TÍTULO DO TFG/TÍTULO DEL TFG:

Paralelización dun algoritmo de aprendizaxe automática multi-vista baseado en grafos

Resumo/Resumen (máximo 350 palabras):

Nos últimos anos, unha tormenta perfecta chegou ao campo da aprendizaxe automática. Aumentos da potencia computacional, afluencia no interese por analizar grandes conxuntos de datos, e innovacións no manexo de información contribuíron a avances en investigación e uso práctico. O incremento na diversidade dos datos estudados levou a un novo paradigma: a aprendizaxe multi-vista. Os algoritmos multi-vista son capaces de explotar datos estruturados pola súa natureza. Este traballo céntrase en aplicar aprendizaxe multi-vista a problemas de agrupamento non supervisado, que ofrecen información sobre patróns ocultos nos datos. En 2020, presentouse unha solución innovadora ao problema: agrupamento multi-vista baseado en grafos [1]. Combinando métodos de agrupamento por grafos e espectral cun novo mecanismo de fusión, este algoritmo supera o estado da arte.

Os requisitos computacionais deste algoritmo sobre grandes conxuntos de datos son prohibitivos. A implementación de referencia funciona baixo o intérprete de MATLAB, e carece de escalabilidade debido ao seu elevado uso de memoria e complexidade computacional cúbica. Neste traballo de fin de grao desenvolveuse HiPGMC, unha implementación construída dende cero en linguaxe C, e deseñada especificamente para o seu uso en sistemas de computación de altas prestacións, facendo uso tanto de técnicas de paralelismo en memoria compartida (OpenMP) e distribuída (MPI). Esta implementación, altamente optimizada, minimiza o uso de memoria e a complexidade computacional do algoritmo (de cúbica a cuadrática). Empregáronse librerías de álxebra lineal de alto rendemento para acelerar a execución. Os requisitos de memoria reducidos e a posibilidade de executar o código en sistemas de memoria distribuída permitiron analizar conxuntos de datos moito maiores, imposibles de tratar anteriormente.

Leváronse a cabo numerosas probas de rendemento sobre un clúster de computación de altas prestacións, usando 288 núcleos de CPU e 2.25TiB de memoria. O salto dende MATLAB a C é responsable de tempos ata 60 veces menores en probas cos datos que os métodos secuenciais poden procesar. Aínda por riba, o proceso de paralelización obtivo aceleracións, de novo, ata 60 veces máis rápidas cas da implementación en C secuencial, coa posibilidade de escalar máis aló dependendo dos recursos dispoñibles.

O software desenvolvido publicouse baixo a licencia aberta MIT.

[1] H.Wang, Y.Yang, and B.Liu, "GMC: Graph-Based Multi-View Clustering", IEEE Transactions on Knowledge and Data Engineering, vol.32, no.6, pp.1116–1129, 2020.

Posibles aplicacións/Posibles aplicaciones (máximo 250 palabras):

Analizar grandes volumes de datos de forma eficiente é crítico para moitas disciplinas. Estes conxuntos de datos esconden patróns complexos co potencial de revelar información importante para os seus dominios. As solucións baseadas en aprendizaxe automática son populares por adquirir coñecemento dos datos minimizando a interacción humana. Mediante aprendizaxe non supervisada, elíminase a necesidade de que expertos preprocesen os datos manualmente; unha grande vantaxe cando o volume de información fai calquera proceso manual prohibitivo.

As técnicas de agrupamento, ou clustering, son moi potentes na clasificación e minería de datos. Tratan de formar grupos de patróns de información, representados mediante conxuntos disxuntos. Os elementos dentro dun mismo clúster son máis similares entre eles que con aqueles noutros clústers. Empregando esta técnica de forma adecuada, ofrece unha nova perspectiva dos patróns que podería pasarse por alto doutra maneira. As metodoloxías de agrupamento tradicionais presentan problemas ao tratar con datos diversos (texto, imaxes, son...), mais algoritmos multivista coma HiPGMC superan estas dificultades. Todo isto fai que esta ferramenta sexa inestimable para a análise estratéxica na industria. Poder traballar de forma eficiente e automatizada con esta información ten o potencial de axilizar fluxos de traballo de alta complexidade.

A versatilidade do algoritmo faino aplicable a diversos problemas:

- Medicina: Interpretación de escáneres médicos, clasificación de estruturas moleculares.
- Web: Manexo de contido xerado por usuarios, algoritmos de recomendacións.
- Bioloxía: Clasificación de organismos, análise de secuencias xenéticas.
- Marketing: Análise de mercado, clasificación de produtos e intereses.
- Socioloxía: Identificación de grupos e tendencias sociais.
- Finanzas: Clasificación de stocks.

Etapas para o seu desenvolvemento futuro/Etapas para su desarrollo futuro (máximo 250 palabras):

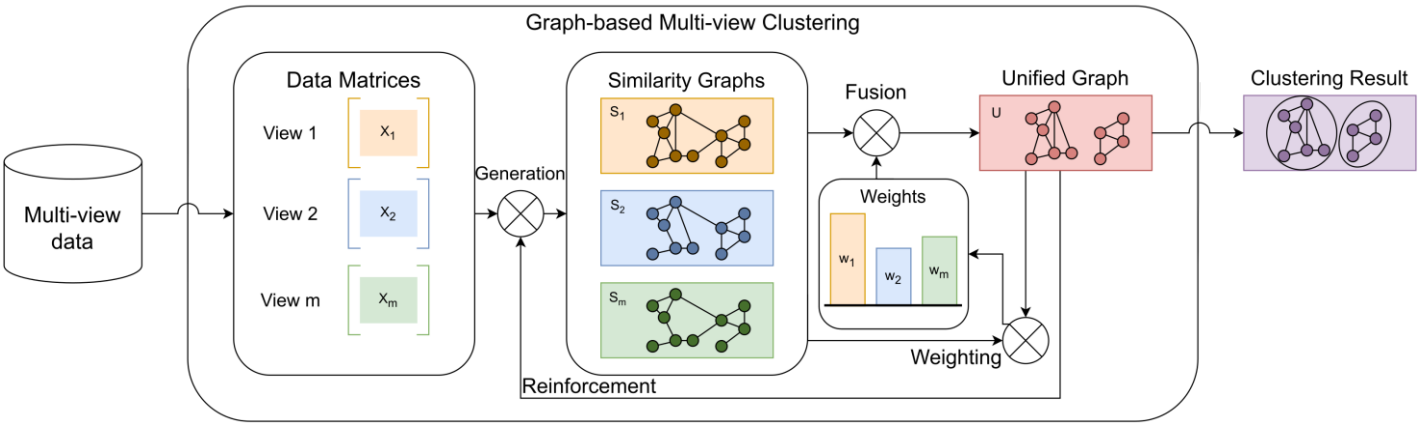
Durante o proxecto, analizáronse as posibilidades de desenvolvemento futuro, así coma potenciais liñas de investigación subsecuentes. Entre elas, destacan:

- Experimentación con outros *eigensolvers*: A investigación no campo da resolución de problemas de autovalores en arquitecturas paralelas é moi activa aínda hoxe en día. Este paso do proceso toma gran tempo e recursos. HiPGMC fai uso da librería ELPA, mais outras eleccións coma ARPACK, BLOPEX, e ChASE tamén teñen un enfoque en arquitecturas clúster e poderían acadar aceleracións aínda superiores.
- Aceleración por GPU: Tecnoloxías coma CUDA permiten exprimir o verdadeiro potencial da computación de propósito xeral en unidades de procesamento gráfico. Algúns dos *eigensolvers* mencionados xa ofrecen aceleracións notables mediante hardware. Esto podería estenderse a outras seccións do algoritmo, nun enfoque híbrido CUDA+MPI.
- Métodos baseados en subespacios: Outras aproximaciones ao agrupamento multivista tamén mostran resultados prometedores, e poderían beneficiarse dun tratamento nas liñas deste traballo. En particular, os algoritmos baseados en subespacios son interesantes pola facilidade coa que manexan datos de alta dimensionalidade.

Por outra banda, dado que o código fonte do proxecto está dispoñible nun repositorio público, calquera persoa cos coñecementos necesarios podería contribuír ou adaptar o software ás súas necesidades.

Imaxes representativas/Imágenes representativas (máximo 2):

Imaxe 1: Diagrama de fluxo de datos a alto nivel do algoritmo. Nel aparecen plasmados o proceso de xeración de grafos a partir dos datos de entrada, así coma a técnica de fusión de vistas e determinación automática de pesos, e o reforzo iterativo na aprendizaxe.



Imaxe 2: Tempos de execución secuenciais comparados coa implementación de referencia. Aínda en casos nos que non facemos uso das capacidades de escalado en contornas clúster, a eficiencia de HiPGMC é remarcable, con tempos de execución ata 60 veces máis rápidos. A implementación de referencia non foi capaz de analizar o conxunto de datos Hdigit sobre o sistema de probas.