Zhang, Yongqing; **Zhang, Danling**; **Mi, Gang**; **Ma, Daichuan**; **Li, Gongbing**; **Guo, Yanzhi**; **Li, Menglong**; **Zhu, Min**

Using ensemble methods to deal with imbalanced data in predicting protein-protein interactions. (English) Zbl 1244.92023

Comput. Biol. Chem. 36, 36-41 (2012).

Summary: In proteins, the number of interacting pairs is usually much smaller than the number of non-interacting ones. So the imbalanced data problem will arise in the field of protein-protein interactions (PPIs) prediction. We introduce two ensemble methods to solve the imbalanced data problem. These ensemble methods combine the based-cluster under-sampling technique and the fusion classifiers. Then we evaluate the ensemble methods using a data set from the Data Base of Interacting Proteins (DIP) with 10-fold cross validation. All the prediction models achieve the area under the receiver operating characteristic curve (AUC) value about 95%. Our results show that the ensemble classifiers are quite effective in predicting PPIs; we also gain some valuable conclusions on the performance of ensemble methods for PPIs in imbalanced data. The prediction software and all data sets employed in the work can be obtained for free at http://cic.scu.edu.cn/bioinformatics/Ensemble_PPIs/index.html.

MSC:

92C40 Biochemistry, molecular biology
62H30 Classification and discrimination; cluster analysis (statistical aspects)
62P10 Applications of statistics to biology and medical sciences; meta analysis

Software:

Cd-hit; Hum-mPLoc

Full Text: DOI

References:

[1] Bader, G.D.; Hogue, C.W.V., Analyzing yeast protein – protein interaction data obtained from different sources, Nature biotechnology, 20, 991-997, (2002)

[2] Breiman, L., Bagging predictors, Machine learning, 24, 123-140, (1996) · Zbl 0858.68080

[3] Charton, M.; Charton, B.I., The structural dependence of amino acid hydrophobicity parameters, Journal of theoretical biology, 99, 629-644, (1982)

[4] Chawla, N.V., SMOTE: synthetic minority over-sampling technique, Journal of artificial intelligence research, 16, 321-357, (2002) · Zbl 0994.68128

[5] Chen, X.; Gerlach, B.; Casasent, D., Pruning support vectors for imbalanced data classification, Proceedings of the international joint conference on neural networks, 1883-1888, (2005)

[6] Cohen, G., Learning from imbalanced data in surveillance of nosocomial infection, Artificial intelligence in medicine, 37, 7-18, (2006)

[7] Das, R.; Sengur, A., Evaluation of ensemble methods for diagnosing of valvular heart disease, Expert systems with applications, 37, 5110-5115, (2010)

[8] Elkan, C., The foundations of cost-sensitive learning, Citeseer, 973-978, (2001)

[9] Enright, A.J., Protein interaction maps for complete genomes based on gene fusion events, Nature, 402, 86-90, (1999)

[10] Gavin, A.C., Proteome survey reveals modularity of the yeast cell machinery, Nature, 440, 631-636, (2006)

[11] Grantham, R., Amino acid difference formula to help explain protein evolution, Science, 185, 862, (1974)

[12] Guo, Y., Using support vector machine combined with auto covariance to predict protein cprotein interactions from protein sequences, Nucleic acids research, 36, 3025, (2008)

[13] Han, J.D.J., Effect of sampling on topology predictions of protein – protein interaction networks, Nature biotechnology, 23, 839-844, (2005)

[14] Hansen, L.K.; Salamon, P., Neural network ensembles, IEEE transactions on pattern analysis and machine intelligence, 12, 993-1001, (1990)

[15] Hart, P., The condensed nearest neighbor rule, IEEE transactions on information theory, 14, 515-516, (1968)

[16] Hopp, T.P.; Woods, K.R., Prediction of protein antigenic determinants from amino acid sequences, Proceedings of the national Academy of sciences of the united states of America, 78, 3824, (1981)

[17] Horton, P., Wolf PSORT: protein localization predictor, Nucleic acids research, 35, W585, (2007)

[18] Huang, K., Learning classifiers from imbalanced data based on biased minimax probability machine, Computer vision and pattern recognition, 2, 558-563, (2004)

[19] Ito, T., A comprehensive two-hybrid analysis to explore the yeast protein interactome, Proceedings of the national Academy of sciences of the united states of America, 98, 4569, (2001)

[20] Jo, T.; Japkowicz, N., Class imbalances versus small disjuncts, ACM SIGKDD explorations newsletter, 6, 40-49, (2004)

[21] Kang, P.; Cho, S., EUS SVMs: ensemble of under-sampled SVMs for data imbalance problems, (2006), Springer, pp. 837-846

[22] Krigbaum, W.; Komoriya, A., Local interactions as a structure determinant for protein molecules: II, Biochimica et biophysica acta (BBA): protein structure, 576, 204-228, (1979)

[23] Krogan, N.J., Global landscape of protein complexes in the yeast \textit{saccharomyces cerevisiae}, Nature, 440, 637-643, (2006)

[24] Kubat, M.; Matwin, S., Addressing the curse of imbalanced training sets: one-sided selection, (1997), Citeseer, pp. 179-186

[25] Li, W.; Godzik, A., Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics, 22, 1658, (2006)

[26] Li, W.; Miao, D.; Wang, W., Two-level hierarchical combination method for text classification, Expert systems with applications, 38, 2030-2039, (2010)

[27] Lippmann, R.P., An introduction to computing with neural nets, Artificial neural network: theoretical concepts, 209, 36-54, (1987)

[28] Liu, J.J., Imbalanced expression of functionally different WT1 isoforms may contribute to sporadic unilateral wilms' tumor, Biochemical and biophysical research communications, 254, 197-199, (1999)

[29] Overbeek, R., Use of contiguity on the chromosome to predict functional coupling, In silico biology, 1, 93-108, (1998)

[30] Raskutti, B.; Kowalczyk, A., Extreme re-balancing for SVMs: a case study, ACM SIGKDD explorations newsletter, 6, 60-69, (2004)

[31] Rose, G.D., Hydrophobicity of amino acid residues in globular proteins, Science, 229, 834, (1985)

[32] Shen, H.B.; Chou, K.C., Hum-mploc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites, Biochemical and biophysical research communications, 355, 1006-1011, (2007)

[33] Tanford, C., Contribution of hydrophobic interactions to the stability of the globular conformation of proteins, Journal of the American chemical society, 84, 4240-4247, (1962)

[34] Tomek, I., Two modifications of CNN, IEEE transactions on systems, man and cybernetics, 6, 769-772, (1976) · Zbl 0341.68066

[35] Uetz, P., A comprehensive analysis of protein – protein interactions in \textit{saccharomyces cerevisiae}, Nature, 403, 623-627, (2000)

[36] Vapnik, V., The support vector method of function estimation, Nonlinear modeling: advanced black-box techniques, (1998), pp. 55-86

[37] Vladimir, V.N.; Vapnik, V., The nature of statistical learning theory, (1995), Springer · Zbl 0833.62008

[38] Wu, G.; Chang, E.Y., Class-boundary alignment for imbalanced dataset learning, (2003), Citeseer, pp. 49-56

[39] Xenarios, I., DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions, Nucleic acids research, 30, 303, (2002)

[40] Yen, S.J.; Lee, Y.S., Cluster-based under-sampling approaches for imbalanced data distributions, Expert systems with applications, 36, 5718-5727, (2009)

[41] Zhou, P., Genetic algorithm-based virtual screening of combinative mode for peptide/protein, Huaxue xuebao, 64, 691-697, (2006)