



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

تأثير 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:

ElhamMosaAbdAljalilMohammed

SUPERVISOR

DR. MurtadaKhalafallahElbashir

OCTOBER 2014

Table of Content

Table of Content.....	iii
List of Figure.....	vii
List of Table.....	vii
الأيـــــــــــــــة.....	viii
الــــــــــــــــــــد.....	ix
DEDICATION	x
ACKNOWLEDGEMENT	xii
Abstract.....	xiii
المستخلص	xiii
References.....	xiv
CHAPTER ONE.....	1
INTRODUCTION.....	1
1.1 Introduction.....	2
1.2 Research scope	4
1.3 Problem statement	4
1.4 Objective of the research.....	4
1.5 Research methodology and tools	5
1.6 Research questions	5
1.7 Research organization.....	5
CHAPTER TWO	6
LITERATURE REVIEW AND RELATED WORK.....	6
2.1.1 Machine Learning	7
2.1.1Supervised learning	7
2.1.1 Unsupervised learning	8

2.1.1 Utility and Loss	8
2.1.1 Zero-one loss	8
2.1.1 General loss functions.....	9
2.1.2 Support Vector Machines.....	Error! Bookmark not defined.
2.1.2 SVMs for Induction	Error! Bookmark not defined.
2.1.3 Kernel logistic regression.....	Error! Bookmark not defined.
2.1.4 K-means Clustering Algorithm	Error! Bookmark not defined.
2.1.4 Assigning points to the closest centroid.....	Error! Bookmark not defined.
2.1.4 Centroid and objective function	Error! Bookmark not defined.
2.1.5 Majority Voting Vector.....	Error! Bookmark not defined.
2.1.6 The Biological Background of the PSSP	Error! Bookmark not defined.
2.1.7 Types of Protein Secondary Structures	Error! Bookmark not defined.
2.1.8 Secondary structure.....	Error! Bookmark not defined.
2.1.9 DSSP algorithm	Error! Bookmark not defined.
2.1.10 STRIDE.....	Error! Bookmark not defined.
2.1.11 DEFINE	Error! Bookmark not defined.
2.1.12 PSSM Generation	Error! Bookmark not defined.
2.2 Related work	Error! Bookmark not defined.
CHAPTER THREE.....	Error! Bookmark not defined.
MATERIAL AND METHODS	Error! Bookmark not defined.
3.1 RS126 Dataset	Error! Bookmark not defined.
3.2 Position Specific Scoring Matrices (PSSMs).....	Error! Bookmark not defined.
3.3 Support Vector Machine (SVM)	Error! Bookmark not defined.
3.3.1 Training and testing	Error! Bookmark not defined.
3.4 kernel Logistic Regression (KLR).....	Error! Bookmark not defined.
3.4.1 Training and testing	Error! Bookmark not defined.

3.5 The proposed approach..... **Error! Bookmark not defined.**
3.6 Clustered model..... **Error! Bookmark not defined.**
3.7 Performance measures **Error! Bookmark not defined.**
CHAPTER FOURE..... **Error! Bookmark not defined.**
RESULT AND DISCUSSION **Error! Bookmark not defined.**
Results and discussion **Error! Bookmark not defined.**
Conclusions and Recommendation **Error! Bookmark not defined.**

List of Figure

- Figure 1: Proteins are polymers of amino acids. **Error! Bookmark not defined.**
- Figure 2: A simple linear support vector machine **Error! Bookmark not defined.**
- Figure 3: Three types of protein secondary structures: α -helix, β -strand, and coil **Error! Bookmark not defined.**
- Figure4 : The architecture of the SVM and KLR inputs in form of PSSMs. **Error! Bookmark not defined.**
- Figure5 : general framework of our proposed approach..... **Error! Bookmark not defined.**
- Figure6 : framework of our proposed approach with clustering as preprocessing step **Error! Bookmark not defined.**
- Figure7 : The ROC curve on KLR without clustering..... **Error! Bookmark not defined.**
- Figure8 : The ROC curve on KLR without clustering..... **Error! Bookmark not defined.**
- Figure9 : The ROC curve of SVM without clustering..... **Error! Bookmark not defined.**
- Figure 10: ROC curve of SVM when using the Clustering..... **Error! Bookmark not defined.**

List of Table

Table 1 : Description of Symbol of the sum of the squared error (SSE)**Error! Bookmark not defined.**

Table 2: List of 126 Proteins in RS126 dataset used in secondary structure prediction**Error! Bookmark not defined.**

Table 3 : Result of Prediction for beta-sheet using different slide windows on SVM**Error! Bookmark not defined.**

Table 4: Result of Prediction for coil using different slide windows on SVM**Error! Bookmark not defined.**

Table 5 : Result of Prediction for α - helix using different slide windows on SVM**Error! Bookmark not defined.**

Table 6: Result of Prediction for coil using different slide windows with clustering on SVM**Error! Bookmark not defined.**

Table 7 Result of Prediction for α –helix using different slide windows with clustering on SVM**Error! Bookmark not defined.**

Table 8: Result of Prediction for α –helix using different slide windows on KLR**Error! Bookmark not defined.**

Table 9 : Result of Prediction for α –helix using different slide windows with clustering on KLR**Error! Bookmark not defined.**

Table 10: Result of Prediction for coil using different slide windows on KLR**Error! Bookmark not defined.**

Table 11: Result of Prediction for coil using different slide windows with clustering on KLR....**Error! Bookmark not defined.**

Table 12: Result of Prediction for Beta-sheet using different slide windows on KLR**Error! Bookmark not defined.**

Table 13: Result of Prediction for Beta-sheet using different slide windows with clustering on KLR.....**Error! Bookmark not defined.**

الآية

قال تعالى: (مُحَمَّدٌ رَسُولُ اللَّهِ وَالَّذِينَ مَعَهُ أَشِدَّاءُ عَلَى الْكُفَّارِ
رُحَمَاءُ بَيْنَهُمْ تَرَاهُمْ تَرْكَعًا سُجَّدًا يَبْتَغُونَ فَضْلًا مِّنَ اللَّهِ وَرِضْوَانًا سِيمَاهُمْ
فِي وُجُوهِهِمْ مِّنْ أَثَرِ السُّجُودِ لَكَ مَثَلُهُمْ فِي التَّوْرَةِ وَمَثَلُهُمْ فِي
الْإِنْجِيلِ كَزَرْعٍ أَخْرَجَ شَطْهُ فَوَرَّهُ فَآسْتَعْلَظَ فَآسْتَوَى
عَلَى سُوقِهِ يُعْجِبُ الزُّرَّاعَ لِيغِيظَ بِهِمُ الْكُفَّارَ وَعَدَّ اللَّهُ الَّذِينَ
ءَامَنُوا وَعَمِلُوا الصَّالِحَاتِ مِنْهُمْ مَغْفِرَةً وَأَجْرًا عَظِيمًا) الفتح 29

الحمد

الحمد لله حمداً كثيراً طيباً مباركاً فيه يليق بجلال وجهه وعظيم سلطانه. الحمد لله الذي بنعمته تتم الصالحات. احمد الله عز وجل أن وفقني إلى إتمام هذا البحث وأسأل 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. Mohmmmed 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 Q_{total} of 86.5%, 77.6%, on α -helix and coil secondary structure respectively when we used SVM method and also we achieved Q_{total} 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), $Q_{predicted}$ and $Q_{observed}$ 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 ببنية 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 كمداخل لنماذج آلات المتجهات (SVM) ونواة الانحدار اللوجستي (KLR). في هذه الدراسة حققنا دقة التنبؤ عالية مقارنة بنتائج الدراسات السابقة وهي Q_{total} من 86.5%، 77.6%، على البنية الثانوية (α -helix) و (coil) على التوالي عندما استخدمنا طريقة SVM وأيضاً حققنا Q_{total} من 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)، $Q_{predicted}$ و $Q_{observed}$ على مجموعات البيانات RS126.

References

- [1] David, Barber (March 9, 2010) "Bayesian Reasoning and Machine Learning."
- [2] MurtadaKhalafallah Elbashir¹, Jianxin Wang^{1*}, 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 Machines in 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 than 70% 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 MohdSaberMohamad (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 CASCADING NEURAL NETWORKS." 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 -turns Prediction." 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."