Detectando Doença de Parkinson - Uma Comparação de Modelos de Aprendizagem de Máquina com Redução de Dimensionalidade Diferencialmente Privada

Manuel E. B. Filho[1], Maria de Lourdes M. Silva[1], Patricia V. da S. Barros[1], César L. C. Mattos[1], Javam C. Machado[1]

[1]Departamento de Computação (DC) – Universidade Federal do Ceará (UFC)
CEP 60440-900 – Fortaleza – CE – Brazil

{edvar.filho, malu.maia, patricia.barros, javam.machado}@lsbd.ufc.br,

cesarlincoln@dc.ufc.br

Abstract. *This paper aims to present a comparison of machine learning models using two dimensionality reduction approaches in data pre-processing, one private and one non-private. The problem is to classify patients as having Parkinson's disease or not. Models were compared based on their ability to diagnose the disease based on a collection of vocal data. The results obtained indicate that the Gaussian and textit Random Forest process models were the best approaches without and with differential privacy restriction, respectively.*

Resumo. *Este artigo visa apresentar uma comparação de modelos de aprendizagem de máquina utilizando duas abordagens de redução de dimensionalidade no pré-processamento dos dados, uma privada e outra não privada. O problema consiste em classificar pacientes como portadores ou não da doença de Parkinson. Modelos foram comparados a partir de suas capacidades em diagnosticar a doença a partir de uma coleta de dados vocais. Os resultados obtidos indicam que os modelos de processo Gaussiano e Random Forest foram as melhores abordagens sem e com restrição de privacidade diferencial, respectivamente.*

1. Introdução

A doença de Parkinson é um distúrbio neurológico causado pela degeneração de uma pequena parte do cérebro, afetando o controle de movimentos de indivíduos. Segundo [Dias and Limongi 2003], alterações na qualidade da voz podem ser detectadas no estágio inicial da doença. Por este motivo, algumas pesquisas visam diagnosticar esse distúrbio a partir da detecção de deficiências vocais [Sakar and Kursun 2010].

Com a finalidade de auxiliar no avanço dos estudos para a realização do diagnóstico da doença de Parkinson, os pacientes devem abrir mão de seus dados de saúde para que o processo possa ser realizado com êxito. No entanto, a divulgação de dados de saúde deve ser feita garantindo a privacidade dos pacientes cujos registros serão publicados. Esse procedimento deve ser feito sem inutilizar os dados aplicados nas análises.

Neste contexto, visto a importância da privacidade dos indivíduos cujos dados de saúde foram publicados, o presente artigo possui o objetivo de elaborar uma metodologia capaz de detectar se um indivíduo possui a doença de Parkinson, além de comparar os

resultados dos modelos de aprendizagem de máquina quando aplicados aos dados que protegem a privacidade dos pacientes e dados originais dos mesmos.

Na Seção 2 abordamos os trabalhos relacionados, em seguida, descrevemos a metodologia adotada para a avaliação dos modelos na Seção 3. Avaliamos os experimentos computacionais na Seção 4 e, por fim, concluiremos e apontaremos direções para investigações futuras na Seção 5.

2. Trabalhos Relacionados

Pesquisas sobre detecção de doença de Parkinson através da análise de coletas vocais estão amplamente disponíveis na literatura. No entanto, a garantia de privacidade dos pacientes cujos padrões são envolvidos na análise ainda encontra-se em estágio inicial.

[Ramani and Sivagami 2011] realizou uma análise comparativa entre regressão logística, máquinas de vetores de suporte (*Support Vector Machine*, SVM), *Random Forest*, entre outros, sem abordar a privacidade dos indivíduos do conjunto de dados. O modelo que obteve melhores resultados foi o *Random Forest*, com a menor taxa de erro comparando-se aos demais modelos abordados, treinados com os atributos de maiores relevâncias para o modelo. O conjunto de dados utilizado foi disponibilizado pela Universidade de Oxford e contém 195 registros vocais e 23 atributos, de 31 pessoas.

Com o mesmo conjunto de dados, [Das 2010] compara regressão logística, árvore de decisão e redes neurais artificiais, mas sem trazer uma abordagem que garanta a privacidade dos usuários. Os resultados obtidos a partir de uma seleção de atributos apresentaram boa acurácia, especialmente com os modelos neurais.

Em [Sakar et al. 2019] é criado um novo conjunto de dados, o mesmo usado no presente trabalho, que realiza três coletas a cada um dos pacientes. Um total de 755 atributos foram gerados a partir das coletas. O trabalho realiza um comparativo entre regressão logística, SVM, *Random Forest* e outros modelos a partir de subconjuntos dos atributos, desconsiderando a privacidade dos pacientes. O modelo *Random Forest* obteve os melhores resultados. A comparação via acurácia, *F1-Score* e o coeficiente de correlação de Matthews indicou que o uso dos atributos básicos apresenta melhores resultados para os modelos e que dentre eles o SVM é aquele que tem melhor comportamento.

Neste trabalho, aderimos práticas dos trabalhos citados, com o diferencial da utilização da redução de dimensionalidade que garante a privacidade dos dados, além de compararmos os experimentos na base de dados com e sem privacidade.

3. Metodologia

O problema considerado trata-se de uma classificação binária, em que um paciente pode ser rotulado como *portador* ou *não portador* da doença de Parkinson. O conjunto de dados utilizado foi extraído da plataforma *Kaggle*[1], e consiste em coletas de 252 pacientes, onde 188 possuem doença de Parkinson, 107 homens e 81 mulheres com idades entre 33 e 87 anos. Os 64 indivíduos saudáveis, 23 homens e 41 mulheres, possuem faixa etária entre 41 e 82 anos. Cada paciente possui três registros referentes ao seu quadro.

Os atributos originais do conjunto de dados foram obtidos através de algoritmos de processamento de sinais de voz aplicados às gravações de fala dos pacientes. Também

[1] https://www.kaggle.com/dipayanbiswas/parkinsons-disease-speech-signal-features

foram incluídos na base o identificador, gênero, idade, e outras informações do paciente relevantes para o problema. Ao todo, somam-se 756 atributos [Sakar et al. 2019].

Os procedimentos de pré-processamento dos dados, redução de dimensionalidade, treinamento e avaliação dos modelos serão descritos nas seções abaixo. Os códigos em Python usados nos experimentos estão disponíveis[2].

3.1. Privacidade Diferencial e Mecanismo Gaussiano

A privacidade diferencial é um conceito formalizado pela matemática Cynthia Dwork [Dwork et al. 2014] para garantir a privacidade de indivíduos em um conjunto de dados usados para a extração de padrões.

Definição 1 (Privacidade Diferencial)**.** Sejam os conjuntos de dados vizinhos D_1 e D_2 quaisquer, que são aqueles que diferem em apenas um registro. Seja o conjunto S contido na variação de resultados de M. Um mecanismo ***M*** é dito (ϵ, δ)-diferencialmente privado se a seguinte condição for satisfeita:

$$Pr[M(D_1) \in S] \leq \exp(\varepsilon) \times Pr[M(D_2) \in S] + \delta, \tag{1}$$

em que ε é um limite para a perda de privacidade de uma consulta e δ uma relaxação da condição.

Este trabalho considera o uso do mecanismo Gaussiano [Dwork et al. 2006], que adiciona ruído Gaussiano com média 0 e variância σ^2 na saída de uma consulta, visando torná-la privada.

Definição 2 (Sensibilidade l_2)**.** Para uma consulta $f : D \rightarrow \mathbb{R}$, a sensibilidade l_2 da função f é definida como:

$$\Delta_2 f = \max \|f(D_1) - f(D_2)\|_2, \tag{2}$$

em que $\| \cdot \|_2$ denota a norma Euclidiana.

Definição 3 (Mecanismo Gaussiano)**.** Dada uma função $f : D \rightarrow \mathbb{R}$ sobre um conjunto de dados D, se $\sigma = \Delta_2 f \sqrt{\frac{2 \times \ln(\frac{2}{\delta})}{\epsilon}}$ e $N(0, \sigma^2)$ for a amostra independente de uma distribuição Gaussiana, o mecanismo M provê (ϵ, δ)-privacidade diferencial quando segue:

$$M(D) = f(D) + N(0, \sigma^2). \tag{3}$$

3.2. Pré-processamento e Redução de Dimensionalidade

A elevada dimensionalidade dos dados pode prejudicar o treinamento dos modelos de aprendizagem, fenômeno conhecido como "maldição da dimensionalidade". Visando reduzir a dimensão original (superior a 750), uma abordagem privada, cuja privacidade é garantida, e outra não privada foram aplicadas em uma etapa de pré-processamento.

Na abordagem não privada, a técnica de Análise de Componentes Principais (*Principal Component Analysis*, PCA) foi usada para reduzir a dimensão dos dados. A projeção foi obtida a partir da seleção dos 100 autovetores, que representam aproximadamente 88% da variância explicada dos dados, relacionados aos seus 100 maiores autovalores calculados a partir da decomposição de valores singulares da matriz de covariância dos dados.

[2] `https://github.com/EdvarFilho/Detectando-Doenca-Parkinson`

No cenário privado, considerou-se o algoritmo Mod_SULQ, proposto em [Chaudhuri et al. 2012], que consiste em gerar uma matriz simétrica a partir dos dados originais por meio de uma transformação matricial e adicionar a esta matriz recém gerada uma matriz ruído Gaussiano com desvio padrão obtido a partir de uma combinação da quantidade de registros e atributos do conjunto de dados e dos parâmetros de privacidade ϵ e δ que foram configurados com o valor 1 para obter os autovalores e autovalores.

3.3. Aprendizagem Supervisionada e Avaliação

Após o pré-processamento, os dados foram separados em conjuntos de treino e teste, sendo 25% dos dados para teste. Nessa etapa, todas as coletas de um mesmo paciente são exclusivamente do conjunto de treinamento ou do conjunto de teste, evitando vazamento de informação entre os conjuntos.

Os modelos de aprendizagem supervisionada utilizados foram: Regressão Logística (RL), Análise do Discriminante Gaussiano (AGD), K-*Nearest Neighbors* (KNN), Árvore de Decisão (AD), *Random Forest* (RF), Máquina de Vetores de Suporte (SVM) e Processo Gaussiano (PG). Os hiperparâmetros dos modelos, quando necessário, foram selecionados automaticamente via *grid search*, pois há modelos que não possuem hiperparâmetros.

As métricas utilizadas para avaliar os modelos são: acurácia (A), precisão (Pr), revocação (R), *F1-Score* (F1), curva ROC (*Receiver Operating Characteristic*) e área sob a curva ROC (AUC). Além disso, matrizes de confusão foram usadas para detalhar o desempenho de cada modelo.

4. Experimentos

Após a etapa de seleção de hiperparâmetros, os valores selecionados foram usados para avaliar os modelos nos dados separados para teste. Os valores escolhidos estão apresentados na Tabela 1.

Tabela 1. Seleção dos hiperparâmetros para os modelos avaliados nas abordagens não privada (ANP) e privada (AP).

Modelos	Hiperparâmetros	Valores testados	Valores selecionados - ANP	Valores selecionados - AP
RL	α	0.1, 0.1, 1	1	1
	Épocas	100, 1000	100	1000
AD	Prof. Máxima	1, ..., 50	1	5
	Índ. Pureza	Entropia e Gini	Entropia	Gini
SVM	Kernel	Polinomial e Gaussiano	Polinomial	Polinomial
	C	2^{-3}, ..., 2^{2}	2^{-3}	2^{2}
	grau	3, 4, 5	3	3
RF	Estimadores	100, ..., 150	146	137
	Prof. Máxima	1, ..., 50	9	8
	Índ. Pureza	Entropia e Gini	Gini	Gini
KNN	k	3, 5, 7, 9, 11	9	5
PG	Kernel	Gaussiano, Materno, QR, Periodico	QR	Materno

Os resultados finais obtidos para as abordagens sem e com privacidade estão apresentados nas Tabela 2, respectivamente. Pode-se perceber que, com relação as métricas

avaliadas, os modelos em ambas abordagens possuem resultados semelhantes, onde a variação dos resultados, independente do modelo, ocorre devido à adição do ruído do mecanismo Gaussiano.

Tratando-se da abordagem sem privacidade, a aplicação do modelo de processo Gaussiano obteve resultados positivos, apresentando a melhor acurácia, precisão, revocação e *F1-Score*. Na abordagem privada, o melhor modelo foi o *Random Forest*, com acurácia, precisão, revocação e *F1-Score* superiores. É interessante perceber que na abordagem sem privacidade garantida, logo após o processo Gaussiano, o modelo *Random Forest* é o que obtém o segundo melhor resultado. Analogamente, na abordagem com privacidade, o segundo melhor modelo, após o *Random Forest*, é o processo Gaussiano.

Tabela 2. Métricas obtidas nos experimentos computacionais.

	Abordagem Não Privada					Abordagem Privada				
Modelo	A	Pr	R	F1	AUC	A	Pr	R	F1	AUC
RL	0.68	0.76	0.68	0.70	0.79	0.67	0.75	0.67	0.69	0.76
AGD	0.76	0.73	0.76	0.71	0.64	0.75	0.71	0.75	0.67	0.62
AD	0.76	0.73	0.76	0.74	0.62	0.75	0.73	0.75	0.74	0.70
SVM	0.80	0.78	0.80	0.78	0.71	0.79	0.79	0.79	0.75	0.70
RF	0.80	0.80	0.80	0.76	0.75	**0.81**	**0.81**	**0.81**	**0.79**	**0.81**
KNN	0.79	0.78	0.79	0.76	0.63	0.79	0.77	0.79	0.75	0.62
PG	**0.81**	**0.82**	**0.81**	**0.78**	**0.83**	0.79	0.78	0.79	0.75	0.81

Ao comparar ambas as abordagens, é possível observar que na maioria das métricas adotadas os valores são similares. As divergências são justificadas pela adição do ruído Gaussiano na redução de dimensionalidade diferencialmente privada e pelos diferentes resultados na etapa de seleção dos hiperparâmetros. Nos dois cenários o modelo de Regressão Logística obteve o pior desempenho, indicando que os dados não são linearmente separáveis.

As matrizes de confusão estão apresentadas na Tabela 3, apresentando os valores de verdadeiros positivos (VP), pacientes com doença de Parkinson testados corretamente, assim como os verdadeiros negativos (VN) são pacientes saudáveis corretamente detectados, valores de falsos positivos (FP) e falsos negativos (FN), que consistem respectivamente, em pacientes saudáveis testados com doença de Parkinson e pacientes com a doença apontados como saudáveis pelo modelo. Note que o último caso constitui o pior cenário para o problema em questão.

Tabela 3. Matrizes de confusão obtidas nas abordagens sem e com privacidade.

	Abordagem Não Privada				Abordagem Privada			
Modelo	VP	VN	FP	FN	VP	VN	FP	FN
RL	93	35	13	48	93	34	14	48
AGD	136	8	40	5	138	4	44	3
AD	126	17	31	15	121	20	28	20
SVM	133	18	30	8	137	13	35	4
RF	138	13	35	3	**137**	**17**	**31**	**4**
KNN	136	14	34	5	135	14	34	6
PG	**138**	**16**	**32**	**3**	137	12	26	4

5. Conclusão

Com a utilização de coletas para diagnósticos de doença de Parkinson, comparamos modelos de aprendizagem em abordagens privada e não privada, esta que ativa a privacidade diferencial durante a redução de dimensionalidade; e comparamos os resultados dos modelos com a finalidade de prevermos se um paciente possui a doença ao mesmo tempo que a privacidade dele não é afetada e analisamos se a capacidade dos modelos mudam quando temos dados privados. Com essa análise é possível notar que podemos utilizar os dados que garantem a privacidade dos pacientes de modo que não tenhamos resultados discrepantes, comparando-os com a aplicação sobre os dados originais.

Os resultados experimentais realizados indicam que o melhor modelo obtido para a abordagem sem privacidade garantida é o processo Gaussiano com função Racional Quadrática como função de *kernel*. Já para a abordagem com privacidade diferencial garantida, o melhor modelo foi o *Random Forest*, com 137 estimadores. De maneira importante, os resultado reportados apontam que a tarefa em questão pode ser realizada mesmo na presença do mecanismo de privacidade diferencial considerado.

Trabalhos futuros envolvem avaliar outros algoritmos de aprendizagem, como redes neurais artificias; aplicar o método *Leave-One-Subject-Out* na etapa de avaliação; verificar o desempenho dos classificadores com diferentes níveis de privacidade; e aplicar técnicas de de redução de dimensionalidade diferencialmente privada alternativas, como proposto em [Chaudhuri et al. 2012].

Referências

Chaudhuri, K., Sarwate, A., and Sinha, K. (2012). Near-optimal differentially private principal components. In *Advances in Neural Information Processing Systems*, pages 989–997.

Das, R. (2010). A comparison of multiple classification methods for diagnosis of parkinson disease. *Expert Systems with Applications*, 37(2):1568–1572.

Dias, A. E. and Limongi, J. C. P. (2003). Tratamento dos distúrbios da voz na doença de parkinson: o método lee silverman. *Arquivos de Neuro-Psiquiatria*, 61(1):61–66.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

Ramani, R. G. and Sivagami, G. (2011). Parkinson disease classification using data mining algorithms. *International journal of computer applications*, 32(9):17–22.

Sakar, C. O. and Kursun, O. (2010). Telediagnosis of parkinson's disease using measurements of dysphonia. *Journal of medical systems*, 34(4):591–599.

Sakar, C. O., Serbes, G., Gunduz, A., Tunc, H. C., Nizam, H., Sakar, B. E., Tutuncu, M., Aydin, T., Isenkul, M. E., and Apaydin, H. (2019). A comparative analysis of speech signal processing algorithms for parkinson's disease classification and the use of the tunable q-factor wavelet transform. *Applied Soft Computing*, 74:255–263.