Diferentes abordagens de Subamostragem para Balanceamento da Base de Dados aplicados ao estudo de caso da Classificação de Absenteísmo de Pacientes Clínicos

Lucas V. Darós[1], **Karin S. Komati**[1], **Leandro C. Resendo**[1]

[1]Programa de Pós-Graduação em Engenharia de Controle e Automação (ProPECAut)
Instituto Federal do Espírito Santo (Ifes) Campus Serra
ES-010, Km-6,5 - Manguinhos, Serra - ES, 29173-087

Abstract. *Among the various problems faced by clinics and doctors' offices, it is worth noting the no-show of patients at their scheduled exams. In a scenario with high volumes of data, where the analysis of patient profiles becomes a non-scalable task to be done manually, a data sampling algorithm was proposed to assist the work of patient classification algorithms, in order to predict patients who will or will not attend the exams based on data from consultations and previously obtained patients. Results pointed out that a balanced sampling and using the proposed algorithm was crucial for a good result achieved by the techniques of machine learning classification when compared to a randomized and random sampling.*

Resumo. *Dentre os diversos problemas enfrentados por clínicas e consultórios médicos, destaca-se o não comparecimento (no-show) de pacientes aos seus exames agendados. Em um cenário com alto volume de dados, onde a análise de perfis de paciente torna-se uma tarefa não escalável para ser feita de forma manual, foi proposto um método de amostragem de dados para o auxiliar o trabalho dos algoritmos de classificação de pacientes, a fim de predizer pacientes que irão ou não comparecer aos exames baseados em dados de consultas e pacientes previamente obtidos. Na análise, as técnicas de aprendizado de máquina para a classificação foram aplicadas em uma amostragem balanceada feita de forma aleatória e utilizado o método proposto. Os resultados indicam que o método proposto foi crucial para um bom resultado da classificação.*

1. Introdução

O absenteísmo, ou *no-show*, é o não comparecimento repentino à um exame médico e pode causar graves consequências, tais como: o aumento dos custos (com subutilização da equipe médica e desperdício de recursos materiais); e a diminuição da receita [Harris et al. 2016].

O absenteísmo tem se tornado um problema crônico em clínicas brasileiras. Em 2017, a rede estadual de saúde do estado de Santa Catarina registrou uma falta a cada cinco consultas médicas agendadas, gerando prejuízo estimado aos cofres públicos de R$ 13,4 milhões de reais [Weiss 2017]. A situação não é diferente na cidade de Vitória-ES, em que o número de faltas em consultas médicas dos postos de saúde alcançou 30% do total de consultas realizadas nos anos de 2014 a 2015, que representou um prejuízo aproximado de R$ 39 milhões aos cofres públicos [Furtado et al. 2016].

Uma estratégia para evitar a ausência de pacientes às consultas médicas é realizar o contato, por telefone ou mensagens, poucos dias antes do exame, para recordar aos pacientes sobre o agendamento. Essa estratégia ajuda na redução do absenteísmo, mas impõem despesas às clínicas [Woods 2011]. Assim, é importante que essa estratégia seja bem direcionada para o perfil de pessoas que provavelmente não comparecerão. Um dos resultados do trabalho de [Furtado et al. 2016] é que as ações de contato são direcionados à 4 diferentes perfis específicos, que são responsáveis por 75% das ocorrências de faltas.

A proposta deste trabalho é usar a classificação binária (*show/no-show*) para auxiliar na predição do absenteísmo [Gama et al. 2011]. Sistemas de classificação são sistemas de aprendizado supervisionado, que é uma sub-categoria de Aprendizado de Máquina. Para se alcançar um bom resultado do classificador é importante que o treinamento seja feita em uma base balanceada, que não é o caso. Isto porque, em geral, o problema apresenta mais ocorrências de comparecimento (*show*) do que de falta (*no-show*). De acordo com [Mountassir et al. 2012], há duas maneiras gerais de se resolver o problema de desbalanceamento de dados: realizar alterações nos algoritmos de classificação, ou no conjunto de dados. Para o segundo caso, existem duas maneiras de se modificar os dados: *under-sampling* e *over-sampling*. A primeira técnica consiste em reduzir o número de ocorrências da classe majoritária, cujo o rótulo possui o maior número de ocorrências. Já a segunda técnica, se destina a aumentar a ocorrência das classes minoritárias.

Este trabalho compara diferentes métodos de *under-sampling* para balanceamento da base de treinamento da classificação de *no-show* para uma base de dados real de uma clínica situada em Vitória-ES, dada as informações do paciente e do agendamento.

2. Metodologia

Para a realização das análises, foi utilizado uma base de dados contendo informações acerca de exames médicos e pacientes entre os anos de 2010 e 2017 de uma clínica localizada no município de Vitória, Espírito Santo. No total, a base de dados é composta por 902.979 registros com 35 atributos. Todos os dados que poderiam identificar o paciente foram retirados antes da disponibilização da base para os experimentos deste trabalho. Todos os dados são orientados pelo agendamento e não pelo paciente, isto significa que é possível haver vários agendamentos de consultas de uma única pessoa.

Cada registro dessa base de dados possui uma classificação (rótulo) binária, que identifica o agendamento da consulta do exame como: *Show* e *No-Show*. O primeiro rótulo se refere à todos os exames cujo os pacientes compareceram, já o segundo, se refere aos pacientes ausentes aos agendamentos. Este trabalho não levou em consideração supostos atrasos dos pacientes aos horários do exames.

Inicialmente, para o armazenamento e manipulação de dados, foi utilizado o *MySQL Community Server*, versão 6.3.9. A partir dos dados extraídos, foi utilizado a ferramenta *Visual Studio 2017 Community* para a construção de um programa em C# para a realização da amostragem dos dados. Com os dados balanceados, foram utilizados algoritmos de classificação provenientes da ferramenta *Weka (The Waikato Environment for Knowledge Analysis)* [Hall et al. 2009], que é uma ferramenta gratuita desenvolvida na linguagem de programação Java, contendo uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Os experimentos foram executados em um computador, com processador 2 GHz Intel Core i5, 8 GB de Memória RAM.

A metodologia deste trabalho foi dividida em três etapas: seleção de características, balanceamento dos dados e classificação. Nas próximas sub-seções, serão descritos com detalhes como cada passo foi realizado.

2.1. Seleção de Características

Na etapa de seleção, foram definidas quais características da base de dados original seriam relevantes para serem utilizadas nos experimentos. Algumas possuíam dados incompletos (devido ao não preenchimento de alguns dados de pacientes e exames no momento do cadastro) e redundantes (devido à erros de grafias também no momento do cadastro). A partir deste cenário, foi realizada uma etapa de pré-processamento dos dados, a fim de selecionar registros válidos para o processo de amostragem.

Foram desconsideradas registros com valores ausentes para os atributos: data de nascimento, sexo, cidade do paciente, profissão, tipo do exame, bairro do paciente. E para os dados redundantes dos atributos nome da cidade e profissão, foi realizada uma avaliação de todos os rótulos distintos, a fim de proporcionar uma melhor categorização de cada característica.

As escolhas das características foram baseadas em opiniões de especialistas sobre a rotina de agendamentos e consultas médicas. Escolhida as características, foram selecionados dois conjuntos de registros iniciais. O primeiro contou com 10 características, sendo quatro associados ao exame: mês, ano, dia da semana e descrição; e seis com relação ao paciente: mês de nascimento, ano de nascimento, sexo, profissão, cidade em que reside, e bairro em que reside. Este conjunto apresentou um total de 524.147 exemplares, sendo 477.438 do tipo *show* (91,08 %) e 46709 do tipo *no-show* (8,92 %). Já o segundo conjunto contou com 7 características, sendo os mesmos quatro da seleção anterior associados ao exame: mês, ano, dia da semana e descrição; e três com relação ao paciente: ano de nascimento, sexo e cidade em que reside. Foram selecionados um total de 572.207 registros, sendo 521.769 do tipo *show* (91,18%) e 50.438 do tipo *no-show* (8,82%). A diferença da quantidade de exemplares, nos experimentos com 7 e 10 características, se deve ao fato de que o experimento com 10 características apresentou algumas inconsistências a mais nos registros, justamente por ter mais características.

Dentre as características selecionadas, a base de dados contou com 1.142 profissões distintas, 6 tipos de exames, 364 cidades e 3.605 bairros. Em ambas seleções, os conjuntos de registros apresentaram um cenário desbalanceado. Foi realizado portanto o balanceamento dos dados, a fim de utilizar a mesma quantidade de rótulos em ambas seleções.

2.2. Balanceamento dos Dados

Inspirado na abordagem *under-sampling* de [Mountassir et al. 2012], foi proposta uma técnica de amostragem nesse trabalho, a fim de selecionar os exemplares do tipo *show* mais significativas. O número de rótulos selecionados se equivale ao mesmo número de registros do tipo *no-show*, balanceando assim o conjunto. Cada registro recebeu um *score*, para determinar seu grau de relevância. Quanto maior o *score*, mais relevância possui o registro. Este grau de relevância foi calculado da seguinte maneira:

$$score(i) = \sum_{j=1}^{n} (Q - K_{ij}) \quad (1)$$

onde *n* corresponde ao número de características; Q corresponde ao total de registros da base; e *K* corresponde ao total de registros que possuem o mesmo rótulo do registro *i* em questão. Portanto, o *score* é calculado a partir da soma da diferença entre o total de registros da base e a quantidade de ocorrências do rótulo das características.

Para exemplificar o cálculo do *score*, segue um exemplo apresentado na Tabela 1. A base de dados possui 5 registos e 3 características (mês, ano e exame). O *score* do registro número 1 seria calculado da seguinte maneira: dado o mês igual à janeiro, existem 2 registros com esse mês. Para o ano, existe apenas o próprio registro com o ano igual a 2013. Para o exame, existem outros dois exames com o mesmo rótulo, totalizando 3 exames A. Dado que a base de exemplo possui 5 registos no total, o cálculo do *score* do registro 1, se daria pela soma das subtrações da quantidade total de registros pelas ocorrências dos rótulos desse respectivo registro, conforme mostrado pela Equação 2.

$$score(1) = (5 - 2) + (5 - 1) + (5 - 3) = 8 \quad (2)$$

Tabela 1. Exemplo de dados para cálculo de amostragem

	mês	ano	exame
1	Janeiro	2013	A
2	Janeiro	2014	B
3	Fevereiro	2015	A
4	Fevereiro	2014	C
5	Março	2014	A

Para isto, foi desenvolvido um programa na linguagem de programação C#, que recebe como parâmetro de entrada os dados dos exames extraídos da base de dados e retorna um subconjunto, onde foram selecionadas os *x* registros de *show* mais relevantes (conforme o *score* mais alto), onde *x* corresponde ao número de exames do tipo *no-show*.

2.3. Classificação

O classificador recebe como parâmetro de entrada, um conjunto de dados com rótulos conhecidos, denominado dados de treinamento. O objetivo é ensinar, a partir dos dados, a tarefa de classificação ao classificador. Após a etapa de aprendizagem, o classificador recebe um segundo conjunto de dados, denominado dados de teste. Este conjunto tem por objetivo avaliar o desempenho da aprendizagem, onde o classificador busca determinar corretamente o rótulo dos respectivos registros desses dados. O algoritmo de classificação deve atribuir uma, e somente uma, classe (rótulo) ao exemplar de teste submetida [Oliveira 2016].

Para a etapa de classificação, foram escolhidos os algoritmos *Naive Bayes* [Lewis 1998] e *K-Nearest Neighbors* (*KNN*) [Hwang and Wen 1998] disponíveis na ferramenta *Weka (The Waikato Environment for Knowledge Analysis)* [Hall et al. 2009].

3. Experimentos

Foram realizados 4 experimentos neste trabalho. Em cada sub-seção a seguir, será detalhado qual seleção foi utilizada, qual tipo de amostragem (aleatória ou o método proposto neste trabalho) foi utilizada, e as técnicas de classificação utilizadas.

Foi utilizada a técnica de *holdout*, em que o conjunto foi dividido em dois grupos: 70% para treinamento e 30% para testes.

3.1. Experimento 1

Neste primeiro experimento, foi utilizada a primeira seleção realizada, descrita na seção 2.1, que contém dez características (quatro associadas ao exame, e seis ao paciente).

Dados os 524.147 registros, sendo apenas 46.709 rotulados como *no-show* (cerca de 8,9%), foram selecionados aleatoriamente 46.709 exemplares *show*, ou seja, sem a utilização da técnica de amostragem proposta. Após o balanceamento de 50% para cada rótulo, foram aplicados os algoritmos de classificação *Naive Bayes* e *KNN*. O arquivo de treinamento contou com 65.392 exemplares (70% de 93.418) e o arquivo de teste com 28.026 registros (30% de 93.418). Ambos os arquivos de treinamento e teste possuíam as mesmas quantidades de rótulos escolhidos aleatoriamente.

3.2. Experimento 2

O segundo experimento teve como objetivo servir de comparação com o primeiro. Foi utilizada a primeira seleção de dados porém, dessa vez, os 46.709 registros *show* foram selecionados utilizando o método de amostragem desenvolvido neste trabalho. Foram utilizados os mesmos algoritmos de classificação do primeiro experimento, assim como também a mesma proporção de registros para treinamento e teste (70% - 30%).

3.3. Experimento 3

Neste terceiro experimento, foi utilizado o segundo conjunto de dados selecionados, descrita na seção 2.1, que contém 7 características (sendo quatro associados ao exame e três com relação ao paciente).

Esta segunda seleção contou com 572.207 registros, sendo apenas 50.438 rotuladas como *no-show* (cerca de 8,82%). Assim como realizado no Experimento 1, este experimento não utilizou o método de amostragem proposto neste trabalho, ou seja, foram escolhidos 50.438 exemplares do tipo *show* aleatoriamente. Os algoritmos de classificação utilizados também foram os mesmos dos experimentos 1 e 2, mantendo também a proporção de registros em 70%-30% para treinamento e teste, e a mesma quantidade de rótulos em cada proporção.

3.4. Experimento 4

Seguindo a mesma ideia de comparação entre os Experimentos 1 e 2, este quarto experimento tem por objetivo ser comparado com o terceiro experimento proposto. Fazendo uso também do segundo conjunto de registros selecionadas neste trabalho, este experimento selecionou 50.438 registros do tipo *show*, a fim de igualar a mesma quantidade de registros do tipo *no-show*.Para a etapa de classificação, este experimento também utilizou os mesmos algoritmos de classificação dos experimentos anteriores: *Naive Bayes* e *KNN*. O percentual de treinamento-teste foi de 70%-30%, contendo a mesma quantidade de rótulos em cada um dos arquivos.

4. Resultados

Foram medidos em cada um dos experimentos: a porcentagem de acertos e erros de cada um dos classificadores. O resultado de cada experimento está descrito na Tabela 2. Nessa tabela: a primeira coluna identifica os experimentos realizados; na segunda coluna é identificada se as classificações são corretas ou incorretas; na terceira e quinta colunas são identificadas as ocorrências de acertos de erros dos algoritmos *Naive Bayes* e *KNN*, respectivamente; e, quarta e sexta colunas os percentuais de acertos e erros dos classificadores. Por exemplo, na primeira e segunda linhas desta tabela são referentes aos resultados do Experimento 1. Considerando a primeira linha, nas terceira e quinta colunas são apresentadas as quantidades de acerto dos algoritmos *Naive Bayes* e *KNN*, respectivamente, e nas quarta e sexta colunas as a porcentagem de acerto dos algoritmos. Analogamente, na segunda linha são apresentadas da ocorrências e taxas de erros dos classificadores. Adicionalmente, foi colocado em destaque as porcentagens de acerto dos dois algoritmos de classificação.

Tabela 2. Resultados dos experimentos

Experimentos	Classificação	Naive Bayes	Naive Bayes (%)	*KNN*	*KNN*(%)
1	Corretas	14023	**50,04%**	13847	**49,41%**
1	Incorretas	14003	49,96%	14179	50,59%
2	Corretas	25925	**92,50%**	23115	**82,48%**
2	Incorretas	2101	7,50%	4911	17,52%
3	Corretas	18678	**61,72%**	17357	**57,35%**
3	Incorretas	11585	38,28%	12906	42,65%
4	Corretas	27506	**90,89 %**	22521	**74,42%**
4	Incorretas	2756	9,11%	7741	25,58%

Observa-se na Tabela 2 um percentual de acerto maior, em ambos algoritmos de classificação utilizados, quando se é utilizado o método de amostragem proposto neste trabalho. Entre o Experimento 1 e 2, houve um aumento de 42,47% de acerto do classificador *Naive Bayes* e de 33,07% para o *KNN*. Na comparação dos Experimentos 3 e 4, também houve um aumento significativo da classificação: 29,17% para *Naive Bayes* e 17,07% para *KNN*.

Outra forma utilizada nesse trabalho para representar o desempenho dos classificadores foi a matriz de confusão [Stehman 1997]. A Tabela 3 apresenta a matriz de confusão dos experimentos para o algoritmo de *Naive Bayes*, enquanto a Tabela 4 apresenta a matriz de confusão dos experimentos para o algoritmo *KNN*. Tomando como exemplo o Experimento 1, na Tabela 3 o valor 5.263 corresponde ao número de registros do tipo *show* que foram preditos corretamente pelo classificador. O valor de 8.750 correspondem aos registros do tipo *no-show* classificados incorretamente como *show*. Na segunda linha, ainda para o Experimento 1, o valor 5.253 corresponde ao número de exemplares do tipo *show* classificados incorretamente como sendo do tipo *no-show*, enquanto o valor de 8.760 corresponde ao número de exemplares do tipo *no-show* preditos corretamente pelo classificador. Nas demais linhas desta tabela estão apresentados os resultados para o *Naive Bayes* nos demais experimentos.

Observou-se também na Tabela 2 que o primeiro conjunto, que utilizou carac-

terísticas selecionadas, quando balanceadas utilizando o método proposto neste trabalho (Experimento 2), obteve melhores resultados de assertividade na classificação quando comparado ao experimento que usou o segundo conjunto de dados, com apenas 7 características (Experimento 4).

Existem dois tipos classificações erradas: o tipo 1, quanto o real é *show* e foi predito *no-show* e o tipo 2 quando o real é *no-show* e o predito foi *show*. Para o caso de uma abordagem de *overbooking*, é ideal que o segundo tipo de erro seja evitado, uma vez que é preciso que se saiba quando de fato o paciente não irá comparecer. Observando as classificações realizadas pelos algoritmos nas Tabelas 3 e 4, para os dois algoritmos de classificação, nota-se que o método proposto nesse trabalho proporcionou uma redução significativa do tipo erro 2 em todos os casos, com exceção do experimento 4, utilizando *KNN*. Além disso, os experimentos utilizando *Naive Bayes* apresentaram o tipo de erro 2 menor que do que os experimentos que utilizaram *KNN*.

Tabela 3. Matriz de Confusão para classificação *Naive Bayes*

	SHOW	NO-SHOW	
Experimento 1	5263	8750	SHOW
Experimento 1	5253	8760	NO-SHOW
Experimento 2	13776	237	SHOW
Experimento 2	1864	12149	NO-SHOW
Experimento 3	7031	8100	SHOW
Experimento 3	3485	11647	NO-SHOW
Experimento 4	14481	650	SHOW
Experimento 4	2106	13025	NO-SHOW

Tabela 4. Matriz de Confusão para classificação KNN

	SHOW	NO-SHOW	
Experimento 1	5147	8866	SHOW
Experimento 1	5313	8700	NO-SHOW
Experimento 2	10451	3562	SHOW
Experimento 2	1349	12664	NO-SHOW
Experimento 3	8276	6855	SHOW
Experimento 3	6051	9081	NO-SHOW
Experimento 4	8253	6878	SHOW
Experimento 4	863	14268	NO-SHOW

Portanto, a partir das diferenças de classificação entre os Experimentos 1 e 2, e os Experimentos 3 e 4 apresentadas nas Tabela 2, 3 e 4, nota-se uma melhora relevante no desempenho da classificação, e na diminuição do tipo de erro mais grave, tipo 2, para o processo de *overbooking* nos experimentos utilizando *Naive Bayes*.

5. Conclusões

Conhecer o perfil e analisar o comportamento de pacientes são tarefas que se tornam cada vez mais complexas na medida em que essas variáveis se expandem. Para a tarefa de classificação, é fundamental que haja não só balanceamento, como também a

representatividade eficiente de dados, a fim de aprimorar a tarefa de classificação. Tal argumento é comprovado pelos resultados apresentados, uma vez em que os experimentos que possuíam o emprego do algoritmo de amostragem, apresentaram resultados bem superiores no ato de classificação frente aos experimentos com amostragem aleatória.

Como trabalhos futuros, há a pretensão de explorar demais algoritmos de classificação fornecidos na literatura, a fim de realizar uma comparação entre os mesmos. Uma outra pretensão é abordar a estratégia de *overbooking* com o emprego das tarefas de classificação, simulando agendamentos de pacientes.

Referências

Furtado, L. P., Fernandes, P. C., and dos Santos, J. H. (2016). Redução de faltas em consultas médicas e otimização dos recursos da saúde pública em vitória-es por meio de mineração de dados e big data. In *Anais do XV Congresso Brasileiro de Informática em Saúde (CBIS 2016)*, pages 23–25. Sociedade Brasileira de Informática em Saúde (SBIS).

Gama, J., Faceli, K., Lorena, A., and De Carvalho, A. (2011). *Inteligência artificial: uma abordagem de aprendizado de máquina*. Grupo Gen - LTC.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Harris, S. L., May, J. H., and Vargas, L. G. (2016). Predictive analytics model for healthcare planning and scheduling. *European Journal of Operational Research*, 253(1):121–131.

Hwang, W.-J. and Wen, K.-W. (1998). Fast knn classification algorithm based on partial distance search. *Electronics letters*, 34(21):2062–2063.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer.

Mountassir, A., Benbrahim, H., and Berrada, I. (2012). An empirical study to address the problem of unbalanced data sets in sentiment classification. In *Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on*, pages 3298–3303. IEEE.

Oliveira, P. H. M. A. (2016). *Detecção de fraudes em cartões: um classificador baseado em regras de associação e regressão logística*. PhD thesis, Universidade de São Paulo.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89.

Weiss, C. E. (2017). Em cada cinco consultas médicas agendadas, um paciente falta e gera prejuízo de R$ 13,4 milhões em SC. Disponível em: ¡http://dc.clicrbs.com.br/sc/estilo-de-vida/noticia/2017/03/em-cada-cinco-consultas-medicas-agendadas-um-paciente-falta-e-gera-prejuizo-de-r-13-4-milhoes-em-sc-9739621.html¿. Acesso em 01 de ago. 2018.

Woods, R. (2011). The effectiveness of reminder phone calls on reducing no-show rates in ambulatory care. *Nursing Economics*, 29(5):278–282.