Classificação de Alto Nível Baseada em Entropia da Rede

Filipe Alves Neto[1], **Liang Zhao**[1]

[1] Departamento de Ciências de Computação
Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo (USP)
Av. Trabalhador São-carlense, 400, 13560-970, São Carlos, SP, Brasil

filipeneto@usp.br, zhao@icmc.usp.br

Abstract. *Traditional data classification is based only on physical features of input data. It is called low level classification. Data classification by considering not only physical attributes but also pattern formation is denominated high level classification. We present here an undergraduate research that proposes a new high level classification technique that calculates the network entropies before and after the insertion of a data item to be classified. Then, we classify it as belonging to the class which results in the largest increase of the entropy. Our method can execute classification tasks according to both similarity and pattern formation of input data. In summary, this technique calculates how significant a data item is for each class performing a new way to classify data.*

Resumo. *Técnicas tradicionais de classificação que baseiam-se apenas em características físicas dos dados são chamadas de classificação de baixo nível. Se consideram, além dos atributos físicos, o padrão de formação, chamamos de classificação de alto nível. Apresenta-se aqui o projeto de iniciação científica que propõe o desenvolvimento de uma nova técnica de classificação de alto nível baseada na medição das entropias da rede antes e depois da inserção de um item a ser classificado. Este item é classificado como pertencente à classe que resultar o maior aumento nas medições. O método pode classificar os dados por sua similaridade e padrão de formação. Em resumo, esta técnica calcula a importância do dado para cada uma das classes.*

1. Introdução

Técnicas de aprendizado de máquina são capazes de construir modelos que podem organizar algum conhecimento ou imitar certo comportamento humano através de representação computacional dos dados obtidos de diversos domínios [Bishop 2006]. Tais técnicas são tradicionalmente dividas em duas classes principais: aprendizado supervisionado e aprendizado não-supervisionado [Mitchell 1997].

O aprendizado supervisionado tem como objetivo obter conhecimento de amostras rotuladas de classes conhecidas. Neste caso, a técnica constrói um mapa entrada-saída baseado na observação dos dados de treinamento. Quando os rótulos são constituídos por valores discretos, o problema é denominado *classificação* e, quando contínuos, *regressão*. No aprendizado não-supervisionado, a principal tarefa é agrupar os dados por algum critério de similaridade. Em seu processo de aprendizado, são analisados apenas as características dos dados, pois não há conhecimento a priori de suas classes [Mitchell 1997, Russell e Norvig 2003].

Redes neurais artificiais, máquina de vetores de suporte e k-vizinhos mais próximos são exemplos de técnicas tradicionais de classificação em que as características físicas dos dados, como distância, similaridade ou densidade, são utilizadas no processo de aprendizagem [Silva e Zhao 2012].

Outras pesquisas buscam considerar no processo de aprendizagem, além das características físicas dos dados, as características de formação dos padrões, que possuem significado semântico. Entre elas, pode-se citar a técnica de co-treinamento e a web semântica [Silva e Zhao 2012].

Classificação baseada nas características físicas dos dados, mas não os padrões de formação da classe, é chamada de classificação de baixo nível, enquanto a classificação que considera, além das características físicas dos dados, o padrão de formação destes, é dita classificação de alto nível [Silva e Zhao 2012].

Este artigo apresenta o desenvolvimento do projeto de iniciação científica, desenvolvida pelo aluno Filipe Alves Neto sob a orientação do Professor Doutor Zhao Liang, que propõe uma nova técnica de classificação de dados de alto nível baseada na extração de características topológicas e dinâmicas de redes complexas através de suas entropias.

A Seção 2 deste artigo elucida as motivações para a escolha do projeto. A Seção 3 apresenta os objetivos do trabalho. A Seção 4 aborda os principais conceitos e técnicas para o entendimento deste projeto. O desenvolvimento do projeto é apresentado na Seção 5. As seções 6 e 7 descrevem, respectivamente, os resultados e as conclusões obtidas durante o desenvolvimento deste.

2. Motivações

As técnicas de classificação de dados, quando aplicadas a conjuntos de dados reais, devem lidar com diversas dificuldades, como ruídos, erros e variações dos dados [Wood 1996]. Uma imagem, por exemplo, pode apresentar variações de rotação, translação ou escala [Tang 2009]. A literatura fornece uma série de técnicas para lidar com estes problemas clássicos.

Entretanto, em alguns casos, apenas a observação das características físicas dos dados não é suficiente para obter bons resultados na classificação dos dados.

A tarefa exemplificada pela Figura 1 é um exemplo destes casos. Pelo padrão visual, conseguimos sem dificuldade diferenciar as duas classes: os pontos verdes marcados por "+" formam um "quadrado", enquanto os pontos azuis marcados por "x" formam uma "circunferência". Deste modo, podemos classificar os pontos vermelhos marcados por "*" como pertencentes à classe da "circunferência". Mas uma técnica tradicional, provavelmente, classificaria estes dados erroneamente, pois cada ponto vermelho possui maior similaridade física com os pontos pertencentes à classe do "quadrado".

Apesar dos estudos relacionados à classificação de alto nível, ainda não há um esquema explícito e geral para lidar com estes problemas na literatura [Silva e Zhao 2012].

3. Objetivos

O objetivo deste projeto é desenvolver e aprimorar nova uma técnica de classificação de dados de alto nível. A técnica deve ser capaz de analisar as características locais e globais

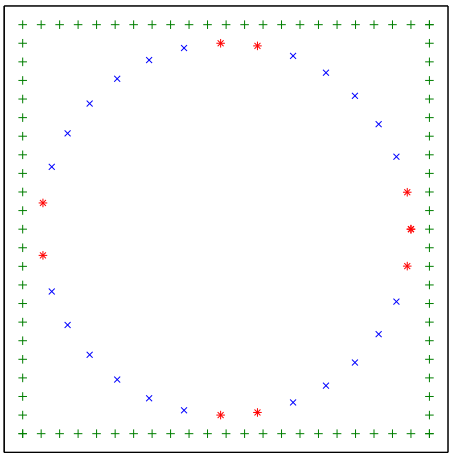


Figura 1. Exemplo de uma tarefa de classificação onde os pontos vermelhos marcados por "*" devem ser classificados em uma das outras duas classes, "+" ou "x".

dos dados fornecidos e com isso obter bons resultados no processo de classificação. A avaliação quantitativa da técnica é feita através da aplicação em bases de dados reais e artificiais.

4. Conceitos e Técnicas Relevantes

Nesta seção, serão apresentados conceitos e técnicas relevantes utilizados ao longo do desenvolvimento do projeto.

4.1. Redes Complexas

Redes complexas são grafos em larga escala com um padrão de conexões não trivial [Boccaletti et al. 2006]. Neste artigo, consideramos os termos grafo e rede como equivalentes.

Diversos fenômenos naturais podem ser representados por uma rede, como estruturas cerebrais, interações sociais e a internet. Em uma rede, cada elemento é representado por um vértice e as conexões entre estes elementos são representadas por arestas [Costa et al. 2005]. Nas estruturas cerebrais, por exemplo, os vértices representam os neurônios e as arestas as conexões físicas entre estes.

4.2. Caminhada Aleatória e Cadeia de Markov

Caminhada aleatória é um processo dinâmico fundamental [Noh e Rieger 2004]. Dada uma rede e um ponto inicial, seleciona-se um vizinho aleatoriamente e move-se até este vizinho; então seleciona-se aleatoriamente um vizinho deste último vértice selecionado e mova-se até este; e assim por diante. A sequência de pontos selecionados é chamada de caminhada aleatória [Lovász 1996].

Cadeia de Markov é um processo estocástico definido por um número contável ou finito de estados onde a probabilidade condicional do processo estar em dado estado no tempo atual depende apenas do estado anterior e não da sequência de estados precedentes a este [Ross 2007]. A matriz formada pelos valores destas probabilidades é chamada de matriz estocástica do processo.

A caminhada aleatória é uma cadeia de Markov finita reversível no tempo, ou seja, toda cadeia de Markov pode ser interpretada como um passeio aleatório em uma rede direcionada com pesos nas arestas [Lovász 1996].

4.3. Entropia da Rede

Entropia da rede pode ser expressa como a entropia de Kolmogorov-Sinai [Kolmogorov 1958] da matriz estocástica associada à matrix de adjacência da rede. Demetrius e Manke (2005) apresentam evidências que esta medida está quantitativamente relacionada com a capacidade da rede suportar modificações aleatórias em sua estrutura.

Além disso, a entropia da rede também caracteriza a multiplicidade dos caminhos internos. Deste modo, esta medida é inversamente proporcional à média dos caminhos mais curtos [Demetrius e Manke 2005].

Como consequência destas propriedades, a entropia da rede captura as características locais e globais dos dados, fazendo desta medida ideal para para a classificação de alto nível.

4.4. Classificação de Dados de Alto Nível baseada em Redes

Silva e Zhao (2012) propõem uma técnica de classificação de alto nível onde a rede caracteriza o padrão de formação dos dados, explorando suas propriedades topológicas. O processo de classificação, inicialmente, constrói uma rede para cada classe. Em seguida, o processo verifica a conformidade entre o padrão de formação da rede e o dado a ser classificado. Esta conformidade é verificada através da comparação de medidas da rede antes e depois da inserção do dado na rede.

Para a formação da rede, são utilizadas as técnicas ϵ-radius e k-vizinhos mais próximos. As medidas utilizadas na comparação das redes são grau médio, coeficiente de clusterização e assortatividade [Silva e Zhao 2012].

5. Desenvolvimento

A técnica proposta nesse projeto analisa os padrões encontrados nas classes através de redes formadas pelos dados de entrada. O processo de classificação da técnica aperfeiçoado ao longo do projeto será abordado nesta seção.

Considere o conjunto de treinamento denotado como $X_{treinamento} = \{(x_1, y_1), ..., (x_n, y_n)\}$, onde o primeiro componente da i-ésima tupla denota os atributos do i-ésimo item de dado. Se existirem d atributos em cada item de dado, dizemos que o conjunto de dados é d-dimensional. O segundo componente da i-ésima tupla caracteriza o rótulo da classe associado ao i-ésimo item de dado. O objetivo no aprendizado de máquina é construir um mapa de x para y. Este mapa é chamado classificador. O classificador construído pode ser verificado através de um conjunto de teste $X_{teste} = \{x_1, ..., x_m\}$, em que os rótulos não são providos e $X_{treinamento} \cap X_{teste} = \emptyset$. Em nosso problema, os rótulos são valores discretos - $y_i \in \{1, ..., L\}, \forall i \in \{1, ..., n\}$ -, então o problema é chamado classificação.

Com base nessas definições, o processo de classificação dos dados é realizado em três etapas. Na primeira fase, são construídas as redes que representarão as classes e

cada um dos itens de dados a serem classificados. Na etapa seguinte, serão calculadas as entropias de cada uma das redes previamente construídas. Por fim, com estas informações, os itens de dados são classificados.

5.1. Fase de Formação das Redes

No primeiro passo do processo de classificação, são construídas redes para cada uma das classes. Estas redes devem ser não direcionadas, com pesos nas arestas e com um único componente conectado. Denomina-se G_l a rede que representa a classe $l \in \{1, ..., L\}$.

Para a construção das redes G_l, cada item de dado do conjunto de treinamento, $X_{treinamento}$, pertencente à classe l é representado como um vértice na rede. Chamamos de $Z(G) = \{z_i\}$ o conjunto de vetores de características, z_i, que representam cada um dos vértices da rede G. Neste caso, $Z(G_l) = \{x_i | x_i \in X_{treinamento}$ e $y_i = l\}$. Haverá uma conexão entre dois vértices se a distância euclidiana entre os vetores de características dos itens de dados associados à estes for menor que determinado valor limite. Além disso, os pesos das arestas devem ser inversamente proporcionais à distância dos vértices. Então, o peso entre dois vértices, i e j, é calculado por

$$a_{ij} = a_{ji} = \begin{cases} 1 & \text{if} \quad ||x_i - x_j|| = 0 \\ (\frac{1}{||x_i - x_j||})^{\alpha} & \text{if} \quad 0 < ||x_i - x_j|| \leq \epsilon \\ 0 & \text{if} \quad ||x_i - x_j|| > \epsilon \end{cases} \tag{1}$$

Deste modo, a rede construída, representada pela matriz de adjacência $A = (a_{ij})$, preserva a similaridade e a topologia dos dados. O parâmetro $\epsilon > 0$ age como o valor limite da distância entre os vértices, enquanto o parâmetro $\alpha \geq 0$ indica o quanto a distância entre os vértices serão consideradas, isto é, se $a \to 0$, o peso das arestas tendem a ter valores binários. Como a rede deve conter um único componente conectado, deve-se escolher um valor para o parâmetro ϵ que reflita esta condição.

Em seguida, são construídas as redes $G_l^{(i)}, \forall l \in \{1, ..., L\}$ e $\forall i \in \{1, ..., u\}$, tal que $Z(G_l^{(i)}) = Z(G_l) \cup \{x_i | x_i \in X_{teste}\}$, seguindo a Equação 1, exceto para os casos em que o vértice extra não possui conexões com os demais. Nestes casos, considera-se $G_l^{(i)} = G_l$.

5.2. Fase de Cálculo das Entropias

Nesta fase, a entropia de cada uma das redes é calculada. Para isto, é necessário obter a matrix estocástica $P = (p_{ij})$ associada a cada uma das redes.

Seja λ o autovalor dominante da matriz A e (v_i) o autovetor correspondente, a matriz estocástica é definida por

$$p_{ij} = \frac{a_{ij} v_j}{\lambda v_i} \tag{2}$$

A partir deste resultado, a entropia da rede é a entropia dinâmica, $H(P)$, do processo estocástico definido pela matriz P, definida por

$$H(P) = \sum_i \pi_i H_i, \text{ onde } H_i = -\sum_j p_{ij} \log p_{ij} \tag{3}$$

Onde H_i é a entropia de Shannon da distribuição $[p_{i1}, ..., p_{iN}]$ e H é a média de todos os estados estacionários [Demetrius e Manke 2005]. Para a próxima etapa do processo de classificação, denomina-se $H(G)$ a entropia da rede G.

5.3. Fase de Decisão

Por fim, define-se o classificador fuzzy C que decide qual classe determinado item de dado pertence. A hipótese implícita no classificador é: se o i-ésimo item de dado do conjunto de teste, X_{teste}, pertence à classe $l \in \{1, ..., L\}$, o aumento proporcional da entropia $\delta_l^{(i)}$, é máximo. Esta ideia é justificada pela propriedade da rede de lidar com modificações aleatórias, deste modo, um item de dado que não pertence à certa classe não gerará grandes mudanças na rede que representa esta classe. Como conclusão, o classificador C é capaz de analisar a importância do dado para cada classe. Matematicamente, a probabilidade $C(i, l)$ do i-ésimo item de dado do conjunto de teste pertencer à classe $l \in \{1, ..., L\}$ é dada por

$$C(i,l) = \frac{\delta_l^{(i)}}{\sum_{k=1}^{L} \delta_k^{(i)}}, \text{ onde } \delta_l^{(i)} = \frac{H(G_l^{(i)})}{H(G_l)} \tag{4}$$

Os valores de $C(i, l)$ variam entre 0 e 1, e $\sum_{k=1}^{L} C(i, k) = 1$.

6. Resultados

Nesta seção, serão apresentadas duas simulações da aplicação da técnica proposta em uma base de dados real e em uma artificial. Na primeira simulação, utiliza-se a base Iris [Frank e Asuncion 2010], que é amplamente estudada na literatura. Em seguida, a técnica é testada em uma base de dados artificiais que simula uma tarefa difícil para classificação de baixo nível.

6.1. Simulação com a Base Iris

A base Iris contém 3 classes, Iris Setosa, Iris Versicolour e Iris Virginica, com 50 instâncias cada. As classes referem-se a um tipo de planta. Uma das classes é linearmente separável das demais, enquanto as outras duas classes não são linearmente separáveis entre si [Frank e Asuncion 2010]. Nesta simulação, a técnica proposta foi testada nesta base.

A Figura 2 mostra os resultados da simulação com os parâmetros ϵ variando entre 1 e 2, e o parâmetro $\alpha = 0, 1, 2$. Neste caso, melhores resultados são obtidos com valores mais baixos dos parâmetros, chegando à 96% de acurácia em validação cruzada 10-fold. É importante destacar que valores $\epsilon < 1$ na construção das redes não geram um único componente conectado nesta base.

Para destacar os detalhes do processo de classificação, a Figura 3 mostra o aumento proporcional da entropia da rede, $\delta_l^{(i)}$, causado pela inserção dos itens de dados do conjunto de teste em cada uma das redes que representam as classes. As redes foram construídas com 66% dos dados no conjunto de treinamento e os parâmetros utilizados foram: $\epsilon = 1$ e $\alpha = 0$.

A Figura 3(a) mostra em detalhes a proporção entre os valores da entropia antes e depois da inserção dos itens da classe Iris Setosa do conjunto de teste nas redes de cada

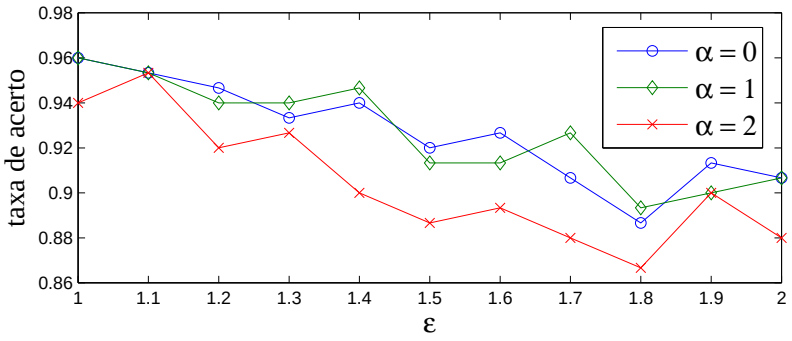
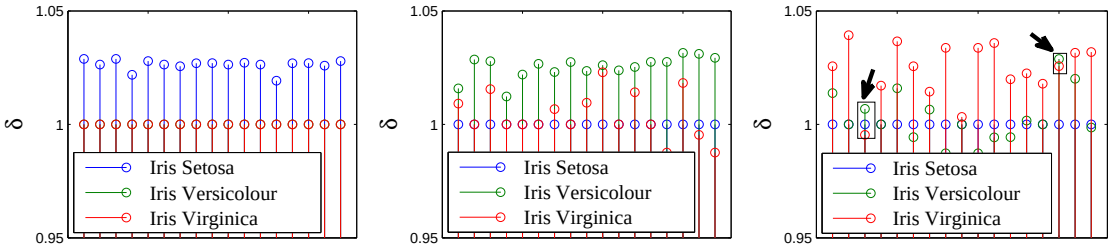


Figura 2. Uma análise da acurácia do classificador em diferentes valores dos parâmetros ϵ e α aplicados na base Iris utilizando o método de validação 10-fold. Foram escolhidos três valores diferentes para α. Na coordenada horizontal variam os valores para o parâmetro ϵ, enquanto na coordenada vertical encontram-se a acurácia obtida em cada caso.



(a) Itens da classe Iris Setosa. (b) Itens da classe Iris Versicolour. (c) Itens da classe Iris Virginica.

Figura 3. Comparação do aumento proporcional da entropia da rede para cada classe ao inserir itens de dados do conjunto de teste. As redes que representam as classes foram construídas com 66% dos dados e com os parâmetros $\epsilon = 1$ e $\alpha = 0$. A Figura 3(a) mostra a inserção dos itens da classe Iris Setosa, a Figura 3(b) dos itens da classe Iris Versicolour e a Figura 3(c) os itens da classe Iris Virginica. Pontos azuis indicam os valores de $\delta^{(i)}_{IrisSetosa}, \forall i$. Pontos verdes e vermelhos indicam, respectivamente, $\delta^{(i)}_{IrisVersicolour}$ e $\delta^{(i)}_{IrisVirginica}, \forall i$.

classe. Nesta figura, observa-se que estes dados aumentam a entropia apenas na rede de mesma classe. Isto acontece porque a classe Iris Setosa é linearmente separável das demais classes. Similarmente, o aumento da entropia causado pela inserção dos itens da classe Iris Versicolour do conjunto de teste é exemplificado na Figura 3(b). Finalmente, a Figura 3(c) apresenta os resultados da inserção dos itens da classe Iris Virginica do conjunto de teste. Enfatizados nesta última figura, encontram-se os itens de dados que foram classificados incorretamente como pertencentes à classe Iris Versicolour.

6.2. Simulação com Grades Parcialmente Sobrepostas

O método de classificação proposto é testado, nesta simulação, em uma base de dados artificial em que técnicas tradicionais não são capazes de extrair informações suficientes para classificar corretamente os dados.

A base de dados artificial, descrita pela Figura 4, consiste em duas classes que formam duas grades parcialmente sobrepostas. Aproximadamente 4% das grades se in-

terseccionam.

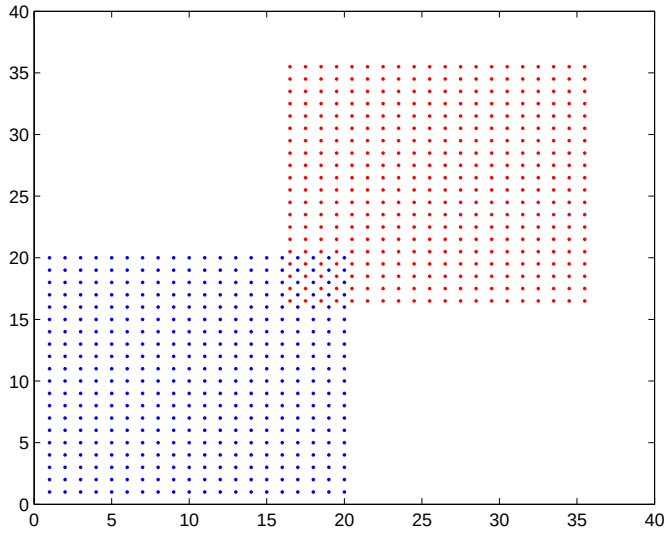


Figura 4. Representação da base de dados artificial construída para avaliar a técnica proposta neste projeto. A base contem duas classes bidimensionais que se intersectam. As classes possuem padrão de formação similar, de fato, ambas as classes são grades com distância 1 entre os itens adjacentes, mas a classe "vermelha" possui itens deslocados em relação aos itens da classe "azul".

Apesar da dificuldade deste problema, uma vez que na região da intersecção os dados possuem mais vizinhos da outra classe, a técnica proposta obteve melhores resultados que as outras técnicas. A Figura 5 demonstra quão bem cada uma das técnicas classificou os dados com diferentes quantidades de dados na conjunto de treinamento. As técnicas tradicionais utilizadas aqui foram Perceptron Multi-camadas (MLP) e Máquina de Vetores de Suporte (SVM).

O classificador MLP obteve acurácia entre 96.3% e 97.5%, enquanto o SVM obteve taxa de acerto entre 90.4% e 96.3%. É importante mencionar que estes valores de acurácia não são bons, pois mais de 92% dos dados neste problema podem ser considerados fáceis de serem classificados.

A técnica proposta, em contraste, pode classificar corretamente mais de 98% dos dados do conjunto de teste utilizando apenas 40% dos dos dados no conjunto de treinamento. Além disso, obteve 100% de acurácia com 90% dos itens no conjunto de treinamento.

7. Conclusões

Neste artigo, foi apresentado o projeto de iniciação científica que propõe o desenvolvimento de uma nova técnica de classificação de dados de alto nível baseada em entropia da rede. Este novo método desenvolvido classifica dados baseado na importância do item de dado na rede que representa uma determinada classe através de suas características físicas e semânticas.

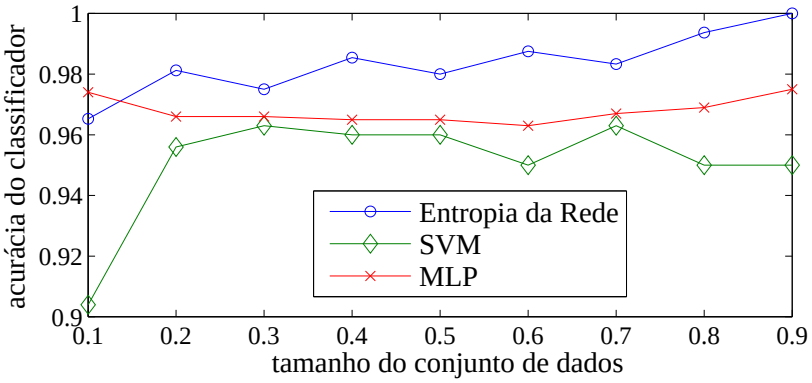


Figura 5. Um estudo detalhado da performance da técnica proposta comparada a outras técnicas tradicionais aplicadas na base de dados artificiais descrita na Figura 4 variando o tamanho do conjunto de treinamento. A linha azul com circunferências indicam os resultados da técnica proposta com parâmetros $\epsilon = \infty$ e $\alpha = 1$. A linha vermelha com x mostra a performance atingida pelo classificador MLP com 4 neurônios na camada de saída, com 500 épocas e taxa de aprendizado 0.3. A linha verde com losangos demonstra os resultados do classificador SVM com base radial.

Em relação à técnica, a escolha dos parâmetros da fase de construção da rede, ϵ e α, é crucial para que o classificador obtenha bons resultados. O parâmetro ϵ influencia o número de conexões entre os vértices da rede, enquanto o parâmetro α decide a importância da distância entre os vértices conectados. Por exemplo, a escolha dos parâmetros na simulação com a base Iris foi $\epsilon = 1$ e $\alpha = 0$ resultou na construção de redes binárias com poucas conexões entre os vértices.

Além disso, a situação apresentada na Subseção 6.2 demonstra a ineficiência de técnicas tradicionais quando a similaridade entre os dados não é suficiente para classificá-los, uma vez que estas técnicas consideram apenas aspectos locais e não globais dos dados. Este problema é facilmente resolvido quando aplicado a técnica proposta. Os próximos passos do desenvolvimento deste projeto consistem em reconhecer e estudar aplicações reais equivalentes a este caso artificial.

Esta técnica, porém, possui algumas limitações, entre elas, o método não é escalável para grandes bases de dados e pode não funcionar em bases desbalanceadas. Pretende-se, em trabalhos futuros, aprimorar a técnica desenvolvida de modo a diminuir estas limitações.

Concluindo, este trabalho mostra-se em vias de aprimorar os resultados atuais das aplicações de classificação de dados reais apresentando inovações para a teoria de aprendizado de máquina.

8. Agradecimentos

Este trabalho é apoiado pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., e Hwang, D. (2006). Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308.

Costa, L. F., Rodrigues, F. A., Travieso, G., e Boas, P. R. V. (2005). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242.

Demetrius, L. e Manke, T. (2005). Robustness and network evolution an entropic principle. *Physica A: Statistical Mechanics and its Applications*, 346(3):682–696.

Frank, A. e Asuncion, A. (2010). UCI machine learning repository.

Kolmogorov, A. N. (1958). A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk SSSR (*N.S.*)*, 119:861–864.

Lovász, L. (1996). Random walks on graphs: A survey. In Miklós, D., Sós, V. T., e Szőnyi, T., editors, *Combinatorics, Paul Erdős is Eighty*, volume 2, pages 353–398. János Bolyai Mathematical Society, Budapest.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition.

Noh, J. D. e Rieger, H. (2004). Random walks on complex networks. *Physical Review Letters*, 92:118701.

Ross, S. (2007). *Introduction to Probability Models*. Academic Press.

Russell, S. J. e Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education.

Silva, T. C. e Zhao, L. (2012). Network-based high level data classification. *Neural Networks and Learning Systems, IEEE Transactions on*, 23(6):954–970.

Tang, Y. (2009). *Wavelet theory approach to pattern recognition*. Series in machine perception and artificial intelligence. World Scientific.

Wood, J. (1996). Invariant pattern recognition: a review. *Pattern Recognition*, 29(1):1–17.