Aprendizado de Métrica por Silhueta para Classificador por Arestas de Suporte

José Geraldo Fernandes
Escola de Engenharia
Universidade Federal de Minas Gerais
Belo Horizonte, Brasil
josegeraldof@ufmg.br

***Resumo*—Este trabalho analisa uma abordagem de aprendizado de métrica a partir da distância de Mahalanobis para minimização da silhueta média do conjunto de dados. Para avaliação objetiva, observa-se o efeito desse procedimento em um Classificador por Arestas de Suporte (CLAS), a partir da acurácia do modelo e o descarte no processo de filtragem como um indicador da sobreposição no conjunto de treinamento.**

***Palavras-chaves*—aprendizado de métrica, métrica de distância, grafo de gabriel, silhueta, mahalanobis**

I. Introdução

A. Grafo de Gabriel

O Grafo de Gabriel [1] é uma construção a partir de um conjunto de pontos $\mathcal{S}$, que define os vértices do grafo, para outro conjunto de arestas $\mathcal{E}$ tal que dois pontos $\boldsymbol{x_i}, \boldsymbol{x_j}$ são adjacentes, definem uma aresta, se não há outro ponto de $\mathcal{S}$ dentro da hiperesfera definida com a distância entre os dois pontos $\boldsymbol{x_i}, \boldsymbol{x_j}$ como diâmetro, como na Equação 1.

$$\begin{gathered}(\boldsymbol{x_i}, \boldsymbol{x_j}) \in \boldsymbol{\mathcal{E}} \leftrightarrow \\ \delta^2(\boldsymbol{x_i}, \boldsymbol{x_j}) \leq [\delta^2(\boldsymbol{x_i}, \boldsymbol{x_k}) + \delta^2(\boldsymbol{x_j}, \boldsymbol{x_k})] \ \forall \boldsymbol{x_k} \in \mathcal{S}\end{gathered} \tag{1}$$

Onde δ é a métrica de distância na construção e comumente definida como a distância euclidiana.

A construção desse grafo é uma técnica da Geometria Computacional e emprestada para os problemas de aprendizado como uma forma de expressar as características de vizinhança da estrutura do conjunto de dados. Note como a definição da métrica de distância ótima pode representar um ganho nessa representação quando a distribuição de classes, em um problema de classificação, é melhor mapeada e discriminada em uma métrica que outra [2].

B. Métricas de Distância

As métricas de distâncias são parte essencial de muitas técnicas de aprendizado de máquina. É proveitoso representar os dados em um espaço tal que amostras similares conjugam uma distância menor que para outras amostras semanticamente díspares. A própria noção de semelhança e disparidade depende também da aplicação, em um mesmo espaço algumas características podem ter relevância superior a depender do contexto.

Considere, por exemplo, um conjunto de dados composto por amostras de fala, dada uma extração de características acústicas. Nesse cenário, dependendo se o interesse é classificar as amostras pela idade do falante, seu gênero ou até identificar aspectos mais culturais como o sotaque ou a emoção do discurso, os atributos extraídos terão relevância diferente no problema.

Uma abordagem para contornar esse obstáculo é determinar a priori o conjunto apropriado de características da extração a partir de conhecimentos específicos do problema. Contudo, esse processo é custoso e perde em generalização, pode-se, portanto, considerar o aprendizado de métrica uma forma de automatizar essa etapa para qualquer problema.

Sobre as métricas de distância, a mais popular na literatura de aprendizado de métrica é a distância Mahalanobis [3]. Proposta como uma medida de distância entre um ponto e uma distribuição, como na Equação 2, para representar o quão distante a observação está do conjunto nos eixos das componentes principais, levando em consideração, portanto, a covariância entre os atributos.

$$d_M^2(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \tag{2}$$

Onde $\boldsymbol{x}$ é a observação, $\boldsymbol{\mu}$ e Σ é a média e a matriz de covariância da distribuição. Todavia, é comum aproveitar essa equação como uma métrica de distância usando a matriz como um parâmetro geral a ser otimizado, como na Equação 3.

$$\delta(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{(\boldsymbol{p} - \boldsymbol{q})^T \boldsymbol{M} (\boldsymbol{p} - \boldsymbol{q})} \tag{3}$$

Note que pode-se considerar essa formulação como uma expansão da distância euclideana, em que $\boldsymbol{M} = \boldsymbol{I}$, e, também, decompor a matriz $\boldsymbol{M}$ como $\boldsymbol{L}^T \boldsymbol{L}$. Nesse formato, o problema se resume a encontrar a distância euclideana em um novo conjunto de dados transformados $\boldsymbol{Lx}$, como na Equação 4.

$$\begin{aligned}\delta(\boldsymbol{p}, \boldsymbol{q}) &= \sqrt{(\boldsymbol{p} - \boldsymbol{q})^T \boldsymbol{M} (\boldsymbol{p} - \boldsymbol{q})} \\ &= \sqrt{(\boldsymbol{p} - \boldsymbol{q})^T \boldsymbol{L}^T \boldsymbol{L} (\boldsymbol{p} - \boldsymbol{q})} \\ &= \sqrt{(\boldsymbol{Lp} - \boldsymbol{Lq})^T (\boldsymbol{Lp} - \boldsymbol{Lq})}\end{aligned} \tag{4}$$

Seja qual for a métrica selecionada, é importante regular uma função de custo que, comumente [4], avalia as seguintes restrições:

$$\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) : \boldsymbol{x}_i \, e \, \boldsymbol{x}_j \text{ devem ser próximos}\}$$

$$\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) : \boldsymbol{x}_i \, e \, \boldsymbol{x}_j \text{ devem ser distantes}\}$$

$$\mathcal{R} = \{(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_k) : \boldsymbol{x}_i \text{ deve ser mais próximo de } \boldsymbol{x}_j \text{ que de } \boldsymbol{x}_k\}$$

Assim, o problema de otimização é descrito como na Equação 5.

$$\min_{M} \, l(M, \mathcal{S}, \mathcal{D}, \mathcal{R}) + \lambda R(M) \tag{5}$$

Onde M é o parâmetro a ser otimizado, l é uma função de custo, R um regularizador e $\lambda \geq 0$ seu parâmetro.

C. Classificador por Arestas de Suporte

Construido o grafo no conjunto de treinamento pode-se aproveitar essa informação estrutural para problemas de classificação. Os Classificadores por Arestas de Suporte (CLAS) [5], por sua vez, são um método de classificação de margem larga que aproveitam do grafo. As Arestas de Suporte são as arestas do grafo que discriminam amostras de classes distintas, vetores geométricos de suporte, semelhante aos vetores de suporte em *Support-vector machines* (SVM) [6].

Dessas arestas e seus pontos médios constrói-se hiperplanos de separação, que são otimizados para maximizar a margem entre as classes. O classificador, então, pondera sobre a resposta de cada hiperplano para rotular a amostra.

Há, no entanto, um revés dessa abordagem ao fenômeno de sobreposição de classes no conjunto de dados. Note que dado esse fenômeno podem existir arestas de suporte distantes da margem de separação. Esse problema é tratado excluindo amostras do treinamento ponderado por um fator de qualidade que depende dos outros pontos que compartilham um vértice. Sujeito ao tamanho da sobreposição, essa exclusão pode prejudicar o desempenho do modelo, contudo é razoável assumir que um mapeamento ótimo do Grafo de Gabriel pode atenuar o corte de amostras [7].

II. Revisão Bibliográfica

O aprendizado de métrica é um processo importante para a utilização de classificadores baseados em distância. Além disso, considerando essa abordagem como um forma de mapeamento otimizado dos dados, é considerável também sua utilidade para redução de dimensionalidade e *clustering*.

Uma abordagem comum, no sentido do mapeamento, é otimizar uma matriz de transformação $\boldsymbol{M}$ como na Equação 3 e formular uma função de custo como na Equação 5, basta, portanto, resolver o problema de otimização.

Em *Local Fisher Discriminant Analysis* [8] os autores combinam duas funções de agrupamento, considerando aspectos globais e locais, na função de custo, adaptando duas técnicas de aprendizado de métrica, *Fisher Discriminant Analysis* e *lLcal-Preserving Projection*.

Já em *Large Margin Nearest Neighbors* (LMNN) [9] utiliza-se uma abordagem de k-vizinhos mais próximos para combinar um incentivo de aproximação de agrupamentos coincidentes e outro de afastamento para dissidentes, modelado, também, na função de custo do problema de otimização da matriz de transformação.

Finalmente, em [7] propõe-se uma modificação no método LMNN para incluir as adjacências do Grafo de Gabriel no algoritmo de definição de vizinhos.

III. Metodologia

A. Base de Dados

Para avaliação do método de aprendizado de métrica seleciona-se um conjunto de 15 *datasets* padrão do repositório UCI [10] em todos os testes. Esses são separados em 10 partições determinadas para validação cruzada *k-fold* [11].

Para confiabilidade do resultado, as bases de dados selecionadas são amplamente utilizadas em aplicações semelhantes. Como pré-processamento, fez-se uma normalização de todos os atributos e conversão dos categóricos em númericos, necessário para o algoritmo CLAS. Todos são de problemas de classificação binária. Segue a descrição da seleção: *Statlog Australian Credit Approval* (australian); *Banknote Authentication* (banknote); *Breast Cancer Wisconsin* (breastcancer); *Breast Cancer Hess Probes* (breastHess); *Liver Disorders* (bupa); *Climate Model Simulation Crashes* (climate); *Pima Indian Diabetes* (diabetes); *Fertility* (fertility); *Statlog German Credit Data* (german); *Gene Expression* (golub); *Haberman's Survival* (haberman); *Statlog Heart Dicease* (heart); *Indian Liver Patient* (ILPD); *Parkinsons* (parkinsons); *Connectionist Bench Sonar, Mines vs. Rocks* (sonar).

A Tabela I mostra as principais características dessa seleção, η representa a proporção entre as classes. Note a alta diversidade dos problemas para atestar a generalização do método.

Tabela I
Características das bases de dados selecionadas.

Dataset	**Amostras**	**Atributos**	$\boldsymbol{\eta}$
australian	690	14	0.44
banknote	1372	4	0.44
breastcancer	683	6	0.65
breastHess	133	30	0.74
bupa	345	6	0.42
climate	540	18	0.91
diabetes	768	8	0.65
fertility	100	9	0.12
german	1000	24	0.70
golub	72	50	0.65
haberman	306	3	0.74
heart	270	13	0.56
ILPD	579	10	0.72
parkinsons	195	22	0.75
sonar	208	60	0.47

B. Aprendizado de Métrica

Em seguida, para o processo de aprendizado de métrica, calcula-se a matriz de transformação $\boldsymbol{M}$ no conjunto de treinamento para cada *fold*. Essa matriz será aplicada na transformação da base de dados como na Equação 4. Dessa forma, o problema de classificação segue sua abordagem tradicional com o *dataset* modificado.

Para o problema de otimização utilizou-se a silhueta [12] média do conjunto de dados. A silhueta é um método de validação de agrupamentos, calcula-se uma relação entre o quão próximo estão as amostras de um *cluster* e o quão distantes estão essas de *clusters* estranhos, para cada amostra. Dessa forma, a acurácia média é um parâmetro utilizado para avaliar a coesão do conjunto.

O problema de otimização é caracterizado como na Equação 6, utilizando a silhueta média como função objetivo e mantendo a restrição da matriz de transformação $\boldsymbol{M}$ como positiva definida.

$$\begin{aligned} \boldsymbol{L}^* = \arg\max_{\boldsymbol{L}} \quad & \text{silh}(\boldsymbol{LX},\ \boldsymbol{Y}) \\ \text{sujeito a} \quad & \boldsymbol{L}^T\boldsymbol{L} \succeq 0 \end{aligned} \tag{6}$$

C. Classificação

O problema de classificação segue o padronizado. Para cada *dataset* separou-se 10 *folds* fixos, desses aplicou-se o método de aprendizado de métrica no conjunto de treinamento para conseguir o *dataset* modificado $\boldsymbol{LX}$, também avaliou-se o *dataset* sem modificação. Aplicou-se o algoritmo CLAS para classificação. Avaliou-se a performance com a área sob a curva ROC (AUC) com validação cruzada nos *folds*, a silhueta média no conjunto de dados obtido e o descarte de dados no processo de filtragem como um indicador de sobreposição das classes.

IV. Resultados

A Tabela II mostra os resultados obtidos com a AUC média de todos os conjunto de dados e sua modificação aprendida.

Tabela II
DESEMPENHO, AUC MÉDIA, PARA CADA BASE DE DADOS.

Dataset	**Base**	**Modificado**
australian	0.85 ± 0.04	0.86 ± 0.05
banknote	0.99 ± 0.01	0.88 ± 0.19
breastcancer	0.96 ± 0.03	0.98 ± 0.01
breastHess	0.81 ± 0.12	0.8 ± 0.1
bupa	0.63 ± 0.1	0.64 ± 0.08
climate	0.84 ± 0.07	0.73 ± 0.16
diabetes	0.72 ± 0.04	0.72 ± 0.07
fertility	0.59 ± 0.26	0.51 ± 0.11
german	0.67 ± 0.04	0.67 ± 0.06
golub	0.77 ± 0.17	0.77 ± 0.15
haberman	0.56 ± 0.09	0.63 ± 0.11
heart	0.8 ± 0.08	0.8 ± 0.07
ILPD	0.57 ± 0.09	0.62 ± 0.07
parkinsons	0.9 ± 0.15	0.81 ± 0.13
sonar	0.88 ± 0.08	0.77 ± 0.1

A comparação entre a silhueta média, do conjunto de dados completo, e a proporção de descarte de dados pelo método de filtragem do algoritmo CLAS é como nas Tabelas III e IV.

V. Discussões

A partir dos resultados de desempenho da métrica aprendida é transparente que a performance foi equivalente ou pior que o conjunto de dados padrão. Além disso, utilizou-se um algoritmo de otimização comum para solucionar o problema da Equação 6 com alto custo computacional.

Tabela III
SILHUETA MÉDIA DO CONJUNTO DE DADOS COMPLETO.

Dataset	**Base**	**Modificado**
australian	0.18	0.48 ± 0.01
banknote	0.24	0.68 ± 0.06
breastcancer	0.57	0.76 ± 0.0
breastHess	0.08	0.34 ± 0.01
bupa	0.0	0.04 ± 0.01
climate	0.03	0.32 ± 0.01
diabetes	0.1	0.22 ± 0.01
fertility	0.01	0.19 ± 0.01
german	0.03	0.14 ± 0.0
golub	0.14	0.33 ± 0.01
haberman	0.01	0.09 ± 0.05
heart	0.13	0.37 ± 0.0
ILPD	0.02	0.06 ± 0.01
parkinsons	0.12	0.3 ± 0.01
sonar	0.03	0.21 ± 0.02

Tabela IV
PROPORÇÃO DE DESCARTE NO PROCESSO DE FILTRAGEM.

Dataset	**Base**	**Modificado**
australian	0.38 ± 0.02	0.34 ± 0.01
banknote	0.0 ± 0.0	0.03 ± 0.02
breastcancer	0.15 ± 0.01	0.09 ± 0.01
breastHess	0.39 ± 0.02	0.35 ± 0.03
bupa	0.48 ± 0.02	0.49 ± 0.03
climate	0.4 ± 0.14	0.16 ± 0.02
diabetes	0.45 ± 0.01	0.44 ± 0.02
fertility	0.43 ± 0.04	0.3 ± 0.03
german	0.47 ± 0.01	0.43 ± 0.01
golub	0.37 ± 0.03	0.42 ± 0.03
haberman	0.46 ± 0.02	0.49 ± 0.03
heart	0.43 ± 0.02	0.39 ± 0.01
ILPD	0.48 ± 0.01	0.48 ± 0.02
parkinsons	0.1 ± 0.17	0.29 ± 0.01
sonar	0.0 ± 0.0	0.38 ± 0.13

Apesar disso, aconteceu o ganho consistente, como esperado, da silhueta média do conjunto. Esperava-se que um conjunto com essa pontuação alta mostrasse baixa sobreposição de classes na região de fronteira e, portanto, um ganho na acurácia do classificador.

Note, contudo, que a proporção de descarte no processo de filtragem do modelo CLAS não seguiu a mesma tendência. Pensando nessa medida como, também, um indicador de sobreposição, coloca-se em dúvida o *score* da silhueta média como função objetivo adequada no processo de otimização.

As Figuras 1, 1 e 3 mostram a distribuição da diferença dessas três medidas do conjunto modificado e base. Não há relação clara a partir de uma inspeção visual, em especial da proporção de descarte e silhueta média, uma associação esperada.

VI. Conclusões

Este trabalho apresentou uma abordagem de aprendizado de métrica com maximizando a silhueta média no problema de otimização. Mostrou-se uma fragilidade dessa função para esse propósito e uma dissociação inesperada com outra medida, proporção de descarte de dados do modelo CLAS.

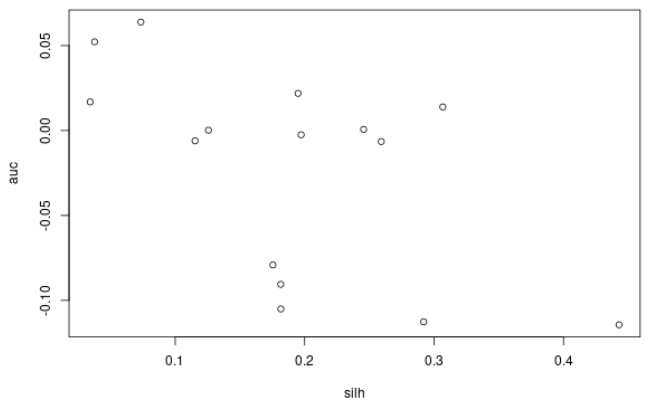

Figura 1. Distribuição da diferença de AUC média e de silhueta média do conjunto modificado e base.

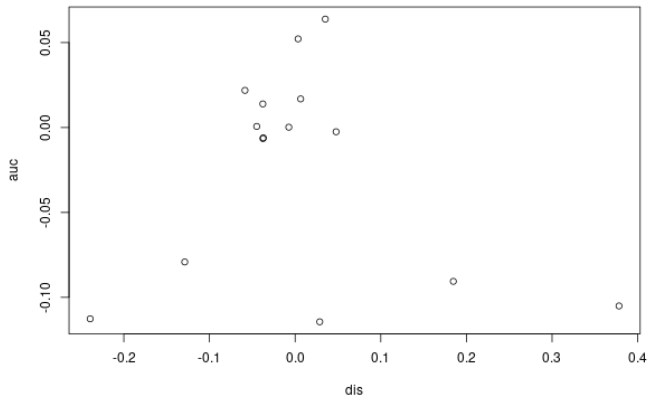

Figura 2. Distribuição da diferença de AUC média e de proporção de descarte do conjunto modificado e base.

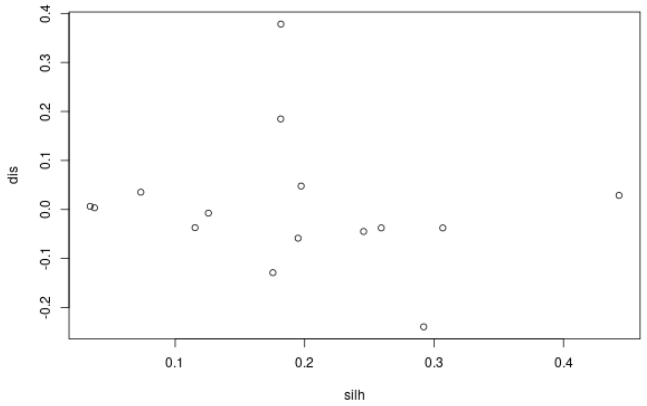

Figura 3. Distribuição da diferença de proporção de descarte e de silhueta média do conjunto modificado e base.

Uma hipótese levantada para explicar esse fenômeno é um possível comportamento multi-modal de classes. Repare que a silhueta tolera pouco esse caso já que amostras de agrupamentos espaciais diferentes são tratados como um só, afinal, já que nesse caso o aprendizado de métrica é supervisionado, são da mesma classe.

Outra questão é considerar a silhueta média do conjunto inteiro. Uma possível abordagem mais vantajosa seria considerar a média das amostras com menor *score* ou dos vetores geométricos de suporte, assim a otimização ataca principalmente a região de fronteira, mais sensível no desempenho do classificador.

AGRADECIMENTO

Este trabalho foi possível pela disponibilização das bases de dados pelo repositório *UCI Machine Learning* [10].

REFERÊNCIAS

[1] K. R. Gabriel and R. R. Sokal, "A new statistical approach to geographic variation analysis," *Systematic zoology*, vol. 18, no. 3, pp. 259–278, 1969.

[2] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, no. 2, p. 4, 2006.

[3] P. C. Mahalanobis, "On the generalized distance in statistics," National Institute of Science of India, 1936.

[4] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv preprint arXiv:1306.6709*, 2013.

[5] L. C. B. Torres, "Classificador por arestas de suporte (clas): Métodos de aprendizado baseados em grafos de gabriel," 2016.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[7] I. P. Gomes, L. C. B. Torres, and A. de Pádua Braga, "Aprendizado de métrica supervisionado para classificador por arestas de suporte,"

[8] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis.," *Journal of machine learning research*, vol. 8, no. 5, 2007.

[9] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification.," *Journal of machine learning research*, vol. 10, no. 2, 2009.

[10] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[11] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, pp. 1137–1145, Montreal, Canada, 1995.

[12] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.