	0.90	0.25	0.39	55.77%
ESTRO	SVM	0.90	0.1	72.00	72.00	159.35	56.65	0.75	7.25	0.91	0.11	0.20	62.09%
NSAID	**LS**	**0.20**	**0.8**	**16.00**	**16.00**	**136.41**	**152.59**	**6.87**	**65.13**	**0.90**	**0.30**	**0.45**	**30.15%**
NSAID	SVM	0.50	0.5	40.00	40.00	194.47	70.53	7.03	40.97	0.85	0.37	0.52	49.73%
NSAID	SVM	0.70	0.5	62.00	62.00	188.49	60.51	4.73	27.27	0.85	0.31	0.46	53.98%
NSAID	SVM	0.90	0.1	72.00	72.00	177.45	48.55	1.40	7.60	0.84	0.14	0.23	60.55%

LS= Label Spreading Algorithm with RBF kernel, **SVM** = soft-marge SVM with polynomial kernel, **Train**= Percentage of training samples, **Validate=** Percentage of training samples, **TN**= True Negative, **FP**= False Positive, **FN**= False Negative, **TP**= True Positive, **WSS**= Work Saving

In experiment 2, we compared the performance of label spreading with RBF kernel with that of the soft-margin semi-supervised learning with polynomial kernel. The results are shown in Table 4. When performing semi-supervised learning, we split the whole training dataset into training and validation sets, based on different partitioning ratios. For instance, the ratio "70:30" shown in Table 5 means that 70% of the dataset was used as the training set and the remaining 30% was used as the validation set, and we conducted cross-validation. Figure 2 summarizes the results shown in Table 5. It shows that even with a small number of

training samples, semi-supervised learning produced reasonable work saving. For example, in AT dataset with 36 training samples semi-supervised learning algorithm was able to obtain work saving of 29.11%, whereas, with 324 training samples, supervised learning was able to obtain work saving of 54.48%. On surface, the difference in work saving might not seems very significant. However, in the case of medical systematic review creation, the identification of training dataset is very expensive. For instance, the development a review presented in (Couch et al. 2008) involved retrieving 12,740 articles out of which only 80 articles are positive samples. Through the random sampling approach, creation of even a single positive sample for training involves reading 160 articles (12,740:80). In such a case, supervised learning technique that needs a considerable amount of training data is not very helpful. Semi-supervised learning can be a viable technique for systematic review creation where training set is not readily available.

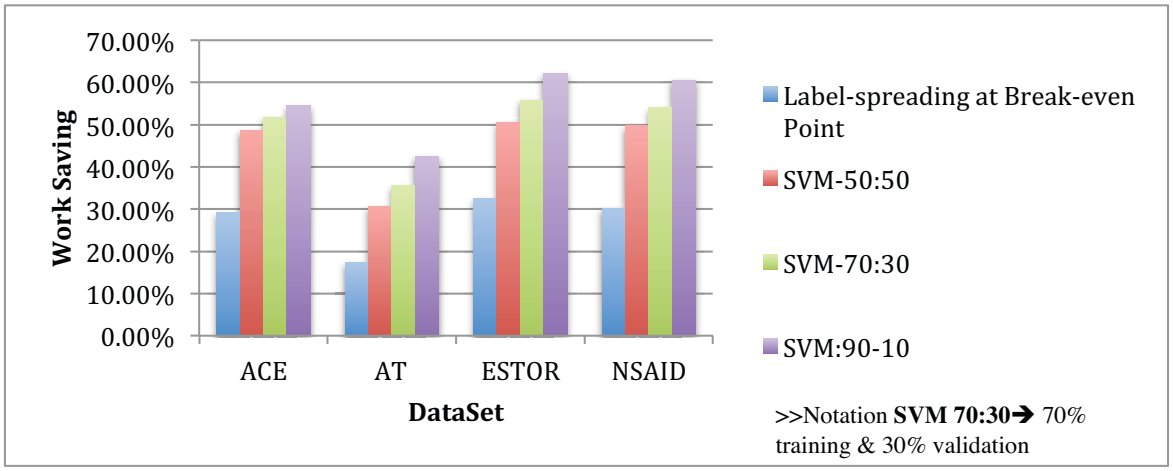


Figure 2: Comparison of semi-supervised learning algorithm (with few samples) as comparison with supervised learning (with complete training samples

8. Conclusion

This research performed an exploratory analysis of semi-supervised learning techniques for the for the article selection procedure in the case of medical systematic review creation. We demonstrated that a less explored machine learning approach, namely semi-supervised learning, is a viable technique for the problem where labeled articles are not available or very costly to obtain during systematic review creation.

From a practical and applied research perspective, this research is expected to result in a significant reduction in the cost of creating and updating systematic reviews. Over 5000 new Systematic Reviews are immediately needed to cover new medical condition. Currently, the substantial cost of SR creation impedes the translation of latest medical evidence into healthcare practice. As a result, the cases of adverse drug events, preventable medical errors, and multiple hospitals visit for same medical problem remain high. This research has potential to optimize SR creation and contribute to the adoption of evidence-based medicine. In summary, this research provides direct impact in the availability of best medical evidence, and consequently, impacts the health and wellbeing of society.

From a theoretical perspective, this research explores the possibility of creating machine learning model with very few labeled instances. In prior research, supervised-learning has been used as the de-facto standard method for article classification for SRs, which however leaves the issue of a small-sized training dataset largely unaddressed. We propose to use semi-supervised learning, which represent a novel approach that to our knowledge, has not been used in the area of SR creation. Among the various semi-supervised learning algorithms, we found label-spreading with RBF kernel outperforms other algorithms when used in the context of systematic review creation. After adding a certain number of training (30-40 training samples with 50% positive ones and 50% negative one for the datasets used in our research), we found supervised learning started to outperform semi-supervised based learning algorithms. The experiences and lessons learned from this research are expected to inform the literature regarding the efficacy of the proposed techniques and the further development and refinement of these techniques.

The research can be further extended and optimized. Currently, the research is still a work-in-progress. We are planning to examine self-learning,

active learning and ensemble techniques to further optimize the work saving. Also, the outcome of semi-supervised learning highly relies on the initial labeled set. Further research is needed to identify good seed information for the training purpose.

References

[1] J. Higgins and S. Green, "Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]," *The Cochrane Collaboration,* 2011.

[2] K. S. Khan, R. Kunz, J. Kleijnen, and G. Antes, "Five steps to conducting a systematic review," *Journal of the Royal Society of Medicine,* vol. 96, pp. 118-121, Mar 2003.

[3] J. S. Lin, E. O'Connor, R. C. Rossom, L. A. Perdue, and E. Eckstrom, "Screening for cognitive impairment in older adults: A systematic review for the U.S. Preventive Services Task Force," *Ann Intern Med,* vol. 159, pp. 601-12, Nov 5 2013.

[4] K. G. Shojania, Margaret Sampson, M. T. Ansari, and C. Garritty, "Updating Systematic Reviews," *AHRQ,* vol. 16, 2007.

[5] J. McGowan and M. Sampson, "Systematic reviews need systematic searchers," *J Med Libr Assoc,* vol. 93, pp. 74-80, Jan 2005.

[6] T. Bekhuis and D. Demner-Fushman, "Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers.," *Artificial intelligence in medicine,* vol. 55, pp. 197-207, 2012.

[7] I. Shemilt, A. Simon, G. J. Hollands, T. M. Marteau, D. Ogilvie, A. O'Mara-Eves*, et al.*, "Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews," *Research Synthesis Methods,* pp. n/a-n/a, 2013.

[8] G. Adeva, P. Atxa, U. Carrillo, and A. Zengotitabengoa, " Automatic text classification to support systematic reviews in medicine," *Expert Systems with Applications,* vol. 41, pp. 1498-1508, 2014.

[9] A. M. C. Cohen, W. R. Hersh, K. Peterson, and P.-Y. Yen, "Reducing Workload in Systematic Review Preparation Using Automated Citation Classification," *JAMIA,* vol. 13, pp. 206-219, 2006.

[10] O. Chapelle and B. Schölkopf, "Semi-Supervised Learning," *The MIT Press,* 2006.

[11] S. Ananiadou, R. Procter, B. Rea, and Y. Sasaki, "Supporting Systematic Reviews using Text Mining," vol. 3, 2009.

[12] A. Cohen, W. Ersh, and K. Eterson, "Reducing Workload in Systematic Review Preparation Using Automated Citation Classification," pp. 206-219, 2006.

[13] O. Frunza, D. Inkpen, and S. Matwin, "Building Systematic Reviews Using Automatic Text Classification Techniques," *Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics,* pp. 303-311, 2010.

[14] M. Song, H. Yu, and W. S. Han, "Combining active learning and semi-supervised learning techniques to extract protein interaction sentences," *BMC bioinformatics,* vol. 12, p. S4, 2011.

[15] Y. Jin, C. Huang, and L. Zhao, "A Semi-Supervised Learning Algorithm Based on Modified Self-training SVM," *Journal of Computers* vol. 6, pp. 1438-1443, 2011.

[16] A. M. Cohen, K. Ambert, and M. McDonagh, "Cross-topic learning for work prioritization in systematic review creation and update," *J Am Med Inform Assoc,* vol. 16, pp. 690-704, Sep-Oct 2009.

[17] A. M. C. Cohen. (2014, April 2, 2014). *Systematic Drug Class Review Gold Standard Data.* Available: http://skynet.ohsu.edu/~cohenaa/systematic-drug-class-review-data.html

[18] G. Lebanon, Y. Mao, and J. Dillon, "The Locally Weighted Bag of Words Framework for Document Representation," *The Journal of Machine Learning Research,* vol. 8, 2007.

[19] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of documentation,* vol. 60, pp. 503-520, 2004.

[20] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Sch¨olkopf, "Learning with Local and Global Consistency," *Max Planck Institute for Biological Cybernetics, 72076 Tuebingen, Germany,* 2004.

[21] J. R. Anderson, "The architecture of cognition. Harvard Univ. press, Cambridge, MA,," 1983.

[22] J. Shrager, T. Hogg, and B. A. Huberman., "Observation of phase transitions in spreading activation networks," *Science,* vol. 236, 1987.

[23] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," *Technical Report CMU-CALD-02-107, Carnegie Mellon University,* 2002.

[24] M. Seeger, *Learning with labeled and unlabeled data*: Institute for Adaptive and Neural Computation University of Edinburgh, 2002.

[25] X. Zhu, "Semi-supervised learning literature survey. TR-1530," *University of Wisconsin-Madison Department of Computer Science,* 2005.

[26] T. Joachims, "Text Categorization with Support Vector Machines : Learning with Many Relevant Features," *Universtat Dortmund,* pp. 1-19, 1998.

[27] H. Liu, S. B. Johnson, and C. Friedman, "Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS," *J Am Med Inform Assoc,* vol. 9, pp. 621-36, Nov-Dec 2002.

[28] Stanford. (2014, June 11, 2014). *Soft margin classification*. Available: http://nlp.stanford.edu/IR-book/html/htmledition/soft-margin-classification-1.html