Designing a learning approach to community-based decision-making:
Using analytical statistics and decision trees to optimize a data structure

Kirk, R.A.
Human-Computer Interaction Program
Iowa State University
1620 Howe Hall
Ames, IA 50010

Danielson, J. A.
Office of Curricular and Student Assessment
Iowa State University College of Veterinary Medicine
2162 Vet. Med.
Ames, IA 50010

Abstract

This paper describes an effort to predict alumni success as measured by employer satisfaction in a veterinary medical education environment. Due to the complexity of the data, this paper sought to augment regression analysis with decision tree analysis to see if the combination of both approaches could result in new insights. Potential predictors of employer satisfaction were characterized as either technical or non-technical skills, and assigned to smaller conceptual item groupings within one of those areas. The internal consistency of each item grouping (concept) was then evaluated using Cronbach’s alpha. In the first approach, regression analysis was used to determine how well each concept predicted overall employer satisfaction with graduates. In the second approach, decision-tree analysis was used to discover the least uncertain hierarchy of this data, with overall employer satisfaction as the outcome. The process and the results from both approaches were examined and discussed along with recommended next steps.

Introduction

In modern communities, the organization of data is a complex and uncertain task. Often, it is difficult to determine how a particular series of events can be parsimoniously represented. Data is now so ubiquitous and disparate that meta-analysis has become difficult and ambiguous. This has become a problem in educational and research environments, where the sheer quantity of available data can complicate the research/assessment process (Fischman, 2011). Nonetheless, effective use of a variety of assessment data holds great promise for the effective design/modification of learning environments.

The goal of this paper is to reflect upon an ongoing, applied method for collecting, organizing and analyzing student performance and satisfaction data for the purpose of curricular improvement. A special focus is given to the organization of data into categories based both upon the characteristics of the data and based upon the ontological structure of the data. We hope to help foster a conversation that will uncover new methods to make sense of the increasingly complex and varied sources of assessment data that are available within our communities. At the Iowa State University College of Veterinary Medicine (ISUCVM), performance and/or satisfaction data from students, faculty, alumni and employers has been collected in order to assess the curriculum. Analysis of assessment data at the ISUCVM is used to guide curricular decision making by administrative and teaching faculty.

One important source of information regarding the success of the curriculum is the satisfaction of employers who hire ISUCVM graduates. In this study we sought to determine what graduate characteristics were most important in determining whether or not employers would be satisfied with their new graduates. This is an important question, because if we can determine what graduate characteristics matter most to employers, we can prepare our students with those characteristics in mind. One common approach to addressing this problem would be to use regression analysis, with student characteristics as predictors, and employer satisfaction as the dependent variable. While this is a beneficial process, we hoped to gain additional insight by also employing decision tree analysis.
Decision trees categorize data using information gain and have been used effectively within veterinary medical situations for making diagnostic decisions (Marsh, W.E., 1993; Ettinger, 2010; Radostits, Gay, Blood and Hinchcliff, 2000; Murray and Arguin, 2000; Shumaker, Corso, Rhyan, Philo, Salman and Gardner, 2010). Decision trees are also useful because they can organize an arbitrarily large amount of information. Decision trees create hierarchical inferences about a data structure through analyzing the uncertainty associated with each concept as it relates to an outcome, effectively reducing the entropy within the data structure (Russel and Norvig, 2008). The output from a decision tree analysis process is a hierarchical tree that organizes a subset of items into a structure useful for making decisions in relation to the outcome(s) used to create the structure. For these reasons, we suspected that decision tree analysis would be a powerful tool with which our stakeholders would also be familiar (Russel and Norvig; García-Almanza and Tsant, 2008).

Method

Since both the internal-consistency/regression and the decision analysis approaches require a well organized data structure, our first step was to organize the data into a structure that had both top-down and bottom-up validity. Top-down validity comes from a theory-driven, concept-based view of the data, while bottom-up validity comes from analyzing the data itself. First we created a hierarchical data structure for all of the assessment data used in the ISUCVM by creating theory-based concepts to which questions from any number of instruments could be mapped. To create the initial list of theory-based concepts, we gathered a list of concepts from a combination of three sources: inter-institutional standards, intra-institutional standards, and more detailed interpretations of data by local experts. In the field of Veterinary Medicine, the American Veterinary Medical Association keeps a list of core competencies that are useful for maintaining a standardized assessment framework across multiple institutions. These inter-institutional standards provided the basic framework for our data structure. As with other institutions, the ISUCVM also has its own organizational structure that naturally developed over time. Such natural structures can be considered a linguistic technology representing the natural ontology of stakeholders and can be incorporated into a data structure to add detail (intra-institutional standards). After applying those two sets of standards to our data structure, if there was still ambiguity regarding how data should be structured, we approached local experts in the field for organizational recommendations. It is important to note that a data structure is not strictly about the natural relationships that occur within our communities. This structure also helps to record the relationships between concepts and events in a way that allows each to have a level of agency upon analysis. In other words the concepts themselves begin to act as agents in the community, even after the students who produced the data are no longer part of it.

This study involved only a part of the data structure created using the process described above. Specifically, we used the results of employer surveys sent to employers of ISU graduates of the classes of 2007, 2008, and 2009. Each survey asked employers to rate their overall satisfaction with their new employees, their overall satisfaction with their new employees’ non-technical skills, and their overall satisfaction with their new employees’ technical skills. Employers also rated their employees’ performance in a number of knowledge or skill areas. We grouped knowledge and skill areas into categories based on their best theoretical fit. For instance, two survey items, “Perform the business related tasks of the position” and “Control expenses and maximize revenue” were categorized together into a “Business Skills” item grouping. The main categories included: Data collection, Data interpretation, Planning, Taking action, Interpersonal skills, Legal issues, Business skills, Making referrals and Problem solving. We then tested the internal consistency of each category using Cronbach’s alpha, and used simultaneous and hierarchical regression analysis to determine how well each item grouping predicted overall employer satisfaction. That full analysis is described elsewhere (Danielson, Wu, Preast and Fales-Williams, In-Press), and we will refer to those results when comparing them to the results of the decision tree analysis.

We used RapidMiner to conduct the decision tree analysis (Mierswa, Wurst, Klinkenberg, Sholz and Euler, 2006). For the regression analysis, we were limited in the number of predictors that we could include in the analysis by the number of subjects involved. Therefore, the regression analysis only used the 9 conceptual item groups (constructs) rather than each individual item. However, decision trees are not limited by the same constraints on numbers of participants. Therefore, the decision tree used overall employer satisfaction as the class label (similar, conceptually, to a dependent variable). As attributes (similar, conceptually to predictor variables) we used the same 9 employer item groupings (constructs) that were also used in the regression analysis. In addition to those 9 item groupings, we also included all individual sub-items that made up those groupings, as well as three additional items from the survey: employer satisfaction with Non-technical skills (overall), veterinary technical skills (overall), and veterinary knowledge (overall). The decision tree analysis classified the likely level of employer satisfaction with a student, given the other information we have about that student.
Results

The Cronbach’s Alpha analysis suggested that item groupings were internally consistent, with all item groupings having a Cronbach’s alpha > 0.7.

All of the veterinary skills analyzed were significantly correlated with the outcome of employer satisfaction. Similarly, all of the non-technical veterinary skills were also positively significantly correlated with employer satisfaction. The two aggregate items: Overall Veterinary Skills and Overall Non-Veterinary Skill, along with the technical and non-technical skills meta-groupings were then entered into a hierarchical regression analysis. The regression analysis revealed that Overall Veterinary Skills alone accounts for over 30% of the variance within employer satisfaction. Adding Overall Non-Veterinary Skills to this analysis explained an additional 21% of the variance. Using the two meta-groupings in a separate hierarchical analysis, the four technical skills explained 25% of the variance and the five non-technical skills together explained an additional 42% of the variance (Danielson, Wu, Kirk, Preast and Fales-Williams, In-Press).

Figure 1 shows the results of the decision tree analysis. In the decision tree analysis, the item Overall Veterinary Skills is referred to as ‘Meta-concept: Veterinary Knowledge’. The two items reporting overall employer satisfaction with ‘Technical Skills’ and ‘Non-technical skills’ are referred to as ‘Meta-concept: Veterinary Technical Skills’ and ‘Meta-concept: Veterinary Non-technical skills’ respectively. The item Overall [Employer] Satisfaction was used as the class variable (the outcome being predicted).

Discussion

The decision tree organized information similarly to the regression analysis approach. For instance, the decision tree automatically chose the meta-concept of Non-Technical Skills as being the single most informative attribute. However, in certain instances, the decision tree chose specific items for classification purposes rather than the item groupings from the statistical analysis. For instance, the single item ‘[Ability to] Control expenses and maximize revenue’ occurred on the second level of the decision tree as an intervening relationship between satisfied and very satisfied outcomes. What is likely happening in situations such as this is that the decision tree is purposefully removing redundant information from the tree. Whereas using the traditional approach, we created a “business skills” concept as an item-group, the decision tree approach saw the relationship between those items as a redundancy and simply reported the one that contributed most to explaining employer satisfaction: Controlling expenses and maximizing revenues. Other similarities include reliance upon non-technical skills to make distinctions between levels of employer satisfaction.
The decision tree analysis gives insight into the data within our community as it relates to employer satisfaction. The most informative features are found closer to the top of the tree. For instance, the item grouping of Non-Technical Skills does a good job at splitting the data into two subsections where one subsection is likely to have high employer satisfaction while the subsection performing poorly in the area of Non-technical skills is likely to have overall lower employer satisfaction. Overall, it appears as though items associated with activities that are traditionally non-technical in nature perform well as predictive indicators for employer success within the context of this analysis.

Conclusion

The data structure that was created effectively organized our complex data situation into a structure that could be used for the statistical and the machine learning analysis techniques. Using statistical methods, we created nine item grouping that helped to simplify the data structure while also quantifying the validity of the resultant structure through the evaluation of the internal consistency of these groupings. The use of decision tree analysis in conjunction with statistical analysis led to the creation of a minimal entropy classification structure for this data source in relationship to these newly created item groupings and the principle outcome of interest, employer satisfaction. Overall, the decision trees analysis has been a useful method to for analyzing large data sets because it analyzes data automatically whereas approaches such as the statistical analysis used here require predetermined inputs. Decision trees also naturally select out redundant information since decision trees use information gain over previous concepts as the criteria for adding new concepts to a model.

Based upon the preliminary reflections of this on-going case study, certain design guidelines can be offered to facilitate discovery from data within a community. These guidelines include the creation and use of a relational structure that allows for both aggregate and event-based comparisons, such as decision-tree analyses. It is believed that data in the format presented will allow valid cross-concept, cross-community comparisons. This belief is based upon the way that the concepts were created, grounded in practice and theory, and then separately validated by both analytical statistics and information gain approaches. Ongoing and future work will strive to assess the validity of this claim through analysis of case studies. It does take time, expertise and money to organize data as outlined above; however, we believe that there is potential value added both in terms of new insights and in terms of increases in curricular quality that may offset this initial investment of resources.

Acknowledgement

Thank you to: Wu, T.-F., Preast, V.A. and Fales-Williams, A. J. for contributing to the statistical analysis component of this research.
References

Fischman, J. (February 11, 2011). Dumped on by data: Scientists say a deluge is drowning research. The Chronicle of Higher Education.


