ANÁLISE DE SENSIBILIDADE DOS PARÂMETROS DO ALGORITMO K VIZINHOS INFORMATIVOS MAIS PRÓXIMOS PARA PROBLEMAS DE CLASSIFICAÇÃO DE DADOS

Vinícius Gomes Pajaro Grande e Leandro Augusto da Silva

Apoio:PIBIC Mackenzie

RESUMO

O problema da parametrização é muito comum nos algoritmos usados em Mineração de Dados, pois a melhor escolha para resolver um determinado problema não é robusta o suficiente para ser generalizada a outras classes de problemas. Em geral, o melhor parâmetro é definido em um processo empírico onde são realizados diversos testes utilizando métricas de avaliação afim de decidir qual é a melhor medida para resolver um problema especifico. O LI-KNN(do inglês Locally Informative - K Nearest Neighbor) é um algoritmo interessante proposto para ser eficiente em situações em que a separação das classes não é bem definida. Porém os autores que propuseram o algoritmo definiram os melhores parâmetros sem uma apresentação da metodologia que pudesse ser reproduzível ou ao menos justificável. Neste artigo é utilizada uma metodologia de literatura proposta para investigar classes sobrepostas e a relação com a escolha de parametrização através de análises quantitativas e qualitativas e do uso de bases reais e artificiais afim de comparar e justificar melhor a escolha destes parâmetros e analisar a sua sensibilidade em situações diversas. Os resultados da melhor parametrização de LI-KNN são contrastados com algoritmos clássicos como Random Forest e SVM. Os resultados para conjuntos de dados sintéticos e conjuntos de dados reais mostram através de gráficos e tabelas a eficiência de LI-KNN em contraste com KNN, Random Forest e SVM.

Palavras-chave: Classificação de dados. Vizinho mais próximo. Informatividade.

ABSTRACT

The parameterization problem is very common in the algorithms used in Data Mining, because the best choice for solving a particular problem is not robust enough to be generalized to other problem classes.In general, the best parameter is an empirical process which various tests are performed using evaluation metrics to decide what is the best measure to solve a specific problem.Locally Informative Nearest Neighbor or simply LI-KNN is an interesting algorithm proposed to be efficient in situations that the classes separations are not well-defined. However, the authors that proposed the algorithm define the best parameters without a presentation of a methodology that is possible to be reproduced or at least justifiable. In this paper is used a literature methodology proposed to investigate overlapped classes versus

parametrization choice through quantitative and qualitative analyzes and the use of real and artificial databases in order to better compare and justify the choice of parameters and analyze their sensitivity in various situations. The results of best parametrization of LI-KNN is contrasted with classical algorithms such as Random Forest and SVM. The results for synthetic datasets and real datasets show through graphs and tables the efficiency of LI-KNN in contrasting to KNN, Random Forest and SVM.

Keywords: Data classification. Nearest neighbor. Informativeness.

1. INTRODUÇÃO

O algoritmo K Vizinhos Mais Próximos ou KNN (do inglês, K Nearest Neighbor) é um dos mais conhecidos e utilizados na literatura de análise preditiva de dados (AHA, 1991;COVER; HART, 1967;HAN, 2011;WU; KUMAR, 2009). A predição com o algoritmo KNN é feita por meio da atribuição de um rótulo a um objeto em análise, de acordo com a classe majoritária dos K objetos mais próximos, sendo K um parâmetro definido pelo usuário e a proximidade estabelecida por uma medida de similaridade, normalmente a distância Euclidiana (Silva, Peres e Boscarioli, 2015).

Uma de suas fraquezas é que a escolha do melhor parâmetro K é variante ao problema em análise, pois isso depende dos dados a serem analisados e precisa ser ajustado caso a caso (Song et al., 2014; Moreira e Silva, 2016; Moreira e Silva, 2017).

Outra desvantagem do KNN se refere ao conjunto de treinamento. O algoritmo considera todos os objetos armazenados neste conjunto com a mesma relevância na construção da fronteira de decisão. Isso pode degradar o desempenho do algoritmo principalmente em situações em que se faz presente objetos com ruídos e com classes sobrepostas (Song et al., 2014).

Com o objetivo de minimizar parte das desvantagens do KNN, principalmente a de dependência do conjunto de dados e do valor K, Song et al. (2014) apresentaram uma medida de similaridade baseada na informatividade dos objetos mais próximos, à qual é superior em desempenho de acurácia ao uso da distância Euclidiana em dados sobrepostos e tem menor sensibilidade a escolha do parâmetro K. O algoritmo foi chamado de LI-KNN, do inglês, Locally Informative - K Nearest Neighbor. No entanto, os autores não deixam claro como os parâmetros deste algoritmo devem ser configurados e os experimentos não são feitos para um conjunto consistente de problemas e os resultados não são comparados com outros algoritmos de classificação de dados.

Sendo assim, o objetivo principal desta pesquisa é analisar a sensibilidade dos parâmetros do algoritmo LI-KNN de forma genérica, ou seja, que sirva para diferentes problemas de classificação de dados. Como objetivo secundário deseja-se preparar um conjunto de dados sintéticos, controlando a sobreposição dos mesmos para analisar a sensibilidade dos algoritmos e, uma vez que se obtenha as melhores

configurações, o LI-KNN será comparado com os algoritmos KNN, Random Forest e SVM com conjuntos de dados públicos na literatura.

Além da introdução, este trabalho apresenta o referencial teórico da pesquisa, com explicação do LI-KNN e os parâmetros, isso na Seção 2. Em seguida se apresenta a metodologia da pesquisa, Seção 3. E os resultados e análises na Seção 4. Por fim, Seção 5, a conclusão e trabalhos futuros são apresentados.

2. REFERENCIAL TEÓRICO

A. Classificação de Dados

A classificação de dados é uma tarefa da mineração de dados que consiste na modelagem de um algoritmo por meio de um processo de aprendizagem (Peres e Boscarioli, 2016;Han, 2011). Esse processo é conduzido a partir de um repertório de dados históricos de alguma área de aplicação. A finalidade deste aprendizado é o uso do algoritmo em um processo de predição, objetivando a inferência de situações futuras, ainda desconhecidas ao problema em análise. O repertório de dados é comumente chamado de conjunto de treinamento e é formado por objetos caracterizados por atributos descritivos e classificatórios. A apresentação do conjunto de treinamento iterativamente à um algoritmo de classificação faz com que esse seja capaz de realizar em situações futuras, onde se tem um objeto novo apenas com os atributos descritivos, a predição do atributo classe (ou rótulo) (Peres e Boscarioli, 2016;Han, 2011). A proposta para o desenvolvimento de algoritmos de classificação de dados baseia-se, de maneira geral, em conceitos de diferentes áreas de pesquisa como, por exemplo, Estatística (MAIMON, 2010), Inteligência Artificial (JAIN, 1996) e Aprendizagem de Máquina (AHA, 1991). Os classificadores têm utilidade em diferentes áreas do conhecimento como, por exemplo, no auxílio à tomada de decisão em clientes que obtêm empréstimo bancário como sendo bom pagador, para aqueles que honram com os compromissos de pagamento em dia, e como mau pagador, os que atrasam os seus pagamentos; entre outros (MAIMON; ROKACH, 2010). Na área da medicina com o auxílio de diagnóstico em exames de imagens como sendo sadio ou patológico (SANTOS et al., 2013). Em redes sociais para predizer a polaridade da mensagem como positiva, negativa ou neutra (LIMA et al., 2015). Em geral, em toda área do conhecimento passível de se gerar dados com atributos descritivos e classificatórios conclui-se que é possível treinar um algoritmo de classificação para uso em um processo de inferência de situações futuras, ainda desconhecidas do negócio em análise (MAIMON, 2010).

B. Algoritmo KNN

Em classificação de dados, o algoritmo KNN de K Vizinhos mais Próximo (do inglês, K Nearest Neighbor) é um dos métodos considerado como referência na literatura (COVER; HART, 1967; WU; KUMAR, 2009). O processo de treinamento desse algoritmo é baseado em distância, isto é, durante o processo de classificação é necessário realizar um cálculo de distância entre o objeto em análise e cada objeto armazenado no conjunto de treinamento, ordenar os objetos armazenados pelas menores distâncias e, finalmente, atribuir como classe do objeto em análise, aquela que é a maioria das classes dos vizinhos mais próximos (WETTSCHERECK, 1994). O vasto uso do classificador KNN deve-se principalmente à sua simplicidade de implementação e a alta acurácia no reconhecimento correto em tarefas de classificação (WU; KUMAR, 2009; MAIMON; ROKACH, 2010). A implementação do KNN depende de dois parâmetros apenas, um valor inteiro K (que deve ser ≥ 1) e uma medida de similaridade (geralmente a distância Euclidiana). Como se pode perceber deste algoritmo, o KNN não realiza efetivamente um aprendizado. Dessa maneira, é necessário fazer uma comparação com o conjunto de treinamento toda vez que um novo objeto é apresentado para a classificação, ou seja, não há existência de um modelo e por isso ele é denominado de Lazy learning - (aprendizado preguiçoso) (WU; KUMAR, 2009; MAIMON; ROKACH, 2010). E ainda, o conjunto de treinamento deve ser mantido armazenado para todo processo de classificação. Tudo isso se manifesta como desvantagem ao KNN, exigindo do mesmo, alto tempo no processo de classificação.

A maioria das aplicações envolvendo KNN utiliza distância Euclidiana como a medida de similaridade. No entanto, um problema com esta medida é tratar os vizinhos que são de baixa relevância com a mesma importância dos demais com relevância, o que pode degradar o desempenho de acurácia. Nesse aspecto é que emerge a possibilidade de explorar a informação exibida pelos vizinhos. Na seção seguinte apresenta-se uma variação deste algoritmo, o qual utiliza como métrica de similaridade a informatividade dos objetos em análise e os armazenados no conjunto de treinamento.

C. Algoritmo LI-KNN

Para entendimento do algoritmo, considere Q como o objeto em análise para classificação, K os vizinhos mais próximos a serem analisados de acordo com a distância Euclidiana e I denota a maioria dos pontos informativos mais próximos calculado a partir da equação (1). Para cada objeto $\mathbf{x}_i$ de um conjunto com N objetos de treinamento, contendo atributos de P dimensão e y_i uma das c classes possíveis para se atribuir ao objeto em análise, especifica-se que:

DEFINIÇÃO 1: um conjunto de objetos de treinamento $\{\mathbf{x}_i,\ y_i\}_1^N$ com $\mathbf{x}_i \in R^P$ e $y_i \in (1,\ \dots c)$, sendo c um número discreto de classes possíveis. Para cada objeto em análise $\mathbf{x}_j$, a informatividade de cada N objeto remanescente, $\{\mathbf{x}_i,\ y_i\}_1^N$ é definida por:

$$P(\mathbf{x}_j, Q = \mathbf{x}_i) = \ -log\left(1 - P(\mathbf{x}_j, Q = \mathbf{x}_i)\right) * (\mathbf{x}_j, Q = \mathbf{x}_i), j = 1, \dots N, j \neq i \quad (1)$$

onde $P(\mathbf{x}_j, Q = \mathbf{x}_i)$ é a probabilidade que o objeto $\mathbf{x}_j$ é informativo a Q, definido como:

$$P(\mathbf{x}_j, Q = \mathbf{x}_i) = \ \frac{1}{Z_i} Pr(\mathbf{x}_j, Q = \mathbf{x}_i)^{\eta} \left(\prod_{n=1}^{N} (1 - Pr(\mathbf{x}_j, Q = \mathbf{x}_i)\, I_{[\mathbf{x}_J \neq \mathbf{x}_n]})\right)^{1-\eta} \quad (2)$$

O primeiro termo da equação 2 $\left(Pr(\mathbf{x}_j, Q = \mathbf{x}_i)^{\eta}\right)$ pode ser interpretado como a probabilidade de $\mathbf{x}_j$ estar próximo à Q, enquanto a segunda parte indica a probabilidade de $\mathbf{x}_j$ estar longe de objetos diferentes. O indicador $I[.]$ equivale a 1 se a condição é atendida e 0 caso contrário. Z_i é um fator de normalização e η é introduzido como um fator de equilíbrio que determina a ênfase do primeiro termo. Intuitivamente, η está configurado como $\frac{N_{\mathbf{x}j}}{N}$, onde $N_{\mathbf{x}j}$ representa o número de pontos na mesma classe de $\mathbf{x}j$.

A lógica da informatividade é que dois objetos são suscetíveis a compartilhar o mesmo rótulo de classe quando sua distância é suficientemente pequena, assumindo que os objetos têm uma distribuição uniforme. Essa ideia é a mesma que a classificação KNN. Por outro lado, comparado com os classificadores tradicionais que medem as distâncias entre pares, objeto de consulta e os vizinhos, a métrica baseada em informatividade também calcula a proximidade entre os objetos vizinhos, ou seja, os objetos informativos também devem ter uma grande distância de objetos diferentes. Isso garante ainda que os locais de outros objetos informativos têm a mesma probabilidade de classificação da mesma classe.

A Figura 1 (à esquerda) fornece um exemplo ilustrativo do algoritmo LI-KNN, em que objeto 1 e o objeto 2 (com o mesmo rótulo de classe) ambos têm a mesma distância *d* para Q (objeto em análise), mas o ponto 1 é mais próximo do limite entre as classes (cores dos objetos). Assim, o objeto 1 é mais provável que pertença à classe preta. Por essa razão, afirmamos que o objeto 1 é menos informativo que o objeto 2 para Q pela DEFINIÇÃO 1. Mais uma vez,

assumindo que a distribuição dos objetos locais seja uniforme, é mais provável que os objetos 3 e 4 têm o mesmo rótulo dos objetos 1 e 2 e se distribuirá melhor em torno do objeto 2 em relação ao objeto 1.

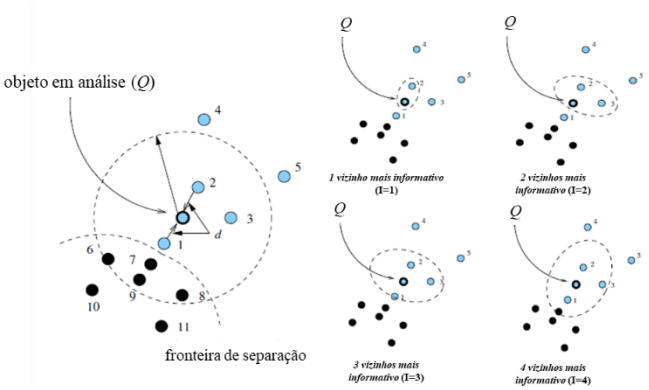

Fig. 1. Uma ilustração de 7-NN e os I objetos informativos correspondentes para o objeto de consulta (Q) A esquerda a classificação do 7-NN e a fronteira de decisão formada pelo conjunto de dados. A direita variação de I (1, 2, 3, 4) objetos informativos para a mesma consulta Q. Adaptado de (Song et al., 2014).

Algoritmo 1: LI-KNN

Entrada: **X** (conjunto de treinamento), *Q* (Objeto em análise), K (vizinhos mais próximos) e I (vizinhos informativos mais próximo)

Saída: Classe de Q

1 Para o objeto em análise Q Faça

2 Encontrar os K vizinhos mais próximos de Q com base em X usando a distância Euclidiana

3 Encontrar os I vizinhos informativos com base nos K objetos mais próximos

4 Decidir por votação majoritária dos vizinhos informativos a classe do exemplar em análise

O algoritmo LI-KNN, descrito no Algoritmo 1, classifica portanto cada objeto de consulta Q individualmente, aprendendo cada um dos objetos informativos separadamente. No entanto, a informatividade desses vizinhos é depois descartada sem ser utilizada para outros pontos de consulta. Na verdade, na maioria dos cenários, diferentes consultas a Q podem render diferentes pontos informativos. No entanto, é razoável esperar que alguns objetos sejam mais informativos do que outros, ou seja, eles poderiam ser vizinhos informativos para vários objetos diferentes. Como resultado, parece razoável colocar mais ênfase nos objetos que são globalmente informativos. Portanto, a proposta deste projeto é utilizar este algoritmo para analisar a sensibilidade de seus parâmetros, a fim de obter o melhor número informativo I e o

parâmetro K mais eficiente de modo a proporcionar o melhor desempenho em problemas de classificação.

Como se pode observar nas variáveis de entrada do Algoritmo 1, os parâmetros K e I devem ser definidos a priori. Song et al. (2014) não apresentam uma discussão na escolha destes parâmetros e justamente esta é a proposta deste trabalho. Ou seja, entender como estes parâmetros se comportam diante a diferentes cenários de dados (distribuição e classes), principalmente se há uma relação de proporcionalidade entre ambas.

3. METODOLOGIA

Para a implementação dos algoritmos será utilizada a linguagem de programação R e o seu ambiente de desenvolvimento RStudio. Além da realização dos experimentos de sensibilidade de parâmetros do algoritmo LI-KNN será feito um estudo comparativo com outros algoritmos: KNN, Random Forest e SVM.

Os conjuntos de dados serão gerados artificialmente, simulando diferentes tipos de distribuições de objetos em duas classes. Isso é importante para analisar a sensibilidade do algoritmo nestes tipos de dados. Adicionalmente, serão utilizados outros conjuntos de dados disponíveis em repositórios públicos como UCI dataset (DUA, 2019).

Após a caracterização do LI-KNN com os melhores parâmetros, o algoritmo deverá ser comparado com outros classificadores da literatura. Os algoritmos a serem comparados são o próprio KNN, Random Forest e SVM (MAIMON, 2010).

O desempenho do estudo comparativo será realizado com uso da metodologia de validação cruzada k-Fold e com apresentação dos resultados com a matriz de confusão (Silva e Del-Moral-Hernandez, 2011;Han, 2011). O indicador que será usado no estudo comparativo é a Acurácia de acertos do algoritmo.

Ressalta-se que os experimentos realizados envolvendo a geração de conjunto de dados artificiais foram previamente generalizados com o método de validação cruzada, no qual reproduziu-se primeiramente alguns exemplos da literatura para analisar os seus resultados, como a sobreposição de classes num conjunto de dados, onde se pode observar como as classes vão se sobrepondo à medida que suas médias vão diminuindo, até chegar a um ponto em que elas estão totalmente sobrepostas.

Um gráfico de linha estabelecendo uma comparação entre as acurácias de NN, Best KNN tem como base a medida F1 (SANCHEZ, 2007) que estabelece o melhor índice (1) quando a sobreposição entre as classes diminui e pior resultado (0) em maiores regiões de sobreposição de classes. (SANCHEZ, 2007); e este mesmo experimento é feito com KNN usando diversos valores do parâmetro K.

Após realizar esses experimentos da literatura com o KNN, foram efetuados testes para averiguar o comportamento do LI-KNN nessas situações. Assim, da mesma forma que no exemplo anterior com o intuito de realizar a parametrização do algoritmo, foi feito um gráfico de linha alternando entre os seus dois parâmetros principais K e I da seguinte forma: primeiro se variou o valor de K, mantendo-se o I fixo. Em um segundo caso, o inverso, variou o valor de I e manteve-se o valor de K fixo. Por fim, ainda com a base de dados artificial, a partir das melhores parametrizações do KNN e LI-KNN, experimentos comparativos com os algoritmos Random Forest e SVM foram realizados, a fim de verificar o comportamento dos mesmos em situações diferentes de sobreposição.

Para uma análise quantitativa, um segundo conjunto de experimentos é realizado com os resultados médios de acurácia sob dois aspectos, um com todas as bases de dados sintéticas e outro com base de dados públicas (DUA, 2019). Neste experimento além do resultado médio da acurácia, também se apresenta o tempo de processamento de cada algoritmo.

4. RESULTADO E DISCUSSÃO

Os experimentos serão apresentados de forma fragmentada. Primeiro se apresenta os experimentos para a base de dados artificial com ênfase em uma análise gráfica (qualitativa) e, depois com experimentos envolvendo também conjuntos de dados reais e uma análise de acurácia e tempo (quantitativa).

A. Análise quantitativa de parametrizações e comparações

Na Fig. 2 encontram-se exemplos dos conjuntos de dados gerados artificialmente, reproduzindo a metodologia proposta por Sanches e Sotoca (2007). Note que as duas classes iniciam com a média igual a 60, ou seja, totalmente sobrepostas uma à outra e a cada iteração se fez o decremento da média1 em 0.25 em relação a outra, até que as duas classes ficassem totalmente separadas, por exemplo com média1 = 55.25 e média2 = 60. Como pode-se observar, quanto mais a

diferença entre a média das duas classes aumenta, menos os dados se encontram sobrepostos.

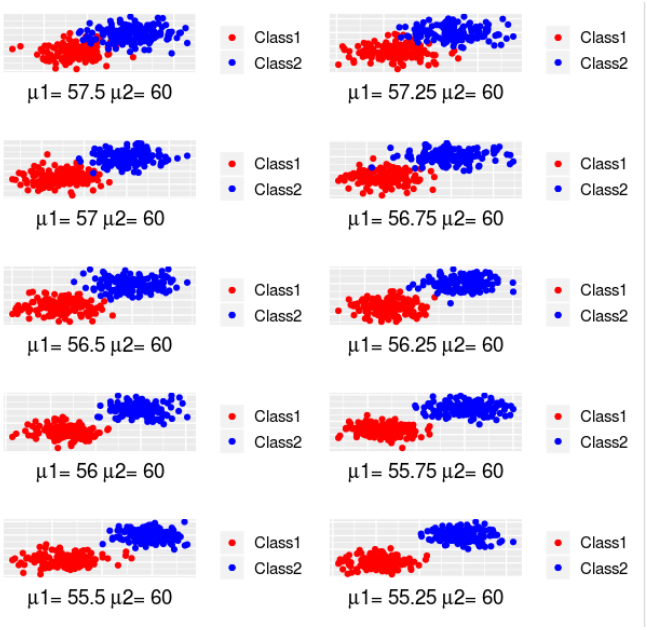

Fig2 – Exemplos dos conjuntos de dados gerados artificialmente.

De modo a verificar como a acurácia do KNN pode variar de acordo com a mudança de seu parâmetro K em termos de sobreposição de classes, a Figura 3 ilustra o desempenho do algoritmo para valores de K de 1 a 11, com apenas escolhas ímpares, para cada diferente nível de sobreposição.O resultado indica que a melhor parametrização é com o valor de K igual a 11.

Para confirmação da escolha, o experimento cujo resultado está ilustrado na Figura 4 foi realizar uma comparação entre as acurácias de 1-NN e do Best KNN (K = 11), segundo o método de validação cruzada k-Fold. Em complemento, o resultado é contrastado também com a medida de complexidade F1. Como era de se esperar, quanto mais os dados estão sobrepostos e a diferença entre a média das duas classes é baixa, menores serão os valores da acurácia; a medida que as classes vão se afastando, a medida F1 aumenta, assim como as acurácias. Neste resultado se pode ver a eficiência do Best.KNN em situações de sobreposição de classes.

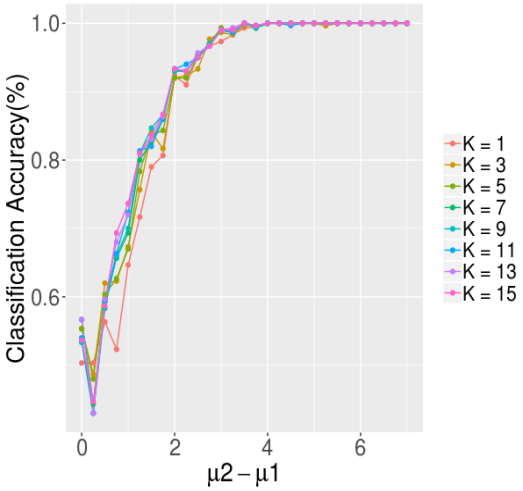

Fig3-Comparação das acurácias variando o valor de *K*.

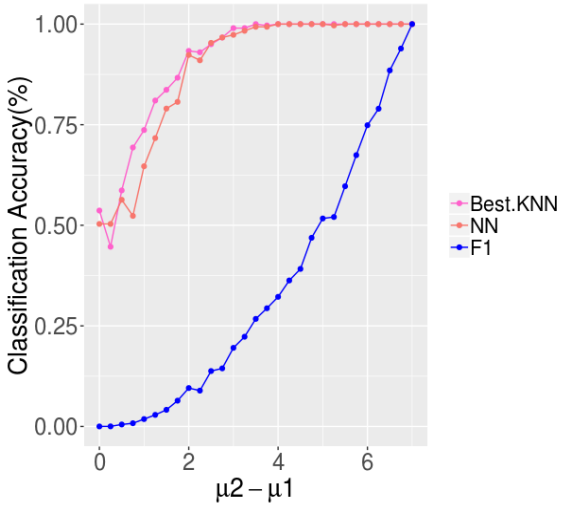

Fig4-Comparação das acurácias de NN e Best KNN com F1.

Analisando agora o algoritmo LI-KNN, o algoritmo tem os parâmetros K e I. Incialmente explora-se K com os valores 5, 10 e 15, mantendo-se constante o parâmetro I como sendo igual a 1. Nota-se, portanto deste experimento, com resultado ilustrado na Fig. 5 que a parametrização igual a 15 é o melhor resultado e nota-se, sobretudo, que o algoritmo é menos sensível a sobreposição, quando se compara com os resultados do KNN (Figura 3 e Figura 4).

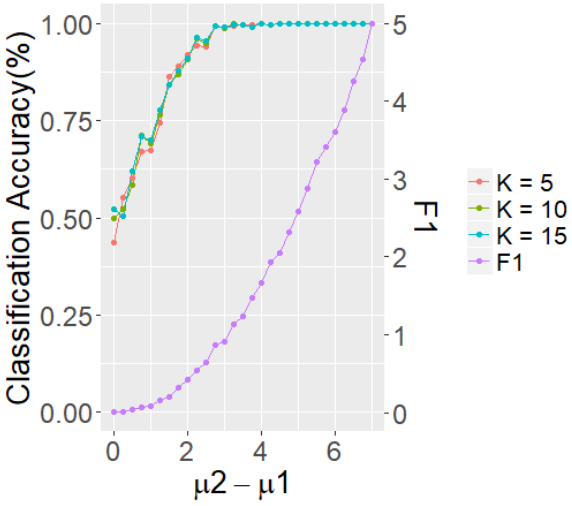

Fig5 - Algoritmo LI-KNN com *I* = 1 e *K* = 5,11,15.

Com a parametrização de *K* definida, agora se verifica o ajuste do parâmetro *I*. Na Figura 6 está o resultado da parametrização de *I* para os valores de 1 até 15, com valores ímpares. Novamente se nota a baixa sensibilidade do algoritmo em classes sobrepostas e a escolha do parâmetro será com o valor de 15, sendo a que quantitativamente apresenta melhor resultado.

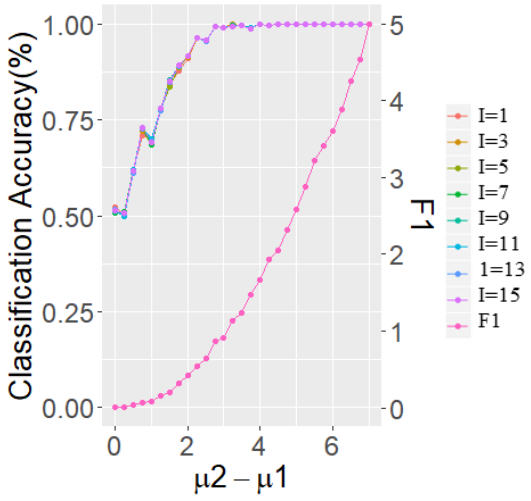

Fig6 - Algoritmo LI-KNN com *I* = 1,3,5,7,9,11,13,15 e 17 e *K* = 15

Pode-se concluir destes experimentos que em regiões mais sobrepostas o LI-KNN possui mehor desempenho que o KNN, enquanto que em regiões afastadas, os ganhos são

praticamente os mesmos, visto que após o parâmetro *K* ultrapassar o valor da sua informatividade no LI-KNN, o algoritmo passa a se comportar como se fosse um KNN.

Como última análise quantitativa e comparativa destes experimentos com bases de dados artificiais, os resultados das melhores parametrizações serão comparados com dois algoritmos da literatura, Random Forest e SVM. O resultado comparativo está ilustrado na Figura 7. Por esta se nota que os resultados são sobrepostos, o que é importante por perceber a eficiência do LI-KNN em relação a eficazes algoritmos da literatura (Random Forest e SVM). Para uma análise mais conclusiva, na seção seguinte são apresentadas análises em termos de média e desvio padrão como resultado, que podem ser melhor comparados.

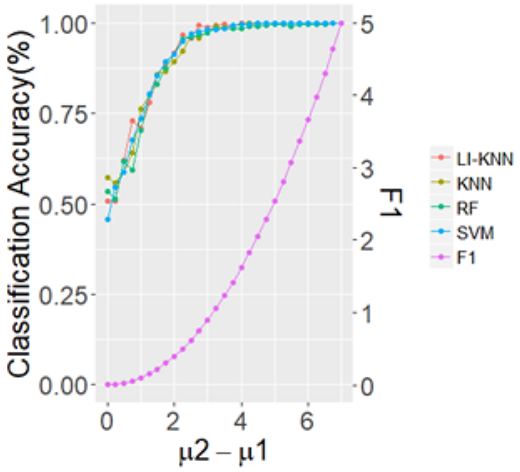

Fig*7* - Comparação entre os algoritmo*s*

B. Análise qualitativa e comparativa

O objetivo desta seção é apresentar análises qualitativas dos resultados feitos de forma comparativa. Além do uso da base de dados artificial, os experimentos também envolvem a comparação com 9 diferentes bases de dados da literatura e o desempenho dos algoritmos apresentados para comparação.

Os resultados da Tabela 1 indicam a comparação em termos de acurácia (média e desvio) e o tempo (treinamento com geração do modelo e teste na predição). Nota-se contudo, que o LI-KNN tem desempenho equivalente ao SVM, que é um algoritmo conhecido na literatura pela eficácia. Por outro lado, o tempo que se leva para a classificação com o LI-KNN é maior que os demais, incluindo para a soma dos tempos para a geração do modelo e a predição.

TABELA 1 COMPARAÇÃO DAS MÉTRICAS DOS ALGORITMOS KNN, LI-KNN, RANDOM FOREST E SVM.

Algoritmos	***Acuracia (Media)***	***Acuracia (Desvio padrão)***	***F1-Score (Media)***	***F1-Score (Desvio padrão)***	***Tempo (Modelo)***	***Tempo Predição)***
KNN	0.906	0.157	0.763	0.128	-	0.092
LI-KNN	***0.910***	0.152	0.773	0.129	-	0.179
Random Forest	0.905	0.148	0.903	0.152	0.008	0.002
SVM	***0.910***	0.147	0.905	0.158	0.002	0.008

TABELA 2 -COMPARAÇÃO ENTRE OS VALORES DE ACURÁCIA DAS BASES DO UCI DATASET,INDICANDO O MELHOR PARÂMETRO DE CADA ALGORITMO

	KNN			**LI-KNN**			**RF**			**SVM**		
	Acc	**Tempo**	**F1**	**Acc**	**Tempo**	**F1**	**Acc**	**Tempo**	**F1**	**Acc**	**Tempo**	**F1**
Iris	**0,973±0,069**	0,001	0,971	**0,973±0,059**	0,033	0,957	0,953±0,018	0,064	0,952	0,946±0,029	0,007	0,945
Hepatitis	0,793±0,084	0,001	0,314	0,748±0,321	0,059	0,344	**0,839±0,039**	0,098	0,500	0,813±0,042	0,008	0,440
Cancer	0,971±0,016	0,02	0,977	**0,972±0,017**	0,376	0,973	0,969±0,014	0,042	0,976	0,966±0,011	0,02	0,974
Glass	0,719±0,103	0,001	0,647	0,719±0,136	0,058	0,669	**0,776±0,047**	0,013	0,689	0,673±0,073	0,013	0,528
Haberman	0,765±0,084	0,002	0,850	**0,768±0,095**	0,097	0,834	0,735±0,069	0,022	0,834	0,742±0,084	0,015	0,839
Wisconsin	**0,973±0,020**	0,002	0,979	**0,973±0,024**	0,386	0,971	0,961±0,021	0,019	0,970	0,969±0,020	0,021	0,977
Mamograpy	0,813±0,042	0,003	0,824	0,537±0,025	0,621	0,698	0,813±0,027	0,037	0,825	**0,821±0,033**	0,071	0,833
Libra	**0,858±0,084**	0,007	0,858	**0,858±0,240**	0,295	0,858	0,794±0,034	0,093	0,786	0,767±0,033	0,09	0,751
Pima	0,747±0,050	0,007	0,814	0,651±0,055	0,295	0,788	**0,760±0,034**	0,093	0,823	0,755±0,019	0,09	0,822

Por fim, os resultados ilustrados na Tabela 2 ilustram a análise comparativa para as diferentes bases de dados reais. Vale ressaltar que a comparação foi feita com as melhores parametrizações encontradas anteriormente. Considerando como um tipo de comparação os resultados de média da acurácia, o LI-KNN apresenta melhor desempenho em 2 de 9 conjuntos de dados e tem empate em três casos. Isto indica o potencial uso do algoritmo e o efeito da parametrização. No entanto, o tempo de classificação do algoritmo é o maior de todas as comparações.

Além das comparações ilustradas nas Tabelas 1 e 2, foi realizado um teste de hipótese utilizando o método bilateral e a distribuição T de Student para verificar se a média real da acurácia de KNN não é igual a média do LI-KNN e assim provar que um resultado foi melhor que o outro.Utilizando os dados da Tabela 1 sobre a acurácia do KNN e do LI-KNN o resultado foi que a hipótese nula foi aceita e a hipótese

alternativa foi rejeitada, visto que a pontuação de z foi de -0,14 e portanto está dentro da região de não rejeição.0

5. CONSIDERAÇÕES FINAIS

O trabalho apresentou uma metodologia para avaliar algoritmos em situações de sobreposição de classes, utilizando-a para validar a escolha dos parâmetros do algoritmo LI-KNN. As melhores parametrizações foram colocadas em análise com outros algoritmos da literatura como o KNN, Random Forest e SVM.

Em complemento a metodologia, o trabalho também fez uso de conjunto de dados reais e amostragem com a metodologia k-Fold, possibilitando a partir dos experimentos realizados com bases de dados reais e sintéticas observar que em problemas com sobreposição de classe o LI-KNN possui um desempenho maior que os outros algoritmos observados no artigo e através das análises quantitativas e dos gráficos exemplificando os experimentos realizados foi definida que a sua melhor parametrização é K = 11 e I = 15.É importante salientar que ao contrário de Song et al. (2014), neste artigo foi apresentado um estudo reproduzível e metodológico a respeito dos parâmetros do algoritmo LI-KNN, demonstrando objetivamente o motivo de sua escolha.

Como principais resultados a serem destacados, o LI-KNN apresenta estabilidade nos resultados em regiões de sobreposição de classes frente aos demais comparados neste artigo. Além disso, em uma análise pela acurácia média, o algoritmo parametrizado conforme a metodologia apresentada, teve a maioria dos melhores resultados de acurácia.

Com base em todos os experimentos demonstrados nesse artigo, pretende-se futuramente substituir a opção do parâmetro K pela saída de um algoritmo de geração de protótipos como proposto por Moreira e Silva (2016). Isto ajuda o algoritmo a diminuir o tempo de classificação que ainda é um problema a ser superado. Além disso, pretende-se explorar outras medidas de sobreposição de classes em bases de dados reais para se ter uma melhor análise do ganho de uso do LI-KNN aos demais algoritmos de classificação de dados da literatura.

6. REFERÊNCIAS

AHA,D. W; KIBLER,D;ALBERT, M. K. Instance-Based Learning Algorithms. Machine Learning,6,pp.37-66,1991

COVER, T; HART, P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, Vol. 13, No.1, pp. 21-27, January 1967.

DUA, D. and GRAFF, C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.

FRIEDMAN, J.: Flexible metric nearest neighbor classification. technical report 113, Stanford university statistics department (1994)

HAN, J. and KAMBER, M. Data mining: concepts and techniques (Elsevier, 2011)

JAIN, A. L; MAO, J; MOHIUDDIN, L. M. Artificial Neural Networks: A tutorial. IEEE Transactions on Neural Networks, pp.31-44,1996

LIMA, ANA CAROLINA E.S; de Castro, L. N; CORCHADO, J. M. A polarity analysis framework for Twitter messages. Applied Mathematics and Computation, v. 270, p. 756-767, 2015.

MAIMON, O; ROKACH, L. Data Mining and Discovery Knowledge Handbook. 2nd edition, Springer, 2010.

MOREIRA,L; SILVA,L. Data Classification combining Self-Organizing Maps and Informative Nearest Neighbor. Internationational Joint Conference on Neural Network, pp. 1-10,2016

SANCHEZ, J.S., MOLLINEDA, R.A., SOTOCA, J.M.: An analysis of how training data complexity affects the nearest neighbor classifiers. Pattern Analysis and Applications10(3),189–201 (2007).

SANTOS, J. B; ALMEIDA JR, J. R; SILVA, L. A. Pattern Recognition in Mammographic Images Used by the Residents in Mammography. Proceedings of IEEE International Conference on Computer Medical Applications, pp. 1-6, 2013.

SILVA L. A., DEL-MORAL-HERNANDEZ E. A SOM combined with KNN for classification task. Proceedings of the 2011 International Joint Conference on Neural Network, IJCNN 2011; August 2011; San Jose, Calif, USA. pp. 2368–2373.

SILVA, L. A; PERES, S. M; BOSCARIOLI, C. Introdução à Mineração de Dados - Com aplicações em R. Elsevier, 2016.

SONG Y., HUANG J., ZHOU D., ZHA H., GILES C. L. IKNN: Informative K-Nearest Neighbor Pattern Classification. Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery; 2007; Berlin, Germany. Springer; pp. 248–264.

WEINBERGER, K.Q., BLITZER, J., SAUL, L.K.: Distance metric learning for large margin nearest neighbor classification. In: NIPS (2005)

WETTSCHERECK, D. A Study of Distance-Based Machine Learning Algorithms. Doctor of Philosophy Dissertation. Oregon State University, 1994.

WU, X; KUMAR, V. The Top Ten Algorithms in Data Mining. Chapman and Hall/CRC Data Mining and Knowledge Discovery Series, April 2009.

WU, X. and KUMAR, V., The top ten algorithms in data mining (CRC Press, 2009)

Contatos: vinygp8@gmail.com e leandroaugusto.silva@mackenzie.br