

Índices de Qualidade de Agrupamento baseados no Grafo de Gabriel

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

Resumo—Este trabalho analisa a relação entre índice de qualidade de agrupamentos e o desempenho de classificadores em problemas supervisionados. Propõe-se indicadores baseados no Grafo de Gabriel e testa-se sua validade a partir de um modelo de regressão linear. Todo o código desenvolvido está disponível em repositório Git.

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 $\mathbf{x}_i, \mathbf{x}_j$ são adjacentes, definem uma aresta, se não há outro ponto de \mathcal{S} dentro da hipersfera definida com a distância entre os dois pontos $\mathbf{x}_i, \mathbf{x}_j$ como diâmetro, como na Equação 1.

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

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.

Uma formulação geral para esse modelo, que explicita as relações de interesse no espaço, é como a função de custo que, comumente [3], avalia as seguintes restrições:

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

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

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

Assim, o problema de otimização é descrito como na Equação 2.

$$\min_M l(M, \mathcal{S}, \mathcal{D}, \mathcal{R}) + \lambda R(M) \quad (2)$$

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

C. Filtro de Sobreposição

Os filtros de sobreposição baseados no Grafo de Gabriel foram adotados como uma forma de realizar regularização no Classificador por Arestas de Suporte (CLAS) [4], [5].

Baseado na diferença de classes entre o vértice e as amostras conectadas por suas arestas, define-se uma grandeza Q que representa a qualidade dessa amostra como na Equação 3, onde V representa o número de arestas e V_{eq} o número dessas para amostras de classe coincidente.

$$Q(\mathbf{x}_i) = \frac{V_{eq}(\mathbf{x}_i)}{V(\mathbf{x}_i)} \quad (3)$$

Amostras na região de fronteira e sobreposição terão sua qualidade afetada quão maior for a mistura. Segue, portanto, um filtro simples de amostras com qualidade inferior a um limiar. Adota-se um limiar dinâmico TC que representa a qualidade média de amostras de classe específica.

Neste trabalho, contudo, é interesse utilizar esses indicadores, qualidade do vértice (Q) e proporção apontada pelo filtro (D_i) como uma forma de avaliar o espaço de um conjunto de dados e um novo espaço aprendido com método de aprendizado de métrica.

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*.

Em *Local Fisher Discriminant Analysis* [6] 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 *LLocal-Preserving Projection*.

Já em *Large Margin Nearest Neighbors* (LMNN) [7] 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.

Para avaliação dessas técnicas, é comum a análise comparativa em função do desempenho na classificação [8]–[10]. Apesar dessa abordagem ser direta ao interesse do problema, é importante considerar um número grande e diverso de conjunto de dados para eliminar um viés de objetivo.

Muitas vezes, a higiene do espaço de representação de dados é tão importante quanto o resultado final do preditor. Nesse sentido, sugere-se índices de qualidade de *clustering* para avaliar essa perspectiva.

Calinski-Harabasz (CH) [11] é um índice que quantifica o quanto as amostras estão distantes dos centroides dos agrupamentos, a variável de coesão, em relação a distância dessas ao centroide global, a variável de separabilidade. Davies-Bouldin (DB) [12] é semelhante em sua formulação, mas difere-se na variável de separabilidade onde computa uma função das distâncias entre centroides. Finalmente, a *Silhouette* (Si) [13] computa a coesão a partir da soma das distâncias de todos os pontos de agrupamento coincidente, enquanto, a separabilidade, a mesma soma em relação ao agrupamento estranho mais próximo.

III. METODOLOGIA

A. Base de Dados

Para avaliação dos índices de qualidade dos agrupamentos seleciona-se um conjunto de 15 *datasets* padrão do repositório UCI [14] em todos os testes. Esses são separados em 10 partições determinadas para validação cruzada k -fold [15].

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 numéricos, 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 Disease* (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	η
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. Classificação

O problema de classificação segue o padronizado. Para cada *dataset* separou-se 10 *folds* fixos, desses aplicou-se os métodos de aprendizado de métrica, LMNN e LFDA, no conjunto de treinamento para conseguir o *dataset* no novo espaço, também avaliou-se o *dataset* sem modificação.

Aplicou-se o algoritmo CLAS e uma SVM com *kernel* RBF para classificação. Avaliou-se a performance com a área sob a curva ROC (AUC) com validação cruzada nos *folds*.

C. Avaliação dos Índices

Para avaliar o espaço aprendido pelos métodos de *metric learning*, utilizou-se os índices de qualidade de agrupamentos: Calinski-Harabasz; Davies-Bouldin; e, a silhueta média. Também avaliou-se os índices propostos: a qualidade média dos vértices; e, a taxa de filtragem de dados por sobreposição. Esses foram utilizados sobre o conjunto de dados completo.

Em seguida, aplicou-se um modelo de regressão linear simples para mensurar a capacidade de previsão que os índices têm da acurácia do classificador.

IV. RESULTADOS

A. Classificação

As Tabelas II e III mostram os resultados obtidos com a AUC média de todos os conjunto de dados e suas modificações aprendidas por ambos classificadores.

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

<i>Dataset</i>	Base	LMNN	LFDA
australian	0.85 ± 0.04	0.86 ± 0.04	0.86 ± 0.03
banknote	0.99 ± 0.01	1.00 ± 0.00	0.97 ± 0.03
breastcancer	0.96 ± 0.03	0.95 ± 0.04	0.97 ± 0.02
breastHess	0.81 ± 0.12	0.79 ± 0.14	0.74 ± 0.11
bupa	0.63 ± 0.10	0.58 ± 0.05	0.72 ± 0.08
climate	0.84 ± 0.07	0.87 ± 0.08	0.88 ± 0.11
diabetes	0.72 ± 0.04	0.72 ± 0.04	0.73 ± 0.06
fertility	0.59 ± 0.26	0.58 ± 0.23	0.59 ± 0.19
german	0.67 ± 0.04	0.69 ± 0.04	0.70 ± 0.05
golub	0.77 ± 0.17	0.68 ± 0.23	0.59 ± 0.17
haberman	0.56 ± 0.09	0.58 ± 0.09	0.57 ± 0.07
heart	0.80 ± 0.08	0.79 ± 0.08	0.84 ± 0.06
ILPD	0.57 ± 0.09	0.57 ± 0.10	0.65 ± 0.07
parkinsons	0.90 ± 0.15	0.91 ± 0.13	0.79 ± 0.11
sonar	0.88 ± 0.08	0.85 ± 0.08	0.76 ± 0.13

Tabela III
DESEMPENHO OBTIDO PELA SVM, AUC MÉDIA, PARA CADA BASE DE DADOS.

<i>Dataset</i>	Base	LMNN	LFDA
australian	0.86 ± 0.04	0.86 ± 0.04	0.86 ± 0.04
banknote	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.01
breastcancer	0.97 ± 0.01	0.97 ± 0.02	0.97 ± 0.01
breastHess	0.78 ± 0.10	0.78 ± 0.12	0.74 ± 0.12
bupa	0.67 ± 0.07	0.67 ± 0.07	0.70 ± 0.04
climate	0.58 ± 0.09	0.68 ± 0.11	0.73 ± 0.11
diabetes	0.71 ± 0.05	0.70 ± 0.05	0.72 ± 0.04
fertility	0.50 ± 0.00	0.50 ± 0.00	0.50 ± 0.00
german	0.67 ± 0.05	0.68 ± 0.04	0.69 ± 0.05
golub	0.82 ± 0.17	0.54 ± 0.12	0.48 ± 0.16
haberman	0.51 ± 0.04	0.50 ± 0.03	0.55 ± 0.07
heart	0.80 ± 0.06	0.82 ± 0.07	0.81 ± 0.06
ILPD	0.50 ± 0.01	0.50 ± 0.01	0.50 ± 0.01
parkinsons	0.81 ± 0.11	0.83 ± 0.12	0.77 ± 0.14
sonar	0.84 ± 0.09	0.87 ± 0.06	0.81 ± 0.09

B. Índices de Qualidade

As Tabelas IV, V e VI mostram os índices de qualidades de agrupamentos calculados com todo o conjunto de dados no formato tradicional e suas modificações.

Dos índices propostos, as Tabelas VII e VIII mostram o resultado encontrado análogo.

C. Regressão Linear

Para cada classificador ajustou-se um modelo de regressão linear dos índices de qualidade para o resultado do desempenho desse. Registrou-se o tamanho do coeficiente, β , e o valor-p, P , da estimativa para cada atributo. As Tabelas IX e X mostram os resultados obtidos.

Dos índices propostos, as Tabelas XI e XII mostram o resultado encontrado análogo.

V. DISCUSSÕES

Como esperado, os métodos de aprendizado de métrica geraram resultados de desempenho ligeiramente melhores. Esperava-se, portanto, uma conclusão mais definitiva a partir

Tabela IV
ÍNDICE DE CALINSKI-HARABASZ DO CONJUNTO DE DADOS COMPLETO.

<i>Dataset</i>	Base	LMNN	LFDA
australian	153.17	136.52	272.60
banknote	439.02	724.80	6286.82
breastcancer	933.04	1374.63	2413.80
breastHess	19.91	36.31	11.06
bupa	3.52	3.90	8.00
climate	8.60	17.56	16.83
diabetes	55.10	52.56	67.75
fertility	1.38	1.51	2.18
german	23.72	26.16	19.98
golub	12.97	24.38	14.28
haberman	1.74	1.52	10.32
heart	39.51	44.45	62.30
ILPD	8.69	12.85	5.74
parkinsons	31.09	42.01	23.18
sonar	7.17	20.36	14.29

Tabela V
ÍNDICE DE DAVIES-BOULDIN DO CONJUNTO DE DADOS COMPLETO.

<i>Dataset</i>	Base	LMNN	LFDA
australian	2.06	2.18	1.47
banknote	1.61	1.25	0.39
breastcancer	0.79	0.64	0.46
breastHess	1.93	1.47	2.46
bupa	8.62	8.19	5.40
climate	4.29	2.94	3.09
diabetes	3.37	3.45	3.01
fertility	5.30	4.50	4.32
german	5.87	5.58	6.34
golub	2.13	1.57	1.35
haberman	10.79	11.49	4.44
heart	2.55	2.40	2.01
ILPD	6.93	5.64	5.13
parkinsons	1.97	1.80	2.34
sonar	5.20	3.04	3.52

Tabela VI
SILHUETA MÉDIA DO CONJUNTO DE DADOS COMPLETO.

<i>Dataset</i>	Base	LMNN	LFDA
australian	0.18	0.16	0.26
banknote	0.24	0.33	0.70
breastcancer	0.57	0.62	0.71
breastHess	0.08	0.19	0.02
bupa	0.00	0.01	0.02
climate	0.03	0.06	0.07
diabetes	0.10	0.10	0.11
fertility	0.01	-0.06	0.05
german	0.03	0.03	0.04
golub	0.14	0.23	0.30
haberman	0.01	0.01	0.08
heart	0.13	0.14	0.17
ILPD	0.02	0.02	-0.08
parkinsons	0.12	0.17	0.11
sonar	0.03	0.10	0.07

Tabela VII
ÍNDICE DE QUALIDADE MÉDIA DOS VÉRTICES DO CONJUNTO DE DADOS COMPLETO.

<i>Dataset</i>	Base	LMNN	LFDA
australian	0.55	0.55	0.55
banknote	0.56	0.56	0.56
breastcancer	0.65	0.65	0.65
breastHess	0.75	0.75	0.72
bupa	0.42	0.43	0.41
climate	0.08	0.07	0.07
diabetes	0.35	0.36	0.35
fertility	0.89	0.86	0.89
german	0.70	0.70	0.70
golub	0.64	0.63	0.67
haberman	0.73	0.74	0.73
heart	0.56	0.56	0.55
ILPD	0.70	0.71	0.71
parkinsons	0.76	0.76	0.76
sonar	0.47	0.48	0.43

Tabela VIII
PROPORÇÃO DE AMOSTRAS AVALIADAS COMO SOBREPOSIÇÃO DO CONJUNTO DE DADOS COMPLETO.

<i>Dataset</i>	Base	LMNN	LFDA
australian	0.48	0.49	0.46
banknote	0.45	0.45	0.45
breastcancer	0.29	0.29	0.29
breastHess	0.36	0.36	0.34
bupa	0.48	0.53	0.50
climate	0.55	0.62	0.62
diabetes	0.52	0.52	0.53
fertility	0.40	0.28	0.43
german	0.47	0.46	0.47
golub	0.47	0.42	0.40
haberman	0.42	0.47	0.52
heart	0.50	0.49	0.48
ILPD	0.45	0.42	0.47
parkinsons	0.36	0.30	0.39
sonar	0.49	0.52	0.52

Tabela IX
REGRESSÃO LINEAR DOS ÍNDICES DE QUALIDADE PARA OS RESULTADOS DE DESEMPENHOS DO CLAS.

Atributo	Base		LMNN		LFDA	
	β	P	β	P	β	P
<i>intercept</i>	0.96	$8e-07$	0.81	$2e-07$	0.94	$1e-06$
CH	0.00	$3e-01$	-0.00	$9e-01$	0.00	$5e-01$
DB	-0.04	$1e-02$	-0.03	$3e-02$	-0.06	$2e-02$
Si	-0.56	$5e-01$	0.40	$3e-01$	-0.04	$9e-01$

Tabela X
REGRESSÃO LINEAR DOS ÍNDICES DE QUALIDADE PARA OS RESULTADOS DE DESEMPENHOS DA SVM.

Atributo	Base		LMNN		LFDA	
	β	P	β	P	β	P
<i>intercept</i>	0.82	$3e-05$	0.68	$5e-06$	0.82	$2e-05$
CH	0.00	$9e-01$	-0.00	$4e-01$	0.00	$1e+00$
DB	-0.03	$9e-02$	-0.01	$3e-01$	-0.04	$1e-01$
Si	0.32	$7e-01$	0.95	$6e-02$	0.26	$4e-01$

Tabela XI
REGRESSÃO LINEAR DOS ÍNDICES DE QUALIDADE PARA OS RESULTADOS DE DESEMPENHOS DO CLAS, INCLUEM-SE OS PROPOSTOS.

Atributo	Base		LMNN		LFDA	
	β	P	β	P	β	P
<i>intercept</i>	1.31	$1e-02$	0.62	$1e-01$	1.45	$2e-03$
CH	0.00	$5e-01$	0.00	$9e-01$	0.00	$4e-01$
DB	-0.04	$2e-02$	-0.03	$5e-02$	-0.05	$7e-02$
Si	-0.41	$6e-01$	0.38	$3e-01$	-0.11	$7e-01$
Q	-0.28	$2e-01$	0.02	$9e-01$	-0.35	$9e-02$
Di	-0.46	$5e-01$	0.41	$5e-01$	-0.72	$2e-01$

Tabela XII
REGRESSÃO LINEAR DOS ÍNDICES DE QUALIDADE PARA OS RESULTADOS DE DESEMPENHOS DA SVM, INCLUEM-SE OS PROPOSTOS.

Atributo	Base		LMNN		LFDA	
	β	P	β	P	β	P
<i>intercept</i>	0.56	$3e-01$	-0.10	$8e-01$	1.42	$7e-03$
CH	0.00	$9e-01$	-0.00	$9e-01$	0.00	$6e-01$
DB	-0.03	$1e-01$	-0.03	$7e-02$	-0.03	$3e-01$
Si	0.41	$7e-01$	0.78	$9e-02$	0.11	$7e-01$
Q	0.05	$9e-01$	0.48	$1e-01$	-0.30	$2e-01$
Di	0.48	$6e-01$	1.23	$7e-02$	-0.98	$2e-01$

dos índices de qualidade de agrupamentos. Seguindo o raciocínio que, apesar da pequena diferença em performance, um espaço mais representativo e com menor sobreposição é benéfico para o tratamento do problema.

De fato, o diferencial desses indicadores é mais ilativo, o que valida os métodos aplicados de *metric learning*. Resta, contudo, uma relação quantitativa desses índices com o interesse do problema, e a validação dos índices propostos.

Essa fase foi assumida pela tentativa dos modelos de regressão linear, infelizmente, no entanto, pouca informação foi obtida por esse processo. Os altos valores de valor-p não invalidam a hipótese nula em uma estimativa razoável para a grande maioria das observações.

VI. CONCLUSÕES

Este trabalho apresentou uma tentativa de aproximação entre o desempenho dos classificadores e índices de qualidade de agrupamentos a partir de um ajuste de regressão linear. Também, propôs dois índices baseados na estrutura do grafo construído no espaço.

Mostrou-se que técnicas de aprendizado de métrica respondem bem aos indicadores mas têm performance tímida no resultado dos classificadores contra o espaço tradicional. A tentativa de relacionar essas grandezas foi frustrada.

Resta a sugestão de um ferramental matemático mais elaborado para validar os índices propostos e justificar seu uso contra o interesse do problema, o resultado do classificador na ponta.

AGRADECIMENTO

Este trabalho foi possível pela disponibilização das bases de dados pelo repositório *UCI Machine Learning* [14].

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 University*, vol. 2, no. 2, p. 4, 2006.
- [3] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *arXiv preprint arXiv:1306.6709*, 2013.
- [4] L. Torres, C. Castro, F. Coelho, F. Sill Torres, and A. Braga, "Distance-based large margin classifier suitable for integrated circuit implementation," *Electronics Letters*, vol. 51, no. 24, pp. 1967–1969, 2015.
- [5] L. C. B. Torres, "Classificador por arestas de suporte (clas): Métodos de aprendizado baseados em grafos de gabriel," 2016.
- [6] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis.," *Journal of machine learning research*, vol. 8, no. 5, 2007.
- [7] 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.
- [8] I. Fehervari, A. Ravichandran, and S. Appalaraju, "Unbiased evaluation of deep metric learning algorithms," *arXiv preprint arXiv:1911.12528*, 2019.
- [9] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3279–3286, 2015.
- [10] P. Moutafis, M. Leng, and I. A. Kakadiaris, "An overview and empirical comparison of distance metric learning methods," *IEEE transactions on cybernetics*, vol. 47, no. 3, pp. 612–625, 2016.
- [11] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.
- [12] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.
- [13] 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.
- [14] D. Dua and C. Graff, "UCI machine learning repository," 2017.
- [15] 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.