

## II Workshop de Dissertações de Mestrado do PPgSI

### FICHA DA PESQUISA<sup>i</sup>

DADOS GERAIS	
<b>Título do projeto de pesquisa</b>	Tratamento do desbalanceamento de classes em problemas de química medicinal.
<b>Orientando</b>	Suzana Gomes Claudino
<b>Orientador(es)</b>	Profa. Dra. Káthia Maria Honório e Profa. Dra. Patrícia Rufino Oliveira
<b>Momento atual</b>	<input type="checkbox"/> 3º semestre <input type="checkbox"/> 4º semestre <input type="checkbox"/> 5º semestre <input checked="" type="checkbox"/> 6º semestre
<b>Qualificação</b>	<input checked="" type="checkbox"/> Qualificação já realizada em: 18/09/2015 <input type="checkbox"/> Qualificação planejada para: ____/____/____
<b>Defesa</b>	Prazo máximo para depósito: 31/01/2017    Depósito planejado para: 20/12/2016
<b>Linha e Área de pesquisa</b>	<input type="checkbox"/> Gestão e desenvolvimento de Sistemas: <input type="checkbox"/> BD <input type="checkbox"/> Gestão de SI <input type="checkbox"/> Eng. de Software <input type="checkbox"/> IHC <span style="margin-left: 100px;"> <input checked="" type="checkbox"/> Inteligência de Sistemas:  <input checked="" type="checkbox"/> IA                                        <input type="checkbox"/> Rec. de Padrões  <input type="checkbox"/> Proc. Gráfico         </span>
<b>Área de aplicação</b>	<input type="checkbox"/> Ambientes Corporativos <input type="checkbox"/> Educação <input type="checkbox"/> Linguagem Natural <input type="checkbox"/> Redes Sociais <input type="checkbox"/> Bioinformática <input type="checkbox"/> Educação a Distância <input type="checkbox"/> Linguística <input type="checkbox"/> Robótica <input type="checkbox"/> Biometria <input type="checkbox"/> Internet <input type="checkbox"/> Processos de Negócio <input type="checkbox"/> Saúde <input type="checkbox"/> Economia <input type="checkbox"/> Jogos <input checked="" type="checkbox"/> Química <input type="checkbox"/> [outro – escrever]

DESCRIÇÃO DO PROJETO DE PESQUISA	
<b>Contextualização / motivação</b>	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.
<b>Problema de pesquisa</b>	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.
<b>Objetivo geral da pesquisa</b>	Aplicar a técnica HC-kNN a fim de tratar o problema de desbalanceamento de classes na classificação de conjuntos de dados de natureza química.
<b>Trabalhos relacionados</b>	<p><i>A Hybrid Coupled k-Nearest Neighbor Algorithm on Imbalance Data</i> – Técnica proposta para classificação de conjuntos mistos com características numéricas e categóricas em conjuntos com dados desbalanceados.</p> <p><i>Influence relevance voting: an accurate and interpretable virtual high throughput screening method.</i> – Técnica proposta para tratamento de desbalanceamento na classificação de dados químicos.</p> <p><i>Machine learning methods for property prediction in chemoinformatics: Quo Vadis?</i> – Apresentar o estado da arte em pesquisas relacionadas a técnicas de aprendizagem de máquina na busca por novos medicamentos.</p>
<b>Justificativa e relevância</b>	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).
<b>Proposta para Solução</b>	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.
<b>Dados</b>	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 <i>Mellitus</i> e síndrome metabólica.
<b>Forma de validação</b>	Serão utilizados a validação cruzada e análise da curva ROC para a validação dos dados.
<b>Limitações</b>	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.

<b>Resultados esperados</b>	Contribuições científicas: Espera-se que a técnica HC-kNN se apresente como uma alternativa para a utilização na busca de novos medicamentos.
	Contribuições tecnológicas: Ajustes no algoritmo para melhora dos resultados obtidos.

MÉTODO DE PESQUISA			
<b>Gênero</b>	<input type="checkbox"/> Pesquisa teórica	<input checked="" type="checkbox"/> Pesquisa prática	<input type="checkbox"/> Pesquisa empírica <input type="checkbox"/> Pesquisa metodológica
<b>Natureza</b>	<input type="checkbox"/> Pesquisa básica/pura <input checked="" type="checkbox"/> Pesquisa aplicada		
<b>Objetivo</b>	<input type="checkbox"/> Pesquisa descritiva	<input checked="" type="checkbox"/> Pesquisa exploratória	<input type="checkbox"/> Pesquisa explicativa <input type="checkbox"/> [outro – escrever]
<b>Abordagem</b>	<input checked="" type="checkbox"/> Pesquisa quantitativa	<input type="checkbox"/> Pesquisa qualitativa	<input type="checkbox"/> Pesquisa mista (quali-quant)
<b>Procedimento(s) técnico(s)</b>	<input checked="" type="checkbox"/> Pesquisa experimental <input type="checkbox"/> Pesquisa bibliográfica <input type="checkbox"/> Pesquisa documental <input type="checkbox"/> Pesquisa <i>ex-post-facto</i> <input type="checkbox"/> Pesquisa de levantamento	<input type="checkbox"/> Pesquisa com <i>survey</i> <input type="checkbox"/> Estudo de caso <input type="checkbox"/> Pesquisa participante <input type="checkbox"/> Pesquisa-ação <input type="checkbox"/> Pesquisa etnográfica	<input type="checkbox"/> Pesquisa netnográfica <input type="checkbox"/> Teoria fundamentada em dados ( <i>grounded theory</i> ) <input type="checkbox"/> Ciência do projeto ( <i>Design science research</i> ) <input type="checkbox"/> [outro – escrever]
<b>Fonte(s) de dados</b>	<input checked="" type="checkbox"/> pesquisa de laboratório	<input type="checkbox"/> pesquisa de campo	<input checked="" type="checkbox"/> pesquisa bibliográfica <input type="checkbox"/> [outro – escrever]
<b>Técnica(s) / Instrumento(s) de coleta de dados</b>	<input type="checkbox"/> medição <input type="checkbox"/> questionário <input type="checkbox"/> entrevista <input type="checkbox"/> grupos focais	<input type="checkbox"/> formulário <input type="checkbox"/> <i>benchmark</i>	<input type="checkbox"/> observação (direta / participante) <input checked="" type="checkbox"/> diário de campo / notas de campo <input type="checkbox"/> análise documental (ou de artefatos) <input checked="" type="checkbox"/> Análise de banco de dados públicos de química medicinal
<b>Técnica(s) de análise de dados</b>	<input checked="" type="checkbox"/> Análise quantitativa: <input checked="" type="checkbox"/> Estatística descritiva <input type="checkbox"/> Estatística inferencial <input type="checkbox"/> [outro – escrever]	<input type="checkbox"/> Análise qualitativa: <input type="checkbox"/> Análise de conteúdo <input type="checkbox"/> Análise do discurso <input type="checkbox"/> [outro – escrever]	

CRONOGRAMA																																									
	2014												2015												2016																
	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12	1	2	3	4	5	6	7	8	9	10	11	12					
Estudo da literatura																																									
Revisão sistemática																																									
Definição da proposta																																									
Execução da proposta																																									
Coleta/obtenção de dados																																									
Análise dos dados																																									
Validação da proposta																																									
Preparo da qualificação																																									
Exame de qualificação																																									
Escrita da dissertação																																									
Escrita de artigo																																									
Depósito da dissertação																																									

<sup>i</sup> 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)”