SUDAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE STUDIES

Effect of Clustering as a Preprocessing Step for Solving Unbalanced Data Set Problem

A Case Study: Protein Secondary Structure Prediction

تاثير التجميع بمثابة خطوة تجهيزية لحل مشكلة مجموعة البيانات غير المتوازنة

دراسة حالة: التنبؤ ببنية البروتين الثانوية

OCTOBER 2014

SUDAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

COLLEGE OF GRADUATE S TUDIES

Effect of Clustering as a Preprocessing Step for Solving Unbalanced Data set Problem.

Case Study: Protein Secondary Structure Prediction

A Thesis Submitted in Partial Fulfillment of the Requirements of Master Degree in Computer Science

BY:

ElhamMosaAbdAljalilMohmmed

SUPERVISOR

DR. MurtadaKhalafallahElbashir

OCTOBER 2014

Table of Content

List of Figure

List of Table

الآيـــــــــــــة

قال تعالي :(مُّحَمَّدٞ رَّسُولُ ٱللَّهِۚ وَٱلَّذِينَ مَعَهُۥٓ أَشِدَّآءُ عَلَى ٱلۡكُفَّار
رُحَمَآءُ بَيۡنَهُمۡۖ تَرَىٰهُمۡ رُكَّعٗا سُجَّدٗا يَبۡتَغُونَ فَضۡلٗا مِّنَ ٱللَّهِ وَرِضۡوَٰنٗاۖ سِيمَاهُمۡ
فِي وُجُوهِهِم مِّنۡ أَثَرِ ٱلسُّجُودِۚ ذَٰلِكَ مَثَلُهُمۡ فِي ٱلتَّوۡرَىٰةِۚ وَمَثَلُهُمۡ فِي
ٱلۡإِنجِيلِ كَزَرۡعٍ أَخۡرَجَ شَطۡـَٔهُۥ فَـَٔازَرَهُۥ فَٱسۡتَغۡلَظَ فَٱسۡتَوَىٰ
عَلَىٰ سُوقِهِۦ يُعۡجِبُ ٱلزُّرَّاعَ لِيَغِيظَ بِهِمُ ٱلۡكُفَّارَۗ وَعَدَ ٱللَّهُ ٱلَّذِينَ
ءَامَنُواْ وَعَمِلُواْ ٱلصَّٰلِحَٰتِ مِنۡهُم مَّغۡفِرَةٗ وَأَجۡرًا عَظِيمَۢا) الفتح 29

الحمـــــــــــــــــــد

الحمد لله حمداً كثيراً طيباً مباركاً فيه يليق بجلال وجهه وعظيم سلطانه. الحمد لله الذي بنعمته تتم الصالحات. احمد الله عز وجل أن وفقني إلي إتمام هذا البحث وأسال الله أن يجعله في ميزان حسناتي وان ينتفع به غيري وان يزدني علماً ...

DEDICATION

To my mother, the first person who care, teach me and for his prayers to me.

To my father, for care, support and his prayers to me

To My Sisters and Brothers

To My Friends

To My Colleagues (My batch)

ACKNOWLEDGEMENT

First, thanks to the supervisor Dr. Murtada Khalafalla Albasher Alfaki for his advice and He did not Stingy any information for me, also I thank Uztaz. Mohmmed Hammed AL Heber very much for his assistance and support, also I will never forget my sister Ustaza. AmelAWad.

Abstract

Protein secondary structure prediction from its sequence of amino acids remains an important issue. Determining the secondary structure of protein in the laboratory is very costly and consumes a lot of time. Development of precise and efficient method for secondary structure prediction is very important. In this research we propose an approach that uses the clustering algorithm as preprocessing steps for machine learning methods for solve unbalanced dataset problem to predict Protein secondary structure and compare the result when using the clustering algorithm, with the result without using it in the prediction. We utilize position specific scoring matrices (PSSMs) as features. The preprocessing for the data will be done using K-means clustering to prepare clusters that can be used as input for a support vector machines (SVM) and kernel logistic regression (KLR) models In this study we achieved high prediction accuracy compared by previous study Qtotal of 86.5%, 77.6%, on α-helix and coil secondary structure respectively when we used SVM method and also we achieved Qtotal of 82.18%, 75.3%and 82.9% on α-helix, coil and extended beta-sheet secondary structure respectively when we used KLR method .Achieves satisfactory performance in predicting secondary structure as measured by the Matthew’s correlation coefficient (MCC), Qpredicted and Qobserved on RS126 datasets.

المستخلص

يبقى التنبؤ ببنية البروتين الثانوية من سلسلة أحماضه الأمينية مسألة هامة .تحديد بنية البروتين الثانوية في المختبر أمر مكلف جدا ويستهلك الكثير من الوقت .تطوير طريقة دقيقة وفعالة للتنبؤ ببنية البروتين الثانوية مهم جدا .في هذا البحث نقترح نهجا يستخدم خوارزمية التجميع باعتبارها خطوات تجهيزية لأساليب تعلم الآلة لحل مشاكل مجموعة البيانات غير المتوازنة للتنبؤ بنية البروتين الثانوية ومقارنة النتيجة عند استخدام خوارزمية التجميع، مع النتيجة دون استخدامه في التنبؤ .نستخدم موضع محدد المصفوفات الدرجات (PSSMs)كسمات .وسوف يتم تجهيز البيانات باستخدام (K-means) لإعداد تجميع الكتل التي يمكن استخدامها كمدخل لنماذج آلات المتجهات (SVM) ونواة الانحدار اللوجستي (KLR) .في هذه الدراسة حققنا دقة التنبؤ عالية مقارنة بنتائج الدراسات السابقة وهي Qtotalمن 86.5٪، 77.6٪، على البنية الثانوية (α-helix) و(coil) على التوالي عندما استخدمنا طريقة SVMوأيضا حققنا Qtotalمن 82.18٪، 75.3٪ و 82.9٪ على البنية الثانوية (α-helix) ،(coil) و(extended beta-sheet) على التوالي عندما استخدمنا طريقة (KLR) . أداء مرضيا في توقع الهيكل الثانوي مقاسا بمعامل ارتباط ماثيو(MCC) ، Qpredicted و Qobserved على مجموعات البيانات RS126 .

References

[1] David, Barber (March 9, 2010) "Bayesian Reasoning and Machine Learning."

[2] MurtadaKhalafallah Elbashir1, Jianxin Wang1*, Fang-Xiang Wu2, Lusheng Wang (4-7 October 2012). "Predicting beta-turns in proteins using support vector machines with fractional polynomials." `.

[3] Simon, Tong ((2001)) "Support Vector Machine Active Learning with Applications to Text Classification" Journal of Machine Learning Research.

[4] Wang, J.-Y. (2002). "Application of Support Vector Machinesin Bioinformatics."

[5] Elbashir, M. K., et al. (2013). "Predicting β-Turns in Protein Using Kernel Logistic Regression."BioMed Research International 2013: 9.

[6] Rost and C. Sander. Prediction of protein secondary structure at better than70% accuracy. Journal of Molecular Biology, 232(2):584–599, 1993.

[7] Pang, Ning, Tan (2005, 2006). "Introduction to data mining”

[8] Singh, M. (2001). "Predicting Protein Secondary and Supersecondary Structure." CRC Press, LLC.

[9] Karypis, G. (July 2005). "Better Kernels and Coding Schemes Lead to Improvements in SVM-based Secondary Structure Prediction."

[10].Maher Maalouf, T. B. Trafalis, ”Robust weighted kernel logistic regression in imbalanced and rare events data,”’ Computational statistics and data analysis, 55, pp. 168–183, 2011.

[11]Wang, L. (2005). "Cancer Diagnosis and Protein Secondary Structure Prediction Using Support Vector Machines." Block S1, Nanyang Avenue, Singapore, 639798.

[12].Chin Yin Fai, R. H., and MohdSaberiMohamad (2012). "Optimized Local Protein Structure with Support Vector Machine to Predict Protein Secondary Structure."Springer-Verlag Berlin Heidelberg 2012.

[13]. Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. Journal of Molecular Biology 232:584–599

[15].Salamov and V. Solovyev. Prediction of protein structure by combining nearest-neighbor algorithms and multiple sequence alignments.J.MolBiol, 247:11{15,1995.

[16].D. Jones. Protein secondary structure prediction based on position-specific scoring matrices Journal of Molecular Biology, 292:195{202, 1999.

[17].King and M. Sternberg. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction.Protein Sci., 5:2298{2310, 1996.

[18] Hua S, Sun Z (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. Journal of molecular Biology 308:397–407.

[19] Maryam Alirezaee1, A. D., 3 and Eghbal Mansoori1 (November 2012). "PREDICTING THE SECONDARY STRUCTURE OF PROTEINS BY CASCADINGNEURALNETWORKS." International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.6, November 2012.

[20].CC C, CJ L, LIBSVM: A library for support vector machines.[http://www.

csie.ntu.edu.tw/~cjlin/libsvm].

[21] Rost, B., Sander, C.: Combining evolutionary information and neural networks to predict protein secondary structure. Proteins: Struct.,Funct., Genet. 19, 55–72 (1994).

[22]MurtadaKhalafallahElbashir, J. W., Fang-Xiang Wu, Min Li. (2012). "Sparse Kernel Logistic Regression for β-turnsPrediction." 2012 IEEE 6th International Conference on Systems Biology (ISB).

[23].Hyun-Chul Kim, S. P., Hong-Mo Je, Daijin Kim∗, Sung Yang Bang (2003). "Constructing support vector machine ensemble."Pattern Recognition Society.

[24].Brunak S, Chauvin Y, Andersen C, Nielsen H: Assessing the accuracy of prediction algorithms: an overview. Bioinformatics 2000, 16:412-424

[25] Schapire, R. (2008) "Theoretical Machine Learning."

[26] Eyal, E. (May 2011). "Secondary structure assignment and prediction."