LOGICLUSTER: um método para agrupamento de dados web categóricos e contínuos usando Regressão Logística

Gustavo Pinheiro, Dilvan Moreira e Dorival Leão

Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, ICMC/USP – Avenida do Trabalhador São-Carlense, 400 - Centro - Cx. Postal 668 - São Carlos - São Paulo - Brasil CEP 13560-970.

gustavo@icmc.usp.br, dilvan@icmc.usp.br, leao@icmc.usp.br

Abstract. *In Distance Learning, teacher and students are physically apart. This characteristic causes a considerable loss in the perception the teacher has of the students learning process. Even worst, if a distance course is provided openly through the Web, the audience could be huge. Therefore, tools that can help a teacher get a better understanding of his students could be very useful. A Web site can be highly monitored. Every visit to a page, every mouse click in a hyperlink and many other online activities can be captured and stored for future analysis. The amount of stored data can reach enormous quantities, which becomes a challenge for data analysis. It is very difficult to extract useful information from this data. Data mining tools, such as clustering methods, could be very useful in dealing with such large amounts of data. But one of the problems with this new research area, Web Usage Mining (WUM), is to deal with categorical and continuous data simultaneously. In this work, a new method of clustering is presented, the LOGICLUSTER. It is based in the Logistic Regression Model and can handle categorical and continuous data simultaneously. The data can be purely categorical or mixed with continuous data.*

Resumo. *No Ensino à Distância via Internet, o professor não tem contato físico com seus alunos e, por isso, perde consideravelmente a percepção da interação destes em relação ao material didático. Além disso, pode-se atingir um número muito grande de usuários caso o meio de ensino seja a Web. Então, é importante dar ao professor ferramentas que o ajudem a conhecer seus alunos e a planejar sua atuação de forma a atender melhor um grande número de alunos. Sabe-se que a Web é um meio que pode ser ricamente instrumentado. Em princípio, cada clique num hyperlink, cada visita e outros dados de atividade online podem ser capturados e armazenados para futura análise. Entretanto, a quantidade de dados que se obtém pode ser imensa, tornando sua análise trabalhosa e demorada. Surge, então, o problema de analisar esses dados a fim de se extrair informações úteis. Pesquisas na área de Mineração de Dados fornecem ferramentas úteis para tratar este problema, sendo que métodos de agrupamento são particularmente interessantes. Uma das dificuldades encontradas nesta nova área, chamada de Web Usage Mining (WUM), é lidar simultaneamente com dados categóricos e contínuos. Neste trabalho desenvolveu-se um novo método de agrupamento, o LOGICLUSTER, baseado no Modelo de Regressão Logística, o qual é adequado para dados categóricos e contínuos, tanto em separado quanto em conjunto.*