AGRUPAMENTO AUTOMÁTICO DE OSCILOGRAFIAS EM REDES DE DISTRIBUIÇÃO

ANDRÉ E. LAZZARETTI* VITOR H. FERREIRA† HUGO VIEIRA NETO‡ CLEVERSON L. S. PINTO§

*Instituto de Tecnologia para o Desenvolvimento – LACTEC
Avenida Comendador Franco, 1341 – Curitiba-PR

†Universidade Federal Fluminense – UFF
Rua Passo da Pátria, 156 – Niterói-RJ

‡Universidade Tecnológica Federal do Paraná – UTFPR
Avenida Sete de Setembro, 3165 – Curitiba-PR

§Companhia Paranaense de Energia – COPEL
Rua José Izidoro Biazetto, 158 – Curitiba-PR

Abstract— The analysis, processing, and information extraction using data from monitoring devices is a recurring subject in the power system sector worldwide. In this sense, the idea of this paper is to propose an original approach for automatic waveform analysis using automatic clustering, for unknown events in electrical distribution networks. We have used two different databases to perform the comparison and validation of the proposed method. The first database contains simulated data, whereas the second, contains real data acquired in a distribution network. The main objective of the automatic clustering is to provide information to the experts, about the possibility of new classes of events, extracting similarities that may exist among patterns in the dataset. Different automatic clustering models are used to perform this task. Here, we have demonstrated that it is possible to use the proposed method to assist experts in the new class identification stage, or even during the multi-class classification applied for known classes, by choosing an appropriate automatic clustering method.

Keywords— Waveform Classification, Novelty Detection, New Class Identification, Automatic Clustering.

Resumo— A análise, o processamento e a utilização de informações geradas por dispositivos de monitoramento de redes elétricas é tema recorrente no setor elétrico em todo o mundo. Nesse sentido, a motivação do presente trabalho é a proposição de uma abordagem original para análise automática de oscilografias previamente desconhecidas em redes elétricas de distribuição, utilizando modelos de agrupamento automático de dados. Para comparação e validação dos métodos são utilizadas duas bases de dados, sendo uma delas formada por dados simulados e outra contendo dados reais de oscilógrafos. O objetivo principal do agrupamento automático é fornecer informações ao especialista sobre a possibilidade de existirem novas classes em um conjunto de padrões previamente desconhecidos, destacando as similaridades existentes entre os padrões desse conjunto. Para realizar esse processo, optou-se por avaliar diferentes modelos de agrupamento automático. Mostra-se que, através de um método de agrupamento automático adequado, é possível utilizar essa metodologia para auxiliar no processo de identificação de novas classes, ou mesmo auxiliar no processo de classificação de classes já conhecidas.

Palavras-chave— Análise de Oscilografia, Detecção de Novidades, Identificação de Novas Classes, Agrupamento Automático.

1 Introdução

Em função das novas mudanças que vêm ocorrendo no setor elétrico de distribuição, principalmente do ponto de vista das características dos consumidores, aumentou-se a preocupação, principalmente por parte das concessionárias de energia, com eventos relacionados com Qualidade de Energia Elétrica (QEE). Esses eventos são caracterizados por alterar a forma de onda de tensão ou corrente, como resultado de curto-circuito, descargas atmosféricas ou má operação de equipamentos instalados em consumidores de energia. Esta alteração nas formas de onda pode, por vezes, resultar em danos a esses consumidores, forçando as concessionárias a monitorar o fornecimento da energia para que esta seja entregue de forma adequada nos vários estágios do sistema elétrico de potência.

Frente à restrição severa em relação aos limites de qualidade e continuidade de energia a serem atendidos, as concessionárias de energia vêm desenvolvendo uma série de medidas visando permitir um amplo monitoramento das suas redes e identificação das variações de forma de onda que possam ocorrer, buscando classificar e mitigar os eventos ocorridos. No presente trabalho, optou-se por utilizar como base para testes e validação da metodologia, o projeto proposto pela Companhia Paranaense de Energia (Copel) em conjunto com o Instituto de Tecnologia para o Desenvolvimento (Lactec), dentro do programa de Pesquisa e Desenvolvimento da ANEEL, intitulado *Instalação Piloto para Avaliação dos Níveis de Sobretensão Atmosférica no Sistema da COPEL Distribuição* (Lazzaretti et al., 2011). O objetivo desse projeto é o desenvolvimento de um sistema automático para a medição e registro de fenômenos eletromagnéticos rápidos em redes de distribuição de energia em média e baixa tensão.

O conjunto de registros obtidos por esses sistemas reflete um cenário interessante e comum a todo registrador que venha a ser utilizado para classificação automática de eventos em redes de distribuição. Uma classificação automática de eventos para um registrador deve levar em conta a identificação (classificação) das classes de eventos já registradas, bem como identificar novas classes de eventos que podem surgir ao longo do tempo em que o registrador encontra-se instalado, ou mesmo registros espúrios que possam ser observados no período. Esse processo de identificação

de novas classes é normalmente denominado de detecção de novidades (Lazzaretti et al., 2013). No contexto desse trabalho, uma forma de onda pode ser caracterizada como novidade em função da dificuldade na interpretação do fenômeno, da falta de registros prévios para caracterização desses eventos ou do desconhecimento com relação à ocorrência de um evento com a forma de onda registrada.

Uma vez identificada uma forma de onda ou um conjunto de oscilografias como novidade, como proposto em (Lazzaretti et al., 2013), o problema passa a ser como proceder com esses padrões. No presente trabalho, a ideia é permitir que informações relevantes possam ser retiradas desse conjunto de dados, fornecendo um procedimento de auxílio na identificação de novas classes e facilitando a inclusão desse conhecimento no classificador proposto.

A dificuldade desse procedimento é que, em princípio, não existirão informações disponíveis sobre os eventos caracterizados como novidades, uma vez que o especialista responsável pela rotulação (identificação) dos padrões não terá conhecimento consolidado sobre esses novos padrões.

Nesse sentido, a utilização de métodos de aprendizado supervisionado, ou mesmo semi-supervisionado (Miller and Browning, 2003), é inviabilizado no contexto proposto. Uma alternativa para prover um auxílio ao especialista no processo de identificação, e até mesmo caracterizar novas classes presentes nos dados classificados como novidades, é aplicar modelos de agrupamento automático (Filippone et al., 2008). Esses modelos, caracterizados por serem fundamentalmente não-supervisionados, permitem que sejam identificados e agrupados eventos com certo grau de similaridade, dentro do conjunto de padrões definidos como novidades.

Sendo assim, a proposta desse trabalho é aplicar modelos de agrupamento automático em oscilografias de tensão de redes de distribuição, utilizando a informação de um dado agrupamento para indicar a existência de um novo conceito, ou classe, dentro de um problema de classificação. Vale ressaltar que nessa modelagem, uma classe não é necessariamente formada por apenas um agrupamento, podendo inclusive, ser representada por inúmeros agrupamentos. No entanto, a identificação de um dado agrupamento pode ser um indício de que existe uma classe, independentemente de quantos agrupamentos formam essa classe, auxiliando o especialista no processo de identificação de novas classes.

Convém ressaltar que essa é uma abordagem não convencional para classificação de formas de onda em sistemas de potência, tal qual é observado em Faiz et al. (2007) e Lazzaretti et al. (2013). Este tipo de abordagem se aproxima de modelos de classificação existentes para problemas de visão computacional, como se observa em Miller and Browning (2003) e Spinosa and Carvalho (2005). Os autores deste trabalho não identificaram trabalhos similares utilizando dados de oscilografia em sistemas elétricos de potência, assinalando a originalidade do trabalho aqui proposto.

2 Fundamentação Teórica

A fundamentação teórica deste trabalho baseia-se, fundamentalmente, em uma revisão bibliográfica dos métodos de agrupamento automático, assinalando as principais diferenças nas modelagens de cada método. Demais detalhes teóricos da metodologia proposta serão indicados por meio de referências ao longo do trabalho.

2.1 Agrupamento Automático

Dentre os diversos métodos de agrupamento automático atualmente utilizados para problemas de aprendizagem de máquina, destacam-se basicamente dois agrupamentos: modelos probabilísticos e modelos baseados em *kernel* (Filippone et al., 2008). Nos métodos probabilísticos, busca-se uma representação dos dados através da sua função densidade de probabilidade. Em geral, o número de funções a serem utilizadas (normalmente Gaussianas) é selecionado de forma automática, bem como o seus respectivos parâmetros (Cheung, 2005; Zeng and Cheung, 2009).

Um modelo de agrupamento automático dessa natureza que merece destaque é o modelo proposto por Cheung (2005), denominado *Rival Penalized Expectation Maximization* (RPEM). Esse modelo é baseado na Maximização da Expectativa para determinação dos parâmetros das funções Gaussianas que caracterizam a função densidade de probabilidade à luz dos dados. No entanto, o autor propõe uma modificação na formulação convencional, visando a seleção automática do modelo. Essa modificação consiste em ponderar a função de verossimilhança (Cheung, 2005), de forma que seja possível realizar uma competição entre cada componente individual da mistura de Gaussianas, atualizando os parâmetros da função Gaussiana vencedora e de todas as demais funções que formam a distribuição. A grande vantagem desse modelo, na sua versão atual (Zhao et al., 2010), é que não é necessário especificar parâmetros *a priori*. Apenas determina-se o intervalo para o número de agrupamentos que serão considerados para realizar a modelagem. No final, é possível descartar os agrupamentos que não contribuem para o resultado, uma vez que todos esses agrupamentos são vazios (não contêm padrões associados a eles).

Já os métodos baseados em *kernel*, do inglês *Support Vector Clustering* (SVC), utilizam o mapeamento (implicitamente definido) para um espaço de características através de um *kernel*, sendo que o processo de agrupamento ocorre nesse novo espaço (Wang et al., 2011). Esses métodos possuem grande flexibilidade para a determinação automática do número de agrupamentos, mas ainda representam forte dependência dos hiperparâmetros associados ao processo de treinamento dos modelos, principalmente em relação ao *kernel*.

Um método de agrupamento que guarda relação com os modelos baseados em *kernel* e merece destaque dentro da possibilidade de ser utilizado no contexto de agrupamento automático é conhecido como método espectral (Filippone et al., 2008). Esses modelos são baseados na teoria de grafos, em que a ideia básica é construir um grafo ponderado utilizando os dados de entrada, de tal forma que cada nó represente o padrão de entrada e as conexões ponderadas representem a similaridade entre cada par de padrões de entrada. Com isso, o problema de agrupamento pode ser visto como um problema de partição de grafos, que pode ser resolvido através do uso da teoria espectral de grafos. Pode-se demonstrar que os métodos espectrais e os métodos baseados em *kernel* possuem a mesma fundamentação matemática (Filippone et al., 2008).

Dentro dos modelos de agrupamento espectral, existe uma proposta que aborda a possibilidade de realizar o agrupamento de forma automática, conhecido como *Sef-Tuning Spectral Clustering* (STSC) (Zelnik-Manor and Perona, 2004). Esse método propõe realizar um processo de escalonamento local para calcular a similaridade entre os pares de padrões de entrada. Além disso, é incluída uma análise da estrutura dos autovetores da matriz de similaridade para inferir o número de agrupamentos. Esse modelo foi testado para problemas de segmentação em imagens e se observou que o desempenho pode ser comparável com demais métodos de agrupamento automático.

Um último modelo que merece destaque aqui, é o modelo *X*-Médias. Esse algoritmo utiliza como base o agrupamento *K*-Médias. No entanto, ele insere a possibilidade de realizar a determinação automática e dinâmica do número de agrupamentos (Pelleg and Moore, 2000). No modelo *X*-Médias, os centros são determinados levando em conta o critério de seleção de informação Bayesiana (BIC). As contribuições do *X*-Médias podem ser resumidas em:

- O *K*-Médias pode ser considerado um algoritmo relativamente lento, no que diz respeito ao tempo que leva para finalizar cada iteração. Nesse contexto, o modelo *X*-Médias propõe a utilização de um procedimento para aceleração de cada iteração, conforme exposto em (Pelleg and Moore, 2000);

- No caso do *K*-Médias, é necessário informar previamente o número de agrupamentos (K) a serem escolhidos pelo algoritmo. Já no modelo *X*-Médias, os centros são determinados de forma automática, levando em conta o critério de seleção *Bayesiano* de modelo (BIC);

- Quando K é mantido fixo ao longo da execução do algoritmo, pode-se demonstrar (de forma empírica) que o resultado tende a um valor ótimo local inferior ao valor obtido quando K é alterado dinamicamente (Pelleg and Moore, 2000). Nesse sentido, o modelo *X*-Médias apresenta um procedimento para alterar o número de agrupamentos dinamicamente, também utilizando o critério de seleção *Bayesiano* de modelo (BIC).

3 Metodologia

A metodologia deste trabalho contempla inicialmente os dados a serem utilizados para avaliação dos modelos de agrupamento. São utilizadas duas bases de dados, sendo uma delas formada por dados simulados no *Alternative Transient Program* (ATP), e outra por dados reais. Posteriormente, são discutidas as formas de pré-processar a extrair as características das formas de onda. Por último, é apresentada a forma de avaliar e comparar o acerto médio na classificação dos modelos de agrupamento.

3.1 Dados Simulados

Inicialmente, foi utilizado o ATP para modelar uma subestação de distribuição escolhida para a análise e gerar a base de dados dos eventos a serem classificados. Esses eventos foram escolhidos com base nos eventos mais comuns (curto-circuitos e eventos associados com qualidade de energia) que ocorrem nos alimentadores dessa subestação e correspondem a um total de 29 tipos diferentes de classes: falta monofásica para a terra (nas fases A, B e C), faltas bifásicas para a terra (nas fases AB, BC e CA), faltas trifásicas para a terra, faltas bifásicas (nas fases AB, BC e CA), faltas trifásicas sem envolver terra, abertura de um alimentador (nas fases A, B, C, AB, BC, CA e trifásica), religamento em um alimentador (nas fases A, B, C, AB, BC, CA e trifásico), chaveamento de banco de capacitores, partida de grandes motores, curto-circuito monofásico a montante da subestação e curto-circuito bifásico a montante da subestação. Mais detalhes desse modelamento podem ser encontrados em Lazzaretti et al. (2009).

Para os conjuntos de treino e teste utilizados nos métodos de agrupamento, foram variados os instantes de ocorrência dos eventos, as curvas de carga dos alimentadores, a distância da subestação em que ocorrem os eventos (de pontos mais próximos até pontos mais distantes) e, no caso das faltas, foi variada a resistência de falta, resultando em um total de aproximadamente 36 instâncias para cada classe para o conjunto de treinamento e 18 instâncias por classe para o conjunto de teste (67% para treinamento e 33% para teste).

3.2 Dados Reais

O sistema de monitoramento utilizado nesse trabalho está fundamentado no trabalho proposto por Lazzaretti et al. (2011). Os registros armazenados forneceram a possibilidade de teste dos métodos de agrupamento automático. Esse sistema permite o registro simultâneo de transitórios decorrentes de sobretensões atmosféricas e demais eventos, em todas as fases da rede de média tensão.

Foram instalados quatro protótipos do sistema de monitoramento, sendo dois deles em Apucarana-PR e dois em São Mateus do Sul-PR. Em aproximadamente um ano de operação deste sistema na rede de distribuição, observou-se a ocorrência de 153 variações significativas nas formas de onda, sendo que aproximadamente 67% dessas variações estão relacionadas com descargas atmosféricas, 14% com eventos de curto-circuito, 8% com manobras no alimentador e 11% com outros eventos, dentre os quais destacam-se os eventos cuja caracterização foi inviabilizada em função da complexidade na análise e a falta de informação sobre as condições que geraram o registro do evento, como será descrito posteriormente.

Devido à grande diferença no número de exemplos por classe, optou-se por incluir dados de curto-circuito e religamento trifásico simulados, de forma que essa diferença fosse reduzida. A utilização desses dois eventos em conjunto com a base de dados real baseia-se na validação realizada para esses eventos no modelo utilizado em (Lazzaretti et al., 2011).

3.3 Pré-Processamento das Formas de Onda

Uma vez estabelecidos os eventos (reais e simulados) e seus respectivos rótulos, os sinais de tensão foram pré-processados utilizando a Transformada *Wavelet Packet* Discreta (TWPD) em quatro níveis para a tensão das três fases na barra da subestação com uma taxa de amostragem de 15360 Hz, com objetivo de extrair o máximo de características dos sinais em análise, minimizando a perda de informação relevante. A *wavelet*-mãe utilizada na decomposição foi a Daubechies-20 ('db20'). A justificativa para a utilização dessa *wavelet*-mãe é que *wavelets* dessa família já são utilizadas para classificar eventos dessa natureza, apresentando bons resultados (Lazzaretti et al., 2009). Com o intuito de reduzir a dimensionalidade da entrada para o estágio de classificação, foi efetuado o cálculo da energia nas várias sub-bandas da TWPD proposto em (Lazzaretti et al., 2013), resultando em um vetor de entrada para os modelos de agrupamento com 48 elementos (16 valores para cada uma das três fases).

3.4 Análise do Acerto Médio de Classificação dos Métodos de Agrupamento

Para avaliar o acerto médio de classificação[1] de um método de agrupamento existem diferentes métricas, conhecidas como índices de validação de agrupamento (Filippone et al., 2008). Dentre as possibilidades, optou-se por utilizar nesse trabalho uma métrica de validação externa baseada na matriz de confusão dos agrupamentos. A justificativa para utilização dessa métrica é que torna-se possível utilizar o padrão-ouro dos dados como referência externa, com o intuito de verificar a capacidade do método de agrupamento em colocar eventos de uma mesma classe em um mesmo agrupamento (Theodoridis and Konstantinos, 2009). Assim, espera-se que para um novo conjunto de dados, o método de agrupamento seja capaz de manter eventos de uma nova classe em um (ou eventualmente mais de um) agrupamento, indicando através disso, a possibilidade de existência dessa nova classe e as similaridades existentes entre os padrões. Mais detalhes desta técnica podem ser encontradas em Theodoridis and Konstantinos (2009).

[1] Acerto médio de classificação no contexto do presente trabalho representa a média entre os acertos obtidos para cada classe, minimizando a influência do desbalanceamento do número de padrões por classe, tal qual se observa ao se utilizar a acurária, por exemplo.

4 Seleção do Método de Agrupamento

Para seleção do método de agrupamento utilizado como parte do processo de identificação de novas classes, foram escolhidos previamente os quatro diferentes modelos: RPEM, STSC, *X*-Médias e SVC. A base de dados inicialmente selecionada consiste de três classes de curto-circuito monofásico, geradas através de simulação no ATP, com 54 padrões por classe de curto-circuito (fase A, B ou C).

O objetivo de trabalhar com uma base de dados com poucas classes é facilitar o processo de interpretação dos resultados em um primeiro momento, apontando as características relevantes e limitações de cada método para um problema cuja separabilidade e possibilidade de agrupamento são maiores, quando comparados a um conjunto com um número elevado de classes.

Utilizando as características acima descritas, foram aplicados os quatro métodos de agrupamento automático na base de dados com as três classes de curto-circuito monofásico. A figura 1 mostra o resultado dessa simulação, onde está representado o acerto médio de classificação dos quatro métodos de agrupamento automático listados anteriormente, para as três classes de curto-circuito monofásico. Além disso, está representado o acerto médio de classificação para o modelo *K*-Médias para diferentes centros, com o intuito de fornecer uma referência para comparação entre os modelos.

Para compreender como a curva da figura 1 pode fornecer uma referência de comparação para os métodos de agrupamento, é necessário primeiramente analisar o seu comportamento, no limite em que o número de centros definidos tende a ser igual ao número de padrões disponíveis para o agrupamento. Conforme descrito anteriormente, a métrica de validação externa baseada na matriz de confusão dos agrupamentos não leva em conta o número de agrupamentos que um dado método produz. Ou seja, no limite em que o número de centros é igual ao número de padrões, o acerto do agrupamento é 100% para essa métrica, conforme é possível observar na figura 1.

Analisando a resposta obtida para o modelo *K*-Médias para diferentes centros, observa-se que existe uma certa saturação no acerto médio de classificação obtido a partir de, aproximadamente, cinco centros. Isso indica que a inserção de novos centros não acrescenta informações relevantes para identificação de no-

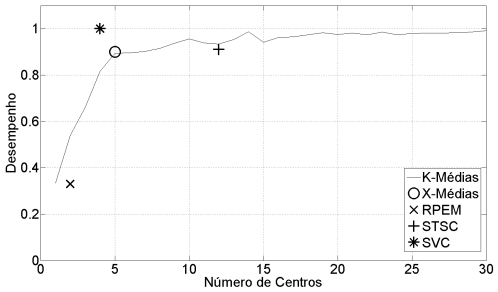

Figura 1: Comparação do Acerto Médio de Classificação para Diferentes Métodos de Agrupamento para as Três Classes de Curto-Circuito Monofásico. Nesse gráfico, fica claro o comportamento assintótico da acurácia na classificação do modelo *K*-Médias, no limite em que o número de agrupamentos tende a ser igual ao número de padrões por classe. Além disso, observa-se que os modelos *X*-Médias, SVC e STSC possuem as melhores acurácias na classificação.

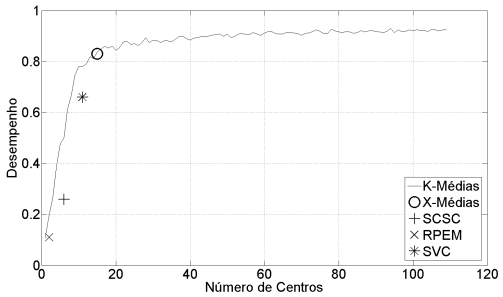

Figura 2: Comparação de Acerto Médio de Classificação para Diferentes Métodos de Agrupamento para todas as Classes de Curto-Circuito. Nesse gráfico, novamente observa-se comportamento assintótico da acurácia do modelo *K*-Médias, no limite em que o número de agrupamentos tende a ser igual ao número de padrões por classe. Nesse caso porém, os modelos *X*-Médias e SVC apresentaram as melhores acurácias na classificação do agrupamento.

vas classes através dos agrupamentos formados, podendo ser inclusive, considerada desnecessária para a representação em questão. Dessa forma, é de se esperar que qualquer modelo de agrupamento automático que apresente um bom acerto médio de classificação no agrupamento tenha o seu resultado localizado na região do início da saturação ('joelho') da curva do modelo *K*-Médias, desde que seja aplicado aos mesmos dados modelados pelo *K*-Médias. Essa região do início de saturação indica o compromisso entre um bom acerto médio de classificação e um número de centros adequado ao número de classes do problema. Nesse sentido, pode-se afirmar que a curva do modelo *K*-Médias serve de referência para a comparação proposta.

Com isso, é possível observar que em princípio, tanto o modelo *X*-Médias, quanto o modelo de agrupamento SVC são candidatos mais adequados ao problema de agrupamento em questão. No entanto, para obter uma ideia mais clara da capacidade de agrupamento de cada método, optou-se por avaliar o acerto médio de classificação para uma base de dados contendo todas as onze classes de curto-circuito, geradas via simulação no ATP (todas com 54 padrões por classe). Os resultados dessa análise estão representados na figura 2, onde novamente pode-se observar que os melhores acertos de classificação são dos modelos *X*-Médias e SVC.

Para detalhar as limitações e as vantagens dos resultados apresentados nas figuras 1 e 2 é necessário analisar cada um dos quatro métodos de forma individual.

Com relação ao método SVC, é possível observar que o aumento do número de classes leva a uma queda considerável no acerto médio de classificação. No caso das três classes de curto-circuito monofásico, é possível obter um acerto médio de classificação de 100% utilizando apenas quatro agrupamentos (vale lembrar que o número de grupos pode não ser exatamente igual ao número de classes). No entanto, quando se consideram todas as onze classes de curto-circuito, o acerto médio de classificação máximo obtido foi de aproximadamente 67%. Observou-se que para esse modelo, a sensibilidade do ajuste dos parâmetros do *kernel* e de controle de complexidade, impactam de forma significativa no formato e na quantidade de agrupamentos gerados, além do fato desse tipo de modelo obter melhores desempenhos para conjuntos com poucas classes. Dessa forma, o resultado final do agrupamento fica comprometido, do ponto de vista de identificação de possíveis classes através dos agrupamentos. Muito embora o modelo determine de forma automática o formato e o número de agrupamentos em uma dada distribuição de dados, a forte dependência da seleção do parâmetro do *kernel* e de controle de complexidade faz com que sua utilização no contexto de agrupamento automático seja dificultada e, para determinados arranjos de classes, inviável.

As mesmas observações apontadas para o método SVC com relação à seleção dos parâmetros, valem para o método STSC. Uma diferença fundamental no entanto, é que esse último permite a determinação automática dos parâmetros envolvidos do processo de agrupamento espectral. Porém, os testes realizados demonstraram que nem sempre a seleção automática proposta pelos autores leva a um acerto médio de classificação superior a uma seleção manual (via validação cruzada, por exemplo). Além disso, a sensibilidade na escolha dos parâmetros impacta diretamente no acerto médio de classificação final. Para ambas as bases de dados avaliadas, os melhores o acertos de classificação obtidos para diferentes parâmetros avaliados estão representados nas figuras 1 e 2.

Já para o modelo RPEM, existem dois fatores que interferem no desempenho final: o número de padrões por classe e a dimensionalidade dos vetores de en-

trada. Para analisar essa questão, optou-se por utilizar a base de dados sintética proposta em (Cheung, 2005), incluindo três funções Gaussianas bem separadas (sem sobreposição) para diferentes dimensionalidades. Sobre essa base de dados foi aplicado um agrupamento automático baseado no modelo RPEM. Para todos os casos, é de se esperar que o modelo forneça a caracterização das três funções Gaussianas, utilizando apenas um centro para cada função, sendo que os demais centros podem ser desprezados na modelagem. A tabela 1 mostra o resultado do agrupamento RPEM, indicando o número de padrões definido por agrupamento, para 200 padrões divididos de forma equilibrada entre as três funções Gaussianas (aproximadamente 66 padrões por classe). Esse processo foi repetido para diferentes dimensionalidades dos dados de entrada.

Tabela 1: Resultado para o Método RPEM com 200 padrões Divididos em Três Agrupamentos. Repara-se que com mais de 20 dimensões o modelo de agrupamento passa a representar todas as três classes por praticamente uma classe.

Número de Dim.	Núm. de Padrões Agrup. 1	Núm. de Padrões Agrup. 2	Núm. de Padrões Agrup. 3
2-D	61	66	73
3-D	61	73	66
4-D	61	73	66
5-D	61	66	73
10-D	61	66	73
20-D	6	0	194
50-D	200	0	0

Observa-se que dados de entrada a partir de vinte dimensões, passam a ser modelados por praticamente uma função Gaussiana. Isso pode ser resultado da esparsidade existente nos dados com alta dimensionalidade, uma vez que o número de padrões por distribuição Gaussiana permanece constante ao longo de todas as simulações. Com um número maior de padrões por classe (aproximadamente 300 padrões por classe), esse problema passa a surgir apenas com um número elevado de dimensões (aproximadamente 50), o que reforça a possibilidade desse resultado estar associado com a esparsidade dos dados em cenários de alta dimensionalidade e a dificuldade em modelar essas distribuições utilizando funções Gaussianas. Essa característica pode justificar a limitação do modelo RPEM em no máximo dois agrupamentos, conforme mostram as figuras 1 e 2.

Por último, o modelo baseado no agrupamento *X*-Médias apresentou o acerto de classificação mais próximo do início da saturação da curva de desempenho do *K*-Médias. De certa forma, esse resultado era esperado, uma vez que o *X*-Médias utiliza como base o agrupamento *K*-Médias. No entanto, uma característica importante é que o modelo *X*-Médias realiza de forma adequada o controle de complexidade, mantendo o número de agrupamentos dentro da região limite esperada para as duas bases de dados utilizadas.

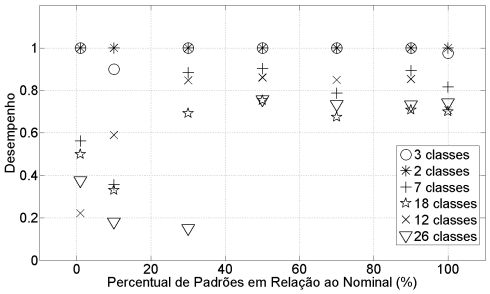

Figura 3: Comparação do Acerto Médio de Classificação para Diferentes Números de Padrões por Classe e Diferentes Classes Utilizando o Método *X*-Médias. Para essa comparação, foram utilizadas bases de dados reduzidas em relação à base de dados original, utilizando os valores de 2%, 10%, 30%, 50%, 70%, 90% e 100% dos padrões, em relação ao número total de padrões disponíveis (base de dados original). Para cada caso, foi avaliada a acurácia para diferentes arranjos e números de classes, dentre as 29 classes de eventos disponíveis por simulação. Com isso, é possível avaliar o impacto do número de padrões por classe, o arranjo e o número de classes na acurácia de classificação do agrupamento.

Analisando os critérios acima discutidos, optou-se por utilizar o modelo *X*-Médias para o agrupamento automático associado ao processo de identificação de classes. Acredita-se que esse modelo seja capaz de atender às restrições impostas para o problema de classificação e agrupamento em questão, principalmente em termos de desempenho, além de não necessitar uma intervenção na seleção de parâmetros para os diferentes arranjos de classes avaliados.

Por fim, existem dois fatores que merecem destaque dentro do escopo proposto, relacionados com a identificação de classes através de métodos de agrupamento: número de padrões por classe e número de classes existentes no momento do agrupamento automático. Nesse sentido, uma análise bastante relevante diz respeito à capacidade do método de agrupamento em identificar os agrupamentos, mesmo em casos em que o número de padrões por classe seja reduzido. A figura 3 mostra o resultado do acerto médio de classificação obtido para diferentes números de padrões por classe (em termos percentuais do valor nominal - 100%) e diferentes arranjos de classes.

Observa-se que para os casos em que existem poucas classes, o número de padrões por classe não interfere no desempenho de classificação do agrupamento. Porém, quando o número de classes aumenta, o acerto médio de classificação passa a ser fortemente afetado, para os casos em que estão disponíveis poucos padrões por classe.

5 Resultados

Aplicando o método *X*-Médias em 100 diferentes arranjos das 29 classes de eventos simulados (tipo e número de classes escolhidos aleatoriamente), obtém-se um acerto médio de classificação de 73,4% ($\pm$ 11,6) para o conjunto de treinamento e 73,4% ($\pm$ 12,5) para o conjunto de teste. Observa-se que o acerto médio de classificação obtido para o conjunto de treinamento é muito próximo ao acerto médio de classificação obtido para o conjunto de teste, uma vez que esses conjuntos foram gerados de forma muito similar para os dados simulados. O teste de Wilcoxon indica que as duas distribuições são equivalentes e que, portanto, os desempenhos podem ser considerados equivalentes. Esse resultado é justificado em função dos dados de treinamento e teste terem sido gerados de forma bastante similar. Se considerado o resultado do agrupamento automático aplicado apenas nas classes de curto-circuito, o acerto médio de classificação para o conjunto de teste é próximo a 83% ($\pm$ 8,7).

Para os dados reais, aplicando o modelo de agrupamento *X*-Médias, o acerto médio de classificação é de 80% para as cinco classes conhecidas desta base de dados. Vale lembrar que, em geral, métodos supervisionados seriam mais adequados para o problema em questão. No entanto, como o objetivo do presente trabalho é avaliar o desempenho de um método de agrupamento quando não se conhecem as classes, optou-se por utilizar um conjunto de exemplos rotulados de modo a facilitar a comparação de desempenho dos modelos de agrupamento. Nesse sentido, o desempenho de 80% nesse caso é bastante relevante considerando que não se conhecem os rótulos dos exemplos durante a etapa de treinamento dos modelos.

A matriz de confusão para esse conjunto de dados está representada na tabela 2. Neste caso, 11 agrupamentos foram necessários para representar as cinco classes, sendo três para a classe Descarga Atmosférica Induzida (DAI) e dois para cada uma das demais.

Para os dados reais, assim como apresentado para os dados simulados, é possível afirmar com base nos acertos de classificação apresentados, que o agrupamento automático pode operar como parte do processo de identificação de novas classes.

Para exemplificar o uso do agrupamento automático no contexto de identificação de classes, é possível utilizar o processo de classificação de DAI, proposto em (Lazzaretti et al., 2011). Naquele trabalho, os autores desconheciam o comportamento desse evento *a priori*, uma vez que o registro deste tipo de evento ainda é pouco realizado em redes de distribuição.

Para realizar a classificação, os autores utilizaram informações de sistemas de localização de descargas atmosféricas. No entanto, em função de problemas de sincronismo, não foi possível utilizar essa correlação para classificar grande parte dos eventos. Com isso, os especialistas responsáveis por classificar as formas de onda levaram em conta o resultado do agrupamento automático. Esse resultado forneceu a possibilidade de identificar outros padrões de DAI através do indicativo de similaridade dos modelos de agrupamento. Dessa forma, os autores puderam extrair ao máximo as informações contidas na base de dados analisada.

Tabela 2: Matriz de Confusão para os Dados Reais. Analisando essa matriz de confusão, fica claro que, na sua grande maioria, os eventos de uma mesma classe foram agrupados nos agrupamentos definidos pelo modelo *X*-Médias, muito embora tenham sido utilizados mais que um grupo por classe em alguns casos. A acurácia global de classificação é definida através da média do acerto de classificação por classe.

Classes	Falta 1f-A	Falta 1f-B	Falta 1f-C	Relig. 3f	DAI
Falta 1f-A	33	0	4	0	0
Falta 1f-B	3	28	7	0	3
Falta 1f-C	3	0	36	0	1
Relig. 3f	0	0	1	20	5
DAI	2	0	5	0	19

6 Conclusões

Quando se realiza um processo de automação da análise de dados armazenados em qualquer base de dados, dá-se um passo importante na direção de transformar dados brutos em informações de uso imediato, ou quase imediato. Essa é a motivação principal da metodologia aqui proposta, que visa identificar similaridades entre padrões de oscilografias de redes de distribuição, através de modelos de agrupamento automático.

Foram utilizadas bases de dados simuladas e reais, com eventos conhecidos, com o intuito de verificar o desempenho da proposta para um cenário no qual se conhecem os rótulos do padrões a serem agrupados por similaridade. Com isso, foi possível verificar que os modelos promovem o agrupamento de padrões pertencentes a uma mesma classe, mesmo que em determinados casos utilizem mais que um agrupamento para definir a classe. Essa característica pode ser fundamental em cenários em que se exige uma análise preliminar da possibilidade de novos eventos geradores de oscilografias (distúrbios), identificando e associando eventos similares, principalmente com o uso em massa de registradores de oscilografia, tal qual vem ocorrendo nas novas Redes Inteligentes, configurando um cenário característico de *Big Data*.

Agradecimentos

Os autores agradecem à Copel, ao LACTEC, a ao programa de Pesquisa e Desenvolvimento da ANEEL pelo apoio financeiro.

Referências

Cheung, Y.-M. (2005). Maximum Weighted Likelihood via Rival Penalized EM for Density Mixture Clustering with Automatic Model Selection, *IEEE Transactions on Knowledge and Data Engineering* **17**: 750–761.

Faiz, J., Lotfi-fard, S. and Shahri, S. H. (2007). Prony-based optimal Bayes fault classification of overcurrent protection, *IEEE Transactions on Power Delivery* **22**: 1326–1334.

Filippone, M., Camastra, F., Masulli, F. and Rovetta, S. (2008). A survey of kernel and spectral methods for clustering, *Pattern Recognition* **41**: 176–190.

Lazzaretti, A. E., Ferreira, V. H., Vieira Neto, H., Riella, R. J. and Omori, J. (2009). Classification of events in distribution networks using autonomous neural models, *Proceedings of the 15th International Conference on Intelligent System Applications to Power Systems*.

Lazzaretti, A. E., Ferreira, V. H., Vieira Neto, H., Toledo, L. F. R. B. and da Silva Pinto, C. L. (2013). A new approach for event classification and novelty detection in power distribution networks, *Proceedings of the IEEE PES General Meeting 2013, IEEE PES Press.*

Lazzaretti, A. E., Ravaglio, M. A., Toledo, L. F. R. B., Teixeira-Júnior, J. A., Rojas, P. M. and Pinto, C. L. S. (2011). Measurements of lightning discharges in overhead distribution feeders, *Anais do XI Simpósio Internacional Proteção Contra Descargas Atmosféricas*.

Miller, D. and Browning, J. (2003). A mixture model and EM-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**: 1468–1483.

Pelleg, D. and Moore, A. (2000). X-means: Extending K-means with efficient estimation of the number of clusters, *Proceedings of the 17th International Conference on Machine Learning*.

Spinosa, E. J. and Carvalho, C. P. L. F. (2005). Combining one-class classifiers for robust novelty detection in gene expression data, *Advances in Bioinformatics and Computational Biology* **3594**: 54–64.

Theodoridis, S. and Konstantinos, K. (2009). *Pattern Recognition*, Academic Press.

Wang, C.-D., Lai, J.-H. and Huang, D. (2011). Kernel-based clustering with automatic cluster number selection, *Proceedings of the 11th International Conference on Data Mining Workshops*.

Zelnik-Manor, L. and Perona, P. (2004). Self-Tuning Spectral Clustering, *Advances in Neural Information Processing Systems 17*, Vol. 2.

Zeng, H. and Cheung, Y.-M. (2009). A new feature selection method for Gaussian mixture clustering, *Pattern Recognition* **42**: 243–250.

Zhao, X.-M., Cheung, Y.-M. and Huang, D.-S. (2010). Analysis of Gene Expression Data Using RPEM Algorithm in Normal Mixture Model With Dynamic Adjustment of Learning Rate, *International Journal of Pattern Recognition and Artificial Intelligence* **24**: 651–666.