UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA
2011.2

Análise comparativa de algoritmos de seleção de protótipos em bases desbalanceadas

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluno: Dayvid Victor Rodrigues de Oliveira (dvro@cin.ufpe.br)
Orientador: George Darmiton da Cunha Cavalcanti (gdcc@cin.ufpe.br)

Recife,13 de Setembro de 2011.

Contexto

Classificadores podem ser definidos como algoritmos que identificam a classe de uma instância baseando-se em suas características. O KNN (*k-Nearest Neighbor*) [1] é um classificador muito popular por ser simples e eficiente. Todavia, o KNN pode ser muito dispendioso, pois utiliza toda a base de dados para fazer uma classificação, e em muitas aplicações o tamanho da base pode tornar o KNN inviável. Com isso, surgiu a ideia de utilizar um conjunto menor, gerado a partir da base de dados original, para treinamento do classificador. O processo de geração deste conjunto é chamado de seleção de protótipos.

Com a seleção de protótipos, pode-se utilizar o *Nearest Prototype Classification*, que consiste na aplicação do KNN, utilizando apenas os protótipos para fazer uma classificação. Assim, a base de dados é reduzida, diminuindo o espaço de armazenamento e o tempo de processamento.

Em várias situações do mundo real, os classificadores precisam ser treinados com bases desbalanceadas, que possuem grande diferença entre a quantidade de instâncias de diferentes classes. Diante de tais bases, técnicas de seleção de protótipos podem gerar instâncias que não representam bem a base original, induzindo o classificador a erros [8].

Existem vários algoritmos de seleção de protótipos, e eles podem ser de síntese ou seleção, determinísticos ou não determinísticos, entre outras características. Entre os algoritmos mais conhecidos estão o ENN (*Edited Nearest Neighbor*) [2], CNN (*Condensed Nearest Neighbor*) [3] , *Tomek Links* [4], OSS (*One-Side Selection*) [5], LVQ (*Learning Vector Quantization*) [6], SGP (*Self-Generating Prototypes*) [7] e CCNN (*Class Conditional Nearest Neighbor*).

A forma como a seleção é feita afeta a disposição dos protótipos, podendo o algoritmo eliminar instâncias de fronteira, remover dados redundantes ou reduzir ruídos. É necessário que seja feita uma avaliação das diferentes técnicas, a fim de se saber quais são mais apropriadas para o caso específico de bases desbalanceadas.

Objetivos

Desenvolver e comparar as principais técnicas de seleção de protótipos em bases desbalanceadas, mostrando as vantagens e desvantagens de acordo com o propósito da aplicação.

Inicialmente, os métodos utilizados no trabalho serão: ENN (*Edited Nearest Neighbor*), CNN (*Condensed Nearest Neighbor*), Tomek Links, OSS (*One-Side Selection*), LVQ (*Learning Vector Quantization*), SGP (*Self-Generating Prototypes*) e CCNN (*Class Conditional Nearest Neighbor*).

Como cada uma das técnicas citadas possui características diferentes, poderão ser identificadas quais abordagens são ideais para cada tipo de aplicação com bases desbalanceadas. Os resultados obtidos poderão ser utilizados para elaboração de novas técnicas de seleção de protótipos.

Cronograma

	Mês															
Atividade	Setembro				Outrubro				Novembro				Dezembro			
1. Levantamento e estudo do material bibliográfico.																
2. Levantamento e estudo dos desafios a serem abordados.																
3. Implementação																
4. Escrita do Relatório Final																
5. Elaboração da apresentação oral																
6. Defesa do TG																

Referências

[1] T.M. Cover, P.E. Hart, "Nearest Neighbor Pattern Classification", *IEEE Transactions on Information Theory*, vol. IT-13, No.1, 1967, pp.21-27

[2] Binay K. Bhattacharya, Ronald S. Poulsen, Godfried T. Toussaint, "Aplication of Proximity Graphs to Editing Nearest Neighbor Decision Rule". *International Symposium on Information Theory*, Santa Monica, 1981.

[3] Hart, P. E. (1968) "The Condensed Nearest Neighbor Rule". *IEEE Transactions on Information Theory* IT-14, pp. 515-516.

[4] Tomek, I. Two Modifications of CNN. *IEEE Transactions on Systems Man and Communications* SMC-6 (1976), 769 772.

[5] Kubat, M., and Matwin, S. (1997) "Addressing the Course of Imbalanced training Sets: One-sided Selection". In *ICML*, pp. 179-186.

[6] T. Kohonen. "Improved Versions of Learning Vector Quantization". In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, vol.1, pags. 545-550, San Diego, California, Junho, 1990.

[7] Fayed, H. A., Hashem, S. R. and Atiya, A. F. (2007). "Self-generating prototypes for pattern classification". *Pattern Recognition*, vol. 40, pp 1498-1509.

[8] H. He, E.A. Garcia, "Learning from imbalanced data", *IEEE Transactions on Knowledge and Data Engineering* 21 (9) (2009) 1263–1284.

Assinaturas

_____

Aluno: Dayvid Victor Rodrigues de Oliveira

_____

Orientador: George Darmiton da Cunha Cavalcanti

Recife, 13 de setembro de 2011.