Universidade Federal Fluminense
Pólo Universitário de Rio das Ostras
Faculdade Federal de Rio das Ostras
Bacharelado em Ciência da Computação

Rodrigo Magalhães Rodovalho

Uso de um método de aprendizado não-supervisionado hierárquico para melhorar a construção de classificadores multirrotulo para bases de dados com grande quantidade de rotulos.

Rio das Ostras-RJ
2016

RODRIGO MAGALHÃES RODOVALHO

USO DE UM MÉTODO DE APRENDIZADO NÃO-SUPERVISIONADO HIERÁRQUICO PARA MELHORAR A CONSTRUÇÃO DE CLASSIFICADORES MULTIRRÓTULO PARA BASES DE DADOS COM GRANDE QUANTIDADE DE RÓTULOS.

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Fluminense, como requisito parcial para obtenção do Grau de Bacharel. Área de Concentração: Mineração de Dados e Inteligencia Artificial.

Orientador: Prof. Dr. FLÁVIA CRISTINA BERNARDINI

Niterói-RJ

2016

RODRIGO MAGALHÃES RODOVALHO

USO DE UM MÉTODO DE APRENDIZADO NÃO-SUPERVISIONADO HIERÁRQUICO PARA MELHORAR A CONSTRUÇÃO DE CLASSIFICADORES MULTIRRÓTULO PARA BASES DE DADOS COM GRANDE QUANTIDADE DE RÓTULOS

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Fluminense, como requisito parcial para obtenção do Grau de Bacharel. Área de Concentração: Mineração de Dados e Inteligencia Artificial.

Aprovada em MÊS de ANO.

BANCA EXAMINADORA

Prof. Dr. FLÁVIA CRISTINA BERNARDINI - Orientador
UFF

Prof. NOME DO PROFESSOR
INSTITUIÇÃO

Prof. NOME DO PROFESSOR
INSTITUIÇÃO

Niterói-RJ
2016

Dedico este trabalho ...

Meu filho heitor

Agradecimentos

Agradecer a Deus agradecer meus pais Agradecer Yasmin agradecer minha orientadora ageadecer todos que me apoiaram

Lista de Figuras

Lista de Tabelas

Sumário

Resumo

O aprendizado multirrótulo tem por objetivo a construção de classificadores que rotulam, com mais de um rótulo, casos ainda não rotulados, como é o caso de diagnóstico de falhar em um equipamento, ou gêneros musicais de uma música. Uma questão importante do aprendizado multirrótulo está relacionado à grande quantidade de exemplos (casos de aprendizado) disponíveis, sendo cada exemplo associado a poucos rótulos, e esses, por sua vez, são oriundos de um grande conjunto de rótulos possíveis. O objetivo deste trabalho é explorar um método de aprendizado não-supervisionado hierárquico como forma de melhorar o processo de aprendizado multirrótulo, como uma técnica de divisão e conquista do problema. Para atingir esse objetivo, utilizaremos as ferramentas Mulan e Weka para apoiar o desenvolvimento do método a ser proposto, e utilizaremos bases de dados naturais para avaliar o desempenho do método a ser proposto.

Palavras-chave: Mineração de Dados. Aprendizado de Máquina.

Abstract

The multi-label learning aims to build classifiers that label, with more than one label, cases not labeled yet, such as diagnostic fails in a device, or genres of a music. An important issue of multi-label learning is related to the large number of examples (cases of learning) available, each instance associated with a few labels, and these, in turn, are derived from a large set of possible labels. The objective of this paper is to explore a non-supervised hierarchical learning method in order to improve the multi-label learning process, as a division and conquer technical of the problem. To achieve this goal, we will use the Mulan and Weka tools to support the development of methods being proposed, and will use natural databases to evaluate the performance of the method being proposed.

Keywords: Data Mining. Machine Learning.

Capítulo 1

Introdução

Aprendizado multirrótulo é uma linha de pesquisa da sub-área de aprendizado de máquina com bastante foco nos últimos tempos. O aprendizado multirrótulo objetiva a construção de classificadores que rotulam, com mais de um rótulo, casos ainda não rotulados, como é o caso de diagnóstico de falhas em um equipamento, ou gêneros musicais de uma música. Uma questão importante do aprendizado multirrótulo está relacionado à grande quantidade de exemplos (casos de aprendizado) disponíveis com poucos rótulos associados, em geral, oriundo de um grande conjunto de rótulos possíveis. O objetivo deste trabalho é investigar o uso de aprendizado de máquina não supervisionado para auxiliar o processo de aprendizado multirrótulo.

[colocar mais coisa]

No capítulo 2, são apresentados os conceitos de aprendizado de máquina. Assim como suas divisões e suas abordagens, aprendizado supervisionado,não supervisionado e aprendizado monorrótulo e multirrótulo, respectivamente. No capítulo 3, é apresentado uma proposta de um método de aprendizado multirrótulo baseado em aprendizado de máquina não-supervisionado hierárquico. No capítulo 4, são mostrados os resultados obtidos pela execução do método proposto e comparações de resulatos com um algoritmos conhecidos da literatura. No capítulo 5, é apresentado uma análise dos resultados obtidos e no capítulo 6, sugestões para trabalhos futuros.

Capítulo 2

Aprendizado de Máquina

Um dos desafios da Inteligência Artificial é construir sistemas que façam a máquina aprender conceitos e/ou se adaptar ao ambiente. A subárea da Inteligência Artificial relacionada a esse tipo de problema é denominada aprendizado de máquina.[1] O aprendizado indutivo, um tipo de aprendizado, tem por objetivo inferir padrões ou conhecimento, representados por uma hipótese, ou modelo, a partir de exemplos fornecidos. O aprendizado indutivo pode ser dividido em dois tipos: aprendizado supervisionado e não supervisionado.

No aprendizado supervisionado, cada exemplo fornecido ao indutor possui um rótulo, oferecido por um supervisor especialista do domínio de onde os dados são provenientes. O objetivo do aprendizado supervisionado é construir uma hipótese, ou modelo, que rotule novos exemplos ainda não rotulados. No problema padrão de aprendizado supervisionado, a entrada do algoritmo consiste de um conjunto de objetos rotulados, ou exemplos, S, com N objetos $T_i, i = 1, ..., N$, escolhidos de um domínio X com uma distribuição $\mathcal{D}$ fixa, desconhecida e arbitrária, da forma $\{(\mathbf{x}_i, Y_i), ..., \mathbf{x}_N, Y_N)\}$ para alguma função desconhecida y = f($\mathbf{x}$). Os ($\mathbf{x}_i$ são tipicamente vetores da forma $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iM})$ com valores discretos ou numéricos, e X_{ij} refere-se ao valor do atributo j, denominado X_j , do exemplo T_i. Os valores y_i referem-se ao valor do atributo Y, frequentemente denominado classe. Os valores de y são tipicamente pertencentes a um conjunto discreto de classes $\mathbf{C} = \{\mathbf{C}_1, ..., \mathbf{C}_R\}$, quando se trata de classificação, ou ao conjunto de números reais em caso de regressão. Neste trabalho, somente serão abordados problemas de classificação. Assim, quando for dito que um exemplo pertence a uma determinada classe $\mathbf{C}_r$, isso significa que o exemplo possui $\mathbf{C}_r$ como valor de y, ou, ainda, o valor $\mathbf{C}_r$ foi associado a y quando o problema é associar uma classe a um exemplo. O problema de associar somente uma classe a cada exemplo é também denominado aprendizado monorrótulo. No aprendizado não supervisionado, os exemplos não são rotulados, e o objetivo desse aprendizado é realizar descoberta de conhecimento por investigação. Nesse caso, analogamente ao aprendizado supervisionado, o conjunto de objetos não-rotulados S é composto por N objetos $T_i, i = 1, ..., N$, escolhidos de um domínio X com uma distribuição $\mathcal{D}$ fixa, desconhecida e arbitrária, da forma $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$, onde $\mathbf{x}_i$ são tipicamente vetores da forma $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iM})$ com valores discretos ou numéricos, e $\mathbf{x}_{ij}$ refere-se ao valor do atributo j, denominado $\mathbf{X}_j$, do exemplo T_i.

[1] Neste trabalho, para efeitos de simplificação, quando o termo aprendizado for utilizado, estará se referindo ao aprendizado de máquina

2.1 Aprendizado de Máquina Não-Supervisionado Hierárquico

Aprendizado não supervisionado pode ser definido como um modo de encontrar padrões nos dados e ignorar de uma certa maneira o que é considerado puramente ruído. Existem técnicas de aprendizado não supervisionado, como por exemplo, clusterização e redução de dimensionalidade. Ambas consideradas pilares do aprendizado não suupervisionado [14].

A técnica de redução de dimensionalidade consiste na redução do número de atributos. O objetivo dessa técnica é obter um conjunto de atributos a partir da redução do espaço de busca pela solução. Portanto o conjunto obtido possui menor dimensionalidade em relação ao conjunto original, entretanto a qualidade da solução final deve ser mantida [23]. Embora a técnica de redução de dimensionalidade seja amplamente utilizada, nesse trabalho o nosso enfoque se concentra na técnica de clusterização.

Clusterização, ou análise de cluster, consiste na busca de agrupar elementos de dados baseando-se na similaridade entre eles [12]. O objetivo é conseguir determinar os grupos, a fim de obter homogeneidade dentro dos grupos e heterogeneidade entre eles. Devido a grande capacidade de armazenamento de dados, esse grande volume de dados podem gerar muitas combinaçoes de grupos, o que dificulta a sua análise, devido ao grande custo computacional.

Afim de resolver esse impasse, foram desenvolvidas várias técnicas que auxiliam na formação dos grupos. Algumas caracterısticas são vitais à essas tecnicas, para que elas consigam um resultado satisfatório, como por exemplo, ser capaz de lidar com dados com alta dimensionalidade, habilidade para lidar com diferentes tipos de dados, entre outros [33]. Ainda nessa seção, serão apresentadas técnicas e métricas que auxiliam nessa análise.

Dois fatores são fundamentais no processo de agrupamento: (1) uma medida de similaridade e (2) uma estratégia de agrupamento. A maneira com que os grupos são obtidos está diretamente relacionada a medida de similaridade escolhida, que determinam como a similaridade entre dois elementos é calculada. Essa escolha depende dos tipos de atributos que representam os exemplos [21]. Para indução de modelos de agrupamento a partir dos dados, são utilizados métodos e algoritmos, que correspondem às estratégias de agrupamento [15].

2.1.1 Medidas de similaridade

Medidas de ”proximidade”ou ”similaridade”são necessárias na maioria das vezes, quando se quer obter de um conjunto de dados complexo, uma estrutura simples de grupos [16]. Essas medidas são calculadas entre os elementos a serem agrupados. A partir da utilização dessas medidas podem ser definidas as relações intra-cluster, ou seja, as relações definidas, para que elementos pertençam a um mesmo cluster. Normalmente as medidas de similaridade são funções de distâncias ou métricas. Ainda nessa seção, serão apresentadas algumas das medidas de similaridade, expressas por funções de distâncias, comumente utilizadas.

Distância Euclidiana

Distância euclidiana é considera uma distância geométrica no espaço multidimensional. Utilizando dois elementos $X = [X_1, X_2, ..., X_p]$ e $Y = [Y_1, Y_2, ..., Y_p]$ ela pode ser calculada através da equação 2.1:

$$d_{x,y} = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + ... + (X_p - Y_p)^2} = \sqrt{\sum_{i=1}^{p}(X_i - Y_i)^2} \tag{2.1}$$

Distância Euclidiana Quadrática

Distância euclidiana quadrática é definida pela equação 2.2:

$$d_{x,y} = (X_1 - Y_1)^2 + (X_2 - Y_2)^2 + ... + (X_p - Y_p)^2 = \sum_{i=1}^{p}(X_i - Y_i)^2 \tag{2.2}$$

Distância Manhattan

Distância Manhattan calcula a distância que seria percorrida para chegar de um ponto de dados para o outro, se um caminho do tipo grade for seguido. A distância Manhattan entre os dois itens é a soma das diferenças dos seus componentes correspondentes, logo pode ser obtida utilizando a equação 2.3:

$$d_{x,y} = (|X_1 - Y_1|) + (|X_2 - Y_2|) + ... + (|X_p - Y_p|) = \sum_{i=1}^{p}(|X_i - Y_i|) \tag{2.3}$$

Distância Chebychev

Distância Chebychev é utilizada, na situação que se deseja diferenciar dois elementos, se houver apenas uma das dimensões diferente, a distância de chebychev é definida pela equação 2.4:

$$d_{x,y} = max(|X_1 - Y_1|) + (|X_2 - Y_2|) + ... + (|X_p - Y_p|) \tag{2.4}$$

2.1.2 Métodos de Agrupamento

Métodos de agrupamento podem ser organizados em dois tipos [24]: agrupamento particional e agrupamento hierárquico. No primeiro tipo, um conjunto de exemplos é dividido em uma partição simples de k grupos. Enquanto no segundo, exemplos são organizados em grupos e subgrupos a partir de uma sequência de partições aninhadas [21].

O método proposto nesse trabalho utiliza a abordagem de clusterização hierárquica, portanto, a seguir iremos apresentar com mais detalhes alguns métodos hierárquicos utilizados para realizar agrupamentos.

Métodos hierárquicos

Os métodos hierárquicos de clusterização consistem em uma série de sucessivos agrupamentos (clusters) ou sucessivas divisões de elementos, onde esses elementos são agregados ou desagregados [12]. A distância entre os clusters é usada como critério para formação dos mesmos [7]. Pode acontecer de um exemplo

pertencer a mais de um grupo, ou até mesmo, ocorrer de cada exempo possuir um grau de pertinência associado aos grupo. A ocorrência de alguma dessas situações é chamada de sobreposição [21].

A clusterização hierárquica pode ser feita de duas formas: aglomerativa – iniciando com tantos clusters quantos objetos e então unindo-os em novos clusters – ou divisiva – iniciando com um cluster apenas e dividindo-o em novos clusters [7]. As duas abordagens são apresentadas a seguir.

Métodos aglomerativos

Os dados são inicialmente distribuídos de modo que cada exemplo represente um cluster e, então, esses clusters são recursivamente agrupados considerando alguma medida de similaridade, até que todos os exemplos pertençam a apenas um cluster [4], representa uma estratégia *bottom-up* de agrupamento. Existem vários métodos aglomerativos, a característica utilizada para classificar esses métodos é o critério utilizado para definir as distâncias entre grupos, ou seja, as distâncias inter-cluster [12].

[colocar figura de clusters do método aglomerativo]

Em geral, os métodos aglomerativos utilizam os passos de um algoritmo padrão. A diferença entre os métodos ocorre no passo, onde a função de distância é definida. Os passos do algoritmo padrão e os métodos aglomerativos são apresentados a seguir.

Algorithm 1: Algoritmo padrão utilizado por métodos aglomerativos.

Entrada: Base de dados com **N** elementos

Saída : Um conjunto de grupos

1 Iniciar com **N** grupos, cada grupo contendo um elemento e uma matriz de similaridade $\mathbf{D_{nXn}}$

2 **repeat**

3 Encontrar a menor distância $\mathbf{d_{uv}}$ (maior similaridade);

4 Atualizar a matriz **D**, removendo os elementos **U** e **V**;

5 Atualizar a matriz **D**, inserindo as novas distâncias do grupo **(U,V)**;

6 **until *N-1, no qual todos elementos estarão em um único grupo***;

Simple Linkage: Conhecido também como, método de ligação por vizinho mais próximo, esse método utiliza a distância de menor valor, distância definida pela expressão 2.5:

$$d_{(UV)W} = min(d_{UW}, d_{VW}) \tag{2.5}$$

Segundo [2], são essas algumas características desse método: Geralmente, grupos muito próximos podem não ser identificados; Possibilita a detecção de grupos que possuem formas não-elípticas; Apresenta pouca tolerância a ruído; Demonstram bons resultados utilizando distâncias euclidianas assim como outras distâncias; Possui a tendência de formar longas cadeias (encadeamento).

Complete Linkage: Conhecido também como, método de ligação do vizinho mais distante, esse método utiliza a distância máxima, que é dada pela expressão 2.6:

$$d_{(UV)W} = max(d_{UW}, d_{VW}) \tag{2.6}$$

Segundo [17], são essas algumas características desse método: Demonstra bons resultados utilizando distâncias euclidianas assim como outras distâncias; Possui a tendência de formas grupos compactos; Os ruídos demoram a serem inseridos ao grupo.

Average Linkage: Conhecido também como, método de ligação por média, esse método utiliza a expressão 2.7, para o cálculo da distância:

$$d_{(UV)W} = \frac{(N_U.d_{UW} + N_V.d_{VW})}{N_U + N_V} \tag{2.7}$$

- onde N_U e N_V são os números de elementos no grupo U e V, respectivamente; d_{UW} e d_{VW} são as distâncias entre os elementos UW e VW, respectivamente.

Segundo [17], são essas algumas características desse método: Menor sensibilidade à ruídos comparado aos métodos de ligação por vizinho mais próximo e vizinho mais distante; Demonstra bons resultados utilizando distâncias euclidianas assim como outras distâncias; Possui a tendência de formar grupos com número de elementos similares

Centroid Linkage: Conhecido também como, método de ligação por centróide, esse método utiliza a expressão 2.8, para o cálculo da distância:

$$d_{(UV)W} = \frac{(N_U.d_{UW} + N_V.d_{VW})}{N_U + N_V} - \frac{N_U.N_V.d_{UW}}{(N_U + N_V)^2} \tag{2.8}$$

- onde N_U e N_V são os números de elementos no grupo U e V, respectivamente; d_{UW} e d_{VW} são as distâncias entre os elementos UW e VW, respectivamente.

Segundo [12], são essas algumas características desse método: Robustez à presença de ruídos; Passível de ocorrência do fenômeno da reversão, isto ocorre quando a distância entre os centróides é menor que a distância entre grupos já formados, consequentemente fará com que os novos grupos se formem ao um nível inferior aos grupos existentes. Esse método não é muito utilizado, pois devido a ocorrência do fenômeno da reversão, o resultado é um dendograma confuso.

Median Linkage: Conhecido também como, método de ligação por mediana, esse método utiliza a seguinte expressão, para o cálculo da distância:

$$d_{(UV)W} = \frac{d_{UV} + d_{VW}}{2} - \frac{d_{UV}}{4} \tag{2.9}$$

- onde d_{UW} e d_{WV} são as distâncias entre os elementos UW e VW, respectivamente.

Segundo [12], são essas algumas características desse método: Demonstra resultados satisfatórios quando os grupos são de tamanhos diferentes; Quando permutado os elementos na matriz de similaridade, pode apresentar resultado diferente; Robustez á presença de *outliers*.

Ward's Linkage: Método de ligação de Ward, nesse método a distância é calculada pea expressão:

$$d_{(UV)W} = \frac{((N_W + N_U).d_{UW} + (N_W + N_V).d_{VW} - N_W.d_{UV})}{N_W + N_U + N_V} \tag{2.10}$$

- onde N_U e N_W são os números de elementos no grupo U e V, respectivamente; d_{UW} e d_{VW} são as distâncias entre os elementos UW e VW, respectivamente.

Segundo [12], são essas algumas características desse método: Demonstra bons resultados utilizando distâncias euclidianas assim como outras distâncias; Se o número de elementos em cada grupo for praticamente igual, o método pode apresentar resultados insatisfatórios; Possui a tendência de combinar grupos com poucos elementos; Sensível à presença de *outliers*.

Métodos divisivos

O processo inicia-se com apenas um agrupamento contendo todos os dados e segue dividindo-o recursivamente segunda alguma métrica até que alcance algum criterio de parada, frequentemente o número de clusters desejados [3].

[colocar figura de clusters do método aglomerativo]

Devido a exigência de maior capacidade computacional, os métodos divisivos são pouco citados na literatura em relação aos métodos aglomerativos [17]. A seguir será apresentado o método divisivo proposto por [19].

MACNAUGHTON-SMITH: O custo computacional demandado por métodos divisivos é alto. O que pode tornar a implementação inviável, caso o número de elementos seja grande e conjunto de divisões possíveis for todo considerado, o número de iterações pode aumentar exponencialmente. Entretanto o método proposto por MacNaughton-Smith, é capaz de evitar esse problema [12]. O pseudo-código do algoritmo de MacNaughton-Smith será apresentado abaixo.

Algorithm 2: Pseudo-código do algoritmo divisivo de MACNAUGHTON-SMITH

Entrada: Base de dados com **N** elementos

Saída : Um conjunto de grupos

j=1
repeat
  Selecionar o grupo $\mathbf{G_j}$ com maior número de elementos $\mathbf{N_j}$;
  Iniciar uma matriz $\mathbf{D_{nj}xD_{nj}}$;
  Calcular a similaridade média $\mathbf{S_m}$ de cada elemento do grupo $\mathbf{G_j}$ em relação aos demais;
  while $\mathbf{S_m} > 0$ **do**
    Remover o elemento **e** com maior $\mathbf{S_m}$ do grupo $\mathbf{G_j}$;
    Armazenar o elemento **e** no grupo $\mathbf{F_j}$;
    (re)Calcular a similaridade média $\mathbf{S_i}$ entre os elementos que restaram no grupo $\mathbf{G_j}$;
    (re)Calcular a similaridade média $\mathbf{S_a}$ entre cada elemento do grupo $\mathbf{G_j}$ e o grupo $\mathbf{F_j}$;
    $\mathbf{S_m} = \mathbf{S_i} - \mathbf{S_a}$;
  end
  j=j+1
until *restarem apenas grupos com dois elementos*;
repeat
  Selecionar o grupo **H** com maior similaridade média;
  Dividir o grupo **H**;
until *que todos grupos sejam divididos*;

2.2 Aprendizado de Máquina Multirrótulo

Tanto o aprendizado de máquina supervisionado quanto o não supervisionado podem ser aplicados nas mais diversas áreas de conhecimento, e nos mais diversos tipos de problemas existentes no mundo real. Como problemas do mundo real, podemos citar a simulação de situações de emergência, jogos, biomedicina, dentre outros [13]. Entretanto, dentre essas áreas de aplicação, há problemas nos quais mais de um rótulo são associados aos exemplos utilizados como treinamento. Exemplos desse tipo de problema são associação de rótulos a imagens (uma imagem pode ter associado vários nomes indicando diferentes objetos na imagem) [27], associação de palavras-chaves a documentos textuais [26, 25], associação de anotações a vídeos [11], associação de gêneros musicais a músicas [18], predição de falhas a equipamentos [5], dentre outros. Para esse tipo de problema, pode ser interessante induzir um classificador que rotule novos exemplos com mais de um rótulo. Para induzir modelos com essa característica, deve ser utilizado o aprendizado de máquina multirrótulo. Para um algoritmo de aprendizado de um modelo multirrótulo, a entrada consiste de um conjunto de exemplos S, com N objetos rotulados, ou exemplos, $T_i, i = 1, ..., N$, escolhidos de um domínio X com uma distribuição $\mathcal{D}$ fixa, desconhecida e arbitrária, da forma $\{(\mathbf{x}_i, Y_i), ..., \mathbf{x}_N, Y_N)\}$. L é o conjunto de rótulos possíveis do domínio D, e $Y_i \subseteq L$, ou seja, Y_i é o conjunto de rótulos associado ao i-ésimo objeto. A saída de um algoritmo de aprendizado supervisionado de modelos multirrótulos é um classificador $\mathbf{h}$, que classifica um exemplo $\mathbf{x_i}$ com o conjunto $\mathbf{Z_i} = \mathbf{h}(\mathbf{x_i})$, o qual é o conjunto de classes preditas por $\mathbf{h}\{\mathbf{x}\}$ para o exemplo $\mathbf{x_i}$. Há diversos métodos propostos na literatura para indução de modelos multirrótulo [31, 6, 1, 10].

2.3 Métodos de Aprendizado Multirrótulo

Existem duas categorias principais que podemos agrupar os métodos de aprendizado multirrótulo [28]. Apesar de serem métodos multirrótulo, a regra para realizar essa categorização está associada ao modo que algoritmos de classificação monorrótulo são utilizados [30].

A primeira categoria é chamada de *Transformação de Problema*. O próprio nome é intuitivo e sugestivo, nessa categoria, problemas de classificação multirrótulo são transformados em um ou mais problemas de classificação monorrótulo [8]. A partir dessa transformação, o processo de classificação é executado da mesma maneira que em problemas de classificação monorrótulo [22]. Isto é, para cada problema monorrótulo transformado, são utilizados algoritmos de classificação monorrótulo para resolvê-los [8]. Métodos pertencentes a categoria de transformação de problema também são chamados de métodos independente de algoritmo [22]. Já a segunda categoria, *Adaptação de algoritmo*, nenhuma transformação é realizada [8]. Portanto, o problema multirrótulo é tratado diretamente, através de modificações em algoritmos existentes[22]. Essa abordagem de criação de métodos específicos para tratar problemas multirrótulo é chamada de abordagem dependente de algoritmo [13].

A seguir serão apresentados alguns exemplos de métodos pertencentes às ambas categorias.

2.3.1 Métodos de Transformação de Problema

Métodos de transformação de problema se utilizam da abordagem independente de algoritmo. Portanto, para resolver o problema pode ser utilizado qualquer algoritmo tradicional. O processo para resolução do problema é feito através da realizacão da transformacão do problema multirrótulo original em um conjunto de problemas de classificação monorrótulo. Essa transformação pode ser baseada em dois tipos[13]:

1 - Baseada nos Rotulos das Classes;
2 - Baseada nos Exemplos;

No tipo de transformação baseada nos rótulos das classes, tomando um problema onde k é o numero de classes, e k classificadores são utilizados. Cada classificador é associado a uma classe e treinado para resolver um problema de classifição binária [13]. Um popular método de transformação de problemas baseado nos rótulos das classes é o *Binary Relevance*(BR)[30]. Por outro lado, no segundo tipo de transformação, o resultado do método não é somente problemas de classificão binária, mas também pode ser produzido problemas multiclasse. Porquanto na transformação baseada nos exemplos, o conjunto de classes associado a cada exemplo é redefinido [13]. Um exemplo de método desse tipo amplamente utilizado é o *Label Powerset* (LP)[30].

A tabela 2.3.1 apresenta um exemplo de conjunto de dados multirrótulo, que serve como base para exibir um exemplo de execução dos métodos BR e LP.

	Y
X_1	$Y_1 = \{y_2, y_3\}$
X_2	$Y_2 = \{y_1, y_3, y_4\}$
X_3	$Y_3 = \{y_4\}$
X_4	$Y_4 = \{y_2, y_3\}$

Tabela 2.1: Exemplo de conjunto de dados multirrótulo

O **Binary Relevance (BR)** é um popular método de transformação de problema, que decompõe um problema de classificação multirrótulo em vários diferentes problemas de classificação binária monorrótulo, um para cada q rótulos diferente no conjunto original S [9]. Por conseguinte são aprendidos pelo BR q classificadores binários. O método BR transforma o conjunto de dados S em $|L|$ conjunto de dados $S_j, j = 1, ..., |L|$ que contém todos os exemplos do conjunto de dados original, rotulados positivamente se o conjunto de rótulos do exemplo original continha o rótulo j e negativamente caso contrário [32].

A tabela TX apresenta o processo de transformação realizado pelo método BR em quatro conjunto de dados monorrotulo.

[fazer tabela nativo do latex]

Para classificação de uma nova instância, BR dá como saída a união dos rótulos λ, que são positivamente preditos pelos q classificadores [32]. O Binary Relevance é muito utilizado e apresenta resutados satisfatórios para diversos problemas, entretanto, sua limitação é nao levar em conta as informações de relacionamento entre rótulos, isto é, a dependência de rótulos [8].

	Y
E_1	$\neg y_1$
E_2	y_1
E_3	$\neg y_1$
E_4	$\neg y_1$

	Y
E_1	y_2
E_2	$\neg y_2$
E_3	$\neg y_2$
E_4	y_2

	Y
E_1	y_3
E_2	y_3
E_3	$\neg y_3$
E_4	y_3

	Y
E_1	$\neg y_4$
E_2	y_4
E_3	y_4
E_4	$\neg y_4$

Figura 2.1: Tabelas monorrótulo resultantes da aplicação do método BR no conjunto multirrótulo ilustrado na Tabela 2.2

Já o **Label Powerset (LP)**, transforma um problema multirrótulo em um problema de classificação multiclasse monorrótulo, onde os possíveis valores para atributos da classe de transformação são o conjunto de únicos e distintos subconjuntos de rótulos presentes no conjunto de treino original. A aprendizagem a partir de exemplos multirrótulo correspondem em encontrar um mapeamento a partir do espaço de características dos conjuntos de rótulos, ou seja, o poder dos conjuntos de todos os rótulos.

A tabela 2.3.1 apresenta o resultado da execução do método LP tornando um conjunto de dados multirrótulo num monorrótulo multiclasse.

	Y
X_1	$y_{2,3}$
X_2	$y_{1,3,4}$
X_3	y_4
X_4	$y_{2,3}$

Tabela 2.2: Conjunto monorrótulo multiclasse resultante da aplicação do método LP no conjunto multirrótulo ilustrado na Tabela 2.3.1

Dado um novo exemplo, o classificador monorrótulo do LP dá como saída a classe mais provável, que é um conjunto de rótulos [22]. Logo, a determinação da classe mais provável é dado de acordo com o valor de predição obtido através de um algoritmo multiclasse treinado com um conjunto de exemplos que foi gerado a partir da transfomação do problema [8]. Diferentemente do metodo BR o método Label Powerset leva em consideração a dependência de rótulos [22]. Uma limitação do LP se dá ao aumento exponencial do número de possíveis conjuntos de dados, por consequência de considerar dependências dos rótulos durante a classificação, quando uma quantidade grande ou moderada de rótulos são considerados. LP pode ter sua performance comprometida caso existirem classes no conjunto de treino que representem muito poucos exemplos. Quando isso de fato ocorre, é conhecido como problema de desbalanciamento de classe [9].

2.3.2 Métodos de Adaptação de Algoritmo

Métodos de adaptação de algoritmo se utilizam da abordagem dependente de algoritmo. Logo, com o intuito de tratar os problemas de classificação multirrótulo diretamente, como um todo, novos algoritmos são propostos. Por se tratar de um algoritmo específico, pode apresentar resultados melhores do que métodos que seguem abordagem independente de algoritmo para um determinado problema de classificação

real [13]. Um exemplo de método de adaptação de algoritmo proposto é o **Hierarchy Of Multi-label classifERs (HOMER)** [29].

O HOMER utiliza a técnica de projeto de algoritmos, dividir e conquistar, com isso, HOMER constrói uma hierarquia de classificadores multirrótulo, onde cada um lida com um conjunto de rótulos muito menor comparado com o conjunto total de rótulos, por conseguinte, um maior balanceamento na distribuição dos exemplos [29]. Na figura 2.2 é apresentado um exemplo simples da hierarquia constrída pelo HOMER para uma tarefa de classificação multirrótulo com 8 rótulos.

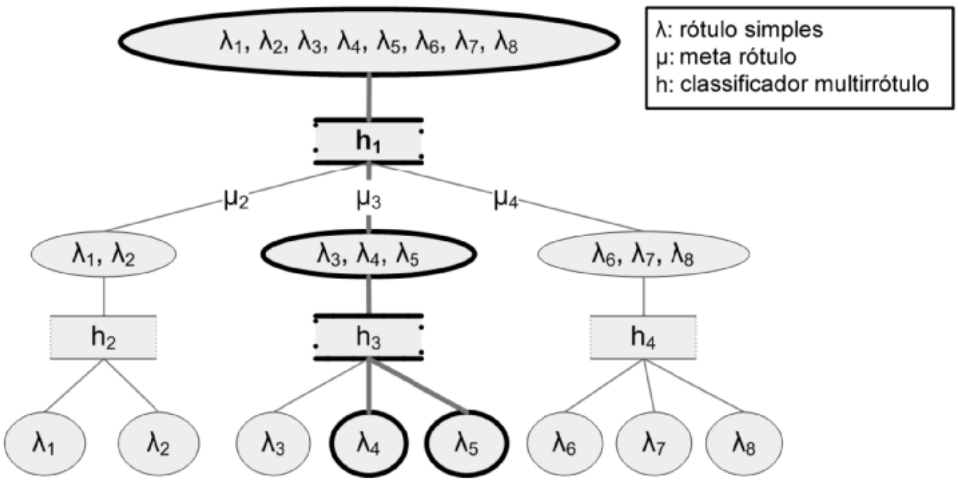

Figura 2.2: Hierarquia de rotulos e classificadores construida pelo HOMER [29]

Um dos principais processos internos do HOMER é a distribuição uniforme de um conjunto de rótulos em subconjuntos disjuntos k de modo que os rótulos semelhantes são colocados juntos e os rótulos que não possuem nenhuma semalhança são colocados aparte. Para execução dessa tarefa o HOMER utiliza um algoritmo chamado de *balanced k-Means* [29]. O algoritmo *balanced k-Means* é uma extensao do algoritmo *k-Means* [20], logo o que difere os dois algoritmos é a utlizacao de uma constante explícita referente ao tamanho de cada cluster e a limitação do número de iterações usando um parâmetro especificado pelo usuário. A peça chave desse algoritmo é que para cada cluster i é mantido uma lista de rotulos, C_i, ordenada em ordem ascendente da distancia dos rótulos para o centroid do cluster c_i. Quando uma inserção de um rótulo na lista ordenada de cluster feita na posição correta, ocasionar em um estouro do tamanho máximo de rótulos permitidos, na lista de rótulos desse cluster, é selecionado o último rótulo, no caso o mais distante. Esse rótulo selecionado é inserido na lista do próximo cluster mais próximo. Isto pode levar a $k-1$ inserções adicionais em cascata no pior caso [29]. O pseudo-código do algoritmo *balanced k-Means* é apresentado a seguir.

Algorithm 3: Algoritmo Balanced k-Means

Entrada: número de cluster k, rótulos L_n, dados de rótulos W_i, iterações it

Saída : k clusters balanceados de rótulos

```
for i ← 1 to k do
    //inicializa clusters e seus centros;
    C_i ← 0;
    c_i ← membro aleatório de L_n;
end
while it > 0 do
    for each λ ∈ L_n do
        for i ← 1 to k do
            d_λi ← distancia(λ, c_i, W_i);
        end
        finalizado ← false;
        v ← λ;
        while not finalizado do
            j ← arg min d_vi;
            Insere ordena(v, d_v) na lista ordenada C_j;
            if |C_j| > [|L_n|/k] then
                v ← remove último elemento de C_j;
                d_vj ← ∞;
            else
                finalizado ← true;
            end
        end
    end
    recalcula centros;
    it ← it − 1;
end
return C_1, ..., C_k;
```

O método HOMER foi construído para lidar com problemas de classifição multirrótulo com domínios que possuam muitos rótulos de maneira efetiva e computacionalmente eficiente. A partir dessa estratégia de hierarquia de classificadores, o HOMER melhora a predição com complexidade linear para treinamento, e logarítma para teste no que se refere a quantidade total de rótulos [29].

2.4 Características de Conjuntos de Dados Multirrótulo

Os conjuntos de dados não são todos igualmente multirrótulo. Em alguns casos, o número de classes de cada exemplo é pequeno se comparado ao número total de exemplos n, enquanto em outros, esse número é grande. Esse número pode ser parâmetro que influencia o desempenho dos diferentes métodos de classificação multirrótulo. Existem duas medidas para avaliar as características de um conjunto de dados: cardinalidade $Card(S)$ e densidade $Dens(S)$ [31].

A cardinalidade de um conjunto de dados multirrótulo S – $Card(S)$ – é dada pelo número médio de rótulos dos exemplos $T_i \in S$, e é independente do número de possíveis rótulos $|L|$ – Equação 2.11. Essa medida é utilizada para quantificar o número de rótulos alternativos que caracterizam os exemplos de um conjunto de dados multirrótulo.

A densidade de um conjunto de dados multirrótulo S – $Dens(S)$ – é dada pelo número médio de rótulos dos exemplos que pertencem a S dividido pelo número total de rótulos $|L|$ – Equação 2.12. A densidade de rótulo leva o número de possíveis rótulos em consideração.

$$Card(S) = \frac{1}{N}\sum_{i=1}^{N}|Y_i| \tag{2.11}$$

$$Dens(S) = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i|}{|L|} \tag{2.12}$$

2.5 Avaliação de Algoritmos de Aprendizado Multirrótulo

Para avaliar os classificadores multirrótulo, existem três grupos de medidas para avaliação induzida: baseada em instâncias, baseadas em rótulos e baseadas em ranking [11]. Neste trabalho, somente são usados os primeiros dois grupos de medidas, pois ranking multirrótulo não é o foco desse trabalho. Do primero grupo, são usadas nesse trabalho *Hamming Loss* (Ham), *Subset Accuracy* (SA) e F, definidas pelas equações 2.13 à 2.15[2], respectivamente. Do segundo grupo, são usados as versões micro e macro da medida $F1$. Medidas baseada em rótulos são calculadas baseadas em falso positivos f_p, falso negativos f_n, verdadeiro positivos t_p e verdadeiro negativos t_n, *i.e.*, medidas do tipo $B(t_p, t_n, f_p, f_n)$ podem ser usadas nesse caso. Dado que t_{p_l}, t_{n_l}, f_{p_l} e f_{n_l} são verdadero positivos, verdadeiro negativos, falso positivos e falso negativos para cada rótulo $l \in L$, a versão micro das medidas B é denotada por B_- é dada pela Eq. 2.16, enquanto que a versão macro das medidas B é denotada por B^- é dada pela Eq. 2.17. In this work, we only use $F1$ as a B measure, and $F1(t_p, t_n, f_p, f_n)$ is given by Eq. 2.18.

[2]Na Eq. 2.13, Δ representa a difereça simétrica entre dois conjuntos.

$$Ham(\mathbf{h}, S) = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \Delta Z_i|}{|L|} \tag{2.13}$$

$$SA(\mathbf{h}, S) = \frac{1}{N} \sum_{i=1}^{N} I(Z_i = Y_i) \tag{2.14}$$

$$F(\mathbf{h}, S) = \frac{1}{N} \sum_{i=1}^{N} \frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|} \tag{2.15}$$

$$B_{-}(\mathbf{h}, S) = \frac{1}{|L|} \sum_{i=1}^{|L|} B(t_{p_i}, t_{n_i}, f_{p_i}, f_{n_i}) \tag{2.16}$$

$$B^{-}(\mathbf{h}, S) = \frac{1}{|L|} B(\sum_{i=1}^{|L|} t_{p_i}, \sum_{i=1}^{|L|} t_{n_i}, \sum_{i=1}^{|L|} f_{p_i}, \sum_{i=1}^{|L|} f_{n_i}) \tag{2.17}$$

$$F1(t_p, t_n, f_p, f_n) = \frac{2 \times f_p}{2 \times t_p + f_n + f_p} \tag{2.18}$$

Capítulo 3

Proposta de um Método de Aprendizado Multirrótulo baseado em Aprendizado de Máquina Não-Supervisionado Hierárquico

Capítulo 4

Experimentos Realizados

Capítulo 5

Conclusão

Capítulo 6

Trabalhos Futuros

Referências Bibliográficas

[1] Everton Alvares-Cherman, Jean Metz, and Maria Carolina Monard. Incorporating label dependency into the binary relevance framework for multi-label classification. *Expert Systems with Applications*, 39(2):1647–1655, 2012.

[2] Michael R. ANDERBERG. Cluster analysis for applications. 1973.

[3] Pavel Berkhin. Survey of clustering data mining techniques, 2002. *Accrue Software: San Jose, CA*, 2004.

[4] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

[5] Flavia Cristina Bernardini, Ana Cristina Bicharra Garcia, and Inhaúma Neves Ferraz. Artificial intelligence based methods to support motor pump multi-failure diagnostic. *Engineering Intelligent Systems*, 17(2), 2009.

[6] Kassio Novaes Calembo, Flavia Cristina Bernardini, and Carlos Bazilio Martins. Proposta de um método de combinaçao de classificadores para construçao de classificadores multi-rótulo. In *Conferência Latinoamericana de Informática—CLEI*, volume 2011, 2011.

[7] Alexandre Xavier Ywata Carvalho, Pedro Henrique Melo Albuquerque, Gilberto Rezende de Almeida Junior, Rafael Dantas Guimarães, and Camilo Rey Laureto. Clusterização hierárquica espacial com atributos binários. 2009.

[8] Everton Alvares Cherman. *Aprendizado de máquina multirrótulo: explorando a dependência de rótulos e o aprendizado ativo*. PhD thesis, Universidade de São Paulo, 2013.

[9] Everton Alvares Cherman, Maria Carolina Monard, and Jean Metz. Multi-label problem transformation methods: a case study. *CLEI Electronic Journal*, 14(1):4–4, 2011.

[10] Patrıcia Pachiega da Gama, Flavia C Bernardini, and Bianca Zadrozny. Rb: A new method for constructing multi-label classifiers based on random selection and bagging. *Learning and Nonlinear Models*, 11(1), 2013.

[11] Anastasios Dimou, Grigorios Tsoumakas, Vasileios Mezaris, Ioannis Kompatsiaris, and Ioannis Vlahavas. An empirical study of multi-label learning methods for video annotation. In *2009 Seventh International Workshop on Content-Based Multimedia Indexing*, pages 19–24. IEEE, 2009.

[12] Marcelo Viana Doni. Análise de cluster: Métodos hierárquicos e de particionamento. Trabalho de Graduação, Universidade Presbiteriana Mackenzie., 2004.

[13] Katti Faceli, Ana Carolina Lorena, João Gama, and ACPLF Carvalho. Inteligência artificial: Uma abordagem de aprendizado de máquina. *Rio de Janeiro: LTC*, 2011.

[14] Zoubin Ghahramani. Unsupervised learning. In Olivier Bousquet, Ulrike von Luxburg, and Gunnar Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 72–112. Springer, 2003.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Unsupervised learning. In *The elements of statistical learning*, pages 485–585. Springer, 2009.

[16] Istvan Karoly Kasznar, Bento Mario Lages Gonçalves, and ML Bento. Técnicas de agrupamento clustering. *Revista Científica e Tecnológica*, 2009.

[17] Leonard Kaufman and Peter J Rousseeuw. Finding groups in data. an introduction to cluster analysis. *Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, New York: Wiley, 1990*, 1, 1990.

[18] Hanna M Lukashevich, Jakob Abeßer, Christian Dittmar, and Holger Grossmann. From multi-labeling to multi-domain-labeling: A novel two-dimensional approach to music genre classification. In *Proc. 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 2009.

[19] P Macnaughton-Smith, WT Williams, MB Dale, and LG Mockett. Dissimilarity analysis: a new technique of hierarchical sub-division. 1964.

[20] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[21] Ricardo Marcondes Marcacini. Aprendizado de máquina com informação privilegiada: abordagens para agrupamento hierárquico de textos. Tese de Doutorado, ICMC/USP.

[22] Hiteshri Modi and Mahesh Panchal. Experimental comparison of different problem transformation methods for multi-label classification using meka. *International Journal of Computer Applications*, 59(15), 2012.

[23] Bruno Magalhães Nogueira. Avaliação de métodos não-supervisionados de seleção de atributos para mineração de textos, March 27 2009.

[24] Lior Rokach. A survey of clustering algorithms. In *Data mining and knowledge discovery handbook*, pages 269–298. Springer, 2009.

[25] Robert E Schapire and Yoram Singer. Boostexter: A boosting-based system for text categorization. *Machine learning*, 39(2-3):135–168, 2000.

[26] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[27] Xipeng Shen, Matthew Boutell, Jiebo Luo, and Christopher Brown. Multilabel machine learning and its application to semantic scene classification. In *International Symposium on Electronic Imaging*, pages 18–22, 2004.

[28] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. In *International Journal Data Warehousing and Mining*, volume 3, page 1–13, 2007.

[29] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Effective and efficient multilabel classification in domains with large number of labels. In *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*, pages 30–44, 2008.

[30] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 1–19. Springer, 2009.

[31] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In *In: (Orgs.) Data Mining and Knowledge Discovery Handbook, 2nd ed.* Springer, 2010.

[32] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079–1089, 2011.

[33] Osmar R. Zaïane, Andrew Foss, Chi-Hoon Lee, and Weinan Wang. On data clustering analysis: Scalability, constraints, and validation. In Ming-Shan Cheng, Philip S. Yu, and Bing Liu 0001, editors, *PAKDD*, volume 2336 of *Lecture Notes in Computer Science*, pages 28–39. Springer, 2002.