

DESCOBERTA DE FALHAS EM MOTORES DE COMBUSTÃO ATRAVÉS DE DATA MINING VIA TÉCNICAS DE CLUSTERIZAÇÃO

Luis E. Zárate
zarate@pucminas.br

Romero Paoliello
romero@dr.com

Thiago Ribeiro
tribeiro@uai.com.br

Laboratório de Inteligência Computacional Aplicada (LICAP)
Pontifícia Universidade Católica de Minas Gerais (PUC-MG)
Av. Dom José Gaspar, 500, Coração Eucarístico
Belo Horizonte, MG, Brasil, 30535-610

Resumo. É possível perceber que sistemas de diagnósticos de falhas e correção on-line são tendências futuras na indústria automobilística. O diagnóstico de falhas nos motores de combustão, é um processo difícil de ser realizado. Dentro dessa perspectiva, técnicas automáticas são cada vez mais necessárias para analisar os dados, que são medidos diretamente do motor, e assim diagnosticar possíveis defeitos. Neste trabalho é proposta a aplicação de uma técnica Data Mining, para diagnosticar defeitos no motor de combustão a partir da emissão dos gases. Com esse objetivo, foi realizada uma coleta de dados referentes à emissão dos gases, em condições de operação normal e com defeito. Isso, com a intenção de estabelecer padrões de comportamento condizentes com as características comportamentais dos veículos.

Por outro lado, através da técnica data mining de clusterização, é possível obter grupos característicos, onde cada grupo apresente uma performance específica de funcionamento quanto à emissão dos gases. O objetivo deste projeto é aplicar a metodologia KDD (Knowledge Discovery in Databases) utilizando a técnica de clusterização de Data Mining para identificar grupos com condições de operação características que permitam auxiliar nos procedimentos de manutenção de um motor de combustão.

Palavras-chave: Data Mining, Kdd, Motores de Combustão, Diagnóstico de Falhas.

1. INTRODUÇÃO

Quando um automóvel apresenta algum tipo de problema e seu usuário leva o veículo para oficina, inicia-se um trabalhoso processo para diagnosticar o defeito do automóvel. O tempo gasto com o trabalho de diagnóstico varia de acordo com a experiência do mecânico e com o tipo de problema. Tradicionalmente, o mecânico usa de conhecimentos tácitos advindos da experiência e do treinamento para diagnosticar uma falha ou defeito no veículo. No processo avaliativo, o mecânico quase que por instinto parte de pressuposições isoladas, desconsiderando as inúmeras possibilidades de um defeito estar relacionado com mais de uma peça do motor. É improvável que o mecânico pense nas *relações* entre as inúmeras peças e fatores ligados com o defeito. A causa mais provável,

do responsável técnico, não ter conseguido visualizar o defeito, em um primeiro momento, é devido à grande quantidade de combinações existentes entre as variáveis do motor (peças e fatores) que poderiam causar aquele defeito.

Atualmente, existe um interesse no desenvolvimento de sistemas ou procedimentos capazes de auxiliar no diagnóstico de falhas de motores de combustão. É possível perceber que sistemas de diagnóstico e de correção on-line são tendências futuras na indústria automobilística e que possivelmente automóveis poderão ser fabricados com sistemas de supervisão e correção embarcados.

O KDD, Knowledge Discovery in Database, (Fayyad, 1996; Pyle, 1999; Han e Kamber 2001) é um procedimento longo, com etapas específicas, tendo como objetivo a descoberta de conhecimento, não óbvio, em banco de dados. O processo do KDD envolve etapas como: seleção, limpeza, transformação, data mining e interpretação. A literatura têm mostra um grande número de aplicações da metodologia KDD em diferentes áreas. Como exemplo, em Zárate, et al (2003) é discutido a aplicabilidade do KDD para o diagnóstico de doenças e em Oliveira, et al (2003) é apresentada a possibilidade da otimização da sequência de testes em circuitos eletrônicos.

O Data Mining é a etapa mais importante do processo KDD. Nessa são utilizados algoritmos complexos que processam o banco de dados de diferentes formas, mas sempre com o mesmo objetivo, a descoberta do conhecimento. Data Mining é uma técnica que define o processo automatizado de captura e análise de grandes conjuntos de dados, visando à descoberta de relações ocultas de grande significado. O Data Mining é composto por tecnologias de inteligência artificial (redes neurais, lógica nebulosa, algoritmos genéricos, entre outras), de estatísticas, de técnicas de modelagem de dados e de tecnologias de banco de dados. Atualmente, a tecnologia do Data Mining, está sendo usada para identificar padrões existentes nos dados e predizer comportamentos futuros. Isso permite o diagnóstico e a tomada de decisões, baseado no conhecimento extraído. Fayyad et al (1997).

Por outro lado, a clusterização (uma das técnicas do Data Mining) corresponde ao processo de agrupamento de dados que possuem características em comum. Esse processo da origem a um ou vários *clusters* ou agrupamentos, onde cada *cluster* é um conjunto de dados que possuem características similares entre si e características não-similares a dados de um outro *cluster*, Han e Kamber et al (2001). O objetivo das técnicas de clusterização são: determinar o número de classes (k) e as características de cada classe (\vec{z}_i , onde $i = 1..k$).

Diversas são as técnicas de clusterização apresentadas na literatura. Por exemplo, em Cai e Kwan et al (1998), é proposta uma técnica baseada no treinamento de redes neurais e na extração de regras nebulosa para classificar as instâncias. Nesse trabalho as funções de pertinência são propostas pelo especialista e isso tende a descaracterizar a procura natural dos agrupamentos (*clusters*). Em Abe et al (1998) é proposta uma técnica para a geração automática de clusters através de regiões elipsoidais e em Zait e Messafa (1997) é apresentado uma comparação de métodos de clusterização baseado na qualidade dos resultados obtidos através de medidas de similaridade como a distância euclidiana.

A utilização da clusterização pode mostrar a distribuição de padrões e correlações interessantes entre os dados. A análise de *cluster* pode ser usada como ferramenta de suporte para observar as características de cada cluster, e para focar em um grupo particular uma análise mais profunda (descoberta de conhecimento).

A aplicação da metodologia do KDD, em processos específicos de oficinas mecânicas pode trazer benefícios aos métodos de manutenção dos veículos automotivos. Como exemplo, através da rápida comparação entre os inúmeros dados relacionados à emissão de gases, parte do processo de manutenção dos veículos pode ser automatizada, com ganhos em eficiência e precisão. O funcionamento dessa automatização, começaria com a instalação de um dispositivo no veículo que coletasse dados da emissão dos gases e a execução de um processamento interno de análise dos dados. Assim, seria possível diagnosticar o tipo de defeito a partir da identificação do padrão obtido com os padrões já identificados pela técnica de clusterização.

Neste trabalho, é mostrado uma aplicação da metodologia KDD para conhecer sua potencialidade na análise de dados e diagnóstico de falhas na indústria automobilística. Neste trabalho, será utilizada a técnica de clusterização para descobrir grupos de comportamento dos gases, ou poluentes, emitidos pelos veículos automotores em condições de funcionamento normal e em condições de funcionamento com defeitos. Assim, através da clusterização dos dados de emissão de gases do motor, identificaremos os regimes de funcionamento que apresentam características comuns quanto à emissão dos gases. A partir daí, poderemos então saber qual o comportamento da emissão de gases do motor em cada situação de defeito específico facilitando assim o diagnóstico das falhas.

Este trabalho é dividido em 5 seções. Na seção 2, é descrito o banco de dados do motor de combustão considerado neste trabalho. Na seção 3, a metodologia do KDD é aplicada, e nas últimas seções os resultados e as conclusões do trabalho são apresentados.

2. BANCO DE DADOS DO MOTOR DE COMBUSTÃO

A definição do espaço problema (Pyle, 1997 e Jones, 1998), a ser investigado, refere-se à descoberta de padrões no funcionamento em um motor 1.3, 16v. O banco de dados considerado é composto por dados referentes à emissão de gases e de dados relativos aos regimentos de funcionamento do motor. A coleta de dados referentes à emissão dos gases do motor foi realizada em condições de operação normal e em condições de defeito. O banco de dados inicial contém 22 atributos e mais de 15 mil registros, o qual foi reduzido a um total de 184 registros após a etapa de pré-processamento, como descrito posteriormente. Na Tabela (1) são apresentadas os atributos considerados.

Tabela 1. Atributos considerados na coleta de dados

Carga (kgf)	Rotação (rpm)	Torque (N.m)
Potência (KW)	PotênciaMédiaEfetiva (MPa)	ConsumoEspecífico
FrequênciaVazão (Hz)	VazãoVolumétrica (l/h)	VazãoMássica (kg/h)
TemperaturaAmbiente (nao.cal)	Termopar 1 (nao.cal)	Termopar 2 (nao.cal)
Termopar 3 (nao.cal)	Termopar 4 (nao.cal)	abertura borboleta (%)
COc (CO corrigido)	Diluição HC	CO ₂
O ₂	CO	Lambda
AFR		

A seleção dos atributos relevantes ao problema é uma tarefa difícil. Esses atributos podem ser *fatos* e *julgamentos* e quanto maior o número de julgamentos maior será a imprecisão do conhecimento descoberto, Abe et al (1998). Portanto, deve ser analisado cada atributo de forma individual e em conjunto. Por exemplo, no nosso caso, todas as variáveis possuem um caráter de exatidão, ou seja, todas as variáveis são fatos, consequentemente não foi necessária a eliminação de nenhuma variável.

De acordo com a metodologia de KDD, Fayyad et al (1996), o banco de dados original precisa passar por um processo de enriquecimento, limpeza e transformação, antes de serem submetidos aos algoritmos de data mining. Na próxima seção a etapa de preparação dos dados, a técnica do data mining de clusterização, bem como a coleta de dados serão descritas.

2.1. Coleta de Dados

A coleta dos dados foi realizada na bancada de testes do laboratório de engenharia automotiva da PUC-MG utilizando um motor 1.3 16v. Nessa coleta, foram realizadas simulações de funcionamento com rotações variando entre 800rpm e 5000rpm. Um dinamômetro foi utilizado para simular condições diferentes de carga (peso) para diferentes condições de rotação. Na prática,

a simulação realizada seria equivalente a dirigir o automóvel com a rotação variando entre 800 e 5000rpm e em diferentes condições de lotação (peso). Através da bancada, várias simulações de condição de funcionamento do veículo foram realizadas em intervalos contínuos de 200 e de 500rpm.

Os dados foram coletados instantaneamente e dinamicamente. Os dados coletados dinamicamente foram os dados de emissão de gases. Os dados coletados instantaneamente estão relacionados com consumo, temperatura, torque, vazão, entre outros, e dizem respeito as características específicas do funcionamento do veículo em dado instante. O mesmo procedimento acima realizado no veículo em regime normal, foi realizado também no veículo com algumas simulações de defeito, são elas: sem a sonda lambda, sem o bico da injeção de número 1, sem o bico da injeção de número 2 e sem os bicos da injeção de números 1 e 2. O Banco de Dados gerado possui dados do veículo em condições normais e anormais, com condições de carga variando entre 0,14 e 15,5 e a rotação variando entre 800 e 5000rpm como dito anteriormente.

A seguir são mostrados dados da coleta como exemplificação:

Tabela 2. Exemplos dos dados coletados em funcionamento normal

Carga (kgf)	Rotação (rpm)	Torque (N.m)	Potência (KW)	Potencia Média Efetiva(MPa)	Consumo Específico	Freqüência Vazão(Hz)	Vazão Volumétrica (l/h)	Vazão Mássica (kg/h)	Temperatura Ambiente (nao.cal)
0,28	1042	1,94	0,21	0,02	5,29	13,19	1,48	1,12	26,33
8,81	1002	61,51	6,43	0,59	0,58	43,98	4,92	3,74	26,83
5,33	1027	37,32	4,01	0,36	0,66	31,10	3,48	2,64	26,32
9,19	1008	64,35	6,79	0,62	0,50	40,11	4,49	3,41	26,31
8,42	1008	58,92	6,22	0,57	0,61	44,33	4,96	3,77	26,67

Termo-par 1 (nao.cal)	Termo-par 2 (nao.cal)	Termo-par 3 (nao.cal)	Termo-par 4 (nao.cal)	abertura borboleta (%)	COc	Diluição	HC	CO ₂	O ₂	CO	Lambada	AFR
6,16	5,79	4,48	4,84	0	2,14	14,70	234	12,60	1,46	2,10	0,99	12,80
5,68	6,20	4,35	4,87	100	2,37	15,47	140	13,10	0,90	2,37	0,96	12,46
5,88	6,19	4,29	4,75	25	2,43	15,03	232	12,60	1,44	2,43	0,98	12,63
5,94	6,72	4,26	4,56	50	2,29	15,39	214	13,10	1,27	2,29	0,98	12,62
5,90	5,60	4,63	4,28	75	1,57	15,17	139	13,60	0,87	1,57	0,99	12,74

Os dados da Tabela 2, são condizentes ao regime de funcionamento normal, ou seja, sem a simulação de defeitos. O número total de registros do banco de dados da coleta foi de 15630, por esse motivo é apresentado apenas um exemplo dos mesmos.

A Tabela 3 mostra os valores máximos e mínimos de rotação e carga atingidos em cada teste realizado na coleta de dados.

Tabela 3. Valores máximos e mínimos de rotação e carga para cada teste

	sem bico de injeção 1		sem bico de injeção 2		sem bico de injeção 1 e 2		funcionamento normal	
	Rotação	Carga	Rotação	Carga	Rotação	Carga	Rotação	Carga
Mínimo	990	0,41	1007	2,09	893	0,15	1002	0,28
Máximo	5000	12,17	4995	9,95	4369	7,05	5000	15,37

3. APLICAÇÃO DO PROCESSO KDD

Existem diversas metodologias KDD, embora todos tratam das mesmas etapas necessárias para realização do processo. Em linhas gerais as etapas do KDD envolvem: a seleção do banco de dados, o pré-processamento, o data mining, a interpretação e a visualização. Cada etapa citada envolve

outros passos que dependem do problema específico sendo tratado. A Fig. (1) mostra as etapas concretizadas neste trabalho.

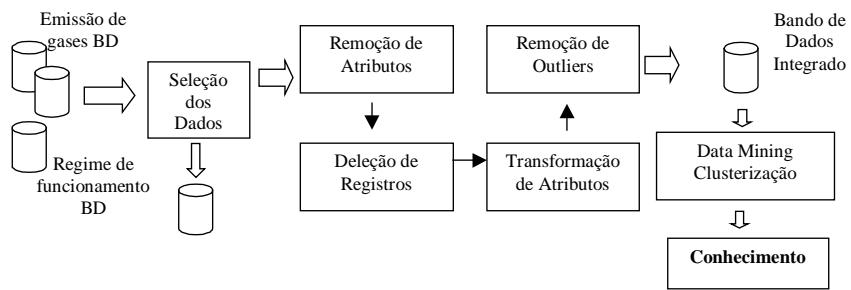


Figura 1. Metodologia KDD aplicada

Na continuação serão descritas as etapas aplicadas ao banco de dados do motor de combustão apresentado na seção 2.

3.1. Pré-processamento do Banco de Dados

3.1.1. Seleção de Registros

Os critérios adotados para seleção de registros de dados, foram baseados na análise do banco de dados e do problema em questão:

- Os dados relativos ao regime transitório de funcionamento são desconsiderados para a realização de uma análise exclusiva e precisa dos gases em estado estacionário.

3.1.2. Eliminação de Registros

O critério adotado para a exclusão de registros do banco de dados foram os seguintes:

- Foram determinados como irrelevantes os registros com valores negativos, provenientes dos dispositivos de medição. Esses dados são desconsiderados tendo em vista a impossibilidade desses existirem.
- Através do cálculo do desvio padrão de cada variável de acordo com os regimes de funcionamento do motor foi possível identificar visualmente e remover os registros que se mostravam muito fora de três desvios padrões (outliers).
- Através do cálculo da média dos dados relativos as emissões de gases, foram eliminados registros que se apresentavam com valores de rotação com um valor 5% superior ou inferior a média do regime de funcionamento a que ele pertencia.

3.1.3. Transformação dos Atributos

As seguintes transformações foram aplicadas na base de dados:

- Os dados foram arredondados de forma a garantir uma precisão de até duas casas decimais.
- Para a realização da análise de forma apropriada, os dados foram normalizados de forma a estabelecer uma unidade padrão para todas as grandezas (Hz, Km, etc) presentes em nosso banco de dados. Essa transformação é realizada devido a característica peculiar da clusterização, que procura estabelecer distâncias entre cada variável sem poder por tanto, ser tendenciosa.

3.2. Data Mining

Nessa etapa, o objetivo é identificar grupos de funcionamento do motor com relação à suas respectivas emissões de gases. Para conseguir esse tipo de classificação, utiliza-se a técnica de clusterização. Com essa técnica os dados são separados em grupos, onde os dados com comportamentos similares pertencem a um mesmo grupo e dados dissimilares pertencem a grupos diferentes.

Em primeira instância, para a clusterização, utilizou-se o algoritmo de *Kohonen*, baseado em redes neurais, e que não foi satisfatório para a análise, devido a que resultados diferentes eram encontrados de acordo com a variação dos pontos centrais iniciais. Neste trabalho foi utilizado o algoritmo *K-means*, o qual apresentou resultados satisfatórios. A continuação uma breve descrição do algoritmo *k-means* será apresentada.

Análise de Cluster k-means:

Como já mencionado, “*clustering*” é o processo de agrupar Instâncias/dados que possuem características em comum. Esse processo dá origem a um ou vários *clusters*. Um *cluster* é um conjunto de Instância/dados que possuem características similares entre si e características não-similares a Instância/dados de um outro *cluster* (Han e Kamber, 2001 e Fayyad e Stolorz, 1997). O objetivo das técnicas de *cluster* são: determinar o número de classes (k) e as características de cada classe (\vec{z}_i , onde $i = 1..k$)

Em geral os dados a serem classificados podem ser representados por uma matriz ($N \times p$) de medições:

$$X = \begin{vmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N1} & \cdots & x_{Nf} & \cdots & x_{Np} \end{vmatrix} \quad (1)$$

onde: N é o número de instâncias e p é o número de medições de cada instância.

Por exemplo, para o presente trabalho, tem-se:

$$\vec{x}_i = \{x_{i,j}, \forall j = 1, \dots, p \wedge x_{i,1} = \text{emissão de HC}; x_{i,2} = \text{emissão de CO}_2; \dots\} \quad (2)$$

Para a Matriz X o objetivo é encontrar o número de agrupamentos (k) e as características de cada classe (\vec{z}_i , onde $i = 1..k$) onde \vec{z}_i representa o centro de um agrupamento no espaço euclidiano.

Conhecido os vetores dos diferentes agrupamentos (ou clusters), o objetivo principal deste trabalho, é obter conhecimento a partir dos vetores \vec{z}_i .

Como \vec{z}_i é um vetor n -dimensional $\vec{z}_i = [z_{i1}, z_{i2}, \dots, z_{ip}]$, é possível identificar os limites das instâncias \vec{x}_j que pertencem a esse agrupamento:

$$\vec{x}_j^{\min} = [x_{j1}^{\min}, x_{j2}^{\min}, \dots, x_{jp}^{\min}] \quad (3)$$

$$\vec{x}_j^{\max} = [x_{j1}^{\max}, x_{j2}^{\max}, \dots, x_{jp}^{\max}] \quad (4)$$

Os valores mínimos e máximos de \vec{x}_j podem ser determinados através da distância euclidiana entre \vec{x}_j , para $j = 1, \dots, N$ e \vec{z}_i . Os vetores \vec{x}_j que pertencem ao agrupamento \vec{z}_i correspondem a aqueles que tiveram uma distância euclidiana abaixo de uma distância mínima estabelecida. Em outras palavras:

$$\vec{x}_j \in z_i, \text{ para } j = 1, \dots, N \quad \text{se} \quad \|x - z\| = \left[\sum_{i=1}^n (x_i - z_i)^2 \right]^{1/2} \leq dist_{\min} \quad (5)$$

A partir das Eqs. (3) e (4) é possível extrair regras do tipo:

$$\text{Se } (x_{j1}^{\min} \leq x_{r2} \leq x_{j1}^{\max}) \text{ e } (x_{j2}^{\min} \leq x_{r2} \leq x_{j2}^{\max}) \text{ e } \dots \dots (x_{jp}^{\min} \leq x_{rp} \leq x_{jp}^{\max}) \text{ Então } \vec{x}_r \text{ é similar a } \vec{z}_i \quad (6)$$

Observe que as regras podem ser utilizadas para identificar o grupo mais apropriado para uma nova instância \vec{x}_r (obtida através de medição dos parâmetros do motor). Note que as técnicas de clusterização permitem agrupar dados referentes a funcionamento normal ou não do motor de combustão.

O algoritmo K-means utiliza o critério da mínima distância para agrupar os dados. Para entender melhor este algoritmo consideremos primeiramente a norma euclideana de um vetor $x = [x_1, x_2, \dots, x_n]^T$ definida como segue:

$$\|x\| = \left[\sum_{i=1}^n x_i^2 \right]^{1/2} \quad (8)$$

Para cálculo da distância entre dois vetores:

$$\|x - z\| = \left[\sum_{i=1}^p (x_i - z_i)^2 \right]^{1/2} \quad (9)$$

Considerando o conjunto de instância: $\{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$. O vetor z , representa o centro do cluster para cada um dos k -clusters, sendo definido como: z_1, z_2, \dots, z_k . Com $S_j = \{x^{(i)} / x^{(i)} \text{ é próximo do Cluster } j\}$ representa todas as instâncias que ficaram próximas do cluster j .

Para o banco de dados, já pré-processado, o melhor número de cluster que encontramos após a análise foi 4. Esses grupos são mostrados graficamente e literalmente a seguir.

A Figura (2) mostra os dados de emissão de gases (HC, CO, CO₂) agrupados, onde registros com a mesma cor pertencem a um mesmo cluster.

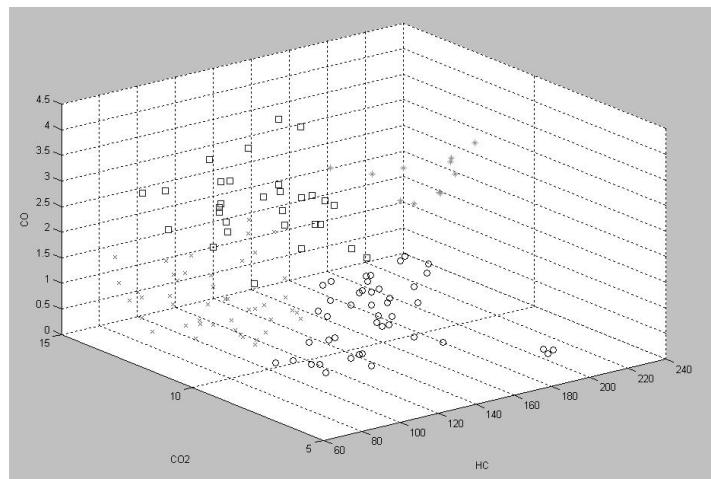


Figura 2. Dados de emissões de gases (HC, CO, CO₂) clusterizados

A Figura (3) mostra os limites (intervalos) de cada cluster encontrado. Através desses intervalos conseguimos saber as características desses grupos e conseguimos mostrar de forma literal o comportamento de cada regime de funcionamento do motor.

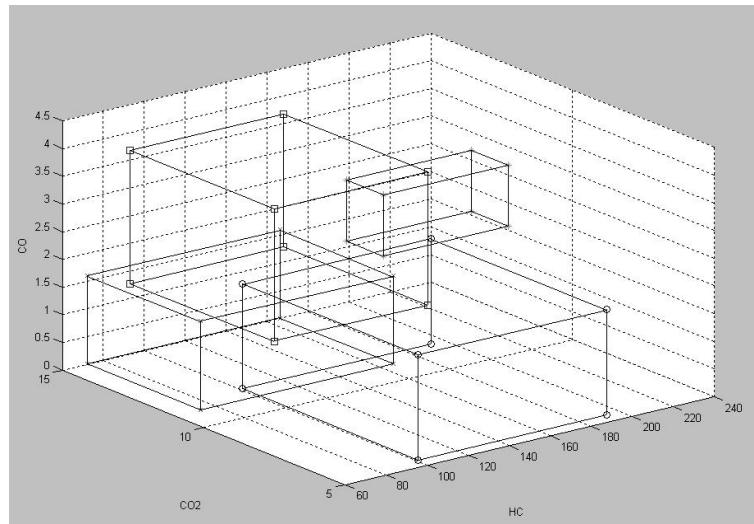


Figura 3. Limites dos clusters

Após a identificação dos clusters e seus respectivos registros conseguimos informações para criação de duas tabelas que são mostradas a seguir:

Tabela 4. Representação dos clusters

CLUSTER 1			CLUSTER 2			CLUSTER 3			CLUSTER 4		
	Quant.	%		Quant.	%		Quant.	%		Quant.	%
sem bico 1 e 2	22	50	sem bico 1 e 2	0	0	sem bico 1 e 2	0	0	sem bico 1 e 2	3	10
sem bico 1	8	18,2	sem bico 1	16	35,5	sem bico 1	0	0	sem bico 1	9	30
sem bico 2	14	31,8	sem bico 2	15	33,3	sem bico 2	0	0	sem bico 2	8	26,7
normal	0	0	normal	14	31,2	normal	11	100	normal	10	33,3

Tabela 5. Valores dos Intervalos dos Clusters

CLUSTER 1		CLUSTER 2		CLUSTER 3		CLUSTER 4	
HC	97 – 189	HC	64 – 158	HC	178 – 239	HC	75 – 150
CO	0,1 – 2,0	CO	0,2 – 1,8	CO	1,6 – 2,7	CO	1,7 – 4,1
CO ₂	5,1 – 11,3	CO ₂	10,4 – 14,4	CO ₂	12,2 – 13,5	CO ₂	8,6 – 13,7

4. RESULTADOS

Os resultados da aplicação do processo de KDD dependem da análise e execução correta de todas as etapas. A etapa de Data Mining é dentre todas a que menos demanda tempo, porém é também a mais importante para que os resultados obtidos possam ser resultados mensuráveis para o problema.

Como finalização do processo de descoberta de conhecimento, os conhecimentos extraídos na etapa de Data Mining devem ser analisados e apresentados de forma a poderem ser entendidos por quem vai utilizá-la para tomar decisões.

As Tabelas (4) e (5), apresentadas como conhecimento extraído na etapa de Data Mining, analisadas devidamente nos leva as seguintes regras:

- Se valores de emissão de HC estão entre 97 e 189, CO entre 0,1 e 2,0, CO₂ entre 5,1 e 11,3; então o motor tem 100% de chances de estar com defeito, sendo que 50% de chances desse defeito estar nos bicos de injeção 1 e 2 simultaneamente, 18,2% de chances de estar no bico 1 e 31,8% de chances de estar no bico de injeção 2.
- Se valores de emissão de HC estão entre 64 e 158, CO entre 0,2 e 1,8, CO₂ entre 10,4 e 14,4; então o motor tem 68,8% de chances de estar com defeito, sendo que não há chances desse defeito estar nos bicos de injeção 1 e 2 simultaneamente, 35,5% de chances haver defeito no bico 1 e 33,3% de chances de estar no bico de injeção 2.
- Se valores de emissão de HC estão entre 178 e 239, CO entre 1,6 e 2,7, CO₂ entre 12,2 e 13,5; então o motor não possui defeitos quanto aos bicos de injeção.
- Se valores de emissão de HC estão entre 75 e 150, CO entre 1,7 e 4,1, CO₂ entre 8,6 e 13,7; então o motor tem 66,7% de chances de estar com defeito, sendo que 10% de chances desse defeito estar nos bicos de injeção 1 e 2 simultaneamente, 30% de chances de estar no bico 1 e 26,7% de chances de estar no bico de injeção 2.

5. CONCLUSÕES

Neste trabalho, foi aplicada a metodologia KDD para a descoberta de conhecimento em um banco de dados de motor de combustão. O diagnóstico de defeitos nos motores de veículos é um trabalho difícil. Por essa razão, pretendeu-se aqui, propor a possibilidade do uso de KDD para diagnosticar falhas nesses motores a partir de suas respectivas emissões de gases. É importante ressaltar que outras abordagens, tais como simulações de outros defeitos e a inclusão de outros gases, para o processo de diagnóstico de falhas nesses motores são válidas.

Foram mostradas as quatro fases necessárias do KDD: a primeira fase correspondeu à coleta de dados, a segunda fase esteve relacionada ao pré-processamento dos dados, a terceira à aplicação do algoritmo de clusterização e a última fase à caracterização do comportamento dos gases, baseada na observação dos resultados apresentados pela clusterização.

Com o aperfeiçoamento da técnica e a extensão do esquema proposto para outros tipos de defeitos, bem como para outros tipos de marcas e modelos automotivos, esse esquema pode servir para automatização do processo de diagnóstico de falhas em motores de combustão de forma viável tecnologicamente e financeiramente. A adoção dessa técnica nas oficinas mecânicas levaria a uma maior eficiência e a redução do tempo gasto no processo de diagnóstico de falhas em motores. As oficinas, caso adotassem essa solução, poderiam gozar de um patamar mais eficiente e eficaz, automatizando parte dos problemas enfrentados cotidianamente.

6. REFERÊNCIAS

- Abe, S. "Dynamic cluster generation for a fuzzy classifier with ellipsoidal Regions", IEEE Trans. On Systems, Man and Cybernetics, Part B, Vol. 28, N° 6, Dec. 1998
- Cai L.Y. e Kwan H.K. "Fuzzy classifications using fuzzy inference network", IEEE Trans. On Systems, Man and Cybernetics, Part B, Vol. 28, N° 3, Jun., 1998
- Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. 1996. Advances in Knowledge discovery and data mining. Menlo Park, CA: AAAI Press/MIT Press.
- Fayyad U. e Stolorz P. "Data Mining and KDD: promise and challenges". Future Generation Computer Systems, 13 pp. 99-115, Elsevier Science, 1997.
- Han, J.W., M. Kamber 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, Inc. San Diego, California.
- Jones, M.D. 14 Powerful Techniques for Problem Solving. Times Books Random House Inc. 1998
- Oliveira, C.G., Zárate, L.E. Vieira, N.J. "Applying Data Mining on Fault Sequencing in Electronic Circuits via Automata"CSITeA2003/ISCA-International Society for Computers and their Applications, 2003, Rio de Janeiro, v1. P. 320-324.
- Pyle, D. 1999. Data Preparation for Data Mining. Morgan Kaufmann Publishers, Inc. San Francisco, California.
- Zait M. e Messaffa H. "A comparativr study of clustering method ", Future Generation Computer Systems, 13 pp. 149-159, Elsevier Science, 1997.
- Zárate, L.E., Alvarenga, P. Paoliello, R. e Ribeiro, T. "Data Mining Applied to the Discovery of Symptom Patterns in Data ase with Nephrolithiasis"AIA2003/IASTED, Benalmādena, Espanhā, v.1, p. 234-239.

KNOWLEDGE DISCOVERY APPLIED TO ENGINES OF VEHICLES THROUGH DATA MINING WITH CLUSTERING TECHNIQUE

Luis E. Zárate & Romero Paoliello & Thiago Ribeiro

zарате@pucminas.br

Applied Computational Intelligence Laboratory (LICAP)

Pontifical Catholic University of Minas Gerais (PUC)

Av. Dom José Gaspar, 500, Coração Eucarístico

Belo Horizonte, MG, Brasil, 30535-610

Abstract. It is possible to notice that systems of flaws diagnoses and on-line correction are future tendencies in the automobile industry. The flaws diagnosis, in combustion motors, is difficult to be detected, and inside of this perspective, automatic techniques are more and more necessary to analyze the data, that are directly measured from the motor, making possible defects diagnose. In this work is proposed the application of techniques of Data Mining to diagnose defects in the combustion motor starting from the emission of gases. With this objective, a data collection was accomplished referring from the gas emission, in conditions of normal operation and with defect. These simulations were made with the intention of establishing patterns of behavior in agreement with the characteristics of the vehicles. Through the clustering technique is possible to obtain characteristic groups, where each group presents a specific performance of operation in agreement with the emission of the gases. The objective of this project is to apply the KDD (Knowledge Discovery in Databases) methodology using the clustering technique of Data Mining to identify groups with characteristic operation conditions that aids in the procedures of maintenance of a combustion motor.

Keywords: Data Mining, Kdd, Motors of Combustion, Diagnosis of Flaws.