DISCRETIZATION: AN ENABLING TECHNIQUE

PINTO, Patrick [1]
ARAÚJO, Wladimir [2]

Resumo:
O artigo intitulado **Discretization: an Enabling Technique** aborda uma revisão de métodos e padronização dos processos de discretização, os quais foram sumarizados em uma estrutura abstrata, de modo a prover uma referência para pesquisa e desenvolvimento futuros.

Introdução:

Os atributos numéricos de um conjunto de dados podem encontrar-se em um formato denominado discreto - quando possui um valor em um intervalo finito de valores; ou contínuo - quando possui um valor entre um conjunto infinito de valores.

Discretização é o processo de quantizar atributos contínuos. Quantizar significa atribuir valores discretos para uma variável cuja amplitude varia entre infinitos valores. Esse processo é necessário dado que isto pode impactar, por exemplo, em um modelo de aprendizagem de árvores/regras de classificação. Em um algoritmo de árvore de decisão, o valor contínuo pode levar a um desempenho ruim, dada a infinita quantidade de valores. Dentre as vantagens de se utilizar valores discretos estão: trabalhar com dados reduzidos e simplificados; sua representação a nível de conhecimento, mais fáceis de entender, utilizar e explicar; muitos algoritmos de classificação somente trabalham com dados discretos e, como reportado em um estudo (Dougherty et al., 1995), a discretização faz com que o aprendizado seja mais rápido e mais preciso, os resultados são geralmente mais compactos, mais curtos e precisos do que os contínuos, portanto, (os resultados) podem ser mais próximos examinados, comparados, usados e reutilizados.

O artigo aqui referenciado aborda uma revisão de métodos e padronização dos processos de discretização, os quais foram sumarizados em uma estrutura abstrata, de modo a prover uma referência futura para pesquisa e desenvolvimento.

[1] Acadêmico da Universidade do Estado do Amazonas (UEA). Email: ptp.cid@uea.edu.br
[2] Acadêmico da Universidade do Estado do Amazonas (UEA). Email: wbgan.cid@uea.edu.br

Metodologia:

Para a realização do trabalho, o primeiro passo foi uma análise dos trabalhos anteriores realizados sobre o processo de discretização. Com esta análise, levantou-se que há abordagens diferentes para que a discretização de uma base de dados seja feita, sendo elas:

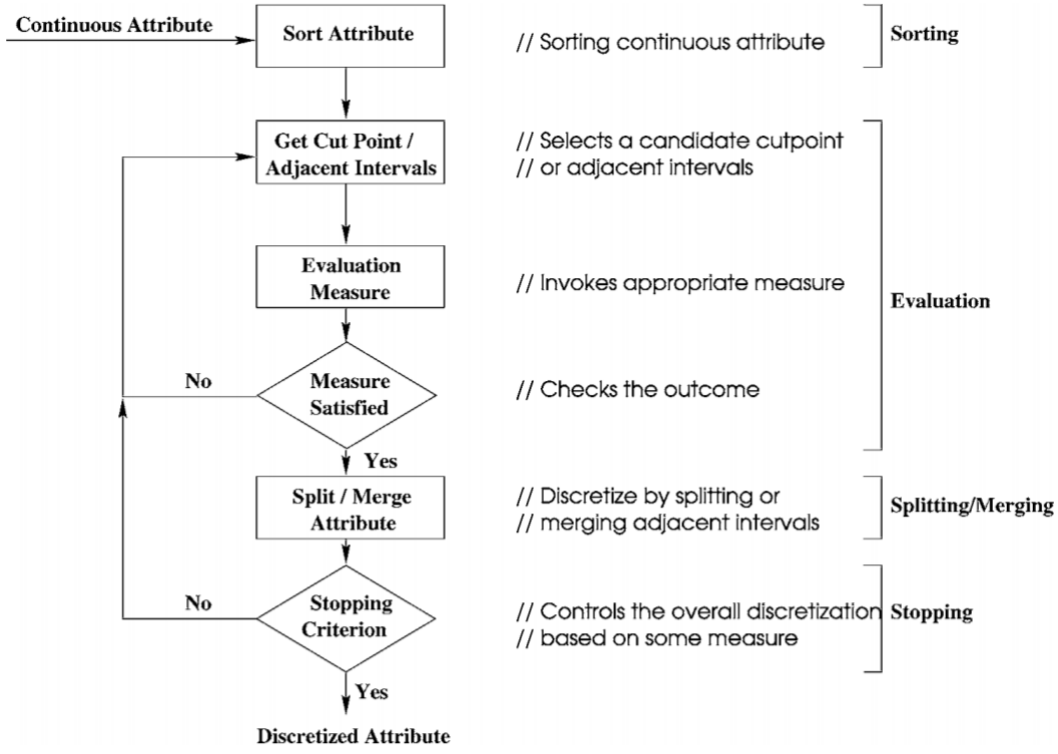
- Supervisionada ou Não-Supervisionada: Onde a discretização leva em conta a presença (Supervisionada) ou ausência (Não-Supervisionada) de informações de classe;
- Dinâmica ou Estática: Onde a discretização é feita durante o processo de classificação (Dinâmica) ou antes deste processo (Estática);
- Local ou Global: Onde a discretização leva em conta apenas um conjunto dos dados (Local) ou leva em conta os dados em sua totalidade (Global);
- Top-Down ou Bottom-Up: Onde a discretização inicia com uma lista vazia de valores de corte e vai acrescentando os mesmos através da "divisão" da base de dados em intervalos (Top-Down) ou inicia com uma lista de valores de corte equivalentes aos dados da base e vai removendo valores ao "fundir" os dados em intervalos (Bottom-Up) ;
- Direta ou Incremental: Onde a discretização ocorre dividindo os valores da base em k intervalos, sendo esse valor suprido pelo usuário (Direta); ou iniciando por uma discretização simples que passa por um processo de melhoria contínua, precisando de critérios adicionais para ter seu ponto de parada (Incremental).

Após a análise, o trabalho concentra-se na descrição de um processo de discretização simplificado. Tal processo é subdividido em quatro etapas:

- Ordenação: Etapa do processo em que os dados provenientes do banco são ordenados através de um algoritmo, seja em ordem crescente ou decrescente. Pela origem complexa do processo, geralmente é utilizado o algoritmo Quicksort nesta etapa, uma vez que seu custo em tempo é de O (n log n).
- Avaliação: Nesta etapa, busca-se o melhor ponto de corte que irá dividir os dados, ou par de intervalos para fusão dos mesmos. Para tanto, são consideradas uma série de funções de avaliação retiradas da literatura, podendo estas serem medidas estatísticas ou de entropia.

- Divisão/Fusão: Como explicado anteriormente, a dicotomia Top-Down / Bottom-Up é utilizada nesta etapa para dividir os dados com base em pontos de corte ou para fundir os dados em intervalos.
- Parada: Nesta etapa, é avaliado o ponto de parada para o processo de discretização, de modo a saber em que momento o processo deve ser encerrado. Para que este ponto seja encontrado, normalmente considera-se o trade-off entre baixa aridade com uma compreensão melhor dos dados, porém com menos acurácia, e alta aridade com uma compreensão pior dos dados, mas alta acurácia.

Estas etapas podem ser visualizadas no gráfico abaixo:



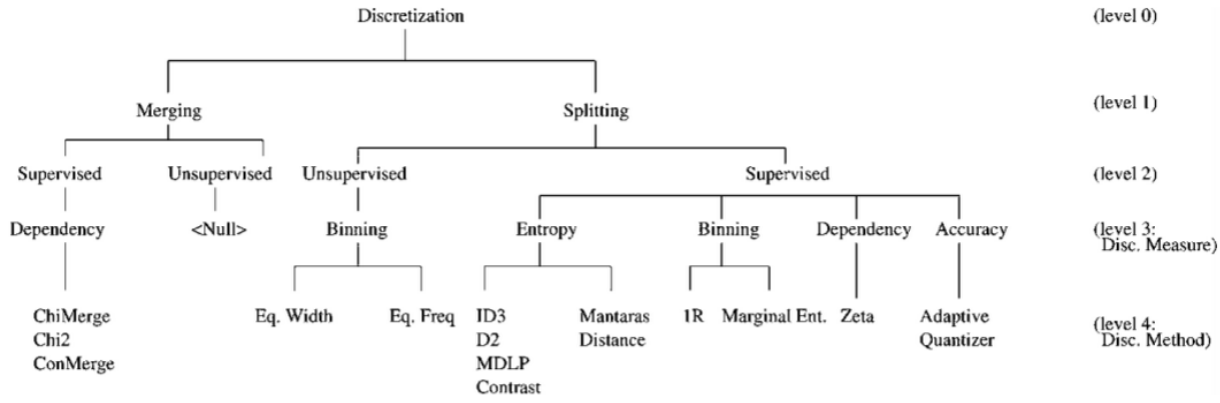
Por ser um processo que pode ter resultados muito variados, devido aos diferentes métodos que podem ser usados em cada etapa, julgar qual a melhor saída é um processo complexo. De modo a facilitar essa decisão, há três dimensões que devem ser consideradas na análise:

- O número total de intervalos, uma vez que quanto menor o número de pontos de corte melhor é a discretização;
- O número de inconsistências causadas pela discretização;

- A acurácia preditiva do processo.

Desta forma, estes três pontos podem ser resumidos em: Simplicidade, Consistência e Acurácia.

De modo a facilitar a escolha de métodos com base na avaliação do resultado dos mesmos, este trabalho produz em seu passo seguinte um arcabouço para o processo de discretização. O arcabouço é construído de forma hierárquica, sendo sistemático e expansível, além de ter como objetivo cobrir todos os métodos existentes a serem aplicados nas etapas do processo de discretização. Sua estrutura pode ser visualizada no diagrama abaixo:



Como pode ser observado, o arcabouço é dividido em níveis, que são os responsáveis por organizar a hierarquia dos métodos de discretização. No nível 1, os métodos são divididos entre os que utilizam Fusão (Bottom-Up) e os que utilizam Divisão (Top-Down) ao lidar com os pontos de corte. No nível 2, os métodos são divididos entre Supervisionados e Não-Supervisionados. No nível 3, os métodos são agrupados pelas medidas de discretização que cada um usa. Por fim, temos os métodos propriamente ditos no nível 4. Com isso, temos a seguinte classificação:

- Fusão:
    - Supervisionado:
        - Dependencia: ChiMerge, Chi2, ConMerge
- Divisão:
    - Não-Supervisionado:
        - Binning (Classificação): Equal Width, Equal Frequency;

- Supervisionado:
    - Entropia: ID3, D2, MDLP, Contrast, Mantaras, Distance;
    - Binning (Classificação): 1R, Marginal Entropy
    - Dependência: Zeta
    - Acurácia: Adaptive Quantizer

Na tabela abaixo temos a discriminação completa dos métodos:

Methods	Global/ local	Supervised/ unsupervised	Direct/ incremental	Splitting/ merging	Static/ dynamic
Equal-width	Global	Unsupervised	Direct	Splitting	Static
Equal-frequency	Global	Unsupervised	Direct	Splitting	Static
1R	Global	Supervised	Direct	Splitting	Static
D2	Local	Supervised	Incremental	Splitting	Static
Entropy (MDLP)	Local	Supervised	Incremental	Splitting	Static
Mantaras	Local	Supervised	Incremental	Splitting	Static
ID3	Local	Supervised	Incremental	Splitting	Dynamic
Zeta	Global	Supervised	Direct	Splitting	Static
Accuracy	Global	Supervised	Direct	Splitting	Static
ChiMerge	Global	Supervised	Incremental	Merging	Static
Chi2	Global	Supervised	Incremental	Merging	Static
ConMerge	Global	Supervised	Incremental	Merging	Static

Com o objetivo de testar a aplicabilidade das técnicas de Discretização, o passo final do trabalho consiste de um teste em uma base de dados consolidada na literatura, a base Iris. A realização deste teste fez uso do algoritmo de aprendizagem por classificação C4.5, uma vez que este consegue lidar bem tanto com variáveis contínuas quanto com variáveis discretas, além de ser de fácil acesso. Para poder avaliar o ganho com o uso das técnicas de Discretização, o algoritmo foi aplicado aos dados duas vezes, uma diretamente e outra com o uso das técnicas.

Os resultados dos testes podem ser vistos nas tabelas abaixo. Quanto aos resultados referentes aos testes sem uso das técnicas de Discretização, estes podem ser vistos na coluna Continuous.

Data	Continuous	Zeta	ChiMerge	Chi2
Australian	15.28 ± 5.84	15.60 ± 4.13	14.42 ± 5.42	13.50 ± 5.14
Breast	4.72 ± 1.25	13.05 ± 5.83	4.92 ± 2.75	5.01 ± 2.43
Glass	1.86 ± 2.28	2.31 ± 4.20	1.90 ± 2.28	3.20 ± 1.22
Heart	22.16 ± 4.14	16.85 ± 4.66	20.21 ± 4.10	20.00 ± 4.12
Vehicle	26.87 ± 4.57	29.00 ± 3.54	30.87 ± 4.57	33.33 ± 2.13
Iris	4.34 ± 2.84	8.24 ± 6.67	5.02 ± 3.48	4.01 ± 3.32
Wine	6.22 ± 6.84	6.82 ± 6.58	7.92 ± 5.80	6.90 ± 4.04
Pima	26.22 ± 2.65	37.70 ± 5.97	27.31 ± 4.43	26.91 ± 3.12
Bupa	33.13 ± 5.70	34.73 ± 6.45	33.99 ± 8.39	32.09 ± 5.55
Thyroid	8.00 ± 4.31	23.77 ± 9.16	8.91 ± 4.43	9.21 ± 2.22
Ionos	9.14 ± 3.78	11.37 ± 6.29	8.94 ± 4.52	8.52 ± 3.22
Average	14.36	18.13	14.95	14.79

Data	Eq-Freq	1R	D2	MDLP	Mantaras
Australian	14.51 ± 6.08	13.00 ± 5.47	14.13 ± 5.96	14.00 ± 5.90	13.82 ± 5.64
Breast	7.65 ± 3.59	13.27 ± 3.32	5.30 ± 3.09	6.37 ± 4.07	7.79 ± 3.03
Glass	22.43 ± 11.09	18.79 ± 8.04	2.79 ± 4.18	2.31 ± 3.08	2.77 ± 4.27
Heart	22.86 ± 6.42	20.00 ± 7.17	22.13 ± 4.71	20.35 ± 5.75	16.07 ± 4.29
Vehicle	31.39 ± 5.00	28.57 ± 3.97	27.90 ± 4.26	29.47 ± 5.03	29.81 ± 6.44
Iris	8.14 ± 5.75	6.07 ± 6.73	5.13 ± 5.79	4.25 ± 4.58	10.23 ± 5.10
Wine	7.96 ± 5.17	6.84 ± 6.67	6.78 ± 5.49	7.95 ± 7.27	7.53 ± 8.35
Pima	27.68 ± 4.67	25.17 ± 4.20	24.42 ± 4.32	25.21 ± 4.23	22.91 ± 8.65
Bupa	43.96 ± 8.96	36.32 ± 6.40	34.32 ± 6.07	34.29 ± 8.27	31.90 ± 8.39
Thyroid	12.69 ± 8.90	6.44 ± 3.08	8.98 ± 6.00	4.23 ± 4.44	7.90 ± 5.11
Ionos	9.67 ± 5.74	11.98 ± 6.23	8.58 ± 5.10	9.15 ± 5.38	10.69 ± 5.32
Average	18.99	16.95	14.59	14.33	14.69

Com os valores obtidos, o resultado atingido é o de que para todas as técnicas de Discretização não há uma variação muito grande com relação ao erro das predições.

Abaixo, seguem duas tabelas com o tempo gasto durante a aprendizagem com a aplicação de cada técnica de Discretização:

Data	Eq-Freq	1R	D2	MDLP	Mantaras	Zeta	ChiMerge	Chi2
Australian	0.87	0.68	1.43	1.51	5.97	0.72	1.16	2.01
Breast	0.79	0.78	0.92	0.74	1.85	0.77	0.72	0.92
Glass	0.29	0.28	0.45	0.71	1.40	0.36	0.37	0.41
Heart	0.33	0.37	0.51	0.45	0.93	0.34	0.46	0.55
Vehicle	1.85	1.88	2.74	1.90	16.55	2.04	2.08	2.10
Iris	0.71	0.76	1.00	0.65	2.25	0.80	0.96	1.02
Wine	0.29	0.30	0.42	0.45	1.76	0.33	0.43	0.55
Pima	0.70	0.75	0.92	0.91	5.62	0.73	0.55	0.61
Bupa	0.24	0.26	0.33	0.33	0.46	0.28	0.29	0.31
Thyroid	0.13	0.15	0.21	0.22	0.80	0.18	0.19	0.21
Ionos	1.62	1.75	2.10	1.87	7.41	1.79	2.08	2.32
Average	0.71	0.72	1.00	0.89	4.09	0.76	0.84	1.00

Data	Continuous	Eq-Freq	1R	D2	MDLP	Mantaras	Zeta	ChiMerge	Chi2
Australian	0.43	0.31	0.27	0.27	0.31	0.26	0.22	0.28	0.10
Breast	0.13	0.06	0.09	0.15	0.10	0.10	0.14	0.13	0.15
Glass	0.10	0.03	0.07	0.04	0.06	0.06	0.06	0.10	0.05
Heart	0.19	0.04	0.08	0.12	0.11	0.09	0.08	0.12	0.04
Vehicle	0.89	0.46	0.57	0.53	0.85	0.54	0.57	0.71	0.82
Iris	0.01	0.02	0.02	0.01	0.01	0.03	0.02	0.01	0.01
Wine	0.12	0.06	0.07	0.05	0.06	0.04	0.04	0.09	0.02
Pima	0.31	0.23	0.21	0.20	0.20	0.30	0.10	0.16	0.11
Bupa	0.11	0.12	0.12	0.07	0.15	0.11	0.04	0.11	0.09
Thyroid	0.05	0.04	0.02	0.06	0.04	0.02	0.02	0.09	0.02
Ionos	1.12	0.75	0.34	0.24	0.34	0.75	0.22	0.20	0.24
Average	0.31	0.19	0.16	0.15	0.20	0.20	0.13	0.18	0.15

Como pode ser visto, o uso de todas as técnicas apresentam ganho em tempo quando comparado à ausência de Discretização no processo de aprendizagem.

Conclusão:

O trabalho de pesquisa e estudo da literatura existente a respeito de processos de discretização resultou na descrição abstrata típica de um processo - ordenação, avaliação, divisão/junção e condição de parada - bem como a proposta de uma estrutura abstrata (*framework*) para categorizar os métodos existentes através uma demonstração sistemática de resultados das aplicações de vários destes métodos utilizando um conjunto de dados de referência: o *Iris dataset*. Foram utilizadas duas medidas de avaliação: número de inconsistências e número de pontos de corte (ponto para dividir os dados em duas partes), e uma relação intuitiva entre elas é que quanto mais pontos de corte, menos inconsistências. Portanto, um bom conjunto de resultados deve ser aquele com valores baixos em ambas as medidas de avaliação. Ressalta-se aqui que estas demonstrações foram efetuadas sem recorrer a nenhum algoritmo de classificação

Escolhidos alguns métodos através deste *framework*, estes foram experimentados em 11 conjuntos de dados levaram em consideração o tempo gasto para a discretização e aprendizado, a taxa de erros dos dados discretizados em comparação aos dados originais e números de nós em uma árvore de decisão.

Em geral, o tempo gasto na discretização leva a uma melhor precisão. Encontrar o método de discretização que melhor se adequa depende da necessidade do usuário e dos tipos de dados. Por exemplo, caso a necessidade seja remover dados irrelevantes ou redundâncias, o melhor método seria o **Chi2**; caso o objetivo seja incorporar a discretização no processo de aprendizagem, o **Contrast** é uma boa opção; se a pretensão é simplesmente discretizar os dados, "ignorando" o restante (dos dados), **Entropy** deve ser seu guia.

Há também que se destacar que o estudo aqui apresentado levou em consideração uma discretização univariada - o processo aplicado a um único atributo - por questão de eficiência. Sabe-se que, em geral, o processo de discretizar leva em consideração mais de um atributo (multivariado), e isso aumenta a complexidade do tempo. Além disso, aborda-se a questão do ruído (conteúdo nas bases de dados que pode prejudicar a qualidade da informação extraída, a partir de qualquer método, seja ele tradicional ou baseado em estratégias mais elaboradas. Destacam-se como ruídos: valores fora do domínio, ausência de valores, inconsistências etc.) e que sua tolerância é uma prática comum nos métodos abordados.

Em suma, este artigo não é sobre uma conclusão de pesquisa de discretização. Em vez disso, trata-se do início de uma nova fase de pesquisa.

Referências Bibliográficas:

Liu et al. Discretization: An Enabling Technique. Data Mining and Knowledge Discovery, 6, 393–423, 2002.