A Best-Fit Model for Concept Vectors in Biomedical Research Grants

Calvin Johnson, PhD[1], William Lau, MS[1], Archna Bhandari, MS[2], Timothy Hays, PhD[2]
[1]Center for Information Technology, [2]Office of Portfolio Analysis and Strategic Initiatives
National Institutes of Health, Bethesda, MD 20892

Abstract

The Research, Condition, and Disease Categorization (RCDC) project was created to standardize budget reporting by research topic. Text mining techniques have been implemented to classify NIH grant applications into proper research and disease categories. A best-fit model is shown to achieve classification performance rivaling that of concept vectors produced by human experts.

Introduction

On a yearly basis, the National Institutes of Health (NIH) reports to the Congress and the public on how much money is allocated to approximately 360 research and disease areas. Congress and the NIH Office of the Director use this data to better understand NIH research spending and priority areas. The RCDC project office has selected the Collexis® software as the platform with which to define disease categories and to classify research grants[1]. The classification is based on a technique known as *fingerprinting*, which extracts from the text a set of concepts using a domain-specific taxonomy[2].

Best-Fit Model for Disease Concept Vectors

The similarity between a grant and a disease is measured via a normalized dot product of the respective concept vectors. The process of creating disease fingerprints, i.e., concept vectors, is necessarily rather complex and involves a series of facilitated meetings with subject-matter experts, thesaurus experts, and others. The complexity of this process motives the potential need for a machine-learning system that can take advantage of previously coded grant data to generate a best-fit model for the disease fingerprint.

In this study, a training set of 1,252 grants and a smaller test set of 472 grants, covering nine disease categories were selected from the tagged set of 2005 NIH grant data. A perceptron neural network was generated from the training data for each of the nine diseases using the Matlab software (MathWorks Inc., Natick, MA). Table 1 demonstrates the performance of the supervised classification task, which compares equivalently with the performance of the expert-generated disease fingerprints in Table 2.

Disease Code	Recall	Precision	F-score
Liver Cancer	0.5844	0.8654	0.6977
Sickle Cell Disease	0.8070	1.0000	0.8932
Down Syndrome	0.5833	0.9333	0.7197
Kidney Disease Research	0.4314	0.6875	0.5301
Spina Bifida	0.5714	0.9231	0.7059
Anthrax	0.8043	0.9737	0.8810
Cervical Cancer	0.6780	0.9756	0.8000
Uterine Cancer	0.5938	0.9048	0.7170
Attention Deficit Disorder	0.8879	0.9450	0.9156
Average	0.6602	0.9120	0.7620

Table 1. Performance of the system-generated concept vectors on a test set of 472 grants.

Disease Code	Recall	Precision	F-score
Liver Cancer	0.5195	0.7407	0.6107
Sickle Cell Disease	0.9497	1.0000	0.9730
Down Syndrome	0.8333	0.8333	0.8333
Kidney Disease Research	0.4314	0.7586	0.5500
Spina Bifida	0.5714	0.8571	0.6857
Anthrax	0.8696	0.9302	0.8989
Cervical Cancer	0.6780	0.8333	0.7477
Uterine Cancer	0.6250	0.8696	0.7270
Attention Deficit Disorder	0.6983	0.9310	0.7980
Average	0.6860	0.9310	0.7980

Table 2. Performance of the official concept vectors on the same test set used in Table 1.

Conclusion

The results presented here demonstrate the feasibility of using neural networks to classify research grants into appropriate disease categories.

References

1. Brynko B, Collexis: The research fingerprint. Information Today. 2008;25:2:21-22.
2. Bhandari A and McGowan C. Using fingerprints to mine and search text, Horizons Perspectives in Healthcare Management and Information Technology, Fall 2005:37-40.