

Diseño de algoritmos basados en independencia para el aprendizaje de modelos probabilísticos gráficos de mejor calidad

Convocatoria

Tipo	Proyectos de Investigación Científica y Tecnológica (2008)
Categoría	Programa de Recursos Humanos}
Tipo de Proyecto	A

Datos generales del proyecto

Áreas Temáticas	Principal : Tecnología Informática de las Comunicaciones y Electrónica Secundarias : Tecnología del Medio Ambiente //
Prioridad Regional	No Aplica
Duración	3 años
Ubicación	Argentina, Mendoza
Idiomas	Español
Investigador Responsable	Nombre y Apellido Facundo Bromberg
Fecha de alta	06-03-2009

Instituciones del Proyecto

Institucion Beneficiaria	Razón Social Universidad Tecnológica Nacional	Teléfono (011) 5371-5600 / 5608 / 5700 / 5702 / 5703 / 5704	Correo Electrónico sec-cyt@utn.edu.ar
Representante Legal	Representante Legal Brotto Héctor Carlos	Cargo Rector	
Dependencia	Razón Social Facultad Regional Mendoza	Teléfono 0261-4239239/4239119	Correo Electrónico decano@frm.utn.edu.ar
Unidad Ejecutora	Razón Social Universidad Tecnológica Nacional - Facultad Regional Mendoza	Teléfono 54-0261-5244500 54-0261-5244531	Correo Electrónico secyt@frm.utn.edu.ar

Conformación de Grupo

Grupo Responsable	Nombre y Apellido Facundo Bromberg	Rol Investigador Responsable
Grupo Colaborador	Nombre y Apellido Dimitris Margaritis Federico Schluter Martin Spertino Diego Sebastian Perez	Rol Investigador Becario Post Grado Estudiante Estudiante

Resumen del Proyecto

Palabras clave

Aprendizaje de máquinas, modelos probabilísticos gráficos, algoritmos basados en independencias, aprendizaje de estructuras

Resumen

Nuestra investigación se enmarca en el problema del aprendizaje de representaciones gráficas de modelos probabilísticos (e.g., redes Bayesianas, redes Markovianas). Estos modelos presentan una importante aumento en la eficiencia del proceso de inferencia estadística y el de aprendizaje automatizado de dichos modelos a partir de conjuntos de datos (Geman and Geman (1984); Besag et. al. 1991 and Anguelov et. al. 2005; Friedman et. al. 2000, otros).

Los modelos gráficos (y el aprendizaje de estos modelos) consisten de dos partes: un grafo (que representa una estructura de independencias) y un conjunto de parámetros numéricos. El aprendizaje de la estructura de los modelos Markovianos ha presentado importantes dificultades computacionales debido a que requiere estimar los parámetros, un problema NP-completo para modelos Markovianos (Barahona 1982). Esta dificultad ha forzado la necesidad de recurrir a expertos para la estimación de las estructuras (e.g., Kindermann and Snell 1980, Geman and Geman 1984, Besag et. al. 1991).

En (Bromberg et al. 2006, 2007, 2007a, y Gandhi et al. 2008) se presentan una serie de algoritmos eficientes para el aprendizaje de estructuras Markovianas, que no requieren el aprendizaje de parámetros. Estos utilizan un enfoque conocido como aprendizaje de estructuras basado en independencias (Spirites et. al. 2000). Este enfoque consiste en la realización de una serie de tests estadísticos de independencias eliminando, en cada iteración, aquellas estructuras inconsistentes con los resultados de estos tests. Desafortunadamente, la enorme ventaja computacional obtenida por estos algoritmos se ve opacada por la sensibilidad en la calidad de los tests estadísticos de independencia - y por lo tanto de las estructuras producidas por estos algoritmos cuando los datos disponibles son insuficientes (Agesti 2002).

Diseño de algoritmos basados en independencia para el aprendizaje de modelos probabilísticos gráficos de mejor calidad

El presente proyecto propone mejorar la calidad de estos algoritmos por medio de dos métodos alternativos: La mejora de la calidad de los test estadísticos de independencia, y el diseño de algoritmos basados en independencias más robustos a errores en los test.

El primer método consiste en generalizar el Argumentative Independence Test (AIT) de (Bromberg and Margaritis 2009). La principal hipótesis del AIT es que es posible mejorar la calidad de tests estadísticos de independencia por medio de la ejecución de tests relacionados y explotar ciertas restricciones que existen entre estos tests para corregir los resultados del test de interés. Generalizaremos el AIT por medio del estudio de metodologías alternativas a la de argumentación para explotar las restricciones. El segundo método consiste en atacar directamente el efecto cascada? (Spirtes et al 2001), que ocurre en los algoritmos basados en independencias. Estos algoritmos restringen los tests a ser realizados según el valor de los tests realizados hasta el momento. De esta manera, errores en los tests producen un efecto cascada desviando al algoritmo de la realización de tests más relevantes e informativos.

Impacto del Proyecto

Subarea identificada

N/A

Impacto sobre el sector socio-económico y/o sector productivo

El proyecto se enmarca en el área del aprendizaje de máquinas, la cual estudia algoritmos para el modelado automatizado a partir de datos. La generación de modelos en forma automatizada acerca herramientas de análisis avanzado de datos a personal no especializado en modelado de datos, o sin la posibilidad de financiar la caracterización manual de estos modelos. En el ámbito privado, gerentes y jefes de departamentos, y en el ámbito público miembros del poder ejecutivo (ministros, gobernantes, secretarios, personal técnico), se han visto beneficiados. Esto ha influenciado en gran medida en la toma de decisiones, detección temprana de fraudes como así de stock, predicción de la lealtad del cliente, y muchas otras predicciones financieras, productivas, logísticas, o de marketing.

El modelado automatizado de modelos probabilísticos no escapa a esta influencia, permitiendo el modelado automatizado de toda información relacionada internamente en forma probabilística. Como ejemplo tenemos aplicaciones en estudios demográficos, agropecuarios, de redes sociales, y estadística espacial en general.

Otra importante contribución es en el ámbito técnico de desarrollo, el cual detallamos abajo en el apartado de "impacto sobre áreas disciplinares o campos de aplicación".

Impacto sobre las capacidades institucionales

El investigador principal del presente proyecto, junto con otros dos investigadores doctorados han sido incorporados recientemente al departamento de Sistemas de Información de la Universidad Tecnológica Nacional. Esta incorporación de doctores consiste en la primera del departamento, resultando en un importante impacto en variadas áreas de interés del departamento como ser la formación de recursos humanos a nivel doctoral y producción de conocimiento científico original. El presente proyecto consiste por lo tanto en una invaluable oportunidad para equipar al departamento y a la facultad regional Mendoza de la UTN con equipamiento propio de investigaciones de este nivel, concretamente sistemas avanzados de computo, inexistentes actualmente en la facultad. Además, proveerá fondos para viáticos a conferencias nacionales e internacionales, también inexistentes actualmente, que permitirán al becario, al investigador principal y/o a alumnos de grado avanzados que colaboren con el proyecto a participar de estas actividades científicas enormemente enriquecedoras.

Impacto sobre las áreas disciplinares o campos de aplicación

El aprendizaje de los modelos probabilísticos gráficos, en particular las redes Markovianas, ha sido utilizado históricamente en ciencias básicas como la Física (e.g., modelo de Ising en magnetismo), sociología, agricultura, estudios demográficos, donde han sido llamados Markov Random Fields. Sin embargo, dificultades computacionales infranqueables han impedido su uso en sistemas altamente inhomogeneos (e.g., redes eléctricas, redes sociales, por citar algunos ejemplos) donde además de parámetros numéricos, se requiere el aprendizaje automatizado de la estructura de independencias (i.e., grafo, red) del modelo. En años recientes miembros del grupo de trabajo han desarrollado los primeros algoritmos computacionalmente eficientes para el aprendizaje de la estructura de estas redes. Estos resultados se ven opacados, sin embargo, por limitaciones en la calidad de los modelos generados. El presente proyecto se focaliza entonces en franquear esta dificultad. Algoritmos eficientes y de alta calidad tendrán entonces un enorme impacto en la variada gama de disciplinas que utilizan el modelado probabilístico.

Además de mejorar la calidad de los modelos producidos, resultados preliminares han resultado (Bromberg 2009) en la mejora en la predicción de tests estadísticos de independencia. El uso de estos tests es generalizado en una enorme gama de aplicaciones.

Datos Complementarios

Investigadores recursados como evaluadores

No se presentan recusaciones.

Salvaguardia ética y ambiental

No es necesaria.

Otras observaciones

Ninguna



Diseño de algoritmos basados en independencia para el aprendizaje de modelos probabilísticos gráficos de mejor calidad

Presupuesto

Año	Rubro	Subsidio	Contraparte	Total
1	Bibliografía	2500.0	0.0	2500.0
1	Equipamiento	8500.0	2200.0	10700.0
1	Insumos	1200.0	0.0	1200.0
1	Personal (Salarios)	0.0	44160.0	44160.0
1	Viajes y viáticos	10230.0	0.0	10230.0
1	GASTOS DE ADMINISTRACION DEL SUBSIDIO	897.2	0.0	897.2
1	TOTAL ANUAL	23327.2	46360.0	69687.2
2	Bibliografía	2500.0	0.0	2500.0
2	Equipamiento	2600.0	0.0	2600.0
2	Insumos	1200.0	0.0	1200.0
2	Personal (Salarios)	0.0	44160.0	44160.0
2	Publicaciones de resultado del proyecto	1000.0	0.0	1000.0
2	Viajes y viáticos	11960.0	0.0	11960.0
2	GASTOS DE ADMINISTRACION DEL SUBSIDIO	770.4	0.0	770.4
2	TOTAL ANUAL	20030.41	44160.0	64190.4
3	Bibliografía	2500.0	0.0	2500.0
3	Insumos	1200.0	0.0	1200.0
3	Personal (Salarios)	0.0	44160.0	44160.0
3	Publicaciones de resultado del proyecto	1000.0	0.0	1000.0
3	Viajes y viáticos	20460.0	0.0	20460.0
3	GASTOS DE ADMINISTRACION DEL SUBSIDIO	1006.4	0.0	1006.4
3	TOTAL ANUAL	26166.4	44160.0	70326.4
	TOTAL GENERAL	69524.01	134680	204204

Detalle del Presupuesto

Año	Rubro	Detalle	Sub.	Cont.	Total
1	Bibliografía	Libros y suscripciones a revistas científicas	2500.0	0.0	2500.0
1	Equipamiento	PC Escritorio+Computo becario	2600.0	0.0	2600.0
1	Equipamiento	PC para computo y almacenamiento de datos	2600.0	0.0	2600.0
1	Equipamiento	UPS	1500.0	0.0	1500.0
1	Equipamiento	Impresora red	1800.0	0.0	1800.0
1	Equipamiento	PC Escritorio investigador principal	0.0	2200.0	2200.0
1	Insumos	Insumos varios para oficina	1200.0	0.0	1200.0
1	Personal (Salarios)	Salario investigador principal	0.0	44160.0	44160.0
1	Viajes y viáticos	Pasaje y viáticos Congresos Internacionales	8500.0	0.0	8500.0
1	Viajes y viáticos	Pasaje y viáticos Congresos Nacionales	1730.0	0.0	1730.0
1	GASTOS DE ADMINISTRACION DEL SUBSIDIO		897.2	0.0	897.2
1	TOTAL ANUAL		23327.2	46360.0	69687.2
2	Bibliografía	Libros y suscripciones a revistas científicas	2500.0	0.0	2500.0
2	Equipamiento	PC escritorio computo para colaboradores	2600.0	0.0	2600.0
2	Insumos	Insumos varios para oficina	1200.0	0.0	1200.0
2	Personal (Salarios)	Salario investigador principal	0.0	44160.0	44160.0
2	Publicaciones de resultado del proyecto	Publicación revista indexada	1000.0	0.0	1000.0
2	Viajes y viáticos	Pasaje y viáticos Congresos Internacionales	8500.0	0.0	8500.0
2	Viajes y viáticos	Pasaje y viáticos Congresos Nacionales	3460.0	0.0	3460.0
2	GASTOS DE ADMINISTRACION DEL SUBSIDIO		770.4	0.0	770.4
2	TOTAL ANUAL		20030.41	44160.0	64190.4
3	Bibliografía	Libros y suscripciones a revistas científicas	2500.0	0.0	2500.0
3	Insumos	Insumos varios para oficina	1200.0	0.0	1200.0
3	Personal (Salarios)	Salario investigador principal	0.0	44160.0	44160.0
3	Publicaciones de resultado del proyecto	Publicación revista indexada	1000.0	0.0	1000.0
3	Viajes y viáticos	Pasaje y viáticos Congresos Internacionales	17000.0	0.0	17000.0
3	Viajes y viáticos	Pasaje y viáticos Congresos Nacionales	3460.0	0.0	3460.0
3	GASTOS DE ADMINISTRACION DEL SUBSIDIO		1006.4	0.0	1006.4
3	TOTAL ANUAL		26166.4	44160.0	70326.4

Justificación

Justificación del subsidio solicitado

Solicitamos financiamiento para bibliografía, tanto en forma de publicaciones periódicas como libros de texto. Ciertas publicaciones periódicas no son proveídas por la biblioteca de la SECYT de la Nación a la UTN-FRM, haciendo necesario una suscripción directa. Además, el rápido avance de nuestra disciplina requiere la renovación permanente de los libros de referencia en el área.

Solicitamos también dinero para financiar la publicación de los resultados generados en revistas indexadas. Muchas de estas revistas cobran un canon por página, o por páginas excedentes.

Siendo el nuestro un grupo incipiente, debemos además equiparnos con herramientas básicas de escritorio, como ser una impresora con conectividad en red y algunas máquinas de escritorio para el uso de nuevos integrantes. Además, la complejidad de muchos de los algoritmos propuestos imposibilitan su evaluación en términos teóricos, requiriendo una evaluación empírica sistemática en forma de experimentos computacionales. Para ello solicitamos la compra de procesadores multi-núcleo para las máquinas de escritorio, permitiendo un uso eficiente de estos recursos como máquina de escritorio y de computo.

Por último, hemos solicitado dinero para financiar viáticos a encuentros científicos nacionales e internacionales. Este



Diseño de algoritmos basados en independencia para el aprendizaje de modelos probabilísticos gráficos de mejor calidad

dinero se utilizará para financiar el viaje al becario doctorando junto con algún otro miembro del grupo (en lo posible) para que exponga los resultados científicos producidos. Tanto en el ámbito local como internacional, esto permitirá a nuestro grupo, y en especial al becario, conocer a la comunidad local de investigadores en el área, sus trabajos e intereses, fortaleciendo así los lazos entre nuestro grupo y la comunidad y permitiendo al becario ampliar sus perspectivas y visión de la disciplina.

Justificación de la beca solicitada

No se ha solicitado beca en este subsidio. El Ing. Federico Schlüter ha solicitado una beca para docentes de la UTN, Convocatoria 2008, de formación de doctores en áreas tecnológicas prioritarias, MINCyT, FONCyT, IP-PRH 2007. Esta beca se le ha sido adjudicada a Federico en la resolución 1590 del 11 de Diciembre del 2008 del Consejo superior de la UTN (ver resolución en documentación adjunta). El programa de investigación propuesto por el becario fue realizado en concordancia con el propuesto en el presente proyecto, y el director propuesto fue el Dr. Bromberg, investigador principal del presente proyecto.

Recursos de la Institución Beneficiaria

Infraestructura y Equipamiento existente en la Unidad Ejecutora

La unidad ejecutora cuenta con un nodo informático que provee servicios web, email y conectividad en general. Además, cuenta con 10 procesadores dedicados, configurados en formato de cluster, para realizar tareas de cómputo. Contamos además cuenta con un espacio físico para el investigador principal y el becario, mobiliario y PCs de escritorio.

La unidad ejecutora cuenta además con una biblioteca básica con bibliografía de apoyo (aunque no de interés directo del proyecto). El investigador principal cuenta con la bibliografía mínima necesaria para estudio y referencia. Por último, la unidad ejecutora cuenta con acceso a la biblioteca on-line de la secyt, a través de la cual es posible consultar bibliografía, revistas y publicaciones periódicas.

Recursos de Otras Instituciones

Otros Recursos afectados para el desarrollo del Proyecto

La unidad ejecutora cuenta con la Fundación UTN, quien posee fondos disponibles para complementar viáticos y equipamiento.

Diseño de algoritmos basados en independencia para el aprendizaje de modelos probabilísticos gráficos de mejor calidad

Investigadores Participantes

Datos Personales	
CUIT/CUIL:	20238491887
Apellidos:	Bromberg
Nombres:	Facundo
Nacionalidad:	AR - Argentina
Fecha de Nacimiento:	28-03-1974
Tipo de documento:	Documento Nacional de Identidad
Numero de documento:	23849188
Estado civil:	Casado/a
Contacto Laboral	Rodriguez 273, Capital (5500) Mendoza, Argentina Tel: 52-261-5244545 Email: secyt@frm.utn.edu.ar
Contacto Particular	Tel: 54-261-5244554 Email: fbromberg@frm.utn.edu.ar

Empleadores				
Fecha Inicio	Fecha Fin	Organización	Cargo	Dedicación (horas)
01-02-2008	00:00:00	Universidad Tecnológica Nacional, Facultad Regional Mendoza	Profesor Asociado	45
14-08-2001	15-12-2007	Iowa State University	Teaching Assistant	20
01-02-1999	31-01-2000	Instituto de Investigaciones Cardiológicas.	Beca FONCYT	40
01-08-1994	18-12-1998	Comisión Nacional de Energía Atómica	Becario	20
00:00:00	00:00:00			

Datos Curriculares
<p>PRODUCCIÓN CIENTÍFICA</p> <p>-----</p> <p>PUBLICACIONES CON REFERATO:</p> <ul style="list-style-type: none"> * Bromberg F. and Margaritis D. "Improving the Reliability of Causal Discovery from Small Data Sets using Argumentation". 10(Feb):310--340, 2009, Journal of Machine Learning Research. * Bromberg F. and Maragaritis D. "Efficient Markov Network Discovery Using Particle Filters". In press, accepted for publication in Computational Intelligence, (c) 2009 Blackwell publishing. Consultas al editor A. Ghorbani and E. Milios, Editors-in-Chief, Computational Intelligence, <ci@cs.ualberta.ca> * Bromberg F., Margaritis D. and Honavar V. "Efficient Markov Network Structure Discovery using Independence Tests". Enviado al Journal of Artificial Intelligence Research, Dic 2008. * Gandhi P., Bromberg F. and Margaritis D.. "Dynamic inference-based learning Markov Network Structure". Proc. of SIAM Data Mining 2008. * Bromberg F. and Margaritis D. "Efficient and Robust Independence-based Markov Network Structure Discovery". Proc of the Inter Joint Conference of Artificial Intelligence (IJCAI-07), Hyderabad, India. pp 2431-2436. * Bromberg F., Margaritis D. and Honavar V. "Efficient Markov Network Structure Discovery from Independence Tests". Proceedings of SIAM Data Mining 2006, Bethesda, Maryland. pp.141-152. <p>THESIS:</p> <ul style="list-style-type: none"> * Bromberg F., Thesis de doctorado: "Markov network structure discovery using independence tests", 2007, Iowa State University. * Bromberg F. Thesis de licenciatura: "Transiciones de Fase Inducidas por Ruido en Osciladores Neuronales Globalmente Acoplados." Grupo de Física Estadística. Centro Atómico Bariloche, S.C. de Bariloche, Argentina. Presentación de poster. <p>PUBLICACIONES SIN REFERATO:</p> <ul style="list-style-type: none"> * Bromberg F. and Honavar, V.. "Multi Agent Learning from Data". Inter Symposium on Modern Computing. Iowa State U. 2003. Poster presentation. * Bromberg F. y Ponce-Hornos J.E. "Modelado Computacional y Simulación de Respuestas Miotérmicas del Musculo Cardíaco". 28va Reunión Anual de la Sociedad Biofísica Argentina. Nov 1999. Presentación de poster. * Bromberg F. y Mato G.. "Transiciones de Fase Inducidas por Ruido en Osciladores Neuronales Globalmente Acoplados.". 83va Reunión Anual de la Asoc de Física Argen- tina. La Plata. Buenos Aires. Argentina. Sept 1998. Presentación de poster.

Diseño de algoritmos basados en independencia para el aprendizaje de modelos probabilísticos gráficos de mejor calidad

PRODUCCIÓN TECNOLÓGICA

Desarrollo, diseño y comercialización de "emusic". Sistema distribuido (LAN) de almacenamiento, administración, y navegación de una biblioteca de archivos de audio. Comercializado en S. C. de Bariloche en los años 1998-2000.

PREMIOS y PARTICIPACIÓN EN COMITES CIENTÍFICOS

Beca de CNEA, Agosto 1994-Dic 1998

Otorgada para el cursado de la carrera de Lic. en Física en el Inst. Balseiro.

Beca de FONCYT, Febrero 1999-Enero 2000

Otorgada durante mi estadía en el Instituto de Investigaciones Cardiológicas, Laboratorio de metabolismo y energética cardíaca, Fac. de Medicina, UBA.

Premio a la excelencia en educación (Teaching Excellence Award), Iowa State University, 2005.

Premio de la Upsilon Pi Epsilon Honorary Society, 2005.

Membrecía en Sociedades Científicas: AAAI (2003) y ACM (2004).

FORMACIÓN DE RECURSOS HUMANOS

* A partir de Marzo 2009, Federico Schlüter trabajará en su doctorado bajo mi dirección con beca de UTN-FONCYT de convocatoria bajo los PRH.

* Dictado de curso postgrado "Inteligencia Artificial Avanzada", Setiembre-Diciembre 2008, Doctorado de Ingeniería, U.N.Cuyo, Mendoza, Argentina.

* "KHIPU: Plataforma de autenticación para acceso a recursos basada en inferencia semántica y confiable", dirección de proyecto final de carrera Ing. en Sistemas de Información, UTN-FRM de Alejandro Edera, Mauricio Pasquier y Marcelo Garsiolo, a defender en Agosto 2009.

OTRAS ACTIVIDADES TÉCNICO-CIENTÍFICAS

Asistencia a congresos:

* Int. Joint Conf of Artificial Intelligence (IJCAI), 2007, Hyderabad, India.

* SIAM Data Mining, Abril 2006, Bethesda, Maryland.

* America's School and Intl. Joint Conference on Autonomous Agents and Multiagent Systems,

Julio 2004, Columbia U., Nueva York. EEUU

* Int. Joint Conf of Artificial Intelligence (IJCAI), 2003, Acapulco, México.

Reviewer (jurado), 27 de Noviembre, 2005

International Conf in Data Mining (ICDM '05) IEEE Workshop on Knowledge Acquisition from Semantically Heterogeneous Data. Houston, Texas, EE.UU.

Formación Académica

Mayor Título Logrado : Doctorado

Licenciado en Física, Diciembre 1998, Instituto Balseiro – Universidad Nacional de Cuyo S. C. Bariloche. Argentina

Director de tesis: Dr. Germán Mato matog@cab.cnea.gov.ar

Area de investigación: Redes Neuronales. Física Estadística.

Doctor en Ciencias de la Computación, Noviembre 2007, Iowa State University, Ames, Iowa. EEUU

Director de tesis: Dr. Dimitris Margaritis dmarg@cs.iastate.edu

Area de investigación: Inteligencia Artificial – Aprendizaje de Máquinas.

Actividades y/o cargos destacados en I+D

* Profesor - Investigador, Febrero 2008 – , Laboratorio y cátedra de Inteligencia Artificial, UTN Facultad Regional Mendoza, Argentina.

* Investigador - doctorando, Agosto 2001 – Diciembre 2007, Laboratorio de Inteligencia Artificial, Ames, Iowa, EEUU

Desarrollo, diseño de algoritmos en el área de aprendizaje de máquinas

Referencia: Dr. Dimitris Margaritis <dmarg@cs.iastate.edu>.

* Investigador – becario (FONCYT), Febrero 1999 – Enero 2000, Laboratorio de Metabolismo y Energética Cardíaca (LMEC). Inst de Investigaciones Cardiológicas (ININCA). Facultad de Medicina. U. de Bs As. Modelado e implementación de simulaciones de respuestas miotérmicas del músculo cardíaco.

Referencia: Dra Patricia Bonazzola <patri@biofis.odon.uba.ar>, directora LMEC.

* Becario, Agosto 1997 – Diciembre 1998, Grupo de Física Estadística, Instituto Balseiro, U.N.Cuyo, S.C. de Bariloche.

Modelado numérico y teórico de la sincronización de osciladores neuronales. Referencia: Dr Germán Mato <matog@cab.cnea.edu.ar>.

Área de actuación en I+D

Inteligencia Artificial, Aprendizaje de Máquinas, Web Semántica.



Diseño de algoritmos basados en independencia para el aprendizaje de modelos probabilísticos gráficos de mejor calidad