Classifying Scientific Performance on a Metric-by-Metric Basis

Eric Bell, Eric Marshall, Ryan Hull, Keith Fligg, Antonio Sanfilippo, Don Daly, and Dave Engel

Pacific Northwest National Laboratory

902 Battelle Boulevard

Richland, WA 99354

{eric.bell, eric.marshall, ryan.hull, keith.fligg, Antonio.sanfilippo, don.daly, dave.engel}@pnnl.gov

Abstract

In this paper, we outline a system for evaluating the performance of scientific research across a number of outcome metrics (e.g. publications, sales, new hires). Our system is designed to classify research performance into a number of metrics, evaluate each metric's performance using only data on other metrics, and to cast predictions of future performance by metric. This study shows how data mining techniques can be used to provide a predictive analytic approach to the management of resources for scientific research.

Introduction

National laboratories and other research institutions track detailed information about resources (e.g. funding) and outcomes (e.g. publications) of the research performed, in order to meet regulatory and accountability requirements by governing agencies and manage research activities more effectively. These data provide a unique resource to monitor and report on scientific progress so as to understand the scientific and societal impact of the research carried out. So far, limited use has been made of the information available, as no systematic practices have been established to infer models from these data capable of establishing performance benchmarks, and identifying and forecasting performance accomplishments. The goal of our study is to address this gap by establishing a statistical approach that makes it possible to transform the information gathered on research resources and performance metrics into training datasets that can be used to infer models of research impact.

Rather than postulating an arbitrary overall impact measure of research, we measure achieved performance for each outcome metric on which data are recorded: intellectual property, publications, staff hires, follow-on sales, and collaboration. Achieved performance for each metric is assessed through classification models that are learned from training datasets that encompass the information gathered on research resources and performance. The evaluation of the classification models developed demonstrates the viability and validity of the approach as the basis for a predictive analytic approach to science policy.

Data Mining for Business Intelligence

Business intelligence (BI) technologies have a critical need for core data mining techniques (Han, Kamber, and Pei 2011). BI systems are created to help transform operational data into valuable knowledge for business decision makers; this is a transformation that often involves significant data mining (Weidong, Weihui, and Kunlong 2010). Unlike much of the recent work in the intersection of these two fields, the work discussed in this paper is not concerned with text analysis, streaming data, or customer satisfaction (Godbole and Roy 2008; Park and Gates 2009). Instead, this work helps provide a means of automatically and continuously measuring Key Performance Indicator (KPI) performance (Kaplan and Norton 1992). In keeping with this orientation, the aim of this study is to mathematically define, quantify, and categorize the current state of research projects. We achieve this goal through evaluating performance through a series of metrics, consistent with the paradigm of metric-driven management (Koudas and Srivastava 2003). This is the first fundamental step towards process improvement. Analyze the current state, describe the future desired state, and implement the changes needed to achieve that future state; this matches the classic pattern seen in Business Process Reengineering (Hammer and Champy 2001).

We set out to develop a decision support system (DSS) that could enable decision-makers to make better use out of limited resources to maximize research objectives among various projects. To accomplish this, we built interfaces for existing enterprise information systems (EIS) to leverage this data in our DSS. In our case, the EIS was a relational database containing various data fields for past resource projects. The DSS would accept a query describing an ongoing research project, and the goal of the DSS was to forecast the future performance of the project. The work

described in this study provides the basis for predictive analytics in the business intelligence work carried out by the users.

Therefore, the data mining challenge of our project was to identify which data records were relevant to a particular query and use statistical analysis as well as machine learning techniques to satisfy the query.

System Description

Our system is a data driven end-to-end process that ingests metadata about research projects performed and communicates useful information about the projects to the user. First the metadata is aligned with five metrics, and clustered into a number of classes for each metric. Next each project is labeled across all of the metrics. Finally those labels are communicated back to the user.

Data Description

The dataset used in this experiment was collected from internally funded research projects at a national laboratory. Collectively, the data can be used to measure project success across a variety of metrics: collaboration, publications, new staff hires, intellectual property, and sales. The strategy pursued is to start with assessing project achievement for each separate metric. The aggregation of these diverse success metrics into a unified description of project impact is a separate effort.

Our dataset includes a total of 224 properties. Of these, there are boolean (186), currency (2), aggregate (28), continuous (1), categorical (5), and date (2) properties that collectively represent records of past performance of laboratory-directed research projects. Many projects receive funding from multiple funding sources, of which there are more than 150 possibilities. Since each funding client is represented by its own flag, we have a very large set of boolean properties.

Each of the five primary metrics identified earlier are determined based upon a predefined subset of these properties. For example, the publications metric is based upon the counts of seven different types of publications (abstracts, journal articles, reports, etc.), a count of all refereed publications, and a count of all published (refereed and non-refereed) documents. The collaboration metric utilizes seven properties, the staff metric six, and the intellectual property and sales metrics each use three.

Clustering Approach

The records are grouped using a two stage clustering process. The first step involves building clusters around records with identical profiles relative to a given metric. The second step transforms attributes of the remaining records to a common scale and then uses consensus clustering to choose the "best" candidate clustering strategy (Daly et al. 2011).

Stage 1: Identical Record Clustering

This stage begins with the removal of attributes that have constant values over the entire collection of records, as these variables are non-discriminating. In the data we received, some of the research projects had attribute values of "0" for all attributes within a given category. In addition to these "0" profile records, some of the other records have identical feature profiles. In this stage, those groups are extracted. The remaining variables are ranked via unbiased variable selection using modified random forests (Strobl et al. 2007, Strobl 2008).

Stage 2: Consensus Clustering

In the second stage of clustering, consensus clustering is used to cluster the remaining contracts. This begins by transforming all of the variables to a common scale. The transformations are dictated by the type and distribution of the variables (ex. binary, multiple category, or currency). Binary variables are transformed using the proportional distribution of zeros and ones and then taking the logit of each case. Multiple category variables are transformed using Gower's distance function for mixed variables (Gower 1971). For non-negative continuous variables, we take the square root of counts and the log. Currency variables are first standardized to a common year.

The remaining contracts are grouped using merged consensus clustering based on consensus matrices from partition around medioids, hierarchical clustering and k-means**.** (Simpson 2010) The selection of clustering method and number of distinct clusters is determined using the area under, and differences thereof, the consensus cumulative distribution curves. Cluster robustness is reported as the 5^{th}, 50^{th}, and 95^{th} quantiles of within-cluster consensus indices. The selection of clustering algorithm, and number of clusters, is made independently of classification performance.

Post-Clustering

Once the clusters have been determined, the cutoffs from the clusters determine the class labels for each metric of each record (e.g. low vs. high sales). These labels are part of the output from the system. In addition to communicating the success against each metric, the resultant cutoffs represent expected performance for ongoing and future research projects.

Evaluation

In order to test our system, the features relating to each metric were used to correctly bin each vector into the appropriate class. The cutoffs for those bins were

determined based on clustering all of the data for each metric according to the clustering approach previously explained. Those labels then served as the gold standard for evaluating classifier performance. The classes for each metric were as follows: Intellectual Property (Zero, Low, Medium, High), Staff Hires (Low, Medium, High), Publications (Zero, Low, Medium, High), Sales (Low, High), and Collaboration (Zero, Low, Medium, High).

Our dataset contained 894 records with 208 features per record.

Overall Performance

To test the performance of the system, five data files were created. Each data file corresponded to one of the five metrics. For the data file corresponding to any given metric, all features related to that metric were removed. The remaining features became the vector used for predicting the appropriate class label for the record. Note that this is in opposition to the clustering originally used to determine the bins; during clustering, all features were available for determining the correct class label. The precision, recall, and F-measure classification results of each individual metric are reported in Figure 1, along with the average across all metrics. All evaluations used 10-fold crossvalidation.

	Precision	Recall	F-measure
Intellectual Prop.	0.935	0.985	0.960
Staff Hires	0.982	0.982	0.982
Collaboration	0.977	0.952	0.964
Publications	0.820	0.850	0.832
Sales	0.852	0.872	0.857
Combined	0.911	0.924	0.915

Table 1 Evaluation of overall classifier performance on a metric by metric basis using Bayesian Networks (results for each metric reflect weighted averages across performance classes)

These results show that the performance in any one metric can be accurately predicted from information about the other metrics Additionally, these results and the experimentation done in assembling these results indicate that standard classifiers such as Naïve Bayes, Bayesian Networks, and Decision Trees can effectively be used to build models that accurately recognize the performance levels of project records for each metric. The classification models developed can help project and program managers evaluateongoing projects and test project improvement strategies by manipulating metric performance values on input project records.

Within Metric Performance

Within each metric, there was not an even division of test instances across the performance classes established (zero, low, medium high).For example, the distribution of instances for the Intellectual Property metric was: Zero = 887 instances, Low = 54 instances, Medium =7 instances, High = 1 instance) Classes with a larger number of test instances tended to form better models, and to perform better overall. Consequently, classification algorithms that have been shown to tolerate an uneven distribution of classes, i.e. Bayesian networks or similar (Daskalaki et al. 2006), are better suited for this task.

In an effort to give a representation of performance that takes into account the uneven distribution of test instances across performance classes, Figure 2 shows overall classifier performance on a metric-by-metric basis where results are not weighted across performance classes. From these results, it is evident that performance varies greatly across performance classes for all metrics, due the uneven distribution of test instances across performance classes. Performance classes with fewer test instances consistently underperform in terms of precision, recall, and F-measure across all five metrics. Those metrics where the instances are most evenly distributed across classes, namely staff hires and sales perform best in this evaluation.

	Precision	Recall	F-measure
Intellectual Prop.	0.392	0.477	0.403
Staff Hires	0.737	0.675	0.661
Collaboration	0.383	0.432	0.391
Publications	0.439	0.465	0.449
Sales	0.733	0.644	0.672
Combined	0.537	0.539	0.515

Table 2 Evaluation of overall classifier performance on a metric by metric basis using Bayesian Networks (results are not weighted across performance classes)

Value of Classification Models to Business

The models derived from the classification algorithms serve multiple purposes. They provide empirical evaluation of the clustering techniques, but they also provide an expectation for future and ongoing projects. The models allow for evaluation of projects against historical data. Looking at the results from the decision tree classifiers, expectations for success in each metric, as well as overall for the project can be communicated to project and program managers to achieve situational awareness of project performance, set goals and expectations of project outcomes, and plan and test improvement strategies.

The evaluation of project performance metric-by-metric helps provide a fairer assessment of project outcomes. Projects may differ markedly in as to their target outcomes. For example, a basic research project may strive to

generate more publications, an applied research project may have a stronger focus on patent applications, and a development/operational project may primarily target sales. Separate evaluation of each metric is therefore needed to assess each project in terms of its target outcomes.

Conclusion and Future Work

We have described a business intelligence application of data mining techniques aimed at managing resources and investments in scientific research. The methods and tools emerging from this work provide significant benefit to the business analysts, funding agencies, and principle investigators on projects. One distinguishing aspect of the approach described is the characterization of project performance metric-by-metric. Such a practice enables the assessment of a project in terms of the project's target outcomes, and thus it provides a fairer assessment of project performance.

The preliminary results presented show that clustering techniques can be profitably used to turn unstructured document collections into training datasets from which viable classification models can be learned. More specifically, mmultidimensional data can be discretized into data bins that represent metrics dimensions of interest (e.g. low, medium, high sales) that can thereafter be used to learn models of scientific research outcomes.

We are currently utilizing the same datasets to infer dynamic models of project performance (e.g. dynamic Bayesian networks) that make use of information about past project performance to assess the performance levels of a given current project and predict its future outcomes. We are also developing techniques to derive an overall measure of project performance from the integration of the five metrics discussed in this paper, which can be effectively tailored to a project target outcomes. These advancements will support the creation of a predictive analytic platform to aid decision making in the management of scientific research.

References

Daskalaki, S., Kopanas, I., and Avouris, N. 2006. *Evaluation of Classifiers for an Uneven Class Distribution Problem.* Applied Artificial Intelligence. Vol. 20, Iss. 5.

Daly, D, D Engel, E Bell and A Sanfilippo (2011) *Classifying Existing Research Contracts to Predict Future Contract Performance*. Unpublished manuscript, Pacific Northwest National Laboratory.

Godbole, S., and Roy, S., 2008. *Test Classification, Business Intelligence, and Interactivity: Automating C Sat Analysis for Services Industry.* Proceeding of the 14th AMC SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 911 919.

Gower, J. 1971. *A General Coefficient of Similarity and Some of its Properties*. Biometrics, 27, 623 637.

Hammer, M. and Champy, J. 2001. *Reengineering the Corporation*. Nicholas Brealey Publishing.

Han, J., Kamber, M., and Pei, J. 2011. *Data Mining: Concepts and Techniques*. Morgan Kaufmann

Koudas, N., and Srivastava, D., 2003. *Data Stream Query Processing: A Tutorial.* Proceedings VLDB Conference. Berlin, Germany.

Laplan, R., and Norton, D., 1992. *The Balanced Scorecard Measures that Drive Performance.* Harvard Business Review 70,1.

Park, Y., and Gates, S., 2009. *Towards Real Time Measurement of Customer Satisfaction Using Automatically Generated Call Transcripts*. Proceeding of the 18th ACM Conference on Information and Knowledge Management. ACM, 1387 1396.

Stobl, C., Boulesteix, A., Zeileis, A. and Hothorn, T. 2007. *Bias in Random Forest Variable Importance Measures: Illustrations, Sources, and a Solution*. BMC Bioinformatics. 8:1:25.

Strobl, C. 2008. *Statistical issues in Machine Learning Towards Reliable Split Selection and Variable Importance Measures*. Ludwig Maximilians University, Munich, Germany. Dissertation.

Weidong, Z., Weihui, D., and Kunlong, Y. 2010. *The Relationship of Business Intelligence and Knowledge Management*. The 2nd IEEE Internation Conference on Information Management and Engineering (ICIME), 26 29.