DEPARTAMENTO: 4. Informática

Título do TFM: Applying Artificial Intelligence to the simulation of pathogenic environments
Titor/a do TFM: Anália Maria Garcia Lourenço
Cotitor/a do TFM (se procede)**:** Nuno Azevedo
Titulación: Mestrado Universitario en Intelixencia Artificial (MIA)
Resumo: **Requisitos:** **- nivel avanzado de programación Java** **- buen nivel de programación Python**

DEPARTAMENTO: 4. Informática

Título do TFM: Improving the flavour of beer through AI
Titor/a do TFM: Anália Maria Garcia Lourenço
Cotitor/a do TFM (se procede)**:** Nuno Azevedo
Titulación: Mestrado Universitario en Intelixencia Artificial (MIA)
Resumo: **- nivel avanzado de programación Java** **- buen nivel de programación Python**

DEPARTAMENTO: 4. Informática

Título do TFM: Deep learning applied to molecule discovery for pathogen recognition
Titor/a do TFM: Anália Maria Garcia Lourenço
Cotitor/a do TFM (se procede)**:** Nuno Azevedo
Titulación: Mestrado Universitario en Intelixencia Artificial (MIA)
Resumo: **- buen nivel de programación Python** **- buen nivel de programación bash**

Universidade Vigo Campus de Ourense Escola Superior de Enxeñaría Informática

DEPARTAMENTO: 4. Informática

Título do TFM: Detección de sobreentrenamiento para entornos de aprendizaje automático, explorando relaciones de correlación en funciones canario

Titor/a do TFM: Manuel VILARES FERRO

Cotitor/a do TFM (se procede)**:** Víctor Manuel DARRIBA BILBAO

Titulación: Mestrado Universitario en Intelixencia Artificial (MIA)

Resumo:

La facilidad de acceso a la información y el incremento de la capacidad de memoria han contribuido decisivamente a la popularización de las técnicas de aprendizaje automático (AA) en tareas de clasificación [Leite et al. 2012] y agrupamiento [Meek et al. 2002]. En este contexto, y con objeto de limitar los costes operativos, es necesario disponer de metodologías de muestreo que nos permitan evaluar la calidad del modelo generado sin por ello tener que hacer un uso efectivo de la totalidad de la base de datos de entrenamiento y sin menoscabo en el rendimiento final del modelo generado. El problema es especialmente grave en el ámbito del procesamiento del lenguaje natural y, en concreto, en la generación de analizadores léxicos. De particular interés resultan las estrategias denominadas adaptativas [Domingo et al. 2002], en las que la talla de la muestra se asocia a un proceso iterativo sujeto a una condición de parada del aprendizaje. En particular, nuestra atención se centra en aquellas de naturaleza activa [Maytal and Provost, 2004], esto es, algoritmos que integran un conocimiento del dominio para el diseño de la secuencia de muestreo [Provost et al. 1999, John and Langley 1996].La propuesta prevé la integración en el entorno gAPOSTA, desarrollado por el grupo COLE de la Uvigo, de diferentes estrategias adaptativas y su comparativa efectiva en el ámbito antes comentado de la generación de analizadores léxicos.[Domingo et al. 2002] Carlos Domingo, Ricard Gavaldà, and Osamu Watanabe. Adaptive sampling methods for scaling up knowledge discovery algorithms. Data Mining and Knowledge Discovery, 6(2):131–152, 2002.

[Provost et al. 1999] Foster Provost, David Jensen, and Tim Oates. Efficient progressive sampling. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 23–32, San Diego, 1999.[John and Langley 1996] George John and Pat Langley. Static versus dynamic sampling for data mining. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pages 367–370, Portland, 1996.[Leite et al. 2012] Rui Leite, Pavel Brazdil, and Joaquin Vanschoren. Selecting classification algorithms with active testing. In Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition, pages 117–131, Berlin, 2012.[Maytal and Provost, 2004] Maytal Saar-Tsechansky and Foster Provost. Active sampling for class probability estimation and ranking. Machine Learning, 54(2):153–178, 2004.[Meek et al. 2002] Christopher Meek, Bo Thiesson, and David Heckerman. The learning-curve sampling method applied to model-based clustering. The Journal of Machine Learning Research, 2:397–418, March 2002.

DEPARTAMENTO: 4. Informática

Título do TFM: Detección diferenciada de inflación y sobreentrenamiento en entornos de aprendizaje automático

Titor/a do TFM: Manuel VILARES FERRO

Cotitor/a do TFM (se procede)**:** Víctor Manuel DARRIBA BILBAO

Titulación: Mestrado Universitario en Intelixencia Artificial (MIA)

Resumo:

El sobreentrenamiento en procesos de aprendizaje automático (AA) se produce cuando el modelo generado pierde capacidad de predicción al ajustarse en exceso a la base de datos de entrenamiento, mientras que la precisión sobre la base de datos de validación disminuye. En consecuencia la capacidad de predicción del modelo se reduce y la necesidad de estudiar el origen del fenómeno es una prioridad en el diseño de entornos prácticos de AA. En este contexto, las estrategias de detección de sobreentrenamiento suelen asociarse a la noción de inflación (bloat) [Silva and Costa 2009]. Esta idea ha justificado algunos de los algoritmos más populares en este ámbito [Prechelt 1997], generalmente englobables en el concepto de función canario [Evett et al. 1998, Foreman and Evett 2005], básicamente en línea con la de fitness usualmente considerada en programación genética [Koza 1992] (PG). Desfortunadamente, existen evidencias de que la presencia de inflación no tiene porque relacionarse necesariamente con el sobreentrenamiento, y vicebersa [Vanneschi et al. 2010]. Ello abre un nuevo marco de trabajo en el que detección de inflaciones y sobreentrenamiento deben diferenciarse en orden a detectar posibles interrelaciones.La propuesta prevé la integración simultánea y diferenciada en el entorno gAPOSTA, desarrollado por el grupo COLE de la Uvigo, de una medida de inflación y de otra de sobreentrenamiento, esta última basada directamente en los principios que inspiran la función fitness en PG.[Evett et al. 1998] GP-based software quality prediction. Evett, M.; Khoshgoftar, T.; der Chien, P. and Allen, E. Proc. of the Third Annual Conf. in Genetic Programming, pp. 60-65, 1998.[Foreman and Evett 2005] Preventing Overfitting in GP with Canary Functions. Foreman, N. and Evett, M. Proc. of the 7th Annual Conf. on Genetic and Evolutionary Computation, pp. 1779-1780, 2005.[Koza 1992] Genetic Programming: On the Programming of Computers by Means of Natural Selection. Koza, J.R. ISBN 0-262-11170-5, MIT Press, 1992.[Prechelt 1997] Automatic Early Stopping Using Cross Validation: Quantifying the Criteria. Prechelt, L. Neural Networks, 11:761-767, 1997.[Vanneschi et al. 2010] Measuring Bloat, Overfitting and Functional Complexity in Genetic Programming. Vanneschi, L.; Castelli, M. and Silva, S. Proc. of the 12th Annual Conf. on Genetic and Evolutionary Computation, pp. 877-884, 2010.[Silva and Costa 2009] Dynamic Limits for Bloat Control in Genetic Programming and a Review of Past and Current Bloat Theories. Silva, S. and Costa, E. Genetic Programming and Evolvable Machines, 10(2):141-179, 2009.

DEPARTAMENTO: 4. Informática

Título do TFM: Estrategias de muestreo en la generación de modelos de análisis léxico

Titor/a do TFM: Manuel VILARES FERRO

Cotitor/a do TFM (se procede)**:** Víctor Manuel DARRIBA BILBAO

Titulación: Mestrado Universitario en Intelixencia Artificial (MIA)

Resumo:

La facilidad de acceso a la información y el incremento de la capacidad de memoria han contribuido a popularizar lel uso de técnicas de aprendizaje automático (AA) en tareas de clasificación [Leite et al. 2012] y agrupamiento [Meek et al. 2002]. Con objeto de limitar los costes operativos, es necesario disponer de metodologías de muestreo que nos permitan evaluar la calidad del modelo generado sin por ello tener que hacer uso de la totalidad de la base de datos de entrenamiento y sin menoscabo en el rendimiento final del modelo generado. El problema es especialmente grave en el ámbito del procesamiento del lenguaje natural y, en concreto, en la generación de analizadores léxicos. De particular interés resultan las estrategias no-activas [Maytal and Provost, 2004], algoritmos que no necesitan integrar un conocimiento específico del dominio para el diseño de la secuencia de muestreo [Provost et al. 1999, John and Langley 1996], facilitando su aplicación. Nuestra atención se centra en aquellas de naturaleza adaptativa [Domingo et al. 2002, Vilares et al. 2020], en las que la talla de la muestra se asocia a un proceso iterativo sujeto a una condición de parada del proceso aprendizaje, y cuya eficacia depende de la información que la propia estrategia de muestreo pueda extraer de dicho proceso. El objetivo es determinar sobre las principales ramas de la familia de lenguas hindo-europeas y arquitecturas de AA en el dominio considerado, la efectividad de las estrategias de muestreo. Con el fin garantizar la fiabilidad de los resultados, ésta será medida a partir de una estimación formal del rendimiento [Vilares et al. 2017] en la generación de los modelos.[Domingo et al. 2002] Carlos Domingo, Ricard Gavaldà, and Osamu Watanabe. Adaptive sampling methods for scaling up knowledge discovery algorithms. Data Mining and Knowledge Discovery, 6(2):131–152, 2002.[Provost et al. 1999] Foster Provost, David Jensen, and Tim Oates. Efficient progressive sampling. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 23–32, San Diego, 1999.[John and Langley 1996] George John and Pat Langley. Static versus dynamic sampling for data mining. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pages 367–370, Portland, 1996.[Leite et al. 2012] Rui Leite, Pavel Brazdil, and Joaquin Vanschoren. Selecting classification algorithms with active testing. In Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition, pages 117–131, Berlin, 2012.[Maytal and Provost, 2004] Maytal Saar-Tsechansky and Foster Provost. Active sampling for class probability estimation and ranking. Machine Learning, 54(2):153–178, 2004.[Meek et al. 2002] Christopher Meek, Bo Thiesson, and David Heckerman. The learning-curve sampling method applied to model-based clustering. The Journal of Machine Learning Research, 2:397–418, March 2002.[Vilares et al. 2017] M. Vilares, V.M. Darriba and F.J. Ribadas. Modeling of learning curves with applications to POS tagging. Computer Speech & Language, 41:1-28. 2017.[Vilares et al. 2020] M. Vilares, V.M. Darriba and J. Vilares. Adaptive scheduling for adaptive sampling in pos taggers construction. Computer, Speech & Languages 60 (2020).

DEPARTAMENTO: 4. Informática

Título do TFM: Estrategias de muestreo en la generación de modelos de análisis sintáctico

Titor/a do TFM: Manuel VILARES FERRO

Cotitor/a do TFM (se procede)**:** Víctor Manuel DARRIBA BILBAO

Titulación: Mestrado Universitario en Intelixencia Artificial (MIA)

Resumo:

La facilidad de acceso a la información y el incremento de la capacidad de memoria han contribuido a popularizar lel uso de técnicas de aprendizaje automático (AA) en tareas de clasificación [Leite et al. 2012] y agrupamiento [Meek et al. 2002]. Con objeto de limitar los costes operativos, es necesario disponer de metodologías de muestreo que nos permitan evaluar la calidad del modelo generado sin por ello tener que hacer uso de la totalidad de la base de datos de entrenamiento y sin menoscabo en el rendimiento final del modelo generado. El problema es especialmente grave en el ámbito del procesamiento del lenguaje natural y, en concreto, en la generación de analizadores sintácticos. De particular interés resultan las estrategias no-activas [Maytal and Provost, 2004], algoritmos que no necesitan integrar un conocimiento específico del dominio para el diseño de la secuencia de muestreo [Provost et al. 1999, John and Langley 1996], facilitando su aplicación. Nuestra atención se centra en aquellas de naturaleza adaptativa [Domingo et al. 2002, Vilares et al. 2020], en las que la talla de la muestra se asocia a un proceso iterativo sujeto a una condición de parada del proceso aprendizaje, y cuya eficacia depende de la información que la propia estrategia de muestreo pueda extraer de dicho proceso. El objetivo es determinar sobre las principales ramas de la familia de lenguas hindo-europeas y arquitecturas de AA en el dominio considerado, la efectividad de las estrategias de muestreo. Con el fin garantizar la fiabilidad de los resultados, ésta será medida a partir de una estimación formal del rendimiento [Vilares et al. 2017] en la generación de los modelos.[Domingo et al. 2002] Carlos Domingo, Ricard Gavaldà, and Osamu Watanabe. Adaptive sampling methods for scaling up knowledge discovery algorithms. Data Mining and Knowledge Discovery, 6(2):131–152, 2002.[Provost et al. 1999] Foster Provost, David Jensen, and Tim Oates. Efficient progressive sampling. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 23–32, San Diego, 1999.[John and Langley 1996] George John and Pat Langley. Static versus dynamic sampling for data mining. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pages 367–370, Portland, 1996.[Leite et al. 2012] Rui Leite, Pavel Brazdil, and Joaquin Vanschoren. Selecting classification algorithms with active testing. In Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition, pages 117–131, Berlin, 2012.[Maytal and Provost, 2004] Maytal Saar-Tsechansky and Foster Provost. Active sampling for class probability estimation and ranking. Machine Learning, 54(2):153–178, 2004.[Meek et al. 2002] Christopher Meek, Bo Thiesson, and David Heckerman. The learning-curve sampling method applied to model-based clustering. The Journal of Machine Learning Research, 2:397–418, March 2002.[Vilares et al. 2017] M. Vilares, V.M. Darriba and F.J. Ribadas. Modeling of learning curves with applications to POS tagging. Computer Speech & Language, 41:1-28. 2017.[Vilares et al. 2020] M. Vilares, V.M. Darriba and J. Vilares. Adaptive scheduling for adaptive sampling in pos taggers construction. Computer, Speech & Languages 60 (2020).

DEPARTAMENTO: 4. Informática

Título do TFM: Comparativa para estrategias de sampling en parendizaje automático

Titor/a do TFM: Víctor Manuel Darriba Bilbao

Cotitor/a do TFM (se procede)**:** Manuel Vilares Ferro

Titulación: Mestrado Universitario en Intelixencia Artificial (MIA)

Resumo:

La facilidad de acceso a la información y el incremento de la capacidad de memoria han contribuido decisivamente a la popularización de las técnicas de aprendizaje automático (AA) en tareas de clasificación [Leite et al. 2012] y agrupamiento [Meek et al. 2002]. En este contexto, y con objeto de limitar los costes operativos, es necesario disponer de metodologías de muestreo que nos permitan evaluar la calidad del modelo generado sin por ello tener que hacer un uso efectivo de la totalidad de la base de datos de entrenamiento y sin menoscabo en el rendimiento final del modelo generado. El problema es especialmente grave en el ámbito del procesamiento del lenguaje natural y, en concreto, en la generación de analizadores léxicos. De particular interés resultan las estrategias denominadas adaptativas [Domingo et al. 2002], en las que la talla de la muestra se asocia a un proceso iterativo sujeto a una condición de parada del aprendizaje. En particular, nuestra atención se centra en aquellas de naturaleza activa [Maytal and Provost, 2004], esto es, algoritmos que integran un conocimiento del dominio para el diseño de la secuencia de muestreo [Provost et al. 1999, John and Langley 1996].

La propuesta prevé la integración en el entorno gAPOSTA, desarrollado por el grupo COLE de la Uvigo, de diferentes estrategias adaptativas y su comparativa efectiva en el ámbito antes comentado de la generación de analizadores léxicos.

[Domingo et al. 2002] Carlos Domingo, Ricard Gavaldà, and Osamu Watanabe. Adaptive sampling methods for scaling up knowledge discovery algorithms. Data Mining and Knowledge Discovery, 6(2):131–152, 2002.[Provost et al. 1999] Foster Provost, David Jensen, and Tim Oates. Efficient progressive sampling. In Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 23–32, San Diego, 1999.[John and Langley 1996] George John and Pat Langley. Static versus dynamic sampling for data mining. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pages 367–370, Portland, 1996.[Leite et al. 2012] Rui Leite, Pavel Brazdil, and Joaquin Vanschoren. Selecting classification algorithms with active testing. In Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition, pages 117–131, Berlin, 2012.[Maytal and Provost, 2004] Maytal Saar-Tsechansky and Foster Provost. Active sampling for class probability estimation and ranking. Machine Learning, 54(2):153–178, 2004.[Meek et al. 2002] Christopher Meek, Bo Thiesson, and David Heckerman. The learning-curve sampling method applied to model-based clustering. The Journal of Machine Learning Research, 2:397–418, March 2002.

DEPARTAMENTO: 4. Informática

Título do TFM: Integración de una condición de parada con umbral de error absoluto en aprendizaje automático.

Titor/a do TFM: Víctor Manuel Darriba Bilbao

Cotitor/a do TFM (se procede)**:** Manuel Vilares Ferro

Titulación: Mestrado Universitario en Intelixencia Artificial (MIA)

Resumo:

Una cuestión fundamental en el tratamiento de procesos de aprendizaje automático es la determinación de un algortimo de parada que garantice que los recursos invertidos sean estrictamente los necesarios para asegurar la calidad predeterminada por el usuario. Habitualmente enfocados como una estimación del compromiso coste/beneficio entendido como utilidad máxima esperable [Meek et al. 2002, Sheng and Ling 2007] y expresada en términos de teoría de decisión [Howard 1966], en la práctica este acercamiento no ofrece garantías [Last 2009] y suelen asumirse costes fijos [Kapoor and Greiner 2005].Alternativamente, otros autores [Vilares et al. 2016] proponen condiciones de parada basadas en el concepto de convergencia funcional uniforme, si bien hasta el momento no había sido posible ofrecer un umbral de error absoluto en las aproximaciones. Al respecto, el grupo COLE de la UVigo ha desarrollado un nuevo algoritmo de aproximación que garantiza este tipo de capacidad y cuya integración y validación en el entorno GAPOSTA constituyen la propuesta.

[Howard 1966] Decision analysis: Applied decision theory. Howard, R.A. Proc. of the 4th Int. Conf. On Operational Research, pp. 55-71, 1966.[Kapoor and Greiner 2005] Learning and Classifying Under Hard Budgets, Machine Learning: ECML 2005, 2005.[Last 2009] Improving Data Mining Utility with Projective Sampling. Last, M. Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 487-496, 2009.[Meek et al. 2002] The Learning-curve Sampling Method Applied to Model-based Clustering. Meek, C,; Thiesson, B. and Heckerman, D. Journal of Machine Learning Research, 2:397-418, 2002.[Sheng and Ling 2007] Partial Example Acquisition in Cost-sensitive Learning. Sheng, V.S. and Ling, C.X. Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 638-646, 2007.[Vilares et al. 2017] Modeling of learning curves with applications to POS tagging. Vilares, M.; Darriba, V.M. and Ribadas, F.J. Computer Speech & Language, 41:1-28, 2017.