O Algoritmo K-médias Geodésico para Agrupamento de Dados

The Geodesic K-Means Algorithm for Data Clustering

Projeto de Pesquisa enviado à FAPESP
MODALIDADE: AUXÍLIO REGULAR

Bolsista:
Antônio Cícero Amorim de Azevedo – azevedoantonio@estudante.ufscar.br
Departamento de Computação - UFSCar

Orientador:
Alexandre Luis Magalhães Levada – alexandre.levada@ufscar.br
Departamento de Computação - UFSCar

8 de janeiro de 2024

Resumo

Algoritmos de aprendizado não supervisionado de métricas utilizam grafos como aproximações discretas para as variedades que representam a estrutura geométrica subjacente dos conjuntos de dados multivariados. Atualmente, a metodologia adotada pelos algoritmos tradicionais de aprendizado de variedades na ponderação das arestas desses grafos ainda é bastante rudimentar, uma vez que adota-se a distância Euclidiana como medida de similaridade. Porém, a distância Euclidiana é extrínseca à variedade em questão e não leva em consideração a noção de curvatura. Por essa razão, esse projeto de pesquisa visa propor métodos matematicamente originais, mais precisos e adequados para a caracterização da similaridade entre amostras vizinhas do grafo KNN. A ideia consiste em utilizar conceitos da geometria diferencial, como a curvatura, que é uma medida intrínseca, para ponderar arestas de caminhos mínimos em tais grafos, uma vez que eles representam aproximações para as verdadeiras distâncias geodésicas entre diferentes pontos pertencentes à variedade. Basicamente, a estratégia consiste em medir as variações dos espaços tangentes conforme nos movemos através de um caminho mínimo no grafo KNN. Com isso, espera-se melhorar o desempenho de diversos algoritmos utilizados na extração de características em problemas de classificação de padrões, como o PCA (*Principal Component Analysis*), ISOMAP (*Isometric Feature Mapping*) e LLE (*Locally Linear Embedding*). Resultados preliminares com o algoritmo ISOMAP mostram um ganho significativo na acurácia da classificação de diversos conjuntos de dados reais em comparação com as versões tradicionais e outros métodos estado da arte, como t-SNE e UMAP. Além disso, espera-se incorporar os grafos baseados em curvatura (*K-graphs*) em modelos para classificação de padrões e agrupamento de dados baseados em grafos.

Abstract

Unsupervised metric learning algorithms use graphs as discrete approximations for the manifolds that represent the underlying geometric structure of multivariate data. At the moment, the usual methodology employed by these algorithms for building the KNN graph edges is quite elementary, since the Euclidean distante is the similarity measure. However, the Euclidean distance is extrinsic to the manifold and it does not take into account the intrinsic notion of curvature. For this reason, this research project aims to propose original and more suitable mathematical methods to characterize the similarity between neighboring samples in the KNN graph. The idea consists in the application of differential geometry concepts, such as the curvature, an intrinsic measure, to weight the edges of shortest paths in the graphs, since they represent discrete approximations to the true underlying geodesic distances between different points belonging to the manifold. Basically, our strategy is to measure the variation of the tangent space as we move along a shortest path in the KNN graph. With the proposed method, we expect to improve the performance of several feature extraction algorithms in pattern classification, such as PCA (*Principal Component Analysis*), ISOMAP (*Isometric Feature Mapping*) and LLE (*Locally Linear Embedding*). Preliminary results with the ISOMAP algorithm show a significant gain in the classification accuracy of several real world datasets in comparison to their regular versions and other state-of-the-art methods, such as t-SNE and UMAP. Moreover, we intend to incorporate the proposed curvature based graphs (*K-graphs*) in graph-based classification and clustering models.

1 Enunciado do problema

test [1, 2, 3, 4, 5, 6]. test [7, 8], (LLE) [9, 10, 11, 12, 13] test[14, 15].

Referências

[1] H. S. Seung and D. D. Lee, "The manifold ways of perception," *Science*, vol. 290, pp. 2268–2269, 2000.

[2] M. Brand, *Charting a manifold*, vol. 15, pp. 961–968. MIT Press, 2003.

[3] L. Cayton, "Algorithms for Manifold Learning," tech. rep., University of California San Diego (UCSD), 2005.

[4] X. Huo, X. Ni, and A. K. Smith, *A Survey of Manifold-Based Learning Methods*, vol. 6 of *Series on Computers and Operations Research*, pp. 691–745. World Scientific, 2008.

[5] H. Yin and W. Huang, "Adaptive nonlinear manifolds and their applications to pattern recognition," *Information Sciences*, vol. 180, no. 14, pp. 2649–2662, 2010.

[6] R. Talmon, I. Cohen, S. Gannot, and R. R. Coifman, "Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 75–86, 2013.

[7] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[8] M. Bernstein, V. de Silva, J. Langford, and J. Tenenbaum, "Graph approximations to geodesics on embedded manifolds," 2000.

[9] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, 2000.

[10] L. Saul and S. Roweis, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[11] L. K. Saul and S. T. Roweis, "An introduction to locally linear embedding," tech. rep., New York University, 2000.

[12] C. Shalizi, "Nonlinear dimensionality reduction i: Local linear embedding," 2009. http://www.stat.cmu.edu/ cshalizi/350/lectures/14/lecture-14.pdf.

[13] D. de Ridder and R. P. Duin, "Locally linear embedding for classification," tech. rep., Delft University of Technology, 2002.

[14] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, pp. 1373–1396, June 2003.

[15] L. Bo, L. Yan-Rui, and Z. Xiao-Long, "A survey on laplacian eigenmaps based manifold learning methods," *Neurocomputing*, vol. 335, pp. 336–351, 2018.

[16] E. Debie and K. Shafi, "Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses," *Pattern Analysis and Applications*, vol. 22, pp. 519–536, 2019.

[17] R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[18] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[19] D. W. Scott, *Multivariate Density Estimation*. John Wiley & Sons, 1992.

[20] J. Hwang, S. Lay, and A. Lippman, "Nonparametric multivariate density estimation: A comparative study," *IEEE Trans. on Signal Processing*, vol. 42, no. 10, pp. 2795–2810, 1994.

[21] G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. on Information Theory*, vol. 14, pp. 55–63, 1968.

[22] D. Li and Y. Tian, "Survey and experimental study on metric learning methods," *Neural Networks*, vol. 105, pp. 447–462, 2018.

[23] F. Wang and J. Sun, "Survey on distance metric learning and dimensionality reduction in data mining," *Data Min. Knowl. Discov.*, vol. 29, pp. 534–564, Mar. 2015.

[24] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*. Springer, 2007.

[25] A. L. M. Levada, "Parametric pca for unsupervised metric learning," *Pattern Recognition Letters*, vol. 135, pp. 425–430, 2020.

[26] A. L. M. Levada, "Pca-kl: a parametric dimensionality reduction approach for unsupervised metric learning," *Advances in Data Analysis and Classification*, 2021.

[27] I. T. Jolliffe, *Principal Component Analysis*. Springer, 2 ed., 2002.

[28] B. Schölkopf, A. Smola, and K. R. Müller, "Kernel principal component analysis," in *Advances in Kernel Methods – Support Vector Learning*, pp. 327–352, MIT Press, 1999.

[29] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[30] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, 2011.

[31] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[32] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[33] A. C. Neto and A. L. M. Levada, "Isomap-kl: a parametric approach for unsupervised metric learning," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 287–294, 2020.

[34] B. O'Neill, *Elementary Differential Geometry*. Elsevier, 2nd ed., 2006.

[35] T. Shifrin, *Differential Geometry: A First Course in Curves and Surfaces*. University of Georgia, 2016.

[36] M. P. do Carmo, *Differential Geometry of Curves and Surfaces*. Dover Publications Inc., 2nd ed., 2017.

[37] J. A. Serret, "Sur quelques formules relatives à la théorie des courbes à double courbure.," *J. de Math*, 1851.

[38] F. Frenet, "Sur les courbes à double courbure.," *Abstract in J. de Math*, 1852.

[39] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*, vol. 88 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 2001.

[40] I. Borg and P. Groenen, *Modern Multidimensional Scaling: theory and applications*. Springer-Verlag, 2 ed., 2005.

[41] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.

[42] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. MIT Press, 3 ed., 2009.

[43] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[44] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American Statistical Association*, vol. 32, no. 200, pp. 675–701, 1937.

[45] P. B. Nemenyi, *Distribution-free multiple comparisons*. PhD thesis, Princeton University, 1963.

[46] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Comp. and Appl. Math.*, vol. 20, pp. 53–65, 1987.

[47] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[48] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 553–569, 1983.