Ogata, H. et al. (Eds.) (2015). Proceedings of the 23rd International Conference on Computers in Education. China: Asia-Pacific Society for Computers in Education

Knowledge Discovery on the Data on Dissolution of Classes of the Ateneo de Davao University

Michelle BANAWAN[a*], Antonio BULAO II [b] , Jerry CANALE[c] & Jocel CATAMBACAN[d]
[a,b,c,d]*Ateneo de Davao University, Philippines*
*mpbanawan@addu.edu.ph

Abstract: Dissolution of classes is a constant dilemma of the Ateneo de Davao University. Although, data on dissolved classes are not directly available, large and distributed data sets on similar and related context are present like data on student registration, and academic classes, enrolment logs, class schedules and effective curriculum has been analyzed to discover patterns that lead to the understanding of class dissolution. Data since 2004 was gathered, processed and analyzed using supervised and unsupervised data mining techniques and methods. The data revealed non-linearity and the nonlinear regression model built and cross-validated gave an R of 0.9929. Running A Priori association also resulted to rules (confidence => 99% and support => 35%) that gave insights to the class dissolution problem. Even with the general tendencies of the data towards non-linearity and dynamism, some order and pattern were derived allowing some control and predictability (using the M5-based Pruned Tree Model). With the knowledge derived from the class dissolution data, key themes of chaos theory were derived.

Keywords: Knowledge discovery, association rules, nonlinear regression, chaos theory

1. Introduction

Pattern recognition and data mining are essential components of the entire process of knowledge discovery from data. Data mining is described as the process that involves the discovery of high-level knowledge from low-level data in large data sets (Fayyad, Piatetsky-Shapiro &Uthurusamy 1996). The Ateneo de Davao University (ADDU) has large datasets on academic classes offered, student registration and enrolment logs. From these data, valuable and high-level knowledge and insights on the phenomenon of class dissolution is derived. Class dissolution has been a prevalent problem of this educational institution that even careful planning and interventions still has remained unsolved and became an inevitable part of the system. The class dissolution phenomenon is suspected to be a chaotic organizational system. A chaotic system is defined to exhibit randomness but patterns may also exist. The combination of data mining techniques and chaos theory have been used as a framework in research studies in the fields of the physical sciences, psychology, education and other fields (Elshorbay, Simonovic & Panu 2002), (Hunt and Madhyashta 2005), and (Solomatine 2002),. Organizational systems have also been studied using key themes of chaos theory like in Levy (1994), McBride (2005), Mingers and White (2010), and Saat & Gleichauf (2009).

1.1. The Class Life Cycle Process Model of the University Enrolment System

In the University Enrolment System, a class goes thru a life cycle, i.e. class request (during projection phase), class creation (during pre-enlistment), class becomes full and viable (during enrolment), class dissolution (after enrolment).

1.2. Research Objectives

This study seeks to derive insights / discover knowledge from the various datasets of university systems to understand and address class dissolution using data mining for quantitative analysis and the lens of chaos theory for qualitative analysis.

2. Methods

2.1 Data Collection and pre-processing

The data sets from different university systems were gathered, cleaned and normalized to produce a single bid data flat file. Pre-processing work include : missing value replacement (using the missing value algorithm in Weka), continuous features were grouped together with the class feature to build a separate dataset for regression and support vector machine models, nominal features were grouped together with the class feature to build a separate dataset for the association rule (A priori) analysis, and the complete dataset consisting of all the features was used for the structured query language processing. Structured query language was used on the (complete) pre-processed data set to derive queries like the percentages of dissolved classes per academic unit and per school year were derived. Central tendencies and behavior of dissolved classes were also established and dependent factors also identified.

2.2 Feature Selection and Data analysis

One-variable statistical tests were used to understand the feature set and establish central tendencies and percentages. The Chi-square probability was used in feature selection to address feature dependencies. Both linear and non-linear regression models were built and compared in terms of performance. The M5 Pruned Tree algorithm was used to build the nonlinear tree model with the minimum number of instances allowed at a leaf node set to 4. Ten-fold cross validation was used to evaluate the performance of the models generated.

3. Results

3.1 Quantitative Analysis Results

SQL-based findings resulted to insightful data that were very useful to understanding the causes and possible prevention of dissolved classes. The nonlinearity of the data was evidenced by the M5P tree model having a correlation coefficient of 0.99. Running the SMO-svm algorithm (using the RBF kernel – which is known to be susceptible to overfitting) resulted to a kappa of 0 (even with 92.88% of correctly classified instances and only 7.12% incorrectly classified instances). Running the A Priori algorithm resulted to meaningful association rules derived (confidence <=99%), and it is observed that from the ten best rules, the common predicate is that : a class is not dissolved when the professor's name is made available as early as the pre-enlistment phase of the enrolment period.

3.2 Qualitative Analysis Results

Having observed the chaotic nature of the data, key themes from chaos theory were evident during the analysis, i.e. sensitivity to initial conditions (the dataset revealed sensitivity to the factors: student enrolment population and student retention data), presence of trajectories and attractors, structural invariance at different scales (the enrolment activity feature revealed that the domain of values are never predicted and have very large variances across time and data sets), irreversibility and bifurcation (given, now known, values of features bifurcation and irreversibility is manifest) , and the presence of strange attractors (features have been identified that result to allowing non-linear prediction of the outcome), negative and positive feedback.

4. Conclusions

This study has found that the data on the dissolution of classes can be better classified using the nonlinear regression model and is not linearly separable (nor also separable by a support vector machine). This study has resulted to decision-support knowledge and insights to help the University address its class dissolution problem. With the key themes of chaos theory, a better appreciation of the class dissolution phenomenon is also achieved, and it is now known that while the specific academic units make decisions on whether to dissolve or offer a class based on information only known to this units (local information), this decision will have unforeseen implication on the bigger (global) picture. Indeed the flap of a butterfly's wings can be instrumental in generating a tornado in another part of the world (Lorenz 2000).

Acknowledgements

We would like to thank all the Ateneo de Davao University thru its University Research Council for the research grant.

References

Elshorbagy, A., Simonovic, S. P., & Panu, U. S. (2002). Estimation of missing streamflow data using principles of chaos theory. *Journal of Hydrology*, *255*(1), 123-133.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery and data mining.

Hunt, E., & Madhyastha, T. (2005, July). Data mining patterns of thought. In*AAAI Workshop on Educational Data Mining* (pp. 31-39).

Levy, D. (1994). Chaos theory and strategy: Theory, application, and managerial implications. *Strategic management journal*, *15*(S2), 167-178.

Lorenz, E. (2000). 7. The Butterfly Effect. *The chaos avant-garde: Memories of the early days of chaos theory*, *39*, 91.

McBride, N. (2005). Chaos theory as a model for interpreting information systems in organizations. *Information Systems Journal*, *15*(3), 233-254.

Mingers, J., & White, L. (2010). A review of the recent contribution of systems thinking to operational research and management science. *European Journal of Operational Research*, *207*(3), 1147-1161.

Saat, J., Aier, S., & Gleichauf, B. (2009). Assessing the complexity of dynamics in enterprise architecture planning-lessons from chaos theory.*AMCIS 2009 Proceedings*, 808.

Solomatine, D. P. (2002, July). Data-driven modelling: paradigm, methods, experiences. In *Proc. 5th international conference on hydroinformatics* (pp. 1-5).