**Notes in ‘Distribution Rules with Numeric Attributes of Interest’**

**Notes in Document**

**'LNAI 4213 - Distribution Rules with Numeric Attributes of Interest':**

Highlight : distribution rules, a kind of as- sociation rules with a distribution on the consequent. Distribution rules are related to quantitative association rules but can be seen as a more fundamental concept, useful for learning distributions *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.1)*

Highlight : An efficient algorithm for the generation of distribution rules is described. We also provide interest measures, visualization techniques and evaluation.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.1)*

Highlight : Machine learning has focused particularly on learning conditional probabilities of one target vari- able y (either numerical or categorical) with respect to a set of input variables X. However, the output of a learning algorithm is typically reduced to associating the most adequate value of y to each combination of values of the variables in X *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.1)*

Highlight : estimation. In this paper, we approach the problems of discovering and presenting important conditional distributions of a target variable with respect to a set of input variables. Our approach is based on association rule discovery [1].  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.1)*

Highlight : Association rules (AR) are highly legible chunks of knowledge that can be discovered from data. On top of that, the process for generating association rules is efficient enough to deal with very large databases, and the intended result is very well defined and free of heuristics *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.1)*

Highlight : devised mainly for descriptive purposes, AR can also be useful in classification [14], clustering [10], regression [17], recommendation and subgroup discovery [11]. Typically, algorithms for the discovery of AR deal with categoric attributes only. Srikant [19] proposed a specific approach for the discretization of numer- ical attributes bearing in mind the descriptive aim of AR *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.1)*

Highlight : , recommendation and subgroup discovery [11].  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.1)*

Highlight : Typically, algorithms for the discovery of AR deal with categoric attributes only. Srikant [19] proposed a specific approach for the discretization of numer- ical attributes bearing in mind the descriptive aim of AR *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.1)*

Highlight : To learn and discover distributions we propose distribution rules (DR). These associate a frequent itemset with an empirical distribution of a numeric attribute of interest without any loss of information. Distribution rules can be used in de- scriptive data mining tasks with the advantage of avoiding pre-discretization of the numeric variable of interest. *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.2)*

Highlight : We provide an efficient algorithm that dis- covers distribution rules and describe how to filter interesting rules, using the statistical distribution of Kolmogorov-Smirnov *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.2)*

Highlight : Distribution rules can be easily visualized as frequency polygons and viewed by a domain expert or data analyst. Besides,  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.2)*

Highlight : DRs can also potentially be used in a predictive setting, and are not fundamentally limited to numeric properties of interest.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.2)*

Highlight : A is a set of items as in a classical association rule, y is a property of interest (the target attribute), and Dy|A is an empirical distribution of y for the cases where A is observed. This attribute y can be numerical or categorical. Dy|A is a set of pairs yj/freq(yj ) where yj is one particular value of y occurring in the sample and freq(yj ) is the frequency of yj for the cases where A is observed. In this paper we will assume y is a numeric variable. Nevertheless, the concept of distribution rules is extended for categorical attributes as well. The attributes on the antecedent are either categorical or are discretized as in [8]. *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.2)*

Highlight : Although distribution rules can be output as text *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.3)*

Highlight : Since the consequent of one distribution rule is an empirical distribution, it can be represented as a frequency polygon *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.3)*

Highlight : The antecedent of each rule (e.g., the leftmost) is displayed as the main title. *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.3)*

Highlight : Some selected measures of the distribution and the name of the property of interest (P.O.I.: MPG) are shown within the plot *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.3)*

Highlight : The x axis is the domain of the P.O.I. and the y axis the estimated probability density. The polygon is drawn by binning the domain of the P.O.I. into a given number of intervals *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.3)*

Highlight : The distribution for the set of cases that satisfy the condition is shown in black, and the default distribution for the whole population is shown in grey *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.3)*

Highlight : In the case of association rules, objective interest measures typically try to assess how much the observed frequency of the conse- quent of the rule, under the conditions imposed by the antecedent, deviate from the frequency that would be expected assuming that antecedent and consequent *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.3)*

Highlight : were independent *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.4)*

Highlight : In the case of distribution rules, objective interest can be measured by as- sessing the difference between the distribution of the consequent and a reference distribution. *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.4)*

Highlight : The difference between two empirical distributions can be assessed through a statis- tical goodness of fit test, such as Kolmogorov-Smirnov [7] *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.4)*

Highlight : of the interest of a distribution rule, we can filter a set of DR’s by selecting the ones with KS-interest above a pre-defined threshold. *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.4)*

Highlight : Distribution rules can be used in descriptive pattern discovery tasks, although they can also be adopted in predictive tasks as well. One immediate advantage of their use in these situations is that it is not required to previously discretize the attribute y.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.4)*

Highlight : Distribution rules can be naturally applied to the data minig task of sub- group discovery [13] both for numeric and categorical properties of interest *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.5)*

Highlight : Distribution rules can also potentially be used in predictive tasks such as regression as in [17] or probability density estimation as in [6]. In this paper we have focused on the fundamental concepts and on the processes of generating, filtering and presenting the rules.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.6)*

Highlight : The algorithm CAREN-DR works by finding frequent itemsets and, simultane- ously, their associated p.o.i. distributions. *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.6)*

Highlight : The algorithm extracts sig- nificant rules by performing a Kolmogorov-Smirnov test between each new rule (Dy|a) and the a priori distribution (Dy|∅). *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.7)*

Highlight : In this section we show how our algorithm CAREN-DR performs on 4 different datasets described in Table 2. The algorithm has been run with different values of minimal support for a minimal KS-interest of 0.95 and with the improvement switch turned off. We can see that the algorithm scales up quite well with the number of examples and the value of minimal support *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.8)*

Highlight : These experiments show that the algorithm is capable of generating a very large number of distribution rules *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.8)*

Highlight : An Artificial Dataset  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.8)*

Highlight : In order to test the ability of the KS test to identify interesting rules, we have generated an artificial dataset with 1000 cases. The values of the attributes were chosen so that specific interesting distribution rules should appear *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.8)*

Highlight : Case Study *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.9)*

Highlight : We have applied distribution rules to the analysis of the main causes of delays in trip time duration for buses in a urban centre. This is a real dataset with about 8000 cases describing trips of a specific bus line. The dataset has 16 attributes plus the property of interest TripTime. *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.9)*

Highlight : We obtain 36 relevant rules *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.9)*

Highlight : We can see for example the difference in the distribution of the time a bus takes to make its route in March (Month=3) and in August (Month=8). Holydays also have a positive impact on trip time *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.9)*

Highlight : This type of rules are being used to attempt to reduce the costs with personnel, since unpredicted delays often force the bus company management to pay for extra labour time. This way, distribution rules can be used both to give managers indications about the most relevant causes of delay and also enable to predict the probability that TripTime will be higher than a certain threshold.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.10)*

Highlight : Distribution rules are mainly related to learning probability distributions [12], subgroup discovery [13] and quantitative association rules (QAR).  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.10)*

Highlight : Aumann and Lindell’s work on QAR uses a z-test to identify rules significance. As already pointed by Webb [20], z-test is inappropriate for small samples. The OPUS-IR authors propose the use of the standard t-test to decide on rules sig- nificance since the t-test tends to the z-test as the number of degrees of freedom increases. However, both z-test and t-test assume normality which in pratice cannot be guaranted. In this sense, using the KS approach is an advantage since no further distribution assumptions need to be considered.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.10)*

Highlight : We have introduced the concept of Distribution Rules as a generalization of asso- ciation rules *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.11)*

Highlight : We also describe how to visualize distribution rules. *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.11)*

Highlight : With classical association rules we would have to pre-discretize the numerical attribute of interest. With quantitative association rules, we would reduce the set of values in the consequent to a summary given by the mean or median. In the case of distribution rules, we keep the whole set of values of the property of interest and use these in graphical representations or post process- ing. Distribution rules can be presented as text or graphically and can be used in tasks of descriptive and predictive knowledge discovery.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.11)*

Highlight : 7. W. J. Conover. Practical Nonparametric Statistics - Third Edition. John Wiley & Sons, New York, 1999.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

Highlight : 8. U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In IJCAI, pages 1022–1029, 1993.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

Highlight : 10. E.-H. Han, G. Karypis, V. Kumar, and B. Mobasher. Clustering based on as- sociation rule hypergraphs. In Research Issues on Data Mining and Knowledge Discovery, 1997.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

Highlight : 11. B. Kavsek, N. Lavrac, and V. Jovanoski. Apriori-sd: Adapting association rule learning to subgroup discovery. In M. R. Berthold, editor, Advances in Intelligent Data Analysis V, volume 2810 of Lecture Notes in Computer Science, pages 230 – 241, Berlin Heidelberg, 2003. Springer-Verlag.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

Highlight : 12. M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie. On the learnability of discrete distributions. In STOC ’94: Proceedings of the twenty- sixth annual ACM symposium on Theory of computing, pages 273–282, New York, NY, USA, 1994. ACM Press.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

Highlight : 13. W. Kl˝osgen. Explora: A multipattern and multistrategy discovery assistant. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, CA, 1996.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

Highlight : 14. B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In KDD ’98: Proceedings of the fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 80–86, New York, NY, USA, 1998. ACM Press.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

Highlight : 16. C. J. Merz and P. Murphy. Uci repository of machine learning database. http://www.cs.uci.edu/∼mlearn, 1996. *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

Highlight : 17. A. Ozgur, P.-N. Tan, and V. Kumar. Rba: An integrated framework for regression based on association rules. In M. W. Berry, U. Dayal, C. Kamath, and D. B. Skil- licorn, editors, SDM ’04: Proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, 2004.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

Highlight : 18. A. Silberschatz and A. Tuzhilin. On subjective measure of interestingness in knowl- edge discovery. In KDD ’95: Proceedings of the First International Conference on Knowledge Discovery and Data Mining, pages 275–281. AAAi Press, 1995.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

Highlight : 19. R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In SIGMOD ’96: Proceedings of the 1996 ACM SIGMOD international conference on Management of data, pages 1–12, New York, NY, USA, 1996. ACM Press.  *(LNAI 4213 - Distribution Rules with Numeric Attributes of Interest, p.12)*

**Notes in Workspace:**