PROGRAMAÇÃO KAIZEN PARA CONSTRUÇÃO DE MODELOS INTERPRETÁVEIS

UMA ABORDAGEM MULTIOBJETIVO PARA REGRESSÃO SIMBÓLICA

DEFESA DE MESTRADO

Aluno: Artur Henrique Gonçalves Coutinho Alves

Orientador: Prof. Dr. Vinícius Veloso de Melo

Programa de Pós-Graduação em Ciência da Computação

Instituto de Ciência e Tecnologia

Universidade Federal de São Paulo – Campus São José dos Campos

13 de abril de 2017

SUMÁRIO

CONTEXTUALIZAÇÃO

- Aumento da complexidade das atividades da sociedade moderna
  - Grandes massas de dados e sistemas integrados
- Aprendizado de máquina supervisionado
  - Regressão e classificação
  - Diversas aplicações, como engenharia, *data mining* e inteligência artificial em jogos
- Programação Kaizen
- Direção automática
  - Controle preditivo baseado em modelo (MPC)

MOTIVAÇÕES

Evolução de Programação Kaizen

Utilização de controle preditivo para direção automática

Aplicação de jogos como plataforma de teste de inteligência computacional

OBJETIVOS

Identificar pontos que podem ser melhorados em Programação Kaizen e propor soluções

Implementar tais soluções e aplicar em problemas reais, comparando seu desempenho com outras técnicas de aprendizado de máquina

Construir uma plataforma de controle preditivo com Programação Kaizen como técnica de modelagem

Integrar este controle a um piloto de simulador de corrida e testar seu desempenho em situações reais

OTIMIZAÇÃO E REGRESSÃO

MODELANDO COMPORTAMENTOS

OTIMIZAÇÃO

Otimização: $min/max\ f(x)$

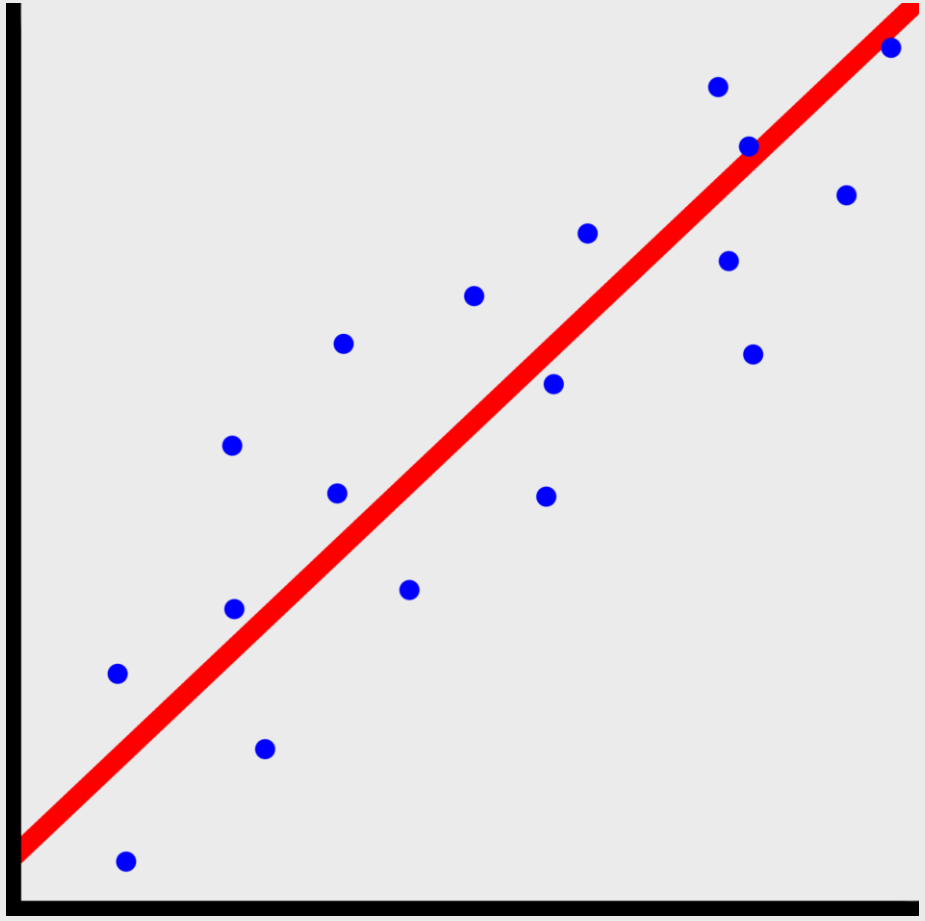
Otimização multiobjetivo

Heurísticas

Meta-heurísticas

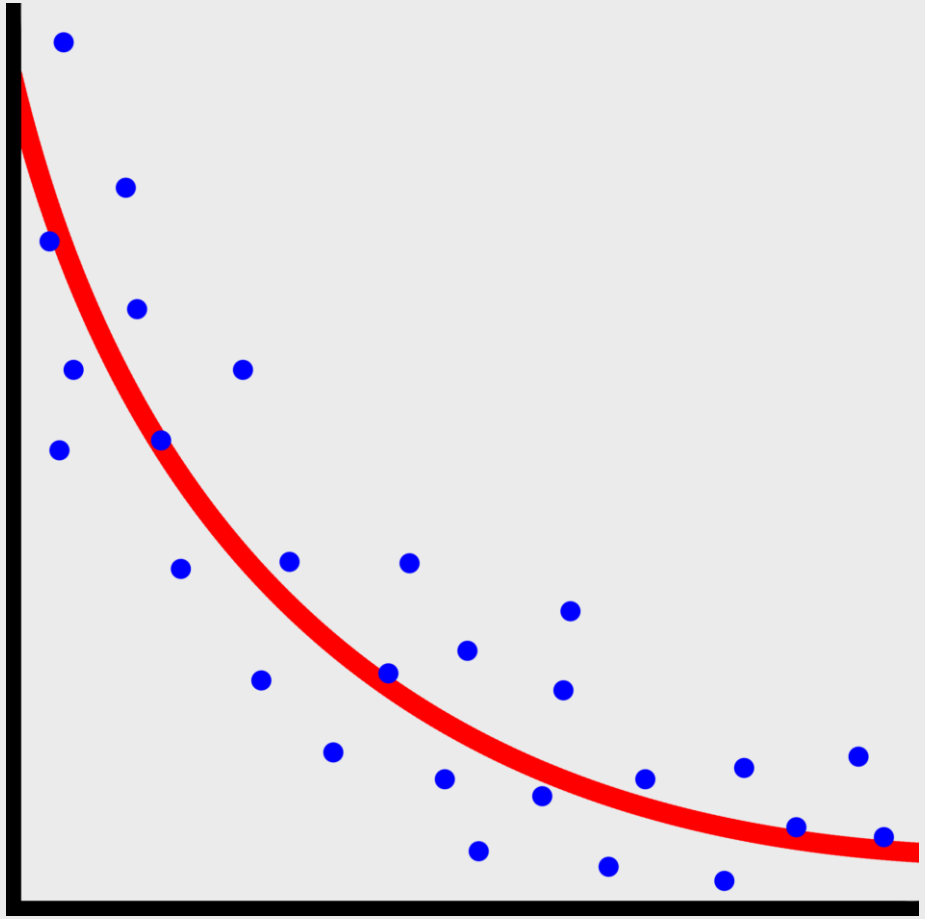
Hiper-heurísticas

OTIMIZAÇÃO
REGRESSÃO LINEAR



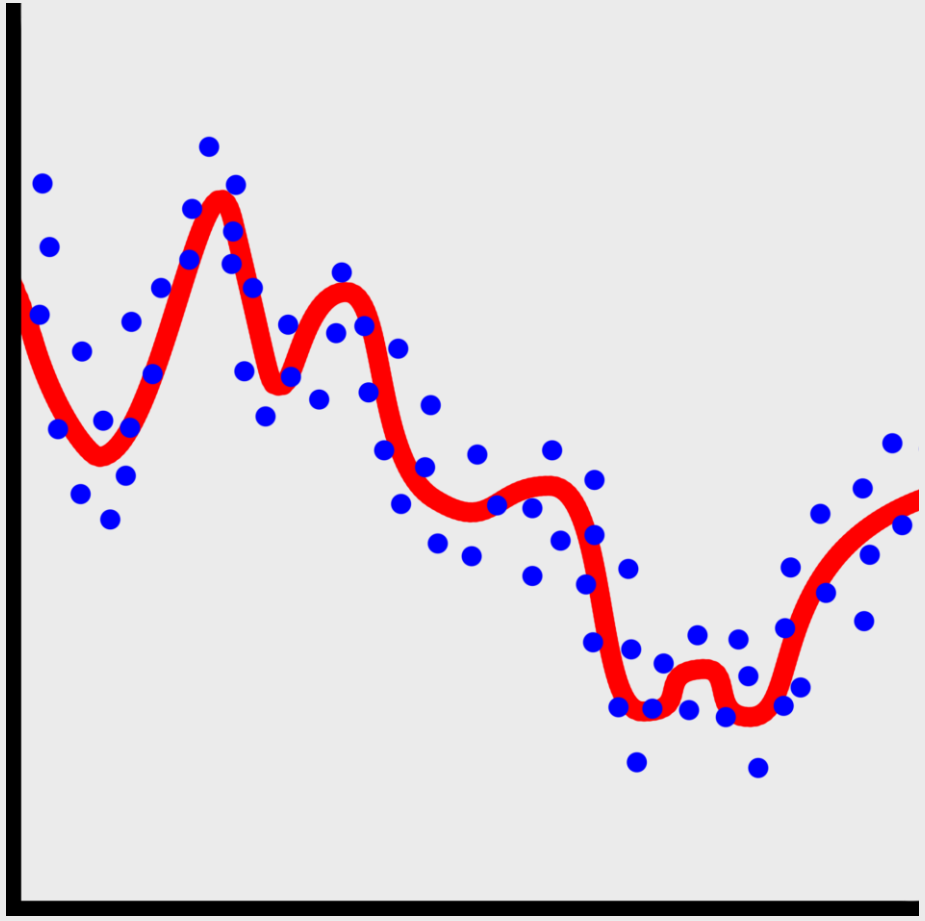
$$y = \boldsymbol{a} + \boldsymbol{b}x$$

OTIMIZAÇÃO
REGRESSÃO NÃO-LINEAR



$$y = \boldsymbol{a}e^{\boldsymbol{b}x}$$

OTIMIZAÇÃO
REGRESSÃO SIMBÓLICA



$$y = ?$$

OTIMIZAÇÃO
REGRESSÃO SIMBÓLICA: TRABALHOS RELACIONADOS

Técnicas com otimização numérica de coeficientes

- MRGP (*Multiple Regression Genetic Programming*)
- GSGP-LSH (*Geometric Semantic Genetic Programming with Local Search – Hybrid*)
- SSR (*Sequential Symbolic Regression*)

Simulated Annealing multiobjetivo para regressão simbólica

Programação Genética para construção de modelos para controle preditivo

- Estudo de Grosman e Lewin
- Estabilização de pêndulo invertido

PROGRAMAÇÃO GENÉTICA

COMPUTAÇÃO EVOLUTIVA PARA REGRESSÃO SIMBÓLICA

PROGRAMAÇÃO GENÉTICA

$$individuo_1 = x_1^2 + x_3$$

$$individuo_2 = \sqrt{x_2} \Big/ 3.14 + \log x_1$$

$$individuo_3 = x_1 + x_2 + x_3$$

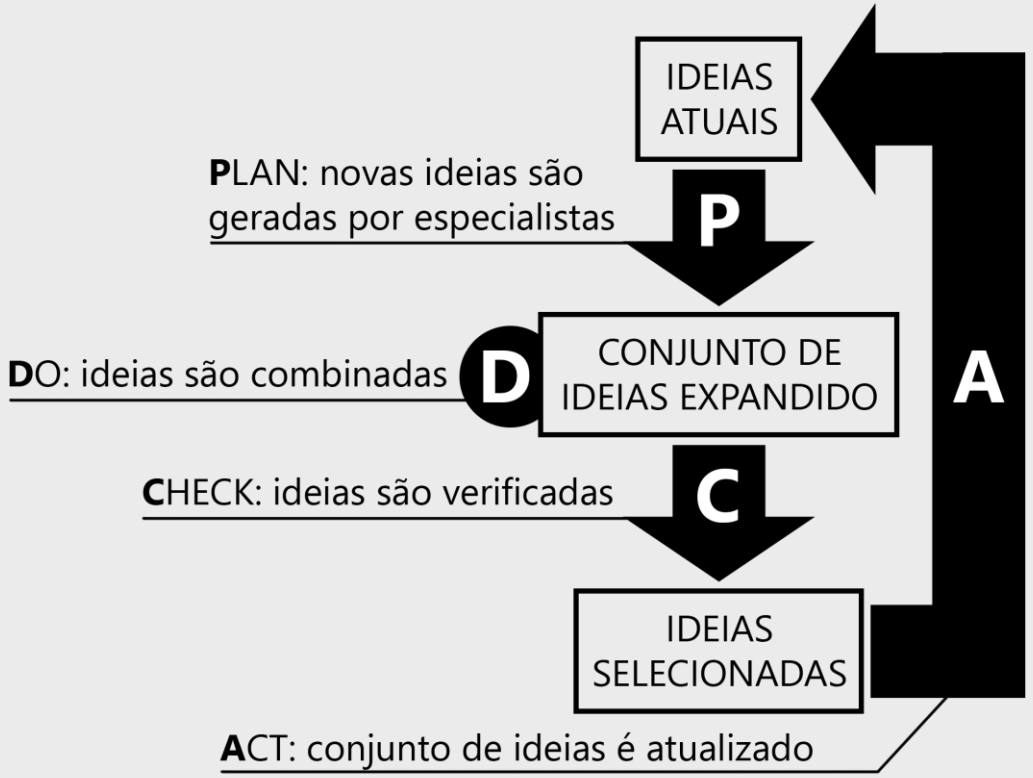
PROGRAMAÇÃO KAIZEN

APLICANDO O PROCESSO DE MELHORIA CONTÍNUA À PROGRAMAÇÃO AUTOMÁTICA

PROGRAMAÇÃO KAIZEN
KAIZEN E PDCA

- Filosofia de trabalho japonesa que busca a melhoria contínua de processos
- Eventos Kaizen
  - Especialistas
  - Metodologia cíclica *Plan-Do-Check-Act* (PDCA)
- Programação Kaizen
  - Aplica conceitos da filosofia Kaizen em inteligência computacional
    - Combina técnicas determinísticas com abordagens aleatórias
  - Evolução *colaborativa*, não competitiva, com indivíduos representando soluções parciais

PROGRAMAÇÃO KAIZEN
KAIZEN E PDCA



$$ideia_1 = x_1^2 + x_3$$

$$ideia_2 = \sqrt{x_2} \Big/ 3.14 + \log x_1$$

$$ideia_3 = x_1 + x_2 + x_3$$

PROGRAMAÇÃO KAIZEN
ESPECIALISTAS PRINCIPAIS

A implementação de KP deste trabalho utiliza oito especialistas; em destaque:

- Criação de novas ideias: $rand(2 * tam)$ terminais, $rand(tam)$ terminais e/ou não-terminais e $rand(2 * tam)$ não-terminais, sendo que $rand$ retorna um valor inteiro de 1 até o parâmetro informado, $tam = rand(maxTam)$ e $maxTam$ é um tamanho máximo definido pelo usuário
- Combinação de ideias: duas ideias existentes são combinadas com um operador de aridade 2
  - Ideias inicialmente pouco importantes podem aumentar a diversidade das soluções
  - A combinação de ideias distintas faz grandes saltos no espaço de busca

PROGRAMAÇÃO KAIZEN
CONSTRUÇÃO DE SOLUÇÕES

Regressão linear múltipla

Busca um hiperplano que aproxime o comportamento de uma variável dependente

Combinação linear das variáveis independentes:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \epsilon_i$$

Variáveis independentes são provenientes das ideias geradas em KP

Combinações não-lineares das variáveis de entrada

Assim, a regressão linear pode aproximar funções não-lineares

$$ideia_1 = x_1^2 + x_3$$

$$ideia_2 = \sqrt{x_2} \Big/ 3.14 + \log x_1$$

$$ideia_3 = x_1 + x_2 + x_3$$

$$y_i = \beta_0 + \beta_1 ideia_{i,1} + \beta_2 ideia_{i,2} + \beta_3 ideia_{i,3}$$

A importância de cada ideia é dada por um teste de significância no modelo

A qualidade do modelo é dada, por exemplo, pelo erro quadrado médio

OTIMIZAÇÃO MULTIOBJETIVO

EVOLUINDO A PROGRAMAÇÃO KAIZEN

OTIMIZAÇÃO MULTIOBJETIVO
OBJETIVOS EM REGRESSÃO SIMBÓLICA

Qualidade: capacidade preditiva de um modelo

Complexidade: dificuldade de interpretação/cálculo de um modelo

Modelos exageradamente complexos podem:

- Apresentar sobreajuste
- Ser muito custosos computacionalmente
- Ser difíceis de interpretar

Maior complexidade, até certo ponto, leva a melhor qualidade

Objetivos são contrários entre si

OTIMIZAÇÃO MULTIOBJETIVO
COMPLEXIDADE

Há diferentes definições para a complexidade em regressão simbólica

- Complexidade estrutural: comprimento das equações
- Complexidade semântica: dificuldade das funções presentes nas equações

$f(x) = 3x^3 + 4x^2 + 2x + 6$ vs. $g(x) = e^{cos\sqrt{x}}$

Neste trabalho, é considerada a complexidade semântica, dada pela não-linearidade das funções

Três etapas do algoritmo consideram a complexidade

- Escolha de ideias quando há alta correlação
- Seleção de melhor padrão ao fim de cada iteração
- Seleção de melhor padrão ao fim da execução

EXPERIMENTOS COM BASES DA LITERATURA

AVALIANDO A PROGRAMAÇÃO KAIZEN MULTIOBJETIVO

EXPERIMENTOS
CONJUNTOS DE DADOS

Abreviação	Nome	Nº de covariáveis	Nº de instâncias
air[1]	Airfoil Self-Noise	5	1.503
bio [131]	Human Oral Drug Biovailability	241	359
con[2]	Concrete Compressive Strength	8	1.030
cpu[3]	Computer Hardware	7	209
enC[4]	Energy efficiency (cooling only)	8	768
enH[5]	Energy efficiency (heating only)	8	768
for[6]	Forest Fires	10	517
ppb [5]	Plasma Protein Binding Levels	626	131
tow [143]	Distillation Tower Problem	25	4.999
wiR[7]	Wine Quality (red only)	11	1.599
wiW[8]	Wine Quality (white only)	11	4.898
yac[9]	Yacht Hydrodynamics	6	768
snk	SnakeOil	80	420

Conjuntos de dados utilizados.

EXPERIMENTOS
CONFIGURAÇÕES DE PROGRAMAÇÃO KAIZEN

	MOKPSA				KPSA				SMORBF	SMOPoly
	iter	*ideas*	*perexp*	*corr*	*iter*	*ideas*	*perexp*	*corr*	*gamma*	*exponent*
air	1.000	10	5	0,9	500	10	2	0,9	100	2
bio	1.000	3	5	0,7	1.000	2	2	0,7	1	1
con	500	10	1	0,7	1.000	10	1	0,9	10	3
cpu	1.000	10	2	0,9	1.000	10	5	0,9	1	2
enC	500	10	5	0,7	1.000	10	10	0,5	10	3
enH	500	3	1	0,9	1.000	10	10	0,5	10	3
for	100	2	3	0,3	100	3	2	0,3	1	3
ppb	N/A				100	3	5	0,5	0,01	1
tow	1.000	2	5	0,9	1.000	10	1	0,5	10	3
wiR	500	10	2	0,7	1.000	5	1	0,9	1	2
wiW	500	10	1	0,9	500	10	1	0,9	100	3
yac	1.000	10	10	0,9	1.000	2	10	0,7	10	3

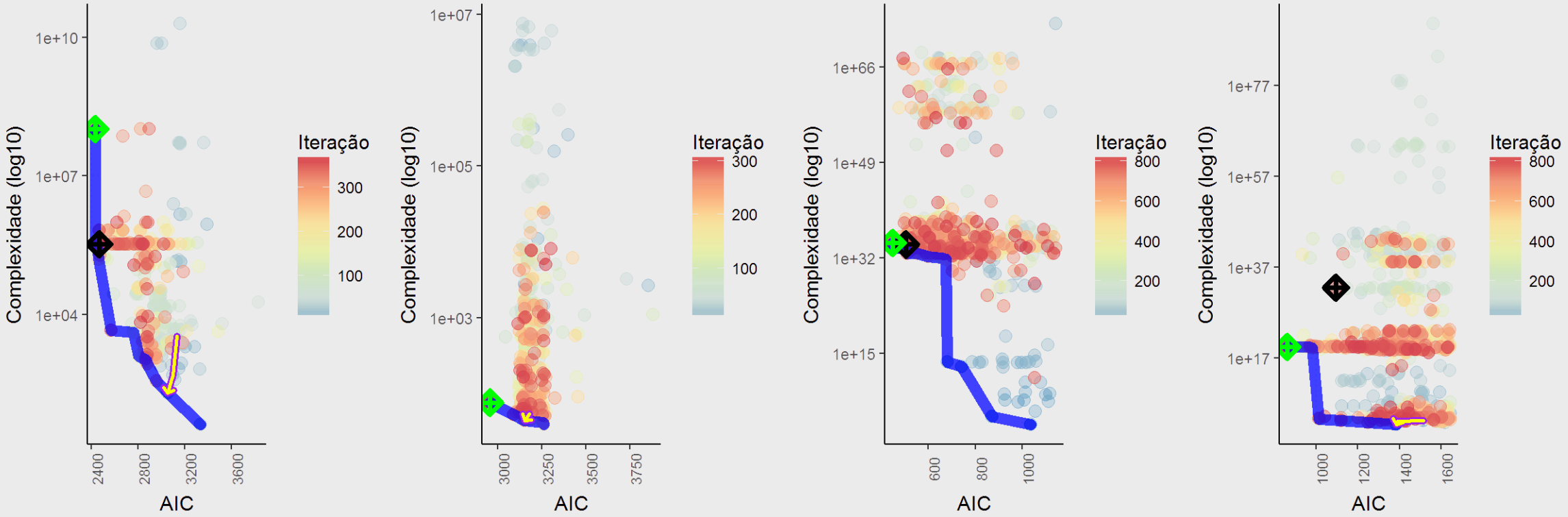
Melhor configuração de cada técnica para cada conjunto de dados de acordo com a mediana do RMSE.

EXPERIMENTOS
CONFIGURAÇÕES DE PROGRAMAÇÃO KAIZEN

	MOKPSA				KPSA			
	iter	*ideas*	*perexp*	*corr*	*iter*	*ideas*	*perexp*	*corr*
#1	1.000	10	2	0,9	500	10	10	0,7
#2	1.000	10	3	0,7	1.000	10	3	0,7
#3	500	10	5	0,7	1.000	10	10	0,7
#4	500	10	2	0,9	1.000	10	3	0,9
#5	1.000	10	5	0,7	1.000	10	1	0,7

Melhores configurações gerais de KPSA e MOKPSA para todos os conjuntos de dados de acordo com a mediana do RMSE.

EXPERIMENTOS
QUALIDADE X COMPLEXIDADE



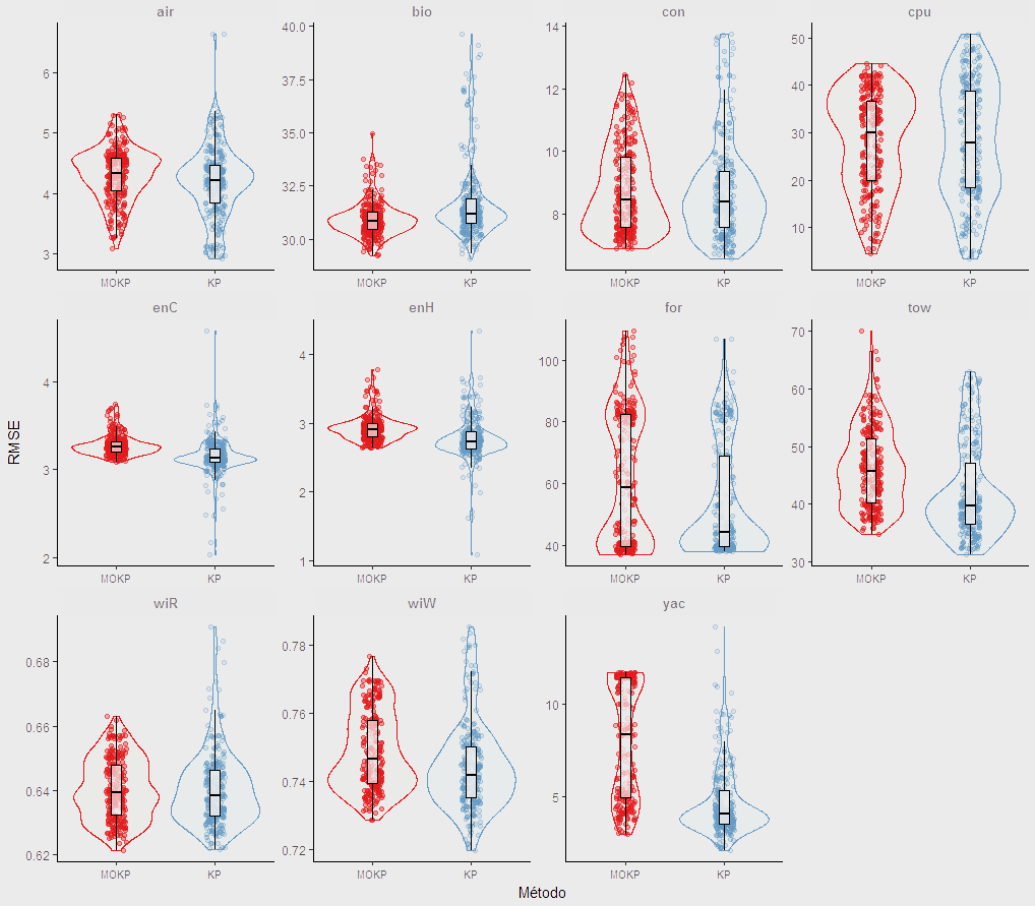
Fronteiras de Pareto das melhores configurações de MOKPSA em enC e yac

EXPERIMENTOS
ABORDAGEM MULTIOBJETIVO

	MOKPSA						KPSA		
	RMSE		R^2		Complexidade		RMSE	R^2	Complexidade
	Absoluto	Relativo	Absoluto	Relativo	Absoluta	Relativa	Absoluto	Absoluto	Absoluta
air	3,09	105,70%	0,79	96,33%	917	52,82%	2,92	0,82	1.736
bio	29,21	100,44%	0,13	133,91%	417	556%	29,09	0,09	75
con	6,91	104,94%	0,83	97,79%	322	9,73%	6,59	0,85	3.308
cpu	4,54	134,47%	1,00	99,92%	688	44,44%	3,38	1,00	1.548
enC	3,08	151,89%	0,89	93,79%	394	8,31%	2,03	0,95	4.741
enH	2,63	243,40%	0,93	94,32%	87	1,59%	1,08	0,99	5.484
for	36,85	97,41%	0,00	56,33%	27	3,81%	37,83	0,01	708
ppb	N/A	N/A	N/A	N/A	N/A	N/A	27,14	0,20	213
tow	34,86	111,54%	0,84	96,32%	708	20,26%	31,25	0,87	3.495
wiR	0,62	99,94%	0,37	101,46%	376	22,25%	0,62	0,37	1.690
wiW	0,73	101,24%	0,32	94,35%	332	9,16%	0,72	0,33	3.625
yac	2,96	141,67%	0,96	98,04%	981	89,02%	2,09	0,98	1.102
Média		126,6%		96,6%		74,31%			
Desvio-padrão		43,12		17,52		161,94			
Mediana		105,7%		96,33%		20,26%			

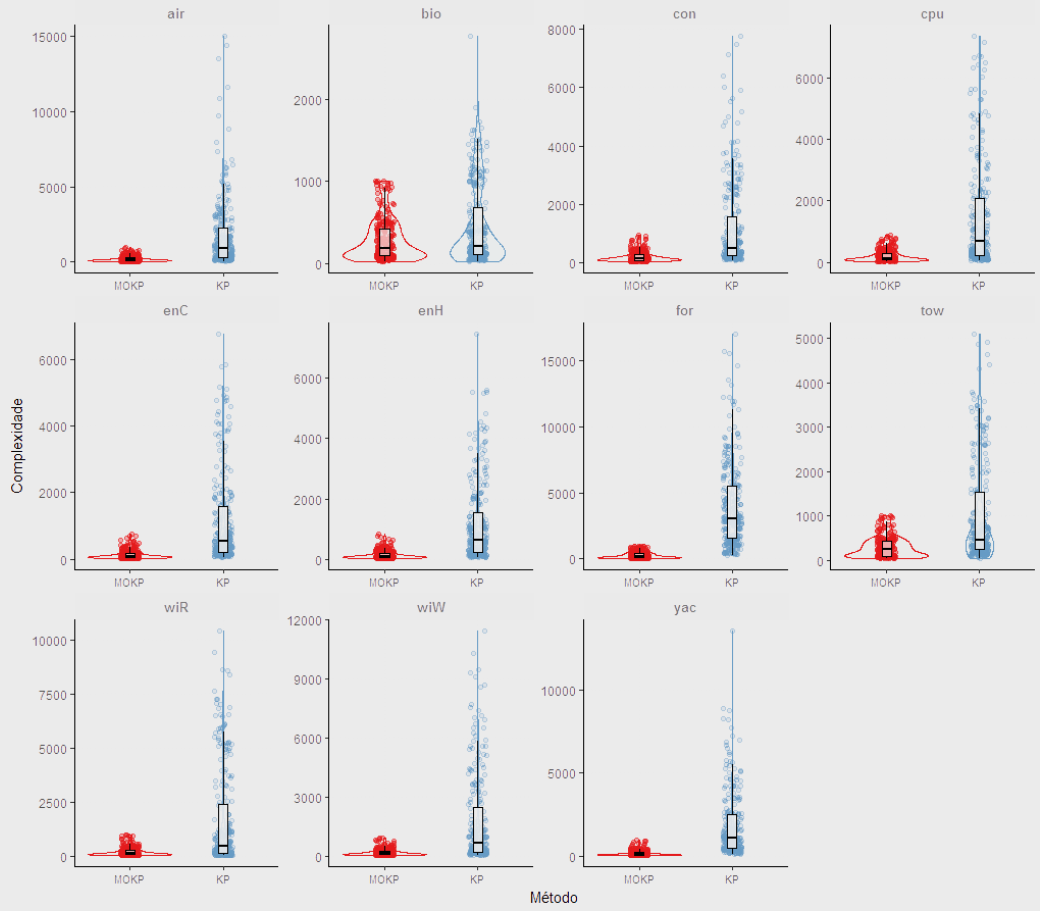
Valores medianos de RMSE, R^2 e complexidade das melhores configurações específicas de MOKPSA e KPSA nos conjuntos de dados da literatura. Os valores relativos em MOKPSA têm como base os respectivos valores absolutos de KPSA.

EXPERIMENTOS
ABORDAGEM MULTIOBJETIVO: QUALIDADE



Distribuição de RMSE da melhor solução apresentada por todas as configurações de MOKP e de KP para cada conjunto de dados da literatura

EXPERIMENTOS
ABORDAGEM MULTIOBJETIVO: COMPLEXIDADE



Distribuição de complexidade da melhor solução apresentada por todas as configurações de MOKP e de KP para cada conjunto de dados da literatura

EXPERIMENTOS
REGRESSÃO SIMBÓLICA: QUALIDADE

	MOKPSA		KPSA		SSR	
	Mediana	IQR	Mediana	IQR	Mediana	IQR
air	3,27	0,85	3,08	0,27	3,06	0,39
bio	31,39	4,84	34,23	14,75	31,21	3,38
con	7,03	2,05	6,99	0,61	7,02	0,62
cpu	4,54	5,5	10,78	16,48	55,26	30,27
enC	3,09	0,32	2,88	0,46	2,38	0,45
enH	2,65	0,29	2,5	1,19	1,83	0,66
for	93	67,29	83,86	74,13	71,23	66,86
ppb	N/A	N/A	N/A	N/A	29,4	7,51
tow	45,68	12,72	37,81	11,43	34,91	3,7
wiR	0,63	0,04	0,65	0,07	0,64	0,03
wiW	0,73	0,03	0,73	0,02	0,73	0,02
yac	3,04	1,24	2,92	1,32	1,88	0,62

Valores de RMSE mediano e interquartil para as melhores configurações gerais de MOKPSA e KPSA e para SSR nos conjuntos de dados da literatura.

EXPERIMENTOS
REGRESSÃO SIMBÓLICA: QUALIDADE

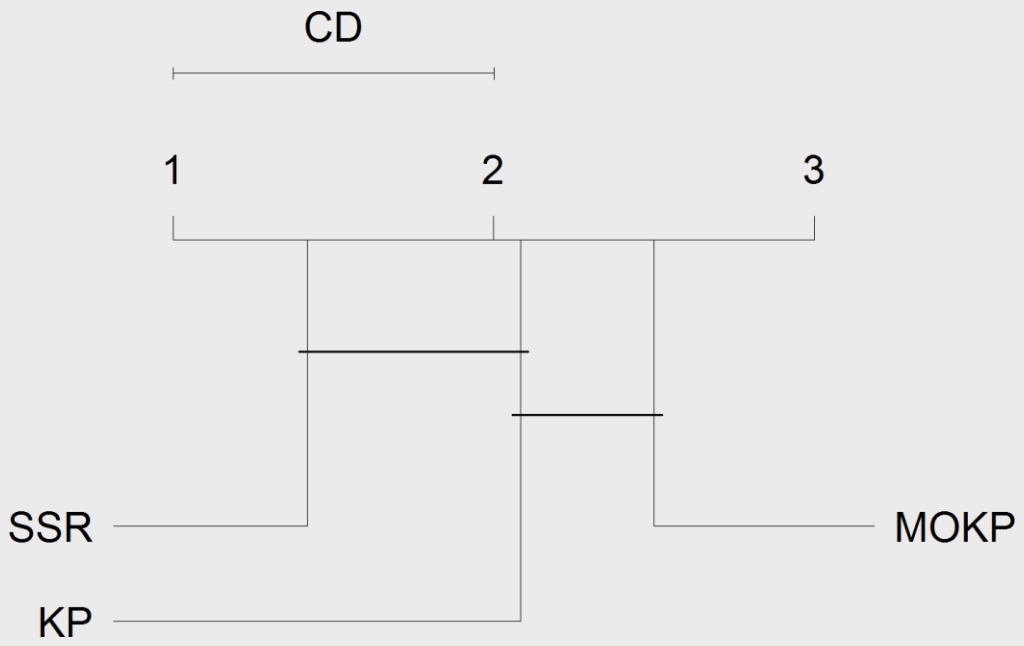


Gráfico de diferença crítica (CD=1). Métodos conectados não apresentam diferença significativa para α=0,05.

EXPERIMENTOS
REGRESSÃO SIMBÓLICA: TAMANHO DE FUNÇÃO

	MOKPSA		KPSA		SSR	
	Mediana	IQR	Mediana	IQR	Mediana	IQR
air	107	30	145	18	641	54
bio	137	110	127	24	252	49,5
con	135	40	149	26	516	75
cpu	101	10	147	34	412	41
enC	85	36	143	46	554	73,5
enH	91	20	143	26	557	87
for	1.214	1.273	195	46	374	59
ppb	N/A	N/A	N/A	N/A	283	97,5
tow	103	36	137	22	619	58,5
wiR	127	56	151	16	437	44,5
wiW	119	58	149	26	540	55
yac	201	140	159	22	595	74,5

Valores de RMSE mediano e interquartil para as melhores configurações gerais de MOKPSA e KPSA e para SSR nos conjuntos de dados da literatura.

EXPERIMENTOS
APRENDIZADO DE MÁQUINA

	MOKPSA		KPSA			SMORBF			SMOPoly			MLP			LinReg		
	RMSE	R^2	RMSE	R^2		RMSE	R^2		RMSE	R^2		RMSE	R^2		RMSE	R^2	
air	3,09	0,79	2,92	0,82	●●●	2,69	0,85	●●●	4,29	0,62	○○○	4,13	0,67	○○○	4,74	0,52	○○○
bio	29,21	0,13	29,09	0,09		28,00	0,14		35,86	0,07		36,92	0,03	○○○	36,27	0,03	○○○
con	6,91	0,83	6,59	0,85	●	5,84	0,88	●●●	7,07	0,82	○○○	7,50	0,83	○○○	10,27	0,62	○○○
cpu	4,54	1,00	3,38	1,00	●●●	16,07	0,99	○○	17,67	0,99	○○○	7,56	1,00	○○○	28,91	0,95	○○○
enC	3,08	0,89	2,03	0,95	●●●	2,02	0,96	●●●	2,49	0,93	●●●	2,42	0,95		3,20	0,89	
enH	2,63	0,93	1,08	0,99	●●●	1,64	0,98	●●●	1,74	0,97	●●●	1,25	0,99	●●●	2,86	0,92	○○
for	36,85	0,00	37,83	0,01		36,94	0,00	○○○	36,89	0,00	○○○	57,11	0,00		36,95	0,00	○
ppb	N/A	N/A	27,14	0,20	N/A	27,51	0,34	N/A	38,90	0,06	N/A	34,67	0,08	N/A	50,94	0,00	N/A
tow	34,86	0,84	31,25	0,87	●●●	14,40	0,97	●●●	15,49	0,97	●●●	20,11	0,96	●●●	33,68	0,85	●●●
wiR	0,62	0,37	0,62	0,37		0,63	0,39	○○○	0,64	0,37		0,73	0,33	○○○	0,65	0,35	
wiW	0,73	0,32	0,72	0,33		0,66	0,43	●●●	0,73	0,33	●●●	0,77	0,32	○○○	0,75	0,28	○
yac	2,96	0,96	2,09	0,98	●	4,16	0,96	○○○	3,07	0,97	○	1,20	1,00	●●●	8,69	0,67	○○○
V/E/D			0/4/7			4/1/6			5/2/4			6/2/3			8/2/1		

Valores medianos de RMSE e R^2 para as melhores configurações específicas das técnicas executadas.

EXPERIMENTOS
APRENDIZADO DE MÁQUINA

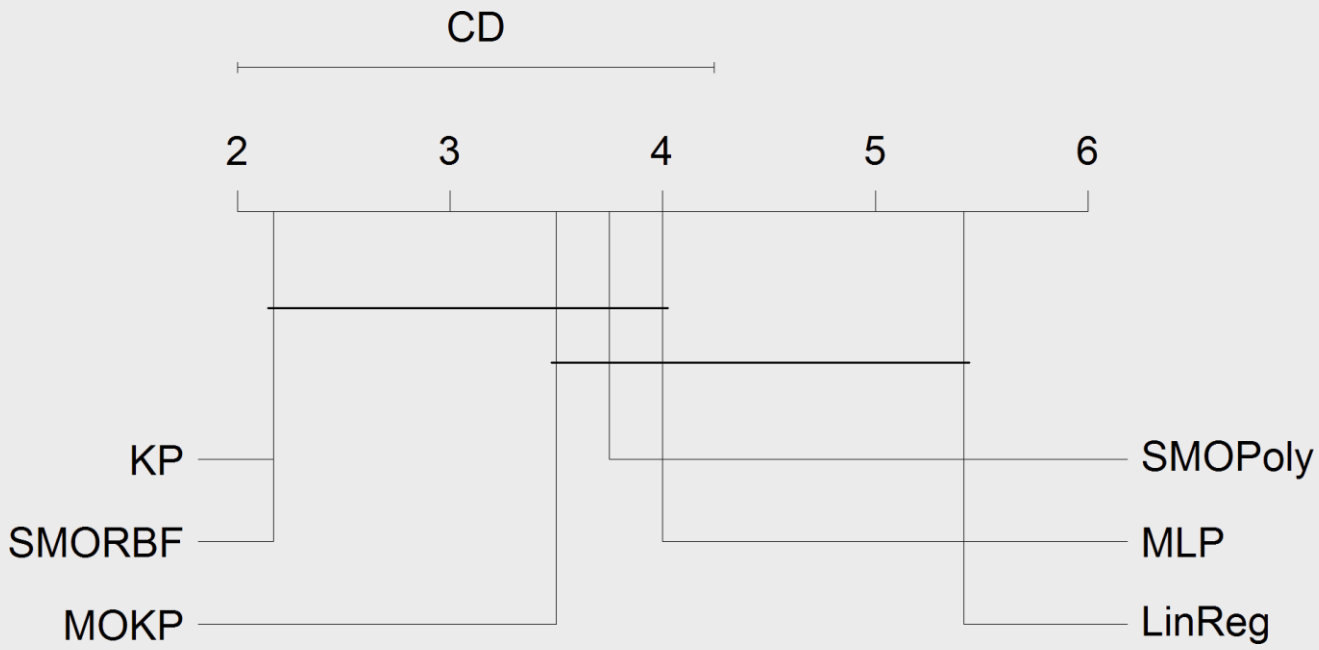


Gráfico de diferença crítica (CD=2,24). Métodos conectados não apresentam diferença significativa para α=0,05.

APLICAÇÃO: DIREÇÃO AUTOMÁTICA

REVOLUCIONANDO OS MEIOS DE TRANSPORTE

DIREÇÃO AUTOMÁTICA

Sistemas autônomos são utilizados para controle na aviação devido à complexidade dos sistemas envolvidos

O transporte terrestre, por sua vez, é muito mais suscetível a erros humanos

- Falta de sistemas autônomos
- Produção em massa de veículos

Iniciativas de pesquisa e desenvolvimento

Simuladores de corrida

DIREÇÃO AUTOMÁTICA
O PROBLEMA

Objetivo: correr o mais rápido possível em uma pista

Maior complexidade em curvas

Simulador: TORCS

Abstração dos dados

- Sensores
  - Internos
  - Externos
- Atuadores

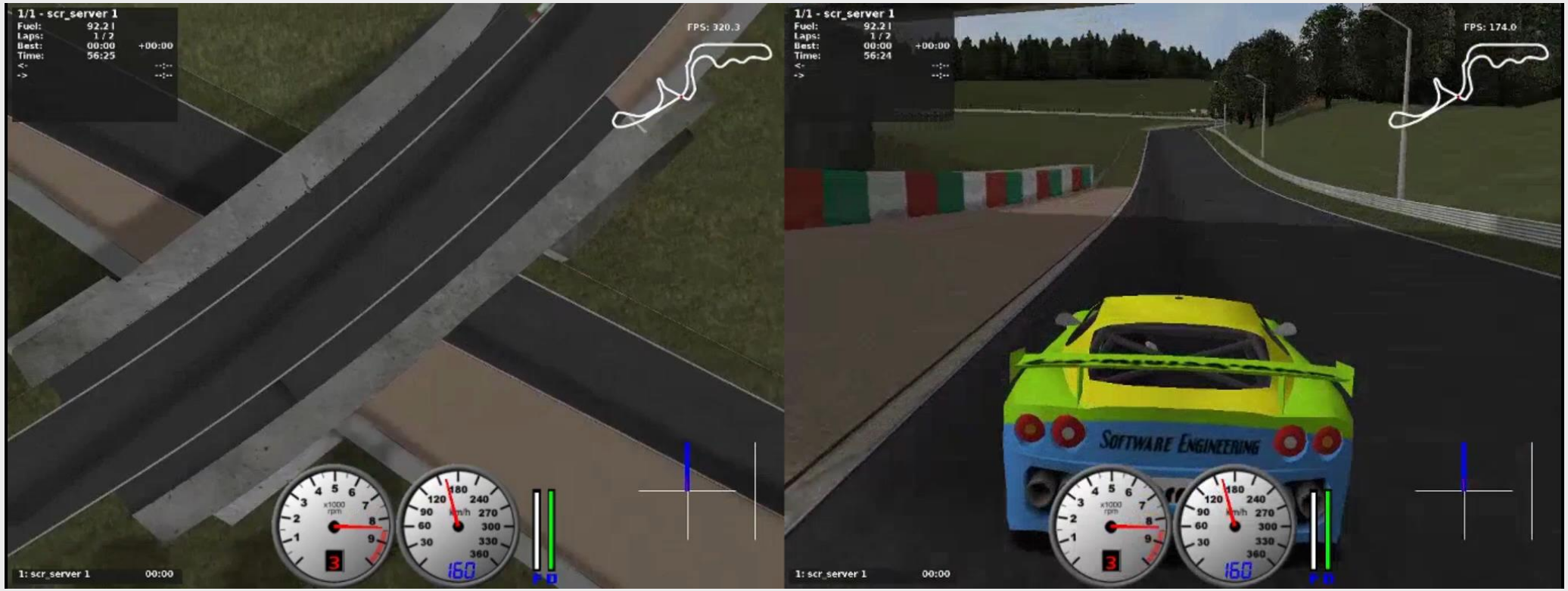
Abordagem: melhoria de um piloto já existente

SnakeOil: *framework* e piloto

Desenvolvido em Python

Modularização

DIREÇÃO AUTOMÁTICA
O PROBLEMA



CONTROLE PREDITIVO BASEADO EM MODELO

APLICANDO A PROGRAMAÇÃO KAIZEN AO PROBLEMA DE DIREÇÃO AUTOMÁTICA

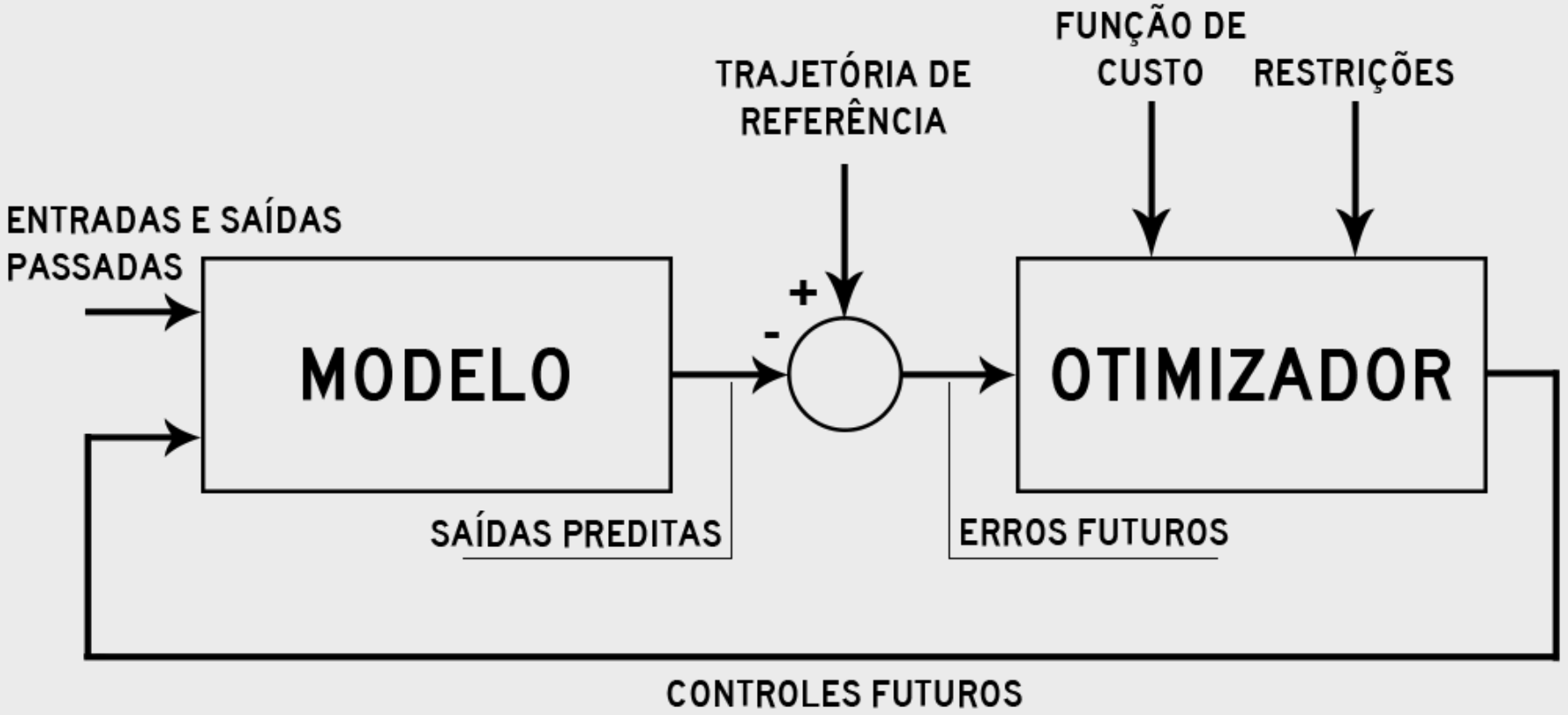
MPC
O QUE É CONTROLE?

Área de estudo de Automação

Aumento de complexidade de sistemas

- Sistemas difíceis de controlar manualmente
- Altos riscos e custos associados
- Necessidade de controle autônomo

MPC
FUNCIONAMENTO



Estrutura básica do MPC.

MPC
OTIMIZAÇÃO

Objetivos

- Maximizar a velocidade do veículo
- Minimizar a distância da referência (centro da pista)
- *Problema multiobjetivo*

MPC
APRENDIZADO

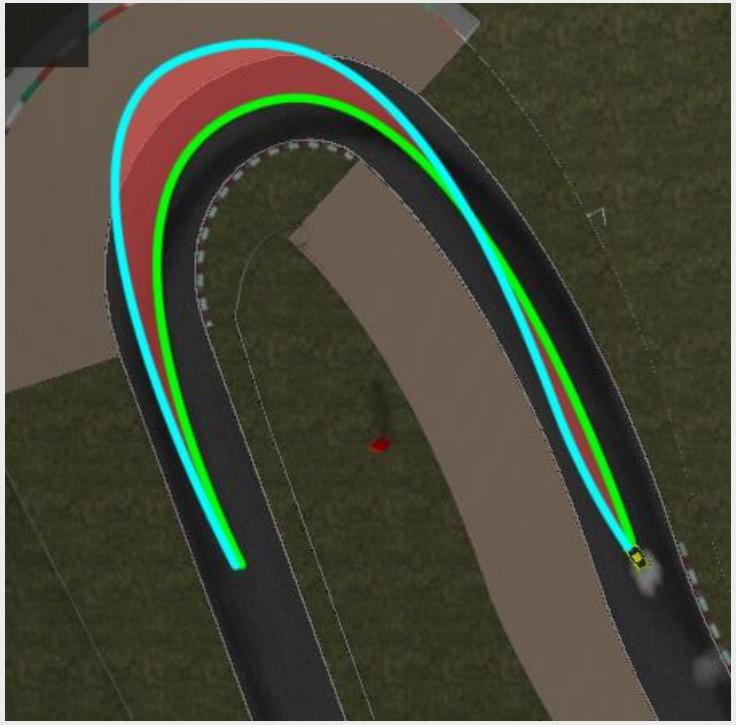
Duas etapas

Construção *offline* dos modelos

- Duas voltas de aquecimento para coleta de dados
- Encerramento do piloto para construção dos modelos
- Modelo do veículo deve prever os sensores escolhidos

Aprendizado *online* do piloto

- Processo contínuo durante a corrida
- *Ticks* de previsão
- Melhoria constante da trajetória
- Aproveitamento de previsões anteriores



Exemplo de predição de trajetória:
Verde – referência
Azul – predição
Vermelho – erro

EXPERIMENTOS COM O SIMULADOR DE CORRIDA

AVALIANDO OS MODELOS NO MPC

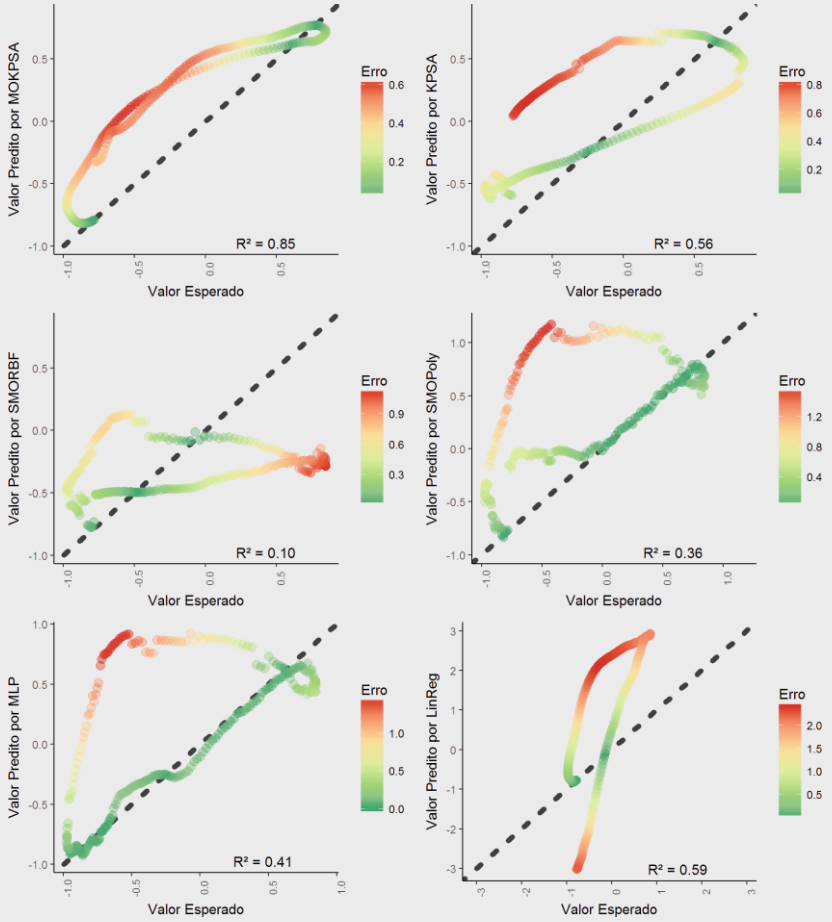
EXPERIMENTOS COM O SIMULADOR
QUALIDADE DOS MODELOS

Considerando que a pista utilizada tem 15m de largura...

MOKPSA		KPSA		SMORBF		SMOPoly		MLP		LinReg	
MAE	Real	MAE	Real	MAE	Real	MAE	Real	MAE	Real	MAE	Real
0,35	2,625m	0,5	3,75m	0,51	3,825m	0.52	3,9m	0.44	3,3m	1.43	10,725m

Estes erros são inaceitáveis.

EXPERIMENTOS COM O SIMULADOR
QUALIDADE DOS MODELOS



Gráficos de dispersão do conjunto de dados do simulador aplicado às melhores configurações específicas das técnicas executadas neste trabalho. Um modelo ideal apresentaria todos os pontos sobre a reta tracejada, onde predito=esperado.

CONCLUSÕES

ANALISANDO OS SUCESSOS E AS FALHAS E PLANEJANDO O FUTURO

CONCLUSÕES
PROGRAMAÇÃO KAIZEN MULTIOBJETIVO

Principal expectativa: modelos de qualidade ligeiramente inferior mas complexidade significativamente reduzida

- Referência: Programação Kaizen original
- Atingida

KPSA e MOKPSA apresentam bons resultados quando comparadas a outras técnicas

- Ajuste de parâmetros e de critérios de complexidade pode trazer melhorias
- Mais iterações *podem* levar a modelos mais simples e com menor correlação

KP constrói modelos lineares

- O uso de estruturas não-lineares apresenta resultados significativamente superiores
- Superou modelos não-lineares de técnicas como MLP

KP pode auxiliar outras técnicas, por construção de novos atributos e seleção de atributos existentes

CONCLUSÕES
CONTROLE PREDITIVO BASEADO EM MODELO

Apesar de a MOKPSA ter apresentado os melhores resultados, estes ainda não foram suficientes para a execução da tarefa

Há muitas causas possíveis

- Por restrições de tempo, não foi possível investigá-las extensivamente
- A metodologia e os resultados apresentados são valiosos para trabalhos futuros

CONCLUSÕES
TRABALHOS FUTUROS

Otimização multiobjetivo na modelagem

Critérios de complexidade

Comparações com aprendizado de máquina

Ajuste fino de parâmetros

Teste de KPSA em outras aplicações

Uso de outros elementos não-terminais

Métodos alternativos de mapeamento de pista

Aplicação de MPC em outros simuladores

CONCLUSÕES
PUBLICAÇÕES

"Training a Multilayer Perceptron to predict a car speed in a simulator: Comparing RPROP, PSO, BFGS and a memetic PSO-BFGS hybrid"

Apresentado no XV Simpósio Brasileiro de Jogos e Entretenimento Digital (SBGames 2016)

Um artigo com os resultados de MOKPSA nas bases de dados tradicionais da literatura está sendo elaborado e será submetido a uma revista

Obrigado!

Perguntas?

MESTRADO
PPGCC
ICT-UNIFESP
2017

Agradecimentos à CAPES pelo auxílio financeiro dado a esta pesquisa