Evaluating the Impact of Pre-clustering and Class Imbalance on Solar Flare Forecasting

Mirelle C. Bueno[1], **Guilherme P. Coelho**[1], **Ana Estela A. da Silva**[1] **and André L. S. Gradvohl**[1]

[1]School of Technology (FT) – University of Campinas (Unicamp)
Limeira – SP – Brazil

mirelle.bueno21@gmail.com, {guilherme, aeasilva, gradvohl}@ft.unicamp.br

Abstract. *Among the phenomena that occur on the surface of the Sun, solar flares may cause several damages, from short circuits in power transmission lines to complete interruptions in telecommunications systems. In order to mitigate these effects, many works have been dedicated to the proposal of mechanisms capable of predicting the occurrence of solar flares. In this context, the present work sought to evaluate two aspects related to machine learning-based solar flare forecasting: (i) the impact of class imbalance in training datasets on the performance of the predictors; and (ii) whether the incorporation of a pre-clustering step prior to the classifiers training contributes to a better prediction.*

Resumo. *Dentre os fenômenos que ocorrem na superfície do Sol, as explosões solares podem provocar desde curtos circuitos em linhas de transmissão de energia até interrupções completas em sistemas de telecomunicações. Visando mitigar tais efeitos, muitos trabalhos têm se dedicado a desenvolver mecanismos capazes de prever a ocorrência de explosões solares. Neste contexto, o presente trabalho buscou avaliar dois aspectos relacionados à previsão deste fenômeno baseada em aprendizado de máquinas: (i) o impacto que o desbalanceamento das classes dos conjuntos de dados de treinamento tem no desempenho dos preditores; e (ii) se a incorporação de uma etapa de pré-agrupamento dos dados, antes do treinamento dos classificadores, contribui para uma melhor previsão.*

1. Introdução

A atividade solar desempenha um papel fundamental na dinâmica terrestre, o que exige o monitoramento constante dos fenômenos que ocorrem no Sol. Dentre tais fenômenos, as chamadas explosões solares (*solar flares*) ocorrem quando a energia magnética acumulada na atmosfera solar é abruptamente liberada [Holman and Benedict 2007], provocando modificações na Ionosfera terrestre. Tais modificações podem provocar efeitos nocivos, que vão desde curtos circuitos em linhas de transmissão de energia elétrica a panes em satélites e sistemas de telecomunicações [Lima 2012].

Diante do potencial nocivo que variações na atividade solar podem provocar na superfície terrestre, vários parâmetros associados ao Sol são constantemente monitorados por radiotelescópios, telescópios ópticos e satélites, e registrados em bases de dados. Estes sensores medem parâmetros como fluxos de raios-X, ultravioleta e gama provenientes do Sol, que permitem a caracterização dos fenômenos solares como um todo [Lima 2012]. Particularmente no caso das explosões solares, valores de pico na intensidade do fluxo de

raios-X (comprimento de onda entre 1 e 8 Angströms) permitem sua classificação nas categorias A, B, C, M e X (de menor para maior intensidade) [Rowlett 2013].

Neste contexto, existe um grande interesse no desenvolvimento de ferramentas e técnicas que permitam a predição da ocorrência de explosões solares com certa antecedência, o que possibilitaria um melhor gerenciamento dos eventuais danos e até mesmo uma redução do impacto causado por tais fenômenos [Argento 2016]. Sendo assim, vários grupos de pesquisa têm se dedicado à compreensão das explosões solares e ao desenvolvimento de modelos e técnicas que possibilitem sua previsão.

Dentre os principais trabalhos da literatura voltados para estudo e previsão de explosões solares, [Liu et al. 2005] analisaram imagens do Sol, dados de fluxo de raios-X e também da classificação magnética das manchas solares durante a ocorrência de explosões de classe X. Os resultados mostraram que há uma relação entre o fluxo de raios-x e as explosões solares, bem como apontaram mudanças das características visuais das manchas solares na iminência da ocorrência das explosões.

Já [Colak and Qahwaji 2009] se propuseram a prever a possível ocorrência de explosões solares e, caso uma explosão fosse prevista, o sistema proposto deveria indicar se tal explosão seria de classe M ou X. Para isso, introduziram um sistema híbrido, baseado em redes neurais artificiais, que mescla processamento de imagens da coroa solar com análises dos dados históricos das manchas e de explosões solares.

A pesquisa de [Li et al. 2011] apresentou um novo método para a previsão de explosões em até 48 horas à frente. Foram combinados métodos não-supervisionados de agrupamento de dados, sendo, em um primeiro momento, utilizado o algoritmo K-means para balancear o conjunto de dados e, posteriormente, aplicado o método *Learning Vector Quantization* (LVQ).

Máquinas de vetores suporte (*Support Vector Machines* – SVMs) foram utilizadas por [Bobra and Couvidat 2015] para indicar, com base em dados históricos do campo magnético solar, em quais regiões ativas da coroa se apresentavam as explosões solares. O objetivo final do trabalho foi a identificação de possíveis ocorrências de explosões solares das classes M e X. Devido à enorme massa de dados (1,5 TB por dia), foi necessária a utilização de técnicas de seleção de atributos, no caso a métrica *Fisher Ranking Score*.

Por fim, [Argento 2016] recorreu às redes neurais artificiais do tipo MLP (do inglês *Multilayer Perceptrons*) e *ensembles* destas redes neurais buscando prever tanto ocorrências futuras de explosões solares quanto a classe a qual essas explosões pertencerão (M ou X). Foram utilizados no trabalho de [Argento 2016] dados de explosões como fluxo de rádio, número das manchas solares, área dessas manchas, fluxo de raio-x integrado, classificação magnética de manchas solares observadas no Sol e a ocorrência ou não de explosões solares nos cinco dias anteriores a cada registro. Ao final, [Argento 2016] obteve um erro médio de classificação de 16% e concluiu que o *ensemble* de redes neurais leva a erros menores que os obtidos pelas redes neurais artificiais utilizadas individualmente.

Analisando a literatura é possível constatar que a tarefa de predição de explosões solares, principalmente as de mais alta intensidade, M e X, não é trivial. Há um grande número de atributos associados ao Sol. Além disso, a massa de dados disponível para análise é considerável e explosões das classes M e X são raras, o que torna os dados dis-

poníveis altamente desbalanceados (número de amostras associadas à não ocorrência do fenômeno é muito maior). Sendo assim, o presente trabalho buscou contribuir para os estudos de previsão de explosões solares em duas frentes: (*i*) avaliando o impacto que a execução de uma estratégia de pré-agrupamento dos dados, antes do treinamento dos preditores, tem na qualidade da previsão; e (*ii*) verificando o impacto que o desbalanceamento dos dados tem na qualidade da previsão. O problema de previsão de explosões solares foi tratado neste trabalho como um problema de classificação de dados, onde a classe positiva corresponde à ocorrência, um dia à frente, de explosões solares de alta intensidade e a classe negativa indica a não ocorrência destas explosões.

A ideia do pré-agrupamento dos dados é identificar grupos com amostras similares, dentro do conjunto de treinamento dos preditores, e, para cada grupo treinar um preditor específico. Com isso, ao final da etapa de treinamento haverá um grupo de preditores disponíveis, sendo cada um voltado para representar características específicas do problema (retratadas em cada grupo de dados). Para esta etapa de pré-agrupamento utilizou-se aqui a abordagem de agrupamento hierárquico, mais especificamente o algoritmo conhecido como *Agglomerative Nesting* (AGNES) [Kaufman and Rousseeuw 1990].

Por fim, para avaliar o impacto que o desbalanceamento dos dados tem na qualidade dos preditores, foram considerados tanto preditores treinados com todas as amostras de treinamento (dados desbalanceados) quanto com um novo conjunto de treinamento cujas classes (positiva e negativa) foram balanceadas via subamostragem aleatória [Rahman and Davis 2013].

Este artigo está estruturado da seguinte forma: na Seção 2 são discutidos os fundamentos teóricos associados a este trabalho, enquanto que na Seção 3 é apresentada a metodologia experimental adotada. Os resultados são apresentados e discutidos na Seção 4. Por fim, a Seção 5 traz as conclusões do trabalho.

2. Fundamentação Teórica

Nesta seção serão discutidos a estratégia de agrupamento de dados utilizada na etapa de pré-agrupamento dos dados e a abordagem de balanceamento dos dados de treinamento.

2.1. Agrupamento de Dados: o Algoritmo AGNES

Existem técnicas de agrupamento de dados (ou *clustering*) na literatura que permitem o particionamento automático dos dados de um problema, conforme suas características intrínsecas. Tais métodos de agrupamento podem ser divididos em duas categorias [Vale 2005, Han and Kamber 2006]:

- **Métodos Particionais:** compreendem os algoritmos chamados exclusivos e os não-exclusivos, que buscam obter simultaneamente todos os grupos (*clusters*), resultando em uma única partição dos dados. Um dos exemplos de algoritmo particional mais conhecido é o K-means [Zhao et al. 2005], que busca minimizar as distâncias entre amostras dos dados e um conjunto de K centros.
- **Métodos Hierárquicos:** englobam os algoritmos aglomerativos e divisivos, que buscam gerar múltiplas partições para os dados do problema. Tais partições são exibidas para o usuário de uma forma hierárquica. Estes métodos têm como principal vantagem a possibilidade de visualização das relações entre os dados em

diferentes níveis de granularidade [Zhao et al. 2005]. Isso permite a realização de análises interativas para identificação, por exemplo, do número de grupos que melhor representa a base de dados em questão.

Diferentemente dos algoritmos particionais, que retornam um particionamento dos dados que divide as amostras em K grupos, os algoritmos hierárquicos de agrupamento não constroem uma única partição, mas exploram todos os valores de K em uma mesma execução. Logo, $K = 1$ irá conter todos os elementos dos dados, enquanto que $K = n$ irá construir n *clusters*, cada um contendo um único elemento dos dados. Entre estas extremidades ($K = 2, 3, \cdots, n - 1$), os algoritmos apresentam diferentes números de *clusters* em um tipo de "transição gradual" [Kaufman and Rousseeuw 1990], que correspondem às consecutivas aglomerações (ou divisões) dos *clusters* previamente identificados.

Neste sentido, os métodos hierárquicos são subdivididos em abordagens *aglomerativas* e *divisivas* [Doni 2004], que constroem tal hierarquia de grupos em direções opostas. Os algoritmos aglomerativos iniciam o processo de agrupamento com *clusters* unitários (uma amostra dos dados em cada grupo), que gradualmente são fundidos em grupos maiores até que se obtenha um único *cluster* formado por todas as amostras [Jain and Dubes 1988]. Já os métodos divisivos seguem a lógica inversa.

Um dos algoritmos de agrupamento aglomerativo mais conhecidos na literatura, e adotado neste trabalho, é chamado de AGNES. Este algoritmo recebe como entrada um conjunto com N amostras de dados (objetos) e avalia o grau de dissimilaridade entre cada par de objetos por meio de uma medida de dissimilaridade. A partir destes valores, é aplicado então algum critério de junção e o processo é repetido para os grupos gerados até que, ao final, tenha-se um grande grupo com todos os objetos do problema [Kaufman and Rousseeuw 1990].

Para implementação do algoritmo AGNES, três aspectos devem ser definidos: a métrica de avaliação de similaridade/dissimilaridade entre amostras dos dados, o critério de junção de grupos e um mecanismo para identificação do número ideal de grupos dentre todos os particionamentos obtidos pelo algoritmo. Estes aspectos serão discutidos nas próximas seções.

2.1.1. Similaridade de Gower

O algoritmo AGNES adotado aqui recebe como entrada uma matriz de dissimilaridade (ou de distância), onde cada entrada D_{ij} desta matriz corresponde a um valor que indica o quão diferentes (distantes) são as amostras i e j do conjunto de dados [Linden 2009].

Existem diversas métricas de distância propostas na literatura, que devem ser escolhidas conforme as características dos dados. Neste trabalho, o conjunto de dados relacionado às explosões solares possui tanto atributos categóricos quanto contínuos, o que exige a utilização de uma métrica de dissimilaridade capaz de lidar com estas diferenças. Sendo assim, foi utilizado aqui o Critério de Similaridade de Gower [Moura et al. 2010], dado pela Eq. 1.

$$S_{ij} = \frac{\sum_{k=1}^{p} W_{ijk} \times s_{ijk}}{\sum_{k=1}^{p} W_{ijk}}, \tag{1}$$

onde p é o número de atributos dos dados, i e j são as amostras analisadas, W_{ijk} é o peso dado à comparação do atributo k dos elementos i e j (para valores válidos de ijk, $W_{ijk} = 1$, caso contrário $W_{ijk} = 0$) e s_{ijk} é a contribuição do atributo k para o valor de similaridade. Para atributos categóricos nominais, s_{ijk} é definido como 1 caso o valor do atributo k seja o mesmo para as amostras i e j, e 0 caso contrário. Para um dado atributo contínuo x, s_{ijk} é dado pela Eq. 2.

$$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}, \tag{2}$$

onde R_k é a diferença entre os valores máximo e mínimo do atributo k e x_{ik} é o valor do atributo k para a $i-$ésima amostra dos dados.

Como o algoritmo AGNES deve receber como entrada uma matriz de dissimilaridade, neste trabalho foi adotada a métrica de dissimilaridade dada por $D_{ij} = 1 - S_{ij}$, onde S_{ij} é o critério de similaridade de Gower dado pela Eq. 1.

2.1.2. Critério de Junção de Grupos: *Complete Linkage*

Um fator que diferencia os algoritmos hierárquicos de agrupamento é o critério ou técnica utilizado para determinar a distância entre grupos e atualizar a matriz de distância sempre que um novo *cluster* for formado. Existem diversas abordagens para isto, sendo que, neste trabalho, foi implementada a estratégia conhecida como *Complete-Linkage* [Jain and Dubes 1988], que define que a distância entre dois grupos G_1 e G_2 é dada pela maior distância entre suas amostras, conforme a Eq. 3.

$$D(G_1, G_2) = \max(D_{ij}), \forall i \in G_1, j \in G_2, \tag{3}$$

onde D_{ij} é a distância entre os elementos i e j dos dados, pertencentes aos grupos G_1 e G_2, respectivamente. O *Complete-Linkage* foi adotado neste trabalho por ser menos sensível à presença de ruído nos dados do que outras estratégias tais como o *Single-Linkage* [Camargos 2016].

2.1.3. Identificação do Número de Grupos: Critério de Mojena

Uma dificuldade ao se utilizar abordagens hierárquicas de agrupamento é que, geralmente, tais abordagens retornam toda a hierarquia de grupos formada em estruturas gráficas conhecidas como dendrogramas. Um dendrograma é uma estrutura de árvore em que cada folha corresponde a um grupo com uma única amostra dos dados. Já os nós internos indicam a junção entre os grupos identificados pelas amostras nas folhas de seus ramos, e o tamanho das arestas corresponde à distância entre estes grupos.

De posse deste dendrograma, cabe ao usuário definir o "ponto de corte" nas arestas, ou seja, o número ideal de grupos para o problema em questão. Tal estatégia para definição do número de grupos é inviável em grandes bases de dados como a estudada neste trabalho, uma vez que a visualização do dendrograma é prejudicada.

A fim de auxiliar na identificação do número de grupos (ponto de corte no dendrograma), algumas técnicas foram propostas na literatura. Uma das mais tradicionais é a estratégia proposta por Mojena [Mojena 1977], adotada neste trabalho.

O Método de Mojena é um método estatístico em que o ponto de corte no dendrograma é dado pela Eq. 4. Esta estratégia tem como objetivo encontrar o "maior salto", ou seja, a maior amplitude dentre as junções do dendrograma [Mojena 1977].

$$h_j > \bar{h} + \phi \cdot \sigma_h, \tag{4}$$

onde h_j é o valor da distância entre dois grupos que foram unidos (distância de junção), $\bar{h}$ e σ_h correspondem, respectivamente, à média e ao desvio padrão de h_j e ϕ é uma constante que deve ser definida pelo usuário. A constante ϕ indica, indiretamente, qual o "salto" que deve ocorrer em uma das arestas do dendrograma para que seja identificado o ponto de corte. Sendo assim, valores maiores de ϕ levam a números menores de grupos, enquanto que valores menores de ϕ resultam em mais grupos.

2.2. Estratégias de Balanceamento de Dados

Um problema enfrentado nos conjuntos de dados que sinalizam a ocorrência de explosões solares de alta intensidade é o desbalanceamento entre amostras das classes positiva e negativa, o que ocorre pois as explosões solares de classes M e X são eventos raros. Neste contexto, é frequente encontrar conjuntos de dados em que 95% das amostras sinalizam a não-ocorrência de explosões solares [Al-Ghraibah et al. 2015].

O efeito do desbalanceamento pode trazer resultados indesejados na classificação de dados, pelo fato de induzirem os algoritmos de aprendizado de máquinas a considerarem prioritariamente as classes majoritárias [Yen and Lee 2006]. Para mitigar este problema, vêm sendo propostas alternativas para o balanceamento de dados, que podem envolver a sobreamostragem (adição artificial de amostras de classes minoritárias) ou a subamostragem (eliminação de amostras da classe majoritária) dos dados [Rahman and Davis 2013].

Neste trabalho foi adotada uma estratégia de subamostragem de dados. Para isto, foi selecionado, aleatoriamente, um subconjunto das amostras da classe majoritária que, ao final, foi combinado às amostras da classe minoritária. Apesar de ser uma estratégia simples que pode levar à eliminação de amostras importantes da classe majoritária, ela tem apresentado bons resultados na literatura [Rahman and Davis 2013].

3. Metodologia Experimental

Os dados de explosões solares utilizados neste trabalho (vide Seção 3.1) correspondem a séries temporais e, para garantir uma maior robustez dos resultados experimentais, foi adotada a metodologia de *Validação Cruzada por Blocos* [Kohavi 1995], descrita na Seção 3.2. Nesta abordagem os dados são divididos em blocos e tais blocos são utilizados para formar múltiplas configurações de subconjuntos de treinamento e teste, utilizados nas repetições dos experimentos.

Para cada configuração de treinamento/teste foi adotada a seguinte metodologia para os experimentos que envolvem o pré-agrupamento:

1. Realização do agrupamento dos dados de treinamento, via algoritmo AGNES associado à métrica de similaridade de Gower, *Complete-Linkage* e definição do número de grupos via critério de Mojena;
2. Treinamento de n classificadores (vide Seção 3.3) com os dados de cada um dos n grupos formados (um classificador por *cluster*);
3. Classificação das amostras do conjunto de testes usando os classificadores previamente treinados. Nesta etapa é calculada a distância de Gower entre os atributos de entrada da amostra s a ser classificada e as amostras de cada *cluster* gerado. Feito isto, identifica-se o *cluster* C_i, que possui a menor distância média entre suas amostras e a amostra s, e utiliza-se o classificador treinado com as amostras de C_i para classificar s.
4. Avaliação dos resultados de classificação do conjunto de testes pelas métricas apresentadas na Seção 3.4.

3.1. Dados de Explosões Solares

O conjunto de dados de explosões solares estudado neste trabalho é composto por 7.320 amostras, coletadas em um período de 20 anos (1997-2017) [Cinto et al. 2018]. Estes dados foram extraídos dos repositórios mantidos pelo *Space Weather Prediction Center*[1] (SWPC), que é um dos nove centros de previsões climáticas dos Estados Unidos. O SWPC promove o monitoramento, em tempo real, dos eventos solares, para posteriormente disponibilizar os dados coletados gratuitamente para fins de pesquisa e estudo.

Cada amostra do arquivo de dados de explosões solares contém 13 atributos:

- **Fluxo de Rádio:** fluxo total de rádio (comprimento de onda igual a 10,7 cm) emitido pelo Sol;
- **Número de manchas solares:** refere-se ao número de manchas solares observadas em um determinado dia;
- **Área das manchas solares:** soma das áreas de todas as manchas solares observadas em um determinado dia;
- **Fluxo de raio-x integrado:** calculado pela combinação do fluxo de raio-x proveniente da Via Láctea e de fora dela. Este parâmetro é obtido pelo satélite GOES;
- **Explosões de classe α, β, γ, $\beta-\gamma$, δ, $\beta-\delta$, $\beta-\gamma-\delta$ e $\gamma-\delta$:** ocorrência ou não destas classes de explosão no dia (um atributo binário para cada classe);
- **Explosão Solar:** ocorrência ou não de explosão solar um dia à frente.

Os quatro primeiros atributos (fluxo de rádio, número de manchas solares, área das manchas e fluxo de raio-x integrado) são do tipo contínuo, ou seja, numéricos, enquanto que os atributos restantes são do tipo categórico nominal. O último atributo ("Explosão Solar") corresponde à saída esperada (classe) a ser gerada pelo preditor.

Com relação ao desbalanceamento dos dados, das 7.320 amostras apenas 1.279 são da classe positiva, o que equivale a $17,4\%$ do total.

3.2. Validação Cruzada por Blocos

Como os dados de explosões solares apresentam uma ordem cronológica, foi adotado aqui o procedimento de Validação Cruzada por Blocos [Kohavi 1995]. Sendo assim, o

[1]Mais informações: `https://www.swpc.noaa.gov`

conjunto de dados utilizado neste trabalho foi dividido cronologicamente em 10 partições com 732 amostras cada.

Ainda mantendo-se a ordem cronológica das amostras, estas partições foram combinadas de forma a se gerar diferentes subconjuntos de treinamento e teste, permitindo a avaliação da metodologia proposta em diferentes subconjuntos de dados. Para isso, foram definidos cinco pares de subconjuntos de treinamento/teste, sendo os subconjuntos de treinamento formados por 4 partições dos dados (2.928 amostras) e os de teste por duas partições dos dados (1.464 amostras). Este processo de particionamento dos dados, ilustrado na Figura 1, permitiu a execução de 5 repetições dos experimentos.

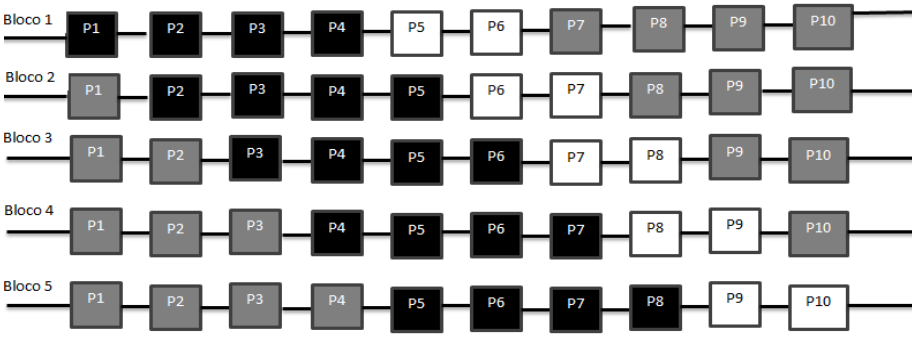

Figura 1. Divisão do conjunto de dados em dez partições (P_i) e criação dos cinco blocos de treinamento e teste. Em cada bloco, os retângulos pretos indicam as partições usadas para o subconjunto de treinamento e os retângulos brancos as partições usadas para o subconjunto de teste.

3.3. Classificadores Utilizados para Previsão de Explosões Solares

Como o foco deste trabalho é avaliar os impactos trazidos pelo pré-agrupamento e balanceamento dos dados, nos experimentos realizados aqui foram utilizados três classificadores bem estabelecidos na literatura: *k–Nearest Neighbors* (k-NN), Naïve Bayes e árvores de decisão (J48) [Han and Kamber 2006].

O algoritmo k-NN armazena os dados de treinamento e, sempre que for necessário classificar uma nova amostra, é calculada a distância entre os atributos de entrada da nova amostra e todas as amostras dos dados de treinamento, escolhidas as k amostras de treinamento mais próximas e definida a classe da nova amostra a partir das classes associadas a estas k amostras (por votação). O único parâmetro a ser definido é o valor de k. Neste trabalho, observou-se que $k = 3$ levou aos melhores resultados (tendo sido adotado nos experimentos que serão discutidos a seguir).

Já o Naïve Bayes é muito utilizado para classificação de dados, mesmo se baseando na suposição ingênua de que a contribuição de um determinado atributo para a classificação não é influenciada pelos demais atributos (atributos independentes). Este classificador retorna as probabilidades de que uma amostra dos dados pertença a cada classe do problema [Han and Kamber 2006].

Já as árvores de decisão, como o algoritmo J48, geram uma estrutura em árvore que possui, em cada folha, a classe que deve ser atribuída a uma determinada amostra, enquanto que os nós internos correspondem aos testes a serem feitos em cada atributo.

3.4. Avaliação dos Resultados

Inicialmente os resultados foram avaliados pela acurácia média, ou seja, pela média da porcentagem de amostras corretamente classificadas ao final das cinco repetições dos experimentos. No entanto, como o conjunto de dados estudado apresenta desbalanceamento de classes, foi utilizada aqui também a métrica conhecida como *F–measure* (Eq. 5), que é menos sensível ao desbalanceamento dos dados [Han and Kamber 2006].

$$F_{measure} = 2 \times \frac{P \cdot R}{P + R}, \tag{5}$$

onde P é a precisão, dada por $\frac{TP}{TP+FP}$, R é a sensibilidade (ou *recall*), dada por $\frac{TP}{TP+FN}$, TP é o número de amostras da classe positiva corretamente classificadas, FP o número de amostras da classe negativa classificadas como classe positiva e FN o número de amostras da classe positiva classificadas como classe negativa.

4. Resultados Experimentais

Neste trabalho, os experimentos que envolvem pré-agrupamento dos dados foram executados tanto com a base completa de dados de explosões solares (dados desbalanceados) quanto com a base de dados balanceada via subamostragem das amostras da classe negativa. Além disso, foram comparados também os resultados obtidos com os classificadores treinados sem o pré-agrupamento dos dados, tanto para os dados balanceados quanto para os dados originais.

Os atributos numéricos dos dados foram normalizados para que apresentassem média $0{,}0$ e desvio-padrão $1{,}0$, através da estratégia *z-score*. Já o parâmetro ϕ de Mojena foi definido empiricamente como $\phi = 18{,}0$ para os experimentos sem subamostragem e $\phi = 10{,}0$ para os experimentos com os dados subamostrados. Estes valores foram definidos de forma a gerar, em média, 4 *clusters*.

Os valores médios e de desvio-padrão de *F–measure* e acurácia, obtidos para cada classificador (k-NN, Naïve Bayes e J48) após as cinco repetições dos experimentos, são dados nas tabelas 1 e 2 para os experimentos com os dados originais (sem balanceamento de classes) e com as classes balanceadas, respectivamente.

Tabela 1. Acurácia e F–measure (média $\pm$ desvio padrão) dos experimentos realizados com os dados originais (sem balanceamento de classes). Os melhores resultados estão em negrito.

Classificador	Pré-agrupamento	F–measure	Acurácia
k-NN	Não	$\mathbf{0{,}53 \pm 0{,}12}$	$\mathbf{0{,}93 \pm 0{,}05}$
	Sim	$0{,}30 \pm 0{,}12$	$0{,}86 \pm 0{,}05$
Naïve Bayes	Não	$0{,}28 \pm 0{,}05$	$0{,}82 \pm 0{,}01$
	Sim	$0{,}36 \pm 0{,}05$	$0{,}84 \pm 0{,}01$
J48	Não	$0{,}35 \pm 0{,}01$	$0{,}87 \pm 0{,}01$
	Sim	$0{,}36 \pm 0{,}01$	$0{,}88 \pm 0{,}01$

Como mencionado na Seção 2.2, para balanceamento dos dados foi utilizada a estratégia de subamostragem aleatória, o que reduziu o conjunto de dados inicial para

Tabela 2. Acurácia e F–measure (média $\pm$ desvio padrão) dos experimentos realizados com os dados balanceados. Os melhores resultados estão em negrito.

Classificador	Pré-agrupamento	F–measure	Acurácia
k-NN	Não	$\mathbf{0{,}53 \pm 0{,}11}$	$\mathbf{0{,}93 \pm 0{,}13}$
	Sim	$0{,}35 \pm 0{,}11$	$0{,}72 \pm 0{,}12$
Naïve Bayes	Não	$0{,}29 \pm 0{,}05$	$0{,}78 \pm 0{,}02$
	Sim	$0{,}38 \pm 0{,}05$	$0{,}82 \pm 0{,}02$
J48	Não	$0{,}40 \pm 0{,}01$	$0{,}69 \pm 0{,}01$
	Sim	$0{,}40 \pm 0{,}01$	$0{,}70 \pm 0{,}01$

um total de 2.558 amostras, sendo 50% destas amostras da classe positiva (ocorrência de explosão solar de alta intensidade um dia à frente) e 50% da classe negativa.

Observando-se os resultados das tabelas 1 e 2, pode-se notar que os melhores resultados tanto para *F–measure* quanto para a acurácia foram obtidos pelo k-NN sem a etapa de pré-agrupamento dos dados. Ainda para o k-NN, pode-se observar que o balanceamento dos dados não influenciou os resultados, o que indica que este tipo de classificador apresenta uma certa robustez frente ao desbalanceamento de classes.

Já as árvores de decisão, obtidas com o algoritmo J48, foram responsáveis pelo segundo melhor desempenho em relação à *F–measure*. No entanto, seu desempenho em acurácia foi o pior quando comparado ao k-NN e ao Naïve Bayes. Para o J48, pode-se dizer que os ganhos apresentados quando a etapa de pré-agrupamento é realizada são insignificantes, tanto para *F–measure* quanto para acurácia. Quanto ao desbalanceamento de dados, a realização da subamostragem levou a pequenos ganhos de *F–measure*, mas a uma piora de cerca de 20% em acurácia. No entanto, é importante destacar que valores de acurácia avaliados em dados desbalanceados geralmente não refletem o desempenho real do classificador para amostras de todas as diferentes classes do problema.

Os classificadores baseados em Naïve Bayes foram os que apresentaram a maior variação diante da realização da etapa de pré-agrupamento de dados: foi observado um ganho de aproximadamente 28% em *F–measure* com a realização do pré-agrupamento, e de cerca de 2% em acurácia. No entanto, mesmo com este ganho, o desempenho do Naïve Bayes ainda é muito inferior ao apresentado pelo k-NN sem pré-agrupamento. Por fim, como observado pra o J48, o balanceamento dos dados levou a ganhos marginais em *F–measure* e a uma piora em acurácia.

5. Conclusões

Neste trabalho buscou-se avaliar dois aspectos relacionados à previsão de explosões solares de alta intensidade via algoritmos de aprendizado de máquinas: o impacto que o desbalanceamento dos dados do problema tem no desempenho de três algoritmos de classificação bem estabelecidos na literatura (k-NN, Naïve Bayes e Árvores de Decisão) e se a inclusão de uma etapa de pré-agrupamento de dados pode melhorar a qualidade final da previsão. O problema de previsão de explosões solares foi tratado como um problema de classificação de dados, sendo a classe positiva a ocorrência de explosões um dia à frente e a classe negativa a não ocorrência de explosões.

Observou-se que o balanceamento dos dados de treinamento tem impacto apenas na métrica de acurácia. Isso é esperado, uma vez que, para dados desbalanceados, esta

métrica não reflete adequadamente o desempenho de um classificador para todas as classes do problema. Já para a *F–measure*, o balanceamento prévio dos dados de treinamento levou a ganhos marginais para todos os classificadores.

Por fim, observou-se também que a inclusão de uma etapa de pré-agrupamento dos dados, antes do treinamento dos classificadores, levou a ganhos, tanto em *F–measure* quanto em acurácia, apenas para o classificador Naïve Bayes. No entanto, mesmo com estes ganhos o desempenho do Naïve Bayes foi muito inferior ao observado para o k-NN sem pré-agrupamento de dados (tanto para *F–measure* quanto para acurácia), que foi o classificador que levou aos melhores resultados para o conjunto de dados estudado neste trabalho. Sendo assim, pode-se dizer que esta etapa de pré-agrupamento de dados pode levar a ganhos para classificadores específicos, mas cabe ao usuário avaliar a efetiva viabilidade da inclusão deste procedimento, uma vez que os resultados obtidos aqui mostraram que os melhores desempenhos foram obtidos por classificadores sem a realização desta etapa adicional.

Como trabalhos futuros pretende-se expandir o conjunto de algoritmos de classificação de dados para técnicas mais avançadas, como máquinas de vetores suporte e redes neurais artificiais, para que se possa obter uma posição definitiva sobre a viabilidade da inclusão desta etapa de pré-agrupamento em previsão de explosões solares. Além disso, pretende-se avaliar também o impacto de diferentes estratégias de balanceamento dos dados.

Referências

Al-Ghraibah, A., Boucheron, L. E., and McAteer, R. T. J. (2015). A study of feature selection of magnetogram complexity features in an imbalanced solar flare prediction data-set. In IEEE, editor, *Proc. of the 15th IEEE International Conference on Data Mining Workshop (ICDMW)*, page 557–564, Atlantic City, USA.

Argento, R. S. V. (2016). Utilização de Ensembles de Redes Neurais MLP para Previsão de Explosões Solares. Dissertação de Mestrado. Faculdade de Tecnologia, Universidade Estadual de Campinas.

Bobra, M. G. and Couvidat, S. (2015). Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm. *The Astrophysical Journal*, 798(2).

Camargos, R. C. (2016). Algoritmos Aglomerativos de Agrupamento Baseados em Teoria de Matrizes. Dissertação de Mestrado em Ciência da Computação. Faculdade Campo Limpo Paulista.

Cinto, T., Gradvohl, A. L. S., Coelho, G. P., and Silva, A. E. A. (2018). Daily solar data and sunspot region summary of 23-24 solar cycle. `http://doi.org/10.5281/zenodo.1307495`. Zenodo. Acessado em: 13-ago-2018.

Colak, T. and Qahwaji, R. (2009). Automated solar activity prediction: A hybrid computer platform using machine learning and solar imaging for automated prediction of solar flares. *Space Weather*, 7(6).

Doni, M. V. (2004). Análise de cluster: Métodos hierárquicos e de particionamento. Trabalho de Graduação. Faculdade de Computação e Informática, Universidade Presbiteriana Mackenzie.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd edition.

Holman, G. D. and Benedict, S. (2007). Solar flare theory. `http://hesperia.gsfc.nasa.gov/sftheory/index.htm`. Acessado em: 01-jul-2018.

Jain, K. A. and Dubes, C. R. (1988). *Algorithms for clustering data*. Prentice Hall, 1st edition.

Kaufman, L. and Rousseeuw, J. P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 1st edition.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. of the International Joint Conference on Artificial Intelligence*, page 1137–1145.

Li, R., Wang, H. N., Cui, Y. M., and Huang, X. (2011). Solar flare forecasting using learning vector quantity and unsupervised clustering techniques. *Science China Physics, Mechanics and Astronomy*, 54(8):1546–1552.

Lima, S. D. S. (2012). Tempestades Geomagnéticas: Origem e Consequência. Trabalho de Conclusão de Graduação em Física. Centro de Ciências e Tecnologia, Universidade Estadual do Ceará.

Linden, R. (2009). Técnicas de agrupamento. *Revista de Sistemas da Informação da FSMA*, (4):18–36.

Liu, C., Deng, N., Liu, Y., Falconer, D., Goode, P. R., Denker, C., and Wang, H. (2005). Rapid change of delta spot structure associated with seven major flares. *Astrophysical Journal*, 1(622):722–736.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal*, 20:359–363.

Moura, M., Gonçalves, L., Sudré, C., Rodrigues, R., and Amaral Jr., A. P. T. (2010). Algoritmo de Gower na estimativa da divergência genética em germoplasma de pimenta. *Horticultura Brasileira*, 28(2):155–161.

Rahman, M. M. and Davis, D. N. (2013). Cluster based under-sampling for unbalanced cardiovascular data. In *Proc. of the 2013 World Congress on Engineering*, volume III, pages 1–6, London, UK.

Rowlett, R. (2013). Solar flare intensity. `http://www.unc.edu/~rowlett/units/scales/solar_flares.htm`. Acessado em: 01-jul-2018.

Vale, N. M. (2005). Agrupamentos de Dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos. Dissertação de Mestrado em Engenharia Elétrica. Pontifícia Universidade Católica do Rio de Janeiro.

Yen, S.-J. and Lee, Y.-S. (2006). Cluster-based sampling approaches to imbalanced data distributions. In Springer, editor, *Lecture Notes in Computer Science (LNCS) - Proc. of the Data Warehousing and Knowledge Discovery Conference*, volume 4081, page 427–436, Krakow,, Poland.

Zhao, Y., Karypis, G., and Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168.