UNIVERSIDADE FEDERAL DE PERNAMBUCO

GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CENTRO DE INFORMÁTICA

2012.1

ISDA.R- INTERVAL SYMBOLIC DATA ANALYSIS FOR R-CRAN

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluno Ricardo Jorge de Almeida Queiroz Filho {rjaqf@cin.ufpe.br}

Orientadora Renata Maria Cardoso Rodrigues de Souza {rmcrs@cin.ufpe.br}

Março de 2012

Índice

1. CONTEXTO **3**

2. OBJETIVOS **4**

3. CRONOGRAMA **5**

4. BIBLIOGRAFIA INICIAL **6**

5. POSSÍVEIS AVALIADORES **7**

6. ASSINATURAS **7**

1. Contexto

Bases de dados estão presentes em praticamente todas grandes companhias e geralmente crescem atingindo um grande tamanho. Associado a esse crescimento das bases de dados surge à dificuldade de manusear e extrair informações dessas bases. É evidente que mesmo em situações que as metodologias tradicionais aparentam ser aplicável, muitas vezes o uso dessas técnicas não são de fato apropriadas.

Uma das maiores razões para as técnicas tradicionais falharem decorre do grande custo computacional necessário para fazer a análise dos dados em base muitos grandes. Mesmo com o poder de processamento crescendo esse problema seguirá ocorrendo, pois as bases de dados acompanharão esse crescimento. Em consequência disso, apesar de os métodos tradicionais servirem bem para bases de dados menores, é necessário aos analistas de dados se utilizarem, de procedimentos que funcionem melhor em bases maiores.

Uma abordagem para contornar esse problema é resumir essas bases de dados muito grandes de maneira que o resultado seja um conjunto de dados manuseável. Nesse sentido, podem ser utilizados dados simbólicos.

Dados simbólicos são denominados de "simbólicos" uma vez que eles não são puramente numéricos, eles podem ser representados por números, intervalos, uma distribuição, um conjunto de pesos, uma sequencia de valores com pesos como um histograma e etc. Esse tipo de dados possui uma estrutura interna que o permite armazenar mais informações que um dado clássico, inclusive os dados simbólicos são capazes de armazenar variações internas e incertezas das variáveis. Um exemplo é quando tratamos a idade de jogadores de futebol através de dados clássicos, teríamos então um valor numérico associado a cada jogador, caso adotemos dados simbólicos agrupando os jogadores por times encontraríamos, por exemplo, que em um determinado time as idades dos jogadores variem entre 19 e 29 anos (intervalo [19,29]).

2. Objetivos

O objetivo do presente trabalho é desenvolver um pacote na linguagem R para tratar de um dado simbólico especifico que é o intervalar. Este pacote que é chamado de ISDA.R permitirá tanto a transformação de variáveis clássicas em variáveis que armazene dados intervales quanto a analise desses dados . Dentre os métodos que estarão presentes nesse pacote estão: histInterval (histogramas para dados intervalares), intervalGraph3D (gráfico tridimensional para três coordenadas intervalares), meanInterval (média para dados intervalares), modeInterval(moda para dados intervalares), percentileInterval (percentil para dados intervalares) ,regInterval (regressão intervalar baseada em centroides), sdInterval (desvio padrão),summary.interval (função que quando chamada para um dado intervalar retorna a média, variância e desvio padrão), tableMulti (transformação de dados clássicos em intervalares) e varianceInterval (variância).

A linguagem R foi escolhida uma vez que ela é amplamente utilizada por analistas de todo mundo e por ser gratuita sendo distribuída sobre a licença GNU GPL. Por sua licença ser GNU ela garante uma série de liberdades ao usuário dentre elas dentre elas estão a: a de executar o programa para qualquer propósito, a de estudar o programa e a possibilidade de modificá-lo de modo a adequar a suas necessidades.

Além disso, o R é extensível e flexível, pois ao contrário da maioria dos softwares de analise de dados que se utilizam de menus "point-and-click" ou procedimentos "caixa-preta", o R é uma linguagem de programação cujo principal propósito é a analise de dados. Portanto, a criação do pacote ISDA.R permitirá que estudantes, cientistas e analistas que tenham interesse em analise de dados simbólicos possam desenvolver seus trabalhos mais rapidamente utilizando ou modificando o pacote de acordo com seus propósitos.

3. Cronograma

Durante a primeira fase do trabalho será realizada a revisão bibliográfica, seguida da implementação e utilização do pacote ISDA.R. A seguir, será concluído o relatório, e será feita a apresentação do trabalho de graduação.

ATIVIDADES	MARÇO			ABRIL			MAIO			JUNHO		
Revisão Bibliográfica												
Implementação												
Elaboração do Relatório												
Elaboração da Apresentação												

4. Bibliografia Inicial

[Diday, 2006] BILLARD, L., DIDAY, E. **Symbolic Data Analysis: Conceptual Statistics and Data Mining**, Wiley, West Sussex, England (2006).

[Diday, 2008] DIDAY, E. and NOIRHOMME-FRAITURE, M**.., Symbolic Data Analysis and the SODAS Software**, John Wesley Sons, ltd, (2008).

[Fagundes et al ,2011] FAGUNDES, R. A. A. ; SOUZA, R. M. C. R.;QUEIROZ FILHO, R.J.A; CYSNEIROS, F. J. A.. **A Robust Regression Method for Large Data Sets using a Symbolic Approach**. In: XII Escola de Modelos de Regressão, Fortaleza. XII Escola de Modelos de Regressão, (2011).

[Diday,2000] Block, H.-H., Diday, E., **Analysis of Symbolic Data** Exploratory Methods for Extracting Statistical Information from Complex Data, Springer-Verlag Berlin, Heidelberg, Germany (2000).

[Lima et al., 2008] E.A. Lima Neto,and F.A.T De Carvalho: **Centre and Range method for fitting a linear regression model to symbolic interval data. Computational Statistics and Data Analysis** 52 , 1500-1515. (2008)

[Fagundes et al., 2009] Fagundes, R.A.A., de Souza, R.M.C.R., Cysneiros, F.J.A. **A Robust Prediction Method for Interval Symbolic Data**, 9th International Conference on Intelligent Systems Design and Applications ISDA (2009).

5. Possíveis Avaliadores

1. Ricardo Bastos Prudêncio

6. Assinaturas

Renata Maria Cardoso Rodrigues de Souza

Orientadora

Ricardo Jorge de Almeida Queiroz

Aluno