II Workshop de Dissertações de Mestrado do PPgSI

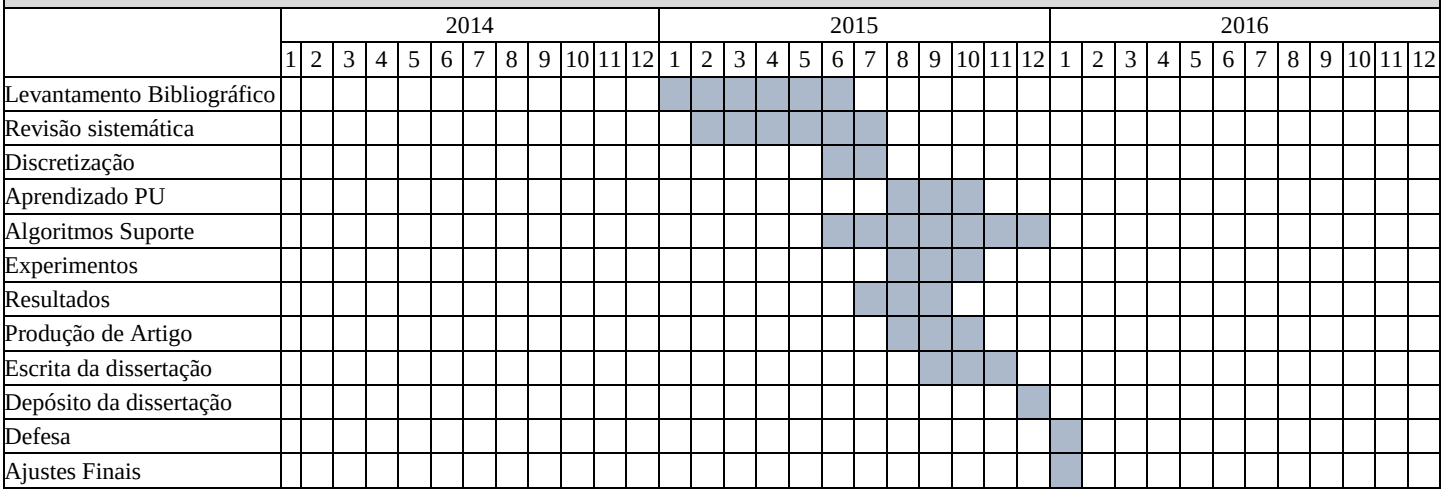
FICHA DA PESQUISA[i]

DADOS GERAIS	
Título do projeto de pesquisa	Aprendizado de dados positivos e não rotulados para o tratamento de incerteza na rotulação de dados de química medicinal
Orientando	João Carlos Silva de Souza
Orientador(es)	Profa. Dra. Patrícia Rufino Oliveira e Profa. Dra. Káthia Maria Honório
Momento atual	[ ] 3º semestre [X] 4º semestre [ ] 5º semestre [ ] 6º semestre
Qualificação	[X] Qualificação já realizada em:17/09/2015 [ ] Qualificação planejada para:____/____/_____
Defesa	Prazo máximo para depósito: _30/01/2017 Depósito planejado para: 19/12/2015
Linha e Área de pesquisa	[ ] Gestão e desenvolvimento de Sistemas: [ ] BD [ ] Gestão de SI [ ] Eng. de Software [ ] IHC [X] Inteligência de Sistemas: [X] IA [ ] Rec. de Padrões [ ] Proc. Gráfico
Área de aplicação	[ ] Ambientes Corporativos [ ] Educação [ ] Linguagem Natural [ ] Redes Sociais [X] Bioinformática [ ] Educação a Distância [ ] Linguística [ ] Robótica [ ] Biometria [ ] Internet [ ] Processos de Negócio [ ] Saúde [ ] Economia [ ] Jogos [ ] Química

DESCRIÇÃO DO PROJETO DE PESQUISA	
Contextualização / motivação	O projeto aborda o problema da incerteza na rotulação de dados e seus efeitos em processos de classificação, mais especificamente, este trabalho se destina a tratar o problema de incerteza em conjuntos de dados medicinais com o principal intuito de melhorar os modelos de classificação criados a partir de dados incertos ou não rotulados.
Problema de pesquisa	Reduzir o impacto na classificação de dados de química medicinal em cenários onde existe incerteza ou ausência na rotulação.
Objetivo geral da pesquisa	Utilizar uma técnica semi-supervisionada de aprendizagem de máquina para estimar uma rotulação confiável para os compostos candidatos a novos fármacos.
Trabalhos relacionados	Machine learning methods for property prediction in chemoinformatics: Quo Vadis? Graph kernels for chemical informatics Genome-wide sequence-based prediction of peripheral proteins using a novel semi-supervised learning technique
Justificativa e relevância	A rotulação incorreta das instâncias utilizadas na fase de treinamento do modelo prejudica a eficiência da classificação de novos compostos. Desse modo, encontrar uma forma de estimar corretamente o rótulo dos candidatos a fármacos, otimiza o processo de descoberta de medicamentos. Apesar de alguns trabalhos terem obtido sucesso na análise da estrutura dos compostos, este trabalho propõe uma análise que não depende de conversão de dados, ao contrário do que acontece principalmente nas análises baseadas em grafos.
Proposta para Solução	Será avaliada uma técnica de aprendizado semi-supervisionado que além dos dados rotulados, utiliza os dados rotulados com incerteza, na tentativa de estimar uma rotulação confiável. Em seguida, a nova rotulação será utilizada para construir um classificador baseado no aprendizado extremo de máquinas.
Dados	Será utilizado a base de dados pública PubChem Bioassay Data Set, na qual existe a possibilidade de extrair o valor do índice de atividade e a rotulação, e uma base de dados contendo dados sobre compostos para o tratamento de diabetes *Mellitus* e síndrome metabólica.
Forma de validação	Após o processamento dos dados pretende-se analisar classificador final comparando casos onde os dados foram estimados no aprendizado PU e sem aprendizado PU.
Limitações	Aquisição de base de dados com quantidade de elementos satisfatórios e definição confiável do limiar de atividade biológica.
Resultados esperados	Contribuições científicas: Espera-se provar que as rotulações estimadas pelo aprendizado PU contribuam de forma positiva na construção de um classificador baseado em aprendizado extremo de máquinas.
	Contribuições tecnológicas: Uso de máquinas de aprendizado extremo na classificação de dados de química medicinal.

MÉTODO DE PESQUISA			
Gênero	[ ] Pesquisa teórica	[X] Pesquisa prática	[ ] Pesquisa empírica [ ] Pesquisa metodológica
Natureza	[ ] Pesquisa básica/pura	[X] Pesquisa aplicada	
Objetivo	[ ] Pesquisa descritiva	[X] Pesquisa exploratória	[ ] Pesquisa explicativa
Abordagem	[X] Pesquisa quantitativa	[ ] Pesquisa qualitativa	[ ] Pesquisa mista (quali-quanti)
Procedimento(s) técnico(s)	[ ]Pesquisa experimental [ ] Pesquisa bibliográfica [ ] Pesquisa documental [ ]Pesquisa *ex-post-facto* [ ] Pesquisa de levantamento	[ ] 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*)
Fonte(s) de dados	[ ] pesquisa de laboratório	[ ] pesquisa de campo	[X] pesquisa bibliográfica
Técnica(s) / Instrumento(s) de coleta de dados	[ ] medição [ ] questionário [ ] entrevista [ ] grupos focais	[ ] formulário [ ] *benchmark*	[ ] observação (direta / participante) [ ] diário de campo/notas de campo [ ] análise documental (ou de artefatos) [X] Análise de banco de dados públicos de química medicinal
Técnica(s) de análise de dados	[X] Análise quantitativa: [ ] Estatística descritiva [ ] Estatística inferencial		[ ] Análise qualitativa: [ ] Análise de conteúdo [ ] Análise do discurso

CRONOGRAMA



[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)"