Abordagens de aprendizado de máquina aplicados aplicadas na classificação da ocorrência de incêndios

Marco Antonio Reichert Boaretto
Programa de Pós-Graduação em Engenharia Elétrica, Departamento de Engenharia Elétrica, Universidade Federal do Paraná (UFPR), Rua Cel. Francisco Heráclito dos Santos, 100, Cep: 81531-980, Curitiba, PR, Brasil
marco.boaretto@hotmail.com

Rafael Busatto
Programa de Pós-Graduação em Engenharia Industrial e Sistemas (PPGEPS), Pontifícia Universidade Católica do Paraná (PUCPR), Imaculada Conceição, 1155, Cep: 80215-901, Curitiba, PR, Brasil
rafael_busatto@yahoo.com.br

Leandro dos Santos Coelho
Programa de Pós-Graduação em Engenharia Industrial e Sistemas (PPGEPS), Pontifícia Universidade Católica do Paraná (PUCPR), Imaculada Conceição, 1155, Cep: 80215-901, Curitiba, PR, Brasil
Programa de Pós-Graduação em Engenharia Elétrica, Departamento de Engenharia Elétrica, Universidade Federal do Paraná (UFPR), Rua Cel. Francisco Heráclito dos Santos, 100, Cep: 81531-980, Curitiba, PR, Brasil
lscoelho2009@gmail.com

RESUMO

Os problemas relacionados com incêndios urbanos acontecem trazendo destruição, perda de vidas e problemas econômicos ao país. Com o intuito de reduzir os riscos especialmente em áreas populosas, a elaboração de sistemas de previsão e/ou classificação de incêndios são necessários para detectar esses eventos. Neste artigo é apresentado um estudo de caso relacionado à ocorrência de incêndios aliado a um conjunto de abordagens de aprendizado de máquina para construir um comitê de máquinas. Referente ao estudo de caso avaliado, dados reais de incêndios dos anos 2013 e 2014 foram obtidos com o Corpo de Bombeiros Militar do Estado do Paraná na cidade de Curitiba-Paraná. Os resultados obtidos contribuíram em ajudar o Corpo de Bombeiros do Estado do Paraná, em especial ao Departamento de Bombeiros de Curitiba, a agilizar o processo de reconhecimento de localização dos incidentes de incêndio, melhorando o tempo de resposta no atendimento.

PALAVRAS CHAVE. Incêndios Urbanos, Aprendizado de Máquina, Comitê de Máquinas.

ABSTRACT

The problems related with urban fires brings destruction, casualties and economic issues to the country. On aiming to decrease the risk specially on populous areas, the development of classification and/or forecasting systems are needed to detect this events. On this article is presented a new case of study related with fire incidents and allied with a group of machine learning approaches to build an ensemble of machine learning techniques. Regarding the case of study, Data from real fire incidents of the years 2013 and 2014 were gathered with the Fire Department of the state of Paraná on the city of Curitiba-Paraná. The results could contribute with the Fire Department of the state of Paraná, specially the Fire Department from Curitiba, to speed up the location reckoning process of the fire incidents, improving the answering time of the calling.

KEYWORDS. Urban Fires, Machine Learning, Ensemble.

1. Introdução

Os incêndios são uma realidade na vida social desde os mais remotos períodos em que se têm registro, podendo causar grandes consequências econômicas ao país afetado. Em 2015, os Estados Unidos da América (EUA) gastaram U$ 14.3 bilhões no combate à incêndios [NFPA, 2016]. Anualmente é gasto nos EUA de 0,813% do PIB (Produto Interno Bruto), na Dinamarca, 0,864%, no Reino Unido o valor é cerca de 0,729%. No Canadá em 2015 o custo devido a incidências de incêndios e suas consequências atingiu o patamar de U$ 730,5 milhões [MCSCS, 2015].

Sem deixar de mencionar um incêndio histórico ocorrido nos EUA no ano de 1989, no Texas. A empresa Philips Petroleum's Pasadena fora incendiada, foi considerado o quarto maior incêndio da história americana até então, com um custo de U$ 750 milhões. Neste caso, a empresa ainda ficou dois anos em reconstrução, o que acarretou um custo de aproximadamente U$ 1 bilhão.

Atualmente fatores relacionados à urbanização acelerada de algumas cidades de modo desorganizado alinhado às condições climáticas tem gerado problemas que envolvem a prestação de serviços públicos adequados quando relacionado ao incêndio urbano, causando danos materiais e econômicos relevantes associados à perda de vidas humanas. Além disso, o tempo resposta de atendimento está ligado ao crescimento urbano, à expansão urbana é um fator importante na influência de questões e resultados de combate a incêndios nos EUA [Lambert *et al.*, 2012], A resposta de emergência do serviço de bombeiros ao incêndio está na base de que quanto mais cedo o incêndio for atacado menor serão as consequências para as pessoas e propriedade [Challands, 2009].

Muitas regiões propensas a incêndios não apenas aquelas com clima mediterrâneo, sofreram perdas significativas de vidas e de bens nos últimos anos devido à crescente urbanização e à frequência de incêndio [Pausas *et al*., 2008; Veblen *et al*., 2008; Gill *et al*., 2013; Penman *et al*., 2013; Keeley *et al*., 2012].

Deve-se enfatizar ainda que a propensão à ocorrência de incêndios aumenta a dedicação na geração sistemas eficientes de prevenção, estratégias de ação e gestão dos incêndios. A detecção de eventos é um problema de previsão, ou, tipicamente, um problema de classificação de dados [Haixiang *et al.*, 2016].

Para o estudo de caso, um conjunto de abordagens de Aprendizado de máquina (*Machine learning*) foram selecionadas dentre uma ampla variedade disponível de técnicas para construir um comitê de máquinas. Neste artigo foram adotados quatro métodos para a concepção do comitê de máquinas, são elas: (i) máquinas de vetor de suporte (do inglês, *Support vector machine*, SVM), (ii) *random forest* (RF), (iii) k-vizinhos mais próximos (K-*nearest neighbors*, KNN), e (iv) extreme gradient boosting (XGBoost).

As SVM [Vapnik, 1995] consistem em técnicas baseadas em aprendizado estatístico e são muito utilizadas em reconhecimento de padrões [Bouzalmat *et al*., 2014; Gonçalves, 2009; José e Ribeiro, 2012; Pradhan, 2012].

O *Random Forest* (RF), desenvolvido por [Breiman, 2001], é um comitê de técnicas que combina a abordagem de reamostragem *bagging* para o treinamento de dados em conjunto com a seleção de características aleatórias para criar um grupo de árvores de decisão com variação controlada [Fawagreh, Gaber, e Elyan, 2014], são robustas a ruído e sobre treinamento, sendo o RF aplicado a problemas de classificação em diferentes áreas do conhecimento[Cabras, Castellanos, e Staffetti, 2016; Cutler *et al*., 2007; Izmirlian, 2004].

O método K-Vizinhos mais próximos (KNN), é um algoritmo de aprendizagem baseado em instâncias introduzido por Cover e Hart [1967], é muito utilizado em diversos problemas de classificação pela sua simplicidade e eficiência tanto em problemas de aprendizado supervisionado como não supervisionado [Athitsos e Sclaroff, 2005; Hwang, 1998; Mejdoub e Ben Amar, 2013; Tan, 2006]

Extreme Gradient Boosting (XGBoost), implementado por [Chen e Guestrin, 2016] é uma abordagem recentemente melhorada baseada nos estudos prévios de [Friedman, 2001]. O XGBoost

venceu várias competições sediadas pelo site *Kaggle* [Chen e Guestrin, 2016; Holloway e Marks, 2016] e provou ser uma eficiente ferramenta de aprendizado de máquina.

A principal contribuição deste artigo está relacionada à abordagem de aprendizado de máquina proposta combinando os quatro métodos mencionados. O desempenho dos métodos é comparado pela sua aplicação isolada como também um comitê de máquinas (ver detalhes sobre o projeto de comitês de máquinas em [Ren, Zhang, e Suganthan, 2016]).

Os comitês de máquina podem ser promissores paradigmas, pois podem obter desempenhos melhores que dos métodos caso estes fossem adotados isoladamente. A abordagem de comitê de máquinas poderá ajudar o corpo de bombeiros de Curitiba a agilizar o processo de reconhecimento de localização dos incidentes, melhorando o tempo de resposta no atendimento. Além disso, em termos de utilização prática dos resultados obtidos com este artigo apresenta uma nova base de dados com dados reais obtidos com o corpo de bombeiros de Curitiba-Pr, que contêm não só informações climáticas da região, mas também informações sobre os incidentes de incêndio. Neste caso, com a ajuda de análises estatísticas e resultados de classificação do comitê de máquinas o corpo de bombeiros de Curitiba pode compreender como os eventos de incêndio se comportam e a reconhecer padrões.

O restante do artigo está organizado da seguinte forma. Na seção 2 são apresentados os métodos e dados. A seguir, na seção 3, os fundamentos das abordagens adotadas de Aprendizado de máquina são detalhados. Os resultados obtidos e a conclusão são discutidos nas seções 4 e 5, respectivamente.

2. Métodos e Dados

O estudo de caso adotado neste artigo está relacionado à cidade de Curitiba, capital do estado do Paraná. Conforme ilustrado na Figura 1, Curitiba tem uma população de aproximadamente 1.893.000 habitantes segundo o Instituto Brasileiro de Geografia e Estatística (IBGE) [IBGE, 2016] com área territorial de 435036 km² [IBGE, 2015] e densidade populacional de 4027,04 hab/km² [IBGE, 2010]. Curitiba foi considerara em 2012 [Exame, 2012] a capital mais desenvolvida do Brasil.

Figura 1 - Curitiba-Paraná-Brasil.
Fonte: adaptado de https://pt.wikipedia.org/wiki/Curitiba

Os dados neste estudo foram obtidos usando o Sistema de Armazenagem de Ocorrências e Estatística do Corpo de Bombeiros do Estado do Paraná (SYSBM) junto com o Sistema Meteorológico do Paraná (SIMEPAR) referentes aos anos de 2013 e 2014.

Como objetivo pretende-se coletar e analisar a viabilidade mensal e dados de temperatura do ar, umidade relativa, precipitação pluviométrica, pressão atmosférica, direção do vento, referentes ao período de 2013 e 2014 relacionando a sucessão dos tipos de clima com as ocorrências de incêndios residenciais na área de Curitiba, assim como localizar e mapear os incêndios.

Os fatores ligados ao clima também fazem relação ao índice de incêndios urbanos nas cidades. A Figura 2 mostra a quantidade de ocorrências registradas na cidade de Curitiba nos anos de 2013 e 2014.

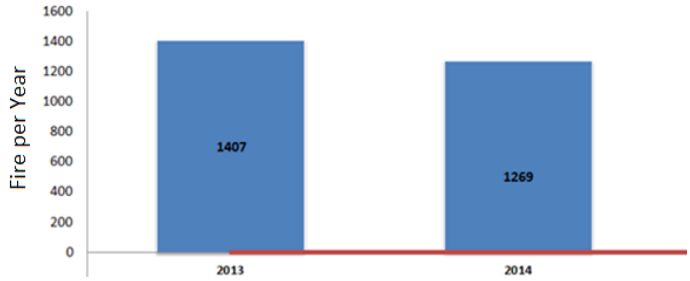

Figura 2 – Total de incêndios em Curitiba.
Fonte: Elaborado pelos autores, 2017.

Os dados apresentados a seguir referem-se à quantidade de ocorrências classificadas por tipo de incêndio, ou seja, pelo tipo: “Incêndio em edificação”, “Incêndio Ambiental” e “Incêndio em meio de transporte”.

No ano de 2013 foram registradas 1407 ocorrências relacionadas a incêndio. Destas, 717 (50,96%) do tipo incêndio em edificação, 508 (36,11%) do tipo incêndio ambiental e 182 (12,94%) do tipo incêndio em meio de transporte.

No ano de 2014, foram 1269 ocorrências relacionadas a incêndio. Destas, 711 (56,03%) do tipo incêndio em edificação, 302 (23,80%) do tipo incêndio ambiental e 256 (20,17%) do tipo incêndio em meio de transporte.

A variação que mais chama atenção dentre os dados estudados, foi o declínio mais acentuado de ocorrências do tipo incêndio ambiental no ano de 2014 em relação aos anos de 2013 (redução de aproximadamente 41%)

3. Fundamentos de Aprendizado de Máquina

A seguir são descritos os métodos de Aprendizado de máquina adotados neste artigo para classificação de ocorrências de incêndios.

3.1 SVM

SVM é uma técnica baseada em aprendizado estatístico desenvolvido por [Vapnik, 1995], com o intuito de resolver problemas de classificação de dados. Desde então vem sido aplicada tanto em problemas de classificação [Bouzalmat *et al*., 2014; Burges, 1998; Gonçalves, 2009; José e Ribeiro, 2012] como em problemas de regressão [Camps-Valls *et al*., 2006; Dutta, *et al*., 2016; Ghaedi *et al*., 2016).

SVM quando aplicado em problemas de classificação que apresentam características linearmente separáveis, utiliza um hiperplano para realizar a separação de classes, o hiperplano é posicionado de forma que a distância entre o mesmo e as classes seja a maior possível [Gonçalves, 2009].

No caso de uma separação de características não-linearmente separáveis, pela sua complexidade não é trivial a aplicação de hiperplano linear. Logo, de acordo com o teorema de Cover [Cover, 1965], em que um problema não-linear que tenha sua dimensionalidade aumentada tem maior chance de se tornar linearmente separável pela utilização de funções Kernel para mapear as características (ver ilustração na Figura 3) [Gonçalves, 2009] o que torna este um algoritmo eficiente. As funções Kernel básicas utilizadas em SVM são lineares, polinomial, sigmoide e de base radial [Bouzalmat *et al*., 2014].

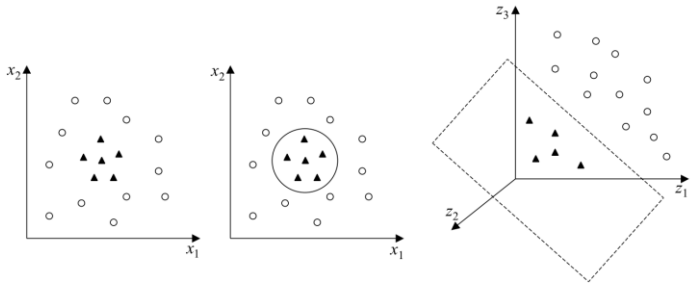

Figura 3 - Teorema de Cover.
Fonte: Adaptado de [Lorena e Carvalho, 2007].

Os parâmetros do algoritmo SVM utilizados nos testes são:

- *kernel:* função kernel a ser utilizada no processo treinamento que pode ser linear, polinomial, base radial ou sigmoide.
- degree: grau da função polinomial.
- gamma: define o quão longe a influência de um único exemplo de treinamento alcança.
- coef0: parâmetro da projeção kernel utilizado em funções polinomial e sigmoide.

3.2 Random Forests

O treinamento de árvores de decisão para ambos problemas de classificação e regressão, como referido no termo CART (*Classification And Regression Trees*) introduzido por Breiman [2001], começa com o nó raiz, e divide binariamente os nós em ramos até atingir as folhas, onde os nós representam o teste sobre uma característica, o ramo representa o valor do resultado do teste e a folha representa a classe.

RF funciona combinando o uso de *bagging* e de seleção aleatória de características desenvolvida por [Ho, 1995], um conjunto de árvores de classificação, em que cada árvore é construída usando uma amostra com reposicionamento selecionada dos dados de treinamento [Fawagreh *et al*., 2014], conforme mostrado na Figura 4, aproximadamente 63% das observações da amostra ocorrem pelo menos uma vez e as observações restantes são denominadas *out-of-bag* [Cutler *et al*., 2007].

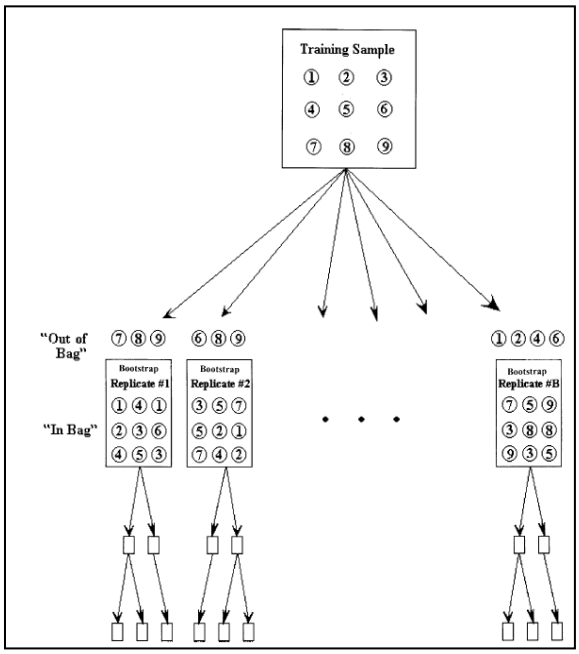

Figura 4 - Reamostragem *bagging*.
Fonte: Adaptado de [Izmirlian, 2004].

No caso da RF é determinada uma amostra para cada árvore gerando um conjunto de classificadores, que por meio de voto majoritário cada classificador vota para a sua classe prevista, e a classe mais votada é usada para classificar a observação [Fawagreh *et al.*, 2014], isso se repete até todas as observações terem sido classificadas. A combinação de vários preditores *bagged* pela técnica RF resulta em um classificador mais preciso com habilidade de modelar iterações complexas entre variáveis preditoras [Cutler *et al*., 2007].

Os parâmetros do algoritmo RF utilizados nos testes são:

- mtry: número de preditores selecionados aleatoriamente.
- predictors: número máximo de preditores.

3.3 K-Nearest Neighbors

KNN é uma técnica de classificação não paramétrica que pode atingir alta precisão de classificação em problemas que possuem distribuição não normal e desconhecida, [Hwang, 1998].

Seu processo de aprendizagem consiste em armazenar todas as instâncias de treinamento com suas respectivas classes, para classificar uma instância desconhecida o classificador ordena as instâncias vizinhas dentre as instâncias de treinamento e usa a classe dos *k* vizinhos mais similares para predizer a classe da nova instância [Tan, 2006].

A distância entre a nova instância e os *k* vizinhos mais próximos é medida e o vizinho mais próximo indica a classe da nova instância, conforme mostrado na Figura 5, onde *k* é um parâmetro definido no começo do algoritmo que representa o número de vizinhos que serão comparadas com a nova instância. O valor de *k* é recomendado ser um número ímpar pelo fato de um número par poder causar empate na decisão da classe a ser escolhida.

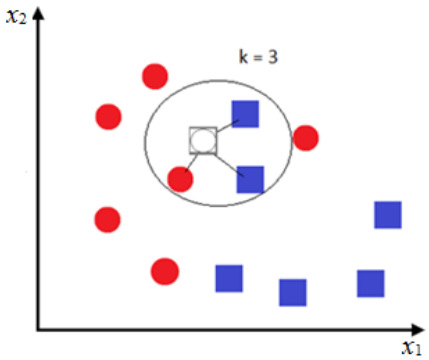

Figura 5 – Algoritmo KNN.
Fonte: Elaborado pelos autores, 2017.

Os parâmetros do algoritmo KNN utilizados nos testes são:

- kmax: número máximo de vizinhos a serem comparados.
- distance: parâmetro da distância máxima entre vizinhos.
- kernel: função kernel utilizada pelo algoritmo pode ser retangular, triangular, epanechnikov, cosseno, inversa, gaussiana, rank e ótima.

3.4 Extreme Gradient Boosting

Gradient Boosting Machine (GBM) produz um competitivo, altamente robusto e interpretável procedimento para ambos classificação e regressão [Friedman, 2001].

GBM é um conjunto de árvores de regressão que utilizam *boosting*, árvores de regressão diferem de árvores de decisão pelo fato de árvores de regressão conterem valores contínuos para cada folha [Chen e Guestrin, 2016]. GBM usa o método *gradient-descent* para construir uma árvore que diminui o objetivo em direção do gradiente.

XGBoost é uma nova implementação do algoritmo GBM, como a natureza do aprendizado aditivo do GBM tende a apresentar um alto risco de sobre treinamento (*overfitting*), XGBoost visa prevenir *overfitting* sem comprometer a eficiência computacional do algoritmo [Chen e Guestrin, 2016].

Os parâmetros do algoritmo XGBoost utilizados nos testes são:

- max.depth: profundidade máxima de uma árvore.
- gamma: especifica a redução mínima de perda requerida pra realizar a divisão de um nó.
- colsample_bytree: fração de colunas a ser aleatoriamente reamostrada para cada árvore.
- lambda: parâmetro que lida com a parte de regularização do algoritmo.
- alpha: termo de regularização, utilizado em caso de alta dimensionalidade para tornar o algoritmo mais rápido.
- min_child_weight: define a soma mínima dos pesos das observações requeridas para cada cria.

3.5 Comitê de máquinas (*ensemble*)

Os conjuntos de técnicas de aprendizado de máquina, consistem em um grupo de técnicas que buscam agregar o conhecimento adquirido dos modelos que os compõem, com o objetivo de atingir uma solução global que resulte em um modelo mais eficiente do que os seus componentes aplicados isoladamente [Wilfredo e Villanueva, 2006].

A aplicação de conjuntos busca melhorar a capacidade de generalização usando os pontos fortes de seus componentes na solução do mesmo problema, o que torna preferível o uso de modelos mais diversificados que possuem características mais distintas.

As técnicas de ensemble compõem-se de no mínimo três componentes, e geralmente se adota uma metodologia de três passos, treinamento (geração), seleção e combinação [Lima, 2004], sendo isto ilustrado na Figura 6.

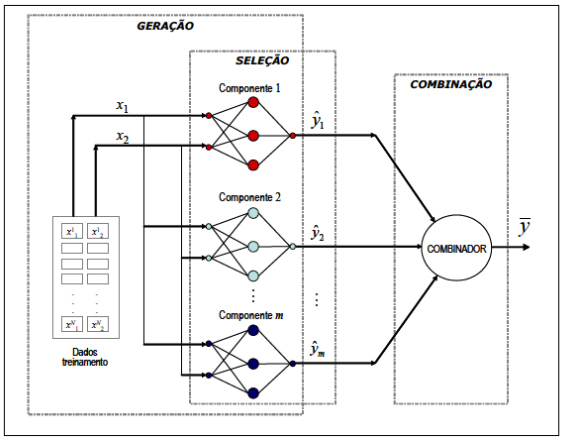

Figura 6 - Etapas da metodologia da criação de um ensemble.
Fonte: Adaptado de [Wilfredo e Villanueva, 2006].

Na etapa de treinamento, os modelos do ensemble são gerados. Na etapa de combinação dos componentes do ensemble o método difere de acordo com o problema no qual vai ser aplicado. Em um problema de classificação pode ser usado uma técnica de votação, na qual a classe que saiu mais vezes é a selecionada. Para um problema de regressão, pode ser aplicada a média ponderada das saídas dos componentes.

Na etapa de seleção, são selecionados os componentes que obtiveram o melhor desempenho do ensemble. Ao utilizar um número maior de componentes é possível que nem todos os componentes contribuam para o desempenho global do ensemble, por tanto é recomendado utilizar técnicas para refinar os componentes seguindo algum critério de seleção, que pode ser uma medida de erro sobre um subconjunto de dados [Wilfredo e Villanueva, 2006].

4. Análise dos Resultados

A seguir, o pré-processamento dos dados e os resultados da aplicação dos algoritmos de Aprendizado de máquina são discutidos.

4.1 Pré-processamento de dados e engenharia de características (*features*)

Referentes aos dados dos incêndios em Curitiba durante os anos de 2013 e 2014, foram adquiridas 1046 observações dos incidentes utilizando as seguintes características:

- *Hora*: Hora dos incidentes;
- *Chefe.Socorro*: Código do chefe de socorro que atendeu a chamada;
- *BM.Autenticador*: Código do corpo de bombeiros onde a chamada foi atendida;
- *Press*: Pressão média do dia do incidente em hPa;
- *Vel_Vento*: velocidade do vento média do dia do incidente em km/h;
- *Umidade*: Umidade média do dia do incidente em %;
- *Preci*: Precipitação média do dia do incidente em mm;
- *Temp_M*: Temperatura média do dia do incidente em ºC.

Com a finalidade de aumentar a complexidade dos dados novas características foram geradas a partir dos dados originais como, *Seasons* baseada a partir do dia dos incidentes, *Rain* que mostra a intensidade da chuva baseada na precipitação média. A partir dos dados originais e das duas novas características criadas pode ser feita uma análise estatística dos eventos de incêndio, conforme mostrado na Figura 7.

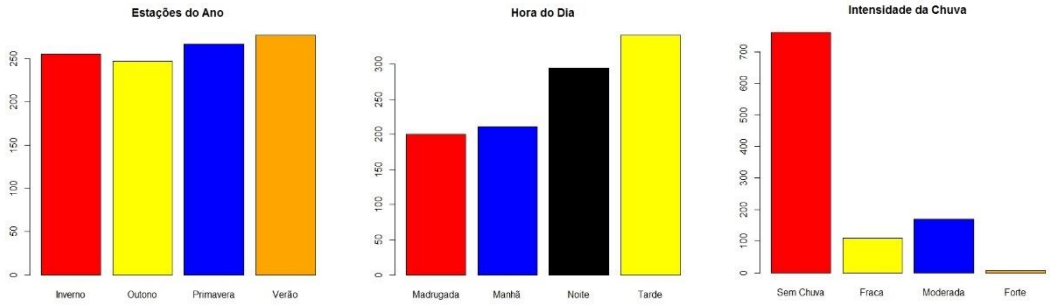

Figura 7 – Características criadas a partir de características originais.
Fonte: Elaborado pelos autores, 2017.

Como pode ser observado pelos histogramas na Figura 10, a maioria dos incêndios ocorreram a noite, durante os dias em que não houve chuva ou no qual os valores de precipitação média apresentam valores menores do que 1.1 mm e durante o verão.

4.2 Treinamento e teste

Os testes foram realizados no sistema operacional Windows 10 64bits, CPU (Central Processing Unit) core i7-4700HQ 2.4 GHz, RAM (Random Acess Memory) de 8 GB e no ambiente computacional RStudio Software.

O estudo de caso é um problema classificação multi-classes com aprendizado supervisionado, onde o classificador a partir das características dos dados realiza a classificação. Neste estudo o foco é qual das 20 sub-regiões de Curitiba melhor caracteriza a observação em questão, com a finalidade de determinar, de acordo com as informações fornecidas, qual seria a região mais provável de ocorrer um incêndio. Como pode ser observado na Figura 8, o mapa da cidade de Curitiba com todos os 1046 eventos de incêndio que ocorreram durante os anos de 2013 e 2014, junto com o mapa de calor dos incidentes.

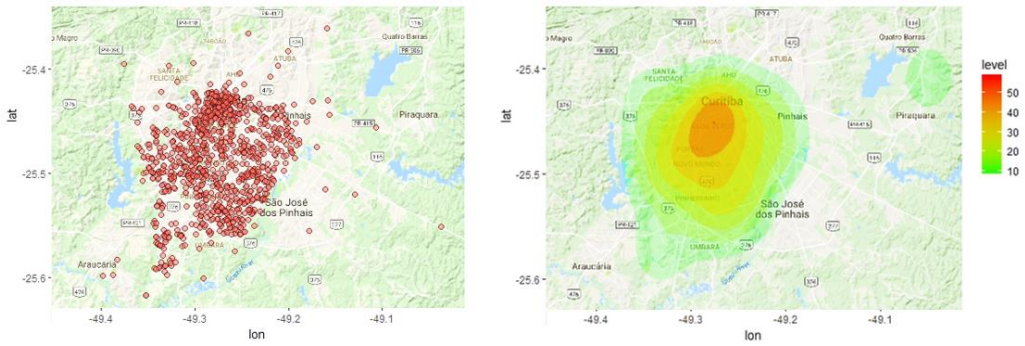

Figura 8 – Mapa de eventos de incêndio e mapa de calor.
Fonte: Elaborado pelos autores, 2017.

Utilizando um algoritmo de agrupamento de dados (*clustering*) (não supervisionado *K-means* proposto originalmente por [MacQueen, 1967] todos os locais dos incidentes foram agrupados em 20 diferentes sub-regiões, como pode ser notado pelo que foi apresentado na Figura 9. Cada uma das 20 sub-regiões serão uma classe diferente para o problema de classificação multi-classes.

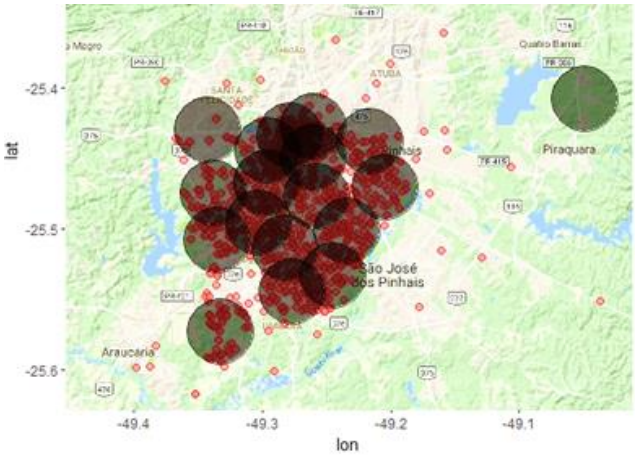

Figura 9 – Divisão de sub-regiões.
Fonte: Elaborado pelos autores, 2017.

Os quatro métodos (XGBoost, SVM, RF e KNN) foram aplicadas isoladamente, e então combinadas em um conjunto de técnicas. A geração de um classificador final combinando os métodos em um comitê de máquinas foi obtida a partir de voto majoritário. Um vetor de pesos multiplicou os classificadores existentes das técnicas isoladas e o valor resultante foi somado em um vetor final das observações classificadas, a média do vetor final é o classificador do conjunto.

4.3 Resultados

Os critérios usados para medir o desempenho de todos os modelos dos classificadores foram a precisão total de classificação (1) e o índice *kappa* (2) da matriz de confusão dos valores classificados de cada modelo.

$$PrecisãoTotal = \frac{VP+VN}{VP+VN+FP+FN} \quad (1)$$

onde, *VP,VN, FP e FN* são respectivamente os valores da matriz de confusão de verdadeiro positivo, verdadeiro negativo, falso positivo e falso negativo.

$$kappa = \frac{PrecisãoTotal - PrecisãoEsperada}{1 - PrecisãoEsperada} \quad (2)$$

onde, *PrecisãoEsperada* é (3):

$$PrecisãoEsperada = \frac{(VN+FP)*(VN+FN)+(FN+VP)*(FP+VP)}{n^2} \quad (3)$$

onde, *n* é o número total de observações.

Os melhores parâmetros definidos para cada algoritmo são mostrados na Tabela 1.

Tabela 1 – Parâmetros de controle dos algoritmos.

XGBoost		**SVM**		**KNN**		**RF**	
Parâmetro	**Valor**	**Parâmetro**	**Valor**	**Parâmetro**	**Valor**	**Parâmetro**	**Valor**
max.depth	4	*kernel*	Polinomial	*kmax*	5	mtry	5
gamma	0,1	*degree*	10	*distance*	10	preditors	9
colsample_bytree	0,3	*gamma*	0,75	*kernel*	*optimal*		
Lambda	0,1	*coef0*	0,0001				
Alpha	0,00001						
min_child_weight	2						

Fonte: Elaborado pelos autores, 2017.

A Tabela 2 mostra os resultados dos modelos, onde estimação e validação representam respectivamente a análise para os dados de treinamento e de teste.

Tabela 2 – Resultados dos algoritmos

Método	**Estimação**		**Validação**		**Tempo (s)**
	Precisão	***Kappa***	**Precisão**	***Kappa***	
XGBoost	0,93471337	0,93021919	0,16267942	0,10558170	67,03
SVM	0,96815286	0,96599118	0,08133971	0,02758307	0,37
KNN	0,97452229	0,97278918	0,10047846	0,04046496	6,81
RF	0,97292993	0,97107010	0,17224880	0,11540884	15,25
Comitê de máquinas	0,99681528	0,99659872	0,11004784	0,06060018	

Fonte: Elaborado pelos autores, 2017.

Os resultados mostram, que todos os modelos apresentam alta variância e um viés (*bias*) baixo, consequentemente apresenta uma má generalização devido ao *overfitting*.

O algoritmo XGBoost possui o maior tempo de treinamento, entretanto ainda teve uma precisão de validação melhor que a maioria das técnicas.

O algoritmo KNN apresentou a segunda maior precisão de estimação, mas apesar desta sua precisão de validação foi muito baixa.

O algoritmo SVM provou ser uma técnica muito rápida, possui uma precisão de estimação maior que a do algoritmo XGBoost, entretanto apresentou a pior precisão de validação dentre todas as técnicas.

O algoritmo RF, apesar de apresentar a terceira maior precisão de estimação o apresentou a maior precisão de validação de todas as técnicas testadas.

A melhor escolha de classificador fica sendo o conjunto de técnicas de aprendizado de máquina, não atingiu o melhor valor de classificação, mas obteve o melhor resultado na precisão de estimação e *kappa* quando comparado com as outras técnicas isoladas.

5. Conclusão

Este artigo apresentou uma nova base de dados para problemas de classificação baseado em dados reais adquiridos com o departamento de bombeiros da cidade de Curitiba.

Junto com a nova base de dados, um conjunto de técnicas combinando diferentes técnicas de diferentes abordagens de aprendizado de máquina na forma de comitê de máquinas mostrou ser um classificador eficiente obtendo um melhor balanço entre precisão de estimação e validação que usando os métodos que o compõem de forma isolada.

Os dados compilados e analisados através das técnicas propostas permitem identificar alguns padrões, tais como: (i) a maioria dos incêndios ocorreram durante a noite; (ii) a maioria dos incidentes de fogo ocorreram em dias em que não estava chovendo, ou que o valor de precipitação média era menos de 1,1; e (iii) a maioria dos incidentes fogo aconteceu durante o verão.

Com algumas informações é possível atuar de forma preventiva, proporcionando melhorias no atendimento das ocorrências de incêndios, pois verifica-se que a análise gerou um padrão na maioria dos incêndios ocorridos.

Com o intuito da continuação desta pesquisa, é sugerido que mais dados sejam coletados pelo Corpo de Bombeiros do Estado do Paraná durante as ocorrências de incêndio para poder melhorar e mais precisamente escrever padrões dos incêndios, deixamos aberto a futuras pesquisas para identificar e melhorar a qualidade na prevenção e combate a incêndios urbanos.

Referências

Athitsos, V., e Sclaroff, S. (2005). Boosting nearest neighbor classifiers for multiclass recognition. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops, (June, p. 1-8.

Bouzalmat, A., Kharroubi, J., e Zarghili, A. (2014). Comparative study of PCA, ICA, LDA using SVM classifier. *Journal of Emerging Technologies in Web Intelligence*, v. 6, n. 1, p. 64–68.

Breiman, L. (2001). Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32.

Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, v. 2, n. 2, p. 121–167.

Cabras, S., Castellanos, M. E., e Staffetti, E. (2016). A random forest application to contact-state classification for robot programming by human demonstration. *Applied Stochastic Models in Business and Industry*, v. 32, n. 2, p. 209–227.

Camps-Valls, G., Bruzzone, L., Rojo-Álvarez, J. L., e Melgani, F. (2006). Robust support vector regression for biophysical variable estimation from remotely sensed images. *IEEE Geoscience and Remote Sensing Letters*, v. 3, n. 3, p. 339–343.

Challands, N. (2009). The Relationships Between Fire Service Response Time and Fire Outcomes, *Fire Technology*, v. 46, p. 665–676.

Chen, T., e Guestrin, C. (2016). XGBoost : Reliable Large-scale Tree Boosting System. KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 785-794, San Francisco, California, USA.

Cover, T.M. (1965). Geometrical and Statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, v. 14, p. 326–334.

Cover, T., e Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., e Lawler, J. J. (2007). Random Forests for Classification in Ecology. *Ecology*, v. 88, n. 11, p. 2783–2792.

Dutta, S., Pal, S. K., e Sen, R. (2016). On-machine tool prediction of flank wear from machined surface images using texture analyses and support vector regression. *Precision Engineering*, v. 43, p. 34–42.

Esposito, F., Malerba, D., Semeraro, G., e Kay, J. (1997). A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 19, n. 5, p. 476–491.

Exame (2012). Exame website Http://exame.abril.com.br/brasil/curitiba-e-a-capital-more-desenvolvida-do-pais-veja-lista/.

Fawagreh, K., Gaber, M. M., e Elyan, E. (2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering*, v. 2, n. 1, p. 602–609.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, v. 29, n. 5, p. 1189–1232.

Gazeta do Povo (2015). Newspapaper Gazeta do Povo website. < http://www.gazetadopovo.com.br/vida-e-cidadania/curitiba-ganha-premio-de-melhor-cidade-do-brasil-7tadh2c7xzcejht1n5aku5s5w.

Ghaedi, M., Rahimi, M. reza, Ghaedi, A. M., Tyagi, I., Agarwal, S., e Gupta, V. K. (2016). Application of least squares support vector regression and linear multiple regression for modeling removal of methyl orange onto tin oxide nanoparticles loaded on activated carbon and activated carbon prepared from Pistacia atlantica wood. *Journal of Colloid and Interface Science*, v. 461, p. 425–434.

Gill, M.A., Scott, L.S., Cary, G.J., (2013). The worldwide "wildfire" problem. *Ecological Application*, v. 23, p. 438-454.

Gonçalves, A. R. (2009). Máquina de Vetores Suporte, Universidade de Campinas, p. 1-18.
Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., e Bing, G. (2016). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, v. 23, i. C, p. 220-239.
Ho, T. K. (1995). Random Decision Forests. In Document analysis and recognition, 1995, Proceedings of the third international conference, pp. 278–282. Montreal, Quebec, Canada.
Holloway, E., e Marks, R. (2016). High Dimensional Human Guided Machine Learning, *Association for the Advancement of Artificial Intelligence*, p. 1-3.
Hwang, W. e Wen, K. (1998). Fast kNN classification algorithm based on partial distant search. *Electronics Letters*, v. 34, n. 21, p. 2062–2063.
IBGE, (2016). Brazilian Statistic and Geographic Institute website . <http://www.ibge.gov.br/home/>
Izmirlian, G. (2004). Application of the Random Forest Classification Algorithm to a SELDI-TPF Proteomics Study in the Setting of a Cancer Prevention Trial. In annals of New York Academy of Sciences 1020, p. 154-174, New York.
José, D., e Ribeiro, S. (2012). Support Vector Machines na Previsão do Comportamento de uma ETAR. Tese de Mestrado em Engenharia Informática da Universidade do Minho, p. 1-160, Braga, Portugal.
Keeley, J.E., Bond, W.J., Bradstock, R.A., Pausas, J.G. e Rundel, P.W. (2012). Fire in Mediterranean Ecosystems. Ecology, *Evolution and Management*, p. 515, Cambridge University Press, New York.
Lambert, T. E., Srinivasan, A. K. e Katirai, M. (2012) Ex-Urban Sprawl and Fire Response in the United States, *Journal of Economic Issues*, v. 46, n. 4, p. 967-988.
Lima, C. A. D. M. (2004). Comitê de Máquinas: Uma Abordagem Unificada Empregando Máquinas de Vetores-Suporte. Tese de Mestrado em Engenharia Elétrica da Universidade Estadual de Campinas. p. 1-378, Campinas.
Lorena, A. C., e de Carvalho, A. C. P. L. F. (2007). Uma Introdução às Support Vector Machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67.
MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Symposium on Math, Statistics, and Probability*, p. 281–297, Berkeley, CA: University of California Press.
MCSCS, Ministry of Community Safety e Correctional Services. Retrieved December 02, 2016, from http://www.mcscs.jus.gov.on.ca/english/FireMarshal/MediaRelationsandResources/FireStatistics/OntarioFires/AllFireIncidents/stats_all_fires.html
Mejdoub, M., e Ben Amar, C. (2013). Classification improvement of local feature vectors over the KNN algorithm. *Multimedia Tools and Applications*, v. 64, v. 1, p. 197–218.
NFPA, National Fire Protection Association. Retrieved December 04, 2016, from http://www.nfpa.org/news-and-research/fire-statistics-and-reports/fire-statistics.
Pausas, J.G., Llovet, J., Rodrigo, A. e Vallejo, R. (2008). Are wildfires a disaster in the Mediterranean basin? A review. *Int. J. Wildland Fire*, v. 17, p. 713-723.
Penman, T.D., Collins, L., Price, O.F., Bradstock, R.A., Metcalf, S. e Chong, D.M.O. (2013). Examining the relative effects of fire weather, suppression and fuel treatment on fire behaviour - a simulation study. *Journal Environmental Management*, v. 131, p. 325-333.
Pradhan, A. (2012). Support vector machine - a Survey. *International Journal of Emerging Technology and Advanced Engineering*, v. 2, n. 8, p. 82–85.
Ren, Y., Zhang, L., e Suganthan, P. N. (2016). Ensemble classification and regression: Recent developments, applications and future directions. *IEEE Computational Intelligence Magazine*, v. 11, n. 1, p. 41–53.
Tan, S. (2006). An effective refinement strategy for KNN text classifier. *Expert Systems with Applications*, v. 30, n. 2, p. 290–298.
Vapnik, V.N. (1995). The Nature of Statistical Learning Theory. 2nd ed. New York: Springer-Verlag.
Veblen, T.T., Kitzberger, T., Raffaele, E., Mermoz, M., Gonzalez, M.E., Sibold, J.S. e Holz, A., (2008). The historical range of variability of fires in the Andean-Patagonian Nothofagus forest region. *Int. J. Wildland Fire*, v. 17, p. 724-741.
Wilfredo, J., e Villanueva, P. (2006). Comitê de Máquinas em Predição de Séries. Tese de Mestrado da Universidade Estadual de Campinas. p. 1-168, Campinas.