MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PIAUÍ
PRÓ-REITORIA DE PESQUISA
COORDENAÇÃO DE INOVAÇÃO TECNOLÓGICA - CITEC
Programa Institucional de Bolsas de Iniciação em Desenvolvimento Tecnológico e Inovação- PIBITI
Campus Universitário Ministro Petrônio Portella, Bloco 06 – Bairro Ininga
CEP: 64049-550 – Teresina-PI – Brasil – Fone (86)3237-1426 – Fone/Fax (86)3235-5560
E-mail: cict@ufpi.edu.br

ANEXO II

Edital PIBITI CNPq e UFPI (2017-2018)

Projeto de Desenvolvimento Tecnológico e Inovação

Dados do Projeto e do Proponente

Proponente (Centro, Departamento).	Prof. Dr. Vinicius Ponte Machado Departamento de Computação – CCN
Título do Projeto	Aplicativo para Rotulação de Clusters
Colaboradores: (Centro, Departamento).	Prof. Dr. Kelson Rômulo Teixeira Aires Prof. Dr. Rodrigo de Melo Souza Verás Departamento de Computação – CCN
Área:	( ) Tecnologia e Inovação em Agropecuária (X) Tecnologia da Informação e Comunicação ( ) Tecnologias Químicas e Novos Materiais ( ) Biotecnologia, Recursos Naturais e Tecnologias da Saúde ( ) Outras tecnologias

Palavras-Chave: ***Rotulação, Dataminning, Descoberta de conhecimento.***

2. Relatório de anterioridade tecnológica / Estudo Prospectivo (até 2 páginas)

O problema de agrupamento (clustering) tem sido considerado como um dos problemas mais relevantes dentre aqueles existentes na área de pesquisa de aprendizagem não-supervisionada (subárea de Aprendizagem de Máquina).

Clustering se refere ao processo de particionar um conjunto de dados (ou objetos) em subconjuntos menores denominados clusters ou simplesmente grupos. Nesse processo, os objetos que possuem similaridades em suas características tendem pertencer a um mesmo cluster enquanto que objetos com características diferentes tendem pertencer a clusters distintos. Conforme Han et al. (2011), o processo de clustering tem sido amplamente utilizado em diversas aplicações como inteligência empresarial, reconhecimento de padrões em imagem, busca na Web, biologia, segurança, entre várias outras. No contexto empresarial, o agrupamento pode ser utilizado para organizar uma grande quantidade de clientes em grupos. Isso facilita o desenvolvimento de estratégias de negócio para uma melhor gestão de relacionamento com o cliente. Na área de reconhecimento de padrões em imagem pode ser utilizado, por exemplo, para aumentar a precisão de sistemas de reconhecimento de escrita. Em buscas na Web, clustering pode organizar a informação em assuntos similares apresentado-a de maneira concisa.

Embora o desenvolvimento e aprimoramento de algoritmos que solucionam esse problema tenha sido o principal foco de muitos pesquisadores o objetivo inicial se manteve obscuro: a compreensão dos grupos formados. Tão importante quanto a identificação dos grupos (clusters) é sua compreensão e definição. Uma boa definição de um cluster representa um entendimento significativo e pode ajudar o especialista ao estudar ou interpretar dados. Portanto, existe a necessidade de descobrir o que caracteriza cada cluster formado. A existência de um rótulo permite a identificação de quais características definem um grupo. Assim, o rótulo pode ser útil para a identificação de quais características necessitam de ações corretivas e até mesmo o quão intensa ela deve ser baseada nos valores das características. Dessa forma, a compreensão de clusters por intermédio de rótulos pode contribuir de diversos modos com a elaboração da solução ou otimização de um problema.

Frente ao problema de compreender clusters – isto é, de encontrar uma definição ou em outras palavras, um rótulo – este projeto propõe uma definição para esse problema, denominado problema de rotulação, além de uma solução baseada em técnicas com aprendizagem supervisionada, não-supervisionada e um modelo de discretização. Dessa forma, o problema é tratado desde sua concepção: o agrupamento de dados. Para isso, um método com aprendizagem não-supervisionada é aplicado ao problema de clustering e então um algoritmo com aprendizagem supervisionada irá detectar quais atributos são relevantes para definir um dado cluster.

A presente proposta apresenta o desenvolvimento um aplicativo (software) que implementa uma abordagem, baseados em algoritmos de aprendizagem de máquina supervisionados, capaz de rotular clusters a fim de esclarecer, orientar e ajudar um especialista. Os rótulos gerados devem ser capazes de identificar as principais características – bem como seus conjuntos de valores – responsáveis pela definição de um determinado cluster.

2.1 Relevância tecnológica e inovadora do projeto.

Nossa proposta é fazer um aplicativo que possa realizar de forma automática a rotulação de dados. Esse aplicativo funcionará em de forma independente (stand-alone[1]) e também em conjunto com a plataforma WEKA[2] (*Waikato Environment for Knowledge Analysis*). Este software começou a ser escrito em 1993, usando Java, na Universidade de Waikato, Nova Zelândia sendo adquirido posteriormente por uma empresa no final de 2006. O WEKA encontra-se licenciado ao abrigo da *General Public License* sendo portanto possível estudar e alterar o respectivo código fonte.

[1] São chamados stand alone, ou stand-alone os programas completamente autossuficientes: para seu funcionamento não necessitam de um software auxiliar, como um interpretador, sob o qual terão de ser executados.

[2] www.cs.waikato.ac.nz/ml/weka/

O WEKA tem como objetivo agregar algoritmos provenientes de diferentes abordagens/paradigmas na sub-área da inteligência artificial dedicada ao estudo de aprendizagem de máquina. Essa sub-área pretende desenvolver algoritmos e técnicas que permitam a um computador "aprender" (no sentido de obter novo conhecimento) quer indutiva quer dedutivamente. O WEKA procede à análise computacional e estatística dos dados fornecidos recorrendo a técnicas de mineração de dados tentando, indutivamente, a partir dos padrões encontrados gerar hipóteses para soluções e no extremos inclusive teorias sobre os dados em questão.

Não há, todavia, nenhum algoritmo no WEKA, que realize o processo de rotulação proposto, apesar dele apresentar diversas implementações de algoritmos da agrupamento não supervisionado.

Ressalta-se ainda que a base de desenvolvimento do software proposto pode ser estendida ou incorporada à outras ferramentas (Matlab, R, Excel, etc.). Isso, graças a metodologia de desenvolvimento a ser empregada, que faz com que a grande parte do software possa ser reaproveitada no desenvolvimento de outros sistemas com a mesma finalidade.

2.2 Pesquisa de anterioridade

O software proposto possui caráter inovador uma vez que não há nenhum software com o mesmo propósito com registro ou pedido de registro na base do INPI[3] (Instituto Nacional da Propriedade Industrial). Nesta base foram realizadas buscas usando as palavras-chaves *clustering, agrupamento, rotulação, weka* e *mineração de dados,* além das combinações entre estas palavras. Apesar de alguns resultados retornados, nenhum deles tem relação com tema proposto.

Contudo, o tema desperta interesse da comunidade de desenvolvimento de software uma vez que foram encontrados diversos trabalhos usam o processo de agrupamento e interpretação dos dados. Um exemplo disso é um software para reconhecimento de faces[4] usando agrupamento Fuzzy. Como este software está sob sigilo não termos maiores detalhes.

Outro exemplo de software com essa mesma temática é o CLASS - Clustering And Analysis Of Scrum Stories[5] que emprega métodos de agrupamento para avaliar as Stories (histórias dos usuários) utilizada para auxiliar no desenvolvimento de sistemas geridos segundo metodologias ágeis.

Ressalta-se que, apesar de trabalhos relacionados ao mesmo tema do sistema proposto, nenhum deles utiliza um processo rotulação proposto neste projeto.

[3] http://www.inpi.gov.br
[4] https://gru.inpi.gov.br/pePI/servlet/ProgramaServletController?Action=detail&CodPedido=20018
[5] https://gru.inpi.gov.br/pePI/servlet/ProgramaServletController?Action=detail&CodPedido=21204

3. Objetivos e Metas (máximo de 1 página)

Dentro desse contexto ainda promissor do tema, este trabalho tem como objetivos:

a) Objetivo Geral

Estudar, conceber, desenvolver e avaliar um software que, dado um certo número de grupos oriundos de algoritmos de aprendizagem de máquina não supervisionados, aplicar outros algoritmos baseados em aprendizagem supervisionada em cada grupo formado para detectar quais atributos – e seus respectivos valores ou faixas de valores – podem ser utilizados para defini-los. Em outras palavras, o método proposto busca apresentar um rótulo, capaz de definir um determinado grupo com base em suas características.

b) Objetivos Específicos

1. Realizar levantamento técnico sobre o cenário da computação aplicada à descoberta de conhecimento em base de dados;
2. Investigar, especificar e implementar elementos de interface homem-máquina que podem facilitar o uso do aplicativo e serviço nas mais diversas situações onde são necessárias;
3. Disponibilizar, em caráter experimental, uma solução (produto) que realize o processo de rotulação.
4. Avaliar, junto a um grupo de usuários, os produtos de software desenvolvidos.

Para alcançar estes objetivos traçamos as seguintes metas:

- Implementar as rotinas de código para o desenvolvimento do aplicativo no modo *stand-alone*.
    - Estudar e fazer uso do SDK (Software Development Kit) nesta implementação.
- Fazer a portabilidade da tecnologia para a plataforma WEKA.
    - Conhecer as limitações de *hardware* dos dispositivos onde será implementada a solução.

4. **Estratégia de Ação** (máximo de 1 página)

A metodologia de desenvolvimento da aplicação. A primeira é o estudo das tecnologias envolvidas. Esta etapa considera uma revisão bibliográfica sobre a descoberta de conhecimento em base de dados. Ainda nesta etapa está prevista a capacitação da equipe na linguagem *Java* (DEITEL, 2005) (utilizada do desenvolvimento das aplicações do sistema)

Na segunda etapa serão feitas os desenvolvimentos dos aplicativos utilizando a linguagem escolhida. Nestas etapas, o desenvolvimento levará em consideração a metodologia de engenharia de software e terão fases de desenvolvimento (implementação) bem definidas:

- Análise de requisitos de software: A extração dos requisitos do software.
- Especificação: Descreve precisamente o software que será escrito, preferencialmente de uma forma matematicamente rigorosa.
- Implementação (ou codificação): A transformação de um projeto para um código deve ser a parte mais evidente do trabalho da engenharia de software, mas não necessariamente a sua maior porção.

A quarta etapa corresponde a Testes e Documentação:

- Criação de um protótipo rotulação automática de clusters.
- Documentação do projeto interno do software para propósitos de futuras manutenções e aprimoramentos.
- Estudo da viabilidade do uso da ferramenta desenvolvida em outro domínio (Weka, Matlab e R)

Por fim, a quinta etapa consiste em registrar o software junto ao INPI e disponibilizar o *produto* desenvolvido para público em geral. As várias plataformas nas quais um desenvolvedor pode escolher para seus aplicativos são mutuamente incompatíveis (ou seja, um aplicativo desenvolvido em uma plataforma não irá executar em outra). Porém, com o uso da plataforma de desenvolvimento na linguagem Java isso não será problema uma vez que ela é executada em diversos sistemas operacionais. Neste caso, pretende-se aproveitar o *know-how* de aplicações anteriormente desenvolvidas oferecendo as novas funcionalidades como novas versões à medida que elas são implementadas. Em resumo, temos o seguinte quadro:

Etapas	**Objetivos**	**Metas**	**Resultados Esperados em cada etapa**
Estudo das Tecnologias	Capacitação da equipe	Entendimento das tecnologias envolvidas e domínio das ferramentas de desenvolvimento	Equipe Capacitada
Desenvolvimento do aplicativo	Implementação dos Serviços	Uso das técnicas de Engenharia de Software no desenvolvimento da aplicação	Protótipo
Teste e Documentação	Criação de um protótipo	Concepção de um aplicativo	Documentação
Registro de Software e Disponibilização ao público	Disponibilização para o público	Entrega de uma versão genérica da Solução	Registro de Software

5. Resultados, Impactos e Planos de proteção tecnológica (máximo de 1 página)

5.1. Resultados Esperados:

Espera-se desenvolver um softwares que *clusters* produzidos através de qualquer algoritmo não-supervisionado, sejam rotulados através da aplicação de um algoritmo com aprendizagem supervisionada em cada grupo formado de forma a detectar quais atributos – e seus respectivos valores ou faixas de valores – podem ser utilizados para defini-los. Em outras palavras, o método apresentado busca apresentar um rótulo, capaz de definir um determinado grupo com base em suas características.

5.2. Impactos Esperados:

O *know-how* adquirido com o projeto propiciará, no futuro, outros produtos relativos a esta mesma tecnologia. Além disso, outros mecanismos de aprendizagem de máquina podem ser incorporados fazendo com que o produto esteja sempre em constante evolução técnica.

Sendo assim, podemos descrever as principais contribuições esperadas em termos qualitativos e quantitativos:

Quantitativos:

- Registro de um software para rotulação de clusters em base de dados;

Qualitativos:

- Estudo e análise dos pontos fortes e fracos das tecnologias atualmente utilizados processo de rotulação;
- Disseminação do conhecimento de tecnologias de aprendizagem de máquina na UFPI;
- Inserção dos alunos de graduação e pós-graduação da UFPI no cenário tecnológico relacionado ao desenvolvimento na linguagem Java.

Ao se definirem as condições para o desenvolvimento deste projeto estaremos abrindo as possibilidades de novas pesquisas na área de desenvolvimento de aplicativos na área de aprendizagem de máquina tanto em nível tecnológico quanto em nível científico. Espera-se contribuir fortemente na melhoria da qualidade de pesquisas na área.

5.3. Plano de depósito ou proteção tecnológica:

Espera-se também que este projeto tenha impacto, principalmente na comunidade tecnológica e científica, através da disseminação dos resultados em *wokshops*, palestras que serão voltadas para a disseminação do desenvolvimento das tecnologias aqui elencadas. Por fim, é importante comentar que a execução deste projeto, de forma indireta, exigirá a criação de um conjunto de ferramentas/softwares que poderão ser utilizadas em outras aplicações e abrirá caminho para novos registros de softwares.

6. Cronograma de execução

O cronograma de atividades descrito a seguir possui 6 fases:

Período / Fases	1º Semestre		2º Semestre	
Fase 1				
Fase 2				
Fase 3				
Fase 4				
Fase 5				
Fase 6				

Fase	Atividades
1	Estudo das Tecnologias
2	Capacitação da equipe de trabalho
3	Implementação do Aplicativo
4	Testes e Documentação
5	Publicação dos resultados

Após estudo sobre as tecnologias envolvidas no projeto e capacitação da equipe, entraremos na fase de desenvolvimento do software onde serão implementadas os planos de trabalhos que consistem no desenvolvimento das funcionalidades aplicativo responsável por realizar o processo de rotulação.

Na fase seguinte, a partir dos testes na interface e nas rotinas de integração de serviços, pretende-se aprimorar as aplicações de modo a disponibilizá-las à comunidade.

Por fim, teremos a compilação da documentação realizada durante todo o processo e a publicação dos resultados. Vale ressaltar que, por questões de tempo algumas fases serão realizadas concomitantemente. Principalmente a implementação das novas funcionalidades.

7. Referências

Han, J., Kamber, M., e Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Lopes, L. A., Machado, V. P., e Rabêlo, R. d. A. L. (2013a). Automatic labeling of groupings through supervised machine learning. Em X Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), pgs. 1–6.

Lopes, L. A., Machado, V. P., e Rabêlo, R. d. A. L. (2013b). Clusters labeling th-rough multi-layer perceptron algorithm. Em XI Simpósio Brasileiro de Automação Inteligente (SBAI), pgs. 1–6.

Chen, M. S., Han, J., e YU, P. S. (1996). Data mining: An overview from a database perspective. IEEE Transactions On Knowledge and Data Engineering, pgs. 866–883.

Deitel, H. M.; Deitel, P. J. Java: Como Programar. 6. Ed. São Paulo: Pearson Education, 2005.

Treeratpituk, P. e Callan, J. (2006). Automatically labeling hierarchical clusters. Em Proceedings of the 2006 International Conference on Digital Government Research, dg.o '06, pgs. 167–176. Digital Government Society of North America.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1):81–106.

Glover, E., Pennock, D. M., Lawrence, S., e Krovetz, R. (2002). Inferring hierarchical descriptions. Em Proceedings of the Eleventh International Conference on Informa- tion and Knowledge Management, CIKM '02, pgs. 507–514. ACM.

Chuang, S.-L. e Chien, L.-F. (2004). A practical web-based approach to generating topic hierarchy for text segments. Em Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, CIKM '04, pgs. 127–136. ACM.

Popescul, A. e Ungar, L. H. (2000). Automatic labeling of document clusters.

Maqbool, O. e Babri, H. (2005). Interpreting clustering results through cluster labeling. Em Proceedings of the IEEE Symposium on Emerging Technologies, 2005., pgs. 429– 434.

Furtado, V. ; Machado, V. P. (2003). Improving Organizational Memory through Agents for Knowledge Discovery in Database. In: Springer-Verlag Heidelberg. (Org.). Lecture Notes in Artificial Inteligence, 2003. 1 ed. Heidelberg: Springer-Verlag , v. 1, p. 164-176.

Machado, V. P.; Furtado, V. (2001). Alimentando a memória organizacional através da descoberta de conhecimento em base de dados. In: 4o. Simpósio Internacional de Gestão do Conhecimento/Gestão de Documentos, 2001, Curitiba. 4o. Simpósio Internacional de Gestão do Conhecimento/Gestão de Documentos.

Machado, V. P. (2003). Uma Arquitetura Multi-Agente para Auxílio à Gestão do Conhecimento. In: KM Ceará 2003, 2003, Fortaleza.

Machado, V.; Dória Neto, A. D.; De Melo, J. D. (2010a) A Neural Network Multiagent

Architecture Applied to Industrial Networks for Dynamic Allocation of Control Strategies Using Standard Function Blocks. IEEE Transactions on Industrial Electronics, v. 57, p. 1823-183.

Machado, V.; Brandão, D.; Dória Neto, A. D.; De Melo, J. D. (2010b). A Multiagent Architecture Based in a Foundation Fieldbus Network Function Blocks for Fault Detection. In: Javier Silvestre. (Org.). Factory Automation. Viena: Intech.

Machado, V.; Brandão, D.; Dória Neto, A. D.; De Melo, J. D.; Ramalho, R.; Medeiros, J. (2008). A Neural Network MultiAgent Architecture Applied to Fieldbus Intelligent Control. In: 13th IEEE International Conference on Emerging Technologies and Factory Automation, 2008, Hamburgo. Proceedings - 13th IEEE International Conference on Emerging Technologies and Factory Automation, p. 567-574.

Machado, V.; Dória Neto, A. D.; De Melo, J. D. (2010a) A Neural Network Multiagent Architecture Applied to Industrial Networks for Dynamic Allocation of Control Strategies Using Standard Function Blocks. IEEE Transactions on Industrial Electronics, v. 57, p. 1823-183.

De Lima, B. V. A.; Machado, V. P. (2012b). Machine learning algorithms applied in automatic classification of social network users. In: 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), 2012, São Carlos. 2012 Fourth International Conference on Computational Aspects of Social Networks (CASoN), p. 58.

Lopes, L. A. ; Machado, V.; Rabêlo, Ricardo, A. L. (2014). Automatic *Cluster* Labeling through Artificial Neural Networks. In: International Joint Conference on Neural Networks 2014, Pequim.