MODELAGEM E PREPARAÇÃO DE DADOS PARA APRENDIZADO DE MÁQUINA:

Pré-processamento

Professor:

Luis E. Zárate

Pré-processamento

- Os dados precisam ser pré-processados, codificados e transformados de forma que possam servir de entrada para os algoritmos de Mineração de Dados, da etapa seguinte.

- Normalmente se faz necessária a transformação ou mudanças de escala dos dados garantindo que se preservem as características dos valores originais.

- A melhor forma de transformar os dados é verificar quais requisitos a solução precisa atender e quais são os requisitos que a técnica de mineração de dados impõe.

Técnicas de Discretização

- Muitos algoritmos como C4.5, Apriori, Naive Bayes utilizam dados categóricos ou nominais ao invés de dados contínuos. A discretização é uma importante e comúm tarefa em Data Mining.

- Muitos problemas do mundo real, contendo dados numéricos contínuos precissam de métodos de discretização que os convertam para dados categóricos.

- O processo de discretização transforma dados quantitativos em dados qualitativos. Por exemplo, transforma atributos numéricos em atributos discretos ou nominais com um número finito de intervalos.

Técnicas de Discretização

- Vantagens:
  - Algoritmo capazes de lidar com dados contínuos, podem ser menos eficientes para compreensibilidade dos resultados.
  - A discretização também pode ser entendida como uma técnica para a redução de dados.

- Desvantagens:
  - Perda de informação
  - Reduzir esta perda é o principal objetivo dos métodos de discretização.

Abordagens de Discretização

- Não-supervisionada:
  - A discretização é realizada sem levar em conta a informação dos grupos a que pertencem as instâncias de treinamento.
- Supervisionada:
  - A discretização é realizada levando em conta os grupos a que pertencem as instâncias no conjunto de treinamento.

Não-Supervisionada: Mapeamento em intervalos

a) Intervalos com tamanho pré-definidos (univariada)

0 a 1 → 0 2 a 5 → 1 6 a 99 → 2

b) Intervalos de igual tamanho

2 intervalos / 5 valores: 0 a 4 → 0 5 a 9 → 1

c) Intervalos com o mesmo número de elementos

{1,2,3,6,7,8,9,13,15,16}

{1,2,3,6,7} →1 {8,9,13,15,16} →2

Não-Supervisionada: Mapeamento em intervalos

d) Número pré-determinado de intervalos uniformes (*equal-interval binning*)

No exemplo (idade):

64, 65, 68, 69, 70, 71, 72, 72, 75, 75, 80, 81, 83, 85

Considerando Bins (caixas) com largura **6**:

60 < x ≤ 66: 64, 65

66 < x ≤ 72: 68, 69, 70, 71, 72, 72

72 < x ≤ 78: 75, 75

78 < x ≤ 84: 80, 81, 83

84 < x ≤ 90: 85

- Como qualquer método não supervisionado, arrisca destruir distinções úteis, devido a divisões muito grandes ou fronteiras inadequadas
- Distribuição de amostras muito irregular, com algumas *bins*com muitas amostras e outras com poucas amostrar.

Não-Supervisionada: Mapeamento em intervalos

e) Número uniforme de amostras por intervalo (*equal-frequency binning*)

Também chamado de equalização do histograma
Cada *bin* tem o mesmo número aproximado de amostras

No exemplo (idade):

64, 65, 68, 69, 70, 71, 72, 72, 75, 75, 80, 81, 83, 85

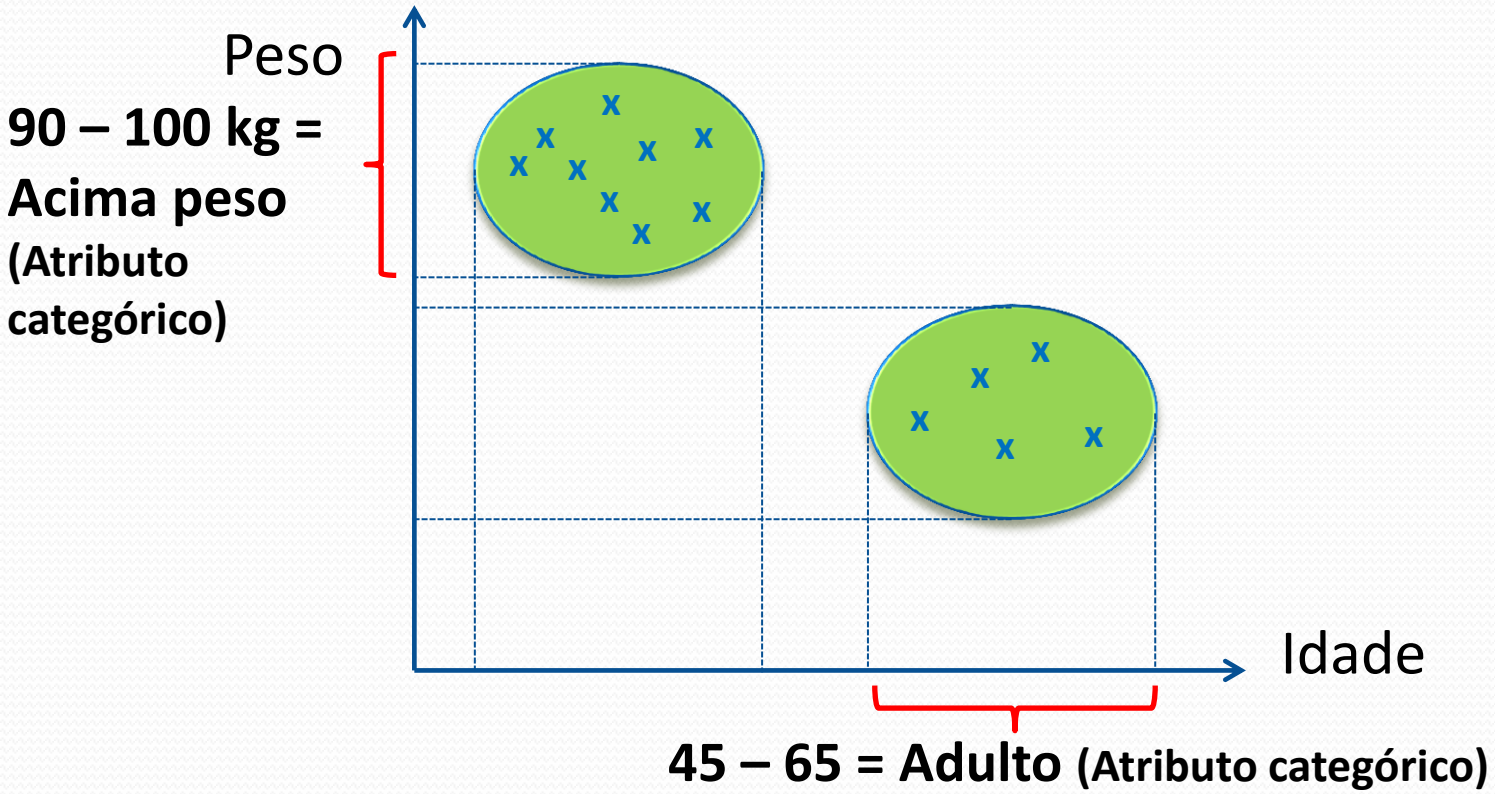
64 65 68 69 | 70 71 72 72 | 75 75 80 | 81 83 85
14 amostras: **4** *Bins*

x ≤ 69,5: **64, 65, 68, 69**
69,5 < x ≤ 73,5: **70, 71, 72, 72**
73,5 < x ≤ 80,5: **75, 75, 80**
x > 80,5: **81, 83, 85**

Não-Supervisionada: Mapeamento em intervalos

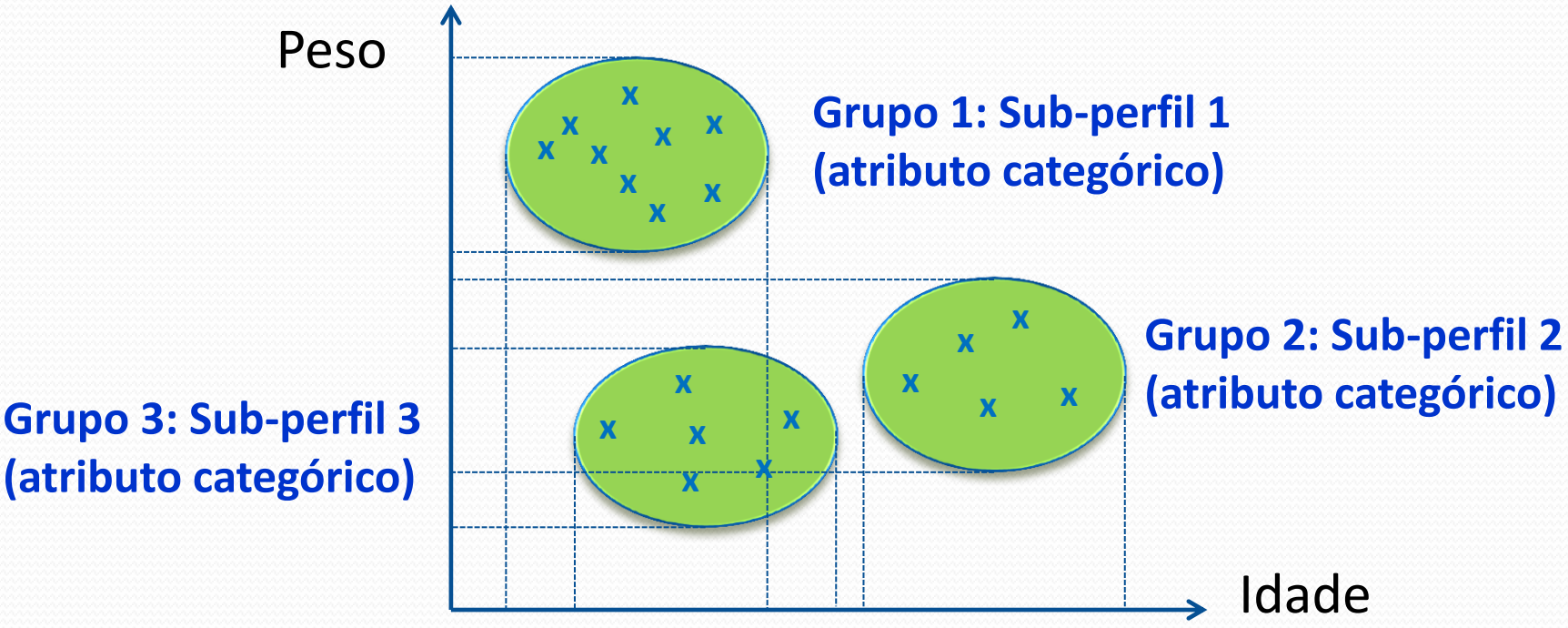
f) Intervalos por meio de clusterização (multivariada)

Utiliza algum algoritmo de agrupamento de dados para descobrir automaticamente a distribuição dos dados



Não-Supervisionada: Mapeamento em intervalos

- A técnica de mapeamento em intervalos pode levar a sobre-posição de variáveis categóricas o que pode levar a um processo de discretização com incertezas. Frente a este problema pode ser dado um valor categórico para todo o grupo



Principais características de um Discretizador

- **Direto versus Incremental:**

- Os métodos de discretização **diretos** dividem os valores do atributo em um **número de intervalos** previamente definido pelo usuário.

- Nos métodos **incrementais**, o processo de discretização é realizado iterativamente até que **um critério de parada** seja alcançado.

Principais características de um Discretizador

- **Estático versus dinâmico:**

-Esta característica refere-se ao momento e à independência que o discretizador opera em relação ao aprendizado.

-Um discretizador **dinâmico** é executado **no momento** da construção do modelo, ou seja, o discretizador é embutido no algoritmo de aprendizagem.

-Um discretizador **estático**é executado **antes** do processo de aprendizagem e independe do algoritmo de aprendizagem.

Principais características de um Discretizador

- **Univariada versus Multivariada:**

- Métodos **univariados** consideram **um atributo** contínuo por vez, não levando em conta a relação entre os atributos.

- **Multivariados** podem considerar **simultaneamente todos os atributos** e a relação de dependência entre eles.

Principais características de um Discretizador

- **Supervisionado versus não-supervisionado:**

- Discretizadores **não-supervisionados** não consideram o rótulo da classe, enquanto os **supervisionados** o fazem.

- A maioria dos discretizadores propostos na literatura é supervisionada e teoricamente, usando informações de classe, deve determinar automaticamente o melhor número de intervalos para cada atributo.

Principais características de um Discretizador

- **Divisão versus Fusão:**

- Métodos de **Divisão** realizam a discretização por meio de um processo iterativo de subdivisão do intervalo de valores inicial que é executado até que uma condição de parada seja satisfeita.

- Os métodos de **Fusão** iniciam com os valores do atributo contínuo particionados e, iterativamente, realizam a junção dessas partições enquanto um critério de parada não é alcançado

- Além disso, alguns discretizadores podem ser considerados **híbridos** devido ao fato de que eles podem alternar as divisões.

Principais características de um Discretizador

- **Global versus Local:**

Para tomar uma decisão, um discretizador pode considerar todos os dados disponíveis do atributo ou usar apenas informações parciais.

- Um discretizador **global** utiliza todas as informações disponíveis.

- Um discretizador **local** faz uso apenas de parte das informações

- Alguns algoritmos seguem o esquema de divisão e conquista e quando é encontrada uma divisão, os dados são recursivamente divididos, restringindo o acesso a dados parciais.

Principais características de um Discretizador

- **Medidas de avaliação:**

- Essas medidas de avaliação são utilizadas pelos discretizadorespara comparar intervalos de valores durante o processo de discretização.

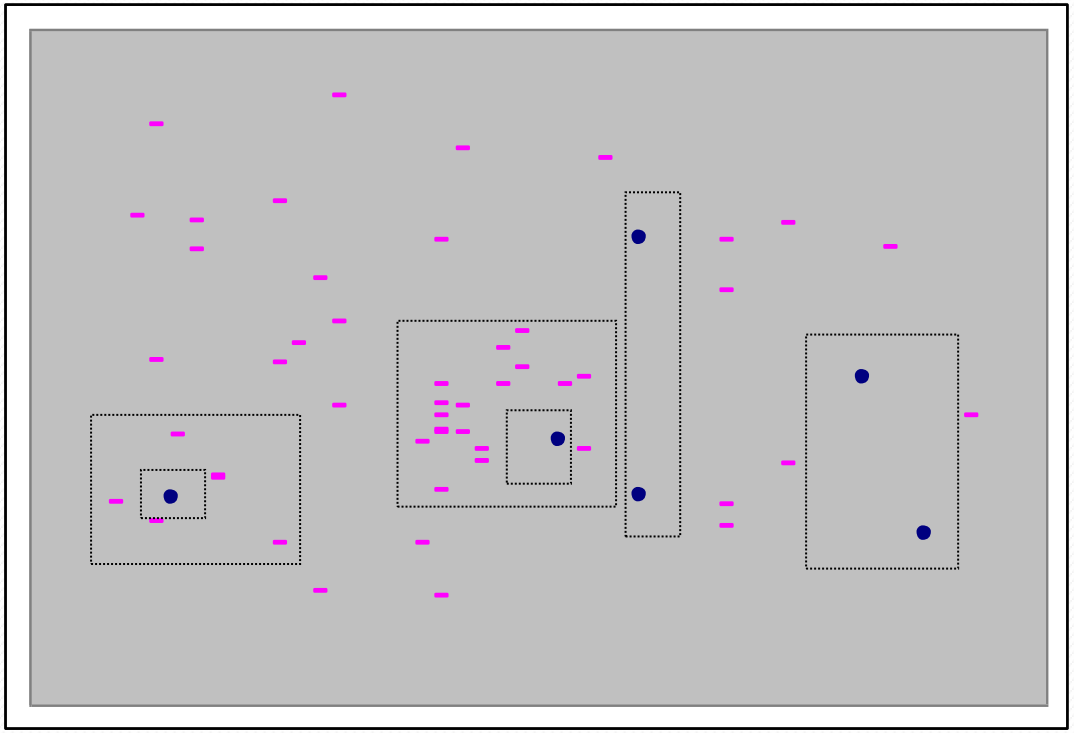
São consideradas 5 medidas:

- **Informação**: utiliza a entropia
- **Estatística**: dependência/correlação entre os atributos
- **Conjuntos aproximados**: uso de medidas de conjuntos aproximados
- ***Wrapper:*** execução de algoritmos de classificação.
- ***Binning:*** quantidade pré-determinada de intervalos

Base de Dados Desbalanceadas

Uma bases de dados é considerada desbalanceada se existe nela objetos (instâncias) pertencentes a uma classe, em menor número em relação a outra.

É considerado base desbalanceada se a relação do número de objetos das classes são da ordem de 1:100, 1:1000 ou 1:10000.



Base de Dados Desbalanceadas

As bases de dados desbalanceadas podem ocorrer por dois motivos:

Motivos Intrínsecos:

Devido à própria natureza do problema. Exemplo, detecção de fraude, gerenciamento de riscos, diagnóstico e monitoramento médico.

Motivos não Intrínsecos:

Devido ao limitado processo da coleta de dados, por razões econômicas ou privacidade.

Base de Dados Desbalanceadas

Existem vários métodos para trabalhar com bases de dados desbalanceadas:

1) Método da Amostragem:

Amostragem sem reposição (**Under-sampling**) diminui os exemplos da maior classe, enquanto a amostragem com reposição (**Over-sampling**) aumenta o número de exemplos da menor classe. Ambas procuram acabar com a raridade da base de dados.

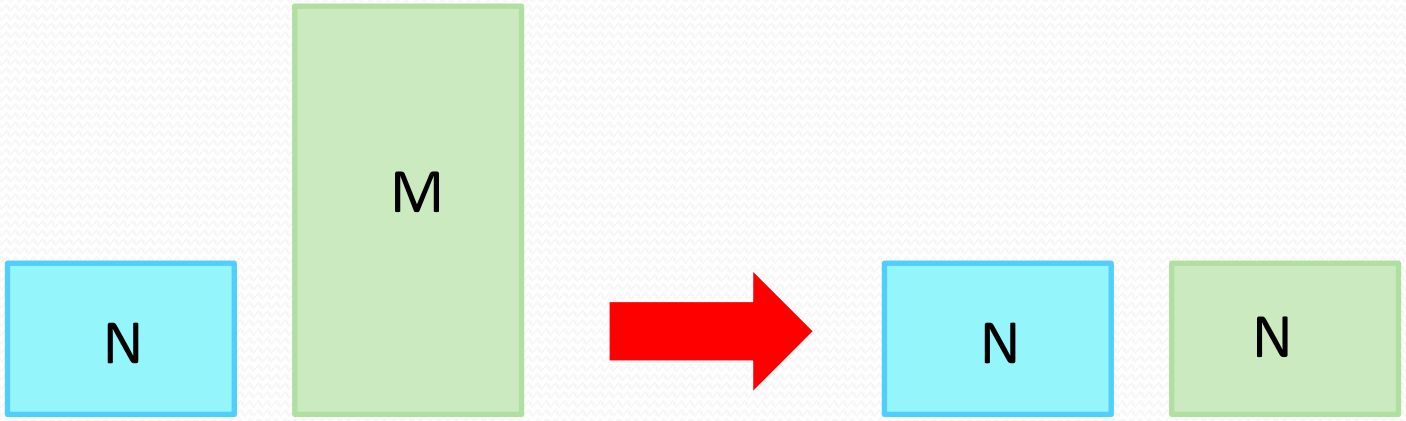
O Over-sampling, tende a duplicar os exemplos da menor classe, não criando novos dados e continuando com a carência dos mesmos.

Base de Dados Desbalanceadas

2) Método seleção aleatória pela menor classe

Dado dois conjuntos de registros com N e M registros (onde N<<M) vinculados a duas classes.

O balanceamento ocorre selecionando de forma aleatória N registros dentro do conjunto contendo M registros.

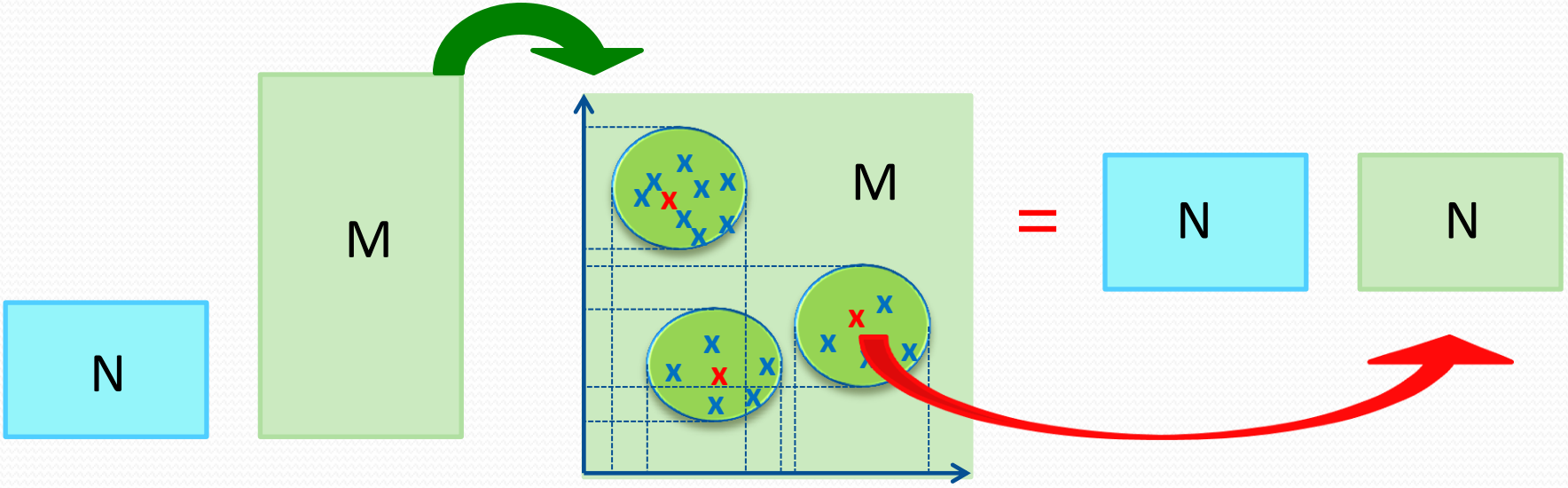


Base de Dados Desbalanceadas

3) Método seleção por agrupamento pela menor classe

Dado dois conjuntos de registros com N e M registros (onde N<<M) vinculados a duas classes.

O balanceamento ocorre selecionando por meio de uma técnica de agrupamento os N registros mais representativos dentro do conjunto contendo M registros.

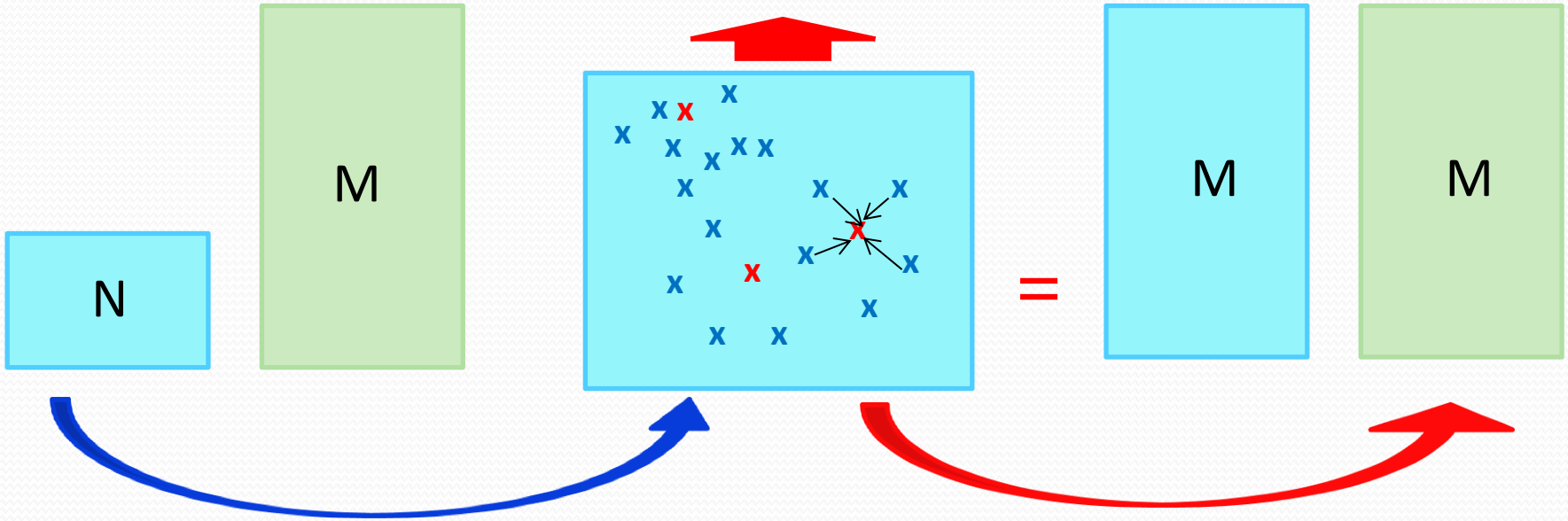


Base de Dados Desbalanceadas

3) Replicação de instâncias – Método Smooth

Dado dois conjuntos de registros com N e M registros (onde N<<M) vinculados a duas classes.

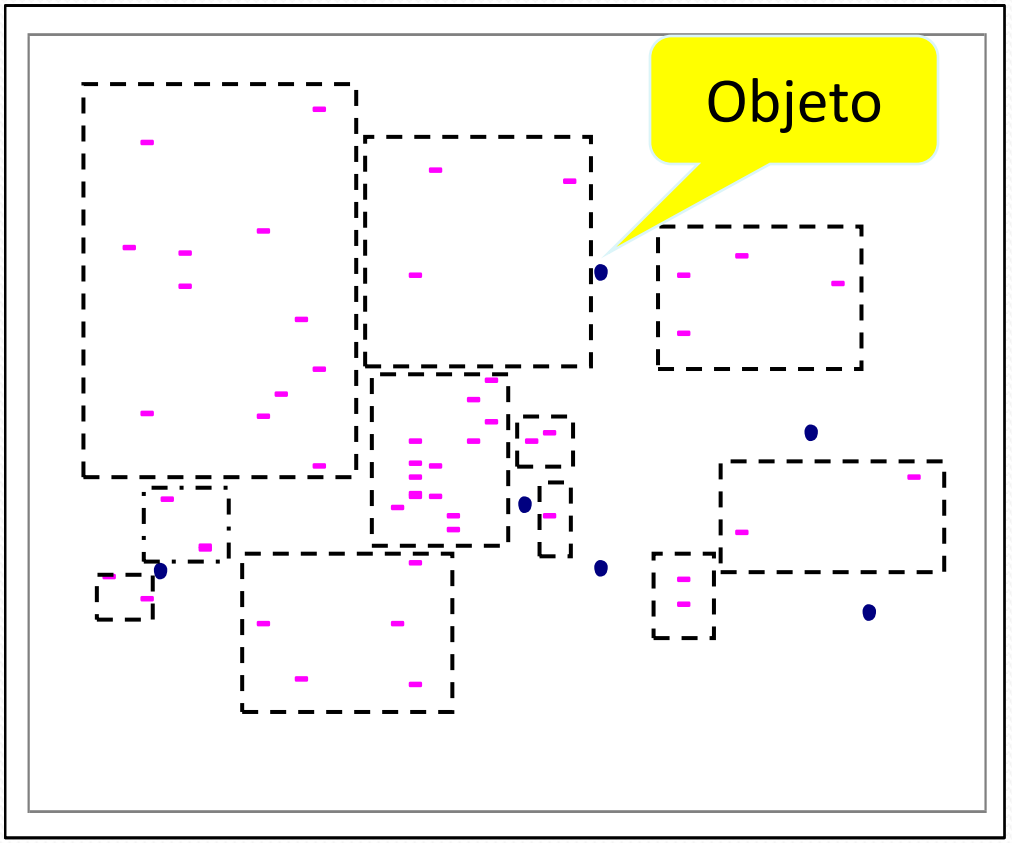
O balanceamento ocorre gerando artificialmente instâncias apartir das instância do conjunto contendo N registros.



Base de Dados Desbalanceadas

4) Aprendizado por meio de Uma-classe mais próxima CNN:

Para determinar se um objeto pertence ou não a um classe já classificada e bem caracterizada, é possível utilizar o modelo desta classe para analisar se um objeto pertence ou não através da similaridade.



Obrigado

Professor:

Luis E. Zárate