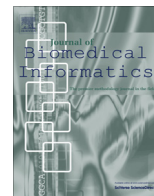




Contents lists available at SciVerse ScienceDirect

## Journal of Biomedical Informatics

journal homepage: [www.elsevier.com/locate/yjbin](http://www.elsevier.com/locate/yjbin)

# Semi-supervised clinical text classification with Laplacian SVMs: An application to cancer case management

Vijay Garla<sup>a,\*</sup>, Caroline Taylor<sup>b</sup>, Cynthia Brandt<sup>a,b,c</sup><sup>a</sup> Yale Center for Medical Informatics, Yale University, 300 George Street, Suite 501, New Haven, CT 06520-8009, United States<sup>b</sup> Connecticut VA Healthcare System, 950 Campbell Avenue, West Haven, CT 06516, United States<sup>c</sup> Department of Emergency Medicine, Yale School of Medicine, 300 George Street, Suite 501, New Haven, CT 06520-8009, United States

## ARTICLE INFO

## Article history:

Received 9 October 2012

Accepted 28 June 2013

Available online 8 July 2013

## Keywords:

Semi-supervised learning

Support vector machine

Graph Laplacian

Natural language processing

## ABSTRACT

**Objective:** To compare linear and Laplacian SVMs on a clinical text classification task; to evaluate the effect of unlabeled training data on Laplacian SVM performance.

**Background:** The development of machine-learning based clinical text classifiers requires the creation of labeled training data, obtained via manual review by clinicians. Due to the effort and expense involved in labeling data, training data sets in the clinical domain are of limited size. In contrast, electronic medical record (EMR) systems contain hundreds of thousands of unlabeled notes that are not used by supervised machine learning approaches. Semi-supervised learning algorithms use both labeled and unlabeled data to train classifiers, and can outperform their supervised counterparts.

**Methods:** We trained support vector machines (SVMs) and Laplacian SVMs on a training reference standard of 820 abdominal CT, MRI, and ultrasound reports labeled for the presence of potentially malignant liver lesions that require follow up (positive class prevalence 77%). The Laplacian SVM used 19,845 randomly sampled unlabeled notes in addition to the training reference standard. We evaluated SVMs and Laplacian SVMs on a test set of 520 labeled reports.

**Results:** The Laplacian SVM trained on labeled and unlabeled radiology reports significantly outperformed supervised SVMs (Macro-F1 0.773 vs. 0.741, Sensitivity 0.943 vs. 0.911, Positive Predictive value 0.877 vs. 0.883). Performance improved with the number of labeled and unlabeled notes used to train the Laplacian SVM (pearson's  $\rho = 0.529$  for correlation between number of unlabeled notes and macro-F1 score). These results suggest that practical semi-supervised methods such as the Laplacian SVM can leverage the large, unlabeled corpora that reside within EMRs to improve clinical text classification.

© 2013 Elsevier Inc. All rights reserved.

## 1. Introduction

The widespread adoption of electronic medical records (EMR) has led to the creation of large repositories of structured and unstructured clinical data. Leveraging this data has the potential to transform biomedical research and the delivery of healthcare. Automated text classification techniques extract structured information from narrative clinical text, empowering novel secondary uses of unstructured data [1,2]. Supervised machine learning based text classification approaches use a labeled training corpus for classifier development. Acquisition of suitably large training corpora for machine learning techniques may be prohibitively expensive, as this requires manual review of notes by trained clinicians. Because of the cost involved in assembling training corpora, they are typically limited in size (between one hundred and several

thousand instances). Furthermore, labeled training corpora represent a tiny fraction of the clinical text corpus: typical EMRs contain between tens of thousands and millions of unlabeled notes. Semi-supervised machine learning algorithms use both labeled and unlabeled data to build classifiers, and may outperform their supervised counterparts. Semi-supervised algorithms have been shown to outperform their supervised counterparts on general-language text classification tasks [3,4]. The goal of this study was to compare supervised and semi-supervised learning algorithms for the classification of clinical text.

The application motivating this study is a cancer case management system. Delays in cancer diagnosis and treatment can result from a failure to follow up abnormal radiological findings [5]. At the Veterans Affairs Connecticut Healthcare System (VACHS), we implemented case management processes and supporting informatics tools to help ensure the timely and appropriate diagnostic workup of patients with suspected cancer. We recently deployed a natural language processing (NLP) system that applies manually defined rules to diagnostic imaging reports to identify patients

\* Corresponding author. Fax: +1 (203) 737 5708.

E-mail address: [vijay.garla@yale.edu](mailto:vijay.garla@yale.edu) (V. Garla).

with potentially malignant lung or liver lesions for follow-up [6]. In this application of supervised and semi-supervised machine learning, we sought to train classifiers that improve upon rule-based methods for the identification of potentially malignant liver lesions from diagnostic imaging reports, and to develop a methodology for the training of machine-learning based classifiers to identify potentially malignant lesions in other organ systems. In this study we focus on the Laplacian SVM, a scalable semi-supervised learning algorithm that has been shown to outperform supervised SVMs and other semi-supervised algorithms on a variety of text classification tasks [3].

This paper is organized as follows: in the background section, we provide an overview of semi-supervised machine learning and an overview of the cancer case management application motivating this study. In the methods section, we describe the construction of our training corpus and our evaluation method. In the results and discussion section, we present the results of different algorithms, and discuss the relevance and practicality of various approaches for clinical text classification in general, and our application in particular.

## 2. Background

### 2.1. Semi-supervised learning algorithms

In supervised learning, an algorithm is given training instances  $(x_1, \dots, x_l)$  from an input space  $X$ , and corresponding targets  $(y_1, \dots, y_l)$  from  $Y$ , and learns a function  $f: X \rightarrow Y$ . In semi-supervised learning, in addition to the labeled training data, the algorithm is presented unlabeled instances  $(x_{l+1}, \dots, x_{l+u})$ , where  $u \gg l$ . To use unlabeled data, semi-supervised learning algorithms make assumptions on the distribution of  $X$ ; these can roughly be divided into the *Low-density separation* (LDS) paradigm, and the *Manifold* paradigm [7]. In the LDS paradigm, it is assumed that points that belong to the same cluster belong to the same class. LDS methods use unlabeled data to identify clusters (high density regions), and seek a classification boundary in a low-density region that avoids cutting clusters.

In this study, we focused on the manifold paradigm, which assumes that data from a high dimensional space lie on a low-dimensional manifold, and that the optimal classification function is ‘smooth’ with respect to the manifold. A manifold can be thought of as a surface embedded in a higher dimensional space; for example, the surface of the earth is approximately a 2-dimensional manifold embedded in a 3-dimensional space. Semi-supervised manifold techniques use unlabeled data to estimate the *geodesic* distance between points according to the *intrinsic* geometry of the manifold. This is in contrast to distance with respect to the *ambient* geometry – the high dimensional space in which the manifold is embedded. Returning to our example, the geodesic distance between London and Sydney corresponds to the path along the 2-d surface of the Earth; the ambient distance is based on the path in the 3-d space through the center of the Earth. Smooth functions along the manifold do not make ‘jumps’ between close points; i.e. nearby points (based on geodesic distance) are assigned the same class label.

Manifold techniques construct sparse graphs in which vertices represent instances from  $X$  and edges represent neighborhood relations. The resulting graph is a discretized approximation to the underlying manifold [8], and is used to formalize the ‘smoothness’ of a function on the manifold. Fig. 1 depicts (a) a 3 dimensional set of points along a 2 dimensional S-shaped manifold; (b) the superimposed graph; and (c) a projection of the data into a 2 dimensional space in which geodesic distances are reflected. Note that points A and D are close in the ambient space, but distant in the intrinsic space.

Let  $G = (V, E)$  be a graph with vertices  $V = \{v_1, \dots, v_n\}$ , where  $n = l + u$ .  $W = (w_{ij})$  is the adjacency matrix of  $G$ , where the weight  $w_{ij}$  represents the similarity between vertices  $v_i$  and  $v_j$ . If  $w_{ij} = 0$ , the vertices  $v_i$  and  $v_j$  are not connected by an edge. A common approach is to create edges between each instance and its  $k$  nearest neighbors, and assign the edges weights based on the Gaussian kernel that assigns instances a similarity based on an exponentially decaying function of distance:

$$k_{\text{gauss}} = e^{\frac{-\|x_i - x_j\|^2}{2\sigma^2}}$$

where the width  $\sigma$  controls the rate of decay as a function of distance.

A function  $f$  on a graph can be viewed as a vector  $(f_1, \dots, f_n)$  where  $f_i$  represents the value assigned to vertex  $v_i$ . A smooth function assigns adjacent vertices the same value, i.e.  $w_{ij}(f_i - f_j)^2 = 0$ . The smoothness of a function on a graph is defined as [10]:

$$S(f) = \sum_{i,j=1}^n w_{ij}(f_i - f_j)^2 = f^T L f$$

Lower values of  $S(f)$  correspond to smoother functions.  $L$  is the graph Laplacian associated with  $W$ , given by  $L = D - W$ , where  $D$  is the diagonal matrix with

$$d_{ii} = \sum_{j=1}^n w_{ij}$$

To quantify the smoothness of a function, the Laplacian is often normalized,  $L_{\text{norm}} = D^{-1/2} L D^{-1/2}$  and iterated to a degree  $p > 1$ , i.e.  $L^p$ .

### 2.2. Laplacian SVM

The support vector machine (SVM) is a supervised classification method that has been applied to a wide range of tasks, including text classification. SVMs project instances into a feature space via a *kernel*, and construct an optimally separating hyperplane in the feature space that discriminates between members of different classes. Popular kernels include the linear kernel, defined simply as the inner product between vectors, and the previously discussed Gaussian kernel. Formally, for a given set of labeled training instances and a kernel  $k$ , SVMs define a hyperplane as follows [3]:

$$f(x) = \sum_{i=1}^l a_i k(x_i, x)$$

$f(x)$  gives the distance of  $x$  from the hyperplane. Points on either side of the hyperplane are assigned the class labels  $-1$  and  $+1$  respectively:

$$g(x) = \text{sign}(f(x))$$

The Laplacian SVM builds upon the standard SVM framework, and constructs a kernel that defines similarity as a function of a mixture of geodesic and ambient distances. The resulting hyperplane is identical in formulation to that of the SVM – the difference lies in the kernel. The Laplacian SVM defines a kernel  $k'$  that uses unlabeled data to ‘deform’ the feature space to reflect the intrinsic geometry [3]:

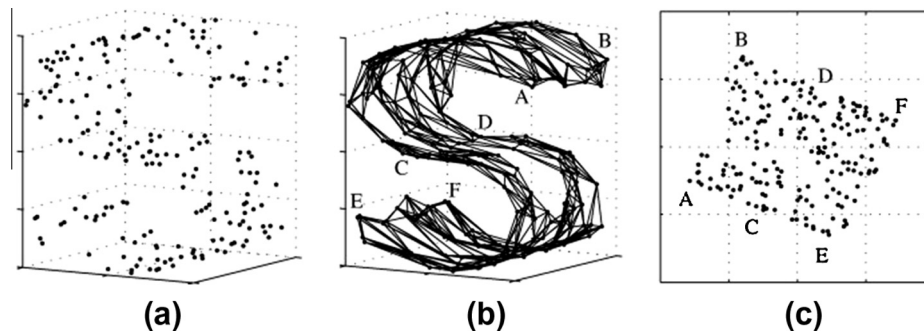
$$k'(x, z) = k(x, z) - k_x(I + MK)^{-1} M k_z$$

where  $M$  is based on the graph Laplacian:

$$M = \frac{\gamma_l}{\gamma_A} L_{\text{norm}}^p$$

and  $k_x$  is a vector of the kernel evaluations of  $x$  against all the training data (labeled and unlabeled).

$$k_x = (k(x_1, x) \dots k(x_n, x))^t$$



**Fig. 1.** (a) The 3-d S-curve, (b) the discretized underlying manifold structure, (c) the 2-d embedding of data reflecting geodesic distances (reproduced from Shao et al. [9]).

The modified kernel deforms the original kernel along a finite dimensional subspace given by the training data. The ratio of the tuning parameters  $\gamma_I$  and  $\gamma_A$  adjusts the strength of the manifold assumption [11].

Many semi-supervised techniques produce a *transductive* solution: a function that is only defined for the provided training data (labeled and unlabeled) [12,4]. One major advantage of the Laplacian SVM as opposed to other semi-supervised techniques is that it produces a classification function that is defined for the entire input space  $X$ .

### 2.3. Cancer case management

Early diagnosis and treatment of cancer significantly improves patient outcomes. Many early stage cancers are discovered incidentally through diagnostic imaging: at the VACHS, 52% of non-small cell lung cancers diagnosed between the years 2005–2010 were incidental findings on imaging obtained for other reasons such as workup of unrelated respiratory symptoms, chest pain, and others (personal communication). In order to prevent the failure to follow up abnormal findings, many radiology services have implemented additional measures beyond the accurate reporting of imaging results in the form of report coding [13]. Radiologists at the VACHS are required to select a diagnostic code when completing a report; these include codes that indicated a suspected malignancy, which we refer to as ‘cancer alerts’. A Cancer Care Coordinator helps ensure the timely and appropriate diagnostic workup of patients with suspected cancer. The majority of the patients managed by coordinators are identified by the ‘cancer alert’ coding of radiology reports. The coordinators refer appropriate cases to the institution’s interdisciplinary tumor boards, and help to ensure that patients who are undergoing tracking of suspicious lesions have serial imaging as recommended [14].

Our internal audits have shown that not all radiology reports of patients with suspected cancers are coded as ‘cancer alerts’, potentially delaying cancer diagnosis and treatment in these patients. For example, an imaging study performed to evaluate for pulmonary embolism, may also result in the incidental detection of an early stage lung cancer. The radiologist will likely assign the report an ‘abnormality’ code to indicate the presence of a clot but may fail to assign the report a secondary ‘cancer alert’ code. There is a risk that the incidental finding will not be addressed, especially when the referring clinician is not the patient’s primary care provider and/or the patient does not have a primary care provider. The lack of a ‘cancer alert’ code prevents cancer care coordinators from identifying appropriate cases. To prevent delays in cancer diagnosis and treatment due to miscoding of radiology reports, we developed a natural language processing (NLP) system to automate lung and liver ‘cancer alert’ coding in addition to the manual coding system already in use.

Approaches to clinical text classification in general, and the classification of radiology reports in particular, include rule-based

and machine learning based approaches [15,16]. The popularity of the rule-based approach stems in part from the scarcity of training data, and the cost involved in obtaining such data. Machine learning approaches to the classification of radiology reports require labor intensive tuning and/or large labeled training corpora [17,18].

Due to the scarcity of labeled training data ( $n = 100/50$  for lung/liver), we initially developed rule-based classifiers. The NLP system works as follows: every night the system retrieves all relevant radiology reports from our EMR; the system then applies to all reports an NLP pipeline that annotates syntactic structure (e.g. sections, sentences, phrases), and performs named entity recognition and negation detection. The system then applies manually developed rules to the radiology reports to automatically code cancer alerts; the system displays NLP-coded cancer alerts to cancer care coordinators, where they enter the same workflow as manually coded cancer alerts. We deployed the system for liver and lung in February and June 2011 respectively.

We provide cancer care coordinators a mechanism to mark as false positives reports erroneously coded as cancer alerts by the NLP system. By using cancer care coordinator feedback obtained as part of routine case management, we have acquired a constantly growing, labeled corpus that we can use to improve the performance of our cancer alert coding system. In this study, we sought to improve the performance of liver cancer alert classification through the use of machine learning algorithms. We also plan to expand this system for cancer alert classification for other organ systems, and therefore sought to develop a generalizable classifier development approach.

## 3. Methods

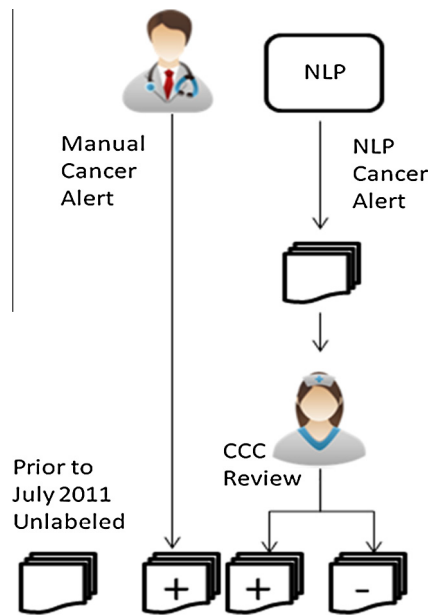
### 3.1. Training data

This study was approved by our local institutional review board and the requirement for patient informed consent was waived. We assembled a labeled reference standard by combining manually and NLP coded liver cancer alerts that included abdominal ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI) reports from July 1, 2011 to July 31, 2012 (Table 1 and Fig. 2). Radiologists assign one or more diagnostic codes to reports upon completion. In addition, all abdominal ultrasound, CT, and MRI reports are processed by the NLP system. Cancer care coordinators review reports automatically coded as ‘cancer alerts’ by the NLP system or manually coded by the radiologist. Cancer care coordinators can mark as false positives reports incorrectly coded as cancer alerts by the NLP system; we use the cancer care coordinators’ judgments to determine the true class label for NLP coded reports. The resulting labeled reference standard contained 1340 instances. We split this into a training reference standard ( $n = 819$ ) and testing reference standard ( $n = 521$ ). We used the training reference standard for parameter optimization via cross-

**Table 1**  
Reference standard composition.

		Coordinator/ radiologist classification	
		Y	N
Training reference standard	Rule-based system classification		
	Y	575 <sup>a</sup>	196 <sup>b</sup>
	N	48 <sup>c</sup>	
Testing reference standard	Rule-based system classification		
	Y	374 <sup>a</sup>	115 <sup>b</sup>
	N	32 <sup>c</sup>	

<sup>a</sup> Classified as a Cancer Alert by Rule Based System, Confirmed by Cancer Care Coordinator.  
<sup>b</sup> Classified as a Cancer Alert by Rule Based System, Marked as false positive by Cancer Care Coordinator.  
<sup>c</sup> Not classified as Cancer Alert by Rule Based System, Coded as a Cancer Alert by Radiologist.



**Fig. 2.** Corpus comprises unlabeled reports, reports manually coded by radiologists, and reports automatically coded by the NLP system and reviewed by CCCs.

validation, and used the testing reference standard for the final evaluation.

We retrieved all abdominal ultrasound, CT, and MRI reports from our EMR created prior to July 1, 2011 ( $n = 79,432$ ) and use these as unlabeled examples (even in cases where the true label is known). Due to the method by which we assembled the reference standard, the prevalence of positive cancer alerts was much higher than in the overall corpus. Based on a review of 100 randomly sampled abdominal and thoracic radiology reports from 1 week in February 2011, we estimate the rate of cancer alerts in the overall corpus at 49% (vs. 77% for our reference standard). The high prevalence of cancer alerts in our population is partially attributable to the cancer screening program: many of the positive cancer alerts are follow-up screenings for patients with a potential malignancy.

3.2. Feature representation

We built upon the preprocessing methods used in the current rule-based system to generate a bag-of-words representation for

radiology reports. We used the clinical Text Analysis and Knowledge Extraction System (cTAKES) version 2.5, and the Yale cTAKES extensions (YTEX) version 0.8 to process reports and annotate sentences, tokens, named entities, and their negation status [19,20]. We configured the YTEX named entity recognition module to map text to body region/location concepts from the Unified Medical Language System (UMLS), and to concepts from our custom vocabulary [21].

The rule based system first identifies all sentences that document findings pertaining to the liver using simple rules based on UMLS body location concepts. It then generates a binary feature vector from the liver sentences; features include mentions of terms suggestive of benign vs. malignant lesions. The system then applies conjunctive rules to the feature vectors to classify sentences. If any sentence is classified as asserting the presence of a malignant lesion, the entire report is classified as a cancer alert.

For our evaluation, we created a binary feature vector for each document that includes features from liver sentences. In addition to the features used by the rule based system, the feature vectors include an entry for each token or UMLS concept present in any liver sentence. The resulting feature vectors had over 15,000 dimensions.

3.3. Evaluation method

We evaluated SVMs and Laplacian SVMs. Purely supervised SVMs that do not use unlabeled data represent the ‘baseline’ to which we compare Laplacian SVMs that use unlabeled data. We (1) optimized all parameters (described below) via 5 runs of a 4-fold cross validation on the training reference standard; (2) trained classifiers on the entire training reference standard using the optimal parameters; and (3) evaluated classifiers on the testing reference standard. We reported Cohen’s Kappa, macro-averaged F1 score, sensitivity, positive predictive value (PPV), specificity, and negative predictive value (NPV) to quantify the agreement between the classifier predictions and the testing reference standard. The F1 score is the harmonic mean of sensitivity and PPV. The macro-averaged F1 score, or macro-F1, is the average of F1 scores across all classes (in this case, we have only 2 classes). The macro-F1 gives equal weight to all classes, ensuring that poor performance on the minority class is not masked by good performance on the majority class.

We used the LibSVM version 3.1 SVM implementation with linear and Gaussian kernels [22]. SVMs have a cost parameter  $c$ . In addition, the Gaussian kernel has a width parameter  $\sigma$ . We optimized  $c$  and  $\sigma$  via cross-validation on the training reference standard.

We used the LapSVM v0.1 Laplacian SVM implementation with the Gaussian kernel [23]. We did not evaluate LapSVM with the linear kernel, as parameter optimization with the linear kernel was too computationally expensive. The Laplacian SVM is controlled by kernel parameters, regularization parameters, parameters used to estimate the intrinsic geometry, and the amount of unlabeled training data. Optimizing all parameters would not be computationally feasible. Here we describe our parameter selection techniques.

3.3.1. Unlabeled data

Computing the graph Laplacian was very memory intensive, making it infeasible to use all the unlabeled data. We simply chose the largest subset of data for which we could compute the graph Laplacian; this included 25% of the unlabeled instances ( $n = 19,845$ ), in addition to all instances from the reference standard. We provided LapSVM only the labels for instances from the training reference standard: the labels for instances from the testing reference standard were not available to LapSVM.

### 3.3.2. Kernel parameters

We evaluated LapSVM with a Gaussian kernel. We used the optimal  $\sigma$  from the Gaussian LibSVM cross-validation.

### 3.3.3. Manifold estimation parameters

We used the LapSVM defaults: we set  $p = 1$ , we added edges for the 6 closest neighbors of each instance ( $k = 6$ ), and computed edge weights using the Gaussian kernel with width ( $\sigma$ ) corresponding to the average distance between neighboring instances.

### 3.3.4. SVM optimization technique

LapSVM implements *primal* and *dual* optimizers [23]. We used the primal optimizer, which computes an approximate solution to the SVM optimization problem and greatly reduces training times.

### 3.3.5. Regularization parameters

The regularization parameters  $\gamma_A$  and  $\gamma_I$  control the ‘smoothness’ of the separating hyperplane with respect to the ambient and intrinsic spaces. Increasing  $\gamma_I$  relative to  $\gamma_A$  forces a solution that is smoother with respect to the intrinsic geometry. These are the only parameters we optimized via cross-validation for LapSVM.

We performed all computations with Matlab 7.11.1 on a 64-bit Windows 2007 server with 36 GB ram and 2.27 GHz Intel Xeon processors.

## 3.4. Post-hoc experiments

To determine the effect of the amount of labeled training data on classifier performance, we trained SVMs and Laplacian SVMs on subsets of the labeled training data. We sampled between 0% and 90% of the labeled data, trained the SVM/Laplacian SVM, and evaluated the classifier on the testing reference standard. We performed this 100 times, generating an empirical distribution of macro-F1 scores that we use to assess the statistical significance of the difference in classifier performance between SVMs and Laplacian SVMs (2-sided *t*-test).

To determine the effect of the amount of unlabeled data on Laplacian SVM performance, and to rule out the possibility that performance was an artifact of our particular random selection of unlabeled data, we randomly sampled between 0% and 25% of the 79,432 unlabeled notes 15 times and evaluated the Laplacian SVM with these subsets. We quantify the effect of unlabeled data on classifier performance by computing the Pearson correlation between number of unlabeled instances used for training and the macro-F1 score on the test reference standard.

## 4. Results

Table 2 presents the performance of classification algorithms on the test reference standard. All algorithms outperformed the manually defined rules. The labeled data is enriched with instances on which the manual rule-based classifier fails, accounting for its low scores.

The Laplacian SVM (LapSVM) achieved the highest performance. The supervised linear and gaussian SVMs do not take advantage of unlabeled data. In contrast, the Laplacian SVM used 19,485 unlabeled instances to define a classification boundary that is smooth with respect to the underlying manifold structure of the data distribution.

### 4.1. Effect of labeled data on performance

Fig. 3 presents the mean macro-F1 score for SVMs (linear kernel) and Laplacian SVMs using between 10% and 90% of the labeled training reference standard, and evaluated against the testing reference standard. For both algorithms, performance increases with the amount of training data. Laplacian SVMs significantly outperformed supervised linear SVMs across all subsets of training data ( $p$ -value  $< 0.0001$  for the difference in mean macro-F1 between LapSVM and SVM).

### 4.2. Effect of unlabeled data on Laplacian SVM performance

Fig. 4 presents the macro-F1 scores for Laplacian SVMs using between 5% and 25% of the 79,432 unlabeled instances, and evaluated on the testing reference standard. The macro-F1 score is significantly positively correlated with the number of unlabeled instances used for training (pearson's  $\rho = 0.529$ ,  $p$ -value  $< 0.0001$ ). Even using a small amount of unlabeled data yielded higher performance than the supervised SVM: by adding just 5% of the unlabeled data (~4000 notes) the Laplacian SVM achieved a mean macro-F1 of 0.756 (compared to 0.741 for the supervised SVM). The mean macro-F1 score for the Laplacian SVM trained with 25% of the 79,432 unlabeled instances was 0.763; this is lower than the score achieved for our evaluation (0.773, Table 1), but higher than that achieved by the linear SVM (0.741). Different random samples of the unlabeled instances result in different estimates of the intrinsic geometry, and hence in different separating hyperplanes for the Laplacian SVM.

Using too little unlabeled data resulted in poorer performance than linear SVMs. Using just 1% of the unlabeled data resulted in a mean macro-F1 of 0.714. Using no unlabeled data resulted in a macro-F1 of 0.649.

### 4.3. Resource utilization

Cross validation and evaluation of the Laplacian SVM with 25% of the unlabeled instances (total 21,147 instances) required 20 min. Intermediate calculations involving the construction of the graph Laplacian consumed the entire system memory (36 GB).

## 5. Discussion

The central assumption of manifold-based semi-supervised learning is that data lies on a low-dimensional manifold, and that the optimal classification function is smooth with respect to the manifold. It is not clear *a priori* if text in general, or clinical text in particular, satisfies this assumption. Our results suggest that for this dataset, the manifold assumption holds. We believe that

**Table 2**  
Performance of classifiers on test reference standard.

Algorithm	Macro-F1	Kappa	F1	Sensitivity	PPV	Specificity	NPV
Manually defined rules	0.418	-0.106	0.836	0.921	0.765	0	0
SVM linear kernel	0.741	0.483	0.894	0.911	0.877	0.548	0.636
SVM Gaussian kernel	0.626	0.287	0.892	0.980	0.819	0.235	0.771
LapSVM	0.773	0.548	0.912	0.943	0.883	0.557	0.736

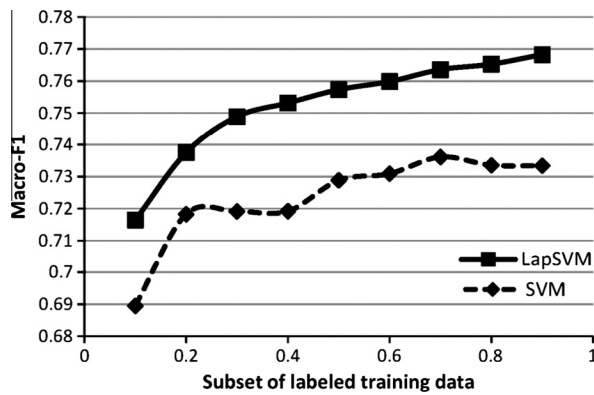


Fig. 3. Macro-F1 score as a function of the amount of labeled training data for SVMs and Laplacian SVMs.

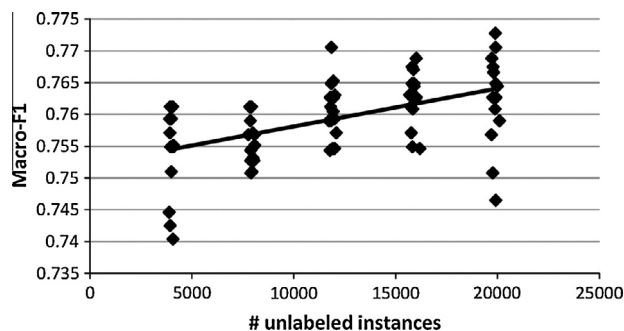


Fig. 4. Macro-F1 score as a function of the number of unlabeled data. Least squares regression line shown.

the manifold assumption may also hold for other clinical text classification tasks, and that semi-supervised learning algorithms that exploit the underlying data's manifold structure will outperform their purely supervised counterparts.

Evaluations of semi-supervised algorithms are often performed with an artificially low number of labeled training instances [8,3]. It is unclear from these evaluations if semi-supervised algorithms would actually outperform their supervised counterparts when evaluated with a larger number of labeled training instances. Some studies showed that the accuracy of supervised algorithms converges with that of semi-supervised algorithms as the number of labeled training examples increases [24,25]. In this study, we evaluated algorithms with the largest labeled training set that could be acquired with the resources available to us; the amount of labeled data likely reflects the limits of many practical clinical text classification applications. The number of labeled instances from our dataset is roughly equivalent in size to that of clinical text classification challenges [26,27]. In this study, we found that the Laplacian SVM's accuracy diverges from that of the supervised SVM: as labeled training data is added, the Laplacian SVM's accuracy increases at a greater rate than the supervised SVM (Fig. 3). These results suggest that, given labeled data of reasonable size for the clinical domain, the semi-supervised Laplacian SVM will outperform the supervised SVM.

Few studies have examined the impact of unlabeled data on semi-supervised classifier performance. For this dataset, Laplacian SVM performance was significantly positively correlated with the number of unlabeled instances used for training (Fig. 4). Using few (<1000) unlabeled instances did not improve performance relative to the supervised SVM. In contrast, using larger numbers of unlabeled instances (>4000) improved performance relative to

the supervised SVM. Unfortunately, the space complexity involved in the computation of the graph Laplacian and kernel matrix poses a limitation on the number of unlabeled instances that can be used. We hope that future advances in this technology will increase the capacity of Laplacian SVMs for learning with larger unlabeled datasets.

Laplacian SVMs build upon the well-established SVM framework and kernel methods. The classification function output by the Laplacian SVM is identical in formulation to that of the traditional SVM, thereby simplifying the application of Laplacian SVMs. Many semi-supervised learning algorithms can only classify the instances used for training. This is acceptable for batch applications, but not for online applications in which new instances must continually be classified. The Laplacian SVM defines a classification function valid for the entire input space, making it a practical solution for a wide range of applications.

One major limitation to the development of machine-learning based classifiers in the clinical domain is the cost involved in assembling a reference standard. By leveraging unlabeled notes, semi-supervised learning improves upon purely supervised techniques, thereby improving classifier performance with small reference standards. *Active learning* techniques, and techniques to learn from positive and unlabeled documents, could be used in conjunction with SVMs and Laplacian SVMs to reduce the number of labeled training instances needed [28–30]. Active learning iteratively constructs classifiers by selecting informative instances (e.g. radiology reports) for labeling by experts (e.g. radiologists or cancer care coordinators). In many general language and clinical text classification problems, a small number of labeled positive instances and a large number of unlabeled instances are available: for cancer alert classification, radiology reports with manually assigned cancer alert codes represent positive labeled instances, and reports with other codes represent unlabeled instances. Approaches for learning from positive and unlabeled data train a classifier through multiple iterations of first automatically selecting strongly negative documents, and then training an SVM classifier on the labeled positives and automatically selected negatives. The manifold approach may complement these methods by improving the identification of informative instances (for active learning) or strong negatives (for learning from positive and unlabeled instances). As part of future work, we will combine semi-supervised with active or positive and unlabeled learning for the initial development of classifiers for cancer alert classification in other organ systems.

The use case motivating this application is of high clinical relevance: the need for automated systems to ensure appropriate follow-up of incidental radiology findings will likely increase as widespread screening for lung cancer in high risk groups is implemented, and due to the increasing prevalence of hepatitis C [31,32].

One limitation of our study is that we only evaluated classifiers on radiology reports from one institution (VACHS). We recently deployed our system at the VA Maine and Ohio Healthcare Systems, and plan deployments at other institutions. As part of future work, we will evaluate machine-learning based classifiers on corpora from other institutions, and plan to develop classifiers for other note types, in particular, pathology notes.

Another limitation of our study is the evaluation of classifiers on a single feature representation: it may be possible that use of a different feature representation may narrow or eliminate the performance gap between linear and Laplacian SVMs. However, we engineered features to optimize the performance of the linear SVMs: in addition to the feature engineering approach used by the rule-based system (limiting to liver sentences, identifying informative concepts/words), we experimented with term frequency weighting and elimination of low frequency features (data not shown), and found that linear SVMs achieved optimal cross-

validation performance with the feature representation described here. Therefore, we believe it is unlikely that the performance improvement achieved by Laplacian SVMs is an artifact of the chosen feature representation.

Another limitation of our study is the imbalanced reference standard, in which positive cancer alerts are overrepresented. The system we developed as part of previous work allows cancer care coordinators to label documents as part of their standard workflow, thereby obviating the need for dedicated document labeling for classifier training. A constantly growing reference standard (from coordinator feedback), and a wider deployment will produce a larger set of negative examples. We believe that retraining classifiers on larger labeled corpora will rectify potential biases due to imbalanced training data.

As part of future work, we plan to refine the SVM models developed as part of this study and deploy them in the production system to automate cancer alert coding. We also plan to implement an automated system to retrain classifiers based on coordinator feedback, thereby enabling classifiers to ‘learn’ from their mistakes.

## 6. Conclusion

Semi-supervised learning algorithms can leverage the unlabeled corpora stored in modern EMRs to improve the classification of clinical text. By leveraging unlabeled radiology reports from our EMR system, the semi-supervised Laplacian SVM significantly outperformed the supervised SVM on the classification of abdominal ultrasound, CT, and MRI reports that document the presence of potentially malignant liver lesions. Semi-supervised learning algorithms may improve classifier performance on other clinical text classification tasks.

## Acknowledgments

We would like to thank all those involved in the VACHS cancer case management program, especially Laura Hunnibell, Anne DeLorenzo, Rosa Mirta, Woody Levin, Steve Steinhardt, and Dr. Tamar Taddei.

**Funding:** This work was supported in part by NIH Grant T15 LM07056 from the National Library of Medicine, CTSA Grant Number UL1 RR024139 from the NIH National Center for Advancing Translational Sciences (NCATS), and VA Grant HIR 08-374 HSR&D: Consortium for Health Informatics.

## References

- [1] Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009;42:760–72.
- [2] Kohane IS. Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet* 2011;12:417–28.
- [3] Sindhwani V, Niyogi P. Beyond the point cloud: from transductive to semi-supervised learning. In: *In ICML*; 2005. p. 824–31.
- [4] Chapelle O, Zien A. Semi-supervised classification by low density separation; 2005.
- [5] Singh H, Hirani K, Kadiyala H, et al. Characteristics and predictors of missed opportunities in lung cancer diagnosis: an electronic health record-based study. *J Clin Oncol* 2010;28:3307–15.
- [6] Garla V, Steinhardt S, Levin F, et al. A natural language processing-based clinical decision support tool improves the management of pulmonary nodules and liver masses. In: *Radiological society of North America annual meeting, Chicago*; 2011. <[http://rsna2011.rsna.org/search/event\\_display.cfm?am\\_id=2&em\\_id=11013797&printmode=Y&autoprnt=N](http://rsna2011.rsna.org/search/event_display.cfm?am_id=2&em_id=11013797&printmode=Y&autoprnt=N)>.
- [7] Chapelle O, Schölkopf B, Zien A, editors. *Semi-supervised learning*. Cambridge (MA): MIT Press; 2006. <<http://www.kyb.tuebingen.mpg.de/ssl-book>>.
- [8] Saul L, Weinberger K, Sha F, et al. Spectral methods for dimensionality reduction. In: Chapelle O, editor. *Semi-supervised learning*. Cambridge (MA): MIT Press; 2006.
- [9] Shao J-D, Rong G, Lee JM. Generalized orthogonal locality preserving projections for nonlinear fault detection and diagnosis. *Chemometr Intell Lab Syst* 2009;96:75–83.
- [10] Belkin M, Matveeva I, Niyogi P. Regularization and semi-supervised learning on large graphs. In: Shawe-Taylor J, Singer Y, editors. *COLT*. Springer; 2004. p. 624–38.
- [11] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *J Mach Learn Res* 2006;7:2399–434.
- [12] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In: *IN ICML*; 2003. p. 912–9.
- [13] American College of Radiology. ACR practice guideline for communication of diagnostic imaging findings. <[http://www.acr.org/SecondaryMainMenuCategories/quality\\_safety/guidelines/dx/comm\\_diag\\_rad.aspx](http://www.acr.org/SecondaryMainMenuCategories/quality_safety/guidelines/dx/comm_diag_rad.aspx)> [accessed 30.04.12].
- [14] Hunnibell LS, Rose MG, Connery DM, et al. Using nurse navigation to improve timeliness of lung cancer care at a veterans hospital. *Clin J Oncol Nurs* 2012;16:29–36.
- [15] Elkin PL, Froehling D, Wahner-Roedler D, et al. NLP-based identification of pneumonia cases from free-text radiological reports. In: *AMIA annu symp proc*; 2008. p. 172–6.
- [16] Mendonça EA, Haas J, Shagina L, et al. Extracting information on pneumonia in infants using natural language processing of radiology reports. *J Biomed Inform* 2005;38:314–21.
- [17] Xu Y, Tsujii J, Chang EI-C. Named entity recognition of follow-up and time information in 20,000 radiology reports. *J Am Med Inf Assoc*. <<http://dx.doi.org/10.1136/amiajnl-2012-000812>> [published online first 06.07.12].
- [18] Dreyer KJ, Kalra MK, Maher MM, et al. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. *Radiology* 2005;234:323–9.
- [19] Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.
- [20] Garla V, Re VL, Dorey-Stein Z, et al. The Yale cTAKES extensions for document classification: architecture and application. *J Am Med Inform Assoc* 2011;18:614–20.
- [21] National Library of Medicine. UMLS® reference manual – NCBI bookshelf. 2009. <<http://www.ncbi.nlm.nih.gov/books/NBK9676/>> [accessed 30.03.11].
- [22] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011;2. 27:1–27:27.
- [23] Melacci S, Belkin M. Laplacian support vector machines trained in the primal. *J Mach Learn Res* 2011;12:1149–84.
- [24] Joachims T. Transductive inference for text classification using support vector machines. In: *Proceedings of the sixteenth international conference on machine learning*. Morgan Kaufmann Publishers Inc.; 1999. p. 200–9.
- [25] Joachims T. Transductive learning via spectral graph partitioning. In: *In ICML*; 2003. p. 290–7.
- [26] Uzuner O. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009;16:561–70.
- [27] Pestian JP, Brew C, Matykiewicz P, et al. A shared task involving multi-label classification of clinical free text. In: *ACL*, editor. *Proceedings of ACL BioNLP*, Prague; 2007.
- [28] Chen Y, Mani S, Xu H. Applying active learning to assertion classification of concepts in clinical text. *J Biomed Inform* 2012;45:265–72.
- [29] Settles B. *Active learning literature survey*. University of Wisconsin–Madison; 2009.
- [30] Yu H, Han J, Chang KC-C. *PEBL*. ACM Press; 2002. p. 239.
- [31] National Comprehensive Cancer Network. The NCCN clinical practice guidelines in oncology. Lung Cancer Screening, Version 1. <[http://www.nccn.org/professionals/physician\\_gls/pdf/lung\\_screening.pdf](http://www.nccn.org/professionals/physician_gls/pdf/lung_screening.pdf)> [accessed 27.02.12].
- [32] Kanwal F, Hoang T, Kramer JR, et al. Increasing prevalence of HCC and cirrhosis in patients with chronic hepatitis C virus infection. *Gastroenterology* 2011;140(1182–1188):e1.