III Workshop de Dissertações de Mestrado do Ppgsi (2016)

FICHA DA PESQUISA[i]

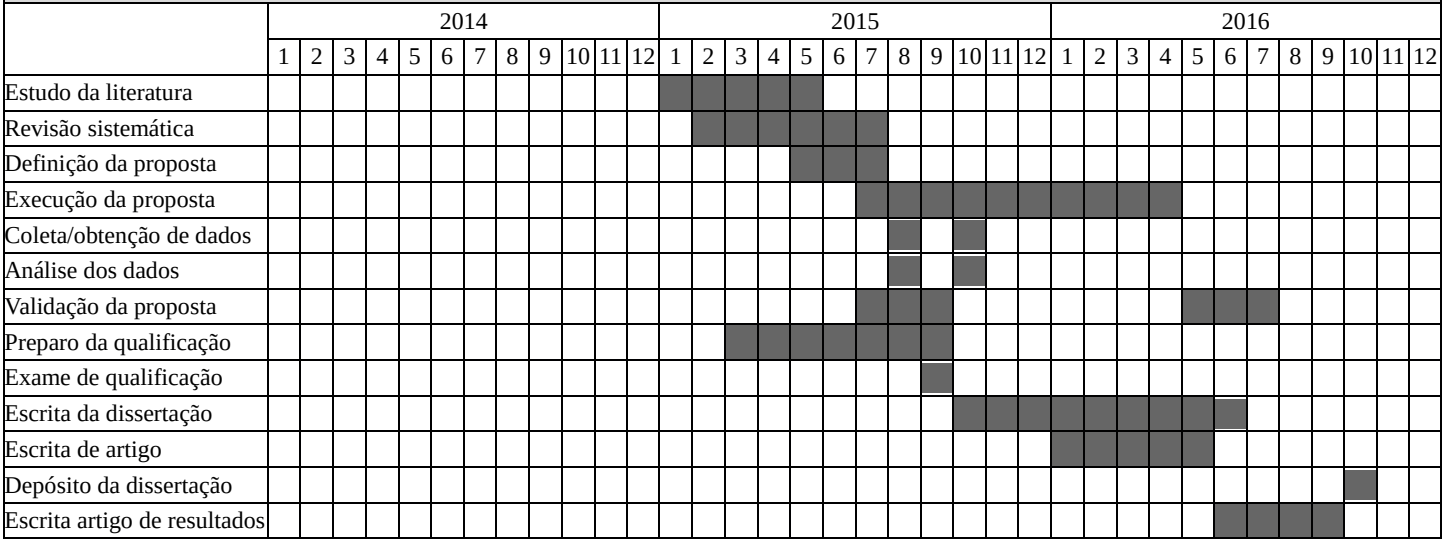
DADOS GERAIS					
Título do projeto de pesquisa	Aprendizado semi-supervisionado para o tratamento de incerteza na rotulação de dados de química medicinal				
Orientando	João Carlos Silva de Souza – N. Usp: 8900551				
Orientador(es)	Profa. Dra. Patrícia Rufino Oliveira e Profa. Dra. Káthia Maria Honório				
Semestre no curso, na data do workshop	[ ] 2º semestre	[ ] 3º semestre	[ ] 4º semestre	[ ] 5º semestre	[x] 6º semestre
Qualificação	[x] Qualificação já realizada em: 17/09/2015 [ ] Realização da qualificação planejada para: dd/mm/aaaa				
Defesa	Prazo máximo para depósito: 30/01/2017 Realização da defesa planejada para: 10/2016				
Linha e Área de pesquisa	Gestão e desenvolvimento de Sistemas: [ ] BD [ ] Gestão de SI [ ] Eng. de Software [ ] IHC	Inteligência de Sistemas: [x] IA [ ] Rec. de Padrões [ ] Proc. Gráfico			
Área de aplicação	[ ] Ambientes Corporativos [ ] Bioinformática [ ] Biometria [ ] Economia	[ ] Educação [ ] Educação a Distância [ ] Internet [ ] Jogos	[ ] Linguagem Natural [ ] Linguística [ ] Processos de Negócio [x] Química	[ ] Redes Sociais [ ] Robótica [ ] Saúde [ ] [outro – escrever]	
Publicações associadas ao projeto de mestrado	Título: Recent Advances for Handling Imbalancement and Uncertainty in Labelling in Medicinal Chemistry Data Analysis / Status: Aceito / "Conferência já realizada. Aguardando publicação e indexação pelo IEEE Explorer conforme informado por contato da conferência". / Veículo: SAI Computing Conference 2016. Título: Handling Label Uncertainty on Chemistry Data Using Weighted Positive and Unlabeled Learning / Status: Em elaboração / Veículo: Ainda não definido.				

DESCRIÇÃO DO PROJETO DE PESQUISA	
Contextualização / motivação	Ele é um caso especifico no tratamento de incerteza na rotulação e desbalanceamento de dados de química medicinal. Estas características refletem negativamente em processos de classificação binária.
Problema de pesquisa	Baixo desempenho no processo de classificação de compostos devido à falta de instâncias que possuem rotulação confiável, métodos não sensíveis a diferente distribuição das classes e fase de treinamento demorada.
Objetivo geral da pesquisa	*Estimar uma rotulação mais confiável para os compostos utilizados no processo de descoberta de novos medicamentos na presença de dados desbalanceados.*
Trabalhos relacionados	Virtual Screening of Bioassy Data: Apresenta resultados utilizando alguns dos conjuntos de dados utilizados. Apesar de contemplar o desbalanceamento das classes, não fornece nenhum tratamento para o problema da incerteza - Genome-wide sequence based prediction of peripheral proteins using a novel semi-supervised learning technique – Aplica a técnica de aprendizado semi-supervisionado, mas não apresenta nenhuma abordagem sensível ao custo de erro de classificação. - Partially Supervised Classification of Text Documents: Não é especificamente de química medicinal, mas apresenta em detalhes a técnica de aprendizado semi-supervisionado.
Justificativa e relevância	A incerteza ou falta de rotulação são comuns nos dados de química medicinal, principalmente para as classes que não são alvos dos estudos. Encontrar uma forma eficiente de tratar esse problema, além de considerar o relevante cenário de desbalanceamento simultaneamente, promete melhorar o desempenho de métodos computacionais relacionados e aprimorar resultados no processo geral de descoberta de novas drogas.
Proposta para Solução	Utilizar a informação contida nos dados positivos e nos dados não rotulados para encontrar os rótulos adequados para as instâncias sob análise. De posse dessa nova rotulação, construir um classificador de rápido treinamento e alto desempenho, levando em consideração a quantidade deferente de exemplos em cada classe.
Dados	*Serão utilizados, para validação, dados de referência presentes na literatura de aprendizagem de máquina, e dados de química medicinal de domínio público presentes no repositório UCI para avaliação do modelo.*
Validação	*Além da comparação gráfica de desempenho, os exemplos do conjunto de treinamento serão identificados antes do processo e após a execução as instâncias com dados alterados serão comparadas com seu estado inicial. Espera-se que somente os exemplos próximos a uma área de sensibilidade sejam alterados.*

Limitações	*Proposta não garante que o fator de custo de classificação seja ótimo e apesar da alteração dos rótulos ser feita em uma região apropriada, somente testes de laboratório podem comprovar os efeitos dos compostos com rótulos modificados.*
Resultados esperados	Contribuições científicas: *Abordagem de estimação de rótulo que leva em consideração o custo do erro de classificação devido ao desbalanceamento e a incerteza presente nos dados.*
	Contribuições tecnológicas: Espera-se que o sistema proposto seja capaz de identificar rotulações confiáveis em diversos processos genéricos de classificação independente da distribuição das classes envolvidas.

MÉTODO DE PESQUISA [para cada item, marque todas as opções adequadas para seu projeto] (basear-se em: http://www.ufrgs.br/cursopgdr/downloadsSerie/derad005.pdf)				
Gênero	[x] Pesquisa teórica	[x] Pesquisa prática	[ ] Pesquisa empírica	[ ] Pesquisa metodológica
Natureza	[ ] Pesquisa básica/pura	[x] Pesquisa aplicada		
Objetivo	[x] Pesquisa exploratória	[ ] Pesquisa descritiva	[ ] Pesquisa explicativa	[ ] [outro – escrever]
Abordagem	[x] Pesquisa quantitativa	[ ] Pesquisa qualitativa	[ ] Pesquisa mista (quali-quanti)	
Procedimento(s) técnico(s)	[ ] Pesquisa experimental [x] Pesquisa bibliográfica [ ] Pesquisa documental [ ] Pesquisa *ex-post-facto* [ ] Pesquisa de levantamento	[x] Pesquisa com *survey* [ ] Estudo de caso [ ] Pesquisa participante [ ] Pesquisa-ação [ ] Pesquisa etnográfica	[ ] Pesquisa netnográfica [x] Teoria fundamentada em dados (*grounded theory*) [ ] Ciência do projeto (*Design science research*) [ ] [outro – escrever]	
Fonte(s) de dados	[ ] pesquisa de laboratório	[ ] pesquisa de campo	[x] pesquisa bibliográfica	[ ] [outro – escrever]
Técnica(s) / Instrumento(s) de coleta de dados	[x] medição [ ] questionário [ ] entrevista [ ] grupos focais	[ ] formulário [x] *benchmark*	[ ] observação (direta / participante) [ ] diário de campo / notas de campo [ ] análise documental (ou de artefatos) [ ] [outro – escrever]	
Técnica(s) de análise de dados	Análise quantitativa: [x] Estatística descritiva [ ] Estatística inferencial [ ] [outro – escrever]		Análise qualitativa: [ ] Análise de conteúdo [ ] Análise do discurso [ ] [outro – escrever]	

CRONOGRAMA [altere conforme necessário]



i Esta ficha é uma adaptação da usada no "VIII Workshop de Teses e Dissertações em Sistemas de Informação (WTDSI 2015)" realizado como parte do "XI Simpósio Brasileiro de Sistemas de Informação (SBSI 2015)"