University of California, Los Angeles

Medical Imaging Informatics Group

BE M227 Final Project

Machine Learning Applications in Medical Informatics

Author:
Nicholas J. Matiasz

Instructor:
Prof. Alex Bui

Submitted:
2013-12-05

1 Introduction

The fields of medical informatics and machine learning overlap considerably; research in the former is often heavily informed by techniques of the latter. While machine learning is generally useful for analyzing various kinds of data, it can be particularly helpful in medical informatics due to constraints on the time of the field's domain experts (i.e., physicians), as well as the growing quantities of medical data that are produced from both research and the routine delivery of care. This paper discusses machine learning's specific utility for medical informatics, as well as challenges that often arise when applying machine learning to this domain; strategies for dealing with these challenges are also presented. Future developments in machine learning that would be a great boon to medical informatics are then discussed. First, a brief overview of basic and salient machine learning concepts is presented for context.

2 Overview of machine learning

Machine learning is a branch of artificial intelligence in which systems are designed to use data to automatically produce algorithms for a given task—e.g., to classify e-mails as either spam or not spam [2]. This broad definition invites a wide variety of applications, and there are multiple ways for this automatic production of algorithms, or machine *learning*, to occur.

In *supervised learning*, a training set of data is prepared in which each instance (e.g., an e-mail) is already labeled per the desired classification task (e.g., spam or not spam). An algorithm is then trained on this labeled data, with the intention that it will generalize

beyond the training data to label new, unlabeled examples accurately. This process requires outside knowledge: namely, the correct labels for the training set, which are often provided by a domain expert who can perform this classification reliably, albeit sometimes slowly. The extent to which the resulting classifier generalizes depends in part on the quality of the training examples. In *unsupervised learning*, this sort of outside knowledge is absent; instead, only raw data is provided to the algorithm, which is then expected to find patterns or structures in the data. *Semi-supervised learning*, as the name suggests, combines the previous two methods in order to accommodate both labeled and unlabeled training data. In *reinforcement learning*, the goal is to arrive at an optimal *policy*, which denotes a set of decision-making rules that an agent or system enacts to achieve a goal. Reinforcement learning is used to teach computers the optimal strategies for playing chess; millions of games can be simulated on a computer, which gleans the optimal strategies [2].

2.1 Performance evaluation of machine learning algorithms

Designing an elegant machine learning algorithm is only half of the story; researchers are also charged with demonstrating how a novel algorithm performs under realistic conditions, and why it is better than existing methods. This second task is not always straightforward, as the methods used to evaluate an algorithm's performance depend largely on the particular domain in which the algorithm is applied; it is also constrained by the quality of labeled data available for testing. The literature is rich with numerous strategies for evaluating machine learning algorithms (e.g., [4]). We focus here on three metrics that are defined by the concepts of *true positive* (TP), *true negative* (TN), *false positive* (FP), and *false*

negative (FN). Figure 2.1 presents a *confusion matrix*, which illustrates the meaning of these four quantities. In the context of evaluating a machine learning algorithm, these values are simply counts of the number of times each type of classification occurs.

Given these quantities, we can define the *accuracy* of a classifier, which is simply the fraction of test instances that were labeled correctly (see Equation 1). This metric is not always sufficiently descriptive; in many cases, there are asymmetric costs of false positive and false negative occurrences. For example, the consequences of labeling an important e-mail as spam are likely more severe than letting a spam message into the inbox. For this reason, additional metrics, defined in Equations 2 and 3, are used to describe an algorithm's performance in greater detail. *Sensitivity* (or *true positive rate*) and *specificty* (or *true negative rate*) characterize an algorithm's ability to correctly identify positive and negative results, respectively. These metrics are especially relevant in medical tests that detect the presence or absence of disease.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

2.2 State-of-the-art machine learning algorithms

As exaplained by Wolpert's "no free lunch" theorem [14], machine learning algorithms cannot be evaluated in a vaccuum; their performance must always be characterized with respect to a specific application whose test data are clearly defined. This limitation makes it impossible

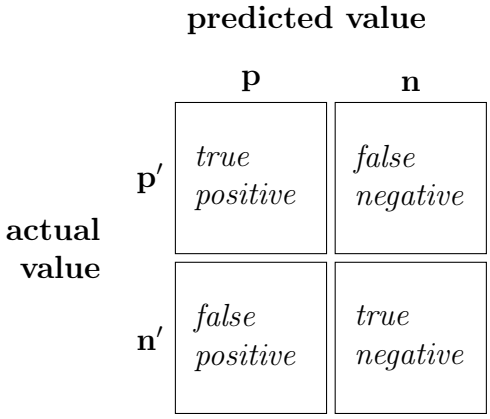

Figure 1: A confusion matrix defining the four outcomes of classification. In the previous example of e-mail/spam classification, labeling a spam e-mail as spam results in a *true positive*; labeling a non-spam e-mail as spam results in a *false positive*; labeling a spam e-mail as not spam results in a *false negative*; labeling a non-spam e-mail as not spam results in a *true negative*.

to provide an absolute ranking of existing machine learning techniques; a given algorithm that performs well in one context may perform terribly in another. Nonetheless, researchers have conducted empirical studies in order to compare the performance of different algorithms; some are consistently more powerful than others, including neural nets, support-vector machines (SVMs), and bagged trees [6]. In medical informatics, the choice of a machine learning strategy is often constrained by characteristics of both medical data and the clinical environment, some of which are discussed below.

3 Machine learning challenges in medical informatics

3.1 Class imbalance and asymmetric costs

Supervised learning algorithms perform best when the training data contains approximately equal quantities of the classes (e.g., a corpus of e-mail messages, half of which is spam, half of which is not). The performance degrades, however, if a "class imbalance" exists in the

training data [8, 9]. Depending on the domain, class imbalance can be either relatively minor (e.g., 10:1) or very severe (e.g., 10000:1) [9]. A common application of machine learning in medical informatics is in the classification of disease states. In such cases, training instances in which the disease is present are given the positive label, and instances in which the disease is absent are given the negative label. It is often the case, however, that there are far fewer people in a population with a certain disease than those without it. Therefore, data sets that are produced from research on such populations will yield far more negative instances than positive ones.

Training an algorithm with imbalanced data leads to poor sensitivity, because—given the overabundance of one of the classes—the training process tends to produce a classifier that is biased toward the majority class [8]. This in turn leads to a high occurrence of false negatives [13]. Note in Equation 2 that raising the value of FN will result in a larger denominator, and therefore a lower value for sensitivity. Although the above strategies can be effective, imbalanced data sets present further complications when the costs associated with false positives and false negatives are unequal.

The cost or penalty of making a false positive prediction is not always equal to the cost or penalty of making a false negative prediction. This fact further complicates training with imbalanced data sets, because additional considerations must be made regarding the biases that may result in the classifier. This problem is especially challenging when researchers are most interested in correctly classifying the minority class, and when the cost of low sensitivity toward the minority class is high [13].

3.1.1 Strategies for addressing class imbalance

Guo, et al. [8] categorize strategies for mitigating class imbalance into four groups, each of which corresponds to a different part of the machine learning process. The first (and perhaps simplest) method is to sample the training instances to achieve greater class balance. This can be achieved by either under-sampling the majority class, over-sampling the minority class, or some combination of the two. Despite the simplicity of this approach, it has been shown to be quite effective under certain conditions [10]. The second method is to perform feature selection, which is a strategy for eliminating irrelevant and redundant features from the training data. The third method is to bias the classier directly to counteract the bias that would occur otherwise due to class imbalance. The fourth and final method is to employ ensemble learning, which combines the output of multiple classifiers according to task-specific rules; examples of this strategy are boosting and bagging [8].

3.2 Scarcity of domain experts, abundance of data

The development of machine learning algorithms is often informed by experts from the domain in which the algorithms will be applied. When applying machine learning techniques to a specific field, researchers must often make considerable use of domain experts' knowledge to inform both the design of algorithms and the correct labels for training data. This holds true when machine learning is applied to medical informatics. A problem, though, is that the domain experts in medical informatics—physicians and other caregivers—are often extremely busy, and thus unable to participate in some of the labor-intensive, time-consuming tasks that can accompany scientific research. Therefore, the ability of machine learning algorithms

to automate various clinical and research tasks holds great promise for physicians who lack sufficient time to complete them. Additionally, when a physician's input is required to inform machine learning work—for example, when screening data sets for algorithm training—advanced machine learning techniques can help to minimize the amount of time the physician needs to dedicate to the training process [13].

One of the most illustrative examples of such a task is the systematic review, in which an exhaustive search of the clinical literature is performed in order to answer a specific medical question [7]. These reviews commonly require trained personnel to review thousands of abstracts to ensure that relevant articles are not excluded arbitrarily from the analysis [13]. Specific machine learning methods can leverage expertise in order to minimize the amount of time that domain experts need to be involved in a task. The following sections present various ways in which medical domain experts can augment machine learning work, as well as machine learning's positive effects on physicians' work.

3.2.1 Strategies for addressing the scarcity of domain experts

Dual supervision is a machine learning strategy in which "domain expert(s) provide explicit information regarding features and their relationship to class labels" [13]. Dual supervision can reduce the amount of work that experts must perform, because, in certain cases, providing additional knowledge about the features is more effective than traditional supervised learning algorithms. Rather than learning certain relationships over the course of numerous training instances, specific aspects of the model (e.g., classifier) are informed directly by an expert. To the extent that dual supervision lessens the requirement for training instances, fewer training data require manual labeling, thereby reducing the time commitment required

of experts during training. For example, a physician who has years of experience in the pathology of cancer can convey this knowledge to a classifier by annotating features in a data set in a machine-readable format.

Active learning is another strategy for reducing various costs associated with training. An assumption of active learning is that an algorithm can be trained more efficiently if the training data is selected strategically. This selection process usually occurs via *queries* that are submitted to an *oracle*, which is generally a human expert who can provide the correct label of a given training instance [12]. Queries can even result in the creation of new training instances that did not appear in the original data set [3]. A problem with active learning is that it generally assumes that its oracle will provide perfectly accurate labels for all queries, and that it will do so for a constant cost; this assumption is simply false in many cases [13].

Dually-supervised active learning combines the previous two approaches and makes use of labeled features (from dual supervision) in an active learning framework. Empirical tests have shown that this strategy is particularly useful for training on data sets with significant class imbalance [13]. Because of its efficient use of domain experts and its accommodation of class imbalance, dually-supervised active learning techniques show great promise for applications in medical informatics. As discussed below, they may also help physicians who interact with machine learning systems to feel more confident in classifiers' results.

4 Future directions

When machine learning is applied to medical informatics, two kinds of expertise are involved: that of the physicians who understand the given feature space, and that of the computer scientists who design the algorithms in light of the physicians' input. The success of this process depends on how well each group can present their knowledge in a way that is understandable by the other. Recent efforts have been made to close this loop more robustly—that is, to give domain experts the ability to interact with the data in a way that both imparts their expertise and directly impacts the resulting machine learning parameters that are used [5, 11]. Such advances are ingenious in that they drastically improve a physicians' ability to improve the training of algorithms directly, as well as providing an intuitive strategy for doing so. One can envision many clinical applications of these interactive technologies, including improved clinical decision support systems. If physicians are given greater and more direct control over the machine learning algorithms, they may have a better intuition for how their domain knowledge constrains the algorithm's design, and may therefore have greater confidence in the output of such systems.

5 Conclusion

It was reported in [1] that, according to a recent IDC Health Insights survey, almost 60 percent of physicians in ambulatory care settings are unsatisifed with EHRs because of the negative effects they have on clinical workflows. It is clear from such statistics such as this that physicians' clinical interactions with machines need to change considerably before there is unanimous agreement that computers are aids—and not obstacles—in healthcare.

Machine learning techniques hold great promise for expediting time-consuming research tasks and for automating overwhelmingly large data mining tasks—in short, for making physicians' lives easier.

One way to frame this relationship is that machine learning techniques afford physicians the ability to embed a part of their expertise in the computing systems that surround them, rather than always having to be a point of contact in the transmission of such information. Machines that have "learned" from a physician's experience can continue to benefit others who utilize the resulting model, and subsequent uses of these algorithms do not require repetitive input from the physician. Methods such as dually supervised active learning and interactive training visualizations continue to minimize the effort required of physicians to impart their knowledge to classifiers. In addition to developing increasingly robust training methods, machine learning researchers would do well to develop interfaces that give non-computer scientists a better intuition for the machine learning workflow, including methods for mapping the parameters of an algorithm to their clinical meaning or significance.

References

[1] Docs blame ehrs for lost productivity. `http://www.healthcareitnews.com/news/docs-blame-ehrs-lost-productivity`. Accessed: 2013-12-03.

[2] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2010.

[3] Dana Angluin. Learning regular sets from queries and counterexamples. *Information and computation*, 75(2):87–106, 1987.

[4] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[5] Eli T Brown, Jingjing Liu, Carla E Brodley, and Remco Chang. Dis-function: Learning distance functions interactively. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 83–92. IEEE, 2012.

[6] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[7] Carl Counsell. Formulating questions and locating primary studies for inclusion in systematic reviews. *Annals of Internal Medicine*, 127(5):380–387, 1997.

[8] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *Natural Computation, 2008. ICNC'08. Fourth International Conference on*, volume 4, pages 192–201. IEEE, 2008.

[9] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.

[10] Nathalie Japkowicz et al. Learning from imbalanced data sets: a comparison of various strategies. In *AAAI workshop on learning from imbalanced data sets*, volume 68, 2000.

[11] Jingjing Liu, Eli T Brown, and Remco Chang. Find distance function, hide model inference. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 289–290. IEEE, 2011.

[12] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[13] Byron C Wallace and Carla E Brodley (Adviser). *Machine learning in health informatics: making better use of domain experts*. PhD thesis, Tufts University, 2012.

[14] D. H. Wolpert. *The Mathematics of Generalization*. 1995.