Apelidos, nome /Apellidos, nombre:	**DNI:**		**Teléfono de contacto:**
Beceiro Fernández, Bieito			
Título: Parallel-FST: Aceleración de algoritmos de selección de características mediante computación paralela			**Mención cursada:** Computación

Resumo / Resumen:

Na actualidade estase a producir un auxe da produción e consumo de grandes cantidades de información (*big data*), que deben procesarse e prepararse para o seu posterior uso. Entre as ferramentas que se utilizan para analizar estes datos atópanse as de aprendizaxe máquina (*machine learning*), o que constitúe outro campo de investigación que gañou importancia nos últimos anos. A pesar dos seus bos resultados, as técnicas de aprendizaxe automática contan cun custo computacional alto, que se incrementa notablemente ao aumentar a cantidade de datos a procesar. Para reducir a dimensionalidade destes datos, existen algoritmos de **selección de características** que, a través de modelos matemáticos, son capaces de eliminar información redundante e innecesaria. Porén, a selección de características tamén é un proceso custoso, pero que pode acelerarse adaptando os algoritmos e técnicas xa existentes para o seu uso en sistemas de computación paralela (coñecidos como sistemas HPC, *High Performance Computing*).

Ao longo dos últimos anos xurdiron moitos traballos de investigación centrados no desenvolvemento de diferentes métodos de selección de características, cada un aplicando uns criterios de cara á devandita selección. Polo xeral, estes criterios deben tentar maximizar a relevancia das características seleccionadas e minimizar a redundancia entre as mesmas, de forma que o subconxunto escollido represente da mellor forma posible ao dataset orixinal. Tamén existen estudos que traballan con varios destes métodos para atopar o grao de conformidade entre os mesmos, para buscar similitudes a nivel de estrutura ou con intención de determinar cal presenta un mellor comportamento en termos de precisión, estabilidade e flexibilidade ante datasets de certas propiedades. Para este tipo de estudos moitas veces é necesario o desenvolvemento de librarías que conteñan os métodos de selección de características a estudar, de forma que se poidan comparar os resultados. Este é o caso de FEAST, unha libraría que conta con oito métodos de selección de características baseada en **información mutua**, entre eles, CondMI, JMI, mRMR e DISR.

Neste Traballo Fin de Grao desenvolveuse Parallel-FST, unha libraría que, mediante o uso de técnicas avanzadas de **computación paralela**, optimiza e adapta os métodos de FEAST para que poidan ser executados e aproveiten as vantaxes dos sistemas HPC. As paralelizacións implementadas desenvolvéronse aplicando unha distribución da carga de traballo entre elementos de procesado. Dado que os sistemas HPC adoitan ser sistemas multinodo con nodos multinúcleo, esta nova versión aproveita as posibilidades que achegan ambos cunha aproximación híbrida baseada na libraría de **paso de mensaxes** MPI e **tecnoloxías multifío**. A estratexia aplicada en ambos niveis foi a descomposición de dominio, i.e. a distribución dos datos cos que traballa o programa para que cada elemento de procesado realice os cálculos sobre un anaco diferente. Deste xeito conseguiuse, por unha parte, reducir o tempo de cómputo; e por outra, posibilitar a análise de datasets de gran tamaño que exceden as limitacións de memoria dos sistemas habituais.

As probas de rendemento realizáronse nun clúster HPC de 16 nodos, con 64GB de memoria e 16 núcleos por nodo (256 núcleos ou unidades de cómputo en total). Os resultados obtidos foron moi satisfactorios, xa que se acadaron unhas aceleracións de ata 229x para catro datasets representativos. A maiores, conseguiuse executar cada algoritmo cun dataset de 512GB de tamaño, o que non sería posible nun único nodo.

A ferramenta desenvolvida neste TFG atópase dispoñible baixo a licenza *BSD 3-Clause* no seguinte repositorio Git: https://gitlab.com/bieito/parallel-fst.

Posibles aplicacións / Posibles aplicaciones:

A ferramenta implementa unha libraría de métodos de selección de características, proceso que está estreitamente ligado á aprendizaxe máquina. En concreto, a selección de características utilízase xunto a algoritmos de aprendizaxe supervisada, como paso previo para reducir o tempo do adestramento e o risco de sobreaxuste (e.g. redes de neuronas artificiais, SVMs), ou como operación incluída no proceso de aprendizaxe (e.g. árbores de decisión).

Polo tanto, as posibles aplicacións de Parallel-FST son moi amplas, xa que se pode utilizar en calquera ámbito no que se empregue a aprendizaxe supervisada. Algúns exemplos con relevancia na actualidade poden ser: bioinformática, xenética, recuperación de información, socioloxía ou análise de redes sociais.

Ademais, actualmente está en auxe a análise de grandes cantidades de datos (*big data*), polo que os tempos de execución da selección de características poden volverse excesivos. Poñendo o foco no carácter HPC deste proxecto, as aplicacións que se poderían beneficiar do seu uso son aquelas con fortes requisitos de tempo e/ou memoria. Por exemplo, cando:

- O tempo de execución é limitado (aplicacións críticas de tempo real): é posible seleccionar un maior número de características respectando o límite de tempo.
- O tempo de execución é moi elevado (aplicacións que necesitan un número fixo de características): redúcese o tempo da execución.
- Os datos a analizar non caben na memoria dun só nodo (aplicacións de *big data*): utilízanse varios nodos para distribuír a carga e análise dos datos.

Etapas para o seu desenvolvemento futuro / Etapas para su desarrollo futuro:

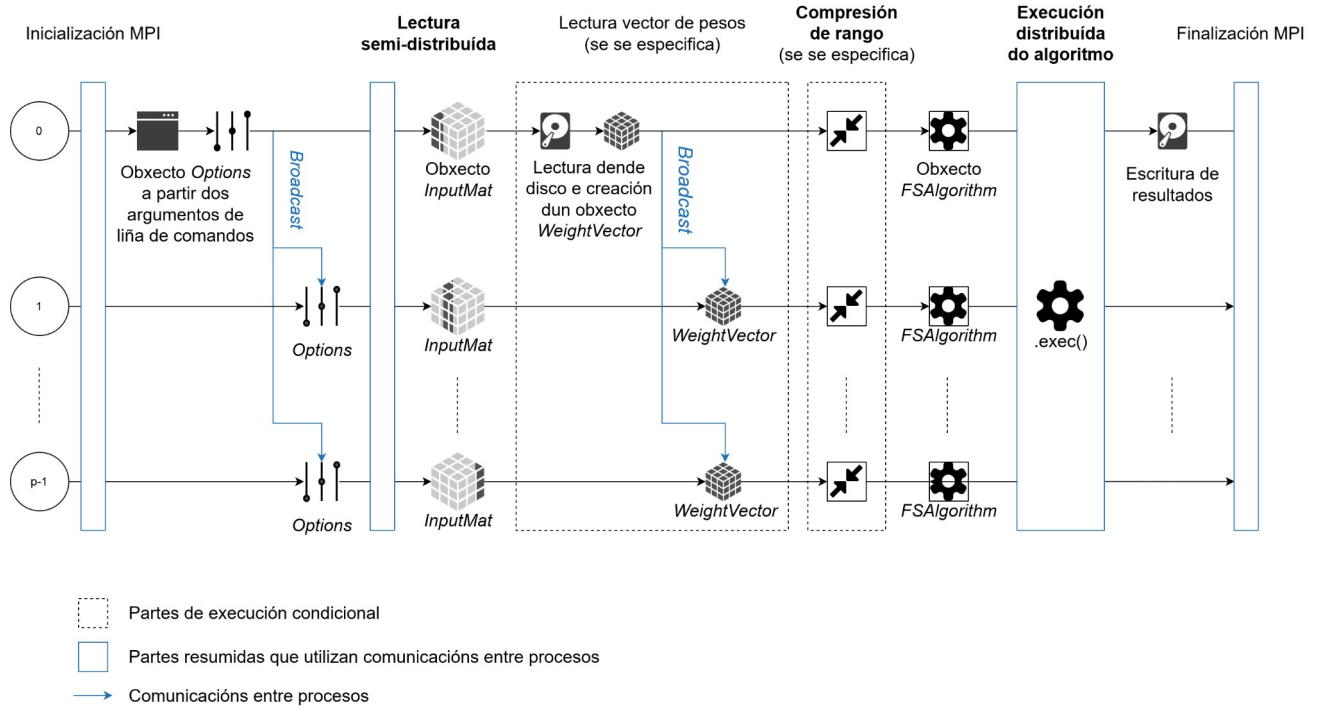
As liñas de traballo futuro polas que podería continuar Parallel-FST son as seguintes:

- Engadir novos métodos de selección de características á libraría.
- Mellorar o rendemento dos algoritmos xa presentes, posibilitando a súa execución sobre novas arquitecturas hardware, como GPUs ou FPGAs, tanto con aproximacións homoxéneas como híbridas.
- Permitir a lectura de datos en novos formatos.
- Desenvolver extensións para poder utilizar a libraría dende outras linguaxes e sistemas de análise de datos, como Python+Scikit Learn, R, Weka ou Matlab.
- Desenvolver unha interface gráfica para facilitar o uso da ferramenta a usuarios con pouca experiencia coa liña de comandos.

Por último, o proxecto deseñouse para ser facilmente ampliable e liberouse baixo unha licenza *open source*, polo que calquera persoa con coñecementos suficientes podería modificalo para adaptalo ás súas necesidades.

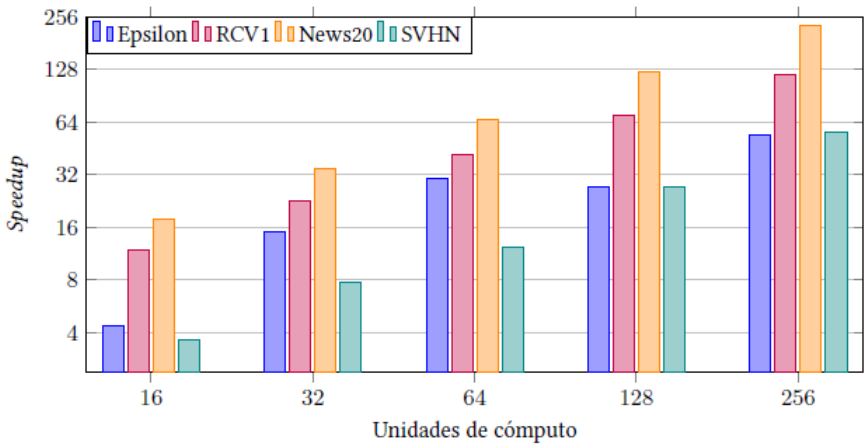
Imaxes representativas / Imágenes representativas:

Esquema xeral do funcionamento distribuído do programa:

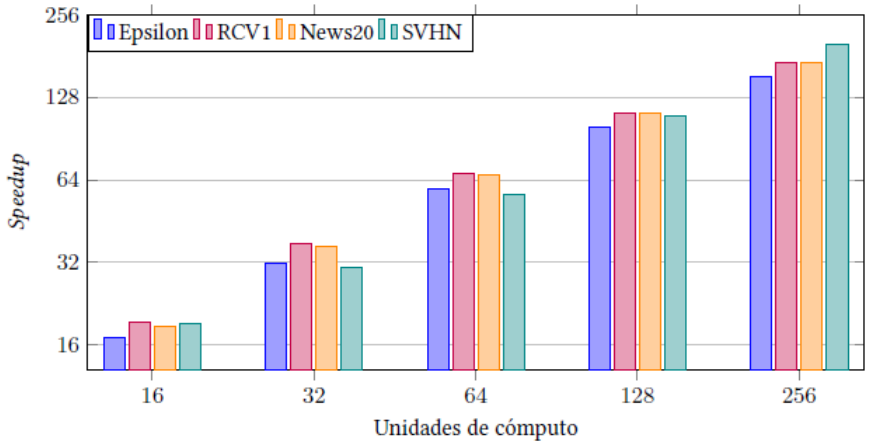


Exemplos da escalabilidade de varios algoritmos incluídos en Parallel-FST e aceleración (*speedup*) con respecto á libraría orixinal FEAST, utilizando datasets de propiedades variadas como entrada (Epsilon, RCV1, News20, SVHN):

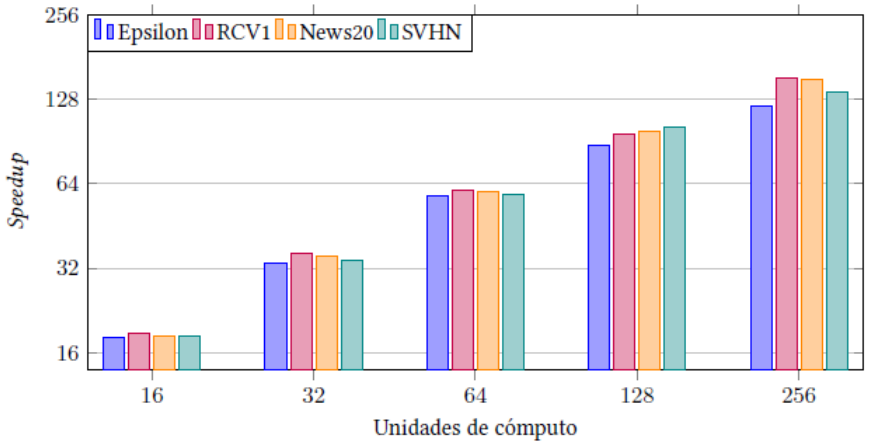
Algoritmo CondMI



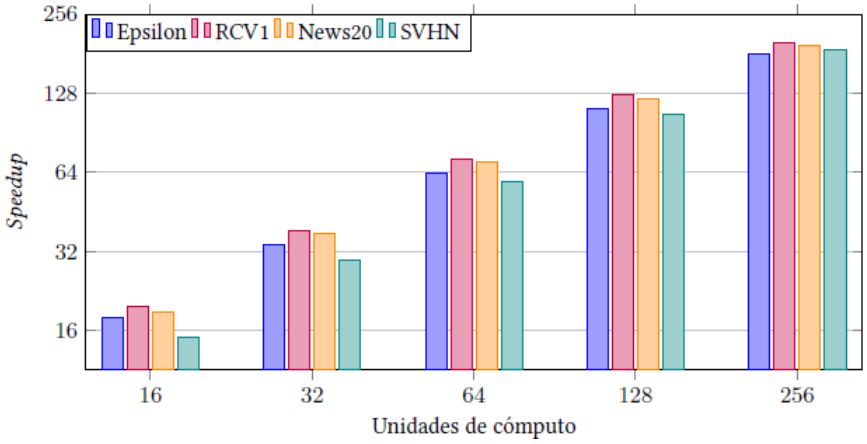
Algoritmo JMI



Algoritmo mRMR



Algoritmo DISR



X	**Autorizo a consulta por parte dos membros da comisión evaluadora da memoria do meu proxecto / Autorizo la consulta por parte de los miembros del tribunal de la memoria de mi proyecto.**