II Workshop de Dissertações de Mestrado do PPgSI

FICHA DA PESQUISA[i]

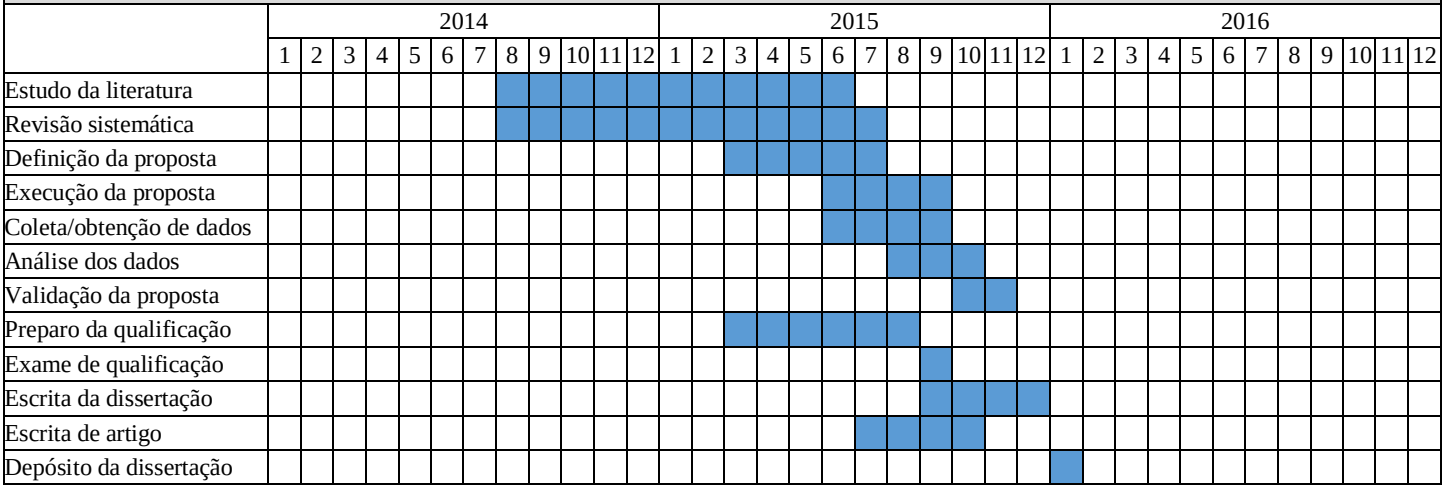
DADOS GERAIS		
Título do projeto de pesquisa	Tratamento do desbalanceamento de classes em problemas de química medicinal.	
Orientando	Suzana Gomes Claudino	
Orientador(es)	Profa. Dra. Káthia Maria Honório e Profa. Dra. Patrícia Rufino Oliveira	
Momento atual	[ ] 3º semestre [x] 4º semestre [ ] 5º semestre [ ] 6º semestre	
Qualificação	[x] Qualificação já realizada em: 18/09/2015	[ ] Qualificação planejada para: ____/____/_____
Defesa	Prazo máximo para depósito: 31/01/2016	Depósito planejado para: 20/12/2015
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 [ ] Bioinformática [ ] Educação a Distância [ ] Linguística [ ] Robótica [ ] Biometria [ ] Internet [ ] Processos de Negócio [ ] Saúde [ ] Economia [ ] Jogos [x] Química [ ] [outro – escrever]	

DESCRIÇÃO DO PROJETO DE PESQUISA	
Contextualização / motivação	A busca de novos medicamentos é uma área importante da química medicinal, porém não é uma tarefa simples. Técnicas de classificação e aprendizagem de máquina podem ser utilizadas para auxiliar algumas tarefas neste processo.
Problema de pesquisa	A classificação de conjuntos de dados de natureza química sofre constantemente com o problema de desbalanceamento de dados, quando se trata de classificação binária entre compostos ativos e inativos frente a um alvo biológico.
Objetivo geral da pesquisa	Aplicar a técnica HC-kNN a fim de tratar o problema de desbalanceamento de classes na classificação de conjuntos de dados de natureza química.
Trabalhos relacionados	*A Hybrid Coupled k-Nearest Neighbor Algorithm on Imbalance Data* – Técnica proposta para classificação de conjuntos mistos com características numéricas e categóricas em conjuntos com dados desbalanceados. *Influence relevance voting: an accurate and interpretable virtual high throughput screening method.* – Técnica proposta para tratamento de desbalanceamento na classificação de dados químicos. *Machine learning methods for property prediction in chemoinformatics: Quo Vadis?* – Apresentar o estado da arte em pesquisas relacionadas a técnicas de prendizagem de máquina na busca por novos medicamentos.
Justificativa e relevância	O desbalanceamento de classes é um problema de classificação, pois muitas técnicas levam em consideração a informação a priori de que todos os compostos têm a mesma importância, o que faz com que a classificação seja tendenciosa para a classe majoritária e no contexto de quimioinformática, a classe de maior interesse é justamente a classe minoritária (compostos que apresentam rotulo "ativo", ou seja, que apresentam atividade biológica frente a um alvo biológico).
Proposta para Solução	Está sendo avaliado o desempenho da técnica HC-kNN na classificação de compostos químicos e realizada a comparação com outras técnicas como o k-NN, variando as medidas de similaridade utilizadas.
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	Serão utilizados a validação cruzada e análise da curva ROC para a validação dos dados.
Limitações	Aquisição de base de dados com quantidade de elementos satisfatórios e definição confiável do limiar de atividade biológica que apresentem o comportamento de desbalanceamento severo.

Resultados esperados	Contribuições científicas: Espera-se que a técnica HC-kNN se apresente como uma alternativa para a utilização na busca de novos medicamentos.
	Contribuições tecnológicas: Será avaliado o desempenho da técnica em classificação binária e com saída de em forma de *ranking* pelo nível de atividade.

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 [ ] [outro – escrever]
Abordagem	[x] Pesquisa quantitativa	[ ] Pesquisa qualitativa	[ ] Pesquisa mista (quali-quanti)
Procedimento(s) técnico(s)	[x] 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 [ ] Teoria fundamentada em dados (*grounded theory*) [ ] Ciência do projeto (*Design science research*) [ ] [outro – escrever]
Fonte(s) de dados	[x] pesquisa de laboratório	[ ] pesquisa de campo	[x] pesquisa bibliográfica [ ] [outro – escrever]
Técnica(s) / Instrumento(s) de coleta de dados	[ ] medição [ ] questionário [ ] entrevista [ ] grupos focais	[ ] formulário [ ] *benchmark*	[] observação (direta / participante) [X] 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: [x] Estatística descritiva [ ] Estatística inferencial [ ] [outro – escrever]		[ ] Análise qualitativa: [ ] Análise de conteúdo [ ] Análise do discurso [ ] [outro – escrever]

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)"