

ANÁLISE DE DADOS APLICADA À PRODUÇÃO: Um estudo de caso na Indústria de Equipamentos Elétricos

Orlando Roque da Silva - UNIP-SP, Brasil
orlandoroque@uol.com.br

Angelo Palmisano - UNIVAG-MT, Brasil
angelopalmisano@uol.com.br

Delvio Venanzi - FATEC-Sorocaba-SP, Brasil
delviosuper@uol.com.br

RESUMO

Um desafio muito comum da aplicação de data analytics na fabricação é a aquisição de dados relevantes. Dada a natureza dinâmica, incerta e complexa específica dos sistemas de manufatura, os algoritmos de ML oferecem a oportunidade de aprender com o sistema dinâmico e adaptar-se ao ambiente em mudança automaticamente até certo ponto. O objetivo desta pesquisa é avaliar o impacto que as tecnologias associadas à indústria 4.0, como big data, machine learning, internet das coisas, rede de sensores, etc., tem sobre as análises dos dados de manufatura. Inicialmente é feito uma avaliação das principais tecnologias e por fim é apresentado um estudo de caso na indústria de equipamentos eletrônicos. O método utilizado foi pesquisa bibliográfica e análise de um caso real. Os resultados estão relatados passo-a-passo na seção de análise. Conclui-se que há um grande caminho para avançar nas pesquisas e nas práticas empresariais, no que se refere às aplicações de data analytics no ambiente de manufatura em vistas de ter novas e inovadoras alternativas.

Palavras-chave: Indústria 4.0, Data Analytics, Machine Learning.

1. Introdução à Análise de Dados

Até recentemente, os pesquisadores que trabalhavam com análise de dados lutavam para obter dados para seus experimentos. Avanços recentes na tecnologia de processamento, armazenamento e transmissão de dados, associados a softwares avançados e inteligentes, reduzindo custos e aumentando a capacidade, mudaram esse cenário. É o tempo da Internet das Coisas, onde o objetivo é ter tudo ou quase tudo conectado. Os dados produzidos anteriormente em papel estão agora online. A cada dia, uma quantidade maior de dados é gerada e consumida. Sempre que você coloca um comentário em sua rede social, carrega uma imagem, música ou vídeo, navega pela Internet ou adiciona um comentário a um site de comércio eletrônico, você está contribuindo para o aumento dos dados. Além disso, máquinas, transações financeiras e sensores, como câmeras de segurança, estão cada vez mais coletando dados de fontes muito diversas e generalizadas.

Em 2012, estimou-se que, a cada ano, a quantidade de dados disponíveis no mundo dobra (Gantz, J. & Reinsel, D., 2012). Outra estimativa, previa que a partir de 2014 até 2020 todas as informações seriam digitalizadas, eliminadas ou reinventadas em 80% dos processos e produtos em relação à década anterior (Laney, D. & White, A., 2014). Em um terceiro relatório, a partir de 2015, foi previsto que o tráfego de dados móveis seria quase 10 vezes maior em 2020 (Cisco Inc., 2016). O resultado de todos esses rápidos aumentos de dados é chamado por alguns de "explosão de dados".

Apesar da impressão que isso pode dar - de que estamos nos afogando em dados - há vários benefícios em ter acesso a todos esses dados. Esses dados fornecem uma fonte rica de informações que podem ser transformadas em conhecimento novo, útil, válido e compreensível pelo homem. Assim, há um interesse crescente em explorar esses dados para extrair esse conhecimento, usando-os para apoiar a tomada de decisões em uma ampla variedade de campos: agricultura, comércio, educação, meio ambiente, finanças, governo, indústria, medicina, transporte e assistência social. Várias empresas em todo o mundo estão percebendo a mina de ouro que possuem e o potencial desses dados para apoiar seu trabalho, reduzir o desperdício e atividades de trabalho perigosas e tediosas e aumentar o valor de seus produtos e seus lucros.

A análise desses dados para extrair esse conhecimento é objeto de uma área vibrante conhecida como análise de dados, ou simplesmente "análise". Você pode encontrar várias definições de análise na literatura. A definição adotada aqui é: a análise é a ciência que analisa dados brutos para extrair conhecimento útil (padrões) deles. Esse processo também pode incluir coleta, organização, pré-processamento, transformação, modelagem e interpretação de dados.

A análise como uma área de conhecimento envolve informações de diversas áreas. A ideia de generalizar o conhecimento a partir de uma amostra de dados vem de um ramo da estatística conhecido como aprendizado indutivo, uma área de pesquisa com uma longa história.

Com os avanços dos computadores pessoais, o uso de recursos computacionais para resolver problemas de aprendizado indutivo se torna cada vez mais popular. A capacidade computacional tem sido usada para desenvolver novos métodos. Ao mesmo tempo, novos problemas surgiram, exigindo um bom conhecimento das ciências da computação. Por exemplo, a capacidade de executar uma determinada tarefa com mais eficiência computacional tornou-se objeto de estudo para pessoas que trabalham em estatística computacional.

Paralelamente, vários pesquisadores sonhavam em reproduzir o comportamento humano usando computadores. Eram pessoas da área de inteligência artificial. Eles também usaram estatísticas para suas pesquisas, mas a ideia de reproduzir o comportamento humano e biológico em computadores era uma fonte importante de motivação. Por exemplo, a reprodução de como o cérebro humano trabalha com redes neurais artificiais tem sido estudada desde a década de 1940. O estudo de como as formigas trabalham com o algoritmo de otimização de colônias vem desde os anos 90. O termo aprendizado de máquina (*Machine Learning*) apareceu nesse contexto como o campo de estudo que dá aos computadores a capacidade de aprender sem serem explicitamente programados (Simon, P., 2013).

Nos anos 90, um novo termo apareceu com um significado diferente: mineração de dados (*Data Mining*). A década de 1990 foi a década do surgimento das ferramentas de inteligência de negócios como consequência das instalações de dados terem capacidade maior e tornarem-se mais baratas. As empresas começam a coletar cada vez mais dados, com o objetivo de resolver ou melhorar as operações

comerciais, por exemplo, detectando fraudes com cartões de crédito, aconselhando o público sobre as restrições da malha rodoviária nas cidades ou melhorando o relacionamento com os clientes, utilizando técnicas mais eficientes de relacionamento. A questão era poder extrair os dados para extrair o conhecimento necessário para uma determinada tarefa. Esse é o objetivo da mineração de dados.

Big data e a Ciência dos Dados

Nos primeiros anos do século XX, o termo *big data* apareceu. O *big data*, uma tecnologia para processamento de dados, foi inicialmente definido pelos "três V's", embora mais V's tenham sido propostos desde então. Os três primeiros Vs permitem definir uma taxonomia de *big data*. São eles: volume, variedade e velocidade. O volume preocupa-se com o como armazenar grandes quantidade de dados, ou seja, como criar repositórios de dados para grandes quantidades de dados. A variedade está preocupada em como reunir dados de diferentes fontes. A velocidade diz respeito à capacidade de lidar com dados que chegam muito rapidamente, em fluxos conhecidos como fluxos de dados.

Outro termo que apareceu e às vezes é usado como sinônimo de *big data* é ciência de dados. Segundo Provost e Fawcett (2013), *big data* são conjuntos de dados muito grandes para serem gerenciados por tecnologias convencionais de processamento de dados, exigindo o desenvolvimento de novas técnicas e ferramentas para armazenamento, processamento e transmissão de dados. Essas ferramentas incluem, por exemplo, MapReduce, Hadoop, Spark e Storm. Mas o volume de dados não é a única caracterização de *big data*. A palavra "grande" pode se referir ao número de fontes de dados, à importância dos dados, à necessidade de novas técnicas de processamento, à rapidez com que os dados chegam, à combinação de diferentes conjuntos de dados para que possam ser analisados em tempo real e sua onipresença, uma vez que qualquer empresa, organização ou indivíduo sem fins lucrativos tem acesso aos dados agora.

Assim, o *big data* está mais preocupado com a tecnologia. Ele fornece um ambiente de computação, não apenas para análise, mas também para outras tarefas de processamento de dados. Essas tarefas incluem processamento de transações financeiras, processamento de dados da web e processamento de dados georreferenciados.

A ciência de dados preocupa-se com a criação de modelos capazes de extrair padrões de dados complexos e o uso desses modelos em problemas da vida real. A ciência de dados extrai conhecimento significativo e útil dos dados, com o suporte de tecnologias adequadas. Ela tem uma estreita relação com análises e mineração de dados. A ciência de dados vai além da mineração de dados, fornecendo uma estrutura de extração de conhecimento, incluindo estatísticas e visualização.

Portanto, embora o *big data* dê suporte à coleta e gerenciamento de dados, a ciência de dados aplica técnicas a esses dados para descobrir conhecimento novo e útil: o *big data* coleta e a ciência de dados descobre. Outros termos como descoberta ou extração de conhecimento, reconhecimento de padrões, análise de dados, engenharia de dados e vários outros também são usados. A definição que usamos da análise de dados abrange todas essas áreas usadas para extrair conhecimento dos dados.

2. Análise de Dados em Projetos Six Sigma

A quantidade de dados gerados e armazenados aumenta todos os dias, assim como a capacidade dos computadores para processar esses dados. A vantagem competitiva pode ser obtida usando esses dados efetivamente (Manyika et al., 2011). Grandes conjuntos de dados são usados para vários objetivos nas organizações, entre eles está a sua aplicabilidade na melhoria das operações.

Uma variedade de metodologias para melhoria das operações foi proposta, como gerenciamento da qualidade total, reengenharia de processos de negócios, gerenciamento lean, teoria das restrições, Six Sigma e Lean Six Sigma. Aqui vamos focar no último: Lean Six Sigma, uma vez que foi provado um método amplamente utilizado e bem-sucedido (Hahn et al., 1999; Pande et al., 2000). O Lean Six Sigma (LSS) consiste em uma estrutura metodológica bem estabelecida para melhorar a eficiência operacional e a eficácia nas organizações (George, 2003). É conhecido principalmente por causa de sua abordagem passo a passo da melhoria, chamada DMAIC, que é um acrônimo para Definir, Medir, Analisar, Melhorar e Controlar. A abordagem é muito parecida com a abordagem científica para a solução de problemas. Como metodologia completa, também estabelece como uma cultura de melhoria contínua eficaz e duradoura pode ser realizada. Por exemplo, fornece diretrizes sobre seleção de projetos, gerenciamento de projetos e implantação.

Os projetos de LSS são orientados por dados: em um primeiro estágio, os dados do processo são coletados e o desempenho da linha de base é quantificado. Em um segundo estágio, fatores de influência vitais são selecionados e seus efeitos quantificados, levando a ações de melhoria baseadas em evidências. Tradicionalmente, os dados coletados em um projeto LSS consistem em algumas características do processo que devem ser aprimoradas (Y) e em alguns fatores de influência (X). Para um projeto bem-sucedido, normalmente são coletadas de 30 a 1000 observações.

Nas últimas décadas, volumes maiores de dados de uma ampla variedade de métricas de processo tornaram-se prontamente disponíveis como resultado de rápidos desenvolvimentos em TI, automação, sensores e mídia de armazenamento. Os conjuntos de dados tradicionais estão cada vez menores em comparação com o conjunto de dados típico que está disponível atualmente. Os métodos estatísticos tradicionais, normalmente usados em projetos de LSS, como o teste t e a regressão linear, são menos eficazes em conjuntos de dados maiores. A análise de grandes conjuntos de dados requer ferramentas diferentes, conhecidas como ferramentas de 'ciência de dados'.

A maioria das pesquisas no campo do LSS se concentra em ferramentas específicas ou técnicas analíticas usadas na abordagem (por exemplo, gráficos de controle, design de experimentos) ou em fatores de sucesso no desenvolvimento de uma melhoria na organização (Lameijer et al., 2016). Algumas idéias iniciais foram compartilhadas na literatura sobre a integração de Lean e Six Sigma, por um lado, e *big data* e ciência de dados, por outro (Dhawan et al., 2014, Auschitzky et al., 2014, Dutcher, 2014).

Ainda pouco se sabe como integrar os dois campos. Isso foi reconhecido por Antony et al. (2017) de que existe uma grande oportunidade para o LSS usar *big data* e ciência de dados. Portanto focamos na capacidade da metodologia de se adaptar a esse desenvolvimento. Vamos colocar o termo LSS de ciência de dados para o Lean Six Sigma que gira em torno de grandes conjuntos de dados (*big data*) e aplicar ferramentas de ciência de dados que estão além dos padrões do Lean Six Sigma.

Antes de discutir a sinergia do LSS e da ciência de dados, é importante entender o que são. Nas duas subseções a seguir, fornecemos alguns

anteriores.

Lean Six Sigma

O LSS é uma estrutura metodológica para estabelecer a melhoria contínua dentro das organizações (De Mast et al., 2016). Embora tenha começado na indústria, tornou-se cada vez mais popular em outros setores, como saúde (De Koning et al., 2006), finanças (De Koning et al., 2008; Delgado, 2010) e outros (Lameijer et 2016).

Para o estudo de LSS usamos a definição da metodologia de Schroeder et al. (2008). A definição é baseada em quatro elementos:

- Existe uma estrutura meso-paralelo, como mostrado na figura 1.
- Um especialista em melhoria lidera o projeto.
- Uma metodologia estruturada é usada, prescrita por um roteiro chamado DMAIC.
- O objetivo é melhorar as métricas de desempenho dentro da organização.

Uma estrutura meso-paralela ajuda a alcançar a melhoria das operações (Raje, 2007), porque a estrutura da organização é usada como modelo para a organização de projetos de melhoria.

A figura 1, baseada em Schroeder et al. (2008) e Akkerhuis et al. (2015), visualiza uma estrutura meso-paralela.



Figura 1: Mesoestrutura paralela em uma organização LSS.

No centro da estrutura organizacional (paralela) do LSS estão os Green e Black Belts (veja a Figura 1). Esses profissionais recebem treinamento extensivo em uma ampla variedade de tópicos e, portanto, são especialistas em melhoria, responsáveis pela execução do projeto. Todo projeto tem um campeão: o proprietário do problema que é responsável pelo projeto. Evidentemente, nem todos os problemas têm a mesma prioridade; portanto, o gerenciamento de programas no nível da alta administração determina a estratégia e define quais problemas devem ser enfrentados primeiro. Finalmente, os green belts estão apoiando projetos de melhoria e são frequentemente encontrados na área operacional (De Mast et al., 2016).

O LSS oferece um método estruturado para melhoria de processos: o ciclo DMAIC, usado pelos líderes do projeto como um roteiro para o sucesso. Suas fases estão resumidas na Figura 2 e Tabela 1. O roteiro prescreve uma trajetória sequencial do projeto na qual todas as fases devem ser discutidas com o líder antes que o projeto possa continuar na próxima fase.

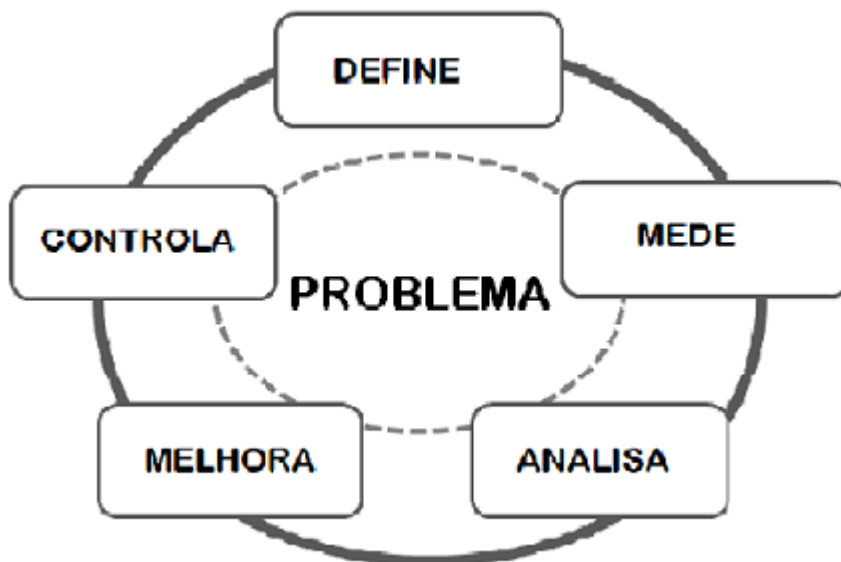


Figura 2: Fases do DMAIC e suas relações para um projeto de melhoria do LSS.

| | |
|-----------------|------------------------------------------------------------------------------|
| Define | Selecione o projeto e o líder do projeto e estabeleça objetivos e condições. |
| Mede | Tornar o problema quantificável e mensurável. |
| Analisa | Analisar e diagnosticar a situação atual. |
| Melhora | Desenvolver e implementar ações de melhoria. |
| Controla | Controle o desempenho aprimorado do processo e feche o projeto. |

Tabela 1: Fases DMAIC de um projeto LSS (De Mast et al., 2016)

Nestas fases do DMAIC, os dados são essenciais. Sem dados, a importância do projeto não pode ser estabelecida e demonstrada e, mais importante, a eficácia das ações de melhoria não pode ser comprovada. No entanto, a quantidade de dados geralmente é limitada, já que os líderes do projeto geralmente coletam dados manualmente para garantir a confiabilidade. Como consequência, trinta observações das métricas do processo são geralmente consideradas suficientes para obter uma boa indicação do desempenho. Como consequência, a complexidade das ferramentas estatísticas é relativamente baixa e a caixa de ferramentas é limitada a ferramentas básicas, como análise de variância e regressão linear simples, e a métodos gráficos, como histogramas e gráficos de Pareto.

Ciência de Dados: definição e estruturas de projetos

Tivemos várias inovações em estatística e modelagem, o que resultou em uma nova terminologia para trabalhar com dados. Como a análise de dados, que tenta fornecer informações extraindo informações dos dados. Nesse sentido, a análise (estatística) faz parte da análise de dados. A análise de negócios (Bartlett, 2013) concentra-se nas idéias que os dados podem fornecer para que a empresa tome melhores decisões. A mineração de dados está focada na aplicação de algoritmos específicos para extrair padrões dos dados (Fayyad et al., 1996).

Finalmente, o *big data* refere-se a técnicas de mineração de dados relacionadas a dados que aderem aos 3 Vs: velocidade, volume e variedade e, às vezes, também ao quarto V (veracidade) (Megahed e Farmer, 2015).

Também se refere à prática física de armazenamento de dados (Sagiroglu e Sinanc, 2013). Um exemplo desse tipo de dado é o registro de cliques em sites, onde a quantidade de informações liberadas não pode ser processada pela maioria das mídias de armazenamento. Concluímos que todos esses termos relacionados a

dados se enquadram no título de ciência de dados. Azevedo e Santos (2008) fornecem uma comparação de vários padrões da indústria para execução de projetos de ciência de dados. A partir da comparação, pode se concluir que o Processo Padrão Intersetorial para Mineração de Dados (CRISP-DM), é uma metodologia amplamente usada e mais abrangente. O CRISP-DM é uma metodologia composta por várias fases que fornecem estrutura para projetos de mineração de dados. Observe que existem loops entre algumas das 6 fases, que são distintas das 5 fases seqüenciais prescritas pelo LSS.

Diferenças entre o LSS de ciência de dados e o LSS tradicional

Estrutura meso-paralela: inclui cientistas de dados

A equipe original, denominada equipe OpEx, ocupa um lugar na organização desde 2007 e estava funcionando no nível gerencial; a gerência sênior atribui tarefas e projetos à equipe do OpEx. Desde 2014, a organização emprega cientistas de dados ao tomar conhecimento da importância dos dados. A partir de então, há uma hierarquia fraca na equipe OpEx, onde os cientistas de dados dão suporte aos Black Belts nos projetos. Durante a execução de um projeto, eles vão para a área de trabalho e solicitam a ajuda e o envolvimento da gerência e do pessoal da linha para análise e implementação de melhorias.

Portanto, concluímos que a hierarquia da equipe é paralela à de toda a organização, onde um cientista de dados trabalha para o Black Belt dentro da equipe de melhoria, ambos são gerenciados pela gerência sênior.

Especialistas em Melhoria: Cientistas de dados gerenciados por Black Belts

Os projetos são liderados por um ou mais Black Belts em período integral. São funcionários especializados e receberam treinamento no método DMAIC, gerenciamento de projetos e política organizacional. Esses Black Belts são especialistas em melhorias reais.

Os cientistas de dados não estão apenas envolvidos na melhoria do processo, mas são usados em toda a organização. Portanto, eles não são especialistas em melhoria em tempo integral. E isso não é necessário, pois eles são gerenciados pelos Black Belts. O trabalho do Black Belt consiste em traduzir questões de negócios em perguntas de pesquisa para os cientistas de dados. Isso

não parece fácil pois os antecedentes dos líderes do projeto e cientistas de dados podem ser bem diferentes.

O cientista de dados é necessário para os projetos de melhorias porque possui habilidades essenciais para lidar com dados, especialmente em lidar com a manipulação de dados e habilidades para lidar com grandes volumes de dados, a aplicação de vários algoritmos e o ajuste e a melhoria iterativos dos modelos de dados são habilidades importantes que um cientista de dados tem em contraste com um profissional de Lean Six Sigma.

Diferenciação entre LSS tradicional e LSS de ciência de dados

Como estamos diferenciando os projetos tradicionais de LSS e LSS com ciência de dados, é útil uma diretriz para selecionar a metodologia apropriada. Para fazer isso, precisamos definir diferenciais entre esses dois tipos de projetos. Voltamos para Hoerl & Snee (2013), que distinguem entre dois eixos em projetos de inovação de processos; ou seja, se uma solução para o processo ou problema é conhecida ou desconhecida e se a complexidade do problema em questão é alta ou baixa.

Para diferenciar entre os projetos LSS e ciência de dados, um terceiro eixo deve ser adicionado: a disponibilidade dos dados. No caso de baixos volumes de dados em mãos, a estrutura tradicional do projeto DMAIC deve ser seguida. No entanto, se altos volumes de dados estiverem disponíveis e devem ser utilizados para o projeto, o projeto deve ser executado de acordo com as fases iterativas do DMAIC e um cientista de dados deve ser incluído na equipe do projeto. Finalmente, observamos que, se a complexidade é baixa, a solução é desconhecida e há muitos dados disponíveis que uma análise de dados provavelmente levaria a uma solução para o problema.

3. Análise de Dados na Programação da Produção

É cada vez mais crescente o uso de Modelagem e Simulação (M&S) para avanços na produção. O uso de M&S foi identificado como uma das principais etapas para alcançar a produção inteligente. Um relatório sobre o avanço da manufatura avançada nos Estados Unidos mostra que é exigido alta fidelidade à M&S para reduzir o projeto ao *lead time* de fabricação e controle e otimização avançados. O relatório também aponta maior uso da análise de dados para avançar

na fabricação e observa links entre análise e simulação.

O *Data Analytics* (DA) foi identificado como a chave para uma maior agilidade para reagir rapidamente às flutuações na demanda ou nos suprimentos do mercado, bem como no controle da produção.

Os aplicativos DA para manufatura, referidos como *Manufacturing Data Analytics* (MDA), podem ser usados para ajudar a melhorar o desempenho da manufatura por meio de informações sobre tendências, padrões, áreas de ineficiência e riscos potenciais para os fabricantes. Os dados para essas análises estão cada vez mais disponíveis com o avanço da tecnologia, incluindo sensores em máquinas e equipamentos, leitores para etiquetas e códigos de barras de identificação por radiofrequência e aplicativos de coleta de dados que rastreiam informações em várias fontes, incluindo transações financeiras, comportamento do mercado e internet.

Atualmente, a maioria das empresas de manufatura não faz bom uso de todos os dados gerados e coletados, como dados de projeto assistido por computador, manufatura auxiliada por computador e sistemas digitais de manufatura.

4. Sistemas Ciberfísicos

O conceito recente de "Indústria 4.0" ou a quarta revolução industrial inclui os Sistemas Ciberfísicos (CPS) como um componente-chave. A função do CPS foi identificada como monitoramento de processos físicos e criação de uma cópia virtual do mundo físico para apoiar a tomada de decisão descentralizada. A "cópia virtual" e os "modelos de planta virtual" discutidos no contexto da Indústria 4.0 correspondem ao conceito de fábrica virtual. A fábrica virtual pode suportar o estágio de projeto da fábrica e, uma vez construída, pode fazer a transição para um componente do CPS e dar suporte ao MDA e à tomada de decisões.

A necessidade identificada na literatura para o MDA baseado em simulação pode ser atendida com habilidade usando as representações da fábrica virtual, conforme sugerido pelos aplicativos relatados. Os aplicativos atuais geralmente utilizam desenvolvimentos personalizados de fábricas virtuais que exigem um grande esforço e conhecimento. O desenvolvimento de uma capacidade de gerar fábricas virtuais em grande parte automaticamente, usando dados de fábricas reais em formatos padrão, conforme proposto neste documento, reduzirá significativamente

os requisitos de esforço e conhecimento.

Concepção e desenvolvimento de conceitos

Esta seção discute os fatores que suportam o desenvolvimento proposto para aumentar os aplicativos de M&S e MDA para manufatura, seguidos pela apresentação do projeto conceitual da fábrica virtual. A primeira subseção estabelece o vínculo entre M&S e MDA. A segunda subseção identifica os recentes avanços na tecnologia de M&S e nos padrões de interface. A terceira subseção apresenta o conceito de fábrica virtual que se baseia em avanços em tecnologia e padrões e visa permitir o uso rápido de M&S e MDA.

Funções de simulação para MDA

Os fabricantes usam a M&S para analisar o projeto e as operações há muito tempo. O aplicativo de simulação inclui análise dos dados de saída para gerar insights e, portanto, a própria simulação é uma ferramenta para o DA. Várias ferramentas de análise estatística estão disponíveis para análise dos dados de entrada da simulação e essas ferramentas também são aplicações DA. Assim, a simulação é suportada por aplicativos DA. A simulação pode ser usada para gerar dados realistas para dar suporte à avaliação de aplicativos de DA e para preencher dados ausentes para uso pelo DA. A simulação suporta, portanto, aplicativos DA. As funções de simulação para MDA podem ser agrupadas em duas categorias, como um aplicativo MDA e como um aplicativo de suporte para outros aplicativos MDA.

O Modelo de Ascendência Analítica do Gartner define quatro aplicações principais do DA, como mostrado na figura 3. Elas incluem descritivo (o que aconteceu?), diagnóstico (por que aconteceu?), preditivo (o que provavelmente acontecerá?) e prescritivo (como podemos fazer isso acontecer)? análises. Uma série de técnicas pode ser usada para as quatro principais aplicações, incluindo mineração de dados, regressão, análise de rede bayesiana, algoritmos de classificação e agrupamento, algoritmos de aprendizado de máquina, otimização e análise de envelopamento de dados. A simulação é útil para três das áreas, a saber, diagnóstico, preditivo e prescritivo, conforme discutido abaixo.

Simulação como um aplicativo de análise de diagnóstico

A análise de sensibilidade dos modelos de simulação permite identificar os fatores que influenciam a medida ou os padrões de interesse. Essa análise pode ser usada para análise de diagnóstico. Por exemplo, pode-se investigar as causas de ciclos de tempos mais longos para um determinado grupo de produtos, variando os fatores identificados como os potenciais contribuintes usando um experimento projetado. Os resultados dessa análise podem ajudar a responder à pergunta de por que o grupo de produtos está passando por longos tempos de ciclo.

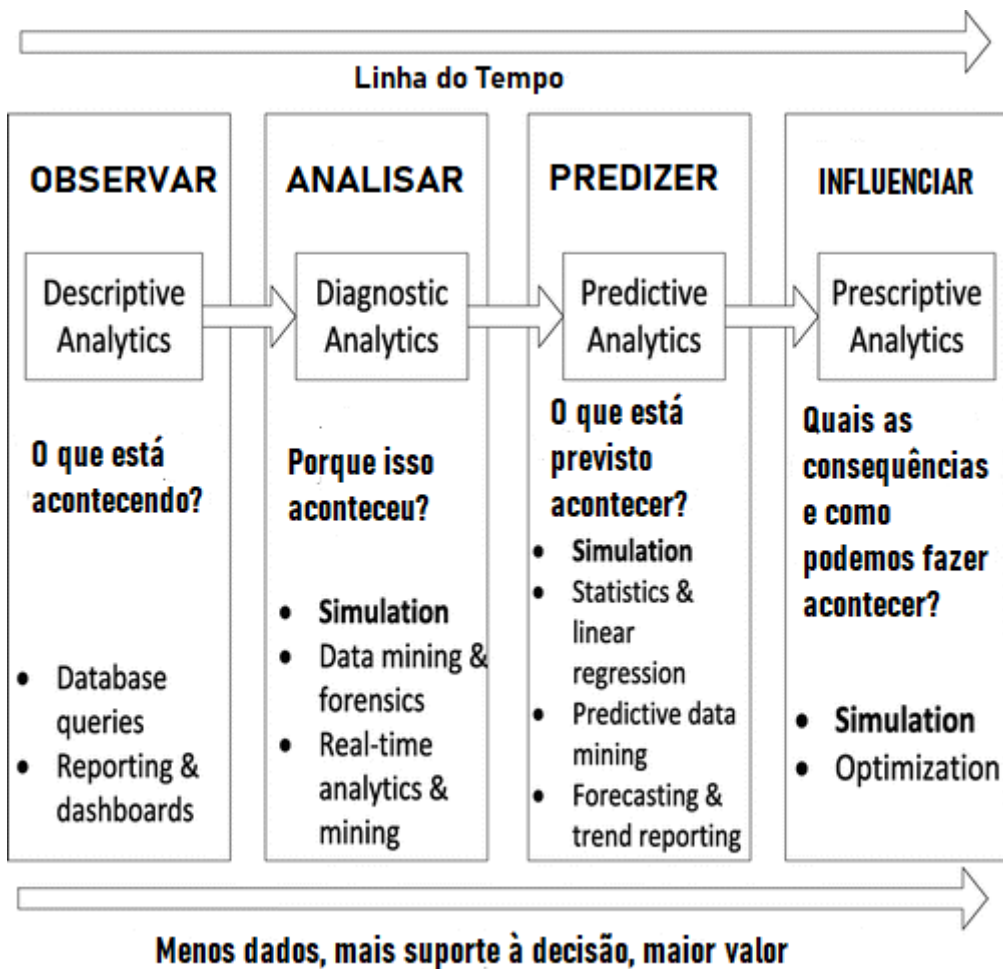


Figura 3. Papel da simulação nas principais aplicações de análise de dados.

Simulação como um aplicativo de análise preditiva

A análise preditiva é necessária para entender o impacto nas medidas de desempenho de interesse de futuras mudanças planejadas e não planejadas, como políticas, mix de produtos, disponibilidade de recursos e demandas. A simulação pode ser usada para avaliar efetivamente esse impacto por meio da modelagem de cenários de mudanças planejadas e não planejadas em um sistema de fabricação. Assim, pode responder à pergunta sobre o que provavelmente acontecerá com as incertezas associadas.

Simulação como aplicativo de análise prescritiva

A análise prescritiva vai além da análise preditiva e identifica as configurações e políticas de parâmetros necessárias que resultarão nas melhorias de desempenho desejadas. A simulação pode contribuir de três maneiras para o papel de um aplicativo de análise prescritiva. Primeiro, os modelos de simulação podem ser exercidos através de vários cenários configurados em um experimento projetado para identificar as configurações de parâmetros que melhoram as medições de desempenho de interesse. Por exemplo, parâmetros como taxas de liberação de lote, regras de envio e recursos por operação podem variar para identificar o conjunto que leva a um desempenho aprimorado da data de vencimento.

Segundo, a simulação pode ser usada para obter a saída de outro aplicativo de análise prescritiva, como uma ferramenta de otimização, e ajustá-la para implementação. Geralmente, é altamente complicado e às vezes não é possível incluir todas as restrições e variabilidades da vida real em um modelo de otimização. Por exemplo, modelos de otimização podem ser usados para representar as principais restrições, como número de recursos principais (máquinas, operadores) e precedências técnicas para gerar um cronograma de fabricação. A simulação pode ser usada para ajustar a solução usando as restrições da vida real restantes, como espaços de buffer e transportadores.

Terceiro, os modelos de simulação podem ser conectados com um procedimento de otimização, como programação matemática, otimização combinatória ou algoritmo genético, para pesquisar iterativamente o conjunto de configurações de parâmetros que fornecem o melhor desempenho possível. A terceira abordagem é geralmente referida como uma abordagem combinada de otimização de simulação. Além disso, a combinação de simulação com otimização, conforme definido na segunda e terceira abordagens, geralmente fornece soluções mais precisas do que um aplicativo de otimização independente.

As tecnologias relevantes para o conceito proposto de fábrica virtual vêm se desenvolvendo rapidamente nos últimos anos. Esta subseção discute brevemente tecnologias e padrões que tornaram possível a implementação do conceito de fábrica virtual.

5. Aplicando Machine Learning na Produção

Atualmente, a indústria de transformação está experimentando um aumento nunca visto nos dados disponíveis (Chand & Davis, 2010). Esses dados apresentam uma variedade de diferentes formatos, semântica, qualidade, por exemplo, dados do sensor da linha de produção, dados ambientais, parâmetros de máquinas-ferramenta, etc. (Davis et al., 2015). A disponibilidade dos dados relacionados à qualidade, por exemplo, oferece potencial para melhorar a qualidade do processo e do produto de maneira sustentável e é uma evidência da importância que o aumento e disponibilidade de dados oferecem à produção (Elangovan, Sakthivel, Saravanamurugan, Nair e Sugumaran, 2015). No entanto, foi reconhecido que muitas informações também podem propor um desafio e podem ter um impacto negativo, por exemplo, negligenciar as principais questões/causalidades ou levar a conclusões atrasadas ou erradas sobre as ações tomadas (Lang, 2007). Em geral, pode-se concluir com segurança, que a indústria de transformação precisa aceitar que, para se beneficiar do aumento da disponibilidade de dados, por exemplo, para iniciativas de melhoria da qualidade, estimativa de custos de fabricação e/ou otimização de processos, melhor entendimento dos requisitos do cliente etc. No entanto é necessário suporte para lidar com a alta dimensionalidade, complexidade e dinâmica envolvidas (Davis et al., 2015; Loyer, Henriques, Fontul, & Wiseall, 2016;

Wuest, 2015).

Novos desenvolvimentos em certos domínios, como matemática e ciência da computação (por exemplo, aprendizado estatístico) e disponibilidade de ferramentas (software) fáceis de usar, frequentemente disponíveis gratuitamente, oferecem grande potencial para transformar o domínio de fabricação e sua compreensão sobre o aumento dos repositórios de dados de fabricação de maneira sustentável. Um dos desenvolvimentos mais empolgantes é na área de aprendizado de máquina (incluindo mineração de dados (DM), inteligência artificial (AI), descoberta de conhecimento (KD) em bancos de dados, etc.). No entanto, o campo de aprendizado de máquina é muito diversificado e muitos algoritmos, teorias e métodos diferentes estão disponíveis. Para muitos profissionais de manufatura, isso representa uma barreira quanto à adoção dessas ferramentas poderosas e, portanto, pode dificultar a utilização de vastas quantidades de dados cada vez mais disponíveis. De acordo com isso, é de se esperar:

- Pela perspectiva da fabricação por que o aprendizado de máquina é uma ferramenta apropriada e promissora para os desafios atuais e futuros;
- Que se introduza a terminologia usada nos respectivos campos;
- Que se apresente uma visão geral das diferentes áreas do aprendizado de máquina e propor uma estrutura geral;
- Que se forneça ao leitor uma compreensão de alto nível das vantagens e desvantagens de certos métodos com relação à aplicação de fabricação.

Os desafios atuais da fabricação nos fornecem uma base para argumentar que o aprendizado de máquina é uma ferramenta apropriada para os fabricantes enfrentarem esses desafios de frente.

Desafios do domínio manufatureiro

A manufatura é uma indústria muito estabelecida, porém sua importância não pode ser classificada como alta o suficiente. Várias economias maduras experimentaram uma redução da contribuição da manufatura para seu PIB nas últimas décadas. No entanto, nos últimos anos, várias iniciativas para renovar o setor manufatureiro foram iniciadas. Exemplos são os EUA por meio de "Ações

executivas para fortalecer a manufatura avançada na América" (Casa Branca, 2014) e a União Europeia com sua iniciativa "Fábricas do futuro" (Comissão Europeia, 2016). Os desafios que a manufatura enfrenta hoje são diferentes dos desafios do passado.

Existem vários estudos disponíveis, propondo os principais desafios da fabricação em nível global. Os principais desafios com os quais a maioria dos pesquisadores concorda (Dingli, 2012; Gordon e Sohal, 2001; Shiang e Nagaraj, 2011; Thomas, Byard e Evans, 2012) são os seguintes:

- Adoção de tecnologias avançadas de fabricação.
- Crescente importância da fabricação de produtos de alto valor agregado.
- Utilizando conhecimentos avançados, gerenciamento de informações e sistemas de IA.
- Manufatura (processos) e produtos sustentáveis.
- Recursos corporativos e cadeias de fornecimento ágeis e flexíveis.
- Inovação em produtos, serviços e processos.
- Estreita colaboração entre indústria e pesquisa para adotar novas tecnologias.
- Novos paradigmas de gerenciamento de fabricação.

Esses principais desafios destacam a tendência contínua do domínio da fabricação de se tornar mais complexa e dinâmica. A aparente complexidade é herdada não apenas nos próprios programas de fabricação, mas cada vez mais no produto a ser fabricado, bem como nos processos (de negócios) das empresas e redes de colaboração (Wiendahl & Scholtissek, 1994). Além do desafio, está o fato de o ambiente de negócios dinâmico das empresas manufatureiras de hoje ser afetado pela incerteza (Monostori, 2003). Analisando especialmente os domínios com maior probabilidade de otimização, por exemplo, monitoramento e controle, programação e diagnóstico, torna-se aparente que a crescente disponibilidade de dados está adicionando outro desafio: além das grandes quantidades de dados disponíveis (por exemplo, dados de sensores), a alta dimensionalidade e variedade dos dados (por exemplo, devido a diferentes sensores ou processos conectados) e da natureza de problemas de otimização da manufatura que sejam do tipo NP Hard,

esse conjunto de condições representam um grande desafio.

Para superar alguns dos principais desafios atuais de sistemas de fabricação complexos, os candidatos válidos são técnicas de aprendizado de máquina. Essas abordagens orientadas a dados são capazes de encontrar padrões altamente complexos e não lineares em dados de diferentes tipos e fontes e transformar dados brutos em espaços de recursos, os chamados modelos, que são então aplicados para previsão, detecção, classificação, regressão ou previsão.

A seguir, primeiro são ilustradas as principais vantagens e desafios das aplicações de aprendizado de máquina em relação à fabricação, seus desafios e requisitos. Em seguida, é apresentado o estado da arte atual do aprendizado de máquina, novamente com foco nas aplicações de fabricação. Nesse contexto, é desenvolvida e apresentada uma estruturação de diferentes técnicas e algoritmos de aprendizado de máquina.

Adequação do aplicativo de aprendizado de máquina com relação aos desafios atuais de fabricação

Antes de analisar a adequação do aprendizado de máquina (ML) com base nos requisitos derivados anteriormente na direção de uma abordagem de solução futura, os termos usados são introduzidos brevemente. O ML é conhecido por sua capacidade de lidar com muitos problemas de natureza NP completa, que frequentemente aparecem no domínio da fabricação inteligente (Monostori, Hornyák, Egresits e Viharos, 1998).

A aplicação de técnicas de ML aumentou nas últimas duas décadas devido a vários fatores, por exemplo a disponibilidade de grandes quantidades de dados complexos com pouca transparência (Smola & Vishwanathan, 2008) e o aumento da usabilidade e poder das ferramentas de ML disponíveis (Larose, 2005). No entanto, a principal definição de ML, permitindo que os computadores resolvam problemas sem serem especificamente programados para fazê-lo (Samuel, 1959), ainda é válida hoje. ML está conectado a outros termos, como DM, KD, AI e outros (Alpaydin, 2010). Hoje, o ML já é amplamente aplicado em diferentes áreas de fabricação, por exemplo otimização, controle e solução de problemas (Alpaydin,

2010; Pham & Afify, 2005).

Muitas técnicas de ML (por exemplo, Support Vector Machine [SVM]) são projetadas para analisar grandes quantidades de dados e capazes de lidar com alta dimensionalidade (> 1000) muito bem (Yang & Trewn, 2004). No entanto, questões que acompanham como possível excesso de ajuste devem ser consideradas (Widodo & Yang, 2007) durante a aplicação. Se a dimensionalidade provar ser um problema, apesar de improvável devido ao poder dos algoritmos, existem métodos disponíveis para reduzir as dimensões. Eles pretendem reduzir o impacto da redução da dimensionalidade nos resultados esperados (Kotsiantis, 2007; Manning, Raghavan, & Schütze, 2009).

A importância do uso do ML, neste caso, o SVM é que a dimensionalidade não é um problema prático e, portanto, a necessidade de reduzir a dimensionalidade é reduzida. Isso implica na possibilidade de ser mais liberal ao incluir informações aparentemente irrelevantes disponíveis nos dados de fabricação que podem vir a ser relevantes em determinadas circunstâncias. Isso pode ter um efeito direto na lacuna de conhecimento existente descrita anteriormente (Alpaydin, 2010; Pham & Afify, 2005).

A aplicação de ML na fabricação pode resultar em um padrão derivado de conjuntos de dados existentes, o que pode fornecer uma base para o desenvolvimento de aproximações sobre o comportamento futuro do sistema (Alpaydin, 2010; Nilsson, 2005). Essas novas informações (conhecimento) podem apoiar os proprietários do processo na tomada de decisões ou ser usadas automaticamente para melhorar o sistema diretamente. No final, o objetivo de certas técnicas de ML é detectar certos padrões ou regularidades que descrevem as relações (Alpaydin, 2010).

Dado o desafio de um ambiente de fabricação dinâmico e em rápida mudança, o ML, fazendo parte da IA e herdando a capacidade de aprender e se adaptar às mudanças permite que o projetista do sistema não precise prever e fornecer soluções para todas as situações possíveis (Alpaydin, 2010). Portanto, o ML fornece um forte argumento do por que sua aplicação na fabricação pode ser benéfica, dada a luta da maioria dos modelos para lidar com a adaptabilidade. Aprender e adaptar-se a ambientes em mudança automaticamente é uma das

principais forças do ML (Lu, 1990; Simon, 1983).

As técnicas de ML são projetadas para derivar conhecimento dos dados existentes (Alpaydin, 2010; Kwak & Kim, 2012). Alpaydin (2010) enfatiza que os dados armazenados só são úteis quando analisados e transformados em informações que podemos usar, por exemplo, para fazer previsões (Alpaydin, 2010). Isso é especialmente verdadeiro para a manufatura, dada a dificuldade de obter dados em tempo real durante um programa de produção ativo, executado com as restrições técnicas, financeiras e de conhecimento. Isso também pode ter um impacto na questão do posicionamento dos pontos de verificação do processo (Wuest, Liu, Lu & Thoben, 2014).

Visto que faz sentido selecionar cuidadosamente pontos de verificação sob a perspectiva de quais dados são úteis e quais não são. Isso pode resultar na capacidade de determinar mais pontos de capturar de dados, ao longo de todo o programa de fabricação. Se isso é benéfico é uma questão em aberto, que precisa ser pesquisada. Dada a capacidade do ML de lidar com dados de alta dimensionalidade, o lado técnico da análise de dados adicionais não oferece nenhum problema. No entanto, em termos de captura de dados, ainda pode ser um problema, especificamente a capacidade de capturar os dados. Uma vez que os dados estão disponíveis, a determinação de fatores de estado em situações de alta dimensionalidade não é considerada problemática, nem é repetida com frequência.

Na tabela a seguir, é apresentado um resumo da capacidade teórica das técnicas de ML para responder aos principais desafios das aplicações de fabricação (requisitos).

No geral, como Monostori, Márkus, Van Brussel e Westkämper (1996) enfatizam, a inteligência está fortemente ligada ao aprendizado, e a capacidade de aprendizado deve ser um recurso indispensável dos Sistemas Inteligentes de Fabricação. ML fornece fortes argumentos quando se trata de limitações e desafia o conceito teórico do estado do produto. Dada a análise acima mencionada, as técnicas de ML parecem fornecer uma solução promissora com base nos requisitos derivados. A maioria dos requisitos identificados é tratada com sucesso pelo ML.

Tabela 2. Adequação das técnicas de ML na aplicação de fabricação.

| REQUISITO DE FABRICAÇÃO | CAPACIDADE TEÓRICA DO ML EM ATENDER AOS REQUISITOS |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Capacidade de lidar com problemas de alta dimensão e conjuntos de dados com esforço razoável. | Certas técnicas de ML (por exemplo, SVM) são capazes de lidar com alta dimensionalidade (> 1000) muito bem. No entanto, questões que acompanham como possível excesso de ajuste devem ser consideradas (Widodo & Yang, 2007; Yang & Trewn, 2004) |
| Capacidade de reduzir a natureza possivelmente complexa dos resultados e apresentar conselhos transparentes e concretos para os profissionais (por exemplo, monitorar XX e o parâmetro YY no ponto de verificação ZZ) | O ML pode derivar padrões a partir dos dados existentes e derivar aproximações sobre o comportamento futuro (Alpaydin, 2010). Essas novas informações (conhecimento) podem apoiar os proprietários do processo na tomada de decisões ou usadas para melhorar automaticamente um sistema |
| Capacidade de se adaptar às mudanças do ambiente, com esforço e custo razoáveis. Idealmente, um grau de adaptação "automatizada" à mudança de condição | Como o ML faz parte da IA e, portanto, é capaz de aprender e se adaptar às mudanças, 'o projetista do sistema não precisa prever e fornecer soluções para todas as situações possíveis' (Alpaydin, 2010). Aprender e adaptar-se a ambientes em mudança automaticamente é um dos pontos fortes do ML (Lu, 1990; Simon, 1983) |
| Capacidade de aprimorar o conhecimento existente, aprendendo com os resultados | O ML pode contribuir para criar novas informações e possivelmente conhecimento, por exemplo, identificação de padrões nos dados existentes (Alpaydin, 2010; Pham & Afify, 2005) |
| Capacidade de trabalhar com os dados de fabricação disponíveis sem requisitos especiais para a captura de informações muito específicas no início | As técnicas de ML são projetadas para derivar conhecimento dos dados existentes (Alpaydin, 2010; Kwak & Kim, 2012). "Os dados armazenados tornam-se úteis apenas quando são analisados e transformados em informações que podemos usar, por exemplo, para fazer previsões" (Alpaydin, 2010) |
| Capacidade de identificar intra-relações e inter-relações relevantes e, idealmente, correlação e / ou causalidade | O objetivo de certas técnicas de ML é detectar certos padrões ou regularidades que descrevem as relações (Alpaydin, 2010) |

No entanto, uma análise mais detalhada das técnicas de ML disponíveis, bem como seus pontos fortes e limitações em relação aos requisitos, devem ser fornecidos. Acima de tudo, a possível compatibilidade com o conceito teórico do estado do produto e sua perspectiva sobre o programa de manufatura deve ser aprofundada antes que um julgamento final possa ser dado. Além disso, há muitas perguntas a serem respondidas, como o modo como as técnicas de ML podem lidar com informações qualitativas.

Na próxima seção, as vantagens e os desafios da aplicação de aprendizado de máquina na fabricação são introduzidos com base nos requisitos apresentados anteriormente.

Vantagens e desafios da aplicação de aprendizado de máquina na fabricação

O ML foi utilizado com sucesso em várias aplicações de otimização de processos, monitoramento e controle na fabricação e manutenção preditiva em diferentes indústrias (Alpaydin, 2010; Gardner & Bicker, 2000; Kwak & Kim, 2012; Pham & Afify, 2005; Susto, Schirru, Pampuri, McLoone e Beghi, 2015). Verificou-se que as técnicas de ML fornecem um potencial promissor para otimização aprimorada do controle de qualidade em sistemas de manufatura (Apte, Weiss & Grout, 1993), especialmente em ambientes complexos de manufatura em que a detecção das causas dos problemas é difícil (Harding, Shahbaz e Kusiak 2006). No entanto, muitas vezes as aplicações de ML são limitadas, concentrando-se em processos específicos, em vez de em todo o programa ou sistema de manufatura (Doltsinis, Ferreira, & Lohse, 2012).

Existem muitos métodos, ferramentas e técnicas diferentes de ML disponíveis, cada um com vantagens e desvantagens distintas. O domínio da ML cresceu para um domínio de pesquisa independente.

Vantagens da aplicação de aprendizado de máquina na fabricação

As vantagens gerais do ML foram estabelecidas nas seções anteriores, afirmando que as técnicas de ML são capazes de lidar com problemas completos de NP, que geralmente ocorrem quando se trata de problemas de otimização de sistemas de fabricação inteligentes (Monostori et al., 1998). A seguir, o foco está na capacidade das técnicas de ML em lidar com dados multidimensionais e de alta dimensão e na capacidade de extrair relacionamentos implícitos em grandes conjuntos de dados em um ambiente complexo e dinâmico, muitas vezes até caótico (Köksal, Batmaz, & Testik, 2011; Yang e Trewn, 2004). "Como a maioria dos problemas de engenharia e fabricação é rica em dados, mas escassa em conhecimento" (Lu, 1990), o ML fornece uma ferramenta para aumentar a compreensão do domínio. Devemos entender que a peculiaridade das vantagens pode diferir dependendo da técnica de ML escolhida.

No geral, é acordado que o ML permite reduzir o tempo e o ciclo de vida e

melhorar a utilização de recursos em certos problemas de fabricação com NP. Além disso, o ML fornece ferramentas poderosas para a melhoria contínua da qualidade em um processo grande e complexo, como a fabricação de semicondutores (Monostori et al., 1998; Pham & Afify, 2005).

Uma vantagem dos algoritmos de ML é a capacidade de lidar com problemas e dados de alta dimensão. Especialmente no que diz respeito à crescente disponibilidade de dados complexos (Yu & Liu, 2003) com pouca transparência na fabricação (Smola & Vishwanathan, 2008), isso provavelmente se tornará ainda mais importante no futuro. No entanto, como é verdade para a maioria das vantagens e desvantagens dos algoritmos de ML, isso não pode ser generalizado. Alguns algoritmos (por exemplo, SVM; Árvore de Decisão Hierárquica Distribuída) podem lidar com alta dimensionalidade melhor do que outros (Bar-Or, Wolff, Schuster e Keren, 2005; Do, Lenca, Lallich e Pham, 2010). Como foi afirmado anteriormente, na fabricação, na maioria dos casos, são aplicáveis os algoritmos de ML capazes de lidar com dados de alta dimensão. Portanto, a capacidade de lidar com alta dimensionalidade é considerada uma vantagem da aplicação de ML na manufatura.

Outra vantagem das técnicas de ML é o aumento da usabilidade da aplicação de algoritmos devido a programas (geralmente de origem) como o Rapidminer. Isso permite a aplicação (relativamente) fácil em muitos casos e, além disso, o ajuste confortável dos parâmetros para aumentar o desempenho da classificação.

Como mencionado anteriormente, uma grande vantagem dos algoritmos de ML é descobrir conhecimento anteriormente desconhecido (implícito) e identificar relacionamentos implícitos em conjuntos de dados. Dependendo da característica do algoritmo ML (supervisionado / não supervisionado ou Aprendizagem por Reforço [RL]), os requisitos para os dados disponíveis podem variar. No entanto, a capacidade geral do algoritmo ML de obter resultados em um ambiente de manufatura foi comprovada com sucesso (por exemplo: Alpaydin, 2010; Filipic & Junkar, 2000; Guo, Sun, Li & Wang, 2008; Kim, Kang, Cho, Lee, & Doh, 2012; Nilsson, 2005).

Dada a natureza dinâmica, incerta e complexa específica dos sistemas de manufatura, os algoritmos de ML oferecem a oportunidade de aprender com o sistema dinâmico e adaptar-se ao ambiente em mudança automaticamente até certo

ponto (Lu, 1990; Simon, 1983). A adaptação é, dependente do algoritmo ML, razoavelmente rápida e em quase todos os casos mais rápida que os métodos tradicionais.

A aplicação de ML na fabricação pode resultar em um padrão derivado de conjuntos de dados existentes, o que pode fornecer uma base para o desenvolvimento de aproximações sobre o comportamento futuro do sistema (Alpaydin, 2010; Nilsson, 2005). Essas novas informações (conhecimento) podem apoiar os proprietários do processo na tomada de decisões ou usadas para melhorar automaticamente o sistema diretamente. No final, o objetivo de certas técnicas de ML é detectar certos padrões ou regularidades que descrevem as relações (Alpaydin, 2010).

Kotsiantis (2007) comparou vários algoritmos de acordo com seu desempenho específico na aplicação de manufatura por diferentes atributos. Mesmo assim, isso apresenta a oportunidade de obter uma primeira impressão, não é recomendável basear a decisão em um algoritmo de ML adequado apenas para comparações apresentadas em uma tabela. Cada problema é diferente e o desempenho de cada algoritmo também depende dos dados disponíveis e do pré-processamento dos dados, bem como das configurações dos parâmetros. O algoritmo de melhor ajuste deve ser encontrado no teste de vários em um ambiente realista. Isso será discutido mais adiante na próxima seção.

Desafios da aplicação de aprendizado de máquina na fabricação

Um desafio muito comum da aplicação de ML na fabricação é a aquisição de dados relevantes. Isso também é uma limitação, pois a disponibilidade, a qualidade e a composição (por exemplo, os metadados são incluídos? São rotulados?), dos dados da manufatura em questão têm uma forte influência no desempenho dos algoritmos de ML. Alguns desafios que o conjunto de dados pode conter são, por exemplo, dados de alta dimensão podem impactar alguns algoritmos ML, ou seja, podem conter um alto grau de informação irrelevante e redundante que pode impactar o desempenho dos algoritmos de aprendizagem (Yu & Liu, 2003). Hoje, a maioria das técnicas de aprendizado de máquina manipula apenas dados com valores contínuos e nominais (Pham & Afify, 2005). A importância da influência depende de vários fatores, incluindo o próprio algoritmo e as configurações dos

parâmetros. Pode ser considerado um desafio geral para a maioria das pesquisas em manufatura e não apenas para a aplicação de ML, obter dados de qualidade devido a, por exemplo, preocupações de segurança ou uma falta básica de captura de dados durante o processo. Embora, na maioria dos casos, o ML permita a extração de conhecimento e gere melhores resultados do que a maioria dos métodos tradicionais, com menos requisitos em relação aos dados disponíveis, alguns aspectos relativos aos dados disponíveis que podem impedir a aplicação bem-sucedida ainda precisam ser considerados. Juntamente com o próximo ponto, isso destaca a crescente necessidade de entender os dados para aplicar o ML. Hoffmann (1990) destaca que, comparado aos métodos tradicionais em que é gasto muito tempo para extrair informações, no ML é gasto muito tempo na preparação dos dados.

Depois que os dados disponíveis são protegidos, os dados geralmente precisam ser pré-processados, dependendo dos requisitos do algoritmo de escolha. O pré-processamento de dados tem um impacto crítico nos resultados. No entanto, existem muitas ferramentas padronizadas disponíveis que suportam os processos mais comuns de pré-processamento, como normalizar e filtrar os dados. Também deve ser verificado se os dados de treinamento estão desequilibrados. Isso pode representar um desafio para o treinamento de certos algoritmos. Na prática de fabricação, é um problema comum que os valores de certos atributos não estejam disponíveis ou estejam ausentes no conjunto de dados (Pham & Afify, 2005). Esses chamados valores ausentes representam um desafio para a aplicação dos algoritmos de ML. Existem certos sistemas de indução práticos disponíveis que podem preencher a lacuna (Pham & Afify, 2005). No entanto, cada problema e o algoritmo ML aplicado posteriormente têm requisitos específicos quando se trata de substituir valores ausentes. Ao substituir os valores ausentes, o conjunto de dados original é influenciado. O objetivo é reduzir o viés e outras influências negativas o máximo possível em relação ao objetivo da análise. Como esta questão representa um desafio muito comum, há uma grande quantidade de literatura e soluções práticas (por exemplo, em R) disponíveis.

Um grande desafio de importância crescente é a questão de qual técnica e algoritmo de ML escolher (seleção do algoritmo de ML). Mesmo assim, houve tentativas de buscar a definição de "técnicas gerais de ML", os diversos problemas e

seus requisitos destacam a necessidade de algoritmos especializados com certa força e fraqueza. Especialmente devido ao aumento da atenção de profissionais e pesquisadores para o campo da ML na fabricação, um grande número de algoritmos diferentes da ML ou pelo menos variações dos algoritmos da ML estão disponíveis.

Adicionando a essa complexidade já existente, combinações de algoritmos diferentes, as chamadas 'abordagens híbridas', estão se tornando cada vez mais comuns, prometendo melhores resultados do que o aplicativo de algoritmo único 'individual' (por exemplo, Lee & Ha, 2009). Muitos estudos estão disponíveis, destacando uma aplicação bem-sucedida de técnicas de ML para problemas específicos. Ao mesmo tempo, os dados de teste não estão disponíveis publicamente em muitos casos. Isso faz uma avaliação neutra e imparcial dos resultados e, portanto, uma comparação final desafiadora. Atualmente, a abordagem geralmente aceita para selecionar um algoritmo de ML adequado para um determinado problema é a seguinte:

- Primeiro, observe os dados disponíveis e como eles são descritos (rotulados, sem rótulo, conhecimento especializado disponível etc.) para escolher entre uma abordagem supervisionada, não supervisionada ou reforço de aprendizagem.
- Em segundo lugar, a aplicabilidade geral dos algoritmos disponíveis em relação aos requisitos de problemas de pesquisa (por exemplo, capazes de lidar com alta dimensionalidade) deve ser analisada. Um foco específico deve ser colocado na estrutura, nos tipos de dados e na quantidade total dos dados disponíveis, que podem ser usados para treinamento e avaliação.
- Terceiro, aplicações anteriores dos algoritmos em problemas semelhantes devem ser investigadas para identificar um algoritmo adequado. O termo "semelhante", neste caso, significa problemas de pesquisa com requisitos comparáveis, por exemplo em outras disciplinas ou domínios.

Outro desafio é a interpretação dos resultados. Deve-se levar em consideração que não apenas o formato ou ilustração da saída é relevante para a interpretação, mas também as especificações do próprio algoritmo escolhido, as configurações dos parâmetros, o resultado planejado e também os dados, incluindo seu pré-processamento. Na interpretação dos resultados, certas limitações mais distintas

(novamente dependendo do algoritmo escolhido) podem ter um grande impacto. Entre esses estão o excesso de ajuste, viés e variância (portanto, troca entre viés e variância).

Estruturação de técnicas e algoritmos de aprendizado de máquina

Como afirmado anteriormente, o ML desenvolveu-se em um amplo e diversificado campo de pesquisa nas últimas décadas. Isso levou a uma variedade de subdomínios, algoritmos, teorias e áreas de aplicação, etc. O relacionamento e a estrutura entre os diferentes elementos não são comumente acordados. Diferentes pesquisadores escolhem diferentes abordagens para estruturar o campo. Na figura 4, é feita uma tentativa de estruturar o domínio ML do DM de acordo com as tarefas de um lado e os algoritmos disponíveis do outro. Essa estrutura destaca a importância da diferenciação da tarefa (qual é o objetivo) e do algoritmo (como esse objetivo pode ser alcançado) dentro do campo ML.

No entanto, a visão geral apresentada na figura 4 está aquém do fato de não refletir a diferenciação comumente aceita dos métodos de ML pelo feedback disponível em aprendizagem por reforço, não supervisionada e supervisionada. Monostori (2003) descreveu as três classes da seguinte forma:

- Aprendizado por reforço: menos feedback é dado, uma vez que não é a ação apropriada, mas apenas uma avaliação da ação escolhida é dada pelo professor;
- Aprendizado não supervisionado: nenhuma avaliação da ação é fornecida, pois não há professor;
- Aprendizado supervisionado: a resposta correta é fornecida por um professor.

Essa estrutura é amplamente aceita, no entanto, ainda existem diferenças no que se refere a elas ou a essas três classes. Por exemplo, Pham e Afify (2005) mapeiam supervisionado, não supervisionado e aprendizagem por reforço como parte de Redes Neurais, como mostrado na figura 5. No entanto, Pham e Afify (2005) também afirmam que se concentraram apenas em métodos de aprendizado de classificação supervisionados. Isso corresponderia que a aprendizagem indutiva pode ser agrupada em aprendizagem supervisionada e não supervisionada. Outros pesquisadores diferenciam entre aprendizado ativo e passivo, afirmando que o aprendizado ativo é geralmente usado para se referir a um problema ou sistema de

aprendizado em que o aprendiz tem algum papel na determinação de quais dados serão treinados enquanto o aprendizado passivo descreve uma situação em que o aluno não tem controle sobre o conjunto de treinamento. Aparentemente, o aprendizado ativo é freqüentemente usado para problemas em que é difícil (caro e / ou demorado) obter dados de treinamento rotulados. A vantagem é ser capaz de obter um bom desempenho, necessitando de menos dados de treinamento do que outros alunos, devido aos exemplos úteis identificados sequencialmente pelo aluno ativo (Cohn, 2011). A aprendizagem ativa é aplicada principalmente em cenários supervisionados de ML, mas também é considerada valiosa em certos problemas de aprendizado por reforço (Cohn, 2011).

Alguns pesquisadores, como Kotsiantis (2007), se concentram apenas nas técnicas de classificação supervisionada e agrupam as redes neurais como um algoritmo de aprendizado como parte do aprendizado supervisionado. No entanto, os algoritmos de redes neurais também podem ser aplicados em aprendizado não supervisionado e aprendizado por reforço (Carpenter & Grossberg, 1988; Pham & Afify, 2005). Isso corresponde basicamente a Pham e Afify (2005), quando a noção no topo da hierarquia é vista como 'ML supervisionado em vez de 'aprendizado de máquina que eles declararam originalmente.

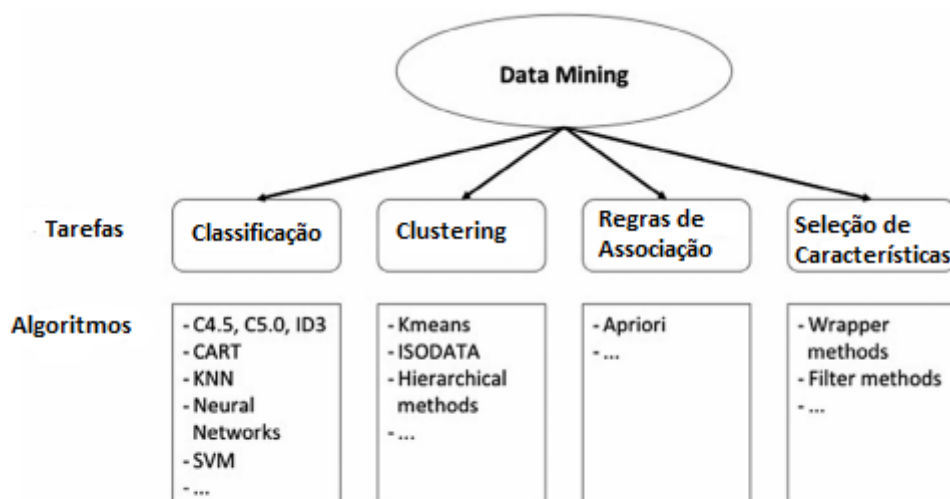


Figura 4. Uma visão geral das tarefas e principais algoritmos no DM (Corne et al., 2012)

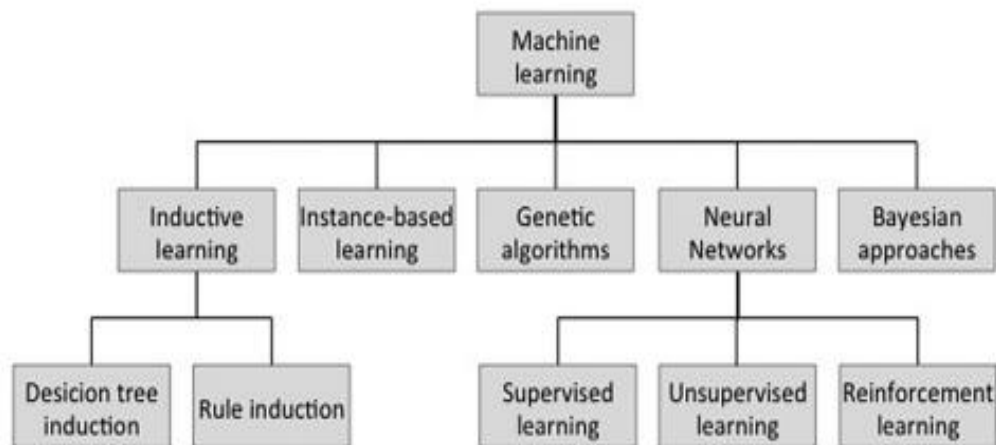


Figura 5. Classificação das principais técnicas de ML de acordo com Pham e Afify (2005).

Os diferentes algoritmos e abordagens combinatórias costumam se adaptar a problemas especiais. Isso dificulta sua comparação especialmente com o poder de classificação para o problema em questão. Uma primeira indicação pode ser a comparação de gráficos, como pode ser encontrado em Kotsiantis (2007). No entanto, uma abordagem mais promissora para selecionar um algoritmo adequado é procurar problemas de natureza semelhante e analisar qual algoritmo de ML foi usado para resolvê-lo e onde estão os resultados. Este é um bom ponto de partida. Depois que o algoritmo é aplicado ao problema e os primeiros resultados estão disponíveis, diferentes métodos podem ser aplicados e os resultados para o problema em questão podem ser comparados. As modernas ferramentas de computador suportam kernels diferentes e tornam a troca (relativamente) confortável.

6. Estudo de Caso: Data Analytics na Indústria de Equipamentos Elétricos

Este estudo de caso ilustra a aplicação de algumas das técnicas discutidas neste capítulo. Este estudo de caso trata da aplicação de ferramentas de *data analytics* numa multinacional alemã que por uma questão de sigilo a empresa será referida como “Empresa de Equipamentos Eletroeletrônicos”.

A Future Factory se reuniu com o Dr. Gunter Beitingger, vice-presidente de fabricação da Empresa de Equipamentos Eletroeletrônicos, para uma analisar de

maneira aprofundada o projeto "*Lean Digital Factory*". Ele conectará mais de trinta plantas à uma plataforma de dados de manufatura (PDM) via MindSphere e implementará uma camada *Industrial Edge* criada para esse fim.

Em outubro de 2017, a Empresa de Equipamentos Eletroeletrônicos lançou seu programa *Lean Digital Factory* (LDF). Combinando um grupo de especialistas de diferentes funções de negócios e unidades de tecnologia, seu objetivo é definir um roteiro conceitual holístico de transformação digital para todas as fábricas da empresa denominado Indústrias Digitais (ID).

Para capturar completamente o valor do uso de *big data* na fabricação, as plantas de ID, precisavam ter uma arquitetura de dados flexível que permitisse que diferentes usuários internos e externos extraíssem o valor máximo do ecossistema de dados. Aqui, a camada *Industrial Edge* entra em cena, que processa dados próximos aos sensores e à fonte de dados.

O conceito *Industrial Edge* e *data lake* permitiu uma solução mais poderosa do que qualquer outro conceito de armazenamento e utilização de dados:

- O PDM será uma área de armazenamento colossal para todos os dados de fabricação e será tremendamente poderoso para todos os níveis de usuário
- A plataforma de dados PDM é uma agregação centralizada e indexada de conjuntos de dados organizados distribuídos
- Os *big data* serão armazenados no PDM independentemente de seu uso posterior, ou seja, dados brutos
- Em combinação com o *Industrial Edge*, o PDM é o pré-requisito para computação em nuvem eficaz e escalável e aprendizado de máquina
- O *Industrial Edge* é usado nessa arquitetura para diversas finalidades, como ingestão de dados, pré-preparação, porta de segurança e decisões em tempo real.
- Funcionalidades de ecossistema altamente integradas, mas baseadas em módulos e serviços.

A métrica da estrela guia foi aplicada à fábrica digital enxuta e seus dados

No ID, pode ser desafiador aproveitar o potencial da digitalização em grande escala devido a soluções de software proprietárias instaladas, processos personalizados, interfaces padronizadas e tecnologias mistas. No entanto, na Empresa de Equipamentos Elétricos, isso não significa que executamos um grande programa de padronização antes de aproveitar as possibilidades de análise de dados e manutenção preditiva em nossas plantas.

Para colocar o pé na estrada em larga escala, exigimos um conceito arquitetônico que nos permitisse desenvolver aplicações, ampliar e transferir soluções de fábrica para fábrica, da engenharia para o chão de fábrica, bem como do fornecedor para o cliente e reutilizar insights de processo identificados de uma aplicação para outro. Durante o nosso programa LDF, usamos o conceito “Estrela Guia” em combinação com os Processos de Referência para descrever exatamente o que pretendemos e quais funcionalidades precisamos para nos beneficiar em troca. O conceito "Estrela Guia" é muito conhecido por inclinar pessoas afins e ajuda a concentrar todas as atividades de implementação.

A ligação do gêmeo digital de produto, produção e desempenho aumentará a eficiência do processo de gerenciamento do ciclo de vida do produto e reduzirá o tempo de colocação no mercado. Enquanto os engenheiros estão projetando, um modelo de custo é calculado em tempo real usando um padrão de relação custo-projeto. Isso é suportado principalmente pelo alto nível de maturidade e interoperabilidade das soluções de Software para as ID baseadas no backbone de dados PLM Teamcenter.

Outra “Estrela Guia” é a visualização digital completa de toda a planta, com suas linhas e células de produção. Isso inclui processos logísticos para materiais e ferramentas por “*Process Simulate*” e “*Plant Simulation*” para projetar e otimizar todos os ativos no chão de fábrica. Todos os dados gerados durante o processo de produção são coletados e analisados por meio de métodos de inteligência artificial para devolvê-los ao processo de design do produto e produção para otimização contínua.

Outro exemplo é a automação total de nosso planejamento, programação e sequenciamento de ordens de produção para garantir a melhor utilização dos

ativos durante a produção. Ao mesmo tempo, queremos chegar o mais próximo possível de um fluxo de peça única e de um tempo takt que seja igual ao tempo takt dos clientes. Um exemplo em que usamos nosso PDM nesse contexto é o balanceamento de capacidade. Na rede da fábrica de ID, temos muitas plantas eletrônicas com as correspondentes linhas de produção de tecnologia montadas na superfície. Para usar essa sinergia, estamos gerando um mercado no qual podemos equilibrar a capacidade através dos limites da fábrica.

No chão de fábrica, veículos guiados autônomos com inteligência de enxame e grupos de robôs formam um sistema cibernético para organizar o fornecimento de material intralogístico e os arranjos de trabalho altamente flexíveis. A inteligência artificial, os algoritmos de aprendizado de máquina e o reconhecimento de padrões oferecem suporte a manutenção preditiva, redução dos esforços de teste e aumento da utilização da máquina, distribuindo informações relevantes às pessoas por meio de dispositivos inteligentes conectados. Em cooperação com a Schmalz, implementamos a utilização da máquina em nossa fábrica da Amberg, onde otimizamos uma máquina de embalagem usando um sensor inteligente do software Schmalz e Empresa de Equipamentos Eletroeletrônicos.

Os dados para todos esses cenários devem ser extraídos de diferentes fontes, como o sistema ERP, "PLM system Teamcenter", nível MES, nível SCADA, "*Industrial Edge*" (dados do chão de fábrica), PLCs Simáticos, sensores e outros que ainda são fontes não conhecidas.

Conhecendo todos os diferentes cenários e estados de destino que estamos buscando, fica claro que a arquitetura tradicional de data warehouse tem seus limites. Os 9Vs de *big data*, como volume, velocidade, variedade, veracidade, valor, volatilidade, visibilidade, viabilidade e validade, agora estão definindo dados. O volume e o fluxo de dados explodiram nas fábricas devido à rastreabilidade, regulamentos legais e garantia de qualidade. Hoje, queremos analisar dados de diferentes fontes: feeds de vídeo, fotografias, dados de processo, dados de teste, arquivos de log e até arquivos de texto. Com isso, surge o desafio: em quais dados confiar, quais devem ser mantidos e quais descartados? Eles precisam ter o mesmo valor em unidades e por quanto tempo esses dados são retidos?

A plataforma de dados de fabricação (PDM)

James Dixon, fundador e CTO da Pentaho foi o primeiro a usar a expressão “*Data Lake*” e a explicou da seguinte maneira:

“Se você pensa no DataMart como uma loja de água engarrafada - limpa, embalada e estruturada para facilitar o consumo - o *data lake* é um grande corpo de água em um estado mais natural. O conteúdo do lago de dados flui de uma fonte para encher o lago, e vários usuários do lago podem vir para examinar, mergulhar ou coletar amostras.”

O LDF PDM do MindSphere permitirá casos de uso orientados a dados, conectar e indexar várias fontes de dados por meio de uma única interface e atuar como uma única fonte de verdade para os dados de fabricação. O PDM fornece tecnologia, capacitação, dados como serviço e, finalmente, análises como serviço para as fábricas da ID e seus clientes:

- Ele suporta todos os tipos de dados e os mantém em sua forma bruta
- Adapta-se a alterações facilmente (sem dependência de hardware) à medida que os requisitos evoluem
- Desenvolvedores e usuários podem acessar todos os dados disponíveis
- Desenvolvedores de aplicativos e cientistas de dados podem aproveitar os dados existentes para criar valor comercial
- Ativação da conversão de requisitos de negócios para consultas de dados e oferta de conjuntos de dados finais para desenvolvedores de aplicativos internos e externos e usuários de negócios (dados como serviço)
- Ativação para oferecer insights profundos e correlações de dados, além dos dados brutos, para desenvolvedores de aplicativos internos e externos e usuários de negócios (análise como serviço)

Os proprietários e desenvolvedores de produtos são suportados por uma "organização PDM" para conectividade e compartilhamento de dados. O foco está na análise de dados e não na conexão técnica.

MindSphere, *Manufacturing Data Platform* e *Industrial Edge* definirão a nova estrutura das fábricas para processamento de dados. Nesse contexto, a computação de borda é usada para otimizar nosso sistema de computação em nuvem, executando o processamento de dados na borda da rede, o que significa que fica próximo à fonte dos dados.

Um dos principais desafios de todos os nossos casos de uso de LDF é: como podemos processar nossos dados a partir de tantos dispositivos diferentes? A computação de borda com o *Industrial Edge* é a opção, mas, diferentemente da arquitetura do PDM, que centraliza o processamento e o armazenamento em um único e maciço data center, o *Industrial Edge* fornece o poder de processamento de dados aos dispositivos de borda. A principal vantagem é que apenas os resultados do processamento de dados precisam ser transportados pelas redes. Isso fornece resultados precisos e em tempo real e consome muito menos largura de banda da rede. O processamento dos dados está próximo da fonte dos dados.

Faz sentido absoluto dividir o processamento entre o *Industrial Edge* e o MindSphere PDM centralizado, mas um fato importante foi considerado: a computação de borda não armazena dados a longo prazo, acaba sendo excluída, o que não é propício para análises de *big data* e modelos de negócios como DaaS (Dados como serviço) ou AaaS (Analytics como serviço).

Se todos os dados coletados precisarem ser armazenados para fins de tomada de decisões de análise cumulativa, como a IA, a computação de borda por si só não é o ajuste certo. É aqui que o PDM baseado na nuvem do MindSphere tem um grande valor. No LDF, estamos criando aplicativos baseados na computação *Industrial Edge* criados especificamente para cada planta em combinação com o PDM, onde mais de 30 plantas serão conectadas. Na camada de borda, haverá aplicativos que colocam o processamento de dados em sensores para processar rapidamente reações a alarmes, como nosso caso de uso do eixo, em que prevemos manutenção para máquinas de despanelamento ou a decisão de raio-x baseada em algoritmo que decide se um circuito impresso A placa deve ser testada ou não na faixa de milissegundos.

A Empresa de Equipamentos Eletroeletrônicos não colocará dados e aplicativos de controle de estoque nos limites - isso resultaria em uma bagunça distribuída, não segura e incontrolável, sem opção de escalabilidade. É por isso que o treinamento de um algoritmo para se tornar mais inteligente será feito na nuvem e os *insights* poderão servir posteriormente para um negócio de AaaS.

Nesse contexto, a computação de borda não substituirá a computação em nuvem via *data lake* de fabricação nem vice-versa, embora as duas abordagens se complementem. Afirmar que a computação de ponta substituirá a computação em nuvem é como dizer que um PC substituirá um datacenter. Ele fornecerá aos cientistas e analistas de dados a possibilidade de provar hipóteses e procurar correlações, além de permitir que os usuários de negócios e serviços, engenheiros e técnicos da oficina explorem os dados, definam casos de uso relevantes na fábrica e criem relatórios por conta própria. O *Edge* em combinação com a computação em nuvem permitirá a escalabilidade dos projetos de IoT a um custo muito menor em comparação aos métodos tradicionais de nuvem. A computação em nuvem *Industrial Edge* e PDM pode e irá funcionar bem em conjunto.

A Empresa de Equipamentos Eletroeletrônicos espera que com a adoção dessas tecnologias possa manter ou até mesmo ampliar sua participação no mercado.

Referências

- [1] Akay, D. (2011). Grey relational analysis based on instance based learning approach for classification of risks of occupational low back disorders. *Safety Science*, 49, 1277–1282. doi:<http://dx.doi.org/10.1016/j.ssci.2011.04.018>
- [2] Akkerhuis, T. S., Spanjaard, C. M., Nuijten, S. E., Berendes, O. G., and Does, R. J. M. M., (2015). “Quality Quandaries: Realizing strategic focal points at a business school”, *Quality Engineering*, Vol. 27 No. 2, pp. 267--273.
- [3] Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). Cambridge, MA: MIT Press.
- [4] Anderson, L. W., Krathwohl, D. R., Airasian, P., Cruikshank, K., Mayer, R., Pintrich, P. and Wittrock, M. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom’s taxonomy*. New York. Longman Publishing.
- [5] Antony, J., Snee, R., and Hoerl, R., (2017). “Lean Six Sigma: yesterday, today and tomorrow”, *International Journal of Quality & Reliability Management*, Vol. 34 No. 7, pp. 1073--1093.
- [6] Artz, AF, & Armour-Thomas, E.(1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. *Cognition and Instruction*, Vol. 9 No. 2, pp. 137--175.
- [7] Arumugam, V., Antony, J. and Linderman, K., (2014).” A multilevel framework of Six Sigma: a systematic review of the literature, possible extensions and future research”, *Quality Management Journal*, Vol. 21 No. 4, pp. 36--61.
- [8] ASQ, (2017). “Certified Six Sigma Black Belt Body of Knowledge”, available at www.asq.org/cert/six-sigma-black-belt (accessed 18 July 2017).
- [9] Auschitzky, E., Hammer, M. and Rajagopaul, A., (2014).” How *big data* can improve manufacturing.” McKinsey & Company. Available at <http://www.mckinsey.com/businessfunctions/operations/our-insights/how-big-data-can-improve-manufacturing> (accessed on 15 August 2017).
- [10] Azevedo, A. and Santos, M. F., (2008). “KDD, SEMMA and CRISP-DM: A parallel overview”. *Proceedings of IADIS European Conference on Data Mining*, pp. 182--185.
- [11] Bar-or, A., Schuster, A., Wolff, R., & Keren, D. (2005). Decision tree induction in high dimensional, hierarchically distributed databases. In *Proceedings SI-AM International Data Mining Conference* (pp. 466–470). Newport Beach, CA.
- [12] Barratt M., Choi T.Y. and Li M., (2011). “Qualitative case studies in operations management: Trends, research outcomes, and future research implications”, *Journal of Operations Management*, Vol. 29 No. 4, pp. 329--342.
- [13] Bartlett, R., (2013). *A practitioner’s guide to business analytics: Using data analysis tools to improve your organization’s decision making and strategy*. McGraw Hill Professional.
- [14] Ben-hur, A., & Weston, J. (2010). A user’s guide to support vector machines. In *Data mining techniques for the life sciences methods in molecular biology* (Vol. 609, pp. 223–239). Totowa, NJ: Humana Press. http://dx.doi.org/10.1007/978-1-60327-241-4_13
- [15] Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York, NY: Springer.

- [16] Borin, A., Ferrão, M. F., Mello, C., Maretto, D. A., & Poppi, R. J. (2006). Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. *Analytica Chimica Acta*, 579, 25–32. doi:<http://dx.doi.org/10.1016/j.aca.2006.07.008>
- [17] Brunato, M., & Battiti, R. (2005). Statistical learning theory for location fingerprinting in wireless LANs. *Computer Networks*, 47, 825–845. doi:<http://dx.doi.org/10.1016/j.comnet.2004.09.004>
- [18] Burbidge, R., Trotter, M., Buxton, B., & Holden, S. (2001). Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Computers & Chemistry*, 26, 5–14.
- [19] Carpenter, G. A., & Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21, 77–88. doi:<http://dx.doi.org/10.1109/2.33>
- [20] Chand, S., & Davis, J. F. (2010, July). What is smart manufacturing? *Time Magazine*.
- [21] Cherkassky, V., & Ma, Y. (2009). Another look at statistical learning theory and regularization. *Neural Networks*, 22, 958–969. doi:<http://dx.doi.org/10.1016/j.neunet.2009.04.005>
- [22] Chinnam, R. B. (2002). Support vector machines for recognizing shifts in correlated and other manufacturing processes. *International Journal of Production Research*, 40, 4449–4466. doi:<http://dx.doi.org/10.1080/00207540210152920>
- [23] Cohn, D. (2011). Active learning (p. 10). Sammut, C. & Webb, G. I. (Eds.) (2011). *Encyclopedia of machine learning* (C. Sammut & G. I. Webb, Eds.) (p. 1058). New York, NY: Springer. doi:<http://dx.doi.org/10.1007/978-0-387-30164-8>
- [24] Cook, D. F., Zobel, C. W., & Wolfe, M. L. (2006). Environmental statistical process control using an augmented neural network classification approach. *European Journal of Operational Research*, 174, 1631–1642. doi:<http://dx.doi.org/10.1016/j.ejor.2005.04.035>
- [25] Corne, D., Dhaenens, C., & Jourdan, L. (2012). Synergies between operations research and data mining: The emerging use of multi-objective approaches. *European Journal of Operational Research*, 221, 469–479. doi:<http://dx.doi.org/10.1016/j.ejor.2012.03.039>
- [26] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.
- [27] Davis, J., Edgar, T., Graybill, R., Korambath, P., Schott, B., Swink, D., & Wetzel, J. (2015). Smart manufacturing. *Annual Review of Chemical and Biomolecular Engineering*, 6, 141–160. doi:<http://dx.doi.org/10.1146/annurev-chembioeng-061114-123255>
- [28] De Koning, H., de Mast, J., Does, R. J. M. M., Vermaat, T. and Simons, S., (2008). “Generic lean Six Sigma project definitions in financial services”. *The Quality Management Journal*, Vol. 15 No. 4, pp. 32–45.
- [29] De Koning, H., Verver, J. P., Heuvel, J., Bisgaard, S. and Does, R. J. M. M., (2006). “Lean Six Sigma in healthcare”. *Journal for Healthcare Quality*, Vol. 28 No. 2, pp. 4–11.
- [30] De Mast, J., Does, R. J. M. M., and De Koning, H., (2016). *Lean Six Sigma for service and healthcare*. Beaumont Quality Publications, Second Edition.

- [31] Delgado, C., Ferreira, M., and Castelo Branco, M., (2010). "The implementation of LSS in financial services organizations". *Journal of Manufacturing Technology Management*, Vol. 21 No. 4, pp. 512--523.
- [32] Dhawan, R., Singh K., and Tuteja A. (2014). "When *big data* goes lean". *McKinsey Quarterly* Vol. 24 No. 2, pp. 97--105.
- [33] Dietterich, T. G. (2000). Ensemble methods in machine learning. *Proceedings of Multiple Classifier Systems: First International Workshop (MCS)* (pp. 1--15). Berlin: Springer. doi:http://dx.doi.org/10.1007/3-540-45014-9_1
- [34] Dingli, D. J. (2012). *The manufacturing industry – Coping with challenges* (Working Paper No. 2012/05). Maastricht: Maastricht School of Management.
- [35] Dutcher, R., (2014). "Linking *Big data* to Big Process Improvement...An Imperative". Capgemini post, available at <https://www.capgemini.com/blog/bpo-thought-process/2014/03/linking-big-data-to-big-process-improvementan-imperative> (accessed on 15 August 2017).
- [36] Dutt, V., & Gonzalez, C. (2012). Making instance-based learning theory usable and understandable: The instance-based learning tool. *Computers in Human Behavior*, 28, 1227--1240. doi:<http://dx.doi.org/10.1016/j.chb.2012.02.006>
- [37] Easton, G.S., and Rosenzweig, E.D., (2012). "The role and experience in six sigma project success: an empirical analysis of improvement projects". *Journal of Operations Management*, Vol. 30 No. 7--8, pp. 481--493.
- [38] Eisenhardt, K., (1989). "Building Theories from Case Study Research". *The Academy of Management Review*, Vol. 14 No. 4, pp. 532--550.
- [39] Elangovan, M., Sakthivel, N. R., Saravanamurugan, S., Nair, B. B., & Sugumaran, V. (2015). Machine learning approach to the prediction of surface roughness using statistical features of vibration signal acquired in turning. *Procedia Computer Science*, 50, 282--288. doi:<http://dx.doi.org/10.1016/j.procs.2015.04.047>
- [40] El-naqa, I., Yang, Y., Wernick, M. N., Galatsanos, N. P., & Nishikawa, R. M. (2002). A support vector machine approach for detection of microcalcifications. *IEEE Transactions on Medical Imaging*, 21, 1552--1563. doi:<http://dx.doi.org/10.1109/TMI.2002.806569>
- [41] European Commission. (2016). *Factories for the future*. Retrieved from http://ec.europa.eu/research/industrial_technologies/factories-of-the-future_en.html
- [42] Evgeniou, T., Poggio, T., Pontil, M., & Verri, A. (2002). Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 38, 421--432. doi:[http://dx.doi.org/10.1016/S0167-9473\(01\)00069-X](http://dx.doi.org/10.1016/S0167-9473(01)00069-X)
- [43] Evgeniou, T., Pontil, M., & Poggio, T. (2000). Statistical learning theory: A primer. *International Journal of Computer Vision*, 38, 9--13. doi:<http://dx.doi.org/10.1023/A:1008110632619>
- [44] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). "From data mining to knowledge discovery in databases". *AI magazine*, Vol. 17 No. 3, pp. 37--54.
- [45] Filipic, B., & Junkar, M. (2000). Using inductive machine learning to support decision making in machining processes. *Computers in Industry*, 43, 31--41. doi:[http://dx.doi.org/10.1016/S0166-3615\(00\)00056-7](http://dx.doi.org/10.1016/S0166-3615(00)00056-7)
- [46] Freund, Y., & Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119--139.

- [47] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., & Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16, 906–914. doi:<http://dx.doi.org/10.1093/bioinformatics/16.10.906>
- [48] Gagliardi, F. (2011). Instance-based classifiers applied to medical databases: Diagnosis and knowledge extraction. *Artificial Intelligence in Medicine*, 52, 123–139. doi:<http://dx.doi.org/10.1016/j.artmed.2011.04.002>
- [49] Gardner, R., & Bicker, J. (2000). Using machine learning to solve tough manufacturing problems. *International Journal of Industrial Engineering-Theory Applications and Practice*, 7, 359–364.
- [50] Gaudard, M., Ramsey, P., and Stephens, M., (2009). “Interactive data mining informs designed experiments”. *Quality and Reliability Engineering International*, Vol. 25 No. 3, pp. 299--315.
- [51] George, M. (2003), *Lean Six Sigma for Service*, McGraw-Hill, New York, NY.
- [52] Gonzalez, C., Dutt, V., & Lebiere, C. (2013). Validating instance-based learning mechanisms outside of ACT-R. *Journal of Computational Science*, 4, 262–268. doi:<http://dx.doi.org/10.1016/j.jocs.2011.12.001>
- [53] Gordon, J., & Sohal, A. S. (2001). Assessing manufacturing plant competitiveness. *International Journal of Operations & Production Management*, 21, 233–253.
- [54] Günther, J., Pilarski, P. M., Helfrich, G., Shen, H., & Diepold, K. (2015). First steps towards an intelligent laser welding architecture using deep neural networks and reinforcement learning. *Procedia Technology*, 15, 474–483.
- [55] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). A gene selection method for cancer classification using Support Vector Machines. *Machine Learning*, 46, 389–422. doi:<http://dx.doi.org/10.1155/2012/586246>
- [56] Hahn, G. J., Hill, W. J., Hoerl, R. W., and Zinkgraf, S. A. (1999). “The impact of Six Sigma improvement—a glimpse into the future of statistics”. *The American Statistician*, Vol. 53 No. 3, pp. 208--215.
- [57] Hansson, K., Yella, S., Dougherty, M., & Fleyeh, H. (2016). Machine learning algorithms in heavy process manufacturing. *American Journal of Intelligent Systems*, 6(1), 1–13. doi:<http://dx.doi.org/10.5923/j.ajis.20160601.01>
- [58] Harding, J. A., Shahbaz, M., & Kusiak, A. (2006). Data mining in manufacturing: A review. *Journal of Manufacturing Science and Engineering*, 128, 969–976. doi:<http://dx.doi.org/10.1115/1.2194554>
- [59] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735– 1780. doi:<http://dx.doi.org/10.1162/neco.1997.9.8.1735>
- [60] Hoffmann, A. G. (1990, August). General limitations on machine learning. In *Proceedings of the 9th European Conference on Artificial Intelligence* (pp. 345–347). Stockholm, Sweden.
- [61] Huang, Z., Chen, H., Hsu, C.-J., Chen, W.-H., & Wu, S. (2004). Credit rating analysis with support vector machines and neural networks: A market comparative study. *Decision Support Systems*, 37, 543–558. doi:[http://dx.doi.org/10.1016/S0167-9236\(03\)00086-1](http://dx.doi.org/10.1016/S0167-9236(03)00086-1)
- [62] Jain, A. K., Murty, M. N., & Flynn, P. (1999). Data clustering: A review. *ACM Computing Surveys*, 31, 264–323. doi:<http://dx.doi.org/10.1145/331499.331504>
- [63] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning*. New York: springer.

- [64] Jang, G. S. and Jeon, J. H. (2009). "A Six Sigma methodology using data mining: a case study on Six Sigma project for heat efficiency improvement of a hot stove system in a Korean steel manufacturing company." In *Cutting-Edge Research Topics on Multiple Criteria Decision Making* (72-80). Springer Berlin Heidelberg.
- [65] Jurkovic, Z., Cukor, G., Brezocnik, M., & Brajkovic, T. (2016). A comparison of machine learning methods for cutting parameters prediction in high speed turning process. *Journal of Intelligent Manufacturing*. doi:<http://dx.doi.org/10.1007/s10845-016-1206-1>
- [66] Kabacoff, R. I. (2011). *Advanced methods for missing data*. In R. I. Kabacoff (Ed.), *R in action: Data analysis and graphics with R* (pp. 352–371). Shelter Island, NY: Manning Publications.
- [67] Kang, P., & Cho, S. (2008). Locally linear reconstruction for instance-based learning. *Pattern Recognition*, 41, 3507–3518. doi:<http://dx.doi.org/10.1016/j.patcog.2008.04.009>
- [68] Kang, P., Kim, D., & Cho, S. (2016). Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing. *Expert Systems with Applications*, 51, 85–106. doi:<http://dx.doi.org/10.1016/j.eswa.2015.12.027>
- [69] Kim, D., Kang, P., Cho, S., Lee, H., & Doh, S. (2012). Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. *Expert Systems with Applications*, 39, 4075–4083. doi:<http://dx.doi.org/10.1016/j.eswa.2011.09.088>
- [70] Köksal, G., Batmaz, İ. and Testik, M. C. (2011). "A review of data mining applications for quality improvement in manufacturing industry". *Expert systems with Applications*, Vol. 38 No. 10, pp. 13448--13467.
- [71] Köksal, G., Batmaz, İ., & Testik, M. C. (2011). A review of data mining applications for quality improvement in manufacturing industry. *Expert Systems with Applications*, 38, 13448–13467. doi:<http://dx.doi.org/10.1016/j.eswa.2011.04.063>
- [72] Koltchinskii, V., Abdallah, C. T., Ariola, M., & Dorato, P. (2001). Statistical learning control of uncertain systems: Theory and algorithms. *Applied Mathematics and Computation*, 120, 31–43. doi:[http://dx.doi.org/10.1016/S0096-3003\(99\)00283-0](http://dx.doi.org/10.1016/S0096-3003(99)00283-0)
- [73] Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31, 249–268.
- [74] Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
- [75] Kwak, D.-S., & Kim, K.-J. (2012). A data mining approach considering missing values for the optimization of semiconductor-manufacturing processes. *Expert Systems with Applications*, 39, 2590–2596. doi:<http://dx.doi.org/10.1016/j.eswa.2011.08.114>
- [76] Lameijer, B., Does, R. J. M. M. and De Mast, J. (2016). "Inter-industry generic Lean Six Sigma project definitions". *International Journal of Lean Six Sigma*, Vol. 7 No. 4, pp. 369--393.
- [77] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. doi:<http://dx.doi.org/10.1038/nature14539>
- [78] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1, 541–555.

- [79] Lee, J., & Ha, S. (2009). Recognizing yield patterns through hybrid applications of machine learning techniques. *Information Sciences*, 179, 844–850. doi:<http://dx.doi.org/10.1016/j.ins.2008.11.008>
- [80] Lee, J., Lapira, E., Bagheri, B., & Kao, H. (2013). Recent advances and trends in predictive manufacturing systems in *big data* environment. *Manufacturing Letters*, 1, 38–41. doi:<http://dx.doi.org/10.1016/j.mfglet.2013.09.005>
- [81] Li, H., Liang, Y., & Xu, Q. (2009). Support vector machines and its applications in chemistry. *Chemometrics and Intelligent Laboratory Systems*, 95, 188–198. doi:<http://dx.doi.org/10.1016/j.chemolab.2008.10.007>
- [82] Li, T.-S., & Huang, C.-L. (2009). Defect spatial pattern recognition using a hybrid SOM–SVM approach in semiconductor manufacturing. *Expert Systems with Applications*, 36, 374–385. doi:<http://dx.doi.org/10.1016/j.eswa.2007.09.023>
- [83] Loyer, J.-L., Henriques, E., Fontul, M., & Wiseall, S. (2016). Comparison of machine learning methods applied to the estimation of manufacturing cost of jet engine components. *International Journal of Production Economics*, 178, 109–119. doi:<http://dx.doi.org/10.1016/j.ijpe.2016.05.006>
- [84] Lu, S. C.-Y. (1990). Machine learning approaches to knowledge synthesis and integration tasks for advanced engineering automation. *Computers in Industry*, 15, 105–120. doi:[http://dx.doi.org/10.1016/0166-3615\(90\)90088-7](http://dx.doi.org/10.1016/0166-3615(90)90088-7)
- [85] Manallack, D. T., & Livingstone, D. J. (1999). Neural networks in drug discovery: Have they lived up to their promise? *European Journal of Medicinal Chemistry*, 34, 95–208.
- [86] Manning, C. D., Raghavan, P., & Schütze, H. (2009). *An introduction to information retrieval*. Cambridge: Cambridge University Press.
- [87] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Report, available at: <http://www.mckinsey.com/business-functions/digital-mckinsey/ourinsights/big-data-the-next-frontier-for-innovation> (accessed on 15 August 2017).
- [88] Margolis, D., Land, W. H., Gottlieb, R., & Qiao, X. (2011). A complex adaptive system using statistical learning theory as an inline preprocess for clinical survival analysis. *Procedia Computer Science*, 6, 279–284. doi:<http://dx.doi.org/10.1016/j.procs.2011.08.052>
- [89] Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183, 1466–1476. doi:<http://dx.doi.org/10.1016/j.ejor.2006.04.051>
- [90] Megahed, F.M. and Jones-Farmer, L.A. (2015). “Statistical Perspective on ‘*Big data*’”. *Frontiers in Statistical Quality Control*, Vol. 11, pp. 29–47. Springer International Publishing.
- [91] Monostori, L. (1993). A step towards intelligent manufacturing: Modelling and monitoring of manufacturing processes through artificial neural networks. *CIRP Annals*, 42, 485–488. doi:[http://dx.doi.org/10.1016/S0007-8506\(07\)62491-3](http://dx.doi.org/10.1016/S0007-8506(07)62491-3)
- [92] Monostori, L. (2003). AI and machine learning techniques for managing complexity, changes and uncertainties in manufacturing. *Engineering Applications of Artificial Intelligence*, 16, 277–291. doi:[http://dx.doi.org/10.1016/S0952-1976\(03\)00078-2](http://dx.doi.org/10.1016/S0952-1976(03)00078-2)
- [93] Monostori, L., Hornyák, J., Egresits, C., & Viharos, Z. J. (1998). Soft computing and hybrid AI approaches to intelligent manufacturing. *Tasks and Methods in Applied*

- Artificial Intelligence Lecture Notes in Computer Science, 1416, 765–774. doi:http://dx.doi.org/10.1007/3-540-64574-8_463
- [94] Monostori, L., Márkus, A., Van Brussel, H., & Westkämper, E. (1996). Machine learning approaches to manufacturing. *CIRP Annals*, 45, 675–712. Nilsson, N. J. (2005). *Introduction to machine learning*. Stanford, CA.
- [95] Okamoto, S., & Yugami, N. (2003). Effects of domain characteristics on instance-based learning algorithms. *Theoretical Computer Science*, 298, 207–233. doi:[http://dx.doi.org/10.1016/S0304-3975\(02\)00424-3](http://dx.doi.org/10.1016/S0304-3975(02)00424-3)
- [96] Oliff, H. and Liu, Y. (2017). “Towards Industry 4.0 Utilizing Data-Mining Techniques: A Case Study on Quality Improvement”. *Procedia CIRP*, Vol. 63, pp. 167--172.
- [97] Pande, P. S., Neuman, R. P. and Cavanagh, R. R. (2000). *The six sigma way*. McGraw-Hill.
- [98] Pham, D. T., & Afify, A. A. (2005). Machine-learning techniques and their applications in manufacturing. *Proceedings of the Institution of Mechanical Engineers. Part B: Journal of Engineering Manufacture*, 219, 395–412. doi:<http://dx.doi.org/10.1243/095440505X32274> Downloaded by [University of Strathclyde] at 03:13 03 August 2016
- [99] Quadrianto, N., & Buntine, W. L. (2011). Regression (pp. 838–842). Sammut, C. & Webb, G. I. (2011). *Encyclopedia of machine learning* (C. Sammut & G. I. Webb, Eds.) (p. 1058). New York, NY: Springer. doi:<http://dx.doi.org/10.1007/978-0-387-30164-8>
- [100] Raje, P. (2007). “Six Sigma maturity model”, available at: <https://www.isixsigma.com/implementation/basics/maturity-model-describes-stages-six-sigmaevolution/> (accessed on 15 August 2017).
- [101] Ribeiro, B. (2005). Support vector machines for quality monitoring in a plastic injection molding process. *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, 35, 401–410. doi:<http://dx.doi.org/10.1109/TSMCC.2004.843228>
- [102] Sagioglu, S. and Sinanc, D., (2013). “*Big data*: A review”. In *Collaboration Technologies and Systems (CTS)*, International Conference on IEEE, pp. 42--47.
- [103] Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal*, 3, 210–229. doi:<http://dx.doi.org/10.1147/rd.33.0210>
- [104] Scheidat, T., Leich, M., Alexander, M., & Vielhauer, C. (2009). Support vector machines for dynamic biometric handwriting classification. In *Proceedings of AIAI Workshops* (pp. 118–125). Thessaloniki, Greece.
- [105] Schroeder, R.G., Linderman, K., Liedtke and C., Choo, A.S. (2008). “Six Sigma: definition and underlying theory”, *Journal of Operations Management*, Vol. 26 No. 4, pp. 536--554.
- [106] Shiang, L. E., & Nagaraj, S. (2011). Impediments to innovation: Evidence from Malaysian manufacturing firms. *Asia Pacific Business Review*, 17, 209–223. doi:<http://dx.doi.org/10.1080/13602381.2011.533502>
- [107] Smola, A., & Vishwanathan, S. V. N. (2008). *Introduction to machine learning*. Cambridge: Cambridge University Press.
- [108] Steel, D. (2011). Testability and statistical learning theory. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Handbook of the philosophy of science* (Vol. 7, pp. 849–861). Amsterdam: Elsevier. doi:<http://dx.doi.org/10.1016/B978-0-444-51862-0.50028-9>

- [109] Stojanovic, N., Dinic, M. and Stojanovic, L., (2015). “*Big data* process analytics for continuous process improvement in manufacturing”, IEEE International Conference on *Big data (Big data)*, Santa Clara, CA, pp. 1398--1407.
- [110] Sammut, C., & Webb, G. I. (Eds.) (2011). Encyclopedia of machine learning (C. Sammut & G. I. Webb, Eds.) (p. 1058). New York, NY: Springer. doi:http://dx.doi.org/10.1007/978-0-387-30164-8
- [111] Sun, J., Rahman, M., Wong, Y., & Hong, G. (2004). Multiclassification of tool wear with support vector machine by manufacturing loss consideration. *International Journal of Machine Tools and Manufacture*, 44, 1179–1187. doi:http://dx.doi.org/10.1016/j.ijmachtools.2004.04.003
- [112] Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, 11, 812–820. doi:http://dx.doi.org/10.1109/TII.2014.2349359
- [113] Sutton, R. S., & Barto, A. G. (2012). Reinforcement learning: An introduction (2nd ed.). Cambridge, MA: MIT Press.
- [114] Tay, F. E. H., & Cao, L. J. (2002). Modified support vector machines in financial time series forecasting. *Neurocomputing*, 48, 847–861. doi:http://dx.doi.org/10.1016/S0925-2312(01)00676-2
- [115] Thomas, A. J., Byard, P., & Evans, R. (2012). Identifying the UK’s manufacturing challenges as a benchmark for future growth. *Journal of Manufacturing Technology Management*, 23, 142–156. doi:http://dx.doi.org/10.1108/17410381211202160
- [116] Voss, C., Tsikriktsis, N. and Frohlich, M., (2002). “Case research in operations management”. *International Journal of Operations & Production Management*, Vol. 22 No. 2, pp. 195--219.
- [117] Wang, K.-J., Chen, J. C., & Lin, Y.-S. (2005). A hybrid knowledge discovery model using decision tree and neural network for selecting dispatching rules of a semiconductor final testing factory. *Production Planning & Control*, 16, 665–680. doi:http://dx.doi.org/10.1080/09537280500213757
- [118] White House. (2014, October 27). FACT SHEET: President Obama announces new actions to further strengthen U.S. manufacturing. Retrieved from <https://www.whitehouse.gov/the-pressoffice/2014/10/27/fact-sheet-president-obama-announces-new-actions-further-strengthen-us-m>
- [119] Widodo, A., & Yang, B.-S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, 21, 2560–2574. doi:http://dx.doi.org/10.1016/j.ymsp.2006.12.007
- [120] Wiendahl, H.-P., & Scholtissek, P. (1994). Management and control of complexity in manufacturing. *CIRP Annals*, 43, 533–540. doi:http://dx.doi.org/10.1016/S0007-8506(07)60499-5
- [121] Wiering, M., & Van Otterlo, M. (2012). Reinforcement learning: State-of-the-art. New York, NY: Springer.
- [122] Wirth, R. and Hipp, J., (2000). “CRISP-DM: Towards a standard process model for data mining”. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pp. 29--39.
- [123] Wu, Q. (2010). Product demand forecasts using wavelet kernel support vector machine and particle swarm optimization in manufacture system. *Journal of Computational and Applied Mathematics*, 233, 2481–2491. doi:http://dx.doi.org/10.1016/j.cam.2009.10.030

- [124] Wuest, T. (2015). Identifying product and process state drivers in manufacturing systems using supervised machine learning (Springer theses). New York, NY: Springer Verlag.
- [125] Wuest, T., Irgens, C., & Thoben, K.-D. (2014). An approach to monitoring quality in manufacturing using supervised machine learning on product state data. *Journal of Intelligent Manufacturing*, 25, 1167–1180. doi:<http://dx.doi.org/10.1007/s10845-013-0761-y>
- [126] Wuest, T., Liu, A., Lu, S. C.-Y., & Thoben, K.-D. (2014). Application of the stage gate model in production supporting quality management. *Procedia CIRP*, 17, 32–37. doi:<http://dx.doi.org/10.1016/j.procir.2014.01.071>
- [127] Yang, K., & Trewn, J. (2004). *Multivariate statistical methods in quality management*. New York, NY: McGraw-Hill.
- [128] Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML2003)* (pp. 8). Washington, DC.
- [129] Zheng, Y., Li, S., & Wang, X. (2010). An approach to model building for accelerated cooling process using instance-based learning. *Expert Systems with Applications*, 37, 5364–5371. doi:<http://dx.doi.org/10.1016/j.eswa.2010.01.020>
- [130] Zhou, Z.-H. (2012). *Ensemble methods – Foundations and algorithms*, Machine Learning & Pattern Recognition Series. Florida, FL: Chapman & Hall/CRC. ISBN: 978-1-4398-3003-1.

DATA ANALYSIS APPLIED TO PRODUCTION: A case study in the Electrical Equipment Industry

Orlando Roque da Silva - UNIP-SP, Brasil
orlandoroque@uol.com.br

Angelo Palmisano - UNIVAG-MT, Brasil
angelopalmisano@uol.com.br

Delvio Venanzi - FATEC-Sorocaba-SP, Brasil
delviosuper@uol.com.br

ABSTRACT

A very common challenge of applying data analytics in manufacturing is the acquisition of relevant data. Given the dynamic, uncertain and complex nature specific to manufacturing systems, ML algorithms offer the opportunity to learn from the dynamic system and adapt to the changing environment automatically to some extent. The objective of this research is to assess the impact that technologies associated with industry 4.0, such as big data, machine learning, internet of things, sensor networks, etc., have on the analysis of manufacturing data. Initially, an assessment of the main technologies is carried out and, finally, a case study is presented in the electronic equipment industry. The method used was bibliographic research and analysis of a real case. The results are reported step-by-step in the analysis section. We conclude that there is a great way to advance in research and business practices, with regard to data analytics applications in the manufacturing environment in order to have new and innovative alternatives.

Keyword: Industry 4.0, Data Analytics, Machine Learning.