



UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
CENTRO DE CIÊNCIAS EXATAS E DA TERRA
DEPARTAMENTO DE INFORMÁTICA E MATEMÁTICA APLICADA
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E COMPUTAÇÃO

USO DE CONFIABILIDADE NA ROTULAÇÃO DE EXEMPLOS EM PROBLEMAS DE
CLASSIFICAÇÃO MULTIRRÓTULO COM APRENDIZADO
SEMISSUPERVISIONADO

Fillipe Morais Rodrigues

Natal – RN

2014

Fillipe Morais Rodrigues

**USO DE CONFIABILIDADE NA ROTULAÇÃO DE EXEMPLOS EM PROBLEMAS DE
CLASSIFICAÇÃO MULTIRRÓTULO COM APRENDIZADO
SEMISSUPERVISIONADO**

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Sistemas e Computação da UFRN (área de concentração: Aprendizado de Máquina e Inteligência Computacional) como parte dos requisitos para a obtenção do grau de Mestre.

Orientador:

Prof. ^a Dr. ^a Anne Magaly de Paula Canuto

Natal – RN

2014

Fillipe Morais Rodrigues

**Uso de confiabilidade na rotulação de exemplos em problemas de classificação
multirrótulo com aprendizado semissupervisionado**

Dissertação de Mestrado apresentada ao
Programa de Pós-Graduação em Sistemas e
Computação da UFRN (área de concentração:
Aprendizado de Máquina e Inteligência
Computacional) como parte dos requisitos para a
obtenção do grau de Mestre.

Prof. ^a Dr. ^a Anne Magaly de Paula Canuto

Banca Examinadora:

Prof. ^a Dr. ^a Anne Magaly de Paula Canuto

Prof. Dr. João Carlos Xavier Júnior

Prof. Dr. Araken de Medeiros Santos

Prof. ^a Dr. ^a Flávia Cristina Bernardini

Defendido em: ____/____/____

À minha família.

Agradecimentos

A Deus, por todas as bençãos e graças que sempre me concede.

À minha esposa, por todo o carinho, apoio e compreensão. Você foi fundamental para essa conquista. Obrigado por sempre estar ao meu lado, por ser minha grande incentivadora e por ter confiado no meu potencial.

A toda a minha família, que tanto sabe do meu esforço para conquistar mais esse objetivo, principalmente aos meus pais, pelo exemplo de vida, determinação e vitória na vida. Vocês são a minha maior inspiração.

A professora Dra. Anne Magaly pela orientação neste trabalho, sempre com muita atenção, paciência e disponibilidade para esclarecer minhas dúvidas. Agradeço, demais, por acreditar na minha capacidade, por ter me motivado quando precisei e pela oportunidade de estudar mestrado em uma universidade federal.

Ao amigo, co-orientador e professor Dr. Araken Medeiros por todo o suporte prestado, pelos inúmeros e-mails respondidos e toda a satisfação em colaborar para a construção desse trabalho.

Ao amigo Daniel Alencar, por TUDO. Para você, meu amigo, serei eternamente grato.

Aos amigos do IFRN, pela torcida, apoio e amizade. Sempre agradecerei pelas palavras positivas para superar a distância de casa.

Enfim, agradeço a todos que colaboraram direta e indiretamente para a realização desse sonho.

“Se você não puder se destacar pelo talento, vença pelo esforço.”

Dave Weinbaum

Resumo

As técnicas de Aprendizado de Máquina são aplicadas em tarefas de classificação para a aquisição de conhecimento através de um conjunto de dados ou informações. Alguns métodos de aprendizado utilizados pela literatura são baseados em aprendizado semissupervisionado; este é representado por pequeno percentual de exemplos rotulados (aprendizado supervisionado) combinados com uma quantidade de exemplos rotulados e não rotulados (não-supervisionado) durante a fase de treinamento, reduzindo, portanto, a necessidade de uma grande quantidade de dados rotulados quando apenas um pequeno conjunto de exemplos rotulados está disponível para treinamento. O problema da escolha aleatória das instâncias é comum no aprendizado semissupervisionado, pois a maioria dos trabalhos usam a escolha aleatória dessas instâncias o que pode causar um impacto negativo. Por outro lado, grande parte dos métodos de aprendizado de máquina trata de problemas unirrótulo, ou seja, problemas onde exemplos de um determinado conjunto são associados a uma única classe. Entretanto, diante da necessidade existente de classificar dados em uma grande quantidade de domínios, ou em mais de uma classe, essa classificação citada é denominada classificação multirrótulo. Este trabalho apresenta uma análise experimental dos resultados obtidos por meio da utilização do aprendizado semissupervisionado em problemas de classificação multirrótulo usando um parâmetro de confiabilidade como auxílio na classificação dos dados. Dessa maneira, a utilização de técnicas de aprendizado semissupervisionado, bem como de métodos de classificação multirrótulos, foram imprescindíveis na apresentação dos resultados.

Palavras Chave: Aprendizado de Máquina; Aprendizado Semissupervisionado; Classificação Multirrótulo; Parâmetro de Confiabilidade.

Abstract

The techniques of Machine Learning are applied in classification tasks to acquire knowledge through a set of data or information. Some learning methods proposed in literature are methods based on semissupervised learning; this is represented by small percentage of labeled data (supervised learning) combined with a quantity of label and non-labeled examples (unsupervised learning) during the training phase, which reduces, therefore, the need for a large quantity of labeled instances when only small dataset of labeled instances is available for training. A common problem in semi-supervised learning is as random selection of instances, since most of paper use a random selection technique which can cause a negative impact. Much of machine learning methods treat single-label problems, in other words, problems where a given set of data are associated with a single class; however, through the requirement existent to classify data in a lot of domain, or more than one class, this classification as called multi-label classification. This work presents an experimental analysis of the results obtained using semissupervised learning in troubles of multi-label classification using reliability parameter as an aid in the classification data. Thus, the use of techniques of semissupervised learning and besides methods of multi-label classification, were essential to show the results.

Keywords: Machine Learning; Semissupervised learning; Multi-label classification; Reliability Parameter.

Sumário

Lista de Figuras	ix
Lista de Tabelas	x
Lista de Abreviaturas e Siglas	xi
Lista de Algoritmos	xiii
1. Introdução.....	1
1.1 Motivação.....	3
1.2 Objetivos	4
1.3 Organização do Trabalho	5
2. Fundamentos Teóricos	7
2.1 Conceitos Básicos de Aprendizado de Máquina.....	8
2.2 Métodos de Adaptação de algoritmo	10
2.2.1 KNN.....	11
2.2.2 SVM.....	12
2.3 Tipos de Aprendizado de Máquina quanto ao grau de supervisão.....	13
2.3.1 Aprendizado Supervisionado.....	14
2.3.2 Aprendizado Não-supervisionado	16
2.3.3 Aprendizado Semissupervisionado.....	17
2.4 Classificação de Dados	20
2.5 Classificação de dados multirrótulos.....	23
2.5.1 Métodos de transformação do problema.....	25
2.5.1.1 <i>BINARY RELEVANCE</i>	25
2.5.1.2 <i>LABEL POWERSET</i>	27
2.5.1.3 <i>RAKEL e RAKELd</i>	29
2.6 Classificação de dados multirrótulos semissupervisionado	31
2.6.1 SSBR.....	32
2.6.2 SSLP	35
2.6.3 SSRAKEL.....	36
2.6.4 SSRAKELd.....	37
2.7 Considerações Finais.....	38
3. Trabalhos Relacionados	39

3.1 Classificação Multirrótulo.....	39
3.2 Aprendizado Semissupervisionado	42
3.3 Aprendizado Semissupervisionado com Classificação Multirrótulo	44
3.7 Considerações Finais.....	44
4. Métodos Propostos	45
4.1 Algoritmo Semissupervisionado Adaptado.....	46
4.2 Fator de Confiabilidade.....	49
4.3 Métodos Semissupervisionados Propostos	49
4.3.1 Modelos SSBR propostos.....	50
4.3.2 Modelos SSLP propostos.....	54
4.3.3 Modelos SSRAKEL propostos.....	55
4.3.4 Modelos SSRAKELd propostos.....	56
4.4 Considerações Finais.....	57
5. Metodologia dos Experimentos.....	58
5.1 Bases de Dados Multirrótulo.....	58
5.1.1 Emotions.....	59
5.1.2 Scene.....	59
5.1.3 Genbase.....	59
5.1.4 Medical	59
5.1.5 Yeast	59
5.1.6 Bibtex.....	60
5.1.7 CAL500	60
5.1.8 Corel5k	60
5.1.9 Enron.....	60
5.1.10 Mediamill.....	61
5.2 Métricas de Avaliação.....	61
5.2.1 Basesadas em Bipartição	62
5.2.1.1 Hamming Loss	62
5.2.1.2 F-Measure.....	62
5.2.1.3 Accuracy.....	63
5.2.2 Basesadas em Ranking	63
5.2.2.1 One-Error	63
5.2.2.2 Average Precision	64

5.2.2.3 Ranking Loss.....	64
5.3 Parâmetro de Confiabilidade.....	64
5.4 Redução de Dimensionalidade – Feature Selection.....	65
5.5 Configuração dos Experimentos	67
5.6 Considerações Finais.....	68
6. Resultados Experimentais	69
6.1 Percentual de 25% dados rotulados iniciais	69
6.1.1 Métodos baseados no SSBR	70
6.1.2 Métodos baseados no SSLP.....	72
6.1.3 Métodos baseados no SSRAkEL	74
6.1.4 Métodos baseados no SSRAkELd	76
6.2 Percentual de 75% dados rotulados iniciais	78
6.2.1 Métodos baseados no SSBR	78
6.2.2 Métodos baseados no SSLP	80
6.2.3 Métodos baseados no SSRAkEL	82
6.2.4 Métodos baseados no SSRAkELd	84
6.3 Análise de desempenho entre métodos Semissupervisionados.....	86
6.4 Resultados Experimentais: Redução de Dimensionalidade	87
6.5 Análise Gráfica.....	90
6.6 Considerações Finais	92
7 Considerações Finais	93
7.1 Trabalhos Publicados	94
7.2 Trabalhos Futuros	94

Referências

Lista de Figuras

1.1	Caracterização das Aplicações de Aprendizado de Máquina.....	1
2.1	Hierarquia do aprendizado indutivo	7
2.2	Impacto do valor de k no algoritmo k -NN.....	11
2.3	Comparação entre um classificador de margem máxima e pequena.....	13
2.4	Estrutura das tarefas de Aprendizado de Máquina	14
2.5	Estrutura das tarefas de Aprendizado de Máquina	15
2.6	Abordagem do Self-Training.....	19
3.1	Processo de Classificação	22
3.2	(a) Típico problema de classificação. (b) Problema de classif. multirrótulo.....	24
3.3	Criação do classificador multirrótulo no <i>BR</i>	26
3.4	Predição do classificador multirrótulo <i>BR</i>	27
3.5	Construção do classificador multirrótulo <i>LP</i>	28
3.6	Construção do classificador multirrótulo <i>RAkEL</i>	30
3.7	Construção do classificador multirrótulo <i>RAkELd</i>	31
5.1	Abordagem do Self-Trainig usando parâmetro de confibilidade	47
5.2	Processo ilustrativo da atuação do parâmetro de confibilidade.....	48
7.1	Análise dos métodos baseados no <i>SSBR</i>	71
7.2	Análise dos métodos baseados no <i>SSLP</i>	74
7.3	Análise dos métodos baseados no <i>SSRAkEL</i>	76
7.4	Análise dos métodos baseados no <i>SSRAkELd</i>	78
7.5	Análise dos métodos baseados no <i>SSBR</i>	80
7.6	Análise dos métodos baseados no <i>SSLP</i>	82
7.7	Análise dos métodos baseados no <i>SSRAkEL</i>	84
7.8	Análise dos métodos baseados no <i>SSRAkELd</i>	86
7.9	Análise de desempenho entre métodos semissupervisionados.....	87
7.10	Gráfico Radar para os Métodos Multirrótulo baseados no <i>BR</i>	91
7.11	Gráfico Radar para os Métodos Multirrótulo baseados no <i>LP</i>	91

Lista de Tabelas

2.1	Conjunto de exemplos no formato atributo-valor.....	16
3.1	Conjunto de dados para avaliar as condições do jogo de tênis.....	20
3.2	Exemplos de um conjunto de dados multirrótulos	23
3.3	Conjunto de exemplos multirrótulos	28
3.4	Conjunto de dados transformados usando <i>RAkEL</i>	30
6.1	Detalhamento da base de dados.....	58
6.2	Tabela de bases multirrótulos com dimensionalidade reduzida	66
7.1	Desempenho dos métodos Multirrótulos para cada <i>SSBR</i>	71
7.2	Desempenho dos métodos Multirrótulos para cada <i>SSLP</i>	73
7.3	Desempenho dos métodos Multirrótulos para cada <i>SSRAkEL</i>	75
7.4	Desempenho dos métodos Multirrótulos para cada <i>SSRAkELd</i>	77
7.5	Desempenho dos métodos Multirrótulos para cada <i>SSBR</i>	78
7.6	Desempenho dos métodos Multirrótulos para cada <i>SSLP</i>	81
7.7	Desempenho dos métodos Multirrótulos para cada <i>SSRAkEL</i>	83
7.8	Desempenho dos métodos Multirrótulos para cada <i>SSRAkELd</i>	85
7.9	Performance dos métodos Multirrótulos para todas as versões do BR	89
7.10	Performance dos métodos Multirrótulos para todas as versões do LP	90

Lista de Abreviaturas e Siglas

Acc	Accuracy
AD	Árvore de Decisão
AM	Aprendizado de Máquina
AvPrec	Average Precision
BR	Binary Relevance
CoAL	Co-Training with Active Learning
Cov	Coverage
C4.5	Algoritmo de AM para construção de Árvore de Decisão
ErrSS	Error Set Size
FM	F-Measure
HLoss	Hamming Loss
HOMER	Hierarchy of Multilabel Classifiers
IA	Inteligência Artificial
IsErr	Is-Error
K-NN	k-Nearest-Neighbor
LCS	Learning Classifier System
LP	Label Powerset
NB	Naive-Bayes
Prec	Precision
RAkEL	Random k-LabElsets
RAkELd	Random k-LabElsets disjoint
Rec	Recall
RLoss	Ranking Loss
SAcc	Subset Accuracy
SOM	Self-Organizing Maps
SSBR	Semi-Supervised Binary Relevance
SSBRc	Semi-Supervised Binary Relevance with confidence
SSBRr	Semi-Supervised Binary Relevance using ranking
SSBRcr	Semi-Supervised Binary Relevance with confidence using ranking
SSLP	Semi-Supervised Label Powerset
SSLPc	Semi-Supervised Label Powerset with confidence

SSLPr	Semi-Supervised Label Powerset using ranking
SSLPcr	Semi-Supervised Label Powerset with confidence using ranking
SSRAkEL	Semi-Supervised Random k-LabElsets
SSRAkELc	Semi-Supervised Random k-LabElsets with confidence
SSRAkELd	Semi-Supervised Random k-LabElsets disjoint
SSRAkELr	Semi-Supervised Random k-LabElsets using ranking
SSRAkELrd	Semi-Supervised Random k-LabElsets disjoint using ranking
SSRAkELdcr	Semi-Supervised Random k-LabElsets disjoint with confidence using ranking
SSRAkELcr	Semi-Supervised Random k-LabElsets with confidence using ranking
SVM	Support Vector Machines
1-Err	One-Error

Lista de Algoritmos

Algoritmo 1	Fluxo básico do <i>SSBR</i> na fase de treinamento	33
Algoritmo 2	Fluxo básico do <i>SSBRc</i> na fase de treinamento	34
Algoritmo 3	Fluxo básico do <i>SSLP</i> na fase de treinamento	35
Algoritmo 4	Fluxo básico do <i>SSLPc</i> na fase de treinamento	36
Algoritmo 5	Cálculo da confiabilidade	47
Algoritmo 6	Fluxo básico do <i>SSBRr</i> na fase de treinamento	51
Algoritmo 7	Fluxo básico do <i>SSBRcr</i> na fase de treinamento.....	53
Algoritmo 8	Fluxo básico do <i>SSLPr</i> na fase de treinamento.....	53
Algoritmo 9	Fluxo básico do <i>SSLPcr</i> na fase de treinamento.....	55

Capítulo 1

Introdução

Uma significativa quantidade de dados e informações vem sendo propagada de maneira acelerada no nosso universo cotidiano. Essa quantidade de dados armazenam informações, muitas vezes sigilosas, sobre algum domínio do qual foram extraídas. No fim da última década, houve uma expressiva preocupação quanto ao desenvolvimento de mecanismos computacionais capazes de processar e armazenar essas informações. Mas, o que se observou com acúmulo dessas informações foi a dificuldade em extrair o conhecimento desses dados. O que muitos estudiosos da área têm em mente hoje é o desenvolvimento de processos automáticos para a aquisição de conhecimento a partir desses dados e informações armazenadas.

Essas dificuldades tem incentivado o desenvolvimento de mecanismos de aprendizado de máquina que, entre outras coisas, permitem a elaboração automática de modelos a partir de um conjunto de dados. O Aprendizado de Máquina (AM) é uma área dedicada ao estudo e desenvolvimento de mecanismos computacionais que permitam o aperfeiçoamento do desempenho em alguma tarefa através de experiências acumuladas (MONARD, 2003; MITCHELL, 1997).



Figura 1.1 Caracterização das Aplicações de Aprendizado de Máquina, (SANTOS,2012).

Além disso, o aprendizado de máquina é representado por uma grande quantidade de algoritmos que propõem soluções ou modelos para diversos tipos de problemas. Para tal,

esses algoritmos são classificados de acordo com diferentes abordagens. Na Figura 1.1 é possível visualizar diferentes abordagens do aprendizado de máquina. No aprendizado supervisionado, também conhecido como classificação e regressão, durante a fase de treinamento os algoritmos recebem como entrada exemplos associados à informação de saída desejada, representando a classe cujo exemplo pertence. Por outro lado, o aprendizado não-supervisionado recebe como entrada exemplos sem a informação de saída desejada, ou seja, não é conhecido, a priori, a qual classe os exemplos do conjunto de treinamento pertencem.

Atualmente, a comunidade científica vem se dedicando aos estudos de uma terceira categoria de aprendizado, surgida através da junção entre o aprendizado supervisionado com o aprendizado não-supervisionado. Essa nova categoria de aprendizado é conhecida como aprendizado semissupervisionado (MATSUBARA; MONARD; BATISTA, 2005). Na sua fase de treinamento, o aprendizado semissupervisionado combina os dados rotulados, ou seja, exemplos que estão associados a uma determinada classe ou categoria e os dados não-rotulados, exemplos sem qualquer classe ou categoria associada, reduzindo, dessa maneira, a necessidade de uma grande quantidade de dados rotulados quando somente um pequeno conjunto de exemplos rotulados está disponível. Essa redução na necessidade de um grande número de dados rotulados é uma das vantagens dos métodos de aprendizado semissupervisionado (SANTOS, 2012).

Em aprendizado de máquina, muitos algoritmos têm como objetivo solucionar problemas de classificação de dados a partir de um conjunto de dados brutos, onde, através desses algoritmos, serão retiradas informações por meio de categorização (SANTOS, 2012). Na Figura 1.1, características quanto à quantidade de rótulos são definidas como unirrótulo ou multirrótulo. Uma abordagem comum e bastante utilizada na literatura é a classificação de dados unirrótulo que trata do aprendizado de um conjunto de exemplos onde cada um desses exemplos estão associados a somente uma classe. Por sua vez, em problemas de classificação multirrótulo, uma grande quantidade de exemplos podem estar associados a mais de uma classe simultaneamente (VALLIM, 2009).

Quanto aos níveis dos rótulos, segundo a literatura, as soluções cuja classificação das classes pertencem ao conjunto finito de classes e estão todas em um mesmo nível possuem, dessa maneira, uma independência entre as classes. Nesse tipo de classificação

não são considerados relacionamentos hierárquicos entre as classes, logo, este tipo de classificação é conhecida como classificação plana. Outrossim, em problemas de classificação mais complexos na qual as classes podem ser divididas em subclasses ou estarem agrupadas em superclasses, estes, por sua vez, são classificados segundo a literatura como problemas de classificação hierárquica.

Levando em consideração as vantagens do aprendizado semissupervisionado, bem como o uso de técnicas de classificação multirrótulo para a resolução de problemas de classificação, este trabalho tem como principal foco a aplicação de aprendizado semissupervisionado combinado a classificação multirrótulo.

1.1) Motivação

Na literatura já podem ser encontrados trabalhos que fazem uso do aprendizado semissupervisionado em problemas de classificação multirrótulo, tais como (METZ, 2011; SANTOS, 2012; RODRIGUES et al, 2013). Porém, alguns destes trabalhos não se preocuparam com o problema relativo ao aprendizado semissupervisionado, que é relacionado ao problema na escolha das instâncias não-rotuladas a serem rotuladas. Em (SANTOS, 2012), por exemplo, foi desenvolvido mecanismos que buscavam auxiliar, de maneira aleatória, na escolha dos exemplos não-rotulados. Porém, parte dos resultados apresentados por (SANTOS, 2012) não foram tão significante quanto estimados, o que, certamente, foi ocasionado pela aleatoriedade nas escolhas das instâncias no aprendizado semissupervisionado.

Outro problema abordado por este trabalho e que também serve como motivação é o problema da dimensionalidade de bases de dados, ou seja, bases de dados com grande quantidade de atributos. Por sua vez, alguns atributos podem ser irrelevantes e/ou redundantes e que, conseqüentemente, podem vir a prejudicar o desempenho durante um processo de aprendizado.

Portanto, a motivação principal para este trabalho é o fato de neste ser aplicada uma maneira mais sistemática na escolha das instâncias, almejando, portanto, alcançar um melhor desempenho a partir da combinação do aprendizado semissupervisionado com classificação multirrótulo. Os dados não-rotulados do conjunto de exemplos serão

classificados como rotulados a partir de um parâmetro que é chamado de confiabilidade onde, por sua vez, auxiliará na escolha desses exemplos no procesos de atribuição de pesos para os rótulos no aprendizado semissupervisionado. Diante da apresentação dos resultados favoráveis à utilização do aprendizado semissupervisionado e as técnicas de classificação multirrótulo que usam pesos no processo de rotulação de instâncias, surge, portanto, a motivação para a ideia desse novo mecanismo afim de minimizar o problema na escolha aleatória de dados. Além dessa abordagem inicial, será abordado nesse trabalho técnicas de redução de dimensionalidade como parâmetro auxiliar no processo de rotulação de instâncias, visto que a grande dimensionalidade de atributos, pode, conseqüentemente, vir a ser um fator complicador quanto a aquisição de bons resultados.

1.2) Objetivos

Como já mencionando, este trabalho tem como objetivo principal desenvolver mecanismos que fazem uso de um parâmetro de confiabilidade em problemas de classificação multirrótulo através da aplicações de técnicas de aprendizado semissupervisionado.

Nesse trabalho, os referidos objetivos a seguir deverão ser alcançados:

- Desenvolvimento e implementação de técnicas de classificação multirrótulo semissupervisionada: Consiste nesse trabalho o desenvolvimento e implementação de novas técnicas de classificação multirrótulos aplicadas em técnicas de aprendizado semissupervisionado, proporcionando, dessa maneira, novas contribuições para os estudos nessa área;
- Proporcionar aos diferentes problemas de classificação, uma aplicação dos modelos propostos: Observa-se a existência de vários problemas de classificação quando estes utilizam modelos de classificação supervisionado existentes, bem como os métodos semissupervisionados propostos. Desse modo, será desenvolvido um estudo comparativo entre as técnicas existentes e as técnicas propostas;
- Avaliar os resultados por meio de diferentes medidas específicas para classificação multirrótulo: Visto as características específicas envolvendo

problemas de classificação multirrótulos, certas medidas específicas de avaliação serão utilizadas para aferir o desempenho de cada classificador sobre cada tipo de problema. Por fim, a avaliação do desempenho dessas medidas é parte fundamental no objetivo do trabalho, pois busca-se a viabilidade da aplicação do aprendizado semissupervisionado em métodos de classificação multirrótulo com o intuito de aproximar os resultados obtidos dos resultados oriundos do aprendizado supervisionado.

1.3) Organização do Trabalho

Além deste Capítulo, essa tese de mestrado está dividida em mais sete Capítulos e organizada da seguinte maneira:

- Capítulo 2: apresenta conceitos básicos de aprendizado de máquina, dando ênfase aos seus algoritmos, bem como os diferentes tipos de aprendizado de máquina existentes. Além disso, esse capítulo apresenta as principais características e algoritmos existentes associados aos diferentes tipos de problemas de classificação de dados. Além disso, mostra a aplicação do aprendizado semissupervisionado em tipos de problemas de classificação de dados;
- Capítulo 3: apresenta os principais trabalhos envolvendo tarefas de classificação multirrótulo, bem como técnicas de aprendizado semissupervisionado;
- Capítulo 4: apresenta os algoritmos propostos para a elaboração de classificadores multirrótulos baseados em parâmetros de confiabilidade através do aprendizado semissupervisionado. Além disso, são apresentadas também as medidas de avaliação usadas neste trabalho;
- Capítulo 5: apresenta a metodologia dos experimentos, os parâmetros utilizados para todo o desenvolvimento do estudo na aquisição dos resultados, bem como as bases de dados multirrótulos usadas nesse experimento;
- Capítulo 6: apresenta os resultados experimentais obtidos a partir de bases de dados multirrótulos, bem como apresenta os resultados a partir de bases de dados multirrótulos reduzidas em sua dimensionalidade

- Capítulo 7: tem como intuito apresentar as considerações finais, trabalhos publicados e trabalhos futuros.

Capítulo 2

Fundamentos Teóricos

O Aprendizado de Máquina (AM) pode ser considerado uma área de pesquisa na Inteligência Artificial (IA) na qual tem o interesse em desenvolver técnicas computacionais a partir de um conhecimento prévio, além de construir sistemas capazes de adquirir conhecimento de forma automática (REZENDE, 2003). Em um modo geral, pode-se considerar diferentes tipos de aprendizado, tais como: indutivo, dedutivo, e o baseado em analogia (TAVARES et al., 2007). O aprendizado estudado atualmente pela literatura é o aprendizado baseado em conceitos de indução. Esse conceito considera fatos ou observações como possível generalização à conclusões, ou seja, esse conceito de indução é a maneira de inferência lógica que possibilita obter conclusões a partir de um conjunto particular de exemplos. Ela é caracterizada pelo raciocínio originado em conceito específico e generalizado, da parte para o todo (REZENDE, 2003). Dessa maneira, a indução vem demonstrando ser uma importante ferramenta na geração de novos conhecimentos. Na Figura 2.1 é ilustrada a hierarquia do aprendizado indutivo.

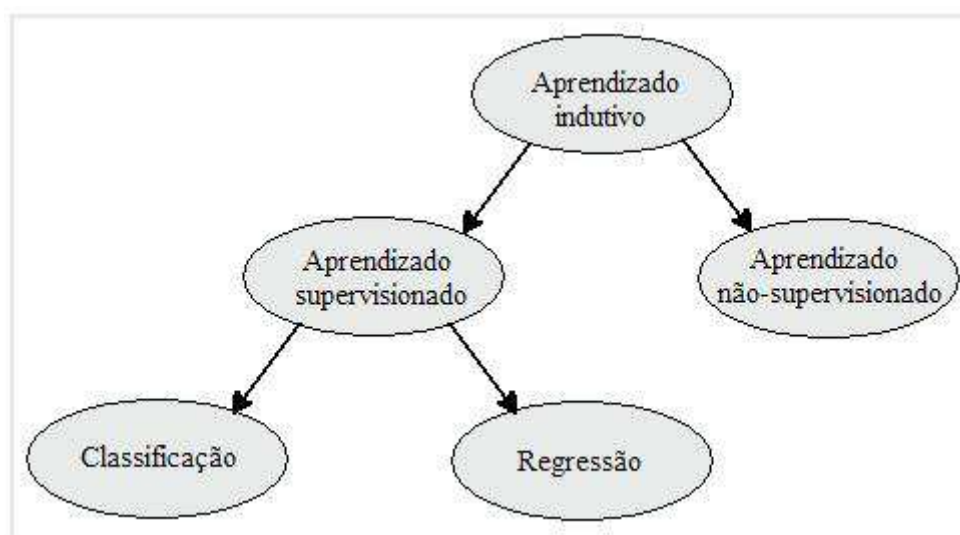


Figura 2.1 Hierarquia do aprendizado indutivo, (SANTOS, 2012).

Podemos observar que na Figura 2.1 o aprendizado indutivo decorre de dois modos de aprendizado distintos: aprendizado supervisionado ou aprendizado não-supervisionado.

No aprendizado supervisionado, um conjunto de dados de treinamento é fornecido ao algoritmo de aprendizado onde o rótulo da classe associada é conhecido. Isso no aprendizado não-supervisionado não ocorre. Apesar dos conceitos introdutórios de aprendizado de máquina, este trabalho apresentará melhor as aplicações do aprendizado supervisionado, semissupervisionado e não-supervisionado na Seção 2.3.

De acordo com a Figura 1.1, Aprendizado de Máquina Indutivo, são apresentadas as perspectivas na qual o aprendizado de máquina se subdivide: grau de supervisão, quantidade de rótulos e nível de rótulos. Quanto ao grau de supervisão, o aprendizado de máquina faz uso de técnicas de aprendizado supervisionado, não-supervisionado e semissupervisionado. Para a quantidade de rótulos, o aprendizado de máquina pode combinar suas técnicas com problemas de classificação unirrótulo ou multirrótulo. E, por fim, quanto ao nível de rótulos, o aprendizado de máquina organiza seus rótulos em um nível plano ou hierárquico. Na Seção 2.3 são apresentados os tipos de aprendizado quanto ao grau de supervisão.

Este capítulo está organizado do seguinte maneira: na Seção 2.1, são apresentados os conceitos básicos associados à área de aprendizado de máquina, bem como, no sub-item da Seção 2.1, serão apresentados os conceitos de alguns algoritmos de aprendizado de máquinas usados nesse trabalho. Já na Seção 2.2, trata-se das categorias de aprendizado indutivo existente acrescido do aprendizado semissupervisionado.

2.1) Conceitos básicos do Aprendizado de Máquina

Na literatura é possível encontrar várias definições sobre conceitos básicos de aprendizado de máquina. Nesta seção, serão apresentados os principais termos referentes à área de aprendizado de máquina, bem como suas definições e características. Alguns termos podem ser aplicados à grande maioria desses métodos. São eles:

- **Atributo:** também chamados campos ou variáveis, são utilizados para descrever características ou aspectos de um exemplo. Geralmente, os atributos se dividem em dois modelos: discretos ou contínuos. Atributos discretos contêm um número finito ou infinito contável de valores. Um especial tipo de atributo discreto é o atributo binário, ou booleano, que representa apenas dois valores de saída, como

0/1, sim/não, ausência/presença, verdadeiro/falso, entre outros. Já os atributos contínuos, por sua vez, podem ser representados por números infinitos de valores. Geralmente esses atributos contínuos são resultados de medidas e representados por números reais, como peso, tamanho e distância. (FACELLI et al., 2011).

- Classe ou Rótulo: também chamado de atributo-alvo, uma classe ou rótulo, descreve o conceito-meta, ou seja, o conceito que se deseja aprender para tornar viável a realização de predições a seu respeito (FACELLI et al., 2011). No aprendizado supervisionado, todo exemplo possui pelo menos um atributo especial denominado rótulo ou classe.
- Exemplo: um objeto único do mundo a partir do qual um modelo será aprendido ou usado, é também conhecido como padrão ou instância. Um exemplo descreve um objeto de uma determinada situação na qual se deseja aprender através de um vetor de valores de características ou atributos (MONARD, 2003). Além do vetor de características, um exemplo pode ter um atributo especial, que indica a que classe ou rótulo esse exemplo pertence.
- Vetor de características: uma lista de características que descreve um exemplo do conjunto de exemplos.
- Conjunto de exemplos: também conhecido como conjunto de dados ou conjunto de instâncias, é composto por um número de exemplos com seus respectivos valores de atributos.
- Conjunto de treinamento: utilizado como entrada pelos algoritmos de aprendizado. É a partir do conjunto treinamento que são construídos classificadores, portanto, ele deve ser representativo na distribuição da população dos dados do domínio.
- Conjunto de teste: utilizado para avaliar o modelo construído. Esse conjunto não deve ser apresentado ao algoritmo de aprendizado durante a elaboração do

modelo. O ideal é que o conjunto de testes não tenha exemplos em comum com os exemplos do conjunto de treinamento.

- Conjunto de validação: utilizado na realização de ajustes no modelo construído pelo algoritmo de aprendizado. Os exemplos do conjunto de teste não são utilizados diretamente na construção do modelo; são utilizados para o seu ajuste. Dessa forma, esses exemplos são perceptíveis de maneira indireta durante o processo de aprendizado, o que obriga aos exemplos de validação que sejam distintos dos exemplos de teste.
- Indutor: é o algoritmo de aprendizado que faz uso de um processo indutivo para gerar a sua hipótese ou modelo. Essa hipótese é utilizada para classificar uma instância ou exemplo.
- Classificador ou hipótese: dado um conjunto de exemplos do conjunto de treinamento, um indutor, ou algoritmo de aprendizado, gera como saída um classificador, ou hipótese, de forma que, dado um novo padrão, ele possa identificar sua classe com a maior precisão possível (MONARD, 2003).

2.2) Tipos de Aprendizado de Máquina quanto ao grau de supervisão

Inteligência é um termo utilizado para descrever as diversas capacidades da mente humana. Tal capacidade possibilita que o homem possa adquirir novos conhecimentos. Segundo, (Dietterich, 1990), aprendizado é definido como:

“Aprendizado é o aumento de conhecimento”

Assumindo essa ideia, o aprendiz humano já possui algum conhecimento previamente adquirido e agregará o novo conhecimento aprendido ao já existente. Na literatura, portanto, encontramos outras definições que explicam essa capacidade do ser humano em aprender. Segundo (Mitchell, 1997; FACELLI et al., 2011), define-se aprendizado de máquina como:

"A capacidade de melhorar o desempenho na realização de alguma tarefa por meio de experiência".

O aprendizado de máquina é, dentre outras coisas, associado à área de inteligência artificial. Entretanto, é válido destacar que o aprendizado de máquina se sub-divide em três tipos de aprendizado. O aprendizado indutivo, representado pelos aprendizados supervisionados e não-supervisionados (MONARD, 2003). A diferença entre o aprendizado supervisionado e o não-supervisionado é a forma como é realizado o processo de generalização do conhecimento, ou seja, o aprendizado supervisionado trabalha com conjuntos de exemplos rotulados e o não-supervisionado com conjuntos de exemplos não-rotulados. Uma nova categoria do aprendizado de máquina e já adotada na literatura é o aprendizado semissupervisionado. Essa nova categoria de aprendizagem surgiu através da junção das técnicas de aprendizado supervisionados com as técnicas de aprendizado não-supervisionado (MATSUBARA; MONARD; BATISTA, 2005).

A seguir, a Figura 2.2 mostra a estrutura das categorias em aprendizado de máquina indutivo:



Figura 2.2 Estrutura das tarefas de Aprendizado de Máquina, (SANTOS, 2012).

2.2.1) Aprendizado Supervisionado

Segundo (HAYKIN, 1999), o aprendizado supervisionado, também conhecido como aprendizagem com professor, é o tipo de aprendizado no qual se tem um conhecimento sobre o ambiente, ou seja, podemos considerar o professor como detentor desse conhecimento prévio e com este conhecimento sendo representado por um conjunto de exemplos de entrada-saída. Em outras palavras, o aprendizado supervisionado requer

algum tipo de supervisão no processo onde possa ser informado os exemplos típicos de cada classe, os quais, dentro de um processo de treinamento, é permitido identificar a classe de novos objetos.

A Figura 2.3 ilustra em diagrama geral do processo de aprendizado supervisionado para classificação, onde P representa o conjunto de exemplos de treinamento, μ representa um algoritmo de aprendizado aplicado a um conjunto de exemplos de treinamento P e, por conseguinte, uma hipótese h é gerada.



Figura 2.3 Estrutura das tarefas de Aprendizado de Máquina.

No aprendizado supervisionado, cada exemplo do conjunto de treinamento é associado, ou rotulado, a uma classe de origem. O rótulo associado a um exemplo, definido como saída desejada, atributo-alvo ou dependente, é fornecido junto aos demais atributos de entrada, estes definidos como atributos independentes. Desse maneira, na fase de treinamento, ao apresentar um exemplo ao algoritmo de aprendizado, além dos atributos que caracterizam o exemplo, um atributo especial também é apresentado, representando a classe à qual este exemplo pertence. Define-se, portanto, esse atributo especial como rótulo (SANTOS, 2012).

Na Tabela 2.1, apresenta-se o formato geral de um conjunto de exemplos com atributos rotulados em N exemplos P_i com $i = 1, \dots, N$, na forma $\{(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n)\}$ para alguma função desconhecida de $y = f(x)$. Os $\vec{x}_i$ são tipicamente vetores da forma $\langle x_{i1}, x_{i2}, \dots, x_{in} \rangle$ com valores discretos ou contínuos. Dessa maneira, x_{ij} refere-se ao valor definido pelo atributo j , chamado X_j , do exemplo P_i . Os valores definidos por y_j referem-se ao valor do atributo Y , representando a classe ou rótulo (SANTOS, 2012).

Baseando-se nessa ideia, o processo de treinamento é finalizado e um novo exemplo não-rotulado é apresentado ao modelo. Tais algoritmos poderão, a partir das experiências adquiridas anteriormente por meio do processo de treinamento, classificar uma classe

para esse novo exemplo, ou seja, rotular esse exemplo como pertencente a uma das classes possíveis, de acordo com o conjunto de dados.

	X_1	X_2	$\dots$	X_M	Y_i
E_1	x_{11}	x_{12}	$\dots$	x_{1M}	Y_1
E_2	x_{21}	x_{22}	$\dots$	x_{2M}	Y_2
$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
E_n	x_{N1}	x_{N2}	$\dots$	x_{NM}	Y_n

Tabela 2.1 Conjunto de exemplos no formato atributo-valor.

Na literatura, há vários métodos de aprendizado supervisionado propostos, dentre os quais podemos destacar: k vizinhos mais próximos (k -NN, do inglês, *k-Nearest Neighbour*) (COVER; HART, 1967), algoritmos de indução de árvore de decisão (DT, do inglês, *Decision Tree*) (QUINLAN, 1993), máquina de vetores suporte (SVM, do inglês, *Support Vector Machines*) (VAPNIK, 1995, 1998; HAYKIN, 1998; CRISTIANINI; SHAW-ETAYLOR, 2000), além de vários modelos de redes neurais artificiais, tais como MLP (do inglês, *Multilayer Perceptron*) (HAYKIN, 1998) e RBF (do inglês, *Radial Basis Function*) (HAYKIN, 1998), dentre outros.

2.2.2) Aprendizado Não-Supervisionado

O Aprendizado não-supervisionado trabalha com conjuntos de exemplos onde o atributo classe é desconhecido. Durante a fase de treinamento, os métodos recebem como entrada exemplos sem a informação de saída, ou seja, não se conhece a princípio a classe a qual as instâncias do conjunto de exemplos pertencem. Assim, a ideia dos métodos de aprendizado não-supervisionado é de aproximar exemplos não-rotulados por meio da identificação de padrões, originando *clusters* distintos, auxiliando, dessa maneira, no entendimento desses dados (COSTA, 1999).

O processo de agrupamento dos algoritmos de aprendizado não-supervisionado pode ser realizado através de duas abordagens: através de uma medida de distância ou de alguma medida de correlação. No agrupamento através de uma medida de distância, os exemplos tendem a se aproximar de outros mais próximos. Outras medidas de distância também podem ser utilizadas como, por exemplo, a distância Euclidiana e a distância de

Mahalanobis (MAHALANOBIS, 1936). Por outro lado, no agrupamento realizado através da medida de correlação, destaca-se a correlação de Pearson (PEARSON, 1901) como a medida de correlação mais conhecida. Ao utilizar essas medidas, seja de distância ou de correlação, o surgimento desses grupos são ocasionados através de uma distribuição linear apresentada entre cada exemplo.

Dessa forma, pode-se dizer que os exemplos pertencentes a um mesmo grupo podem apresentar um alto grau de semelhança entre si do que com exemplos em grupos diferentes. Como exemplo de métodos de aprendizado não-supervisionado pode ser citado, dentre outros: redes neurais do tipo Mapa Auto-Organizáveis (SOM, do inglês, *Self-Organizing Maps*) (KOHONEN, 1982), *k*-Médias (do inglês, *k-Means*) (JAIN; DUBES, 1988) e Agrupamento Hierárquico (SANGER, 1989).

2.2.3) Aprendizado Semissupervisionado

A construção de um modelo de classificação por meio de algoritmos de aprendizado supervisionado muitas vezes somente é possível através da existência de uma quantidade expressiva de exemplos rotulados, ou seja, instâncias do conjunto de dados onde a classe é conhecida.

Todavia, algumas técnicas de atribuição de rótulos de exemplos podem ser tarefas custosas e requerem auxílios de um especialista. Essa técnica é conhecida como atribuição de rótulos manual. A atribuição de rótulos manual, além de outras características, depende da disponibilidade e conhecimento prévio do domínio por um especialista para executar manualmente os exemplos de treinamento.

Devido às limitações apresentadas pelo aprendizado supervisionado, vários pesquisadores buscam alternativas para rotular os exemplos de treinamento durante o processo de aprendizado, quando a disponibilidade dos números de exemplos rotulados é pequena. Esse modelo estudado por pesquisadores recebe o nome de “aprendizado semissupervisionado”. A ideia principal das técnicas de aprendizado de máquina dessa abordagem é combinar, dentro de um conjunto de exemplos, os poucos exemplos rotulados e os não- rotulados durante o processo de aprendizado (METZ, 2011).

Conforme mencionado, situações onde existam exemplos cujos rótulos das classes são conhecidos previamente, é possível, portanto, usar a técnica de aprendizado supervisionado para induzir classificadores a partir desses exemplos. Em contrapartida, por não se conhecer previamente os rótulos das classes dos exemplos, utiliza-se o aprendizado não-supervisionado com o objetivo de encontrar os *clusters*. Entretanto, há situações em que o conjunto de exemplos possui uma quantidade pequena de exemplos rotulados, enquanto a quantidade de exemplos não-rotulados é alta; nessa circunstância, pode-se utilizar o aprendizado semissupervisionado.

Uma das ideias principais do aprendizado semissupervisionado é a preocupação em rotular mais exemplos e, dessa maneira, incrementar esse exemplos rotulados ao conjunto de dados de treinamento. Assim, esse conjunto de dados de treinamento incrementado será apresentado a um algoritmo de aprendizado supervisionado para que este, por sua vez, possa induzir classificadores mais precisos. Dentre as vantagens na aplicação da técnica de aprendizado semissupervisionado, destaca-se o potencial em reduzir a necessidade de um grande percentual de dados rotulados, em circunstâncias onde apenas um pequeno grupo de exemplos rotulados esteja disponível (SANTOS, 2012).

Alguns exemplos das técnicas de aprendizado semissupervisionado utilizadas são o *Self-Training* (ROSENBERG; HEBERT; SCHNEIDERMAN, 2005) e o *Co-Training* (BLUM; MITCHELL, 1998). A ideia do *Self-Training* é, primeiramente, criar um classificador a partir de uma pequena quantidade de exemplos rotulados. Após isso, o classificador é usado para classificar os exemplos não-rotulados do conjunto de dados. Dessa maneira, os exemplos não-rotulados, bem como os rótulos preditos anteriormente pelo classificador, serão incluídos no conjunto de treinamento. Cada classificador usa suas próprias previsões para se auto-orientar. Por fim, ao término do processo, o classificador será treinado outra vez repetindo todo o processo. Por outro lado, o *Co-Training* (BLUM; MITCHELL, 1998) é uma abordagem do aprendizado semissupervisionado baseada na ideia da multidescrição, ou seja, os exemplos de treinamento são descritos por dois ou mais conjuntos de atributos disjuntos. Na Figura 2.4 é apresentada a estrutura procedural da abordagem *Self-Training*.

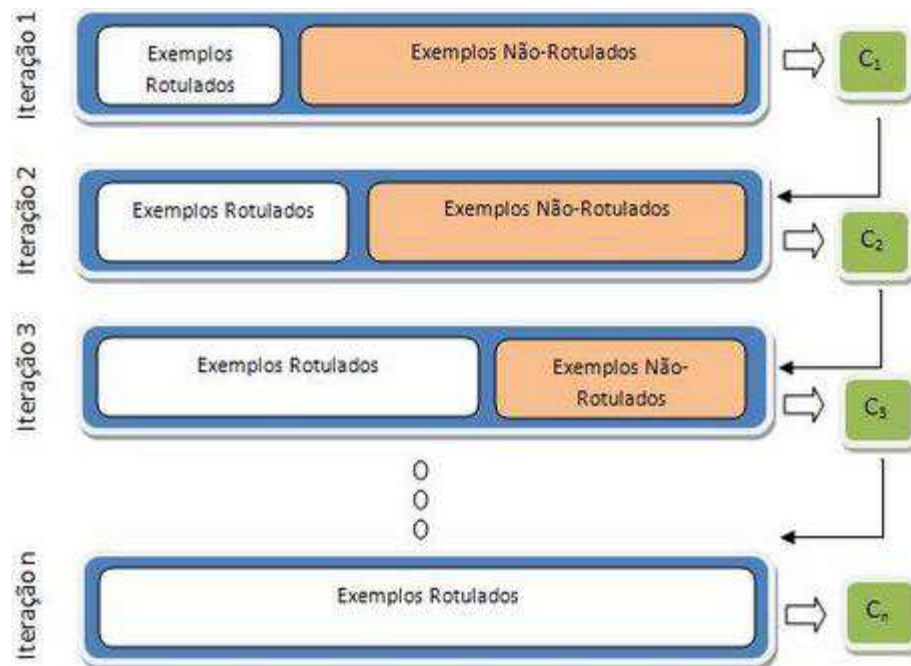


Figura 2.4 Abordagem do *Self-Training*, (SANTOS, 2012).

No *Co-Training*, dois classificadores são criados inicialmente, cada classificador usa uma descrição diferente dos exemplos rotulados. Em seguida, é usado para cada classificador, o mesmo algoritmo de aprendizado supervisionado para induzir as duas hipóteses. Será gerado, portanto, um classificador a partir dos mesmos exemplos, mas cada classificador é induzido através de diferentes descrições. Por fim, os exemplos não-rotulados serão apresentados aos dois classificadores que, por sua vez, terão suas saídas combinadas. Os exemplos com maior grau de confiança de classificação serão incluídos ao conjunto de exemplos rotulados e, a partir de então, dois novos conjuntos classificadores são criados e todo o processo se repetirá (SANTOS, 2012).

2.3) Métodos de Adaptação de algoritmo

Para este trabalho, serão aplicados métodos de adaptação, cujos resultados gerados são de uma extensão de um determinado algoritmo de aprendizado unirrótulo, possibilitando assim a manipulação de dados multirrótulo diretamente. Na literatura, há uma variedade de trabalhos que tratam da adaptação das mais variadas técnicas de aprendizado de máquina, como por exemplo: Algoritmos baseados na teoria do aprendizado estatístico, as máquinas de vetores de suporte e o algoritmo do k vizinhos mais próximos (ZHANG; ZHOU, 2007).

2.3.1) KNN

O k -NN (k -Nearest-Neighbor), ou o algoritmo dos k vizinhos mais próximos, é uma técnica de aprendizado de máquina baseado em instâncias. No k -NN, as bases de dados são fornecidas como entrada durante a etapa de treinamento e, também, usadas para avaliar a distância de cada instância para a instância que está sendo classificada. Utiliza-se, portanto, uma medida de distância entre as instâncias do conjunto de instâncias como parâmetro do algoritmo. Uma instância desconhecida é classificada através da análise das distâncias da instância desconhecida para as demais e selecionadas as k instâncias mais próximas dela, sendo k um parâmetro do algoritmo. A classe da instância é definida através do processo de voto majoritário, onde cada voto é o rótulo das k instâncias mais próximas (WITTEN; FRANK, 2005).

Na Figura 2.5, por exemplo, deve ser considerado um problema no qual é tratado um conjunto de dados onde será classificado um objeto como pertencente à classe "doente" ou classe "saudável". Para $k = 3$, o objeto de teste será caracterizado como pertencente à classe "doente", enquanto o $k = 5$ o objeto de teste será classificado como pertencente à classe "saudável".

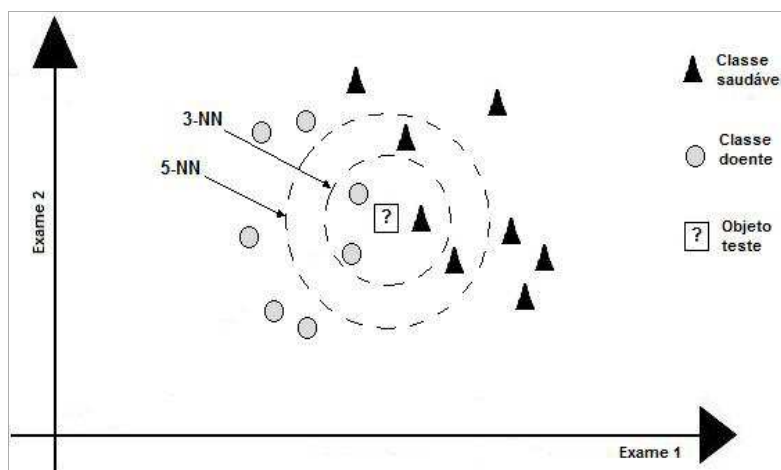


Figura 2.5 Impacto do valor de k no algoritmo k -NN, (CERRI, 2010 – adaptado).

Em (LUO; ZINCIR-HEYWOOD, 2005), são apresentados dois sistemas de classificação de documentos multirrótulo, baseado no algoritmo de aprendizado k -NN. A contribuição principal deste trabalho é no estágio de pré-processamento para a efetiva representação dos documentos. No momento da classificação de uma nova instância, o sistema

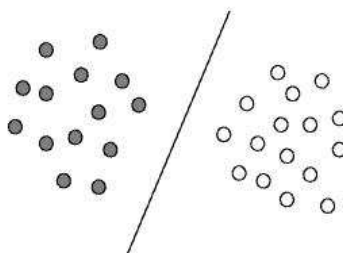
inicializa a busca por exemplos mais próximos. Portanto, a cada manifestação de um rótulo em cada um destes exemplos, um contador correspondente a este rótulo é incrementado (TSOUMAKAS; KATAKIS, 2007).

Como mencionado anteriormente sobre a aplicação dos métodos de adaptação de algoritmos, o k -NN foi adaptado a técnicas de classificação multirrótulo possibilitando, assim, a manipulação de dados multirrótulo diretamente.

2.3.2) SVM

Baseadas na teoria do aprendizado estatístico, as máquinas de vetores de suporte (*support vector machines*), ou SVMs, vêm se destacando na comunidade de aprendizado de máquina nos últimos anos. Os resultados obtidos através dessa técnica são comparáveis e muitas vezes superiores aos resultados conquistados por outros algoritmos populares de aprendizado, por exemplo, os resultados obtidos em uma Rede Neural.

Os algoritmos de aprendizado de máquina baseados em SVM podem ser definidos em um hiperplano no espaço de exemplos que tenta separar os exemplos que estão em treinamento das classes positivas e negativas. Além disso, o hiperplano deve estar o mais distante possível do conjunto de exemplo de treinamento, criando o que se chama de “*margem máxima*”, ver Figura 2.6. Essa técnica foi nominada dessa maneira por causa dos exemplos mais próximos do hiperplano, chamados de vetores de suporte (*support vectors*). A minimização de erros de classificação dos exemplos, bem como a maximização da margem são argumentos que limitam os erros do classificador em exemplos não detectados na fase de treinamento (BRAGA, 2010).



Exemplo linearmente separável

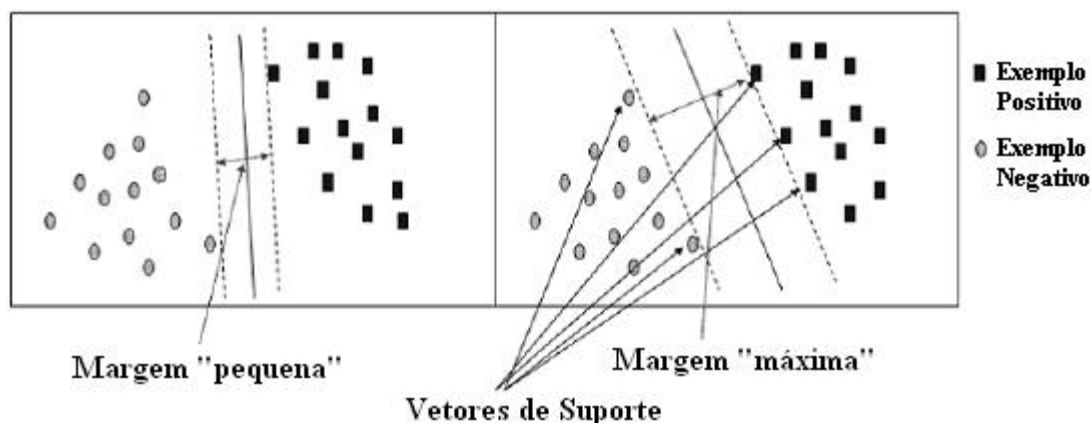


Figura 2.6 Comparação entre um classificador de margem máxima e pequena (BRAGA, 2010).

Em (ELISSEEFF; WESTON, 2001), é proposta uma variação do SVM para manipular dados multirrótulo. Neste algoritmo, busca-se minimizar o número de vezes que rótulos irrelevantes estão posicionados acima dos rótulos relevantes no *ranking*.

2.4) Classificação dos Dados

Em um processo de classificação de dados, a partir de um conjunto de dados brutos, é necessário que as informações sejam extraídas através de categorização, ou seja, através da atribuição de características chamada de rótulos. A partir dessa extração é possível classificar ou rotular os dados de tal maneira que esses rótulos confirmem informações aos dados categorizados pelo mesmo rótulo (COSTA, 2008).

Nos últimos anos, o interesse em estudos voltados para classificação de dados vem aumentando consideravelmente e ampliando, também, pesquisa através do aprendizado de máquina. Várias dessas pesquisas proporcionaram novos conhecimentos sobre os mais diversos tipos de classificação em aprendizado de máquina, tais como em (CANUTO et al., 2009; ALMEIDA; LUDERMIR, 2010; LORENA; CARVALHO, 2010; OLIVEIRA; CANUTO; SOUTO, 2007; SANTANA; CANUTO; ABREU, 2006).

Os problemas de classificação de dados são subconjunto dos problemas de aprendizado supervisionado. Dessa maneira, a fase de classificação ocorrerá por meio de um algoritmo de aprendizado, na qual seu objetivo será gerar um classificador a partir de dados de entrada rotulados.

Tabela 3.1 Conjunto de dados para avaliar as condições do jogo de tênis.

<i>Exemplo</i>	<i>Tempo</i>	<i>Temperatura</i>	<i>Umidade</i>	<i>Vento</i>	<i>Jogar Tênis</i>
<i>Ex1</i>	ensolarado	quente	alta	fraco	não
<i>Ex2</i>	ensolarado	quente	alta	forte	não
<i>Ex3</i>	nublado	quente	alta	fraco	sim
<i>Ex4</i>	chuvoso	regular	alta	fraco	sim
<i>Ex5</i>	chuvoso	fria	normal	fraco	sim
<i>Ex6</i>	chuvoso	fria	normal	forte	não
<i>Ex7</i>	nublado	fria	normal	fraco	sim
<i>Ex8</i>	ensolarado	fria	normal	fraco	sim
<i>Ex9</i>	ensolarado	fria	normal	fraco	sim
<i>Ex10</i>	chuvoso	regular	normal	forte	sim
<i>Ex11</i>	ensolarado	regular	normal	forte	sim
<i>Ex12</i>	nublado	regular	alta	forte	sim
<i>Ex13</i>	nublado	quente	normal	fraco	sim
<i>Ex14</i>	chuvoso	regular	alta	forte	não

Com base nos exemplos da Tabela 3.1, para a formação de um classificador, é necessário que um conjunto de exemplos da base de dados seja rotulado. Esse conjunto de exemplos rotulados também é conhecido como conjunto de treinamento. Cada exemplo desse conjunto de dados é caracterizado por atributos. Esses atributos de entrada estão associados a uma das propriedades de cada exemplo. Além desses atributos classificadores, também é fornecido o rótulo de cada exemplo. A Tabela 3.1 é representada por um conjunto de dados que avaliam as possíveis condições para um atleta jogar tênis em um determinado dia. Ou seja, cada linha representa um desses exemplos do conjunto de dados e as colunas representam o valor das variáveis de cada atributo. O último atributo, denominado atributo-alvo, definirá a qual rótulo esse exemplo pertencerá, ou seja, se aquele dia, com tais condições, terá possibilidade de jogo.

A construção de um classificador é tida juntamente com o início da fase de pré-processamento dos dados. Esta fase tem o intuito de transformar os dados brutos, na qual

são pertencentes ao domínio sobre o qual será aplicado o algoritmo de classificação, em dados classificados adequadamente, possibilitando a sua utilização na fase de treinamento. Concluída a fase de pré-processamento, é iniciada a fase de treinamento desses dados. Cada exemplo E_i , representado por um par $(\vec{x}_i, y_i)$, tal que $i = 1, 2, \dots, N$ é o número de exemplos de treinamento, $\vec{x}_i$ é vetor de atributos que descrevem cada exemplo e y_i é o valor do atributo-alvo Y , também conhecido como rótulo ou classe. O objetivo final da etapa de treinamento é rotular cada exemplo para sua respectiva classe. Tal procedimento é dado através de uma função de mapeamento, na qual é encontrada por meio de ajustes dos parâmetros livres do modelo, para mapear cada exemplo para seu rótulo correspondente (SANTOS, 2012).

Por fim, obtêm-se um classificador, após a fase de treinamento, que seja capaz de realizar o processo de predição da classe aos quais novos exemplos de entrada pertencerão; tal predição é dada pela generalização dos exemplos do conjunto de treinamento.

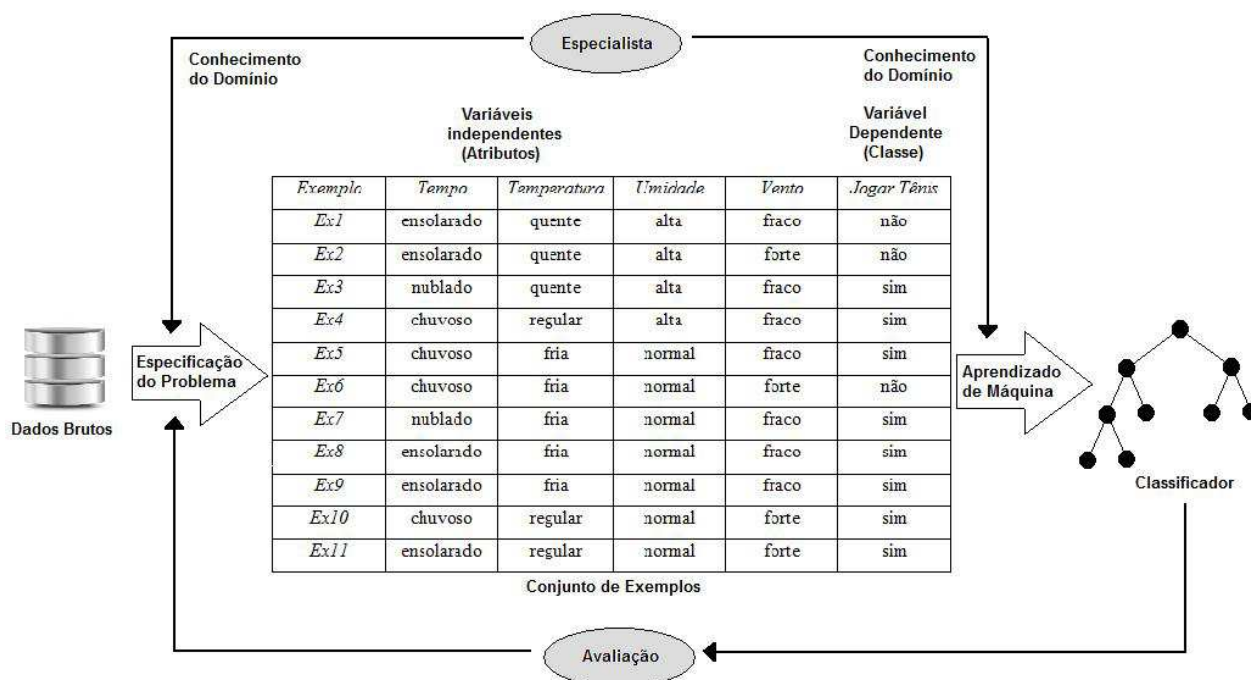


Figura 3.1 Processo de Classificação, (CERRI, 2010 – adaptado).

Segundo mostra a Figura 3.1, na fase de classificação, os dados são organizados em um conjunto de exemplos no qual são pertencentes ao domínio do problema. Serão aplicados sobre dados um algoritmo de classificação, para que os dados possam ser preparados para serem representados de forma adequada para o processamento. Os dados são organizados

e representados por seus atributos, sejam atributos de entrada para representar as características de cada exemplo (variáveis independentes) e induzir o classificador, ou pelos atributos de saída para representar as classes que são associadas aos exemplos (variável dependente).

Este capítulo é organizado da seguinte forma: a Seção 3.1 apresenta a classificação de dados multirrótulo e os principais algoritmos supervisionados. Na Seção 3.2, é apresentada uma recente área de classificação de dados, resultante da combinação da classificação multirrótulo com o aprendizado semissupervisionado.

2.5) Classificação de dados multirrótulos

O aprendizado supervisionado destaca-se por considerar que os exemplos (instâncias) de um conjunto de treinamento estejam associados a uma única classe (rótulo) de um conjunto de classes Y , onde $|Y| > 1$. Essa tarefa de classificação é chamada de unirrótulo (do inglês, *single-label classification*). Mesmo sabendo da grande utilidade da classificação unirrótulo, hoje existem diferentes domínios de aplicação nos quais os exemplos estão associados a um conjunto de classes L , tal que $L \subseteq Y$, ou seja, cada instância poderá estar associada a mais de um rótulo simultaneamente. Dá-se o nome para essa tarefa de classificação multirrótulo (do inglês, *multi-label classification*). (TSOUMAKAS; KATAKIS; VLAHAVAS, 2008). Alguns dos exemplos típicos de objetos que podem ser classificados a várias classes simultaneamente são, por exemplo, diagnóstico médico, artigos em jornais, e filmes. Em filmes, por exemplo, estes podem ser rotulados a partir das categorias que um exemplar pode estar associado.

Formalizando, seja $L = \lambda_j : j = 1, \dots, M$ um conjunto finito de rótulos em um aprendizado multirrótulo e $D = \{(\vec{x}_i, L_i), i = 1, \dots, N\}$ o conjunto de instâncias de treinamento multirrótulo, onde $\vec{x}_i$ é o vetor de características de cada instância e $L_i \subseteq Y$ o conjunto de classes da i -ésima instância (TSOUMAKAS; KATAKIS; VLAHAVAS, 2009).

Para exemplificar, vejamos o exemplo da Tabela 3.2. São apresentados cinco exemplos de filmes que estão associados a mais de uma classe simultaneamente. É apresentado para cada filme de um a quatro tipos de rótulos.

EXEMPLOS	CLASSES
Ex1	$C = \{y2, y4\}$
Ex2	$C = \{y1\}$
Ex3	$C = \{y1, y2\}$
Ex4	$C = \{y2, y3, y4\}$
Ex5	$C = \{y1, y3\}$

CLASSE	CATEGORIA
y1	TERROR
y2	COMÉDIA
y3	ROMANCE
y4	AÇÃO

Tabela 3.2 Exemplos de um conjunto de dados multirrótulos.

Grande parte das pesquisas sobre classificação de dados é voltada para a classificação de dados unirrótulos. Entretanto, nos últimos anos, houve um crescimento por parte dos pesquisadores pela área de classificação de dados multirrótulos. Tal motivação é perceptível diante das diversas pesquisas sobre esse tipo de classificação, por exemplo trabalhos na área de diagnósticos médicos, onde pacientes podem sofrer de mais de um tipo de doença ao mesmo tempo (TSOUMAKAS; KATAKIS, 2007); categorização de músicas através de emoções (LI; OGIHARA, 2003, 2006; WIEC-ZORKOWSKA; SYNAK; RA, 2006; TROHIDIS et al., 2008); classificação de imagens (BOUTELL et al., 2004); função genômica (BLOCKEL et al., 2006; BARUTCUOGLU; SCHAPIRE; TROYANSKAYA, 2006; ELISSEFF; WESTON, 2001; CLARE; KING, 2001); além de trabalhos nas áreas de Bioinformática (CLARE; KING, 2001; ZHANG; ZHOU, 2005; ELISSEFF; WESTON, 2001).

A Figura 3.2 apresenta uma comparação entre um problema de classificação convencional, onde os exemplos podem ser associados a apenas uma classe, e um problema de classificação multirrótulo, onde os exemplos podem pertencer a uma ou mais classes. A Figura 3.2 (a) ilustra um problema de classificação cujo documento pertence ou à classe "Diabetes" ou à classe "Câncer"; nunca em ambas as classes. Já na Figura 3.2 (b) é ilustrado um exemplo de classificação multirrótulo, ou seja, exemplos pertencentes simultaneamente às classes "Diabetes" e "Câncer". Esses exemplos multirrótulos são classificadas na classe "DiaCan".

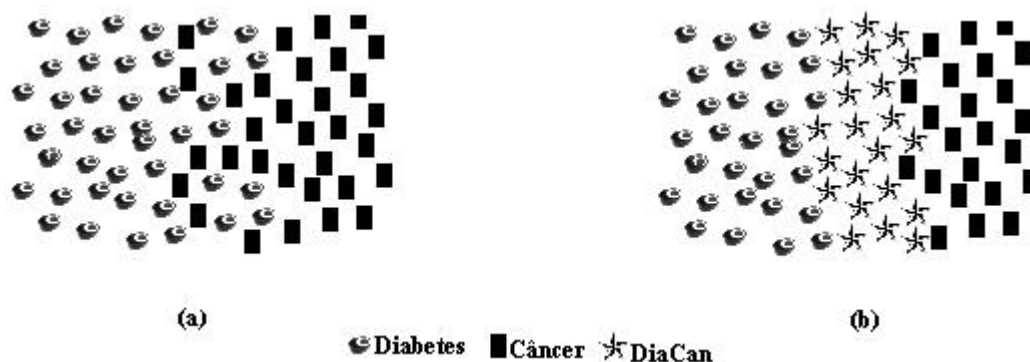


Figura 3.2 (a) Típico problema de classificação. (b) Problema de classificação multirrótulo, (CERRI, 2010 – adaptado).

Na literatura, algumas técnicas tratam problemas de classificação multirrótulo, outras tratam problemas, como classificação unirrótulo por exemplo, que podem ser combinadas com problemas de classificação multirrótulo. Além disso, outros algoritmos também podem surgir para atuar em problemas de classificação multirrótulo (CERRI, 2010). Em contrapartida, existem técnicas resultantes da modificação de algoritmos de classificação unirrótulo, de maneira que, através da adaptação de seus mecanismos internos, torna-se possível a sua utilização em problemas de classificação multirrótulo.

2.5.1) Métodos de transformação do Problema.

A ideia dos métodos de transformação do problema é decompor um problema multirrótulo existente em um ou mais problemas unirrótulos. Esses métodos utilizam algoritmos de classificação unirrótulo para resolução de cada problema de maneira distinta. É importante considerar a independência de cada método sobre o algoritmo base. Assim, há a possibilidade de qualquer algoritmo de aprendizado unirrótulo poder construir individualmente os classificadores.

Existem na literatura diversos métodos capazes de transformar problemas de classificação multirrótulos em problemas de classificação unirrótulo, dentre eles os métodos “*Binary Relevance*” “*Label Powerset*” e “*RAkEL*” (do inglês, *R*ANdom *k*-*Lab*ELsets) (TSOUMAKAS; KATAKIS; VLAHAVAS, 2010).

2.5.1.1) *BINARY RELEVANCE*

Em problemas multirrótulo, o método BR (BR, do inglês *Binary Relevance*) é um dos métodos de transformação mais popular (TSOUMAKAS; KATAKIS, 2007). É com o BR que é tida a visão de dividir os problemas em L problemas de classificação unirrótulo binário, sendo um para cada rótulo contido em L . Dessa maneira, o métodos considera que, de maneira independente, cada rótulo tem sua predição.

O conjunto de exemplos unirrótulos gerados a partir da divisão do problema multirrótulo em L problemas de classificação binária serão replicados e seus atuais rótulos serão modificados. Cada exemplo do conjunto replicado que contém o unirrótulo y_j será identificado como positivo; para os exemplos que não contém esse unirrótulo, estes serão classificados como negativos. A partir desses classificadores unirrótulos, uma nova instância tem sua predição multirrótulo formada pela união dos rótulos associados preditos positivamente pelos classificadores binários, ver Figuras 3.3 e 3.4.

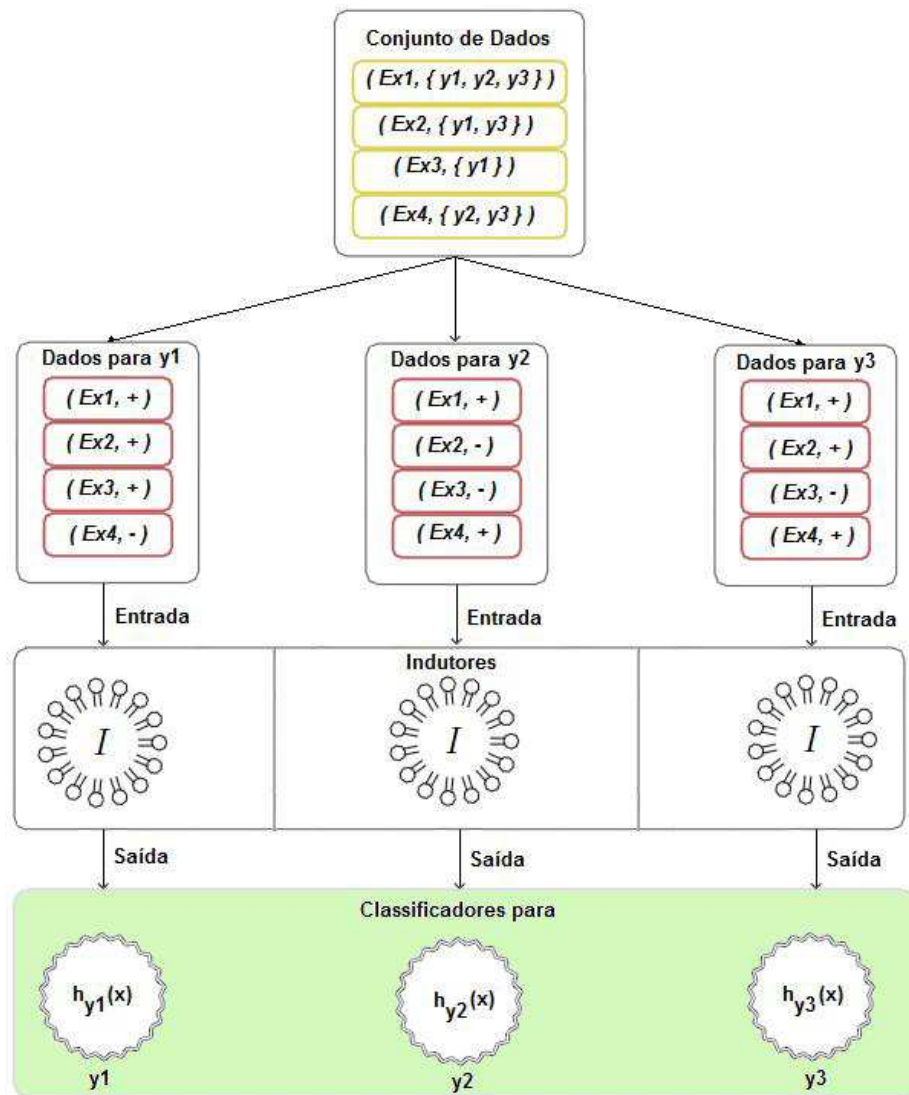


Figura 3.3 Criação do classificador multirrótulo no BR, (SANTOS, 2012 – adaptado).

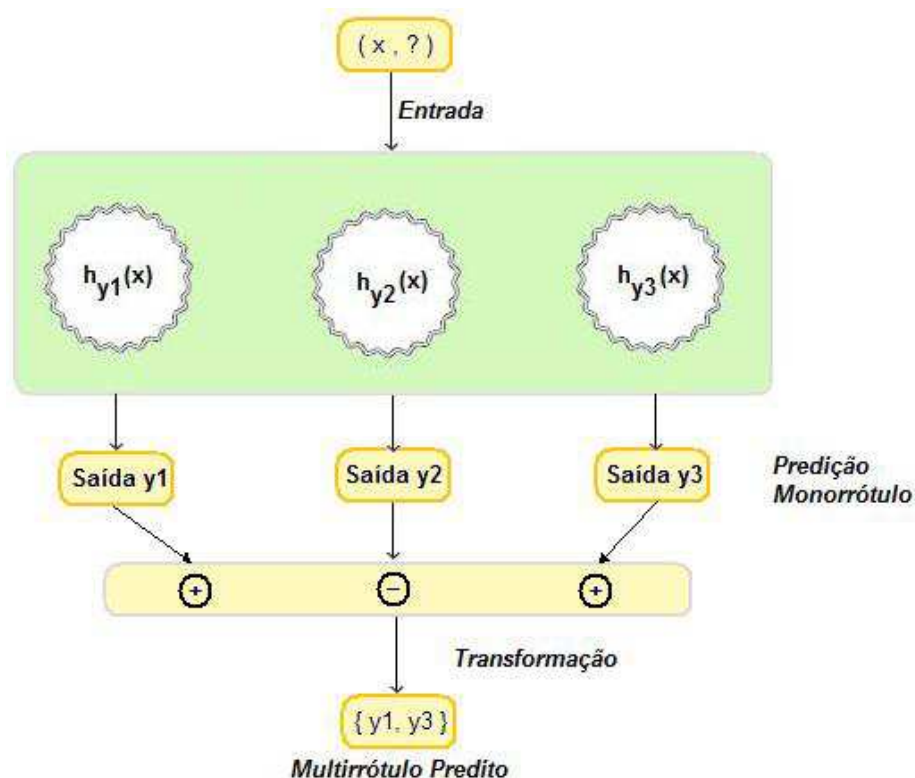


Figura 3.4 Predição do classificador multirrótulo BR, (SANTOS, 2012 – adaptado).

Uma das deficiências do método *Binary Relevance* é por não serem consideradas correlações entre instâncias (rótulos) durante a construção do modelo de classificação multirrótulo, proporcionando, de certa maneira, uma perda nas informações relevantes à tarefa de classificação. Além disso, esse método propicia um modelo de classificação com pouca generalização e legibilidade. Isso se deve à construção de um único modelo para cada rótulo correlacionado a um exemplo de teste (SANTOS, 2012).

2.5.1.2) LABEL POWERSET

O método LP (do inglês *Label Powerset*) procura transformar um problema multirrótulo em distintos subconjuntos do problema. Para tal feito, é criado, através dessa transformação, um conjunto unirrotulo multiclasse, na qual os subconjuntos, ou conjuntos únicos de rótulos, estão presentes no conjunto de treinamento e representados por valores unitários do atributo classe, a partir dos quais um classificador multiclasse é construído (METZ, 2011).

O exemplo apresentado pela Tabela 3.3 ilustra um conjunto multirrótulo, no qual é composto por quatro exemplos e $L = \{y_1, y_2, y_3\}$ como conjunto de unirrótulos.

Exemplo	Y_i
E_1	$Y_1 = \{y_1, y_2, y_3\}$
E_2	$Y_2 = \{y_1, y_3\}$
E_3	$Y_3 = \{y_1\}$
E_4	$Y_4 = \{y_2, y_3\}$

Tabela 3.3 Conjunto de exemplos multirrótulos.

No método *Label Powerset*, os exemplos multirrótulos são transformados em novos tipos de dados, agora unirrótulos. Essa transformação ocorre por meio da concatenação dos múltiplos rótulos para a criação de um novo unirrótulo. Essa transformação pode ser realizada através da criação de um índice único para cada agrupamento de dados multirrótulos no conjunto de treinamento. Por exemplo, o dado multirrótulo $\{y_1, y_3\}$ poderia estar associado a um índice apresentado pela letra “A”. A partir de então, cada vez que o unirrótulo “A” fosse utilizado, automaticamente seriam recuperados os múltiplos rótulos originais. Por fim, os exemplos multirrótulos transformados são apresentados a um algoritmo de aprendizado para a construção de um classificador unirrótulo multiclasse (METZ, 2011).

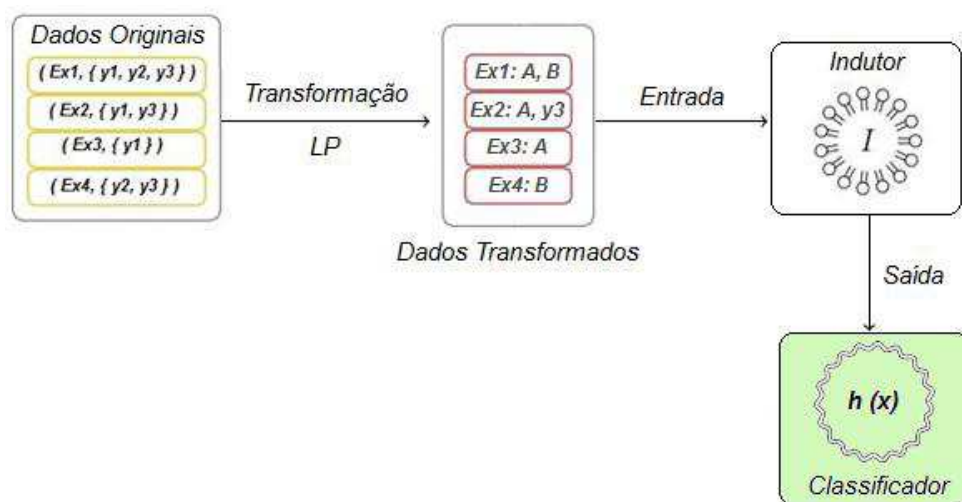


Figura 3.5 Construção do classificador multirrótulo LP, (SANTOS, 2012 – adaptado).

Ao contrário do método de classificação *Binary Relevance* (BR), o *Label Powerset* tem a vantagem de considerar as correlações entre os rótulos, ver Figura 3.5. Entretanto, pode

haver um crescimento desproporcional do número de rótulos de uma classe caso haja um grande número de subconjuntos de rótulos, resultando, dessa maneira, em muitas classes com poucos exemplos associados. Isso porventura ocasionará em um alto custo computacional do método *Label Powerset* e a diminuição da acurácia dos classificadores. Por outro lado, o método de classificação LP apresenta-se por ser um método simples, menos comum, mas bastante utilizado (SANTOS, 2012; TSOUMAKAS et al., 2008).

2.5.1.3) *RAkEL* e *RAkELd*

O método de classificação *RAkEL* (do inglês, *RANdom k-LabELsets*) é construído por meio de comitês de classificadores LP. Um dos seus objetivos é minimizar os problemas oriundos do LP, pois mesmo considerando a correlação entre os rótulos, busca-se coibir o problema da suscetibilidade ao surgimento de muitas classes com uma quantidade mínima de exemplos. Foi proposta uma abordagem em (TSOUMAKAS et al., 2010) onde um parâmetro definido k especifica o tamanho do subconjunto de rótulos que devem ser considerados.

No *RAkEL*, cada classificador do comitê de classificadores LP é treinado usando um pequeno subconjunto aleatório de *labelsets*. Portanto, é possível afirmar que nesse método, os classificadores unirrótulos, mesmo considerando o correlacionamento dos rótulos, são aplicados em subtarefas cujo número de rótulos e exemplos por rótulos sejam adequados. (NASIERDING et al., 2009; TSOUMAKAS et al., 2009).

A Figura 3.6 a seguir apresenta o resultado da criação do classificador através do *RAkEL*. Além disso, Tabela 3.4 expõe o resultado da transformação do conjunto de dados originais usando o método *RAkEL*.

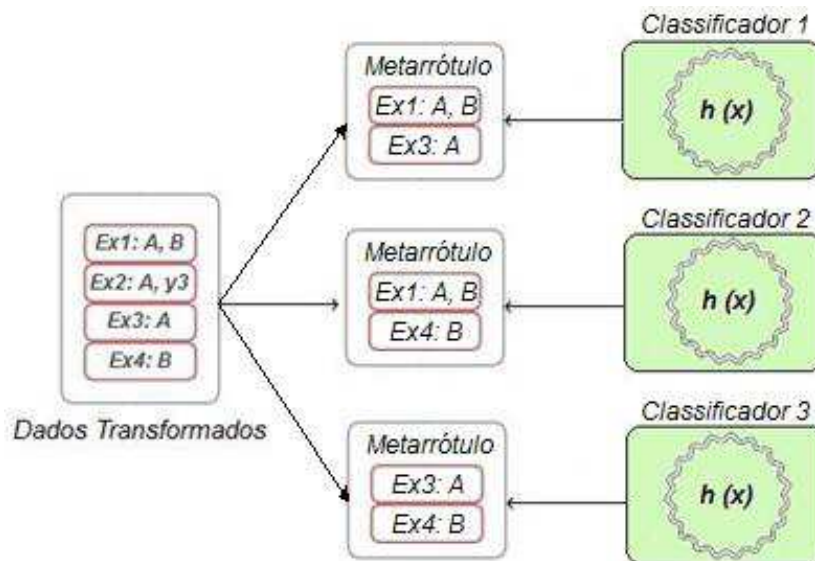


Figura 3.6 Construção do classificador multirrótulo RAKEL, (SANTOS – adaptado).

Exemplo	Y_i	Metarrótulo
E_1	$Y_1 = \{y_1, y_2, y_3\}$	$= M1$
E_2	$Y_2 = \{y_1, y_3\}$	$= M2$
E_3	$Y_3 = \{y_1\}$	$= M3$
E_4	$Y_4 = \{y_2, y_3\}$	$= M4$

Tabela 3.4 Conjunto de dados transformados usando RAKEL.

Enquanto no *Random k-Labelsets* é permitida a sobreposição dos *labelsets*, em (TSOUMAKAS et al., 2010), uma extensão do RAKEL, chamada RAKELd (do inglês, *Random K-LabELsets disjoint*), considera que os *labelsets* de cada classificador do comitê sejam disjuntos. Na Figura 3.7, tem-se a representação da criação do classificador usando o RAKELd. Também percebe-se que estudos foram realizados para entender melhor o comportamento dos dois métodos de classificação. A partir dos experimentos, foi percebido que as duas abordagens conseguiram ter um aumento significativo do desempenho comparado ao método do LP, principalmente quando se tinha um domínio do problema com grande número de dados rotulados. A ideia de sobrepor os *labelsets* atingiu um melhor desempenho comparado à ideia de manter os *labelsets* disjuntos.

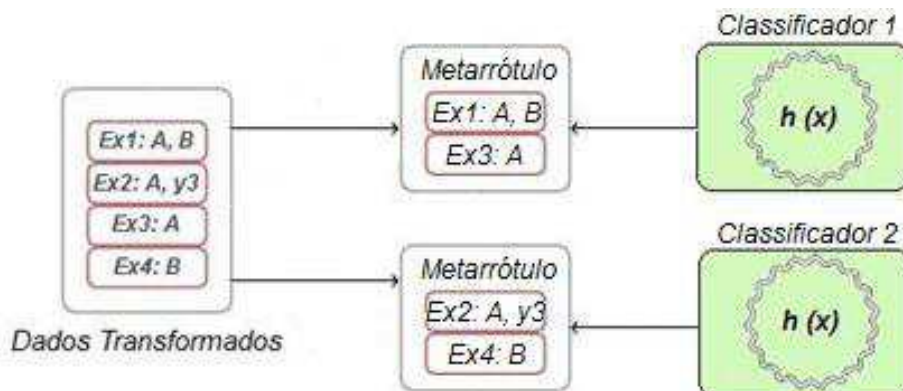


Figura 3.7 Construção do classificador multirrótulo RAKELd, (SANTOS – adaptado).

2.6) Classificação de dados multirrótulo semissupervisionado

Em (SANTOS, 2012) foi proposta a construção de classificadores multirrótulos através da aplicação de técnicas de aprendizado semissupervisionado em métodos de classificação existentes. A ideia desse estudo era avaliar a capacidade de aproximação dos resultados obtidos pelos métodos construídos comparados aos resultados oriundos do aprendizado supervisionado.

Um questionamento apresentado por (SANTOS, 2012) quanto à utilização das técnicas de aprendizado semissupervisionado foi o primeiro passo para compreender como controlar o processo de atribuição automática de rótulos, de tal maneira que essa atribuição assegurasse que os rótulos atribuídos a um determinado exemplo contribuiriam de maneira positiva para o processo de aprendizado. Baseando-se nesse contexto, os algoritmos de classificação multirrótulos semissupervisionados propostos permitiam a utilização de um parâmetro para controlar essa atribuição automática de rótulos denominado fator de confiança.

A ideia de usar o fator de confiança no processo de atribuição de rótulos era para que, em cada exemplo não-rotulado, apenas os rótulos que apresentassem um valor de saída maior ou igual ao fator de confiança, seria a ele atribuído rótulos. Além do uso do fator de confiança, em (SANTOS, 2012), há a citação em que o aprendizado semissupervisionado é iterativo e incremental, ou seja, os exemplos não-rotulados são rotulados e, por conseguinte, adicionados ao conjunto de treinamento aos poucos após cada iteração. Para esse processo, foi necessário o uso de parâmetros que definissem a

quantidade de exemplos (instâncias) não-rotuladas que seriam adicionados a cada iteração.

Abaixo os métodos de classificação multirrótulo semissupervisionada em (SANTOS, 2012).

- o SSBR (do inglês, *Semi-Supervised Binary Relevance*);
- o SSLP (do inglês, *Semi-Supervised Label Powerset*);
- o SSRaKEL (do inglês, *Semi-Supervised Random k-Labelsets*);
- o SSRaKELd (do inglês, *Semi-Supervised Random k disjoint Labelsets*);

2.6.1) SSBR

O método SSBR (do inglês, *Semi-Supervised Binary Relevance*) é uma extensão do algoritmo supervisionado BR. O SSBR é um método semissupervisionado na qual, durante a fase de treinamento, o SSBR busca incrementar a quantidade de exemplos do conjunto de dados de treinamento. A partir desse conceito, fica definido, portanto, as ideias adotadas pelos métodos BR e SSBR, haja vista que, no processo usado durante a fase de treinamento, cada método tem uma aplicabilidade diferente (SANTOS, 2012).

Considerando, para os algoritmos de aprendizado semissupervisionado apresentados nesse trabalho, um cenário onde conjunto de exemplos de treinamento é definido por $D = \{D_l \cup D_u\}$, onde o conjunto de exemplos rotulados é representado por $D_l = \{E_{l1}, \dots, E_{lN}\}$; $D_u = \{E_{u1}, \dots, E_{uK}\}$ apresenta-se como o conjunto de dados não-rotulados onde $L = \{\lambda_1, \dots, \lambda_M\}$ representa o conjunto de rótulos, N representa o número de exemplos rotulados, K é o número de exemplos não-rotulados e M representa o número de rótulos possíveis (SANTOS, 2012).

Seguindo a mesma lógica do método BR, o SSBR é iniciado com a transformação dos exemplos. Nessa etapa, os exemplos do conjunto de dados rotulados D_l são transformados em M subconjuntos de dados, sendo um para cada rótulo do conjunto L . (SANTOS, 2012).

Terminada a fase de transformação dos dados, é dado início à fase de treinamento desse conjunto de dados D_l . No SSBR, a quantidade de iterações a ser realizadas por esse método dependerá da proporção de exemplos não-rotulados constantes no conjunto de dados disponível para serem rotulados em cada iteração, permitindo, dessa maneira, que a fase de treinamento do SSBR seja realizado em uma ou mais iterações. Diferentemente de como é o processo da fase de treinamento do SSBR, no algoritmo do BR a fase de treinamento é realizada em uma única iteração, pois trata-se de um processo supervisionado.

No algoritmo SSBR, as iterações em sua fase de treinamento são divididas em duas sub-etapas: na primeira delas, um classificador multirrótulo é gerado através do conjunto de exemplos rotulados e, na segunda etapa, os exemplos do conjunto de exemplos não-rotulados serão rotulados através do classificador gerado, incrementando, assim, o número de exemplos do conjunto de exemplos rotulados. Para melhor definir a quantidade de exemplos a ser rotulado é necessário parametrizar a proporção ideal de exemplos não-rotulados que serão rotulados a cada iteração, pois quanto maior a proporção desses exemplos não-rotulados, menor será o número de iterações (SANTOS, 2012).

A sequência abaixo explica como funciona o algoritmo do SSBR em cada iteração na fase treinamento:

Algoritmo 1 Fluxo básico do SSBR na fase de treinamento em (SANTOS, 2012).

- 1: Dados $D_{l_1}, D_{l_2}, \dots, D_{l_M}$ e D_u
- 2: **enquanto** $D_u \neq \emptyset$ **faça**
- 3: Selecionar aleatoriamente p exemplos de D_u e movê-los para o conjunto $D_{u_{temp}}$, tal que $1 \leq p \leq |D_u|$
- 4: Criar uma cópia do conjunto $D_{u_{temp}}$ para cada rótulo $\lambda_j \in L$, chamada D_{u_j}
- 5: Descartar o conjunto $D_{u_{temp}}$
- 6: **para** Cada $\lambda_j \in L$ **faça**
- 7: Treinar o classificador C_j utilizando o conjunto de exemplos D_{l_j}
- 8: **para** Cada exemplo $E_t \in D_{u_j}$, tal que $1 \leq t \leq |D_{u_j}|$ **faça**
- 9: Apresentar E_t ao classificador C_j

```

10:      Utilizar a saída do classificador  $C_j$  para rotular  $E_t$  (1, caso  $E_t \in \lambda_j$ , 0,
caso contrário)
11:      fim para
12:      Mover os  $p$  exemplos recém-rotulados do conjunto  $D_{u_{temp}}$  para o conjunto  $D_l$ 
13:      fim para
14: fim enquanto

```

Mesmo com a proposição do algoritmo SSBR, estudos realizados em (SANTOS, 2012) levantaram questionamentos sobre a capacidade de diminuição do processo de atribuição automática de rótulos em deteriorar o processo de aprendizado, ao invés de melhorar. Embasado por essa discussão, foi proposto uma extensão para o algoritmo do SSBR, chamada SSBRc (do inglês, *Semis-Supervised Binary Relevance with Confidence*). Esse método faz uso de um fator de confiança, $conf$, onde $0 \leq conf \leq 1$ que é capaz de controlar o processo de atribuição automática de rótulos, tomando como base o valor de saída do classificador gerado.

A seguir, é apresentado a lógica do algoritmo do SSBRc:

Algoritmo 2 Fluxo básico do SSBRc na fase de treinamento em (SANTOS, 2012).

```

1: Dados  $D_{l_1}, D_{l_2}, \dots, D_{l_M}$  e  $D_u$ 
2: enquanto  $D_u \neq \emptyset$  faça
3:   Selecionar aleatoriamente  $p$  exemplos de  $D_u$  e movê-los para o conjunto  $D_{u_{temp}}$ , tal
que  $1 \leq p \leq |D_u|$ 
4:   Criar uma cópia do conjunto  $D_{u_{temp}}$  para cada rótulo  $\lambda_j \in L$ , chamada  $D_{u_j}$ 
5:   Descartar o conjunto  $D_{u_{temp}}$ 
6:   para Cada  $\lambda_j \in L$  faça
7:     Treinar o classificador  $C_j$  utilizando o conjunto de exemplos  $D_{l_j}$ 
8:     para Cada exemplo  $E_t \in D_{u_j}$ , tal que  $1 \leq t \leq |D_{u_j}|$  faça
9:       Apresentar  $E_t$  ao classificador  $C_j$ 
10:      se Saída do classificador  $C_j \geq conf$  então
11:        Atribuir rótulo 1;
12:      senão
13:        Atribuir rótulo 0;
14:      fim se
15:    fim para
16:    Mover os  $p$  exemplos recém-rotulados do conjunto  $D_{u_{temp}}$  para o conjunto  $D_l$ 
17: fim para

```

18: fim enquanto

Nos métodos SSBR e SSBRc, terminada a fase de treinamento, é iniciada a fase de teste. Nessa nova fase, todos os classificadores intermediários gerados durante a fase de treinamento são descartados por esses métodos, bem como as informações contidas neles. Considera-se, portanto, apenas o último classificador C_j gerado para cada rótulo λ_j .

2.6.2) SSLP

O método SSLP (do inglês, *Semi-Supervised Label Powerset*) é uma extensão do algoritmo supervisionado LP. O métodos semissupervisionado SSLP considerando as combinações entre os rótulos, criando, assim, um metarrótulo q para cada combinação, visto que $1 \leq q \leq |L|$. Dessa maneira, no método SSLP, a transformação dos exemplos do conjunto de dados rotulados D_l resulta na geração do conjunto de dados $D_{l_{tr}}$. Em seguida, cria-se uma cópia do conjunto de dados não-rotulados D_u , nominado por $D_{u_{tr}}$ (SANTOS, 2012).

Haja vista a atribuição de rótulos para todos os exemplos do conjunto de dados $D_{u_{tr}}$, o classificador gerado C é descartado utilizando o conjunto de dados rotulados temporário $D_{l_{tr}}$, onde é composto por todos os exemplos do conjunto D_l , bem como os exemplos do conjunto D_u e seus respectivos rótulos.

A sequência abaixo explica como funciona o algoritmo do SSLP na fase treinamento:

Algoritmo 3 Fluxo básico do SSLP na fase de treinamento em (SANTOS, 2012).

- 1: Dados $D_{l_{tr}}$ e $D_{u_{tr}}$
- 2: **enquanto** $D_{u_{tr}} \neq \emptyset$ **faça**
- 3: Selecionar aleatoriamente p exemplos de $D_{u_{tr}}$, tal que $1 \leq p \leq |D_{u_{tr}}|$, movendo-os para um conjunto $D_{u_{temp}}$
- 4: Treinar o classificador C_i utilizando o conjunto de dados $D_{l_{tr}}$
- 5: para Cada exemplo $E_t \in D_{u_{temp}}$, tal que $1 \leq t \leq |D_{u_{temp}}|$ **faça**
- 6: Apresentar E_t ao classificador C_i

- 7: Utilizar a saída do classificador C_i para atribuir um dos metarrótulos q
 - 8: **fim para**
 - 9: Mover os exemplos do conjunto $D_{u_{temp}}$, juntamente com os metarrótulos resultantes, para o conjunto $D_{l_{tr}}$
 - 10: **fim enquanto**
-

Assim como no método SSBR, em (SANTOS, 2012), também, foi proposta uma variação através do uso de um fator de confiança $conf$, tal que $0 \leq conf \leq 1$. Essa variação foi chamada de SSLPc (do inglês, *Semi-Supervised Label Powerset with Confidence*), com a objetivo de diminuir a inserção de ruídos durante a etapa de atribuição de rótulos aos exemplos do conjunto não-rotulados D_u .

A seguir, é apresentado o procedimento utilizado na fase de treinamento do SSLPc:

Algoritmo 4 Fluxo básico do SSLPc na fase de treinamento em (SANTOS, 2012).

- 1: Dados $D_{l_{tr}}$ e $D_{u_{tr}}$
 - 2: **enquanto** $D_{u_{tr}} \neq \emptyset$ **faça**
 - 3: Selecionar aleatoriamente p exemplos de $D_{u_{tr}}$, tal que $1 \leq p \leq |D_{u_{tr}}|$, movendo-os para um conjunto $D_{u_{temp}}$
 - 4: Treinar o classificador C_i utilizando o conjunto de dados $D_{l_{tr}}$
 - 5: **para** Cada exemplo $E_t \in D_{u_{temp}}$, tal que $1 \leq t \leq |D_{u_{temp}}|$ **faça**
 - 6: Apresentar E_t ao classificador C_i
 - 7: **para** Cada t **faça**
 - 8: **se** Saída do clasificador $C_i \geq conf$ **então**
 - 9: Atribuir rótulo 1;
 - 10: **senão**
 - 11: Atribuir rótulo 0;
 - 12: **fim para**
 - 12: **fim para**
 - 12: Mover os exemplos do conjunto $D_{u_{temp}}$, juntamente com os metarrótulos resultantes, para o conjunto $D_{l_{tr}}$
 - 14: **fim enquanto**
-

2.6.3) SSRAkEL

Em (SANTOS, 2012), o método semissupervisionado do RAkEL, chamado SSRAkEL (do inglês, *Semi-Supervised Random k-Labelsets*), faz uso da mesma ideia do método

RAkEL para construir um comitê de classificadores LP. Como mencionado anteriormente, o principal objetivo do RAkEL é formar comitês (*ensemble*) de classificadores LP, de maneira que cada classificador seja treinado utilizando diferentes subconjuntos *randômico* de *labelsets* (NASIERDING et al., 2009; TSOUMAKAS et al., 2009). O que diferencia cada método (RAkEL e SSRAkEL) são os tipos de classificadores gerados a partir da fase de treinamento. Enquanto o RAkEL constrói classificadores LP, o SSRAkEL gera classificadores SSLP.

Seguindo a mesma ideia da variação proposta em SSBR e SSLP com o intuito de reduzir a inserção de ruídos durante o processo de treinamentos dos rótulos aos exemplos do conjunto D_u , foi criado o método SSRAkELc (do inglês, *Semi-Supervised Random k-Labelsets with Confidence*). A principal disparidade entre os métodos SSRAkEL e SSRAkELc é no tipo de classificador gerado, ou seja, enquanto o classificador SSLP é gerado no SSRAkEL, o classificador SSLPc é gerado no SSRAkELc.

A respeito do fluxo do algoritmo que representa o método do SSRAkEL, este, por sua vez, é representado pela mesma estrutura do fluxo do Algoritmo 3. O que vai diferenciar o fluxo do algoritmo do SSRAkEL para o fluxo do algoritmo do SSLP é, somente, o classificador gerado.

2.6.4) SSRAkELd

Como mencionado anteriormente, em (TSOUMAKAS et al., 2010) é proposto uma variação do RAkEL, chamada de RAkELd. Esse método leva à disjunção dos *labelsets*. O método proposto em (SANTOS, 2012), definido como SSRAkELd, (do inglês, *Semi-Supervised Randomk-Labelsets disjoint*), faz uso da mesma proposição para construir um comitê de classificadores SSLP a partir do SSRAkEL. A diferença entre os métodos (SSRAkEL e SSRAkELd) durante a fase de seleção metarrótulos de cada *labelsets* é que, enquanto no SSRAkEL é possível a sobreposição de metarrótulos, no método SSRAkELd os *labelsets* são disjuntos.

Seguindo a mesma ideia da variação proposta em SSRAkELc, cujo intuito é minimizar a inserção de ruído durante o processo de atribuição de rótulos aos exemplos do conjunto

de não-rotulados D_u , em (SANTOS, 2012) é apresentado o método SSRaKEldc, (do inglês, *Semi-Supervised Random k-Labelsets disjoint with Confidence*), onde a utilização é pré-definida por um fator de confiança para controlar o processo de atribuição automática de rótulos.

Quanto ao fluxo do algoritmo que representa o método do SSRaKEld, este, por sua vez, é representado pela mesma estrutura do fluxo do SSRaKEl. O que vai diferenciar o fluxo do algoritmo do SSRaKEld para o fluxo do algoritmo do SSRaKEl é, somente, a sobreposição de metarrótulos, pois é permitido SSRaKEl e no SSRaKEld eles são disjuntos.

2.7) Considerações Finais

Neste capítulo foram apresentados os conceitos de classificação multirrótulo, bem como os fundamentos de alguns dos principais métodos de transformação do problema, BR, LP, RAkEL e RAkELd. Ademais, foram apresentados os conceitos, algoritmos e exemplos dos métodos de classificação de dados multirrótulo semissupervisionado que foram propostos pelo trabalho de (SANTOS, 2012).

Além disso, foram apresentados os conceitos de classificação multirrótulo, bem como os fundamentos de alguns dos principais métodos de transformação do problema, BR, LP, RAkEL e RAkELd. Ademais, foram apresentados os conceitos, algoritmos e exemplos dos métodos de classificação de dados multirrótulo semissupervisionado que foram propostos pelo trabalho de (SANTOS, 2012).

Capítulo 3

Trabalhos Relacionados

Este capítulo apresenta os principais trabalhos relacionados a este. Cita os trabalhos que estudam, dentre outras coisas, as técnicas de classificação multirrótulo associados aos estudos do aprendizado semissupervisionado. Como citado anteriormente, as aplicações de classificação multirrótulo consideram a possibilidade de alguns exemplos, pertencentes a um conjunto de dados, estarem rotulados a mais de uma classe simultaneamente (TSOUMAKAS et al., 2008). Para tal, têm-se os algoritmos de classificação multirrótulos para predizer, de maneira precisa e eficiente, a qual rótulo cada exemplo pertence. Surge, portanto, a ideia da associar a classificação de dados (*instâncias*) multirrótulos com aprendizado semissupervisionado. A principal característica dessa técnica de aprendizado de máquina é combinar, dentro de um conjunto de exemplos, os poucos exemplos rotulados e os não-rotulados durante o processo de aprendizado (Zhu et al., 2003; Zhou & Li, 2010). Ficou concluído, citando os principais trabalhos que, apesar de existir trabalhos que abordam aplicações envolvendo aprendizado semissupervisionado em tarefas de classificação multirrótulo, nenhum deles parte do princípio que a seleção dos dados pertencentes a um conjunto de dados não-rotulados possam ser selecionados para serem rotulados a partir de um parâmetro de confiabilidade. O parâmetro de confiabilidade, neste caso, será calculado o desvio padrão dos valores de confiança de cada instância.

Este capítulo está organizado da seguinte forma: na Seção 4.1 são apresentados os principais trabalhos envolvendo classificação multirrótulo, enquanto na Seção 4.2 trata dos principais trabalhos envolvendo aprendizado semissupervisionado.

3.1) Classificação Multirrótulo

Na literatura encontramos diferentes trabalhos que tratam dos problemas de classificação multirrótulo, como por exemplo: (VALLIM; FREITAS, 2009; TSOUMAKAS et al., 2008; TSOUMAKAS; KATAKIS, 2007). Dentre os trabalhos, (TSOUMAKAS; KATAKIS; VLAHAVAS, 2008) propõe um modelo de classificação multirrótulo além

dos modelos apresentados pelos demais trabalhos relacionados. Este modelo de classificação multirrótulo hierárquico, por sua vez, apresenta um algoritmo chamado HOMER (do inglês, *Hierarchy Of Multilabel classifiERs*), capaz de construir uma hierarquia de classificadores multirrótulo. Nesse modelo, uma de suas características é a transformação de tarefas de classificação multirrótulo que possui um alto número de rótulos L em uma hierarquia no formato de árvore de tarefas de classificação multirrótulo. Além disso, um dos processos realizados por esse método é o agrupamento do conjunto de rótulos em subconjuntos *disjuntos*, de tal maneira que os rótulos semelhantes sejam alocados em um mesmo subconjunto. Seguindo o paradigma “dividir para conquistar”, no HOMER, cada nó interno da árvore contém um subconjunto *disjunto* de rótulos do conjunto inicial, portanto, um classificador multirrótulo é gerado para cada nó interno e a classificação multirrótulo dos nós é executada de maneira recursiva segundo o estilo *top-down*, ou seja, é iniciado na raiz pela classificação de um ou mais rótulos e, caso os rótulos estejam entre os rótulos preditos pelo classificadores pais, estes vão seguindo pelos nós-filhos. Os demais trabalhos apresentam técnicas e conceitos sobre classificação multirrótulo em diferentes tipos de aplicação.

Em (VALLIM, 2009) é apresentado uma proposta de LCS (do inglês, *Learning Classifier System*) para problemas de classificação multirrótulo, na qual se pretende induzir um conjunto de regras de classificação que possa gerar um resultado eficaz e que, também, seja capaz de ser comparável com outras técnicas de classificação. Os LCS, *Learning Classifier System*, são representado por técnicas de indução de regras de classificação, cujo principal mecanismo de busca é um algoritmo genético. Essa técnica almeja encontrar um conjunto de regras que disponha de uma alta precisão de classificação. Na literatura, os estudos sobre os LCS são temas muito bem fundamentados e vêm se mostrando capazes de atingir resultados satisfatórios com classes mutuamente exclusivas. Apesar de existirem trabalhos de classificadores evolutivos para problemas multirrótulo (CHAN; FREITAS, 2006), o trabalho proposto surgiu através do pouco conhecimento tido a respeito de um LCS aplicado a problemas dessa natureza.

Em (METZ; FREITAS, 2009) é proposto um modelo hierárquico de classificação, onde cada nó da hierarquia de classes recebe um classificador binário plano. Esses classificadores, por conseguinte, decidem se cada instância de teste será vinculada à classe definida com classificador correspondente ou não. Portanto, embasado por essa

definição, uma instância poderá ser facilmente declarada a mais de uma classe a cada nível de classificação, ou seja, tornando essa classificação do tipo multirrótulo. A intenção do método proposto por (METZ; FREITAS, 2009) é tratar a tarefa de classificação hierárquica em um cenário cuja quantidade de dados disponíveis previamente classificados é muito escassa. Neste contexto, torna-se difícil a construção de um bom modelo de classificação, uma vez que as informações disponíveis sobre o espaço de dados são pobres.

Visando minimizar essa deficiência, é possível utilizar informações parcialmente supervisionadas para orientar o processo de aprendizagem e impulsionar o aumento de evidências a respeito do problema. Em algumas aplicações, por exemplo, o processo de rotulagem de dados é uma tarefa considerada difícil, haja vista a rotulagem dos dados geralmente é feita por um especialista; o que pode caracterizar uma tarefa custosa, bem como demorada. Em contrapartida, a aquisição dos dados não-rotulados é uma tarefa simples comparada à aquisição de dados rotulados, pois essa tarefa envolve basicamente a coleta dos dados (*instâncias*) de uma base de dados. Para esse impasse, é ideal o uso do aprendizado semissupervisionado, pois esse tipo de aprendizado trata de situações onde a quantidade de dados não-rotulados é menor e a quantidade de dados rotulados é maior.

Os métodos de classificação multirrótulo estão cada vez mais sendo exigidos por modernas aplicações que classificam dados. O trabalho de (TSOUMAKAS; KATAKIS, 2007) propõe realizar uma tarefa de comparação dos resultados experimentais de alguns métodos de classificação, além de contribuir, também, para a definição de conceitos para a quantificação multirrótulo do conjunto de dados. A classificação *single-label*, por exemplo, está focada na aprendizagem cujos exemplos estão associados a um único rótulo l a partir de um conjunto de rótulos *disjoint* L , onde $|L| > 1$. Caso $|L| = 2$. Problemas desse tipo de aprendizagem é conhecido como classificação binária, enquanto que, se $|L| > 2$, são problemas de classificação multiclases. Na classificação multirrótulo, os dados (*instâncias*) são associados a um conjunto de rótulos $Y \subseteq L$, ou seja, cada instância poderá estar associada a mais de um rótulo simultaneamente.

Como mencionado anteriormente, os métodos multirrótulos são cada vez mais exigidos por novas aplicações, como a função de classificação de proteínas em (ZHANG; ZHOU,

2005), a categorização de música em (LI; OGIHARA, 2003), e, também, a classificação semântica de imagens em (BOUTELL et al., 2004). Esses trabalhos caracterizam bem a aplicabilidade dos métodos de classificação multirrótulo, pois são apresentadas situações em que exemplos das bases de dados podem estar associados a mais de uma classe conceitual ao mesmo tempo. Por exemplo, em (LI; OGIHARA, 2003) a categorização da música é definida pela classificação dos dados a um ou mais gêneros musicais, ou seja, várias músicas de sucesso de uma conhecida banda de rock podem ser caracterizadas tanto como rock quanto eletrônica, ou até mesmo outra classificação. Portanto, o trabalho proposto em (TSOUMAKAS; KATAKIS, 2007) serve como ponto inicial e de referência para estudos na área de classificação multirrótulo. A definição de um conceito para a quantificação de um conjunto de dados multirrótulos, a comparação de resultados experimentais sobre certos rótulos em diferentes tipos de métodos, além de um programa estruturado que apresenta pontos fortes e fracos sobre os métodos de classificação multirrótulos, são exemplos de contribuições que foram apresentadas por esse trabalho.

3.2) Aprendizado Semissupervisionado

Como mencionado anteriormente, os algoritmos de aprendizado semissupervisionado aprendem a partir de uma combinação de dados (*instâncias*) rotulados e não-rotulados. Assim, esses algoritmos de semissupervisão podem ser aplicados em problemas cujos exemplos rotulados são escassos e uma vasta quantidade de exemplos não-rotulados estão disponíveis.

Na literatura, podemos encontrar trabalhos que abordam técnicas bastantes utilizadas no aprendizado semissupervisionado como, por exemplo, a técnica do *Co-Training* em (BLUM; MITCHELL, 1998) e o *Self-Training* em (ROSENBERG et al., 2005; METZ, 2011). Além destes, podem ser citados, também, o trabalho proposto por (BRAGA, 2010) na qual se aplica o aprendizado semissupervisionado multidescrição em classificação de textos, bem como em (WAGSTAFF et al., 2001) e (BASU et al., 2002) que propõe uma variação do algoritmo não-supervisionado *COP-k-means* e *Seeded-k-means*. É sabido, portanto, que os algoritmos semissupervisionados podem atingir um desempenho superior aos algoritmos supervisionados treinados nos mesmos poucos exemplos rotulados (BRAGA, 2010). Vejamos as características dos trabalhos relacionados a seguir.

Na literatura, a técnica de aprendizado semissupervisionado multidescrição ainda não foi bem explorada no que diz respeito às várias possíveis maneiras de se descrever bases de textos. Em (BRAGA, 2010) é proposto um estudo que analisa o desempenho de algoritmos semissupervisionados multidescrição na classificação de textos. Considera-se a priori o algoritmo multidescrição *Co-Training*, pois são propostas alterações a fim de resolver o problema dos pontos de contenção. Além disso, é proposto o algoritmo *CoAL*, este, por sua vez, pode melhorar ainda mais o algoritmo *Co-Training* através da incorporação de aprendizado ativo como uma forma de resolver problemas de contenção. No trabalho proposto por (BRAGA, 2010) foi usado unigramas e bigramas para compor duas descrições distintas de documentos textuais. Com base nisso, foram analisados os resultados obtidos pelos dois algoritmos multidescrição apresentados. Os resultados mostram que o algoritmo *CoAL*, usando unigramas como descrição das bases textuais e bigramas como uma outra descrição, atingem um desempenho superior comparado a um algoritmo semissupervisionado monodescrição, o que caracteriza que o uso de unigramas e bigramas como duas descrições distintas de bases de textos podem ser bastante compensadores.

Em (WAGSTAFF et al., 2001) é proposta a criação de um algoritmo chamado *COP-k-means*. Este, por sua vez, é uma versão modificada do algoritmo não-supervisionado *k-means* (MACQUEEN, 1967). O que diferencia o algoritmo *COP-k-means* do método não-supervisionado *k-means* é a forma como é usada o conhecimento prévio (*background knowledge*), ou seja, a forma como é usada o conhecimento prévio sobre o domínio do problema descrito na forma de relacionamentos entre os exemplos, cuja utilização será na fase de formação dos *clusters*. Outro trabalho que expõe a ideia de modificação de algoritmo é o (BASU et al., 2002). Neste trabalho é proposto o *Seeded-k-Means*, um algoritmo gerado a partir da aplicação do algoritmo não-supervisionado *k-means* em um conjunto de dados em *k clusters*, cuja principal característica é a utilização de exemplos previamente rotulados como centróides iniciais dos *clusters*, ou seja, estes exemplos definidos por *seeds* e o processo de escolha não ser aleatória. Além das menções citadas anteriormente, em (BASU et al., 2002) é proposto um algoritmo melhorado a partir do *Seeded-k-means*. Porém, apesar da similaridade, há uma diferença entre ambos. Os exemplos que fazem parte do conjunto de *seeds*, e que foram associados inicialmente a um dado *cluster*, não poderão ser associados posteriormente a outro *cluster*. Nesse sentido, apenas os exemplos não selecionados como *seeds* terão o *cluster* correspondente

reestimado. Por fim, leva-se em consideração a associação inicial dos *seeds*, pois no *Seeded-k-means*, as *seeds* podem vir a pertencer a diferentes *clusters* daqueles previamente associados, caso que no *Constrained-k-means* isso não acontece. Com base nesse contexto, o método *Constrained-k-means* é definido como mais adequado quando as *seeds*, relacionadas aos exemplos rotulados, estão ausentes de ruídos.

3.3) Aprendizado Semissupervisionado com Classificação Multirrótulo

Na literatura é possível encontrar alguns métodos que combinam técnicas de aprendizado semissupervisionado com classificação multirrótulo, às vezes classificação multirrótulo plana, outras hierárquica. Em (METZ; FREITAS, 2009), por exemplo, foram propostos métodos que avaliam algoritmos multirrótulos e hierárquico, os quais podem definir estratégias para identificar os *multirrótulos majoritários*, que são dados utilizados para calcular os valores *baseline* das medidas de avaliação.

Além do trabalho acima citado, outros trabalhos presentes na literatura, que também estão relacionados com problemas de classificação multirrótulo em aprendizado semissupervisionado, abordam técnicas de aprendizado semissupervisionado com classificação multirrótulo; estes são (RODRIGUES et al, 2013; SANTOS, 2012). Por não ter, em algumas situações, alcançado resultados mais significantes. O trabalho proposto por (SANTOS, 2012) servirá como base para esse trabalho de dissertação que será apresentado no capítulo 5.

3.4) Considerações Finais

Neste capítulo foram apresentados diversos trabalho referente às abordagens semissupervisionada e classificação multirrótulo. Essas abordagens vem apresentando resultados favoráveis a sua utilização através de intensas pesquisas realizadas pela comunidade científica e, por isso, tornou-se interessante desenvolver estudos que combinassem as duas técnicas, unindo os pontos fortes e buscando, dessa maneira, soluções eficazes em problemas de classificação.

Capítulo 4

Métodos Propostos

Como anteriormente mencionado, um dos objetivos desse trabalho é tentar melhorar as técnicas de classificação propostas por (SANTOS, 2012) através da atribuição de confiabilidade no processo automático da escolha de instâncias a serem rotuladas. Dessa maneira, é proposto neste capítulo alguns algoritmos que geram os classificadores multirrótulos através da aplicabilidade de técnicas de aprendizado semissupervisionado a partir de um parâmetro de confiabilidade. Para esse trabalho o parâmetro de confiabilidade definido a partir do cálculo do desvio padrão dos valores da confiança de cada exemplos. Antes de iniciar a descrição de cada método proposto, é imprescindível definir de maneira mais precisa o problema multirrótulo semissupervisionado.

No trabalho proposto por (SANTOS, 2012), foi constatado que, em alguns casos, o uso da confiabilidade dos exemplos pertencentes a um conjunto de dados não apresentavam resultados satisfatórios o quanto eram estimados, ou seja, esses resultados não atingiram os pontos máximos e mínimos de acordo com as métrica de avaliação. Uma das possíveis causas para esse fato ter ocorrido pode ter sido através da maneira como os exemplos a serem rotulados foram escolhidos, ou seja, a forma aleatória que esses exemplos foram escolhidos pode ter contribuído de maneira negativa na obtenção dos resultados, pois pressupõe-se que por meio dessa escolha aleatória é possível que, dentre um grupo de exemplos de uma mesma região do espaço ou que não estejam tão confiantes para a rotulação. Portanto, para minimizar essa possibilidade, esse trabalho propõe que sejam escolhidos padrões que estejam mais confiantes para a rotulação em diferentes região do espaço, para que o classificador a ser gerado possa cobrir todo o espaço disponível de exemplos. Nesse contexto, serão usados como parâmetro de confiabilidade os valores obtidos por meio do cálculo dos desvios padrões das confidências de cada exemplo para que, dessa maneira, seja suprimido o problema da escolha de exemplos similares. Espera-se, portanto, que os exemplos com alto valor do desvio padrão sejam menos similares comparados com os exemplos de menor valor do desvio padrão.

Este capítulo está organizado da seguinte forma: na Seção 5.1 apresenta os algoritmos semissupervisionados adaptados; na Seção 5.2 é apresentada a conceituação do uso do fator de confiança nos métodos semissupervisionados com parâmetro de confiabilidade; na Seção 5.3 são apresentados os novos métodos de classificação multirrótulo semissupervisionado baseados no valor de confiança; na Seção 5.4 serão apresentadas algumas considerações acerca de cada método proposto

4.1) Algoritmo Semissupervisionado adaptado

Para desenvolver as adaptações no algoritmo semissupervisionado, foi necessário modificar o processo de treinamento desses algoritmos, no caso o *Self-Training*. Considere que um cenário, cujos elementos de um espaço de classificação multirrótulo é representado por um conjunto de exemplos de treinamento $D = \{D_l \cup D_u\}$, conjunto de exemplos rotulados representado por $D_l = \{E_{l_1}, \dots, E_{l_n}\}$, $D_u = \{E_{u_1}, \dots, E_{u_k}\}$ representando o conjunto de dados não-rotulados, conjunto de rótulos sendo definido por $L = \{\lambda_1, \dots, \lambda_m\}$, os números de exemplos rotulados e não-rotulados respectivamente e, por fim, M representando o número de rótulos possíveis. Ademais, cada rótulo do conjunto de exemplos rotulados pode receber r rótulos, onde r varia de 1 até a quantidade de rótulos possíveis M .

Aplicando a técnica do *Self-Training*, descrita na seção 2.3.3, como uma abordagem do aprendizado semissupervisionado, para este trabalho foram feitas adaptações com relação ao uso dessa técnica na aplicação dos métodos propostos. A cada iteração e após a criação do classificador inicial, exemplo C_1 , é aplicada a técnica do parâmetro de confiabilidade como mecanismo auxiliar na rotulação dos exemplos não-rotulados. Desta forma, o método *Self-Training* adaptado obedecerá aos seguintes passos:

1. Serão selecionados todas as instâncias não-rotulados do conjunto de dados;
2. Cada instância possui valores de confiança;
3. A partir desses valores de confiança, é calculado o desvio padrão desses valores;
4. Por fim, serão ordenados em um vetor do maior para o menor valor de desvio padrão.

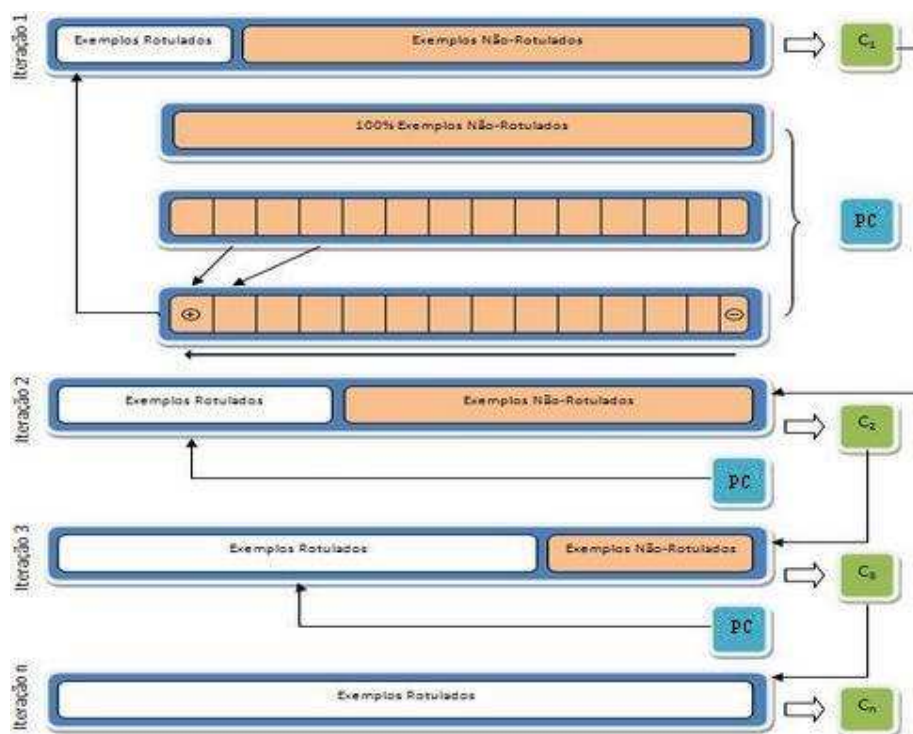


Figura 5.1 Abordagem do *Self-Training* usando o parâmetro de confiabilidade.

Para facilitar o entendimento sobre o cálculo da confiabilidade primeiro é preciso esclarecer que, durante o processo de aprendizado semissupervisionado, a partir da ferramenta MULAN, cada instância imprime valores conhecidos como de valores de confiança, ranking e bipartição. O cálculo da confiabilidade é alcançado a partir dos valores de confiança dessas instâncias onde, a partir deles, é possível dizer o quão confiante é certa instância para um determinado rótulo. Dessa maneira, um algoritmo genérico que usa o critério do parâmetro de confiabilidade como auxílio no processo de rotulação das instância através da atribuição de pesos é apresentado no Algoritmo 5.

Algoritmo 5 Parâmetro de Confiabilidade

- 1: Dados D_{l_M} e D_u
- 2: instâncias p
- 3: **enquanto** $D_u \neq \emptyset$ **faça**
- 4: Gerar *classificadores* a partir de D_{l_M}
- 5: Apresentar em $D_{u_{temp}}$ os *valores de confiança* dos p exemplos
- 6: Calcular o Desvio Padrão dos *valores de confiança* dos p exemplos
- 7: Selecionar um percentual x de p exemplos do conjunto $D_{u_{temp}}$ com os maiores valores do Desvio Padrão
- 8: Atribui *labels* aos p exemplos do conjunto $D_{u_{temp}}$ selecionados

9: Insere p rotulado em D_{l_M}

10: **fim enquanto**

Nesse algoritmo é possível entender como é usado o parâmetro de confiabilidade no processo de rotulação dos dados ainda não-rotulados. Inicialmente são gerados os classificadores a partir do conjunto de dados D_{l_M} (Linha 4). Uma vez feito isso, em um outro conjunto $D_{u_{temp}} = \{(X_{i_1}, X_{i_2}, \dots, X_{i_n}) | i = 1, \dots, N\}$, é apresentado os valores de confiança de p instâncias que servirá para o calculado do desvio padrão desses valores. Após calculado o valor do desvio padrão (Linha 6) dos valores de confiança dessas instâncias, são selecionados as instâncias p que possuem o maior valor de desvio padrão a partir da ordenação dos elementos p que foram organizados (Linha 7).

O parâmetro de confiabilidade pode ser apresentado a partir do ranking ordenado de maneira decrescente do desvio padrão dos valores de confiança das instâncias ainda não-rotuladas. Para os métodos propostos a seguir, e que fazem uso do parâmetro de confiabilidade como mecanismo auxiliar no processo de rotulação das instância, estes apresentarão a nomenclatura *ranking* como referência ao ranking ordenado das instâncias não-rotuladas. Abaixo segue um exemplo ilustrativo de como é definido o processo de confiabilidade durante a escolha dos padrões em um determinado conjunto de dados.

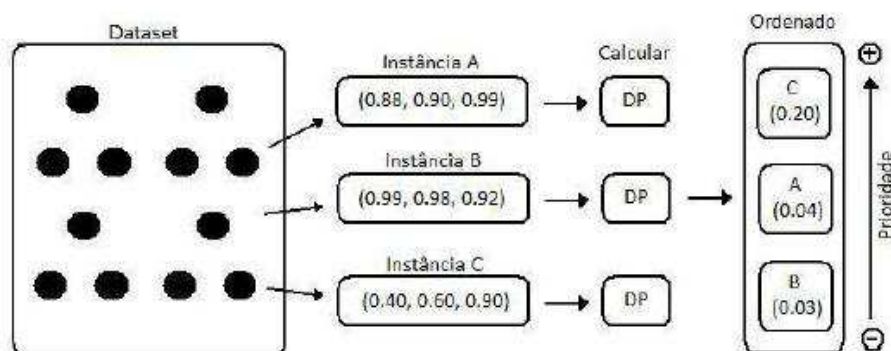


Figura 5.2 Processo ilustrativo da atuação do parâmetro de confiabilidade.

Na Figura 5.2, é observado que as instâncias são ordenadas de acordo com o nível de prioridade para rotular, ou seja, serão escolhidas primeiramente as instâncias que apresentarem a menor similaridade entre os padrões. Como mencionado anteriormente, cada instância possui valores de confiança e, a partir desses valores, é calculado o desvio padrão dos valores de confiança dessas instâncias. Assim, as instâncias que

apresentarem o maior valor do desvio padrão serão menos similares e, portanto, colocado como prioridade para rotular. Sabendo disso e analisando a Figura 5.2, os padrões da instância C apresentaram o maior desvio padrão, ou seja, são menos similares, seguidos pelos desvios padrões das instâncias A e B, respectivamente.

Um questionamento levantado antes da criação desses métodos era em relação à garantia de que os exemplos escolhidos estariam contribuindo positivamente para a obtenção dos resultados, pois, em (SANTOS, 2012), existiam chances de algumas instâncias serem escolhidas aleatoriamente e com isso os resultados apresentados não contribuiriam positivamente. Nesse contexto, cada algoritmo proposto permite que o processo de atribuição dos rótulos seja definido por um parâmetro de confiabilidade, ou seja, cada algoritmo atribuirá rótulos através do valor da confiabilidade calculado a partir do valor de cada instância e não mais pela atribuição automática de rótulos. Assim, depois de selecionado os exemplos com maior valor do desvio padrão, ou seja, selecionado aqueles exemplos que apresentaram menor similaridade, estes serão rotulados e, após isso, será gerado um classificador. Além disso, a abordagem semissupervisionada é considerada iterativa e incremental, ou seja, os exemplos não-rotulados selecionados serão rotulados e, a posteriori, adicionados ao conjunto de treinamento aos poucos, normalmente, em várias iterações. Ademais, outro parâmetro necessário utilizado na abordagem semissupervisionada é a quantidade de exemplos não-rotulados que serão inseridos a cada iteração.

4.2) Fator de confiança

No trabalho proposto por (SANTOS, 2012), há uma discussão sobre como reduzir a possibilidade do processo de atribuição automática de rótulos fragilizar o processo de aprendizado, ao invés de melhorar. Com o intuito de amenizar este problema, foi proposto por (SANTOS, 2012) a utilização de um parâmetro, chamado fator de confiança, que busca controlar o processo de atribuição automática de rótulos, baseado no valor da saída do classificador gerado. Dessa maneira, apenas os rótulos cuja saída esteja acima do valor do fator de confiança seriam atribuídos ao conjunto de exemplos.

A justificativa para a escolha dos valores utilizados como fator de confiança neste trabalho reside no fato de, segundo (SANTOS, 2012), uma vez que a confiança é a

probabilidade de o exemplo pertencer a um determinado rótulo, espera-se que quanto maior o valor do fator de confiança utilizado, maior a probabilidade do exemplo ser rotulado corretamente. Esse fator de confiança também será aplicado nos seguintes métodos propostos (SSBRcr, SSLPcr, SSRAkELcr e SSRAkELdcr). Em (SANTOS, 2012), foi feita uma análise inicial dos métodos que fizeram uso deste fator de confiança, e que, neste trabalho, essa análise foi aprofundada a partir da utilização deste fator de confiança em conjunto com a confiabilidade nos métodos propostos.

4.3) Métodos Semissupervisionado propostos

Os métodos propostos utilizam os algoritmos semissupervisionados apresentados na seção anterior com aprendizado de máquina e classificação multirrótulo.

A seguir serão apresentados quatro algoritmos usando *ranking* e os quatro algoritmos usando *confidência* e *ranking*:

- SSBRr (do inglês, *Semi-Supervised Binary Relevance using ranking*);
- SSLPr (do inglês, *Semi-Supervised Label Powerset using ranking*);
- SSRAkELr (do inglês, *Semi-Supervised Random k-Labelsets using ranking*);
- SSRAkELdr (do inglês, *Semi-Supervised Random k-Labelsets using ranking disjoint*);

- SSBRcr (do inglês, *Semi-Supervised Binary Relevance with confidence using ranking*);
- SSLPcr (do inglês, *Semi-Supervised Label Powerset with confidence using ranking*);
- SSRAkELcr (do inglês, *Semi-Supervised Random k-Labelsets with confidence using ranking*);
- SSRAkELdcr (do inglês, *Semi-Supervised Random k-Labelsets with confidence using ranking disjoint*);

4.3.1) Modelos SSBR propostos

Como mencionado anteriormente, o método semissupervisionado SSBR, criado a partir do algoritmo BR, tem como objetivo incrementar a quantidade de exemplos do conjunto de dados durante a fase de treinamento, selecionando, de maneira a

aleatória, os dados não-rotulados para rotular e incluir ao conjunto de dados rotulados. Nesta subseção, é proposto uma extensão do método SSBR, chamada SSBRr (do inglês, *Semi-Supervised Binary Relevance using ranking*), na qual, durante a fase de treinamento, essa técnica de aprendizado semissupervisionado objetiva incrementar a quantidade de exemplos não-rotulados, que foram previamente selecionados por meio do parâmetro de confiabilidade, ao conjunto de dados rotulados. Dessa maneira, o SSBR e o SSBRr diferem apenas na maneira como são selecionados os exemplos para o processo de rotulagem dos dados. Um método seleciona os dados de maneira aleatória e o outro seleciona os dados baseando-se pelo valor do desvio padrão dos valores de confiança de cada exemplo.

Após a etapa de transformação dos dados, é iniciada a fase de treinamento. Diferentemente do método BR, no qual toda a fase de treinamento é realizada em uma única iteração, no SSBRr, extensão do método SSBR proposta neste trabalho, a fase de treinamento pode ser realizada, também, em uma ou mais iterações. O número de iterações da fase de treinamento do SSBRr dependerá, primeiramente, da quantidade de exemplos não-rotulados pertencentes ao conjunto de exemplos disponíveis a serem rotulados a partir dos valores obtidos pelos desvios padrões das confidências de cada exemplo a cada iteração. Assim, cada iteração que ocorre na fase de treinamento do método SSBRr está dividida em duas etapas: A primeira delas condiz com criação de um classificador multirrótulo utilizando o conjunto de exemplos rotulados disponível. Na segunda, os exemplos do conjunto de dados não-rotulados passarão pelo processo de seleção através do *ranking* do desvio padrão dos valores de confiança das instâncias e, por conseguinte, serão rotulados utilizando o classificador gerado, com o objetivo de incrementar o número de exemplos do conjunto de exemplos rotulados. Nesse contexto, é sabido que um importante parâmetro, cujo ajuste será imprescindível, é o quantitativo de exemplos não-rotulados que devem ser rotulados a cada iteração, pois quanto maior a proporção de exemplos não-rotulados utilizados a cada iteração, menor será o número de iterações na fase de treinamento. O Algoritmo 6 descreve toda a fase de treinamento do método SSBRr.

Algoritmo 6 Fluxo básico do SSBRr na fase de treinamento

- 1: Dados $D_{l_1}, D_{l_2}, \dots, D_{l_M}$ e D_u
- 2: **enquanto** $D_u \neq \emptyset$ **faça**

- 3: Selecionar os p exemplos de D_u e movê-los para o conjunto $D_{u_{temp}}$, tal que $1 \leq p \leq |D_u|$
 - 4: Criar uma cópia do conjunto $D_{u_{temp}}$, chamada $D_{u_{temp}DP}$
 - 5: **Calcular** a Confiabilidade a partir do Algoritmo 5
 - 6: Criar uma cópia do conjunto $D_{u_{temp}DP}$ para cada rótulo $\lambda_j \in L$, chamada D_{u_j}
 - 7: Aplicar o percentual x de p exemplos selecionados ao conjunto D_{u_j}
 - 8: Descartar os conjuntos $D_{u_{temp}}$ e $D_{u_{temp}DP}$
 - 9: **para** Cada $\lambda_j \in L$ **faça**
 - 10: Treinar o classificador C_j utilizando o conjunto de exemplos D_{l_j}
 - 11: **para** Cada exemplo $E_t \in D_{u_j}$, tal que $1 \leq t \leq |D_{u_j}|$ **faça**
 - 12: Apresentar E_t ao classificador C_j
 - 13: Utilizar a saída do classificador C_j para rotular E_t (1, caso $E_t \in \lambda_j$, 0, caso contrário)
 - 14: **fim para**
 - 15: Mover os p exemplos recém-rotulados do conjunto $D_{u_{temp}DP}$ para o conjunto D_l
 - 16: **fim para**
 - 17: **fim enquanto**
-

Como apresentado no algoritmo anterior, a cada iteração da fase de treinamento do SSBRr, p exemplos do conjunto de exemplos não-rotulados D_u são selecionados e movidos para o conjunto $D_{u_{temp}}$, tal que $1 \leq p \leq |D_u|$. Em seguida, é executado o algoritmo 5 para calcular o desvio padrão dos valores de confidências das instâncias não-rotuladas a partir dos dados do conjunto $D_{u_{temp}}$. Após isso, para cada rótulo $\lambda_j \in L$, é criada uma cópia do conjunto $D_{u_{temp}DP}$, chamada D_{u_j} . Será aplicado um percentual x dos p exemplos selecionados ao número de exemplos que estará contido em D_{u_j} , e os conjuntos $D_{u_{temp}}$ e $D_{u_{temp}DP}$ serão descartados. Para cada rótulo $\lambda_j \in L$ é treinado um classificador C_j utilizando o conjunto de exemplos D_{l_j} . Para cada classificador C_j gerado, os exemplos do conjunto $D_{u_{temp}}$ são apresentados a C_j e rotulados de acordo com sua saída, de modo que, caso pertença a λ_j , é atribuído rótulo 1 e 0, caso contrário. Por fim, os p exemplos selecionados e recém-rotulados são movidos do conjunto $D_{u_{temp}DP}$ para o conjunto D_l . Ao fim de cada iteração, o processo será repetido, dando início a uma nova iteração, até que $|D_u| = \emptyset$.

Como dito anteriormente, em (SANTOS, 2012) foi proposto uma variação para os métodos semissupervisionados, chamada *Confidence*, com o intuito de amenizar o problema da atribuição automática de rótulos, haja vista o deterioramento do processo de aprendizado. Este trabalho propõe uma variação do SSBRr, chamada SSBRcr (do inglês, *Semi-Supervised Binary Relevance with confidence using ranking*), na qual é utilizado um fator de confiança, $conf$, onde $0 \leq conf \leq 1$ para controlar o processo de rotulagem automática, baseando-se pelo valor de saída do classificador gerado.

No Algoritmo 7 é apresentado a lógica para o algoritmo SSBRcr:

Algoritmo 7 Fluxo básico do SSBRcr na fase de treinamento

```

1: Dados  $D_{l_1}, D_{l_2}, \dots, D_{l_M}$  e  $D_u$ 
2: enquanto  $D_u \neq \emptyset$  faça
3:   Selecionar os  $p$  exemplos de  $D_u$  e movê-los para o conjunto  $D_{u_{temp}}$ , tal que
    $1 \leq p \leq |D_u|$ 
4:   Criar uma cópia do conjunto  $D_{u_{temp}}$ , chamada  $D_{u_{temp}DP}$ 
5:   Calcular a Confiabilidade a partir do Algoritmo 5
6:   Criar uma cópia do conjunto  $D_{u_{temp}DP}$  para cada rótulo  $\lambda_j \in L$ , chamada  $D_{u_j}$ 
7:   Aplicar o percentual  $x$  de  $p$  exemplos selecionados ao conjunto  $D_{u_j}$ 
8:   Descartar os conjuntos  $D_{u_{temp}}$  e  $D_{u_{temp}DP}$ 
9:   para Cada  $\lambda_j \in L$  faça
10:    Treinar o classificador  $C_j$  utilizando o conjunto de exemplos  $D_{l_j}$ 
11:    para Cada exemplo  $E_t \in D_{u_j}$ , tal que  $1 \leq t \leq |D_{u_j}|$  faça
12:      Apresentar  $E_t$  ao classificador  $C_j$ 
13:      se Saída do classificador  $C_j \leq conf$  então
14:        Atribuir rótulo 1;
15:      senão
16:        Atribuir rótulo 0;
17:      fim se
18:    fim para
19:    Mover os  $p$  exemplos recém-rotulados do conjunto  $D_{u_{temp}DP}$  para o conjunto  $D_l$ 
20:  fim para
21: fim enquanto

```

4.3.2) Modelos SSLP propostos

Outro método semissupervisionado chamado SSLP, criado a partir do algoritmo LP, tem como primeiro passo a transformação do conjunto de dados de treinamento, considerando as combinações de rótulos, gerando um metarrótulo q para cada combinação.

Algoritmo 8 Fluxo básico do SSLPr na fase de treinamento

- 1: Dados $D_{l_{tr}}$ e $D_{u_{tr}}$
 - 2: **enquanto** $D_{u_{tr}} \neq \emptyset$ **faça**
 - 3: Selecionar os p exemplos de $D_{u_{tr}}$ e movê-los para o conjunto $D_{u_{temp}}$, tal que $1 \leq p \leq |D_{u_{tr}}|$
 - 4: Criar uma cópia do conjunto $D_{u_{temp}}$, chamada $D_{u_{temp}DP}$
 - 5: **Calcular** a Confiabilidade a partir do Algoritmo 5
 - 6: Aplicar o percentual x de p exemplos selecionados ao conjunto $D_{u_{temp}TR}$
 - 7: Treinar o classificador C_i utilizando o conjunto de dados $D_{l_{tr}}$
 - 8: **para** Cada exemplo $E_t \in D_{u_{temp}TR}$, tal que $1 \leq t \leq |D_{u_{temp}TR}|$ **faça**
 - 9: Apresentar E_t ao classificador C_i
 - 10: Utilizar a saída do classificador C_i para atribuir um dos metarrótulos q
 - 11: **fim para**
 - 12: Mover os p exemplos recém-rotulados do conjunto $D_{u_{temp}}$, juntamente com os metarrótulos restantes, para o conjunto $D_{l_{tr}}$
 - 13: **fim enquanto**
-

Nesta subseção, é proposta uma extensão do método SSLP, chamada SSLPr (do inglês, *Semi-Supervised Label Powerse using ranking*), na qual, durante a fase de treinamento, p exemplos do conjunto $D_{u_{tr}}$ são selecionados e movidos para o conjunto $D_{u_{temp}}$, tal que $1 \leq p \leq |D_{u_{tr}}|$. Em seguida, é executado o Algoritmo 5 para calcular o desvio padrão dos valores de confidências das instâncias não-rotuladas a partir dos dados do conjunto $D_{u_{temp}}$. Esse percentual x , por sua vez, será aplicado ao conjunto que consta os maiores valores do desvio padrão dos exemplos previamente selecionados, chamado conjunto $D_{u_{temp}TR}$. O algoritmo 8, que representa o método SSLPr, mostra os seguintes passos da fase de treinamento desse algoritmo.

Dessa maneira, uma vez que tenham sido atribuídos rótulos a todos os p exemplos definidos pelo valor do desvio padrão das instâncias e, também, pertencentes ao conjunto de dados não-rotulados $D_{u_{tr}}$, o classificador C_i é retreinado utilizando seu recente conjunto de dados $D_{l_{tr}}$, composto de todos os exemplos do conjunto D_l , bem como com os exemplos do conjunto D_u e seus respectivos rótulos.

Como proposto ao método SSBRr, uma variação também ao método SSLPr foi proposta, chamada SSLPcr (do inglês, *Semi-Supervised Label Powerset with confidence using ranking*), com o objetivo de reduzir o surgimento de ruídos durante o processo de atribuição de rótulos aos exemplos do conjunto de dados não-rotulados D_u .

Algoritmo 9 Fluxo básico do SSLPcr na fase de treinamento

```

1: Dados  $D_{l_{tr}}$  e  $D_{u_{tr}}$ 
2: enquanto  $D_{u_{tr}} \neq \emptyset$  faça
3:   Selecionar os  $p$  exemplos de  $D_{u_{tr}}$  e movê-los para o conjunto  $D_{u_{temp}}$ , tal que
    $1 \leq p \leq |D_{u_{tr}}|$ 
4:   Criar uma cópia do conjunto  $D_{u_{temp}}$ , chamada  $D_{u_{tempDP}}$ 
5:   Calcular a Confiabilidade a partir do Algoritmo 5
6:   Aplicar o percentual  $x$  de  $p$  exemplos selecionados ao conjunto  $D_{u_{tempTR}}$ 
7:   Treinar o classificador  $C_i$  utilizando o conjunto de dados  $D_{l_{tr}}$ 
8:   para Cada exemplo  $E_t \in D_{u_{tempTR}}$ , tal que  $1 \leq t \leq |D_{u_{tempTR}}|$  faça
9:     Apresentar  $E_t$  ao classificador  $C_i$ 
10:    para cada  $t$  faça
11:      se Saída do classificador  $C_i \leq conf$  então
12:        Atribuir rótulo 1;
13:      senão
14:        Atribuir rótulo 0;
15:      fim se
16:    fim para
17:  fim para
18:  Mover os  $p$  exemplos recém-rotulados do conjunto  $D_{u_{temp}}$ , juntamente com os
  metarrótulos restantes, para o conjunto  $D_{l_{tr}}$ 
19: fim enquanto

```

4.3.3) Modelos SSRaKEL propostos

O método SSRaKEL, criado a partir de um algoritmo muito utilizado chamado RAKEL, é construído através de comitês (*ensembles*) de classificadores SSLP, na qual cada classificador é treinado usando um diferente subconjunto aleatório de *labelsets*, ou seja, cada classificador SSLP criado é aplicado em subtarefas com um número ideal de rótulo e número adequado de exemplos por rótulo. Dessa maneira, o método SSRaKEL tenta minimizar a vulnerabilidade à aparição de muitas classes com poucos exemplos do SSLP.

Nesta subseção é proposto uma extensão do método SSRaKEL, chamada SSRaKELr (do inglês, *Semi-Supervised Random k-Labelsets using ranking*) que usa os mesmos passos para construir um comitê de classificadores SSLP do SSRaKEL. O que diferencia o método SSRaKEL do SSRaKELr é apenas na maneira como são selecionados os exemplos para o processo de rotulagem de dados. Na fase de treinamento do SSRaKEL, cada classificador gerado é treinado usando um diferente subconjunto aleatório de *labelsets*; por sua vez, na fase de treinamento do SSRaKELr, os classificadores gerados são treinados a partir de um subconjunto de *labelsets* cujos valores do desvio padrão de cada exemplo foram imprescindíveis para a escolha dos exemplos a serem rotulados.

Assim como nos métodos anteriores, é proposta uma variação, chamada SSRaKELcr (do inglês, *Semi-Supervised Random k-Labelsets with confidence using ranking*), cujo objetivo é reduzir a inserção de ruído durante o processo de atribuição de rótulos aos exemplos do conjunto de dados não-rotulados D_u . Além disso, o que torna diferente os métodos SSRaKELr e o SSRaKELcr são os tipos de classificadores gerados, ou seja, o classificador SSLPr é gerado a partir do método SSRaKELr e, em contrapartida, o classificador SSLPcr é gerado a partir do método SSRaKELcr.

4.3.4) Modelos SSRaKELd propostos

Como mencionado anteriormente, o algoritmo RAKEL proposto em (TSOUMAKAS et al., 2010) é formado por um comitê (*ensemble*) de classificadores LP, no qual cada classificador é treinado usando um diferente conjunto aleatório de *labelsets*. O algoritmo RAKEL possui uma variação chamada de RAKELd, cuja principal diferença entre os dois métodos é o tratamento que cada algoritmo tem com os *labelsets*, ou seja, o algoritmo

RAkEL leva à sobreposição dos *labelsets* e o RAkELd leva à disjunção dos *labelsets*. A partir dessa ideia em (SANTOS, 2012), essa variação, chamada disjunção, será aplicado ao algoritmo SSRAkEL proposto. Essa variação transformará o método SSRAkEL no método SSRAkELd (do inglês, *Semi-Supervised Random k-Labelsets disjoint*). Os dois algoritmos propostos em (SANTOS, 2012) diferem apenas na forma como são utilizados na fase de seleção dos metarrótulos de cada *labelsets*, pois no SSRAkEL é permitida a sobreposição de metarrótulos e no SSRAkELd ocorre a disjunção dos *labelsets*.

Nesta subseção é proposto uma extensão do método SSRAkELd, chamada SSRAkELdr (do inglês, *Semi-Supervised Random k-Labelsets disjoint using ranking*) que usa os mesmos passos para construir um comitê de classificadores SSLP do SSRAkEL. O que diferencia o método SSRAkELd do SSRAkELdr é apenas na forma como são selecionados os exemplos para o processo de rotulagem de dados. O SSRAkELd promove uma seleção aleatória dos exemplos a cada iteração. Já o SSRAkELdr, por sua vez, irá atacar a escolha dos exemplos baseando-se no valor do desvio padrão de cada instância.

Com o intuito de reduzir a inserção de ruído durante o processo de atribuição de rótulos aos exemplos do conjunto D_u é proposto nesse trabalho uma variação do SSRAkELdr, chamada SSRAkELdcr (do inglês, *Semi-supervised Random k-Labelsets disjoint with confidence using ranking*), na qual é utilizado um fator de confiança para controlar o processo de atribuição automática de rótulos também baseada no valor do desvio padrão de cada exemplo.

4.4) Considerações Finais

Neste capítulo, foram apresentados algoritmos que combinam técnicas de aprendizado semissupervisionado com métodos de classificação multirrótulo. Além disso, durante a fase de treinamento, as escolhas dos exemplos a serem rotulados são feitas a partir dos resultados obtidos pelos maiores valores do desvio padrão de cada instância ou dos valores de confiança, ou seja, será selecionado um percentual de exemplos com maior dissimilaridade. Tais algoritmos foram chamados de SSBRr, SSLPr, SSRAkELr, SSRAkELdr, SSBRcr, SSLPcr, SSRAkELcr, SSRAkELdcr.

Capítulo 5

Metodologia dos Experimentos

Este capítulo descreve os procedimentos adotados no trabalho para a realização dos experimentos e está organizado da seguinte maneira: Na seção 6.1 são apresentadas as bases de dados multirrótulos. As subseções a seguir apresentam uma breve descrição de cada base de dados usada por essa proposta, configurações do experimento, além do conceito do parâmetro de confiabilidade.

5.1) Bases de Dados Multirrótulo

Para fins de experimentais, esse trabalho faz uso de dez diferentes domínios de aplicação. A seguir, são descritas as características e especificações das bases de dados utilizadas no trabalho.

Na Tabela 6.1 são apresentadas algumas características das bases de dados multirrótulo utilizadas, tais como o número de exemplos, o número de atributos numéricos, o número de atributos discretos e, por fim, o número de rótulos.

Base	Nº Exemplos	Nº Atrib. Num	Nº Atrib. Disc	Rótulos
Emotions	593	72	0	6
Scene	2712	294	0	6
Genbase	662	0	1186	27
Medical	978	0	1449	45
Yeast	2417	103	0	14
Bibtex	7395	0	1836	159
CAL500	502	68	0	174
Corel5k	5000	0	499	374
Enron	1702	0	1001	53
Mediamill	43907	120	0	101

Tabela 6.1 Detalhamento da base de dados.

5.1.1) *Emotions*

A base de dados *Emotions* (WIECZORKOWSKA, 2006) está relacionada com a classificação de músicas de acordo com as emoções provocadas por elas. Esta base de dados contém 593 instâncias, sendo 72 atributos contínuos, com possibilidade de associação a 6 rótulos (classes).

5.1.2) *Scene*

A base de dados de *Scene* (BOUTELL et al., 2004) está relacionada à indexação semântica de cenas. Essa base possui 2712 instâncias, onde cada instância é descrita por 294 atributos contínuos e, também, associadas com 6 conceitos (rótulos), tais como praia, montanha e campo.

5.1.3) *Genbase*

A base de dados *Genbase* (DIPLARIS et al., 2005) representa dados biológicos a qual estão relacionados à classificação funcional de proteínas. Essa base de dados possui um quantitativo de 662 instâncias, descritas por 1186 atributos discretos, com possibilidade de associação a 27 rótulos (classes).

5.1.4) *Medical*

Essa base de dados é composta de documentos que descrevem histórias de sintomas e prognósticos que são usados para prever códigos seguros e está disponível em <http://www.computationalmedicine.org/challenge/index.php>. A composição desta base de dados são 978 instâncias, descritas por 1449 atributos discretos, com possibilidade de associação a 45 rótulos (classes).

5.1.5) *Yeast*

Essa base de dados biológicos *Yeast* (CLARE; KING, 2001) está relacionada à classificação de funções de proteínas. Em sua composição encontramos microvetores de

expressões e perfis filogenéticos de 2417 genes *yeast*, descritos por 103 atributos contínuos. Cada é associado a um subconjunto de 14 categorias.

5.1.6) *Bibtex*

A base bibtex utiliza um arquivo em formato de texto simples, que pode ser criado e modificado através de um arbitrário editor de texto. Essa base de dados é composta por 7395 instâncias, 1836 atributos discretos e 159 rótulos. Essa base está disponível em <http://www.bibtex.org/>.

5.1.7) *CAL500*

Essa base de dados é a composição em um conjunto de 1.700 anotações musicais do ser humano que, gerados, descrevem 500 faixas musicais populares do ocidente. A CAL500 é definida por suas 502 instâncias, 68 atributos numéricos, bem como seus 174 rótulos. Ela está disponível em <http://cosmal.ucsd.edu/cal/projects/AnnRet/>.

5.1.8) *Corel5k*

A corel5k é uma base de dados representada por um diretório no qual contém dados utilizados para reconhecimento de objetos como *Machine Translation*, de Pinar Duygulu, Kobus Barnard, Nando Freitas e David Forsyth. Essa base é composta por dados redundantes e com algumas anomalias. Essa base é formada por 5000 instâncias, 499 atributos discretos e 374 rótulos.

5.1.9) *Enron*

A base enron representa um conjunto de dados que foi coletado e preparado pelo projeto CALO – assistência cognitiva que aprende e organiza –. Essa base contém os dados de 150 usuários organizados em pastas. Essa *dataset* é formada por 1702 instâncias, 1001 atributos discretos e 53 rótulos.

5.1.10) *Mediamill*

Essa base de dados deu origem no laboratório de estudos em sistemas inteligentes da Universidade de Amsterdam e tem como objetivo traduzir pixel para texto com o propósito de recuperação de imagens. A base mediamill é composta por 43907 instâncias, 120 atributos numéricos e 101 rótulos.

5.2) Métricas de Avaliação

Na literatura é possível encontrar várias medidas de classificação multirrótulo. Segundo (TSOUMAKAS et al., 2009), estas medidas propostas podem ser desmembradas em dois grupos: As medidas baseadas em *bipartição*; que avaliam as *bipartições* sobre todos os exemplos da base de dados de avaliação, além de, em outros casos, decompor o processo de avaliação em avaliações separadas para cada rótulo. O outro grupo refere-se às medidas baseadas em *ranking*; que avaliam os *rankings* com respeito à precisão da base de dados multirrótulo.

Sabe-se, por isso, que a maneira como são avaliados os classificadores multirrótulo diferem da forma como são avaliados os classificadores unirrótulos. Segundo (SANTOS, 2012), ao contrário dos problemas de classificação unirrótulo, cuja classificação dos exemplos estão entre correto ou incorreto, em um problema de classificação multirrótulo a classificação de um exemplo pode ser considerada parcialmente correta ou incorreta, ou seja, quando um classificador corretamente atribui um exemplo a ao menos uma das classes que ele pertence, mas não atribui a todas as classes que de fato ele é pertencente. Aproveitando o contexto, segundo (CERRI; CARVALHO, 2009), o classificador também pode atribuir um exemplo a uma ou mais classes a que ele não pertence.

Portanto, segundo (SANTOS, 2012), para definir estas medidas, uma base de dados de avaliação de exemplos multirrótulo será denotada como $(x_i, y_i), i = 1, \dots, N$, onde $Y_i \subseteq L$ representa o conjunto de rótulos verdadeiros e $L = \{\lambda : j = 1, \dots, M\}$ representa o conjunto de todos os rótulos. Dado um exemplo x_i , o conjunto de rótulos que são preditos por um

classificador multirrótulo é representado por Z_i , enquanto o *ranking* predito para um rótulo λ é representado por $r_i(\lambda)$.

Quanto maior a relevância dos rótulos, maior será o *ranking* atribuído ao rótulo. Quanto menor a relevância dos rótulos, menor será o *ranking* atribuído ao rótulo.

A seguir, serão apresentadas algumas medidas de avaliação usadas pela literatura.

5.2.1) Métricas baseadas em *Bipartição*

Este trabalho faz uso das seguintes medidas de *Bipartição*: *Hamming Loss* (HLoss), *F-Measure* (FM) e *Accuracy* (Acc).

5.2.1.1) *Hamming Loss*

O *Hamming Loss* avalia a porcentagem correta de rótulos não preditos e rótulos preditos incorretamente. Dessa maneira, esta medida considera a frequência com que um par exemplo/rótulo é classificado de maneira incorreta, ou seja, cada exemplo é associado a um rótulo errado ou um rótulo não predito. Esta medida define que o melhor desempenho é alcançado quando o resultado é igual a zero, isto é, quanto mais o valor se aproxima a zero, melhor será seu desempenho.

$$HLoss = \frac{1}{N} \sum_{i=1}^N \frac{|Y_i \Delta Z_i|}{M}$$

Δ é a diferença simétrica entre os dois conjuntos

Equação da fórmula matemática da métrica *Hamming Loss*.

5.2.1.2) *F-Measure*

F-Measure é uma métrica que representa a combinação entre Precisão e Reconhecimento. Sua representação é definida pela média harmônica das duas métricas, *Precision* e *Recall*, e é usada como um *score* de desempenho agregado.

$$FM = \frac{1}{N} \sum_{i=1}^N \frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|}$$

Equação da fórmula matemática da métrica *F-Measure*.

5.2.1.3) *Accuracy*

Essa métrica avalia a porcentagem de verdadeiros positivos fora do total de verdadeiros positivos, falso positivos e negativos.

$$Acc = \frac{1}{N} \sum_{i=1}^N \frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}$$

Equação da fórmula matemática da métrica *Accuracy*.

5.2.2) Métricas baseadas em *Ranking*

Este trabalho faz uso das seguintes medidas de *Ranking*: *One-Error* (1-Err), *Average Precision* (AvPrec) e *Ranking Loss* (RLoss).

5.2.2.1) *One-Error*

One-Error, também conhecida por *1-Err*, avalia a frequência de rótulos no topo do ranking que não estavam no conjunto de rótulos (*relevantes*) corretos. Nessa métrica, o melhor desempenho é alcançado quando essa medida vai se aproximando a zero. Portanto, quanto menor o valor alcançado, melhor o desempenho obtido.

$$1 - Err = \frac{1}{N} \sum_{l=1}^{|N|} \delta(\arg_{\lambda \in L} \min r_l(\lambda))$$

Onde:

$$\delta(\lambda) = \begin{cases} 1 & \text{se } \lambda \in Y_i \\ 0 & \text{caso contrário} \end{cases}$$

Equação da fórmula matemática da métrica *One-Error*.

5.2.2.2) *Average Precision*

A métrica *Average Precision*, conhecida por *AvPrec*, avalia a fração média de rótulos rankeados acima de uma rótulo particular $l \in Y_i$ que está em Y_i . O melhor resultado é alcançado quando a precisão média é igual a um. Ou seja, essa medida de avaliação representa a precisão média tomada para todos os possíveis rótulos.

$$\text{avgprec}_s(H) = \frac{1}{N} \sum_{i=1}^N \frac{1}{|Y_i|} \sum_{l \in Y_i} \frac{|\{l' \in Y_i | \text{rank}_f(x_i, l') \leq \text{rank}_f(x_i, l)\}|}{\text{rank}_f(x, l)}$$

Equação da fórmula matemática da métrica *Average Precision*.

5.2.2.3) *Ranking Loss*

A *Ranking Loss*, ou *RLoss*, expressa o número de vezes que os rótulos irrelevantes são classificados em uma posição mais alta no *ranking* comparado aos rótulos relevantes. Ou seja, essa medida de avaliação expressa uma média de pares de rótulos ordenados de maneira incorreta.

$$\text{R-Loss} = \frac{1}{N} \sum_{i=1}^N \frac{1}{|Y_i| |\overline{Y_i}|} \cdot |R_i|$$

Onde: $\overline{Y_i}$ é o conjunto complementar de Y_i com seu respectivo rótulo L . E

$$R_i = \{(\lambda_a, \lambda_b) : r_i(\lambda_a) > r_i(\lambda_b), (\lambda_a, \lambda_b) \in Y_i \times \overline{Y_i}\}$$

Equação da fórmula matemática da métrica *Ranking Loss*.

5.3) Parâmetro de Confiabilidade

Como mencionado na Seção 1.1, a principal motivação para esse trabalho é encontrar a solução para o problema da aleatoriedade na escolha dos dados para o processo de atribuição de rótulos. Em (SANTOS, 2012), o uso dos valores de confiança nos exemplos mostraram que parte dos resultados alcançaram uma significância subestimada. Uma possível causa para isso pode ter sido a maneira aleatória como os dados foram escolhidos para rotular. É possível, portanto, que a escolha aleatória dos dados pode ter

escolhido exemplos de uma mesma região do espaço de dados o que, conseqüentemente, pode ter contribuído de maneira negativa na obtenção dos resultados. Buscando minimizar o problema na escolha de exemplos similares, surgiu a ideia do parâmetro de confiabilidade como mecanismo usado no auxílio da classificação de dados multirrótulos através da atribuição de pesos. Essa atribuição de pesos dá-se através do cálculo do desvio padrão dos valores de confiança das instâncias onde espera-se, portanto, que os exemplos com alto valor do desvio padrão sejam menos similares comparados com os exemplos com menor valor do desvio padrão. Tomando como base a relevância do uso do parâmetro de confiabilidade, na Seção 8.1 um artigo foi publicado apresentando os resultados obtidos a partir de experimentos iniciais.

5.4) Redução de Dimensionalidade (*Feature Selection*)

A partir de técnicas de mineração de dados, o qual é constituído por três etapas: pré-processamento, extração de padrões e pós-processamento (HAN; KAMBER, 2006), a Seleção de Atributos (do inglês *Feature Selection*) é uma tarefa presente na etapa de pré-processamento. A Seleção de Atributos pode contribuir para a geração de modelos de dados que permitem representar com maior fidelidade padrões inerentes aos dados. Além disso, a Seleção de Atributos busca atributos que se destacam a partir de medidas ou critérios de importância, oferecendo, dessa maneira, um potencial de redução nos efeitos prejudiciais da chamada “maldição da dimensionalidade” (LIU; MOTODA, 2008).

Uma vantagem dessa abordagem é a eficiência para a seleção de atributos em dados multirrótulos, visto que a existência de um único subconjunto de atributos do conjunto de dados elimina a mensuração dos atributos descartados.

Por fazer uso de bases de dados com grande dimensionalidade de atributos, percebeu-se um potencial problema quanto a aquisição dos resultados satisfatórios, haja vista a grande quantidade de informações, muitas vezes desnecessárias, presentes nessas bases. Para esse trabalho, portanto, foram aplicados os conceitos de redução da dimensionalidade em base de dados a partir dos trabalhos elaborados por (SPOLAÔR; CHERMAN; MONARD, 2011) e (SPOLAÔR, CHERMAN; MONARD; LEE, 2013). Conceitos esclarecedores foram imprescindíveis como auxílio na aplicação do algoritmo de seleção de atributos chamado *ReliefF* para bases de dados com grande complexidade.

O algoritmo *ReliefF* busca definir um *ranking* final a partir da média dos valores da medida de importância para cada atributo em cada conjunto de dados multirrótulo. A avaliação dos resultados obtidos é dada a partir dos melhores atributos, dado um limiar.

Outro objetivo do algoritmo *ReliefF*, em problemas de classificação, pode ser especificado como a seleção de atributos que exibam valores distintos para exemplos próximos de um grupo de classes c distintas e valores similares para exemplos de mesma classe (SIKONJA, 2003). O *ReliefF* é usualmente aplicável em problemas de classificação multirrótulos, pois apresenta maior robustez a dados ruidosos e incompletos (DERNSAR, 2010). Além disso, é também um algoritmo aplicável em problemas multirrótulos de acordo com a transformação do problema, visto que o conjunto de dados multirrótulos, em análise, seja extraído em m conjuntos de dados monorrótulos.

O uso do algoritmo *ReliefF* neste trabalho tem como objetivo reduzir a dimensionalidade desses conjuntos de dados. A ideia da redução de dimensionalidade dessas bases surgiu a partir da escolha das mesmas, o que, possivelmente, minimizaria o tempo de processamento para a aquisição dos resultados, além de eliminar atributos não relevantes que comprometem o alcance de resultados satisfatórios. Como mencionado anteriormente, esse trabalho faz uso dos conceitos desenvolvidos por (SPOLAÔR; CHERMAN; MONARD, 2011) e (SPOLAÔR, CHERMAN; MONARD; LEE, 2013) como critério para a redução das seguintes bases de dados: *bibtex*, *CAL500*, *corel5k*, *enron*, *mediamill*.

Após a aplicação da redução de dimensionalidade das bases multirrótulos citadas anteriormente, segue a tabela ilustrativa com os novos valores para os atributos e rótulos.

Base	N° Atributos Numéricos		N° Atributos Discretos	
	Antes	Depois	Antes	Depois
Bibtex	0	0	1836	781
CAL500	68	57	0	0
Corel5k	0	0	499	148
Enron	0	0	1001	376
Mediamill	120	79	0	0

Tabela 6.2 Tabela de bases de dados multirrótulos com dimensionalidade reduzida

5.5) Configuração dos experimentos

Neste trabalho foram utilizados os seguintes parâmetros para cada métodos de classificação multirrótulo proposto: SSBRr, SSLPr, SSRAkELr, SSRAkELdr, SSBRcr, SSLPcr, SSRAkELcr e SSRAkELdcr.

- Classificadores Base
 - o K-NN e SVM
- Percentual de dados rotulados inicialmente
 - o 25% e 75%
- Percentual de dados não-rotulados para rotular
 - o 17% e 33.3%
- Fator de confiança
 - o 0.6, 0.7, 0.8 e 0.9
- Metodologia de validação cruzada
 - o *10-fold cross-validation*
- Medidas de Avaliação
 - o Bipartição
 - Hamming Loss, F-Measure e Accuracy
 - o Ranking
 - One-Error, Average Precision e Ranking Loss
 - o Nível de Confidência
 - 95%
 - o Teste Estatístico
 - T-Test
 - o Limiar de importância para o *ReliefF*
 - 0.01

Com o objetivo de analisar os resultados apresentados pelos diferentes métodos propostos, esse trabalho fez uso de testes estatísticos. Os testes estatísticos são bastante utilizados na literatura como mecanismos que aprovam/validam experimentos e comparações dos resultados. O teste estatístico utilizado por este trabalho foi o T-Teste, na qual realiza a comparações entre duas amostras (LUO; ZINCIR, 2005), que neste caso irá comparar a significância dos resultados alcançados a partir das versão com e sem parâmetro de confiabilidade.

5.6) Considerações Finais

Neste capítulo foram apresentados algumas das principais bases de dados multirrótulo utilizada na literatura. Para o desenvolvimento deste trabalho, bem como os experimentos, foram usadas as seguinte bases: Emotion, Scene, Genbase, Medical, Yeast, Bibtex, CAL500, Corel5k, Enron e Mediamill. Também, neste capítulo, foram apresentadas algumas das principais métricas de avaliação, três baseadas em bipartição e três baseadas em ranking. Ademais, conceitos e algoritmo de redução de dimensionalidade foram apresentados neste trabalho para auxiliar na compreensão desses mecanismo. E, por fim, foi apresentado as configurações dos experimentos que foram aplicadas para este trabalho.

Capítulo 6

Resultados Experimentais

Como mencionado anteriormente, esse trabalho usa valores de um parâmetro de confiabilidade em problemas de classificação multirrótulo com aprendizado semissupervisionado. O objetivo é possibilitar a análise do impacto desse parâmetro no comportamento dos métodos de classificação multirrótulo na abordagem semissupervisionada, cuja classificação das instâncias seja baseada a partir do *ranking* dos valores de confiança das instâncias. Além das diferentes abordagens de aprendizado, bem como os diferentes métodos de classificação utilizados por esse trabalho, os experimentos fizeram uso de diferentes parâmetros, tais como: percentual inicial de exemplos rotulados, percentual de exemplos não-rotulados a cada iteração, fator de confiança. Dessa maneira, uma análise empírica foi realizada sobre as perspectivas gerais e é dividida em oito partes, onde os métodos propostos serão comparados com os métodos correspondentes que não usam parâmetros de confiabilidade.

Nessa primeira parte, as proporções de exemplos não-rotulados que deverão ser rotulados a cada iteração são de 17% e 33%. Cada representação será dada a partir dos percentuais de dados rotulados inicialmente que são de 25% e 75%. Para cada parte desta análise, serão apresentados dois gráficos estatísticos a fim de definir qual método multirrótulo semissupervisionado baseado no valor de confiança, ou não, que apresentou o melhor desempenho.

6.1) Percentual de 25% de dados rotulados iniciais

Nessa seção será apresentado os resultados para os métodos semissupervisionados derivados do BR, LP, RAkEL e RAkELd com o percentual de dados rotulados inicialmente de 25%.

HLss↓	0,202	0,193	0,202	0,194	0,206	0,175	0,173	0,206
F-M↑	0,494	0,523	0,548	0,546	0,575	0,581	0,577	0,570
Acc↑	0,490	0,523	0,490	0,522	0,493	0,502	0,498	0,490
1-Err↓	0,253	0,238	0,251	0,230	0,228	0,252	0,256	0,281
AvPr↑	0,736	0,759	0,736	0,762	0,748	0,745	0,748	0,744
RLoS↓	0,209	0,178	0,209	0,176	0,186	0,185	0,183	0,184

Tabela 7.1 Desempenho dos métodos Multirrótulos, para cada versão do SSBR.

Pode-se observar a partir dos resultados extraídos pela Tabela 7.1 que a utilização do parâmetro de confiabilidade no processo semissupervisionado teve um impacto positivo no desempenho de seus métodos. Para o método SSBRr, este, por sua vez, alcançou 13 entre 30 casos possíveis e, conseqüentemente, foi o método que teve o melhor desempenho com cerca de 43% dos melhores resultados. Em segundo lugar encontra-se o método SSBRcr com 40% dos melhores resultados, ou seja, alcançou 12 entre 30 casos possíveis. Os métodos sem parâmetros de confiabilidade alcançaram cerca de 20% dos melhores casos, para o SSBR, e 13% dos melhores casos para o SSBRc. Destaca-se, portanto, que dos seis conjuntos de dados, os métodos baseados em parâmetros de confiabilidade alcançaram a maioria dos melhores resultados nas bases *Emotions* e *Medical*, por exemplo, o uso do valor de confiança resultou em uma melhoria no desempenho do SSBR em todos os seis parâmetros de avaliação. Esses resultados já eram esperados, uma vez que os métodos SSBRr e SSBRcr fazem uso de uma técnica mais elaborada na forma de identificar os exemplos não-rotulados.

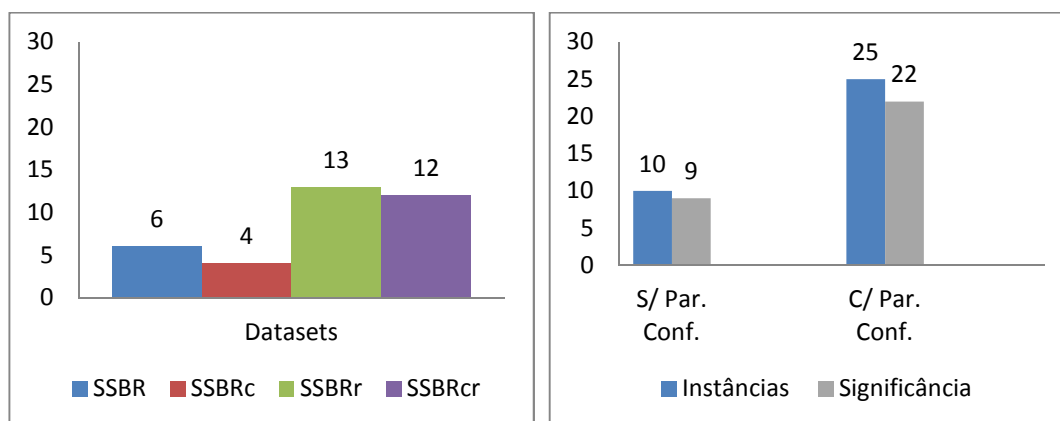


Figura 7.1 Análise dos métodos baseados no SSBR.

Analisando o percentual de melhores casos alcançados pelos métodos que fizeram ou não uso do parâmetro de confiabilidade, pode-se destacar que os métodos que fizeram uso do

parâmetro de confiabilidade alcançaram um percentual de 83% (SSBRr e SSBRcr) dos melhores casos possíveis comparado com 33% dos melhores casos para os métodos que não fizeram uso do parâmetro de confiabilidade (SSBR e SSBRc). Quanto aos resultados alcançados a partir dos percentuais de exemplos não-rotulados que deverão ser rotulados a cada iteração, que são de 17% e 33.3%, a maioria dos métodos analisados obtiveram os melhores resultados a partir do percentual de 33.3% dos exemplos não-rotulados que deverão ser rotulados a cada iteração com um percentual de 67%, ou seja, 80 dos 120 possíveis casos comparados com o percentual de 17% que foi 24%, ou seja, 29 dos 120 possíveis casos.

6.1.2) Métodos baseados no SSLP

A Tabela 7.2 ilustra o desempenho dos métodos SSLP, SSLPc, SSLPr e SSLPcr usando todos os cinco diferentes conjuntos de dados. Como mencionado anteriormente, todos estes métodos foram avaliados utilizando-se seis parâmetros de avaliação diferentes, além de um percentual de 25% de dados não rotulados. Para esta análise, os melhores resultados obtidos em cada métrica estão destacados em negrito e, para definir o melhor resultado entre os métodos em análise, a célula vencedora será destacada com a cor cinza. Além disso, os resultados estatisticamente mais significantes foram marcados com sublinhado.

Analisando a Tabela 7.2, pode-se observar que, assim como no método proposto anterior, os métodos com parâmetro de confiabilidade alcançaram resultados superiores comparados aos resultados atingidos pelos métodos sem parâmetro de confiabilidade. Tal constatação é observada pelo método que melhor desempenho teve, que foi o SSLPcr com 37% dos melhores casos com a maioria dos casos em 11 dos 30 casos possíveis. Em seguida, e de maneira bem equilibrada, vieram os métodos SSLPc, SSLPr e SSPL com 30%, 24% e 20% dos melhores casos respectivamente. Destaca-se, portanto, o bom desempenho alcançado pelo método SSLPc na base de dados *Scene* por atingir a maioria dos casos. Outro aspecto interessante a ser observado é que, para essas cinco bases de dados multirrótulo analisadas, os métodos baseados no SSLP que tem valores de confiança, ou seja, SSLPc e SSLPcr alcançaram os melhores dos 30 casos possíveis.

Analisando, também, o percentual de melhores casos alcançados pelos métodos que fizeram ou não uso do parâmetro de confiabilidade, pode-se destacar que os métodos que fizeram uso do parâmetro de confiabilidade alcançaram um percentual de 60% (SSLPr e SSLPcr) dos melhores casos comparado com 40% dos melhores casos para os métodos que não fizeram uso do parâmetro de confiabilidade (SSLP e SSLPc). Quanto aos resultados alcançados a partir dos percentuais de exemplos não-rotulados que deverão ser rotulados a cada iteração, que são de 17% e 33.3%, a maioria dos métodos analisados obtiveram os melhores resultados a partir do percentual de 33.3% dos exemplos não-rotulados que deverão ser rotulados a cada iteração com um percentual de 72%, ou seja, 87 dos 120 possíveis casos comparados com o percentual de 17% que foi 24%, ou seja, 29 dos 120 possíveis casos.

	SSLP		SSLPc		SSLPr		SSLPcr	
	17%	33.3%	17%	33.3%	17%	33.3%	17%	33.3%
Measure	EMOTIONS							
HLss↓	0,272	0,235	0,274	0,242	0.235	0.245	0.229	0.233
F-M↑	0,572	0,611	0,580	0,625	0.622	0.615	0.638	0.635
Acc ↑	0,441	0,520	0,444	0,500	0.512	0.502	0.527	0.519
1-Err↓	0,431	0,366	0,450	0,394	0.740	0.728	0.733	0.726
AvPr↑	0,680	0,723	0,672	0,708	0.770	0.770	0.765	0.776
RLoS↓	0,314	0,277	0,325	0,286	0.617	0.616	0.632	0.614
	GENBASE							
HLss↓	0,002	0,003	0,006	0,002	0.004	0.004	0.002	0.005
F-M↑	0,982	0,990	0,982	0,988	0.734	0.732	0.753	0.726
Acc ↑	0,972	0,984	0,947	0,977	0.731	0.731	0.751	0.726
1-Err↓	0,026	0,026	0,040	0,012	0.007	0.014	0.003	0.003
AvPr↑	0,974	0,981	0,955	0,983	0.933	0.919	0.932	0.926
RLoS↓	0,015	0,011	0,029	0,019	0.013	0.011	0.095	0.096
	MEDICAL							
HLss↓	0,029	0,021	0,028	0,020	0.013	0.014	0.016	0.015
F-M↑	0,500	0,641	0,700	0,723	0.729	0.706	0.680	0.690
Acc ↑	0,401	0,567	0,425	0,583	0.519	0.586	0.404	0.481
1-Err↓	0,589	0,545	0,559	0,544	0.474	0.495	0.445	0.442
AvPr↑	0,452	0,472	0,467	0,475	0.263	0.249	0.295	0.297
RLoS↓	0,265	0,339	0,270	0,346	0.529	0.530	0.471	0.466
	SCENE							
HLss↓	0,127	0,096	0,124	0,094	0.108	0.107	0.111	0.109
F-M↑	0,730	0,748	0,741	0,748	0.782	0.785	0.772	0.780
Acc ↑	0,626	0,715	0,633	0,720	0.781	0.783	0.770	0.772
1-Err↓	0,381	0,316	0,369	0,307	0.402	0.391	0.396	0.395
AvPr↑	0,723	0,766	0,731	0,771	0.702	0.709	0.708	0.710
RLoS↓	0,251	0,219	0,241	0,215	0.288	0.279	0.277	0.274
	YEAST							
HLss↓	0,263	0,249	0,265	0,245	0.233	0.233	0.231	0.232
F-M↑	0,481	0,507	0,567	0,475	0.592	0.594	0.596	0.598
Acc ↑	0,448	0,468	0,442	0,474	0.486	0.488	0.487	0.785
1-Err↓	0,331	0,329	0,322	0,319	0.868	0.860	0.874	0.857
AvPr↑	0,647	0,657	0,645	0,663	0.374	0.375	0.368	0.377
RLoS↓	0,325	0,313	0,329	0,308	0.245	0.281	0.254	0.284

Tabela 7.2 Desempenho dos métodos Multirrótulos, para cada versão do SSLP.

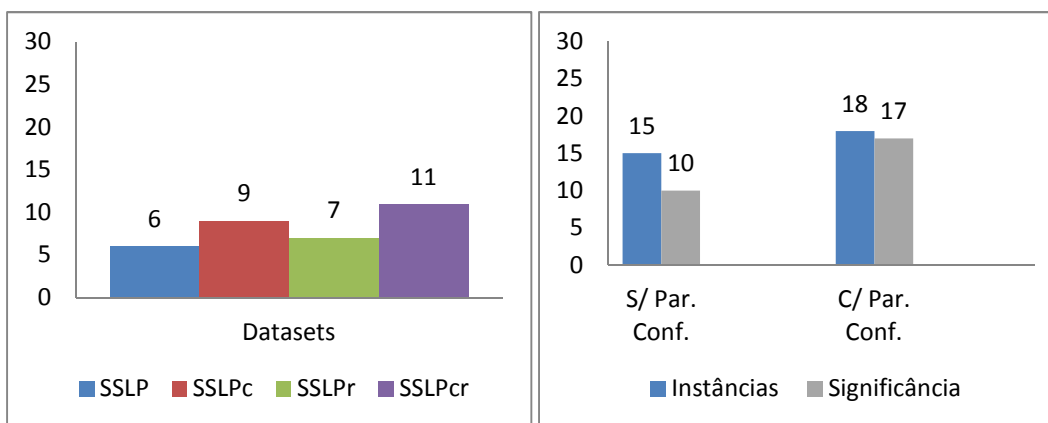


Figura 7.2 Análise dos métodos baseados no SSLP.

6.1.3) Métodos baseados no SSRAkEL

A Tabela 7.3 apresenta o desempenho do SSRAkEL, SSRAkELc, SSRAkELr e SSRAkELcr, também usando cinco diferentes conjuntos de dados. Mais uma vez, todos os resultados foram obtidos através do uso de seis métricas de avaliação diferentes, além do percentual de 25% de dados inicialmente rotulados. Para essa comparação, os melhores resultados obtidos em cada métrica estão em **negrito** e, para definir o melhor resultado entre os métodos em análise, a célula vencedora será destacada com a cor cinza. Além disso, os resultados estatisticamente mais significantes foram marcados com sublinhado.

Ao analisar o desempenho dos métodos derivados do SSRAkEL na Tabela 7.3, é possível observar que os melhores resultados na maioria dos casos, 17 dos 30 casos, foi obtido pelo método SSRAkELr baseados no parâmetro de confiabilidade usando *ranking*, ou seja, um percentual de 57% dos melhores casos. Além disso, outro método baseado no parâmetro de confiabilidade, o SSRAkELcr, também alcançou resultados satisfatórios, atingindo o percentual de 24% dos melhores casos, ou seja, 7 dos 30 casos possíveis. Já o método SSRAkEL e SSRAkELc, ambos alcançaram o mesmo percentual de melhores casos que foi representado por 16%, ou seja, 5 dos 30 casos possíveis foram alcançados por esses dois métodos.

Analisando o percentual de melhores casos alcançados pelos métodos que fizeram ou não uso do parâmetro de confiabilidade, pode-se destacar, portanto, que os métodos que

fizeram uso do parâmetro de confiabilidade alcançaram um percentual de 80% (SSRAkELr e SSRAkELcr) dos melhores casos comparado com 33% dos melhores casos para os métodos que não fizeram uso do parâmetro de confiabilidade (SSRAkEL e SSRAkELc). Quanto aos resultados alcançados a partir dos percentuais de exemplos não-rotulados que deverão ser rotulados a cada iteração, que são de 17% e 33.3%, a maioria dos métodos analisados obtiveram os melhores resultados a partir do percentual de 33.3% dos exemplos não-rotulados que deverão ser rotulados a cada iteração com um percentual de 67%, ou seja, 80 dos 120 possíveis casos comparados com o percentual de 17% que foi 39%, ou seja, 46 dos 120 possíveis casos.

	SSRAkEL		SSRAkELc		SSRAkELr		SSRAkELcr	
	17%	33.3%	17%	33.3%	17%	33.3%	17%	33.3%
Measure	EMOTIONS							
HLss↓	0,257	0,248	0,240	0,242	0,234	0,233	0,237	0,229
F-M↑	0,418	0,625	0,610	0,722	0,642	0,641	0,622	0,638
Acc ↑	0,471	0,492	0,498	0,502	0,520	0,522	0,508	0,541
1-Err↓	0,406	0,396	0,393	0,398	0,332	0,340	0,372	0,344
AvPr↑	0,697	0,702	0,712	0,708	0,734	0,722	0,718	0,721
RLos↓	0,300	0,294	0,257	0,289	0,275	0,270	0,279	0,264
	GENBASE							
HLss↓	0,020	0,019	0,019	0,019	0,010	0,011	0,019	0,018
F-M↑	0,800	0,800	0,800	0,800	0,821	0,797	0,793	0,811
Acc ↑	0,528	0,544	0,545	0,544	0,599	0,599	0,592	0,597
1-Err↓	0,462	0,439	0,439	0,441	0,378	0,378	0,387	0,371
AvPr↑	0,558	0,571	0,572	0,571	0,640	0,643	0,637	0,638
RLos↓	0,370	0,364	0,363	0,363	0,458	0,460	0,462	0,461
	MEDICAL							
HLss↓	0,140	0,173	0,027	0,027	0,023	0,025	0,023	0,024
F-M↑	0,100	0,200	0,003	0,003	0,220	0,222	0,205	0,216
Acc ↑	0,189	0,230	0,003	0,003	0,215	0,219	0,202	0,208
1-Err↓	0,758	0,717	0,952	0,952	0,726	0,725	0,733	0,735
AvPr↑	0,292	0,320	0,116	0,115	0,317	0,330	0,308	0,318
RLos↓	0,387	0,374	0,322	0,545	0,371	0,372	0,375	0,370
	SCENE							
HLss↓	0,579	0,636	0,120	0,115	0,177	0,160	0,108	0,118
F-M↑	0,700	0,758	0,433	0,433	0,669	0,664	0,658	0,670
Acc ↑	0,607	0,669	0,428	0,451	0,502	0,508	0,625	0,632
1-Err↓	0,356	0,298	0,475	0,454	0,224	0,224	0,259	0,258
AvPr↑	0,748	0,784	0,663	0,679	0,817	0,822	0,840	0,841
RLos↓	0,214	0,187	0,295	0,277	0,219	0,217	0,217	0,218
	YEAST							
HLss↓	0,023	0,022	0,288	0,289	0,225	0,227	0,227	0,229
F-M↑	0,532	0,538	0,265	0,057	0,547	0,540	0,545	0,546
Acc ↑	0,346	0,350	0,089	0,072	0,399	0,399	0,396	0,398
1-Err↓	0,332	0,338	0,738	0,746	0,264	0,260	0,261	0,255
AvPr↑	0,581	0,580	0,468	0,462	0,610	0,617	0,608	0,608
RLos↓	0,415	0,417	0,493	0,500	0,396	0,399	0,400	0,417

Tabela 7.3 Desempenho dos métodos Multirrótulos, para toda versão do SSRAkEL.

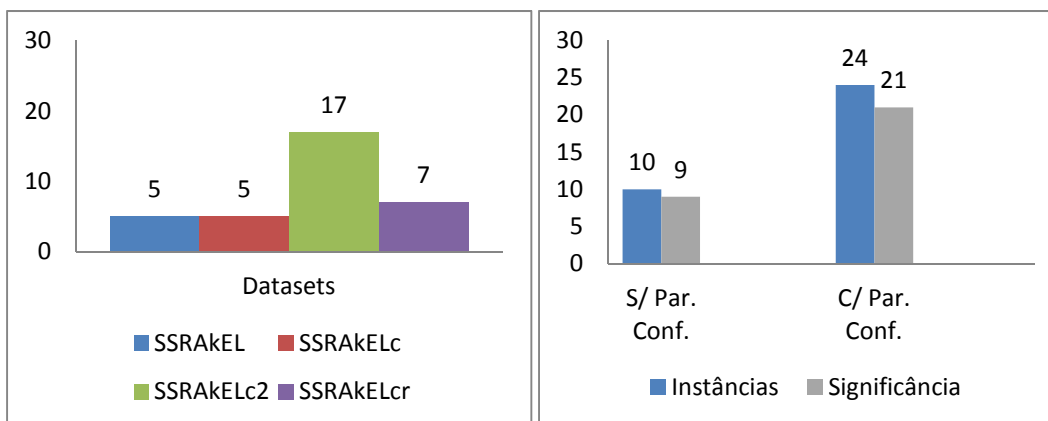


Figura 7.3 Análise dos métodos baseados no SSRaKEl.

6.1.4) Métodos baseados no SSRaKEld

A Tabela 7.4 apresenta o desempenho do SSRaKEld, SSRaKEldc, SSRaKEldr e SSRaKEldcr, também usando cinco diferentes conjuntos de dados. Mais uma vez, todos os resultados foram obtidos através do uso de seis métricas de avaliação diferentes, além do percentual de 25% de dados inicialmente rotulados. Para essa comparação, os melhores resultados obtidos em cada métrica estão em **negrito** e, para definir o melhor resultado entre os métodos em análise, a célula vencedora será destacada com a cor cinza. Além disso, os resultados estatisticamente mais significantes foram marcados com sublinhado.

Ao analisar o desempenho dos métodos baseados SSRaKEld na Tabela 7.4, é possível observar que os melhores resultados na maioria dos casos, 13 dos 30 casos, foi obtido pelo método SSRaKEldcr baseados no parâmetro de confiabilidade usando *ranking*, ou seja, um percentual de 43% dos melhores casos. Além disso, outro método baseado no parâmetro de confiabilidade, o SSRaKEldr, também alcançou resultados satisfatórios, atingindo o percentual de 27% dos melhores casos, ou seja, 8 dos 30 casos possíveis. Já o método SSRaKEld, este, por sua vez, alcançou o percentual de 30% dos melhores casos com 9 dos 30 possíveis casos e, por último, o métodos que apresentou o pior desempenho foi o SSRaKEldc com apenas 13% dos melhores casos, ou seja, 4 dos 30 casos possíveis.

	SSRAkELd		SSRAkELdc		SSRAkELdr		SSRAkELdcr	
	17%	33.3%	17%	33.3%	17%	33.3%	17%	33.3%
Measure	EMOTIONS							
HLss↓	0,258	0,246	0,267	0,250	0,225	0,227	0,229	0,223
F-M↑	0,618	0,627	0,590	0,608	0,555	0,548	0,697	0,545
Acc ↑	0,473	0,489	0,460	0,489	0,473	0,477	0,489	0,455
1-Err↓	0,397	0,376	0,399	0,379	0,539	0,540	0,593	0,587
AvPr↑	0,694	0,708	0,689	0,707	0,733	0,777	0,771	0,778
RLos↓	0,305	0,296	0,304	0,295	0,353	0,352	0,446	0,476
	GENBASE							
HLss↓	0,013	0,011	0,019	0,191	0,016	0,016	0,010	0,018
F-M↑	0,984	0,990	0,937	0,833	0,739	0,740	0,724	0,722
Acc ↑	0,966	0,981	0,545	0,544	0,738	0,741	0,723	0,722
1-Err↓	0,192	0,193	0,439	0,442	0,181	0,197	0,237	0,231
AvPr↑	0,974	0,988	0,572	0,570	0,828	0,833	0,782	0,788
RLos↓	0,014	0,016	0,363	0,364	0,007	0,007	0,086	0,093
	MEDICAL							
HLss↓	0,025	0,020	0,027	0,365	0,021	0,023	0,022	0,020
F-M↑	0,333	0,294	0,003	0,001	0,350	0,350	0,339	0,341
Acc ↑	0,338	0,241	0,002	0,001	0,343	0,343	0,334	0,334
1-Err↓	0,560	0,435	0,015	0,970	0,567	0,599	0,564	0,569
AvPr↑	0,512	0,612	0,950	0,114	0,509	0,512	0,519	0,560
RLos↓	0,155	0,112	0,544	0,545	0,123	0,121	0,110	0,113
	SCENE							
HLss↓	0,116	0,107	0,115	0,122	0,108	0,107	0,108	0,106
F-M↑	0,863	0,765	0,516	0,466	0,658	0,658	0,658	0,659
Acc ↑	0,625	0,653	0,468	0,421	0,652	0,650	0,652	0,678
1-Err↓	0,355	0,329	0,429	0,467	0,359	0,359	0,359	0,360
AvPr↑	0,762	0,670	0,699	0,671	0,740	0,739	0,742	0,747
RLos↓	0,170	0,189	0,258	0,287	0,117	0,219	0,215	0,222
	YEAST							
HLss↓	0,260	0,254	0,289	0,281	0,212	0,209	0,211	0,207
F-M↑	0,527	0,528	0,057	0,083	0,597	0,602	0,604	0,611
Acc ↑	0,450	0,459	0,068	0,084	0,498	0,507	0,504	0,515
1-Err↓	0,406	0,450	0,886	0,854	0,450	0,477	0,482	0,484
AvPr↑	0,639	0,645	0,445	0,441	0,560	0,563	0,561	0,566
RLos↓	0,304	0,283	0,508	0,509	0,224	0,381	0,348	0,331

Tabela 7.4 Desempenho dos métodos Multirrótulos, para cada versão do SSRAkELd.

Analisando o percentual de melhores casos alcançados pelos métodos que fizeram ou não uso do parâmetro de confiabilidade, pode-se destacar, portanto, que os métodos que fizeram uso do parâmetro de confiabilidade alcançaram um percentual de 70% (SSRAkELr e SSRAkELcr) dos melhores casos comparado com 43% dos melhores casos para os métodos que não fizeram uso do parâmetro de confiabilidade (SSRAkEL e SSRAkELc). Quanto aos resultados alcançados a partir dos percentuais de exemplos não-rotulados que deverão ser rotulados a cada iteração, que são de 17% e 33.3%, a maioria dos métodos analisados obtiveram os melhores resultados a partir do percentual de 33.3% dos exemplos não-rotulados que deverão ser rotulados a cada iteração com um percentual de 59%, ou seja, 71 dos 120 possíveis casos comparados com o percentual de 17% que foi 47%, ou seja, 57 dos 120 possíveis casos.

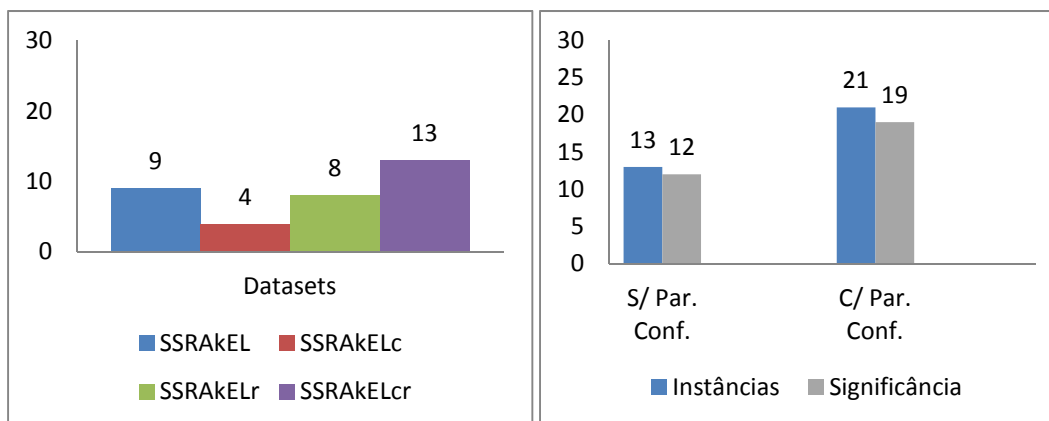


Figura 7.4 Análise dos métodos baseados no SSRaKELd

6.2) Percentual de 75% de dados rotulados iniciais

Assim como na Seção 7.1, esta seção apresentará os resultados para os métodos semissupervisionados derivados do BR, LP, RAkEL e RAkELd com o percentual de dados rotulados inicialmente de 75%.

6.2.1) Métodos baseados no SSBR

A Tabela 7.5 mostra o desempenho do SSBR, SSBRc (primeira abordagem) além do SSBRr e o SSBRcr (segunda abordagem) usando todos os cinco conjuntos de dados. Nesta tabela, os resultados são apresentados usando seis medidas de avaliação diferentes. O melhor resultado obtido em cada medida está em negrito e, para definir o melhor resultado entre os métodos em análise, a célula vencedora será destacada com a cor cinza. Além disso, os resultados estatisticamente mais significantes foram marcados com sublinhado.

Estes resultados são válidos para o percentual de 75% dos dados inicialmente rotulados.

	SSBR		SSBRc		SSBRr		SSBRcr	
	17%	33.3%	17%	33.3%	17%	33.3%	17%	33.3%
Measure	EMOTIONS							
HLss↓	0,308	0,308	0,308	0,308	<u>0.189</u>	0.243	0.196	0.195
F-M↑	0,015	0,009	0,012	0,009	0.622	0.562	0.627	0.623
Acc↑	0,018	0,009	0,012	0,009	0.564	0.441	0.541	0.539
1-Err↓	0,077	0,381	0,398	0,384	0.275	0.431	0.261	0.276
AvPr↑	0,736	0,726	0,711	0,727	0.801	0.675	0.801	0.800
RLoS↓	0,242	0,252	0,275	0,255	<u>0.159</u>	0.316	0.162	0.161
	GENBASE							
HLss↓	0,001	0,002	0,001	0,004	0.005	0.004	0.002	0.001

F-M↑	0,992	0,989	0,993	0,989	0,987	0,992	0,980	0,995
Acc ↑	0,984	0,981	0,987	0,981	0,986	0,992	0,978	0,995
1-Err↓	0,006	0,004	0,004	0,001	0,014	0,008	0,001	0,005
AvPr↑	0,990	0,991	0,992	0,991	0,980	0,977	0,980	0,999
RLoS↓	0,003	0,002	0,002	0,002	0,005	0,004	0,003	0,004
MEDICAL								
HLss↓	0,021	0,021	0,021	0,021	0,013	0,014	0,017	0,016
F-M↑	0,559	0,557	0,580	0,461	0,753	0,622	0,408	0,593
Acc ↑	0,460	0,459	0,491	0,466	0,685	0,580	0,407	0,481
1-Err↓	0,459	0,456	0,418	0,442	0,274	0,301	0,309	0,312
AvPr↑	0,574	0,572	0,606	0,585	0,771	0,777	0,767	0,780
RLoS↓	0,207	0,211	0,187	0,205	0,091	0,081	0,054	0,079
SCENE								
HLss↓	0,093	0,093	0,091	0,092	0,091	0,087	0,091	0,087
F-M↑	0,641	0,741	0,747	0,650	0,667	0,772	0,667	0,772
Acc ↑	0,629	0,625	0,652	0,637	0,662	0,659	0,668	0,677
1-Err↓	0,241	0,243	0,242	0,236	0,242	0,255	0,248	0,266
AvPr↑	0,854	0,853	0,853	0,858	0,853	0,855	0,851	0,859
RLoS↓	0,087	0,088	0,088	0,084	0,088	0,087	0,089	0,084
YEAST								
HLss↓	0,207	0,209	0,195	0,201	0,249	0,228	0,195	0,207
F-M↑	0,581	0,566	0,647	0,614	0,573	0,577	0,608	0,623
Acc ↑	0,430	0,411	0,516	0,473	0,432	0,451	0,524	0,525
1-Err↓	0,241	0,244	0,231	0,236	0,409	0,418	0,236	0,244
AvPr↑	0,759	0,753	0,764	0,762	0,602	0,645	0,761	0,750
RLoS↓	0,177	0,185	0,173	0,174	0,339	0,287	0,176	0,182

Tabela 7.5 Desempenho dos métodos Multirrótulos, para cada versão do SSBR.

Pode-se observar a partir dos resultados extraídos pela Tabela 7.5 que a utilização do parâmetro de confiabilidade no processo semissupervisionado teve um impacto bastante positivo no desempenho de seus métodos. Para o método SSBRr, este, por sua vez, alcançou 12 entre 30 casos possíveis e, consequentemente, foi um dos métodos que teve um desempenho satisfatório a partir do percentual de 40% dos melhores resultados alcançados. Em seguida, podemos destacar o melhor desempenho para o método SSBRcr com 47% dos melhores resultados, ou seja, alcançou 14 entre 30 casos possíveis. Os métodos sem parâmetros de confiabilidade alcançaram cerca de 36% dos melhores casos, para o SSBRc, e 6% dos melhores casos para o SSBR. Destaca-se, portanto, que dos seis conjuntos de dados, os métodos baseados em parâmetros de confiabilidade alcançaram a maioria dos melhores resultados nas bases *Emotions*, *Genbase*, *Medical* e *Scene*. Esses resultados já eram esperados, haja vista que os métodos SSBRr e SSBRcr fazem uso de uma técnica mais elaborada na forma de identificar os exemplos não-rotulados.

Analisando o percentual de melhores casos alcançados pelos métodos que fizeram ou não uso do parâmetro de confiabilidade, pode-se destacar que os métodos que fizeram uso do parâmetro de confiabilidade alcançaram um percentual de 87% (SSBRr e SSBRcr) dos melhores casos possíveis comparado com 43% dos melhores casos para os métodos que

não fizeram uso do parâmetro de confiabilidade (SSBR e SSBRc). Quanto aos resultados alcançados a partir dos percentuais de exemplos não-rotulados que deverão ser rotulados a cada iteração, que são de 17% e 33.3%, a maioria dos métodos analisados obtiveram os melhores resultados a partir do percentual de 17% dos exemplos não-rotulados que deverão ser rotulados a cada iteração com um percentual de 65%, ou seja, 78 dos 120 possíveis casos comparados com o percentual de 33.3% que foi 40%, ou seja, 49 dos 120 possíveis casos.

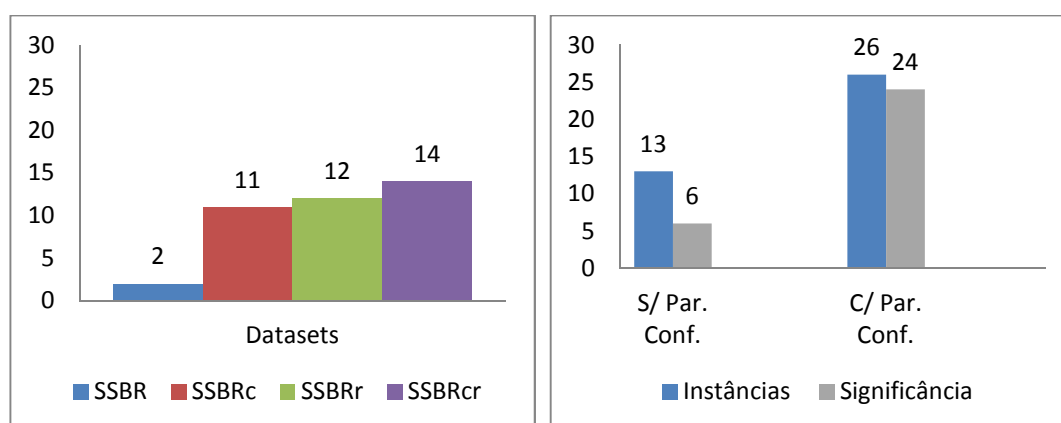


Figura 7.5 Análise dos métodos baseados no SSBR.

6.2.2) Métodos baseados no SSLP

A Tabela 7.6 ilustra o desempenho dos métodos SSLP, SSLPc, SSLPr e SSLPcr usando todos os cinco diferentes conjuntos de dados. Como mencionado anteriormente, todos estes métodos foram avaliados utilizando-se seis parâmetros de avaliação diferentes, além de um percentual de 75% de dados não rotulados. Para esta análise, os melhores resultados obtidos em cada métrica estão destacados em **negrito** e, para definir o melhor resultado entre os métodos em análise, a célula vencedora será destacada com a cor cinza. Além disso, os resultados estatisticamente mais significantes foram marcados com sublinhado. Estes resultados são válidos para o percentual de 75% dos dados inicialmente rotulados.

Ao verificar a Tabela 7.6, pode-se observar que, assim como no método proposto anterior, os métodos com parâmetro de confiabilidade alcançaram resultados superiores comparados aos resultados atingidos pelos métodos sem parâmetro de confiabilidade. Tal constatação pode ser observada pelo método que melhor desempenho teve, que foi o SSLPr com 50% dos melhores casos com a maioria dos casos em 15 dos 30 casos

possíveis. Em seguida, e de maneira equilibrada, vieram os métodos SSLPcr, SSLP e SSPLc com 30%, 20% e 13% dos melhores casos respectivamente..

Analisando o percentual de melhores casos alcançados pelos métodos que fizeram ou não uso do parâmetro de confiabilidade, pode-se destacar que os métodos que fizeram uso do parâmetro de confiabilidade alcançaram um percentual de 80% (SSLPr e SSLPcr) dos melhores casos comparado com 33% dos melhores casos para os métodos que não fizeram uso do parâmetro de confiabilidade (SSLP e SSPLc). Quanto aos resultados alcançados a partir dos percentuais de exemplos não-rotulados que deverão ser rotulados a cada iteração, que são de 17% e 33.3%, a maioria dos métodos analisados obtiveram os melhores resultados a partir do percentual de 33.3% dos exemplos não-rotulados que deverão ser rotulados a cada iteração com um percentual de 55%, ou seja, 66 dos 120 possíveis casos comparados com o percentual de 17% que foi 41%, ou seja, 50 dos 120 possíveis casos.

	SSLP		SSPLc		SSLPr		SSLPcr	
	17%	33.3%	17%	33.3%	17%	33.3%	17%	33.3%
Measure	EMOTIONS							
HLss↓	0,252	0,231	0,254	0,248	0,210	0,212	0,214	0,215
F-M↑	0,660	0,625	0,686	0,550	0,684	0,681	0,668	0,672
Acc ↑	0,484	0,513	0,475	0,491	0,568	0,572	0,551	0,545
1-Err↓	0,414	0,389	0,398	0,406	0,667	0,691	0,649	0,653
AvPr↑	0,693	0,713	0,695	0,701	0,711	0,729	0,718	0,722
RLoS↓	0,302	0,277	0,310	0,291	0,579	0,581	0,566	0,570
	GENBASE							
HLss↓	0,001	0,002	0,001	0,004	0,005	0,004	0,005	0,003
F-M↑	0,992	0,892	0,990	0,990	0,946	0,946	0,982	0,996
Acc ↑	0,983	0,976	0,983	0,981	0,946	0,945	0,948	0,995
1-Err↓	0,022	0,022	0,018	0,021	0,049	0,037	0,048	0,037
AvPr↑	0,977	0,975	0,980	0,977	0,939	0,985	0,940	0,971
RLoS↓	0,015	0,015	0,012	0,014	0,047	0,042	0,045	0,041
	MEDICAL							
HLss↓	0,021	0,025	0,024	0,027	0,019	0,018	0,020	0,018
F-M↑	0,004	0,008	0,004	0,003	0,610	0,617	0,599	0,608
Acc ↑	0,004	0,008	0,004	0,003	0,594	0,603	0,586	0,591
1-Err↓	0,951	0,949	0,952	0,953	0,622	0,667	0,650	0,611
AvPr↑	0,117	0,119	0,115	0,116	0,401	0,425	0,373	0,407
RLoS↓	0,543	0,541	0,544	0,545	0,428	0,433	0,451	0,442
	SCENE							
HLss↓	0,145	0,145	0,135	0,136	0,096	0,099	0,097	0,098
F-M↑	0,283	0,100	0,200	0,300	0,719	0,722	0,714	0,725
Acc ↑	0,242	0,247	0,324	0,325	0,717	0,709	0,712	0,716
1-Err↓	0,678	0,686	0,633	0,634	0,329	0,331	0,330	0,335
AvPr↑	0,522	0,514	0,550	0,552	0,754	0,755	0,753	0,758
RLoS↓	0,422	0,435	0,409	0,407	0,234	0,237	0,234	0,238
	YEAST							
HLss↓	0,245	0,246	0,246	0,246	0,220	0,223	0,220	0,224
F-M↑	0,517	0,498	0,501	0,499	0,623	0,625	0,626	0,620
Acc ↑	0,477	0,479	0,474	0,470	0,517	0,508	0,518	0,522
1-Err↓	0,320	0,316	0,320	0,327	0,832	0,835	0,834	0,842

AvPr↑	0,662	0,665	0,662	0,661	0,393	0,402	0,390	0,415
RLo↓	0,310	0,307	0,309	0,308	0,630	0,639	0,634	0,646

Tabela 7.6 Desempenho dos métodos Multirrótulos, para cada versão do SSLP.

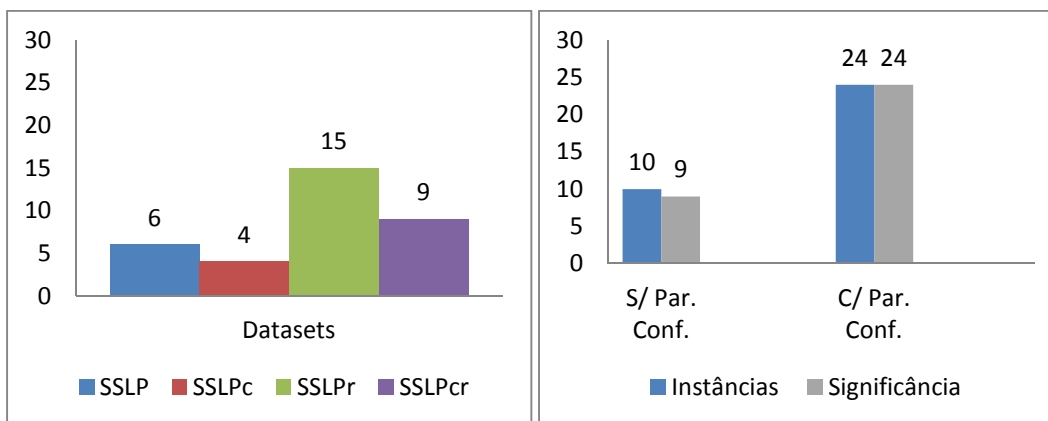


Figura 7.6 Análise dos métodos baseados no SSLP.

6.2.3) Métodos baseados no SSRaKEL

A Tabela 7.7 apresenta o desempenho do SSRaKEL, SSRaKELc, SSRaKELr e SSRaKELcr, também usando cinco diferentes conjuntos de dados. Mais uma vez, todos os resultados foram obtidos através do uso de seis métricas de avaliação diferentes, além do percentual de 75% de dados inicialmente rotulados. Para essa comparação, os melhores resultados obtidos em cada métrica estão em negrito e, para definir o melhor resultado entre os métodos em análise, a célula vencedora será destacada com a cor cinza. Além disso, os resultados estatisticamente mais significantes foram marcados com sublinhado. Estes resultados são válidos para o percentual de 75% dos dados inicialmente rotulados.

A partir do processo de análise dos resultados, o desempenho dos métodos derivados do SSRaKEL, na Tabela 7.7, atingiram os melhores resultados na maioria dos casos em, 15 dos 30 casos dos possíveis casos, pelo método SSRaKELcr, ou seja, 50% dos melhores casos, enquanto que o método SSRaKELr obteve, em 7 dos 30 possíveis casos, um percentual de 23% dos melhores casos. Esses métodos fizeram uso do parâmetro de confiabilidade baseado no *ranking* dos valores de confiança das instâncias. Após esses resultados, destaca-se, também, o bom desempenho do método SSRaKELc, pois este

alcançou um desempenho próximo ao método baseado em parâmetro de confiabilidade. O resultado alcançado pelo SSRaKElc foi em 6 dos 30 possíveis casos, o que significa um percentual de 20% dos melhores casos. Por último temos o método do SSRaKEl com 2 dos 30 melhores casos e um percentual de 6% dos melhores casos.

	SSRaKEl		SSRaKElc		SSRaKElr		SSRaKElcr	
	17%	33.3%	17%	33.3%	17%	33.3%	17%	33.3%
Measure	EMOTIONS							
HLss↓	0,267	0,267	0.241	0.256	0.206	0.208	0.207	0.209
F-M↑	0,578	0,578	0.631	0.670	0.685	0.663	0.659	0.641
Acc ↑	0.458	0,459	0.501	0.479	0.564	0.538	0.558	0.543
1-Err↓	0.419	0,418	0.408	0.418	0.315	0.329	0.421	0.398
AvPr↑	0,686	0,686	0.701	0.692	0.753	0.723	0.692	0.707
RLos↓	0,315	0,311	0.288	0.307	0.242	0.279	0.300	0.308
	GENBASE							
HLss↓	0,020	0,020	0.081	0.083	0.021	0.025	0.019	0.024
F-M↑	0,800	0,700	0.830	0.827	0.275	0.271	0.281	0.279
Acc ↑	0,530	0,531	0.990	0.984	0.266	0.265	0.269	0.267
1-Err↓	0,454	0,455	0.008	0.009	0.679	0.677	0.668	0.653
AvPr↑	0,560	0,561	0.981	0.984	0.985	0.972	0.970	0.966
RLos↓	0,369	0,368	0.377	0.378	0.356	0.357	0.352	0.355
	MEDICAL							
HLss↓	0,158	0,158	0.027	0.028	0.021	0.020	0.021	0.019
F-M↑	0,100	0,200	0.051	0.039	0.275	0.277	0.281	0.290
Acc ↑	0,210	0,211	0.052	0.039	0.265	0.266	0.269	0.271
1-Err↓	0,735	0,734	0.825	0.832	0.679	0.680	0.668	0.670
AvPr↑	0,310	0,312	0.365	0.341	0.364	0.359	0.370	0.362
RLos↓	0,376	0,378	0.185	0.207	0.358	0.367	0.352	0.363
	SCENE							
HLss↓	0,578	0,578	0.175	0.119	0.157	0.122	0.096	0.108
F-M↑	0,558	0,581	0.023	0.477	0.726	0.730	0.716	0.781
Acc ↑	0,606	0,683	0.022	0.587	0.561	0.604	0.707	0.723
1-Err↓	0,363	0,363	0.733	0.382	0.368	0.311	0.260	0.239
AvPr↑	0,738	0,717	0.527	0.728	0.758	0.799	0.811	0.834
RLos↓	0,232	0,231	0.325	0.227	0.181	0.183	0.161	0.179
	YEAST							
HLss↓	0,021	0,022	0.245	0.246	0.255	0.246	0.219	0.222
F-M↑	0,490	0,492	0.473	0.473	0.419	0.425	0.462	0.449
Acc ↑	0,349	0,348	0.479	0.464	0.345	0.347	0.410	0.408
1-Err↓	0,334	0,335	0.422	0.462	0.300	0.272	0.259	0.255
AvPr↑	0,580	0,582	0.647	0.658	0.598	0.602	0.618	0.729
RLos↓	0,418	0,419	0.370	0.265	0.392	0.390	0.387	0.377

Tabela 7.7 Desempenho dos métodos Multirrótulos, para toda versão do SSRaKEl.

Analisando o percentual de melhores casos alcançados pelos métodos que fizeram ou não uso do parâmetro de confiabilidade, pode-se destacar, portanto, que os métodos que fizeram uso do parâmetro de confiabilidade alcançaram um percentual de 73% (SSRaKElr e SSRaKElcr) dos melhores casos comparado com 26% dos melhores casos para os métodos que não fizeram uso do parâmetro de confiabilidade (SSRaKEl e SSRaKElc). Quanto aos resultados alcançados a partir dos percentuais de exemplos não-

rotulados que deverão ser rotulados a cada iteração, que são de 17% e 33.3%, a maioria dos métodos analisados obtiveram os melhores resultados a partir do percentual de 17% dos exemplos não-rotulados que deverão ser rotulados a cada iteração com um percentual de 59%, ou seja, 71 dos 120 possíveis casos comparados com o percentual de 33.3% que foi 48%, ou seja, 58 dos 120 possíveis casos.

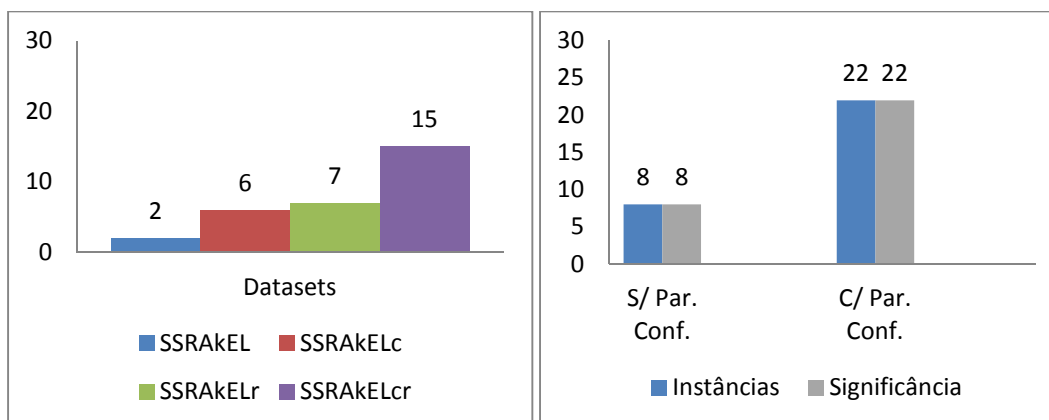


Figura 7.7 Análise dos métodos baseados no SSRAkEL.

6.2.4) Métodos baseados no SSRAkELd

A Tabela 7.8 apresenta o desempenho do SSRAkELd, SSRAkELdc, SSRAkELdr e SSRAkELdcr, também usando cinco diferentes conjuntos de dados. Mais uma vez, todos os resultados foram obtidos através do uso de seis métricas de avaliação diferentes, além do percentual de 75% de dados inicialmente rotulados. Para essa comparação, os melhores resultados obtidos em cada métrica estão em **negrito** e, para definir o melhor resultado entre os métodos em análise, a célula vencedora será destacada com a cor cinza. Além disso, os resultados estatisticamente mais significantes foram marcados com sublinhado. Estes resultados são válidos para o percentual de 75% dos dados inicialmente rotulados.

Ao analisar o desempenho dos métodos baseados SSRAkELd na Tabela 7.8, é possível observar que os melhores resultados na maioria dos casos, 13 dos 30 casos, foi obtido pelo método SSRAkELdr, ou seja, com um percentual de 43% dos melhores casos. Além disso, outro método baseado no parâmetro de confiabilidade, o SSRAkELdcr, também alcançou resultados satisfatórios, atingindo o percentual de 27% dos melhores casos, ou seja, 8 dos 30 casos possíveis. Já o método SSRAkELdc, este, por sua vez, alcançou o

percentual de 23% dos melhores casos com 7 dos 30 possíveis casos e, por último, o métodos que apresentou o pior desempenho foi o SSRaKEld com apenas 16% dos melhores casos, ou seja, 5 dos 30 casos possíveis.

	SSRAkELd		SSRAkELdc		SSRAkELdr		SSRAkELdcr	
	17%	33.3%	17%	33.3%	17%	33.3%	17%	33.3%
Measure	EMOTIONS							
HLss↓	0,259	0,259	0,241	0.245	0.207	0,206	0,203	0.205
F-M↑	0,613	0,602	0.634	0,783	0,659	0.655	0.644	0,646
Acc↑	0,470	0,476	0,502	0.494	0,555	0.552	0.544	0,547
1-Err↓	0,422	0,436	0.395	0,381	0,421	0.429	0,430	0.440
AvPr↑	0,651	0,686	0,713	0.702	0,692	0.685	0.677	0,680
RLos↓	0,281	0,300	0,264	0.274	0,200	0.312	0,319	0.333
	GENBASE							
HLss↓	0,004	0,002	0,002	0.003	0,004	0.007	0,004	0.008
F-M↑	0,955	0,982	0,991	0.893	0.950	0,952	0.945	0,951
Acc↑	0,943	0,967	0,975	0.017	0.950	0,961	0.953	0,956
1-Err↓	0,010	0,017	0,017	0.016	0,021	0.024	0,008	0.009
AvPr↑	0,974	0,977	0,985	0.975	0.970	0,977	0.972	0,983
RLos↓	0,012	0,017	0,009	0.014	0.010	0,009	0.014	0,013
	MEDICAL							
HLss↓	0,022	0,024	0,027	0.029	0,021	0.023	0,021	0.026
F-M↑	0,357	0,300	0.030	0,035	0,336	0.325	0,281	0.274
Acc↑	0,381	0,371	0.030	0,035	0,324	0.311	0,281	0.272
1-Err↓	0,533	0,528	0,790	0.802	0.557	0,595	0.777	0,498
AvPr↑	0,530	0,537	0.350	0,353	0,541	0.487	0,370	0.344
RLos↓	0,152	0,154	0.192	0,187	0,121	0.205	0,352	0.370
	SCENE							
HLss↓	0,109	0,122	0.136	0,128	0,092	0.117	0,922	0.945
F-M↑	0,554	0,581	0.383	0,650	0,711	0.682	0,696	0.688
Acc↑	0,554	0,583	0.552	0,587	0,707	0.660	0,694	0.674
1-Err↓	0,389	0,395	0,412	0.739	0.292	0,283	0.297	0,291
AvPr↑	0,714	0,722	0,714	0.400	0.794	0,799	0.797	0,798
RLos↓	0,242	0,254	0.240	0,181	0.170	0,164	0.125	0,120
	YEAST							
HLss↓	0,259	0,252	0.245	0,241	0.205	0,202	0.191	0,184
F-M↑	0,456	0,466	0,538	0.475	0,631	0.627	0,600	0.598
Acc↑	0,451	0,467	0.483	0,484	0.525	0,526	0.531	0,533
1-Err↓	0,419	0,433	0.408	0,394	0,337	0.352	0,369	0.377
AvPr↑	0,647	0,671	0.658	0,659	0,639	0.632	0.627	0,635
RLos↓	0,293	0,299	0,298	0.289	0,297	0.301	0,305	0.313

Tabela 7.8 Desempenho dos métodos Multirrótulos, para cada versão do SSRaKEld.

Seguindo o mesmo paradigma de análise, observa-se que o percentual de melhores casos alcançados pelos métodos que fizeram ou não uso do parâmetro de confiabilidade, pode-se destacar, portanto, que os métodos que fizeram uso do parâmetro de confiabilidade alcançaram um percentual de 70% (SSRAkELr e SSRAkELcr) dos melhores casos comparado com 40% dos melhores casos para os métodos que não fizeram uso do parâmetro de confiabilidade (SSRAkEL e SSRAkELc). Quanto aos resultados alcançados a partir dos percentuais de exemplos não-rotulados que deverão ser rotulados a cada iteração, que são de 17% e 33.3%, a maioria dos métodos analisados obtiveram os

melhores resultados a partir do percentual de 17% dos exemplos não-rotulados que deverão ser rotulados a cada iteração com um percentual de 54%, ou seja, 65 dos 120 possíveis casos comparados com o percentual de 33.3% que foi 46%, ou seja, 56 dos 120 possíveis casos.

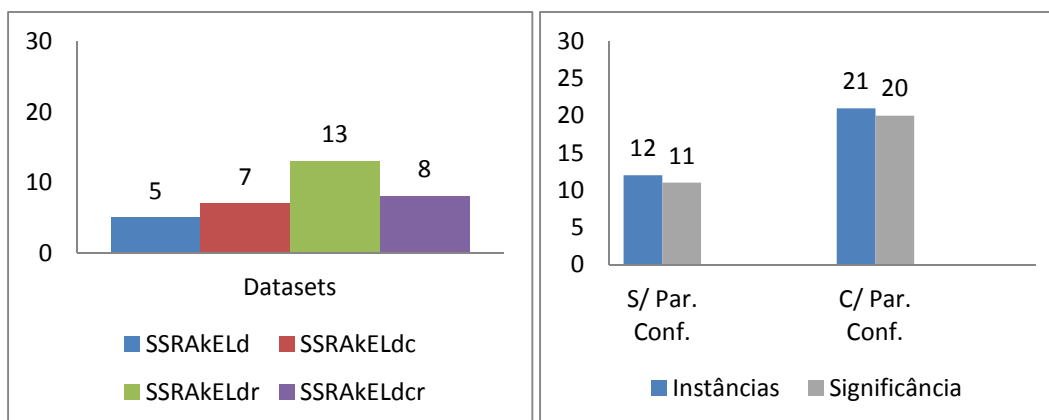


Figura 7.8 Análise dos métodos baseados no SSRAkELd

6.3) Análise de desempenho entre métodos Semissupervisionados

Com o objetivo de também analisar o desempenho dos métodos a partir de uma nova perspectiva – quanto ao percentual de dados inicialmente rotulados –, é apresentado neste tópico a análise gráfica dos resultados obtidos pelos métodos semissupervisionados derivados do SSBR, SSLP, SSRAkEL e SSRAkELd que fizeram, ou não, uso do parâmetro de confiabilidade.

Segundo os resultados apresentados nos gráficos abaixo, é possível observar que, de um modo geral, os métodos que fazem uso de parâmetros de confiabilidade concentram a maioria dos melhores resultados, tanto em um percentual de 25% de dados inicialmente rotulados, quanto o limiar de 75% de dados também inicialmente rotulados. Isso significa, portanto, que os métodos apresentados neste trabalho possuem um potencial bastante positivo quanto a classificação de dados multirrótulos em problemas de aprendizado semissupervisionado.

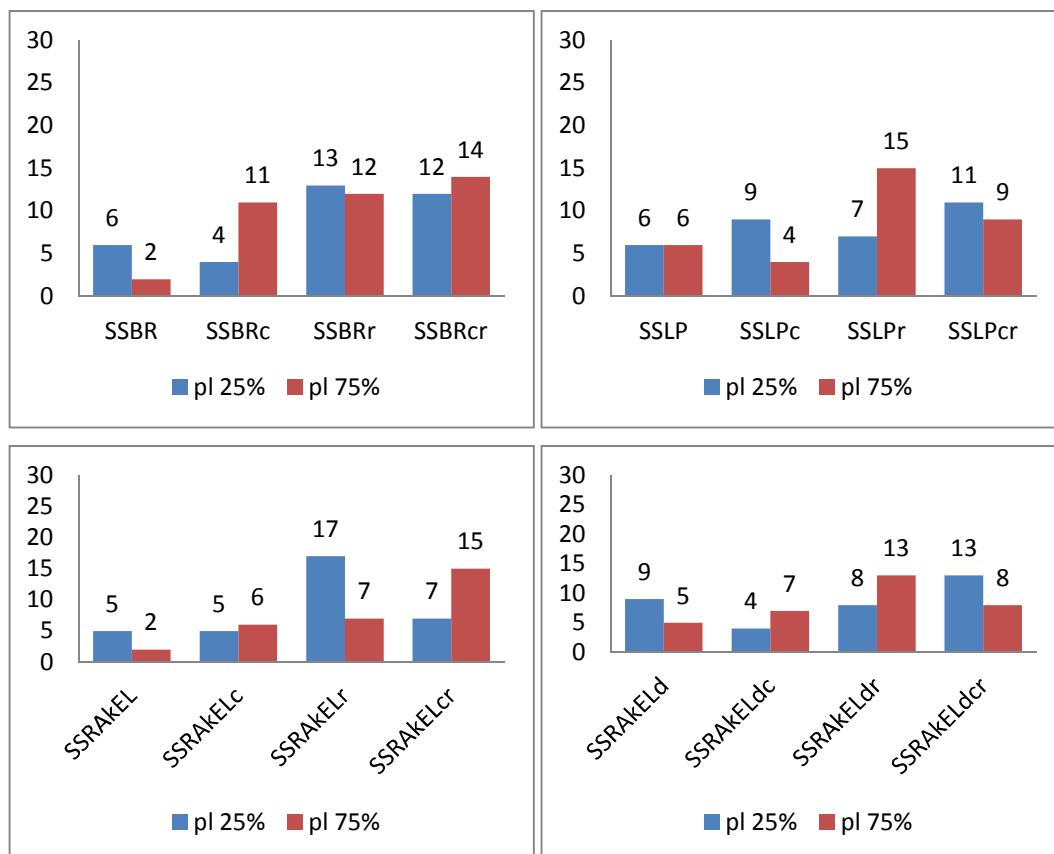


Figura 7.9 Análise de desempenho entre métodos Semissupervisionados

Os gráficos apresentados mostram que, a partir da comparação entre os percentuais de dados inicialmente rotulados, os métodos baseados nas versões com fator de confiança, SSBRc, SSBRcr, SSRAkELc e SSRAkELcr, alcançaram os melhores resultados quando o percentual de dados inicialmente rotulados era 75%. Já os métodos baseados nas versões sem fator de confiança, SSBR, SSBRr, SSRAkEL e SSRAkELr, alcançaram melhores resultados quando o percentual de dados inicialmente rotulados era 25%. Para as versões dos métodos SSLP e SSRAkELd, estes, por sua vez, apresentaram resultados bem equilibrados quanto aos percentuais de 25 e 75%. Visto tais diferenças, percebe-se, portanto, que o fator de confiança c tem melhor desempenho quando, dentre um conjunto de dados multirrótulo semissupervisionado, o percentual de dados inicialmente rotulados é maior.

6.4) Resultados experimentais: Redução de dimensionalidade

Nesta seção será apresentado uma análise comparativa entre métodos semissupervisionados que fazem uso do parâmetro de confiabilidade proposto por este trabalho no qual serão aplicados às bases de dados multirrótulos. Para iniciar essa análise, bases de dados farão parte do conjunto de *dataset* multirrótulos já utilizadas nesses experimentos, aumentando, portanto, o número de resultados obtidos a partir dos novos mecanismos de classificação. Essas bases de dados, originalmente, apresentam uma grande dimensionalidade de atributos o que, conseqüentemente, podem vir a ser um potencial problema quanto a aquisição de resultados satisfatórios. Como mencionado anteriormente, foi pensado em uma técnica conhecida como seleção de atributos ou redução de dimensionalidade, com o objetivo de reduzir a dimensionalidade dos dados das novas bases multirrótulos escolhidas. O objetivo principal deste capítulo é possibilitar a análise do impacto desses métodos em bases de multirrótulos a partir dos métodos semissupervisionados com parâmetros de confiabilidade.

Na segunda parte dessa análise empírica, os métodos propostos serão comparados com métodos multirrótulos já existentes, propostos em (SPOLAOR; RODRIGUES et al, 2013). Essa análise comparativa é realizada usando apenas dois métodos semissupervisionados multirrótulo, o SSBR e o SSLP e cinco bases de dados. Por questões de limitação nos resultados presentes no trabalho de (SPOLAOR, 2013), somente será possível realizar uma análise comparativa entre os métodos semissupervisionados BR e LP.

Nas Tabelas 7.9 e 7.10 foram apresentados os resultados para os métodos BR e LP, respectivamente. Nessas tabelas, para cada conjunto de dados, a primeira e a segunda linha representam os métodos propostos por este trabalho. O objetivo principal dessa comparação é investigar se a abordagem desenvolvida por este trabalho alcança resultados mais significantes comparado com a abordagem proposta por (SPOLAOR, 2013). Além disso, a terceira linha presente nas tabelas representa os resultados referente ao método de seleção de característica proposto em (SPOLAOR, 2013). Para uma melhor compreensão na análise, os melhores resultados obtidos em cada métrica estarão destacados na cor cinza, o que representa a célula vencedora.

Como pode ser observado, na Tabela 7.9, o método SSBRr apresentou os melhores resultados em 8 casos dos 20 casos analisados. Os melhores resultados para o SSBRcr e RF-BR foram 6 e 7 dos 20 casos analisados, respectivamente. Em outras palavras, para os métodos analisados, podemos afirmar que um dos métodos da metodologia proposta apresentou um desempenho um pouco melhor comparado com os outros dois métodos. Portanto, é possível obter resultados satisfatórios na abordagem proposta para o BR.

EMOTIONS				
	HLss ↓	SAcc ↑	Acc ↑	F-M ↑
SSBRr	0.189	0.273	0.564	0.622
SSBRcr	0.196	0.266	0.541	0.627
RF-BR	0.220	0.260	0.530	0.620
GENBASE				
SSBRr	0.005	0.938	0.981	0.989
SSBRcr	0.002	0.927	0.978	0.980
RF-BR	0.001	0.940	0.970	0.980
MEDICAL				
SSBRr	0.013	0.612	0.685	0.753
SSBRcr	0.017	0.633	0.407	0.408
RF-BR	0.010	0.640	0.710	0.730
SCENE				
SSBRr	0.090	0.669	0.662	0.667
SSBRcr	0.091	0.688	0.688	0.667
RF-BR	0.120	0.640	0.660	0.670
YEAST				
SSBRr	0.249	0.144	0.432	0.573
SSBRcr	0.195	0.147	0.524	0.608
RF-BR	0.240	0.120	0.480	0.590

Tabela 7.9 Performance dos métodos multirrótulo para todas as versões do BR.

Nesta parte do trabalho será analisado a performance dos métodos baseados no LP da Tabela 7.10. A partir dessa tabela é possível observar que o método proposto, SSLPr, teve uma performance significativamente superior comparado com os outros dois métodos analisados, SSLPcr e RF-LP. O SSLPr alcançou os melhores resultados em 9 dos 20 possíveis casos, enquanto o SSLpcr obteve os melhores resultados em 7 casos dos

20 casos possíveis. Por fim, o RF-LP atingiu os melhores resultados em 4 dos 20 possíveis casos.

Com base na Tabela 7.10, é possível analisar que a abordagem proposta teve um impacto significativo na performance dos métodos baseados no LP. Além disso, o uso da redução de dimensionalidade teve um impacto positivo na performance dos métodos semissupervisionados, uma vez que, para esse tipo de comparação, o método do RF-LP não teve resultados superiores comparados com os métodos propostos pela abordagem inicial.

EMOTIONS				
	HLss ↓	SAcc ↑	Acc ↑	F-M ↑
SSLPr	0.210	0.308	0.568	0.684
SSLPcr	0.214	0.299	0.551	0.668
RF-LP	0.220	0.280	0.540	0.630
GENBASE				
SSLPr	0.005	0.966	0.946	0.946
SSLPcr	0.005	0.967	0.996	0.995
RF-LP	0.001	0.940	0.970	0.980
MEDICAL				
SSLPr	0.019	0.544	0.594	0.610
SSLPcr	0.020	0.552	0.586	0.599
RF-LP	0.020	0.560	0.670	0.700
SCENE				
SSLPr	0.096	0.581	0.717	0.719
SSLPcr	0.097	0.583	0.712	0.714
RF-LP	0.110	0.640	0.670	0.680
YEAST				
SSLPr	0.220	0.148	0.517	0.623
SSLPcr	0.222	0.134	0.518	0.626
RF-LP	0.222	0.150	0.510	0.620

Tabela 7.10 Performance dos métodos multirrótulos para todas as versões do LP.

6.5) Análise gráfica

A fim de dar suporte a esta análise comparativa, gráficos do tipo radar, baseados na performance das métricas de avaliação proposta por este trabalho, representam os resultados alcançados pelos métodos sob a perspectiva de métricas de avaliação, onde o maior valor indica a melhor performance para as seguintes métricas (F-Measure, Accuracy, Subset Accuracy) e o menor valor indica a melhor performance para o Hamming Loss. Portanto, os melhores resultados apresentados pelos gráficos a seguir estão representados pelos valores que mantiveram-se distantes do centro quando as métricas eram F-Measure, Accuracy, Subset Accuracy e Hamming Loss quando os valores alcançados estavam próximo ao centro do gráfico. Na Figura 7.10 ilustrada pelo gráfico Radar para os métodos baseados no BR, foi possível observar que, em algumas bases, os resultados foram mais significantes para as abordagens propostas quando comparados com o método RF-BR. Da mesma maneira, na Figura 7.11 ilustrada, foi apresentado os resultados obtidos pelos métodos baseados no LP e que, para algumas bases, as abordagens propostas inicialmente apresentaram alguns resultados mais significantes quando comparado com o RF-LP.

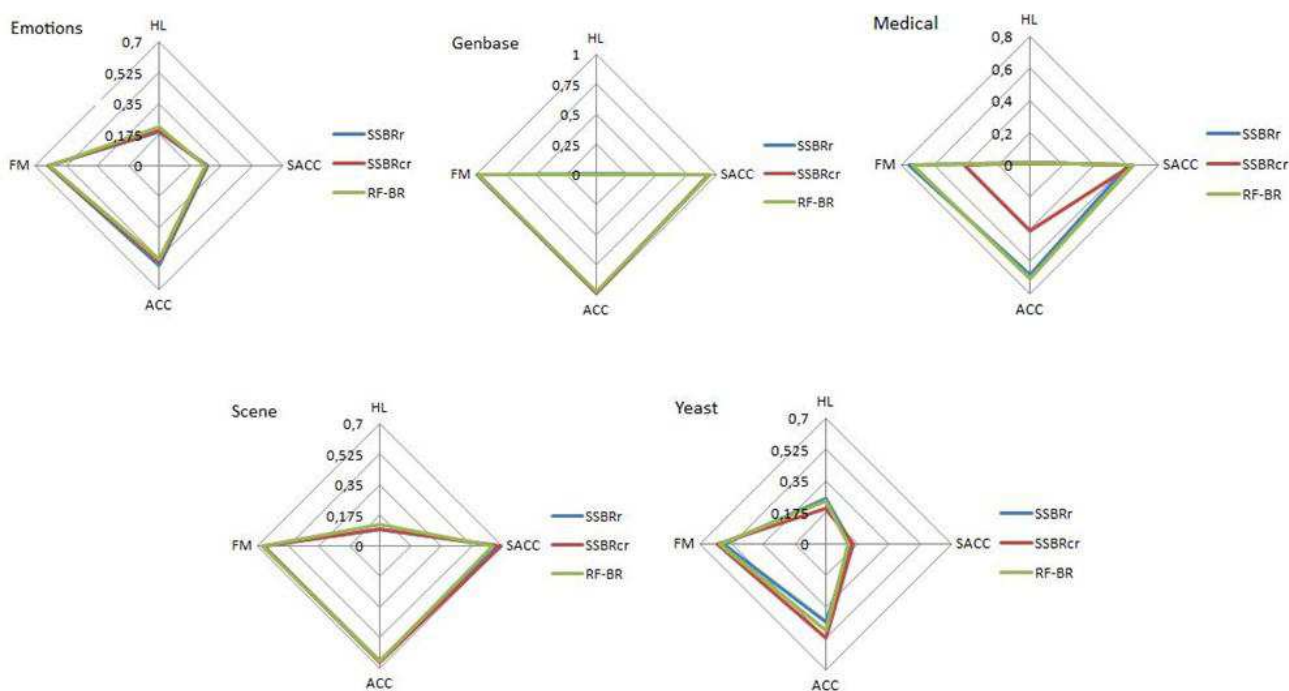


Figura 7.10 Gráfico Radar para os métodos multirrótulo baseados no BR.

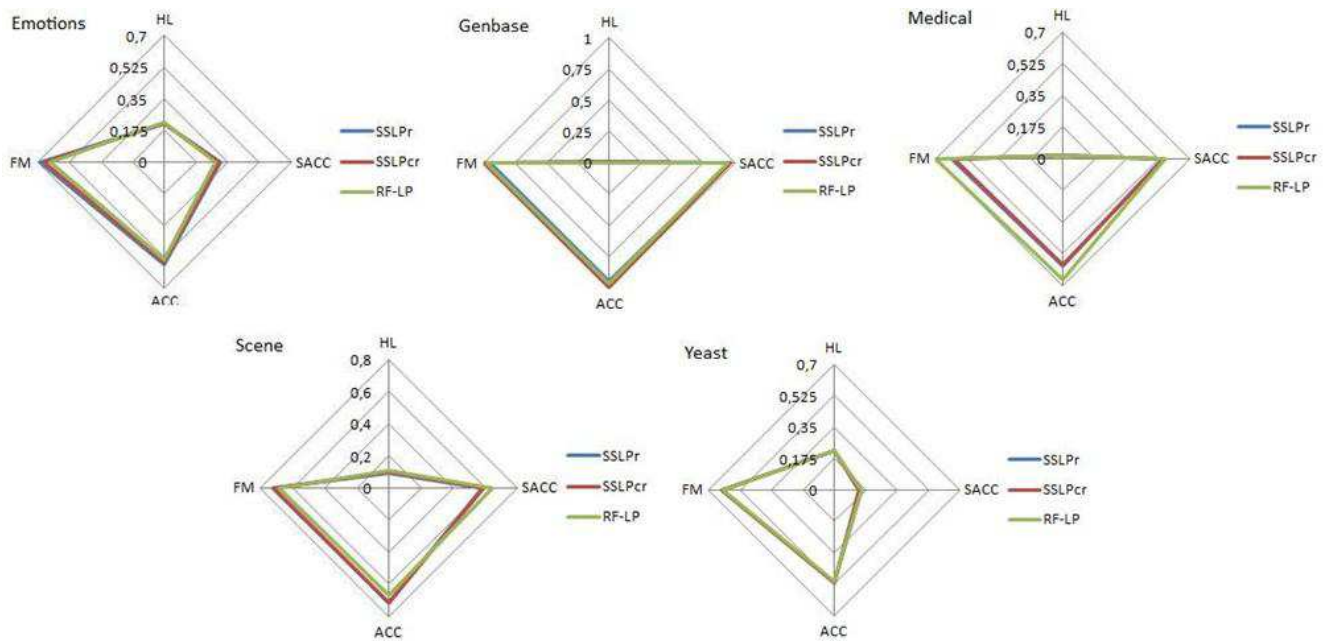


Figura 7.11 Gráfico Radar para os métodos multirrótulos baseados no LP.

6.6) Considerações Finais

Neste capítulo foram apresentados os resultados experimentais obtidos pelos métodos de classificação multirrótulo utilizando a abordagem do aprendizado semissupervisionado com confiabilidade; outra abordagem resultou em uma análise comparativa entre os métodos propostos na primeira abordagem deste trabalho e os métodos propostos pelo trabalho de (SPOLAOR, 2013); ambos fizeram uso de bases de dados multirrótulo, objetivando um impacto positivo na obtenção dos resultados finais. Inicialmente, os métodos da primeira abordagem deste trabalho foram organizados em dois grupos com diferentes percentuais de dados a serem rotulados inicialmente. Parte dos resultados foram analisados com um percentual de 25% dos dados rotulados e outra análise foi feita a partir do percentual de 75% dos dados rotulados inicialmente. Já na segunda abordagem, foram analisados sob um percentual de 75% os resultados alcançados a partir dos métodos semissupervisionados multirrótulos com confiabilidade e métodos baseados no *ReliefF*. Neste contexto, pode-se afirmar que, de modo geral, a utilização da abordagem proposta por esse trabalho apresentou resultados satisfatórios.

Capítulo 7

Considerações Finais

Neste trabalho foi realizado um estudo sobre a aplicação de técnicas de aprendizado semissupervisionado combinado com tarefas de classificação multirrótulo, de modo a avaliar os resultados obtidos e identificar suas vantagens e desvantagens.

Esse estudo foi motivado pela busca de resultados eficientes a partir da escolha de exemplos em diferentes regiões do conjunto de dados, baseando-se, portanto, pelo valor do desvio padrão das confidências que cada exemplo apresenta. Em estudos anteriores, foi constatado que uma das possíveis causas para os resultados apresentados terem contribuído de maneira negativa era devido a forma aleatória como esses exemplos haviam sido escolhidos. Para tal motivação, os dados não-rotulados do conjunto de exemplos serão rotulados a partir de um *ranking* do desvio padrão das confidências desses exemplos.

Além disso, é notável um crescente aumento no interesse pela área de aprendizado de máquina pela comunidade científica, principalmente pelos estudos sobre tarefas de classificação multirrótulo e aprendizado semissupervisionado; seja pela diversidade de problemas encontrados quanto a classificação dos dados a uma ou mais classes, classificação multirrótulo, ou quanto a maneira como são tratados os dados rotulados e não-rotulados no aprendizado semissupervisionado.

Foram investigados as combinações entre cada algoritmo de classificação multirrótulo com a técnica de aprendizado semissupervisionados baseadas no *ranking* do desvio padrão das confidências ou melhor, a partir de parâmetros de confiabilidade. Tais algoritmos são: SSBRr, SSLPr, SSRAkELr, SSRAkELdr, SSBRcr, SSLPcr, SSRAkELcr e SSRAkELdcr.

Nos resultados experimentais, é possível observar que o uso do parâmetro de confiabilidade no aprendizado semissupervisionado teve um efeito positivo em mais de 73% dos casos analisados, 88 em 120 casos, a partir dos métodos *r* e *cr* comparados com

os métodos propostos por (SANTOS, 2012) no qual não fazem uso do parâmetro de confiabilidade. Dentre as duas abordagens, r e cr , ou seja, métodos com parâmetro de confiabilidade, a utilização da segunda abordagem, cr , apresentou um desempenho superior comparado com a primeira abordagem. Os resultados obtidos a partir desta análise são promissores, uma vez que o desempenho dos métodos multirrótulos semissupervisionados podem ser melhorados a partir do cálculo do parâmetro de confiabilidade das instâncias. Além disso, pode-se concluir que a redução na dimensionalidade das bases de dados apresentadas neste trabalho podem proporcionar aos métodos semissupervisionados com confiabilidade resultados satisfatórios quando comparado com métodos que usam apenas a técnica baseada no *ReliefF*.

7.1) Trabalhos Publicados

Nesta seção são listados os trabalhos publicados e submetidos para publicação como resultado do estudos propostos:

- IJCNN 2013 – Using Confidence Values in Multi-label Classification Problems with Semi-Supervised Learning. Rodrigues, F. (IFRN); Santos, A. (UFERSA); Canuto A. (UFRN); Dallas – TX – USA.
- IJCNN 2014 – Confidence Factor and Feature Selection for Semi-supervised Multi-label Classification Methods. Rodrigues, F. (IFRN); Santos, A. (UFERSA); Canuto A. (UFRN); Beijing – CHINA.

7.2) Trabalhos Futuros

A análise realizada neste trabalho ainda pode ser expandida. Dessa maneira, alguns trabalhos futuros possíveis são:

- Definir e implementar novos métodos que serão empregados no aprendizado semissupervisionado multirrótulo a partir do *Co-Training*;
- Propor novos mecanismos para classificação de dados hierarquicamente multirrótulo a partir da confiabilidade;

- Expandir os estudos referente aos métodos redução de dimensionalidade e aplicá-los em problemas de classificação multirrótulo semissupervisionado;
- Utilização de novas bases de dados no processo de classificação multirrótulo semissupervisionado.

Referências Bibliográficas

- ALMEIDA, L. M.; LUDERMIR, T. B. A multi-objective memetic and hybrid methodology for optimizing the parameters and performance of artificial neural networks. *Neurocomputing*, v. 73, n. 7-9, p. 1438-1450, 2010.
- BARUTCUOGLU, Z.; SCHAPIRE, R. E.; TROYANSKAYA, O. G. Hierarchical multi-label prediction of gene function. *Bioinformatics*, v. 22, n. 7, p. 830-836, 2006.
- BASU, S.; BANERJEE, A.; MOONEY, R. Semi-supervised clustering by seeding. *Proc. of the 19th International Conference on Machine Learning. (ICML-02)*, p. 19-26, 2002.
- BLOCKEEL, H.; SCHIETGAT, L.; STRUYF, J.; DZEROSKI, S.; CLARE, A. Decision trees for hierarchical multilabel classification: A case study in functional genomics. *Machine Learning*, p. 18-29, 2006.
- BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with co-training. In: New York, NY, USA: Morgan Kaufmann Publishers, p. 92-100, 1998.
- BOUTELL, M. R.; LUO, J.; SHEN, X.; BROWN, C. M. Learning multi-label scene classification. *Pattern Recognition*, v. 37, n. 9, p. 1757-1771, 2004.
- BRAGA, I. A. Aprendizado semissupervisionado multidescrição em classificação de textos. Dissertação (Mestrado). Instituto de Ciências Matemáticas e de Computação – ICMC. Universidade de São Paulo – São Carlos, 2010.
- CANUTO, A.; SANTOS, A.; VARGAS, R. Ensembles of artmap-based neural networks: an experimental study. *Applied Intelligence*, Springer Netherlands, 2009.
- CERRI, R.; CARVALHO, A. C. P. L. F. de. Comparing local and global hierarchical multilabel classification methods using decision trees. In: *Proceedings of the Brazilian Symposium on Bioinformatics, Lecture Notes on Bioinformatics*. Porto Alegre, Brazil: Springer-Verlag, p. 109-120, 2009.
- CERRI, R. *Técnicas de classificação hierárquica multirrótulo*. Dissertação (Mestrado). Universidade de São Paulo - São Carlos, 2010.
- CHAN, A.; FREITAS, A. A. A New Colony Algorithm for Multi-Label Classification with Applications in Bioinformatics. In: *Proceedings of the Genetic and Evolutionary Computation Conference*, p. 27-34, 2006.
- CHEESMAN, P. & STUTZ, J. Bayesian classification (*autoclass*): theory and results advances in knowledge discovery and data mining, *American Association for Artificial Intelligence*, Menlo Park, CA, USA, p. 153-180, 1996
- CLARE, A.; KING, R. D. *Knowledge discovery in multi-label phenotype data*. In: *RAEDT, L. D.; SIEBES, A. (Ed.). PKDD*. Freiburg, Germany: Springer, 2001. (Lecture Notes in Computer Science, v. 2168), p. 42-53, 2001.

COSTA, E. Investigação de técnicas de classificação hierárquica para problemas de bioinformática. Dissertação (Mestrado). Universidade de São Paulo - São Carlos, 2008.

COSTA, J. A. Classificação automática e análise de dados por redes neurais auto-organizáveis. Dissertação (Doutorado). Universidade de São Paulo - São Carlos, 1999.

COVER, T.; HART, P. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, v. 13, p. 21- 27, 1967.

CRISTIANINI, N.; SHAW-TAYLOR, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. New York, NY, USA: Cambridge University Press, 2000.

DIETTERICH, T. G. Machine Learning. *Annual Review of Computer Science*, cap. 4. p. 255-306, 1996.

DIPLARIS, S.; TSOUMAKAS, G.; MITKAS, P.; VLAHAVAS, I. Protein Classification with Multiple Algorithms. In: *Proceedings of the 10th Panhellenic Conference on Informatics (PCI 2005)*. Volos, Greece: Springer-Verlag, p. 448-456, 2005.

DERNSAR, J. Algorithms for subsetining attribute values with Relief, *Machine Learning*, 78, pp. 421-428, 2010.

ELISSEEFF, A.; WESTON, J. A kernel method for multi-labelled classification. In: *NIPS - Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic]*. Vancouver, British Columbia, Canada: MIT Press, p. 681-687, 2001.

FACELLI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. *Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina*. Rio de Janeiro: LTC, 2011.

HAN, J.; M. KAMBER. *Data mining: concepts and techniques*. 2nd edition, 2006.

HAYKIN, S. Redes Neurais. *Princípios e Prática*. 2^aed, 1999, p. 88. Disponível em: <http://books.google.com.br/books?id=lBp0X5fyjUC&printsec=frontcover&hl=pt%20BR&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false> Acessado em: 8 out. 2012, 14:30:00.

JAIN, A. K.; DUBES, R. C. *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

KOHONEN, T. *Self-organized formation of topologically correct feature maps*. *Biological Cybernetics*, Springer, v. 43, n. 1, p. 59-69, 1982.

LI, T.; OGIHARA, M. Detecting emotion in music. In: *Proceedings of the International Symposium on Music Information Retrieval*. Washington D.C., USA: ISMIR'03, p. 239-240, 2003.

LI, T.; OGIHARA, M. Toward intelligent music information retrieval. *IEEE Transactions on Multimedia*, v. 8, n. 3, p. 564-574, 2006.

LIU, H.; MOTODA, H. *Computational methods of feature selection*, Chapman & Hall/CRC, 2008.

LORENA, A. C.; CARVALHO, A. C. P. L. F. de. *Building binary-tree-based multiclassifiers using separability measures*. Neurocomputing, 2010.

LUO, X.; ZINCIR-HEYWOOD, A. N. Evaluation of two systems on multi-class Multi-label document classification. In: *Proceedings of the 15th International Symposium on Methodologies for Intelligent Systems*. Berlin: Springer, 2005.

MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: CAM, L. M. L.; NEYMAN, J. (Ed.). *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. California: University of California Press, v. 1, p. 281-297, 1967.

MAHALANOBIS, P. C. *On the generalised distance in statistics*. In: *Proceedings National Institute of Science, India*. India: National Institute of Science, v. 2, n. 1, p. 49-55, 1936.

MATSUBARA, E. T.; MONARD, M. C.; BATISTA, G. E. A. P. A. Multi-view semi-supervised learning: An approach to obtain different views from text datasets. In: *LAPTEC*. Himeji, Japan: IOS Press, p. 97-104, 2005.

MCCALLUM, A. K. Multi-label text classification with a mixture model trained by em. In: *99 Workshops on Text Learning*. Pittsburg: AAAI, 1999.

METZ, J.; FREITAS, A. Extending hierarchical classification with semi-supervised learning. In: *UK Workshop on Computational Intelligence*. United Kindom: UK, p. 1-6, 2009.

METZ, J. *Abordagens para aprendizado semissupervisionado multirrótulo e hierárquico*. Dissertação (Doutorado) – Instituto de Ciências Matemáticas e de Computação – ICMC, Universidade de São Paulo (USP), São Carlos, 2011.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997.

MONARD, M. C. In: *Sistemas inteligentes: fundamentos e aplicações*. Barueri /SP: Manole. cap. Aprendizado de Máquina, 2003.

NASIERDING, G.; TSOUMAKAS, G.; KOUZANI, A. Z. Clustering based multi-label classification for image annotation and retrieval. In: *Proc. 2009 IEEE International Conference on Systems, Man, and Cybernetics SMC'09*. Texas, USA: IEEE, p. 4514-4519, 2009.

OLIVEIRA, D. F. de; CANUTO, A. M. P.; SOUTO, M. C. P. de. Investigating the use of an evolutionary agent-based system for classification tasks. In: IJCNN. Orlando, FL: IEEE, p. 1266-1271, 2007.

PARENTE, R. R. *Abordagem de Construção de Arquitetura Homogênea para Comitês via Meta-aprendizagem*. Dissertação (Mestrado) – Programa de Pós-Graduação em Sistemas e Computação, Universidade Federal do Rio Grande do Norte (UFRN), Natal, 2012.

PEARSON, K. *On lines and planes of closest fit to systems of points in space*. Philosophical Magazine, v. 2, n. 6, p. 559-572, 1901.

QUINLAN, J. R. C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning). 1. ed. San Francisco, California: Morgan Kaufmann, paperback, 1993.

REZENDE, S. O. *Sistemas Inteligentes: Fundamentos e aplicações*, Baurueri - SP: [s.n.], São Paulo, 2003.

RODRIGUES, F. M.; SANTOS, A. M.; CANUTO, A. M. P. *Using Confidence Values in Multi-label Classification Problems with semi-supervised Learning*. In: IJCNN. Dallas, TX: IEEE-IJCNN, 2013.

ROSENBERG, C.; HEBERT, M.; SCHNEIDERMAN, H. Semi-supervised self-training of object detection models. In: *Seventh IEEE Workshop on Applications of Computer Vision*. Breckenridge, CO, USA: IEEE, p. 29-36, 2005.

SANGER, T. D. *Advances in neural information processing systems 1*. In: San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, cap. An optimality principle for unsupervised learning, p. 11-19, 1989.

SANTANA, L. E. O.; CANUTO, A. M. P.; ABREU, M. C. C. *Analyzing the performance of an agent-based neural system for classification tasks using data distribution among the agents*. In: IJCNN. Vancouver, BC: IEEE. p. 2951-2958, 2006.

SANTOS, A. M. *Investigando a Combinação de Técnicas de Aprendizado Semissupervisionado e classificação Hierárquica Multirrótulo*. Dissertação (Doutorado) – Programa de Pós-Graduação em Sistemas e Computação, Universidade Federal do Rio Grande do Norte (UFRN), Natal, 2012.

SIKONJA, M.; KONONENKO, I. *Theoretical and empirical analysis of ReliefF and RReliefF*, Machine Learning 53, pp. 23-69, 2003.

SPOLAÔR, N.; CHERMAN, E.; MONARD, M. *Uso do ReliefF para Seleção de Atributos em Dados Multirrótulo*, 2011.

SPOLAÔR, N.; CHERMAN, E.; MONARD, M.; LEE, H. A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach. *Sciverse ScienceDirect* 292. pp 135-151, 2013.

TAVARES, L. G., H. S. LOPES & C. R. E LIMA, *Estudo comparativo de métodos de aprendizado de máquina na detecção de regiões promotoras de genes de escherichia coli*, Anais do I Simpósio Brasileiro de Inteligência Computacional, p. 8-11, 2007.

TROHIDIS, K.; TSOUMAKAS, G.; KALLIRIS, G.; VLAHAVAS, I. Multilabel classification of music into emotions. In: *Proc. 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, PA, USA, 2008. Germany: Springer, 2008.

TSOUMAKAS, G.; KATAKIS, I. Multi label classification: An overview. *International Journal of Data Warehouse and Mining, Idea Group Publishing*, v. 3, n. 3, p. 1-13, 2007.

TSOUMAKAS, G.; KATAKIS, I.; VLAHAVAS, I. Effective and efficient multilabel classification in domains with large number of labels. In: *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD'08)*. Antwerp, Belgium: Springer, 2008.

TSOUMAKAS, G.; KATAKIS, I.; VLAHAVAS, I. Mining Multi-label Data. Unpublished book chapter, 2009.

TSOUMAKAS, G.; KATAKIS, I.; VLAHAVAS, I. Random k-labelsets for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering, IEEE Computer Society*, Los Alamitos, CA, USA, v. 99, n. PrePrints, 2010.

UEDA, N.; SAITO, K. Parametric mixture models for multi-labeled text. In: *In Advances in Neural Information Processing Systems 15*. Vancouver, Canada: MIT Press,. p. 721-728, 2003.

VALLIM, R. M. M. *Sistemas Classificadores Evolutivos para Problemas Multirrótulo*. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação (ICMC) – Universidade de São Paulo (USP), São Carlos, 2009.

VAPNIK, V. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

VAPNIK, V. N. *Statistical Learning Theory*. USA: Wiley, 1998.

WAGSTAFF, K.; CARDIE, C.; ROGERS, S.; SCHROEDL, S. Constrained k-means clustering with background knowledge. In: *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. p. 577-584, 2001.

WIECZORKOWSKA, A.; SYNAK, P.; RA $\pm$, Z. Multi-label classification of emotions in music. In: Berlin, Heidelberg: Springer Berlin / Heidelberg. p. 307-315, 2006.

WITTEN, I. H. & FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques*, second edition, Elsevier: San Francisco (Morgan Kaufmann) WEKA, 2005.

ZHANG, M.-L.; ZHOU, Z.-H. A k-nearest neighbor based algorithm for multi-label classification. In: *IEEE International Conference on Granular Computing - GrC 2005*. Beijing, China: IEEE. p. 718-721, 2005.

ZHANG, M.; ZHOU, Z. *ml-knn: A lazy learning approach to multi-label learning*. Pattern Recognition, Elsevier Science Inc., New York, NY, USA, v. 40, n. 7, p. 2038-2048, 2007.

ZHOU, Z.-H. & Li, M. Semi-supervised learning by disagreement. *Knowledge and information systems*, 24:415-439, 2009.

ZHU, X., GHAHRAMANT, Z., & LAFFERTY, J. D. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML 2003: Proceedings of The Twentieth International Conference on Machine Learning*. p. 912-919. AAAI Press, 2003.