Applying the Portfolio Learning Tool to Prevention Research

Patricia Mabry[1], Jennifer Villani[1]*, Jessica Wu[1], Jocelyn Lee[1], Ranell Myles[1], Fai Chan[2], B. Ian Hutchins[2], Carole Christian[2], George Santangelo[2], Calvin Johnson[3], William Lau[3], David M. Murray[1]

[1]NIH Office of Disease Prevention, [2]NIH Office of Portfolio Analysis, [3]NIH Center for Information Technology

*Corresponding author: villani@nih.gov

Goal

To present a case study using two different algorithms in the Portfolio Learning Tool (PLT) to categorize the NIH prevention research portfolio.

Background

One of the strategic priorities of the NIH Office of Disease Prevention (ODP) Strategic Plan for 2014-2018 is *to systematically monitor NIH investments in prevention research and assess the progress and results of that research*. In support of this priority, the long-term goal of this project is to develop a computerized tool capable of classifying NIH Type 1 R01 grants within each of eight categories of study characteristics. The ODP is collaborating with the NIH Office of Portfolio Analysis (OPA) to develop a machine-learning algorithm that can accurately identify prevention research grants. This poster presents early results from this collaboration.

The Portfolio Learning Tool

The PLT is a prototype application for classifying NIH research grants, contracts, intramural projects, and interagency agreements. The initial version of PLT was developed by Center for Information Technology (CIT) in response to requests by IC staff for a flexible text-mining tool to classify grant abstracts according to user-specified characteristics. CIT and OPA collaborated to more fully develop the PLT and OPA will soon make the beta version of this tool available to NIH staff. The PLT allows users to train a model of a research category of interest and then retrieve a list of all grants in IMPAC II that are deemed to be relevant to that category. Users supply the PLT with a training set of NIH grants consisting of positive and negative examples of the category of interest. Grants are represented in the PLT as indexed vectors of the scientific concepts extracted from the title, abstract, and specific aims sections of the application. A screenshot of the graphical user interface for the PLT is shown in Figure 1.

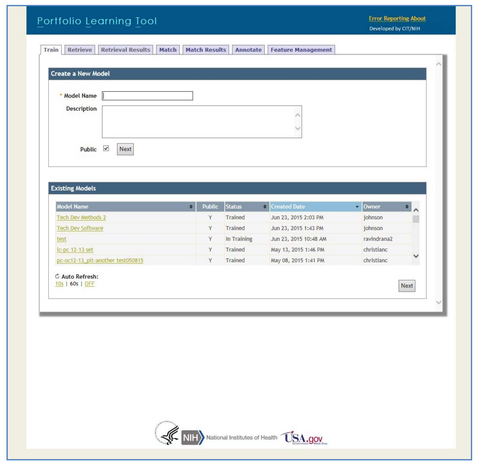


Figure 1. A screenshot of the PLT user interface.

The PLT provides a common implementation to the widely-used support vector machine (SVM) as well as an enhancement to the standard SVM. The common implementation is a library known as libSVM.[1] The enhancement is the Sampled, Augmented Ensemble SVM (SAE-SVM) that was designed by CIT to yield improved classification accuracy on the NIH research portfolio. Both learning methods analyze text from grant applications, recognize patterns, and classify grants according to pre-defined criteria.

When building a retrieval set, the PLT compares the "feature vector" representation of a candidate grant with the model representation and provides the user with a score representing the distance of that grant to the decision boundary. SAE-SVM is able to improve classification accuracy by automatically expanding the training set and by using an ensemble of classifiers to calculate relevance based on number of votes received from the individual classifiers in the ensemble. SAE-SVM generates a trained model from the training set, and the model is used to calculate classification threshold values which are used to classify abstracts from IMPAC II.

ODP Definition of Prevention

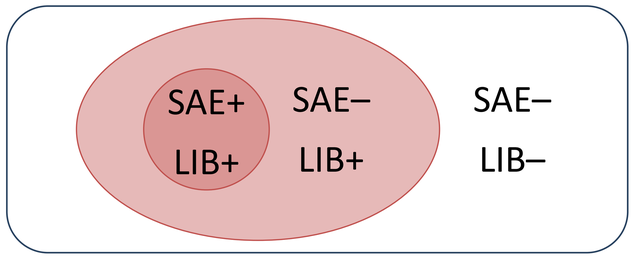
The ODP definition of prevention research includes the following categories of research, encompassing both primary and secondary prevention and excluding treatment:

- Preventing a new health condition, promoting health in the general population, or identifying risk factors for a new health condition
- Preventing progression of disease, preventing recurrence in those with a known health condition, or identifying risk factors for progression or recurrence
- Screening for a risk factor
- Screening for early disease
- Methods research, such as developing and/or evaluating different study design options, statistical analysis approaches, measurement approaches, survey instruments, new biomarkers, or screening methods.

Methods

The ODP used a novel team-based coding approach[2] to identify a set of 181 Type 1 R01 prevention exemplars (118 positive and 63 negative examples). These exemplars were used to train the PLT using both the libSVM and the SAE-SVM algorithms. These two algorithms established models that were then used by the PLT to classify Type 1 R01 grants. Each grant examined was assigned a score determined by how closely that abstract matched the model. A default score threshold was used to determine whether the abstract fell within the category of interest (prevention research). All grants above the threshold and 200 grants below the threshold were retrieved for possible validation. Over 10,000 Type 1 R01 grants from FY2007 to FY2014 were identified. To determine the sensitivity and specificity of each algorithm's ability to classify according to the ODP prevention definition, a random sample of the output was collated by OPA for validation. This set included grants identified as prevention positive or negative by each algorithm, and ODP was blinded to which algorithm was used and to whether the grants were identified as prevention positive or negative.

The sample consisted of 50 randomly sampled grants from among those deemed prevention-positive by both the libSVM and the SAE-SVM models, 50 randomly sampled grants from those deemed prevention-negative by both the libSVM and the SAE-SVM models, and 100 randomly sampled grants from those deemed prevention-positive by the libSVM but prevention-negative by the SAE-SVM model. Due to the fact that SAE-SVM is derived from the libSVM, by definition there are no grants that SAE-SVM would classify as prevention-positive that libSVM would classify as prevention-negative (see Figure 2).



Composition of Sample of Output Validated by the ODP

SAE + and LIB + : 50 randomly sampled abstracts

SAE – and LIB + : 100 randomly sampled abstracts

SAE – and LIB – : 50 randomly sampled abstracts

Figure 2. Comparing libSVM and SAE-SVM.

ODP subject matter experts (JW, JV) performed the validation and were blind to the results of the two algorithms to ensure an unbiased evaluation. Using the ODP Prevention Taxonomy and protocol for coding prevention research, two ODP subject matter experts independently read each grant abstract and judged each to be prevention-positive or prevention-negative. A third ODP subject matter expert (PM) reviewed all of the determinations and resolved any discrepancies through discussion. The classification status for the two algorithms was compared to the ODP validation results to calculate sensitivity and specificity for each model, with the ODP decision used as the gold-standard.

Results

Using libSVM, the PLT retrieved 8,996 Type 1 R01 grants that were classified as prevention-positive and 200 Type 1 R01 grants that were classified as prevention-negative. With SAE-SVM, the PLT retrieved 1,556 Type 1 R01 grants that were classified as prevention-positive and 200 Type 1 R01 grants that were classified as prevention-negative. These retrieval sets were based on the default threshold set by the models generated from the respective algorithms; the retrieval sets automatically included 200 grants that fell below the default threshold.

In the random sample, of the 150 grants classified as prevention-positive by libSVM, 95 grants were validated by ODP as prevention-positives (see Figure 3). Of the 50 grants classified as prevention-negative by libSVM, 44 grants were judged by ODP as prevention-negatives. These results show that the libSVM is able to accurately classify grants according to the ODP definition of prevention with a sensitivity of 94% and a specificity of 44%.

The random sample had 50 grants classified as prevention-positive by SAE-SVM (see Figure 3). Of those, 45 grants were validated by ODP as prevention-positives. Of the 150 grants classified as prevention-negative by SAE-SVM, 94 grants were judged by ODP as prevention-negatives. Therefore, SAE-SVM has a sensitivity of 45% and a specificity of 95%.

libSVM

	ODP validation +	ODP validation –
+	95	55
–	6	44

Sensitivity = 94%
Specificity = 44%

SAE-SVM

	ODP validation +	ODP validation –
+	45	5
–	56	94

Sensitivity = 45%
Specificity = 95%

Figure 3. Comparing the sensitivity and specificity of libSVM and SAE-SVM.

Conclusion and Future Directions

This poster summarizes our early efforts to develop a computerized tool to objectively and accurately deeply characterize NIH's investments in prevention research. The initial phase of our work, reported here, has focused on training PLT models to differentiate prevention research from non-prevention research, according to ODP's definition of prevention research. We reported our first sensitivity and specificity results of two machine learning algorithms: libSVM and SAE-SVM. The results indicate that libSVM is excellent at identifying prevention research (94% sensitivity over 200 abstracts), but does so at the expense of poor specificity (44%), indicating that libSVM is overly inclusive with respect to prevention research. Conversely, SAE-SVM has high specificity, rarely making a mistake when classifying a grant as prevention research (95% specificity over 200 grant abstracts), but also fails to include many grants that are indeed prevention-focused, reflecting poor sensitivity of this algorithm (45%).

These results indicate that neither algorithm is ready for routine use without further refinement. One way to achieve better results is to provide additional training abstracts, which is part of our future work, or to explore using the two algorithms sequentially. Moreover, for our next iteration of training and validation, we will use an updated version of PLT, which will optimize the threshold used by the algorithm for the specific set of grants it was trained on.

In addition to training the PLT to further classify grants according to prevention research subtype, we plan to train the PLT to classify grants within each of the remaining seven categories of interest (study rationale, independent and dependent variables, entities studies, study setting, population focus, and study design). Ultimately, this work will facilitate the identification of patterns and trends in NIH prevention research funding, as well as research areas that may benefit from targeted investments by the NIH Institutes and Centers.

References

[1]Chang C-C. & Lin C-J. *LIBSVM: A library for support vector machines*. ACM Transactions on Intelligent Systems and Technology, 2(3):27:1--27:27, 2011.

[2]Mabry P.L., Murray D., Villani J., Wu J., Lee J., Myles R., Carter-Nolan P., Yousif, L., Panzer R., Hamrick J., & Udo I. *Systematic Characterization of the NIH Prevention Research Portfolio.* NIH Portfolio Analysis Symposium, Bethesda, MD. July 21, 2015.