Representação de Conjuntos de Dados utilizando Redes Neurais Artificiais

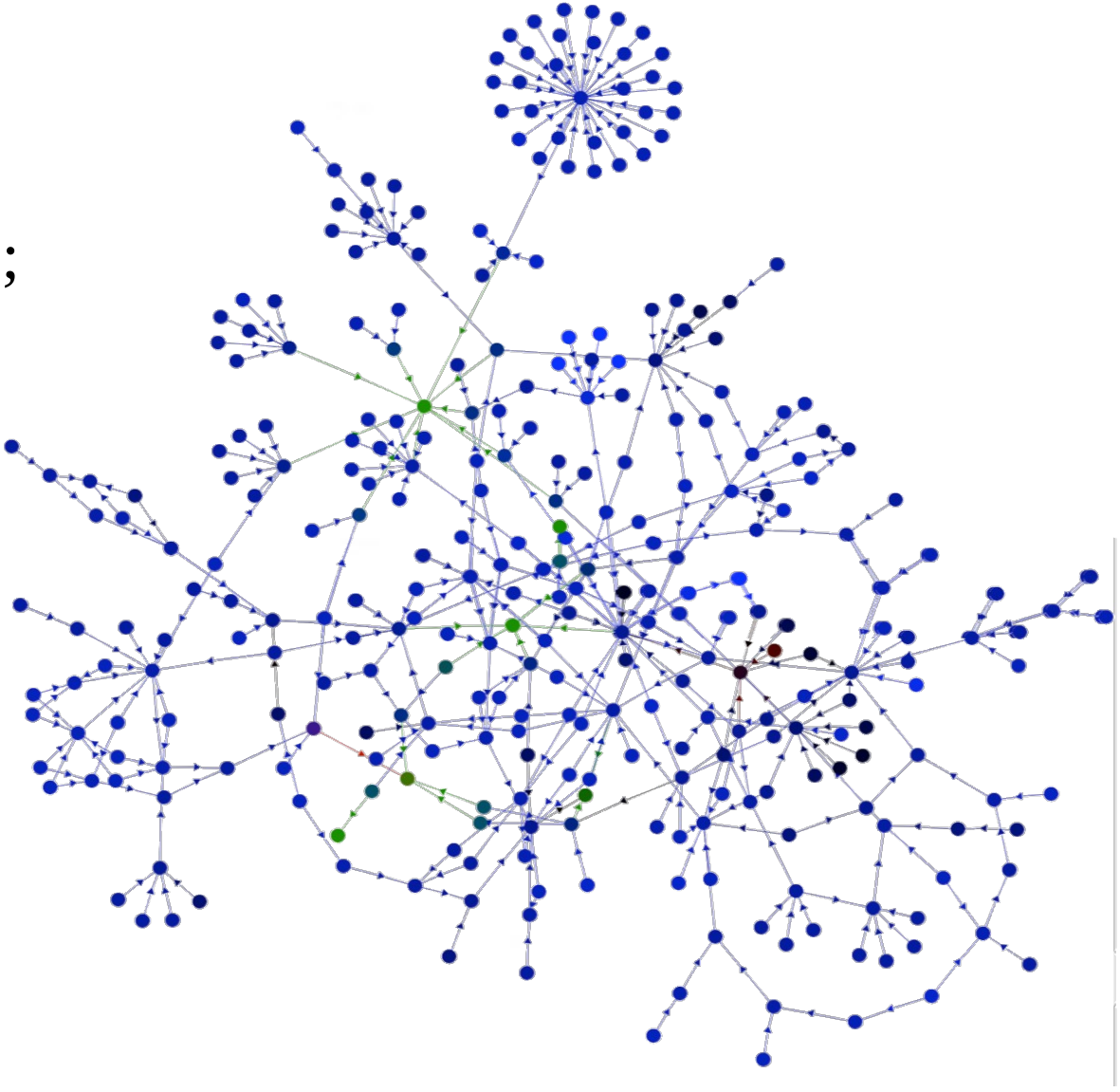
Victor De Cia Costa

Orientador: Prof. Dr. Marcos Gonçalves Quiles

4 de Julho, 2016. São José dos Campos, SP

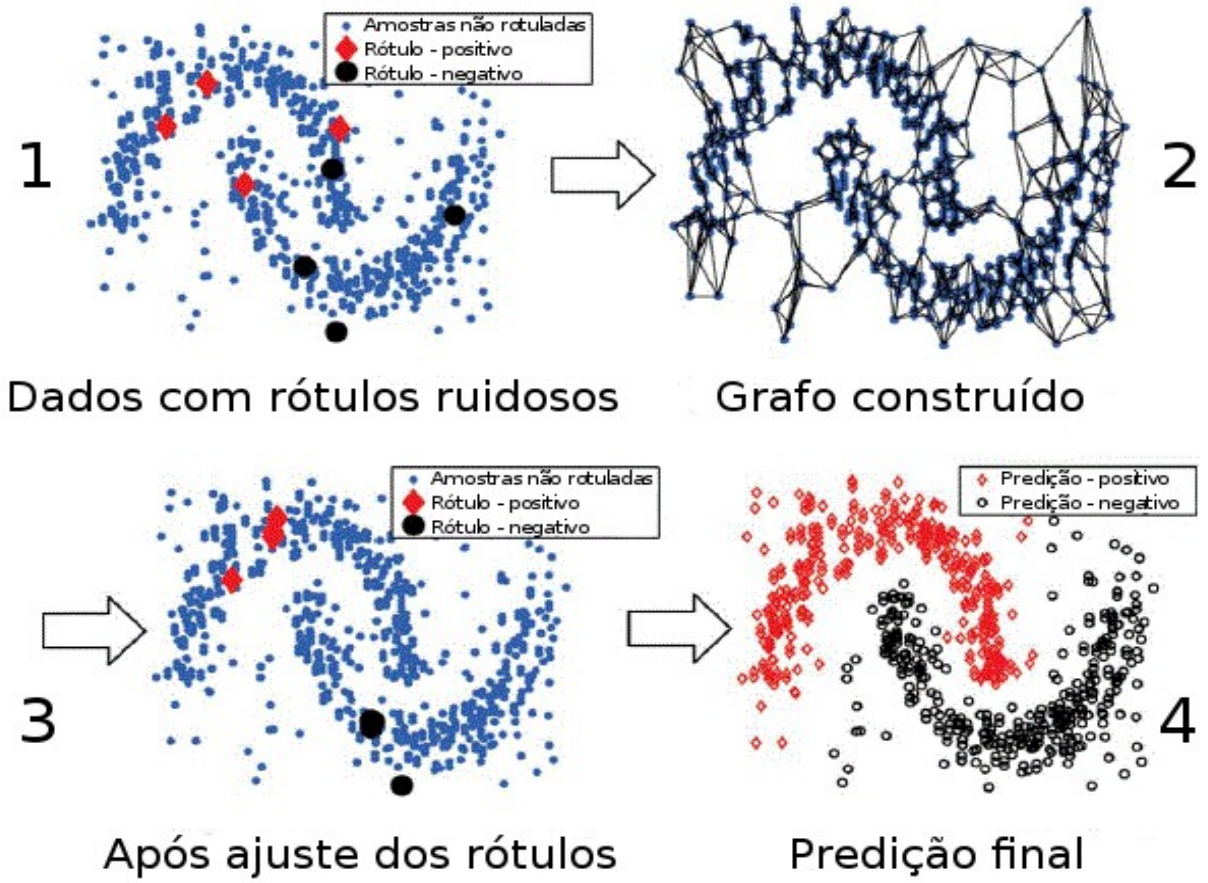
Introdução

- O problema:
  - Dados reais;
  - Alta dimensionalidade;



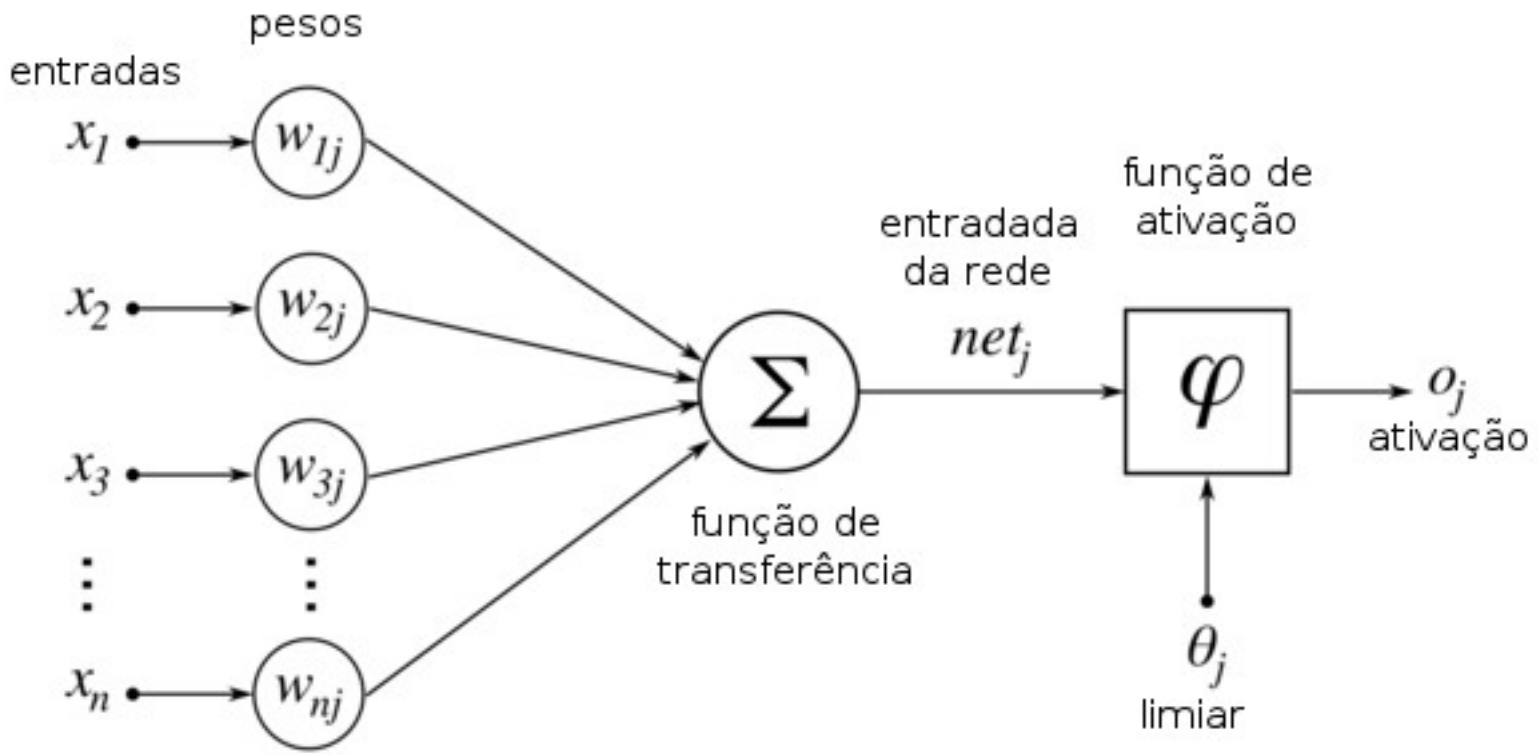
Introdução (continuação)

- O problema:
  - Geração de uma rede que represente fielmente dados de uma base;
- Objetivos;



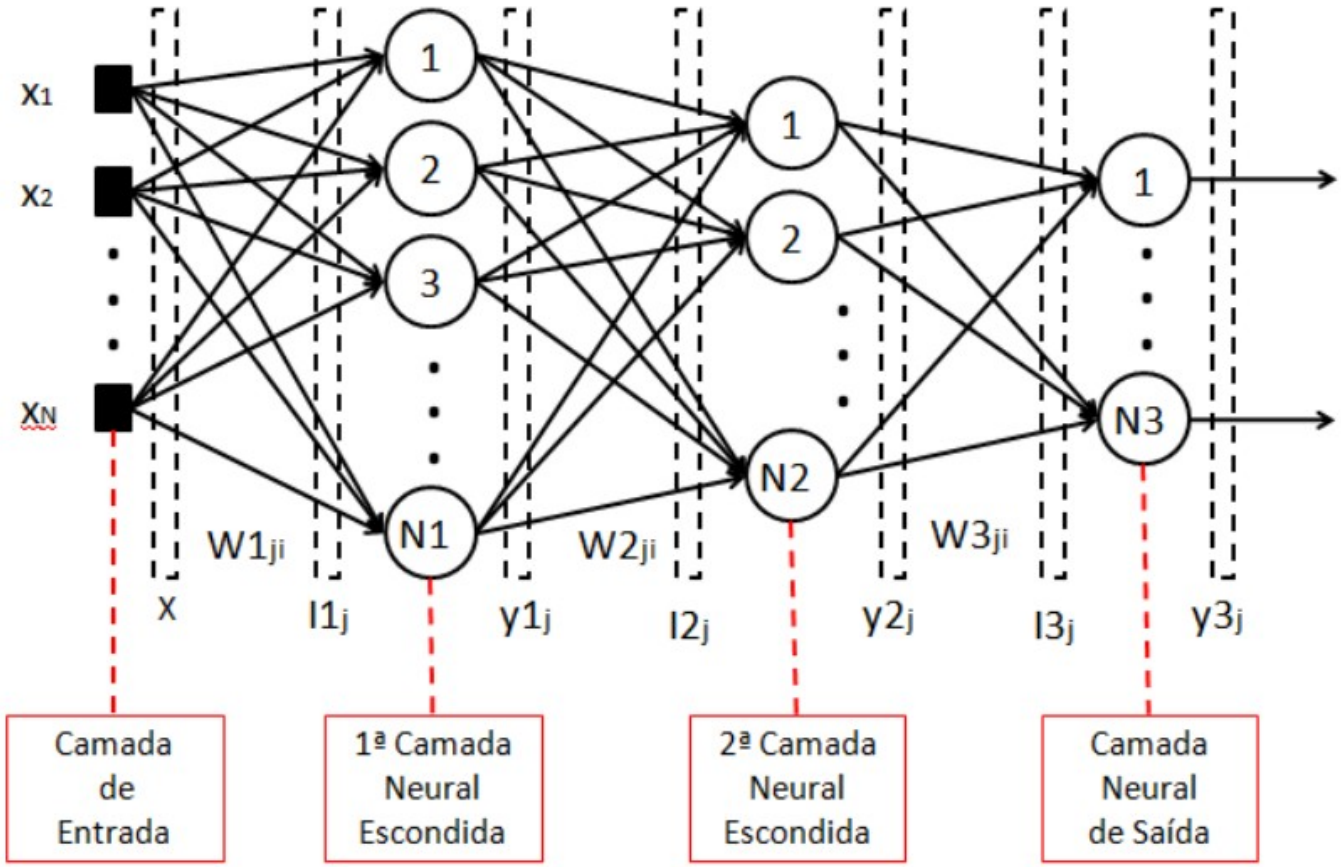
Revisão Bibliográfica

- Redes Neurais Artificiais;
  - Perceptron;



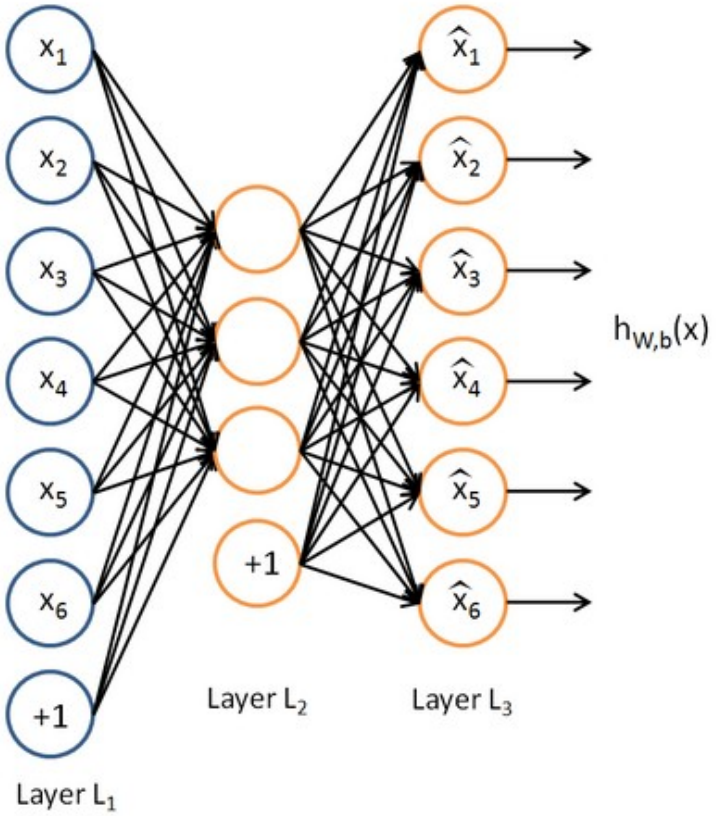
Revisão Bibliográfica (continuação)

- Redes Neurais Artificiais;
  - MLP;



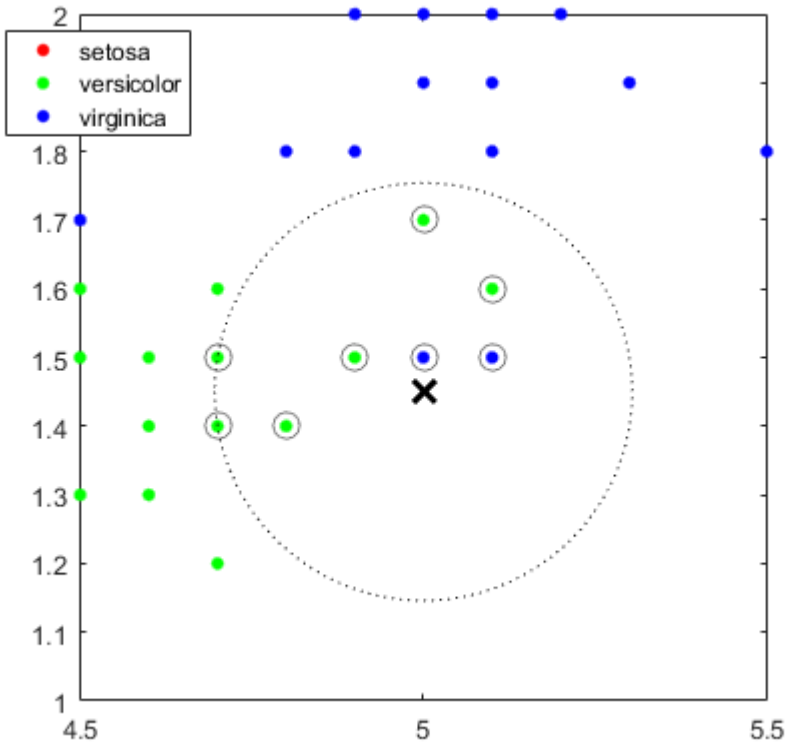
Revisão Bibliográfica (continuação)

- Redes Neurais Artificiais;
  - Autoencoders;



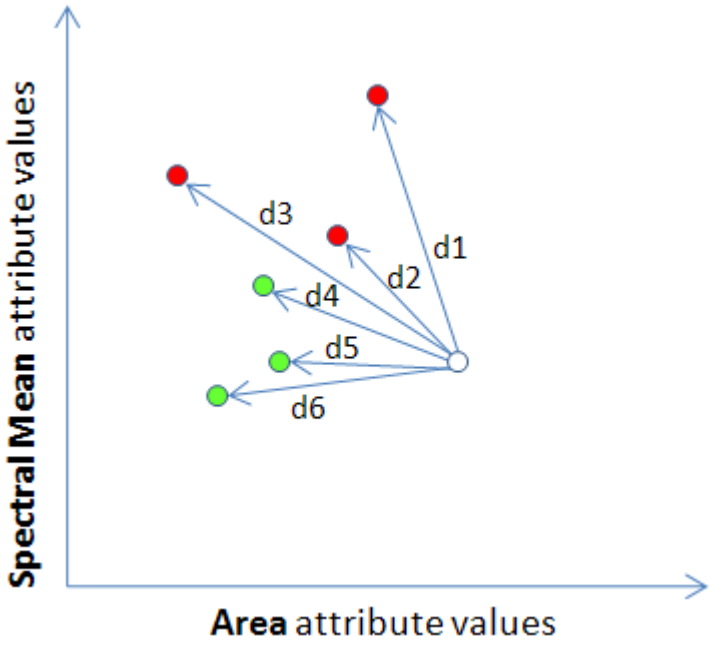
Revisão Bibliográfica (continuação)

- Representação de dados através de grafos;
  - *K-nn*;



Revisão Bibliográfica (continuação)

- Representação de dados através de grafos;
  - Corte epsilon;



Revisão Bibliográfica (continuação)

- Representação de dados através de grafos;
  - Técnicas Híbridas:

    Corte epsilon + *K-nn*: pode-se utilizar o corte epsilon para gerar uma rede inicial e o *K-nn* como uma ferramenta de pós-processamento.

    – Essa abordagem pode reduzir a quantidade de exemplos isolados facilmente observados ao utilizar a técnica corte epsilon isoladamente.

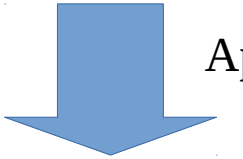
Metodologia

- Uso de autoencoders para geração de novas representações;

  O autoencoder foi utilizado como uma ferramenta de pré-processamento dos dados.

Tabela 1 – Exemplo de tabela atributo-valor com amostras do conjunto de dados Iris da UCI.

sepal lenght	**sepal width**	**petal lenght**	**petal width**	**Class**
5,1	3,5	1,4	0,2	Iris-setosa
7,0	3,2	4,7	1,4	Iris-versicolor
6,3	3,3	6,0	2,5	Iris-virginica

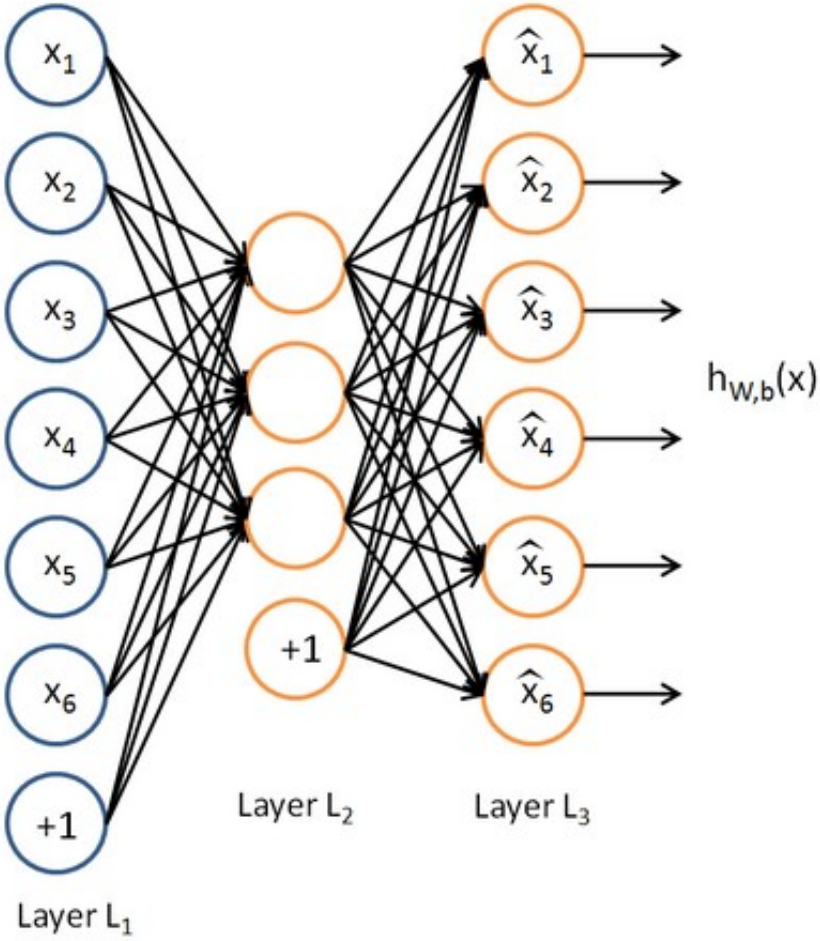


Aplica-se o autoencoder

Atributo 1	Atributo 2
0,67	0,28
0,03	0,21
0,42	0,27

Metodologia (continuação)

- Conjunto de características;
- Vantagens: correlação reduzida entre os novos atributos, valor normalizados entre 0 e 1 (ou -1 e 1, conforme a função de ativação utilizada) e número reduzido de características;



Metodologia (continuação)

- Geração da rede com as novas representações;

1. Um vértice é gerado para cada exemplo do conjunto de dados;
2. Uma função de similaridade é definida;
3. Os parâmetros de similaridade são ajustados;
4. Se a similaridade entre um par de exemplos é superior a um determinado limiar (épsilon) ou se eles estão entre os k vizinhos mais próximos, uma aresta é criada entre seus respectivos vértices representantes;
5. As arestas podem ou não ser ponderadas;

Metodologia (continuação)

- Experimentos computacionais e avaliação dos resultados;

- Comparações:
  - Funções Harmônicas (ZHU et al., 2003);
  - **Consistência Local e Global (ZHOU et al., 2004);**
  - Técnica de classificação baseada em Competição de Partículas (BREVE; ZHAO; QUILES, 2015);

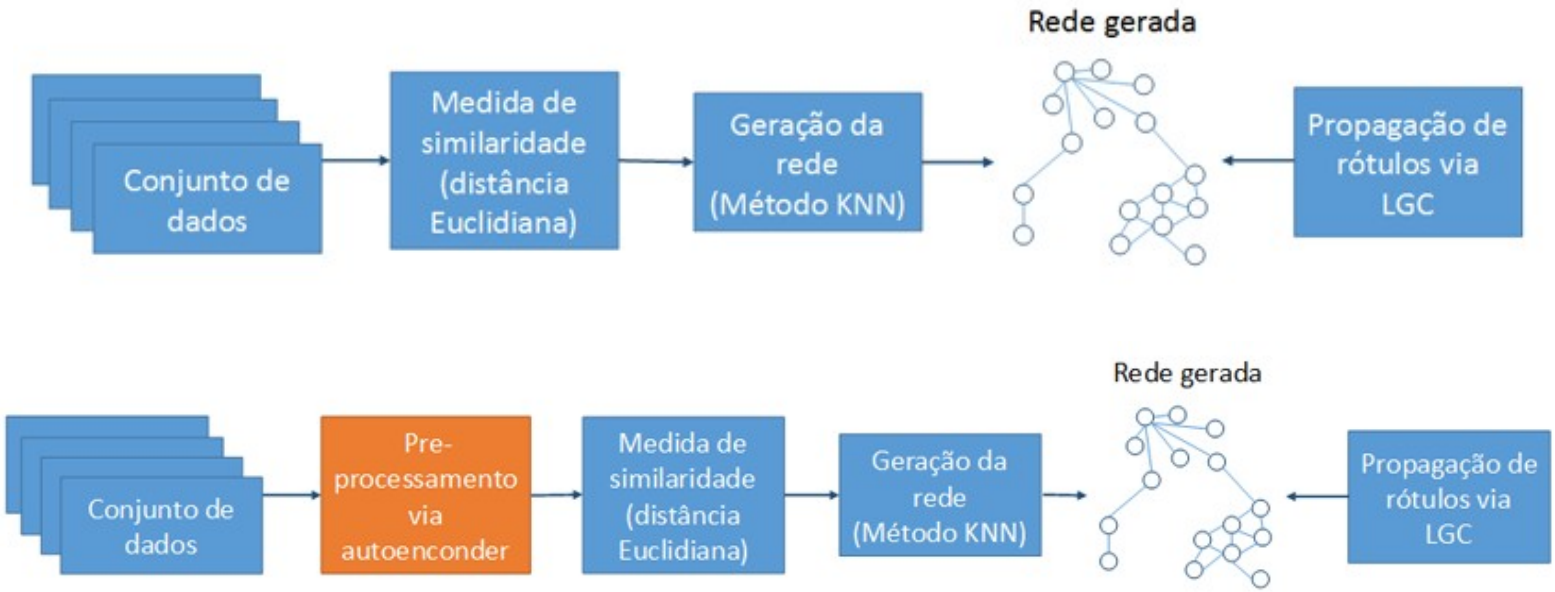
Metodologia (continuação)

- Experimentos computacionais e avaliação dos resultados;
- Bases de dados:
  - Testes foram realizados com dados selecionados da UCI;

Metodologia (continuação)

- Experimentos computacionais e avaliação dos resultados;

  Para os testes e comparações foram comparadas as redes geradas com os novos atributos obtidos pelo autoencoder com as redes geradas a partir dos dados originais sem pré-processamento.



Experimentos e Resultados

Bases de dados utilizadas nos experimentos e as quantidades de objetos, atributos e classes em cada uma delas.

Nome	Objetos	Atributos	Classes
glass	214	9	7
iris	150	4	3
wine	178	13	3
seeds	210	7	3
parkinsons	195	22	2

Experimentos e Resultados (continuação)

Valores utilizados para os parâmetros.

Parâmetro	Valor
k	1...10
Γ	100000
μ	2...(Quantidade de atributos do conjunto de dados) - 1
δ	10

Γ: utilizado no autoencoder para definir o máximo de épocas que a rede poderia alcançar em sua fase de treinamento;

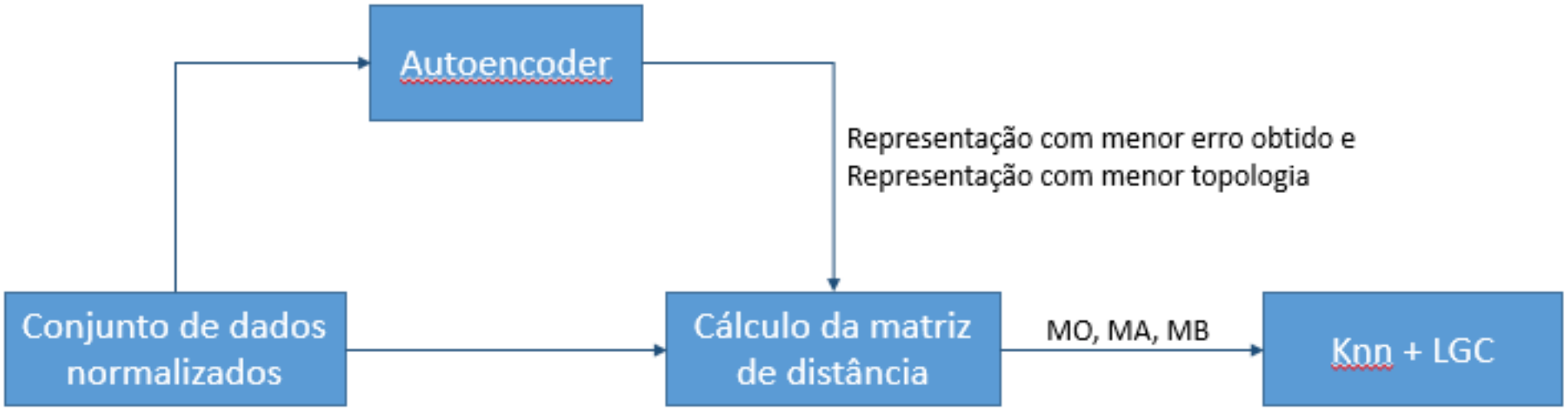
μ: o número de neurônios da camada oculta do autoencoder;

δ: a quantidade de folds utilizados na validação cruzada;

k: utilizados nos métodos de geração da rede kNN;

Experimentos e Resultados (continuação)

Resumo dos experimentos realizados:



Estatísticas dos valores obtidos da execução do algoritmo LGC 100 vezes com k variando de 1...10, para cada representação do conjunto de dados *glass*.

		Conjunto de dados		
		GLASS		
k		Original	Autoencoder Menor Erro	Autoencoder Menor Topologia
1	Média	0.995259	0.992984	0.994842
	Des. Pad.	1.000274	0.998007	0.999854
	Mediana	0.994845	0.994845	0.994845
2	Média	0.995205	0.945255	0.969951
	Des. Pad.	1.000219	0.963155	0.979096
	Mediana	0.994845	0.994845	0.994845
3	Média	0.980913	0.967444	0.953305
	Des. Pad.	0.986862	0.974927	0.967270
	Mediana	0.994845	0.994845	0.994845
4	Média	0.979859	0.994843	0.968613
	Des. Pad.	0.985568	0.999855	0.979070
	Mediana	0.994845	0.994845	0.994845
5	Média	0.975427	0.984143	0.955600
	Des. Pad.	0.982306	0.989628	0.968569
	Mediana	0.994845	0.994845	0.994845
6	Média	0.975136	0.991657	0.972010
	Des. Pad.	0.981850	0.996721	0.980485
	Mediana	0.994845	0.994845	0.994845
7	Média	0.988391	0.940674	0.968572
	Des. Pad.	0.993614	0.956846	0.977562
	Mediana	0.994845	0.994845	0.994845
8	Média	0.993759	0.995152	0.994637
	Des. Pad.	0.998770	1.000166	0.999648
	Mediana	0.994845	0.994845	0.994845
9	Média	0.985238	0.993396	0.994843
	Des. Pad.	0.990679	0.998406	0.999854
	Mediana	0.994845	0.994819	0.994845
10	Média	0.994274	0.995616	0.993605
	Des. Pad.	0.999283	1.000633	0.998616
	Mediana	0.994819	0.994845	0.994845

k		Conjunto de dados		
		IRIS		
		Original	Autoencoder Menor Erro	Autoencoder Menor Topologia
1	Média	0.989593	0.984785	0.992699
	Des. Pad.	0.994628	0.990066	0.997701
	Mediana	0.992593	0.992593	0.992593
2	Média	0.934571	0.963581	0.935606
	Des. Pad.	0.957434	0.973491	0.957819
	Mediana	0.992593	0.992647	0.992593
3	Média	0.941609	0.965203	0.944042
	Des. Pad.	0.962205	0.974566	0.963148
	Mediana	0.992647	0.992647	0.992647
4	Média	0.942199	0.939709	0.907034
	Des. Pad.	0.962427	0.952901	0.934339
	Mediana	0.992647	0.992593	0.992593
5	Média	0.898054	0.974878	0.856759
	Des. Pad.	0.930518	0.982874	0.920393
	Mediana	0.992593	0.992647	0.992593
6	Média	0.961267	0.975248	0.946327
	Des. Pad.	0.975833	0.983117	0.972486
	Mediana	0.992647	0.992647	0.992647
7	Média	0.961267	0.963445	0.910672
	Des. Pad.	0.975833	0.972795	0.952009
	Mediana	0.992647	0.992647	0.992647
8	Média	0.963670	0.981695	0.949371
	Des. Pad.	0.972943	0.987094	0.964132
	Mediana	0.992647	0.992593	0.992647
9	Média	0.947723	0.986420	0.924166
	Des. Pad.	0.960309	0.991538	0.947334
	Mediana	0.992593	0.992593	0.992593
10	Média	0.956851	0.988493	0.970646
	Des. Pad.	0.966594	0.993538	0.977371
	Mediana	0.992593	0.992593	0.992593

Estatísticas dos valores obtidos da execução do algoritmo LGC 100 vezes com k variando de 1...10, para cada representação do conjunto de dados *iris*.

Estatísticas dos valores obtidos da execução do algoritmo LGC 100 vezes com k variando de 1...10, para cada representação do conjunto de dados *parkinsons*.

		Conjunto de dados		
		PARKINSONS		
k		Original	Autoencoder Menor Erro	Autoencoder Menor Topologia
1	Média	0.994856	0.861414	0.922460
	Des. Pad.	0.999870	0.865993	0.927828
	Mediana	0.994350	0.864407	0.943662
2	Média	0.980687	0.653287	0.735222
	Des. Pad.	0.986333	0.681652	0.769491
	Mediana	0.994350	0.723164	0.627119
3	Média	0.968116	0.594160	0.729614
	Des. Pad.	0.978575	0.630512	0.776687
	Mediana	0.994350	0.677966	0.795455
4	Média	0.969187	0.603274	0.766037
	Des. Pad.	0.977843	0.643237	0.801116
	Mediana	0.994350	0.681818	0.954545
5	Média	0.988148	0.603341	0.766099
	Des. Pad.	0.994082	0.641872	0.806275
	Mediana	0.994334	0.681818	0.954545
6	Média	0.969812	0.572492	0.687376
	Des. Pad.	0.977851	0.624064	0.731573
	Mediana	0.994350	0.680797	0.619318
7	Média	0.955838	0.573430	0.733096
	Des. Pad.	0.966363	0.621965	0.778736
	Mediana	0.994350	0.684505	0.801136
8	Média	0.979642	0.570112	0.717579
	Des. Pad.	0.986318	0.609790	0.757468
	Mediana	0.994350	0.677966	0.627119
9	Média	0.972793	0.614630	0.745882
	Des. Pad.	0.980452	0.641346	0.786133
	Mediana	0.994318	0.681564	0.803346
10	Média	0.964249	0.614250	0.779955
	Des. Pad.	0.972568	0.642903	0.809717
	Mediana	0.994350	0.680669	0.954545

		Conjunto de dados		
		SEEDS		
k		Original	Autoencoder Menor Erro	Autoencoder Menor Topologia
1	Média	0.991266	0.987995	0.993527
	Des. Pad.	0.996324	0.992979	0.998541
	Mediana	0.994737	0.989446	0.994737
2	Média	0.951833	0.866398	0.967237
	Des. Pad.	0.974918	0.901388	0.979548
	Mediana	0.994737	0.989418	0.994737
3	Média	0.956318	0.933308	0.943812
	Des. Pad.	0.979573	0.953179	0.962431
	Mediana	0.994737	0.989474	0.994737
4	Média	0.920261	0.918500	0.944051
	Des. Pad.	0.955031	0.941396	0.962539
	Mediana	0.994737	0.989474	0.994737
5	Média	0.943412	0.856063	0.961657
	Des. Pad.	0.969897	0.889735	0.975420
	Mediana	0.994737	0.989418	0.994737
6	Média	0.888065	0.938336	0.962628
	Des. Pad.	0.938572	0.951039	0.975879
	Mediana	0.994737	0.989474	0.994737
7	Média	0.870638	0.921865	0.949512
	Des. Pad.	0.928027	0.940325	0.960929
	Mediana	0.994737	0.989474	0.994737
8	Média	0.898272	0.926578	0.956644
	Des. Pad.	0.943981	0.944103	0.968302
	Mediana	0.994737	0.989418	0.994723
9	Média	0.898065	0.927151	0.994262
	Des. Pad.	0.943955	0.946061	0.999273
	Mediana	0.994737	0.989474	0.994737
10	Média	0.877321	0.938839	0.993467
	Des. Pad.	0.933093	0.955015	0.998478
	Mediana	0.994723	0.989418	0.994723

Estatísticas dos valores obtidos da execução do algoritmo LGC 100 vezes com k variando de 1...10, para cada representação do conjunto de dados *seeds*.

Estatísticas dos valores obtidos da execução do algoritmo LGC 100 vezes com k variando de 1...10, para cada representação do conjunto de dados *wine*.

		Conjunto de dados		
		WINE		
k		Original	Autoencoder Menor Erro	Autoencoder Menor Topologia
1	Média	0.992823	0.992017	0.992017
	Des. Pad.	0.997830	0.997034	0.997034
	Mediana	0.993827	0.993827	0.993827
2	Média	0.918900	0.918771	0.925339
	Des. Pad.	0.958177	0.958167	0.958841
	Mediana	0.993827	0.993789	0.993789
3	Média	0.924163	0.894448	0.889502
	Des. Pad.	0.958676	0.942770	0.942408
	Mediana	0.993789	0.993789	0.993789
4	Média	0.900879	0.936550	0.936861
	Des. Pad.	0.943595	0.968446	0.968461
	Mediana	0.993789	0.993827	0.993827
5	Média	0.935991	0.902551	0.896801
	Des. Pad.	0.968426	0.943881	0.943028
	Mediana	0.993827	0.993789	0.993789
6	Média	0.887711	0.938662	0.906632
	Des. Pad.	0.942338	0.964791	0.944678
	Mediana	0.993789	0.993827	0.993789
7	Média	0.926780	0.956164	0.941556
	Des. Pad.	0.963285	0.975041	0.965477
	Mediana	0.993827	0.993827	0.993827
8	Média	0.948450	0.954988	0.956597
	Des. Pad.	0.973729	0.974762	0.975150
	Mediana	0.993827	0.993827	0.993827
9	Média	0.947978	0.962019	0.953623
	Des. Pad.	0.973692	0.979467	0.974475
	Mediana	0.993827	0.993827	0.993827
10	Média	0.955560	0.913781	0.960901
	Des. Pad.	0.978676	0.949562	0.979252
	Mediana	0.993827	0.993789	0.993827

Experimentos e Resultados (continuação)

Quantidade de atributos obtidos pelas representações geradas pelo autoencoder por conjunto de dados.

	Número de atributos		
	Original	**Representação MA**	**Representação MB**
glass	9	3	2
iris	4	3	2
parkinsons	22	19	5
seeds	7	6	5
wine	13	9	7

MA: Maior Acurácia;
MB: Menor topologia de acurácia aceitável;

Experimentos e Resultados (continuação)

Valores das acurácias obtidas ao treinar a rede autoencoder para as representações MA e MB. Foi utilizada a validação cruzada com método *k-fold* utilizando 10 *folds*.

	Acurácias	
	MA	**MB**
glass	0.002566	0.002859
iris	0.003857	0.004241
parkinsons	0.010058	0.011330
seeds	0.008254	0.009627
wine	0.008728	0.011347

MA: Maior Acurácia;
MB: Menor topologia de acurácia aceitável;

Conclusões

As duas representações testadas geradas pela rede possuiam menos atributos do que a representação original e obtiveram no geral, os melhores resultados quando aplicado o algorítimo kNN e LGC. Para a base de dados parkinsons, a base de maior número de atributos, observou-se que os melhores resultados foram obtidos em maioria esmagadora através das representações geradas pela técnica. Reduzindo de 22 atributos até 5, maximizando os resultados.

Conclusões (continuação)

Os resultados obtidos mostraram que foi possível utilizar um autoencoder como ferramenta de pré-processamento do conjunto de dados, gerando melhores resultados, ou equivalentes, do que da representação sem esse processamento. Notou-se também que em todos os conjuntos de dados, foi vantajoso ter seus atributos reduzidos, como já era esperado.

Referências

BREVE, F. A.; ZHAO, L.; QUILES, M. G. Particle competition and cooperation for semi-supervised learning with label noise. Neurocomputing, Elsevier, v. 160, p. 63–72, 2015.

HAUSSLER, D.; WELZL, E. -nets and simplex range queries. Discrete & Computational Geometry, Springer, v. 2, n. 1, p. 127–151, 1987.

HEIN, M.; MAIER, M. Manifold denoising as preprocessing for finding natural representations of data. In: MENLO PARK, CA; CAMBRIDGE, MA; LONDON; AAAI PRESS; MIT PRESS; 1999. PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE. [S.l.], 2007. v. 22, n. 2, p. 1646.

JEBARA, T.; SHCHOGOLEV, V. B-matching for spectral clustering. In: Machine learning: Ecml 2006. [S.l.]: Springer Berlin Heidelberg, 2006. p. 679–686.

JEBARA, T.; WANG, J.; CHANG, S.-F. Graph construction and b-matching for semi-supervised learning. In: ACM. Proceedings of the 26th Annual International Conference on Machine Learning. [S.l.], 2009. p. 441–448.

KUBAT, M. Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7. [S.l.]: Cambridge Univ Press, 1999.

LIU, W.; CHANG, S.-F. Robust multi-class transductive learning with graphs. In: IEEE. Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. [S.l.], 2009. p. 381–388.

OLSHAUSEN, B. A. et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, v. 381, n. 6583, p. 607–609, 1996.

QUILES, M. G. et al. Label propagation through neuronal synchrony. In: IEEE. Neural Networks (IJCNN), The 2010 International Joint Conference on. [S.l.], 2010. p. 1–8.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning internal representations by error propagation. [S.l.], 1985.

SOUSA, C. A. R. d. Impacto da geração de grafos na classificação semissupervisionada. Tese (Doutorado) — Universidade de São Paulo, 2013.

ZHOU, D. et al. Learning with local and global consistency. Advances in neural information processing systems, v. 16, n. 16, p. 321–328, 2004.

ZHU, X. et al. Semi-supervised learning using gaussian fields and harmonic functions. In: ICML. [S.l.: s.n.], 2003. v. 3, p. 912–919.

ZHU, X.; GOLDBERG, A. B. Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning, Morgan & Claypool Publishers, v. 3, n. 1, p. 1–130, 2009.

ZHU, X.; LAFFERTY, J.; ROSENFELD, R. Semi-supervised learning with graphs. [S.l.]: Carnegie Mellon University, Language Technologies Institute, School of Computer Science, 2005.