Actas

Sessão de Estudantes

(EPIA'03)

Sessão Aberta a Estudantes sobre Desenvolvimentos em Inteligência Artificial

Beja, Portugal

December 4-7, 2003

A sessão é aberta a estudantes de Mestrado ou Doutoramento.

Visa promover a divulgação, o debate, a discussão, troca de ideias e contactos sobre o trabalho em IA desenvolvido em Portugal.

Os trabalhos a apresentar podem ser teses de mestrado, teses de doutoramento, ou trabalhos apresentados em conferências internacionais (com sistema de avaliação).

EPIA'03 - 11th Portuguese Conference on Artificial Intelligence

Edited by: João Gama

INDICE

Inferência de Gramáticas Regulares usando Algoritmos de Fusão de Estados com Procura

Autor: Miguel Mourão Fialho Bugalho

Orientador: Arlindo Manuel Limede de Oliveira

Instituição: INESC-ID

Palavras-chave: Temporal Data Mining, Inferência gramatical, Inferência de DFAs, Fusão de estados, EDSM, Procura em feixe, Abbadingo One

Resumo

A área de Temporal Data Mining tem merecido bastante atenção por parte de investigadores, como se pode verificar pelos artigos publicados nas conferências da especialidade. Tal como em todos os problemas de Data Mining, nos problemas de Temporal Data Mining pretende-se projectar e analisar algoritmos para a extracção automática de informação a partir de grandes quantidades de dados. A área de Temporal Data Mining foca-se nos problemas onde os dados possuem informação temporal relevante.

A inferência de gramáticas é uma área específica da área de Temporal Data Mining, em que os conceitos que se pretendem descobrir a partir dos dados podem ser representados como gramáticas. Neste trabalho são utilizados autómatos finitos deterministas para representar as gramáticas regulares a inferir. O objectivo dos algoritmos descritos neste trabalho é descobrir o autómato mais pequeno que consiga representar um conjunto de sequências rotuladas.

A inferência de autómatos finitos deterministas mínimos a partir de um conjunto de sequências rotuladas é um problema NP-difícil. Neste trabalho são analisados algoritmos para a resolução deste tipo de problemas, sendo dada especial atenção aos algoritmos de fusão de estados pelo facto de constituírem actualmente o estado da arte. São ainda analisadas técnicas de procura em feixe que permitem melhorar o resultado dos algoritmos de fusão de estados. Em particular, é avaliada uma nova técnica de procura em feixe para algoritmos de fusão de estados que permite melhores resultados que as técnicas actualmente utilizadas.

Apesar das gramáticas regulares estarem entre as mais simples, estas conseguem mesmo assim modelar alguns conceitos interessantes. As técnicas de fusão de estados aqui estudadas podem também ser utilizadas no problema mais genérico de inferência de DFAs estocásticos. As abordagens analisadas neste trabalho podem assim ser utilizadas não só em problemas representados por gramáticas regulares mas também em problemas probabilísticos como os encontrados em áreas como a bioinformática.

Extracção de Elementos Relevantes em Texto/Páginas da World Wide Web

João Paulo da Costa Cordeiro
Orientador: Pavel Brazdil
FCUP - LIACC

Palavras chave: {Information Extraction}

Resumo

Nos últimos anos, o aumento da quantidade de informação digital disponível é um facto irrefutável, nomeadamente a que se encontra na *World Wide Web*, sob a forma de documentos de texto. Nunca na história da humanidade houve um tão elevado volume de informação acessível. Apesar dos aspectos positivos que isto representa e do potencial que permite, existe uma nova problemática que surge e consiste na necessidade de ferramentas eficazes na pesquisa e extracção de informação. O trabalho desenvolvido, no âmbito desta dissertação, enquadra-se neste contexto.

O principal objectivo deste trabalho consiste em aplicar um conjunto de técnicas da *Inteligência Artificial* (IA), nomeadamente da área da Extracção de Informação, para a criação de um sistema capaz de identificar e extrair certos elementos de texto, considerados relevantes, em documentos. Embora o alvo tenha sido o de implementar uma metodologia genérica, adaptável a qualquer domínio, fixamos a nossa atenção num domínio concreto, de modo demonstrar a essa mesma metodologia. Esse domínio consistiu nos anúncios (em Português) relativos à venda de habitações.

O sistema desenvolvido utiliza *Aprendizagem Supervisionada*, para que possa ser treinado, com uma colecção de documentos anotados e assim "aprenda" a reconhecer/extrair os elementos relevantes no texto. Uma das preocupações foi que o processo de treino produzisse conhecimento simbólico, de maneira que para além de poder ser aplicado, pudesse também ser analisado. Assim, no processo de treino são induzidas regras lógicas de extracção dos elementos relevantes, satisfazendo esta exigência.

A metodologia proposta foi devidamente avaliada, mostrados os resultados obtidos e feita alguma comparação com outros sistemas. O sistema obteve resultados muito satisfatórios, no domínio fixado, abrindo assim novas possibilidades para futuras aplicações interessantes.

UNIVERSIDADE DE ÉVORA

GNU Prolog to Java

A study on how to connect the two programming environments

Abstract

This work is intended to study and put up a bidirectional interface between GNUProlog and the Java language. The purpose of this tool is the possibility to use the power of logic programming within a cross platform environment. This meaning to write prolog programs invoking java methods and Java programs calling prolog predicates.

Java and Prolog are an ideal pair for delivering useful intelligent applications with state-of-the-art user interfaces deployed over several operating systems and media. Mixed with Java this intelligence benefits from all of the design characteristics of this language like platform independence, security, type safety, exception handling, and so on, to create such tools as servers for diagnosing problems, spider and robot applications that transparently wander the net, mobile intelligent agents attending requests from other reasoning agents, human or not.

One of the primary objectives, though, intended to be achieved is to integrate a full blown, ISO Prolog compliant, open source Prolog with the many IDEs and tools in the Java momentum.

Dissertação de

Mestrado em Inteligência Artificial Aplicada

Orientador: Prof. Salvador Pinto de Abreu

Por: David José Murteira Mendes

Palavras-chave: GNU Prolog, Java, JNI

Modelos de Previsão da Carga Crítica e Tensão Última de Estruturas de Engenharia Civil, Utilizando Técnicas de Inteligência Artificial

Hélder Quintela[1], Manuel Filipe Santos[2], Paulo Cruz[3]

Resumo

Na concepção e projecto de qualquer estrutura de engenharia civil devem ser ponderados factores da mais variada índole, tais como *estética*, *funcionalidade*, *deformabilidade*, *estabilidade*, *durabilidade*, *resistência* e *custo*. Na generalidade dos casos esse exercício está condicionado à busca da solução mais segura e mais económica. Esta preocupação, aliada à evolução das propriedades dos materiais e dos meios de cálculo, tem conduzido à utilização de critérios de dimensionamento cada vez mais refinados.

No caso de estruturas metálicas com secções muito esbeltas, o erro apresentado pelas fórmulas de previsão de carga crítica de vigas sujeitas a cargas concentradas é significativo, ficando a dever-se ao grande número de parâmetros que influenciam o comportamento e ao número insuficiente de dados experimentais que permitam efectuar uma análise paramétrica completa e calibrar, convenientemente modelos simplificados.

Neste trabalho são apresentados dois casos de estudo para geração de modelos de previsão, um da carga crítica de estruturas metálicas com secções muito esbeltas, e outro de previsão da tensão última de vigas com perfil em "I" de inércia variável, utilizando técnicas de *Inteligência Artificial*.

A aplicação destas técnicas a problemas de *Engenharia Civil* tem conhecido um interesse crescente, motivado pelas características íntrinsecas destas técnicas que permitem a resolução de problemas complexos, comuns na área da *Engenharia Civil*.

No primeiro caso de estudo, os modelos gerados com base em *redes neuronais artificiais* provaram ser mais eficazes que os métodos convencionais. Todavia, no segundo caso, a performance dos modelos gerados não revelou melhorias significativas quando comparada com outros modelos alternativos, apontando para a necessidade de realização de trabalho complementar (e.g. *Análise de Sensibilidade* da importância das variáveis de entrada dos modelos).

No futuro, pretende-se o desenvolvimento de modelos de suporte à decisão para a manutenção de estruturas de engenharia civil (e.g., *obras de arte*).

Palavras-chave: *Carga Crítica de Vigas de Aço, Tensão Última de Vigas, Engenharia Civil, Descoberta de Conhecimento em Bases de Dados, Metodologias, Data Mining, Redes Neuronais Artificias.*

[1] Mestrando, Assistente Convidado, Departamento de Sistemas de Informação, Universidade do Minho, Azurém, 4800-058 Guimarães, hquintela@dsi.uminho.pt

[2] Orientador, Professor Associado, Departamento de Sistemas de Informação, Universidade do Minho, Azurém, 4800-058 Guimarães, mfs@dsi.uminho.pt

[3] Orientador, Professor Auxiliar, Departamento de Engenharia Civil, Universidade do Minho, Azurém, 4800-058 Guimarães, pcruz@civil.uminho.pt

Genetic Vehicle Representation: An Overview

Jorge Tavares
Universidade de Coimbra
Polo II, Pinhal de Marrocos
3030 Coimbra, Portugal
jast@dei.uc.pt

Francisco B. Pereira
ISEC
Quinta da Nora
3030 Coimbra, Portugal
xico@dei.uc.pt

Ernesto Costa
Universidade de Coimbra
Polo II, Pinhal de Marrocos
3030 Coimbra, Portugal
ernesto@dei.uc.pt

Keywords: evolutionary algorithms, vehicle routing

1 The Problem

The Vehicle Routing Problem (VRP) is a complex combinatorial optimization problem, which can be described as follows: given a fleet of vehicles with uniform capacity, a common depot, and several customer demands, find the set of routes with overall minimum route cost which service all the demands. All the itineraries start and end at the depot, and they must be designed in such a way that each customer is served only once and just by one vehicle. The VRP is NP-Hard and the most general version is the Capacitated Vehicle Routing Problem (CVRP).

An important extension to the problem is the Vehicle Routing Problem with Time Windows (VRPTW) which adds several time constraints to the previous definition. Associated with each customer there is a time window during which it has to be served.

2 Evolutionary Algorithm

Initial attempts shows that the representation is a key issue in the application of EC techniques to the VRP. This lead to the proposal of a new representational scheme, Genetic Vehicle Representation (GVR)[Tavares, 2003], which deals efficiently with the two levels of information that a candidate solution must encode: clustering of the demands (i.e., allocation of all the demands to different vehicles) and specification of the delivery ordering for each one of the routes. It is important to notice that our methodology does not use any specific heuristic.

A candidate solution to an VRP instance must specify the number of vehicles, the partition of the demands through all these vehicles and also the delivery order for each route. In GVR, the genetic material of an individual contains several routes, each of them is composed by an ordered subset of the customers. All demands belonging to the problem being solved must be present in one of the routes.

Two categories of operators are considered: crossover and mutation. They must be able to deal with the two levels of the representation. Thus, they should be capable to change the delivery order within a specific route and to modify the allocation of demands to vehicles. In this last situation, they can, not only switch customers from one route to another, but also modify the number of vehicles belonging to a solution (adding and removing routes).

The crossover operator used in our approach does not promote a mutual exchange of genetic material between two parents. Descendants resulting from crossover can be subject to mutation. We consider four operators, based on proposals usually applied to order based representations: swap, inversion, insertion and displacement. All genetic operators described have a specific probability of application to a single individual.

We performed several experiments with some well-know benchmarks, that enabled us to discover new best solutions.

3 Conclusion

We presented an evolutionary approach to the VRP. The proposed two-level representational scheme shows to be effective, since it was able to discover the best solutions found and even improving some of them.

References

[Tavares, 2003] Tavares, J. (2003). Uma Abordagem Evolucionária ao Problema do Encaminhamento de Veículos. Master's thesis, Universidade de Coimbra.

Verificação e Especificação de Conhecimento Temporal em Sistemas Inteligentes

Jorge Santos (doutorando)[1], Zita Vale (orientador)[1],
Carlos Ramos(co-orientador)[1] e António Almeida Vale (co-orientador)[2]

[1] GECAD – Grupo de Engenharia do Conhecimento e Apoio à Decisão
Instituto Superior de Engenharia do Porto; 4200-072 Porto – Portugal
{ajs; csr; zav}@isep.ipp.pt

[2] Faculdade de Engenharia da Universidade do Porto
Departamento de Engenharia Electrotécnica e Computadores; 4200-465 Porto – Portugal
avale@fe.up.pt

Palavras-chave: Ontologias, sistemas inteligentes, verificação automática

O tempo, tal como o espaço, é uma categoria fundamental à cognição humana. Pelas suas características é ortogonal à generalidade dos domínios de conhecimento sendo a capacidade de representar e raciocinar sobre o tempo um aspecto crucial na vida humana. Este tópico desperta grande interesse desde há décadas, designadamente nas áreas de planeamento, escalonamento, interpretação de linguagem natural, sistemas multi-agente, raciocínio qualitativo e de senso comum.

Para além da produção de um estado da arte sobre os tópicos investigados, designadamente desenvolvimento de ontologias, verificação de sistemas inteligentes, raciocínio e representação temporal, o trabalho desenvolvido compreende duas partes distintas e complementares na integração de conhecimento temporal em sistemas inteligentes, a especificação e a verificação.

A verificação é um processo que visa garantir que um determinado sistema foi desenvolvido com recurso as técnicas e métodos adequados e satisfaz os requisitos do utilizador. Devido às especificidades dos sistemas inteligentes, as técnicas desenvolvidas pela disciplina de sistemas de informação não se mostraram eficazes na verificação destes sistemas. Assim, foram investigadas técnicas mais efectivas neste tipo de sistema, com particular destaque para a detecção de anomalias do conhecimento. Sendo que uma anomalia é sintoma que permite por sua vez detectar potenciais erros. A este respeito foi desenvolvido um sistema de detecção automática de anomalias em sistemas que integram conhecimento temporal designado VERITAS. Este sistema foi aplicado com sucesso no SPARSE, um sistema pericial desenvolvido para auxiliar os operadores da Rede Eléctrica Nacional (REN) no diagnóstico de incidentes e produção de conselhos para reposição do serviço.

A especificação do tempo em aplicações inteligentes é um processo naturalmente complexo em parte devido à necessidade de seleccionar e implementar diferentes aspectos ontológicos, como topologia, primitivas temporais, granularidade, modelos densos *vs* discretos, bem como aspectos relacionados com a eficácia e exequabilidade da solução a desenvolver e adequação aos requisitos do utilizador. A esse respeito foram investigadas técnicas que visam o desenvolvimento de ontologias complexas com recurso à combinação de ontologias. A combinação distingue-se da junção de ontologias, pois, enquanto esta pretende criar uma ontologia a partir de ontologias sobre domínios similares, a combinação pretende combinar conceitos ortogonais a um determinado domínio, como o tempo e o espaço. As técnicas investigadas foram materializadas através de uma ferramenta – FONTE - que foi aplicada com sucesso na engenharia de conceitos temporais em diversos domínios de conhecimento, prevendo-se para trabalho futuro a integração dos aspectos espaciais.

Towards A Language For Beliefs-Knowledge Representation

Juan Carlos Acosta Guadarrama
Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Quinta da Torre 2829-516
Caparica, Portugal
jcag@fct.unl.pt

Abstract

A logic language based on stable models is visualized to make the difference between beliefs and knowledge, as well as a global view of its repercussion. This proposal has a basic part on belief revision that grants the agent a degree of autonomy in the sense of automatically redefining new specifications and even contradictions from its changing environment. On the other hand, formal interaction among other agents provides the potential to define a society aimed at the solution of a particular problem. As a result, this solution would generate knowledge from individual beliefs of specialized agents. Should a beliefs-knowledge generalization be achieved in a language, an agent society shall be able to take over knowledge from a specific field, construct it and represent it by means of negotiating individual beliefs. It is about an area of great impact on human reasoning towards automatic process of heterogeneous and abundant information.

Keywords: Stable semantics; belief revision; dynamic knowledge representation; updates; agent society; argumentation

1 Identification

Name: Juan Carlos Acosta Guadarrama

Course: Doctorate

Institution: Departamento de Informática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Supervisor: Dr. Luís Moniz Pereira

Co-supervisor:

Area: Computer Science

Subarea: Computer Science

Research Line: Logic Programming

Análise da pesquisa heurística de um algoritmo para a determinação da estrutura de proteínas

Marco Correia

21 de Novembro de 2003

Orientador: Pedro Barahona
Instituição: Faculdade de Ciências e Tecnologia - UNL
Palavras-chave: programação por restrições, biologia computacional

Resumo

As proteínas são sequências compostas de um alfabeto de 20 aminoácidos que em grande parte determinam a sua forma tridimensional (conformação). A conformação das proteinas por sua vez descreve a sua função (e.g. enzimas, imunoglobulinas, etc), uma informação obviamente muito importante.

Um dos maiores problemas ainda por resolver no campo da biologia computacional consiste em determinar a conformação a partir de uma dada sequência de aminoácidos. Existem várias técnicas para abordar este problema: dinâmica molecular, minimização da energia livre, homologia, cristalografia e raio-X, e métodos assistidos por NMR.

Esta tese descreve a análise da pesquisa heurística usada no PSICO, um algoritmo que utiliza programação por restrições para resolver o problema da determinação da estrutura de proteínas a partir do conjunto de pares de distâncias gerados por NMR [2]. Testes preliminares mostram que, no referido algoritmo, a relação entre as heurísticas de variável e valor e a qualidade dos resultados globais obtidos não é directa [1]. Este trabalho foi motivado pela necessidade de perceber o efeito que estes parâmetros têm neste problema específico, para que melhores alternativas possam ser encontradas.

Referências

[1] Marco Correia. A profiler for the CSP solver of project proteina. Technical report, FCT/UNL, 2002.

[2] Ludwig Krippahl. *Integrating Protein Structural Information.* PhD thesis, FCT/UNL, 2003.

1 Título

Programação por Restrições em Sistemas de Informação Heterogéneos
(Constraint Programming in Heterogeneous Information Systems)

2 Autor

Vitor Manuel Beires Pinto Nogueira

3 Orientadores

- Salvador Pinto Abreu (Universidade de Évora.)
- Gabriel Torcato David (Universidade do Porto.)

4 Palavras Chave

Sistemas de Informação Heterogéneos; Programação: em Lógica, com Restrições e Orientada por Objectos.

5 Plano de Trabalho

1. *Survey* das abordagens actuais no âmbito dos sistemas de informação heterogéneos (SI).
2. *Survey* das técnicas e linguagens de programação com restrições, nomeadamente programação em lógica com restrições, restrições concorrentes e restrições sobre conjuntos.
3. *Survey* dos formalismos e das técnicas de modularização e extensões que incluam conceitos de programação orientada por objectos, para linguagens de Programação em Lógica.
4. Desenvolvimento de uma linguagem de programação com restrições. A linguagem pretendida servirá de "coluna vertebral" para a construção e manuseamento de SIs e especificará a semântica da implementação a desenvolver no ponto 5.
5. Especificação e desenvolvimento duma implementação para a linguagem definida.
6. Utilização da implementação para a elaboração de restrições de integridade e *queries* inteligentes em SI.
7. Aplicação do sistema desenvolvido a centros de tratamento de informação (centros de informática institucionais, ...), com especial relevo para aqueles que recebam informação que varie quer na sua génese quer na tecnologia utilizada.

Mapeamento de ontologias usando uma abordagem orientada por serviços multi-dimensionais

Nuno Silva (doutorando)[1], João Rocha (orientador)[1] e José Cardoso (co-orientador)[2]

[1] GECAD – Grupo de Engenharia do Conhecimento e Apoio à Decisão
Instituto Superior de Engenharia do Porto
4200-072 Porto – Portugal
Nuno.Silva@dei.isep.ipp.pt

[2] Engenharias, Universidade de Trás-os-Montes e Alto Douro
Apartado 202, 5001-911 Vila Real - Portugal

Mapeamento de ontologias é um processo que consiste na definição a nível ontológico de relações de equivalência semântica entre entidades da ontologia de origem e a ontologia de destino, sendo essas relações posteriormente aplicadas na transformação das instâncias da ontologia de origem em instâncias da ontologia de destino. Não existe uma definição universalmente aceite de ontologia, mas a definição de Gruber - "ontologia é uma especificação explícita duma conceptualização" – tende a ser genérica e abstractamente aceite. No contexto deste trabalho de doutoramento, além dos elementos incluídos na definição anterior, ontologia inclui duas características suplementares: (i) a abstração ontológica por contraponto à especificação lógica do modelo de informação, e (ii) a capacidade de partilha e raciocínio sobre a ontologia por diferentes comunidades de informação. Torna-se portanto possível relacionar ontologias distintas, determinando equivalências, diferenças e semelhanças entre elas e assim relacionar bases de conhecimento ou informação com elas conformes.

Mapeamento de ontologias é considerada uma tecnologia fundamental em cenários em que troca e partilha de informação sejam essenciais. Interoperabilidade entre sistemas de informação, Web Semântica, negócios electrónicos, migração de dados, "data wharehouse clean & transform" e evolução dos modelos de dados subjacentes aos sistemas de informação, são alguns desses cenários.

A qualidade de mapeamento de ontologias depende não só das ontologias mas também da capacidade em identificar, especificar e posteriormente executar as transformações necessárias entre entidades ontológicas. É necessário portanto dotar o sistema de apoio a este processo com capacidades de evolução e adaptação às diferentes necessidades impostas pela interligação entre as ontologias. Sugere-se então um sistema de apoio ao processo de mapeamento entre ontologias baseado no conceito de serviço multi-dimensional. O mapeamento de ontologias é um processo moroso, complexo e iminentemente subjectivo, pelo que a sua completa automatização nem sempre é possível. O sistema de apoio deve portanto fornecer ao perito mecanimos sofisticados que lhe permitam o controlo e promoção de iterações ao longo do processo.

O sistema é composto por um conjunto central de módulos representando as diversas fases do processo, complementado por módulos externos e dinamicamente integráveis no sistema base, denominados serviços. Cópia de instâncias de entidades, concatenação e "separação" de atributos, cópia de relações entre instâncias de entidades, mapeamento entre unidades de medida, são alguns dos serviços tipicamente necesários em qualquer cenário. Denominam-se serviços multi-dimensionais porque são várias as competências fornecidas por cada um ao sistema central em diferentes fases do processo e não apenas na fase de transformação, como é tradicional. As competências mais comuns são a medição de equivalências semânticas entre (grupos de) entidades de duas ontologias, especificação automática de relações semânticas, validação de relações semânticas especificadas manualmente, interligação automática entre relações semânticas, gestão da evolução ontológica sobre as relações semânticas e apoio na negociação de relações semânticas entre comunidades de informação.

Definição de *kernels* para problemas de bioinformática

Autor: Orlando Miguel Cruz da Anunciação

Orientador: Arlindo Manuel Limede de Oliveira

Instituição: INESC-ID

Palavras-chave: support vector machines, kernels, bioinformática, splice sites, redes neuronais, árvores de decisão, naive Bayes, classificação.

Resumo

Um dos problemas importantes na pesquisa do ADN e que é tratado neste trabalho, está relacionado com a previsão numa dada sequência de bases (Adenina, Citosina, Timina e Guanina vulgarmente abreviadas respectivamente por A, C, T, G) onde se encontram as zonas de codificação de proteínas. Se uma sequência possui uma zona de codificação de proteínas, é porque existe algures uma ou mais transições (*splice sites*) entre zonas codificantes e zonas não codificantes (respectivamente exões e intrões). Considera-se ainda como problemas diferentes o reconhecimento de zonas de início de tradução (deteccão de sinais *start*) e de zonas de terminação da tradução (detecção de sinais *stop*).

Neste trabalho aplicaram-se algoritmos de aprendizagem (*support vector machines*, redes neuronais, árvores de decisão e classificador *naive Bayes*) ao problema da detecção de transições entre regiões codificantes e não codificantes. A criação de classificadores baseados em *support vector machines* que se adaptem com sucesso ao problema exige que se definam *kernels* apropriados. A aplicação e comparação do desempenho de vários *kernels* é também um dos objectos de estudo deste trabalho.

O desempenho das *support vector machines* superou o desempenho dos restantes métodos utilizados, sendo este resultado bastante significativo estatisticamente em diversos casos. Para além dos *kernels* tradicionais (polinomial, gaussiano e sigmoidal), codificou-se ainda o *kernel* de correlações locais (*locality-improved kernel*). O uso dos *kernels* tradicionais revelou-se suficiente para superar o desempenho dos outros métodos (redes neuronais, árvores de decisão e *naive Bayes*). No entanto nos quatro problemas considerados os melhores resultados foram obtidos usando o *kernel* de correlações locais codificado. Contudo estas diferenças de desempenho não se revelaram estatisticamente significativas.

Accurate Decision Trees for Mining High-speed Data Streams

Authors: João Gama, Ricardo Rocha, Pedro Medas.

Adviser: João Gama.

Institution: LIACC – Laboratório de Inteligência Artificial e Ciências de Computadores.

Key words: Data Streams, Incremental Decision Trees, Functional Leaves.

Abstract: In this paper we study the problem of constructing accurate decision tree models from data streams. Data streams are incremental tasks that require incremental, online, and any-time learning algorithms. One of the most successful algorithms for mining data streams is VFDT. In this paper we extend the VFDT system into directions: the ability to deal with continuous data and the use of more powerful classification techniques at tree leaves. The proposed system, VFDTc, can incorporate and classify new information online, with a single scan of the data, in time constant per example. The most relevant property of our system is the ability to obtain a performance similar to a standard decision tree algorithm even for medium size datasets. Under a bias-variance analysis we observe that VFDTc in comparison to C4.5 is able to reduce the variance component.

Resumo de poster para EPIA'03

Título: Dynamic Maximum Tree Depth – A Simple Technique for Avoiding Bloat in Tree-based GP

Autores: Sara Silva[1], Jonas Almeida[1,2]

Orientador(*): Jonas Almeida

Instituições(*):
[1]Instituto de Tecnologia Química e Biológica, Universidade Nova de Lisboa
[2]Department of Biometry and Epidemiology, Medical University of South Carolina

Palavras Chave: bloat, code growth, tree depth, dynamic limit

(*)Notas:
Este trabalho foi realizado no ano passado e apresentado no GECCO-2003. Desde então, mudei de instituição e de orientador, encontrando-me neste momento a iniciar de novo um doutoramento em computação evolutiva, sob a supervisão do Prof. Ernesto Costa na Universidade de Coimbra.

Abstract

We present a technique, designated as dynamic maximum tree depth [1], for avoiding excessive growth of tree-based GP individuals during the evolutionary process. This technique introduces a dynamic tree depth limit, very similar to the Koza-style strict limit [2] except in two aspects: it is initially set with a low value; it is increased when needed to accommodate an individual that is deeper than the limit but is better than any other individual found during the run. The results show that the dynamic maximum tree depth technique efficiently avoids the growth of trees beyond the necessary size to solve the problem, maintaining the ability to find individuals with good fitness values. When compared to lexicographic parsimony pressure [3], dynamic maximum tree depth proves to be significantly superior. When both techniques are coupled, the results are even better.

References

[1] Silva S, Almeida J (2003): Dynamic Maximum Tree Depth – A Simple Technique for Avoiding bloat in Tree-based GP. In Cantú-Paz E, Foster JA, Deb K, *et al.*, editors, Proceedings of GECCO-2003. Springer Verlag, 1776–1787.

[2] Koza J (1992): Genetic programming – on the programming of computers by means of natural selection, Cambridge, MA. MIT Press.

[3] Luke S, Panait L (2002): Lexicographic Parsimony Pressure. In Langdon WB, Cantú-Paz E, Mathias K, *et al.*, editors, Proceedings of GECCO-2002, San Francisco, CA. Morgan Kaufmann, 829–836.