Invited Session *Quality Measures in Data Mining*,
Conference ASMDA 07, Chania, Crète, Grèce, Mai 2007

Organisateurs
Philippe Lenca, GET - ENST Bretagne – Dép. LUSSI, CNRS UMR 2872 TAMCIC, France
Stéphane Lallich, Université de Lyon, Laboratoire ERIC, Lyon 2,

Interactive Exploration of Decision Tree Results

Khang N. Pham, IRISA, Rennes, France, pnguyenk@irisa.fr
Nghi T. Do, INRIA Futurs, L.R.I., University Paris-Sud Orsay, dtnghi@lri.fr
François Poulet, ESIEA Ouest, Laval, France, poulet@esiea-ouest.fr
Annie Morin, IRISA, Rennes, France, amorin@irisa.fr

Abstract. Our investigation aims at interactively exploring the decision tree results obtained by the machine-learning algorithms like C4.5. We propose an interactive graphical environment using the new radial tree layout, zoom/pan techniques and some existing visualization methods like explorer-like, hierarchical visualization, interactive techniques to represent large decision trees in a graphical mode more intuitive than the results in output of usual decision tree algorithms. The interactive exploration system on one hand can preserve the global view in a large representation of radial layout, zoom/pan techniques and on the other hand, it also provides a very good performance for an interesting sub-tree in the explorer-like view with simplicity, speed of task completion, ease of use and user understanding. The user can easily extract inductive rules and prune the tree in the post-processing stage. He has a better understanding of the obtained decision tree models. The numerical test results with real datasets show that the proposed methods have given an insight into decision tree results.
Keywords: Post-processing decision trees, Interactive exploration, Visual data mining.

Comparison of interestingness measures applied to textual taxonomies matching

Jérôme David, Polytechnic School of Nantes University, LINA FRE CNRS 2729, France, jerome.david@univ-nantes.fr
Fabrice Guillet, Polytechnic School of Nantes University, LINA FRE CNRS 2729
Régis Gras, Polytechnic School of Nantes University, LINA FRE CNRS 2729
Henri Briand, Polytechnic School of Nantes University, LINA FRE CNRS 2729

Abstract. This paper presents an experimental comparison of Interestingness Measures (IMs), in the context of an approach designed for matching textual taxonomies. This extensional and asymmetric approach makes use of association rule model for matching entities issued from two textual hierarchies. We select 6 IMs and we perform two experiments on a benchmark composed of two textual taxonomies and a set of reference matching relations between the concepts of the two structures. The first test concerns a comparison of matching accuracy with each of the selected measures. In the second experiment, we compare how each IM evaluates reference relations by studying their values distributions. Results show that the implication intensity delivers the best results.
Keywords: association rule, interestingness measures, textual taxonomy matching.

Random simulations of a datatable for efficiently mining reliable and non-redundant itemsets

Martine Cadot, UHP/Loria, Nancy, Martine.Cadot@loria.fr
Pascal Cuxac, INIST, Nancy, Pascal.Cuxac@inist.fr
Alain Lelu, INRA, Crebi, Jouy en Josas et UFC/Laseldi, Besançon, Alain.Lelu@jouy.inra.fr

Abstract. Our goal is twofold: 1) we want to mine the only statistically valid 2-itemsets out of a boolean datatable, 2) on this basis, we want to build the only higher-order non-redundant itemsets compared to their sub-itemsets. For the first task we have designed a randomization test (Tournebool) respectful of the structure of the data variables and independent from the specific distributions of the data. In our test set (959 texts and 8477 terms), this leads to a reduction from 126, 000 2-itemsets to 13, 000 significant ones, at the 99% confidence interval. For the second task, we have devised a hierarchical stepwise procedure (MIDOVA) for evaluating the residual amount of variation devoted to higher-order itemsets, yielding new possible positive or negative high-order relations. On our example, this leads to counts of 7,712 for 2-itemsets to 3 for 6-itemsets, and no higher-order ones, in a computationally efficient way.

Keywords: Text Mining, Randomization Tests, Significant Itemsets, Statistical Interaction, Multiple Comparison.

Construction of an off-centered entropy for supervised learning

Stéphane Lallich, Université de Lyon, Laboratoire ERIC, Lyon 2, stephane.lallich@univ-lyon2.fr
Philippe Lenca, GET - ENST Bretagne - Département LUSSI, CNRS UMR 2872 TAMCIC, France philippe.lenca@enst-bretagne.fr
Benoît Vaillant, UBS - IUT STID Vannes, Laboratoire VALORIA, France, benoit.vaillant@univ-ubs.fr

Abstract. In supervised learning, many measures are based on the concept of entropy. A major characteristic of the entropies is that they take their maximal value when the distribution of the modalities of the class variable is uniform. To deal with the case where the a priori frequencies of the class variable modalities are very imbalanced, we propose an off-centered entropy which takes its maximum value for a distribution fixed by the user. This distribution can be the a priori distribution of the class variable modalities or a distribution taking into account the costs of misclassification.

Keywords: supervised learning, entropy, imbalanced class.

A clustering approach for analysing association rules,

Jérôme Azé, LRI, Université Paris-Sud 11, Orsay, jerome.aze@lri.fr

Abstract. On of the difficult task when dealing with association rules is the analysis of the huge amount of rules. To help the user in this task, many measures of quality have been proposed in the past ten years. One of the difficulties is now to be able to select the good measure to use.

We propose, in this paper, a new way of analysing association rules. Rather than selecting a single measure of quality, we propose to use many measures to describe the set of rules to analyse and then, to apply a clustering algorithm on these rules in order to find coherent clusters. The user can then analyse the closest rule to the centroid of each cluster, instead of the whole set of rules.

Our approach is presented and tested it on five well know database of the UCI repository. As it is difficult to find a expert for each database, we choose to show that the clusters found by the clustering algorithm can reduce the error of classification made by decision trees.

Keywords: Clustering, Association rules, Measure of quality