PISSN: (1681-6870); EISSN: (2790-2293)

مجلة كلية الرافدين الجامعة للعلوم

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

AL- Rafidain University College

استخدام اساليب الجوار الاقرب k والغابة العشوائية لتصنيف بيانات للإشعاع الشمسي

أ.م.د. أسامة بشير شكر	أ.م.د. مثنى صبحي سليمان
drosamahannon@uomosul.edu.iq	muthanna.sulaiman@uomosul.edu.iq

قسم الاحصاء والمعلوماتية - كلية علوم الحاسوب والرياضيات - جامعة الموصل، نينوى، العراق

المستخلص

ان دراسة الاحوال المناخية والتقلبات الجوية وتأثيراتها مهم جدا لتشخيص الملامح البيئية والمناخية وتأثيراتها على مختلف المجالات التي تخص حياة الانسان والكائنات الحية الاخرى. في هذه الدراسة سيتم دراسة وتصنيف متغير الاشعاع الشمسي solar radiation (SLR) الكلي من خلال الاعتماد على متغيرات الانحدار الذاتي Auto-regression (AR) بعد تشخيص العلاقة بين تلك المتغيرات رياضيا من خلال استخدام اسلوب الجوار الاقرب K- (K) Nearest Neighbor (KNN) والغابة العشوائية Random Forest (RF). تم اخذ بيانات عراقية من محطة الانواء الجوية الزراعية في مدينة الموصل واستخدامها كحالة حقيقية في هذه الدراسة. مع هكذا بيانات فان من اهم اسباب عدم دقة التصنيفات هو وجود عدة عوائق ومشاكل مثل عدم الخطية وعدم التأكدية بالنسبة للبيانات المدروسة. اظهرت النتائج من خلال المقارنة التفوق المتبادل بين اسلوبي RF و KNN لتصنيف متغير SLR بالنسبة لكلا فترتي التدريب والاختبار بينما ادى كلا الاسلوبين بنتائج تصنيفية عالية الدقة. وكاستنتاج فان اسلوبي RF وKNN من الممكن استخدامهما لتصنيف بيانات SLR والحصول على نتائج دقيقة.

معلومات البحث

تواريخ البحث

تاريخ تقديم البحث: 2022/12/15

تاريخ قبول البحث: 2023/3/3

تاريخ رفع البحث على الموقع: 2023/12/31

الكلمات المفتاحية

الاشعاع الشمسي SLR، الجوار الاقرب KNN (K)، الغابة العشوائية RF، التصنيف.

للمراسلة:

أ.م.د. أسامة بشير شكر

drosamahannon@uomosul.edu.iq

https://doi.org/10.55562/jrucs.v54i1.586

المقدمة

في هذه الدراسة تم التطرق الى دراسة التصنيف لاحد اهم متغيرات الانواء الجوية اذ تكمن أهمية هكذا تنبؤات من خلال معرفة مدى تأثيرها على الانسان والحيوان والنبات وسائر الكائنات الحية والتخطيط لمستقبل خال من مشاكل التأثيرات السلبية لمتغيرات الانواء الجوية المختلفة وغني بتأثيراتها الإيجابية. سيتم استخدام بيانات السطوع الشمسي SLR وتصنيفه اعتماداً على مبدأ الانحدار الذاتي بتشخيص علاقة رياضية بينه وبين المتغيرات الذاتية التنبؤية من خلال أنموذج خاص بأسلوبي RF و KNN لتصنيف متغير SLR من حيث أن السطوع عالٍ أو أنه منخفض. سيستخدم اهم أساليب تعلم الآلة متمثلا أسلوبي RF و KNN لتصنيف بيانات السطوع الشمسي SLR اعتماداً على متغيرات الانحدار الذاتي. تعد معظم بيانات الانواء الجوية وملوثات الهواء من نوع غير الخطي ولذلك فإن استخدام بعض الأساليب والنماذج الخطية قد يؤدي بالتالي الى نتائج قليلة الدقة وذلك ان بيانات الانواء الجوية تعد بشكل عام أحد أنواع السلاسل الزمنية التي تحتوي على العديد من المتغيرات الموسمية وكذلك الدورية التي قد تؤثر سلبا في جعل هذا النوع من البيانات غير متجانسة وكذلك تؤثر في نتائج التنبؤ ودقتها. صنفت البيانات الى نوعين من التصنيفات وفقا لطبيعة الأجواء في محافظة نينوى. المجموعة الأولى من البيانات صنفت حسب حد العتبة للمعدل السنوي للسطوع الشمسي الكلي في حين تضم المجموعة الثانية فاستخدم فيها حد عتبة لكل شهر تبعاً للمعدلات الشهرية المتنوعة لتجاوز اختلاف التباينات بين معدلات السطوع الشمسي الشهرية[1-5].

في هذه الدراسة سيقترح استخدام اسلوبي RF و KNN بوصفه أسلوباً حديثاً لتحسين نتائج التصنيف لمتغير السطوع الشمسي SLR اعتماداً على متغيرات الانحدار الذاتي إذ إنه أسلوب يجمع أسلوبين مهمين هما شجرة الانحدار Regression Tree (RT) وشجرة التصنيف Classification Tree (CT) في أسلوب واحد مع تعدد تلك الاشجار مشابها لأساليب التعلم العميق يستخدم لتصنيف متغير السطوع الشمسي SLR بدقة متميزة.

تعتبر نماذج الغابة العشوائية Random Forest طريقة دقيقة وقوية للغاية في التصنيف بسبب اعتمادها في اتخاذ القرار على العديد من أشجار القرار حيث تكون أشجار القرار هذه غير مترابطة وكل منها تؤدي الى قرار مستقل وفي نهاية الامر فأن القرار النهائي لأسلوب الغابة العشوائية RF سيكون بالغالبية المطلقة لقرارات أشجار الانحدار التي تتكون منها الغابة العشوائية مما يجعل من اسلوب الغابة العشوائية اسلوباً حصينا ضد عدم خطية البيانات وكذلك عدم تجانسها.

تعد خوارزمية الجوار الاقرب k، التي غالبًا ما يتم اختصارها KNN ، احد اساليب تعلم الالة تقترح لتصنيف البيانات التي تقدر مدى احتمال أن تكون نقطة البيانات تنتمي الى مجموعة دون الأخرى وفقًا للمجموعة التي توجد بها نقاط البيانات الأقرب إليها .يعد الجوار الاقرب مثال لخوارزمية "التعلم البطيء" مما يعني أنه لا ينشئ نموذجًا لمجموعة البيانات. الحسابات الوحيدة التي تستخدمها الطريقة هي تنفيذ استطلاع آراء من خلال جارات نقطة البيانات. هذا يجعل تطبيق KNN سهلًا للغاية ورصيناً للتنقيب عن البيانات وتصنيفها.

المفاهيم النظرية

الغابة العشوائية هي احدى خوارزميات التعلم الخاضعة للأشراف Supervised أي ان مخرجات الغابة العشوائية يجب ان تتطابق مع متغيرات الهدف وبمقارنتها تنتج أخطاء التنبؤ وتعتمد على مبدأ تقنيات أشجار التصنيف والانحدار ومن مميزاتها انها دقيقة حسابياً وتعمل بسرعة وذلك عبر بيانات كبيرة نسبياً وهي من التقنيات الحديثة حيث يتم استخدامها في العديد من التطبيقات في مجالات متنوعة لاعتمادها على مبدأ التصنيف والانحدار فهي عبارة عن مخطط لمجموعة أشجار تستخدم لبناء أنموذج يعطي تنبؤات من خلال اوراقها الناتجة عن مساحات وتفرعات مختارة عشوائياً من البيانات بمبدأ مشابه لبديهيات أشجار الانحدار[5]. الشكل (1) يوضح هيكلية الغابة العشوائية كأحد أنواع أشجار الانحدار والتصنيف.

شكل (1):هيكلية الغابة العشوائية كأحد أنواع أشجار الانحدار والتصنيف

كل تفرع في الشجرة في الشكل (1) يمثل نقطة قرار تم اتخاذها على أساس جملة شرطية وهكذا تستمر التفرعات لحين الوصول الى القرارات النهائية المتمثلة بعقد الأوراق حيث ان كل ورقة تعتبر كعقدة منفصلة من قرار منفصل عن باقي الأوراق وان هذه الأشجار تعطي تطابقا امثل بين المخرجات المتمثلة بالتصنيفات بالمقارنة مع المتغير الأصلي الذي تم اعتباره كمتغير هدف, أي سيتم تطوير أسلوب التنبؤ والحصول على تصنيفات مثلى بأقل أخطاء للتصنيف عند استخدام أسلوب (RF) كأحد تقنيات أشجار التصنيف مقارنة بالأساليب التقليدية. توفر نمذجة السلاسل الزمنية باستخدام الغابات العشوائية قدرة تصنيفية معززة وأكثر دقة مقارنة بالنماذج التقليدية للتصنيف خصوصاً ببيانات الأرصاد الجوية وغيرها كثيرة على العموم. يتم اعتماد مبدأ التعبئة (bagging principle) الذي اساسه هو عملية المعاينة التمهيدية (Bootstrap Sampling) اذ تعمل طريقة التعبئة على تحسين أداء أشجار التصنيف وتجعل (RF) أكثر حصانة عند تجميعها مع بعضها. يتم معالجة ذلك بجعل الاشجار في الغابة العشوائية غير مترابطة مع بعضها (مختلفة). بعد تحويل الأشجار في الغابة العشوائية من مترابطة الى غير مترابطة (مختلفة) مما سيضمن زيادة ملحوظة في دقة التصنيف باستخدام الغابة العشوائية.

يتم بناء خوارزمية الغابة العشوائية باستخدام الخطوات الثلاث ادناه:

1. من بيانات التدريب يتم استخراج B من العينات التمهيدية والتي هي في الأصل مترابطة فيما بينها اذ ان B تمثل حجم الغابة او عدد الأشجار المتعددة المشار اليها في الشكل (2.3)

2. لكل مجموعة من مجموعات البيانات B فأن نمو الشجرة T_b سيتم باتباع خطوات متسلسلة في كل عقدة من عقد الشجرة لحين الوصول الى n_{min} والتي تمثل الحد الأدنى من أوراق الأشجار او عدد العقد وكما يلي:

أ- اختيار m والتي تمثل العدد المختار عشوائياً من التنبؤات في كل قسم من العدد الكلي للمتغيرات p.

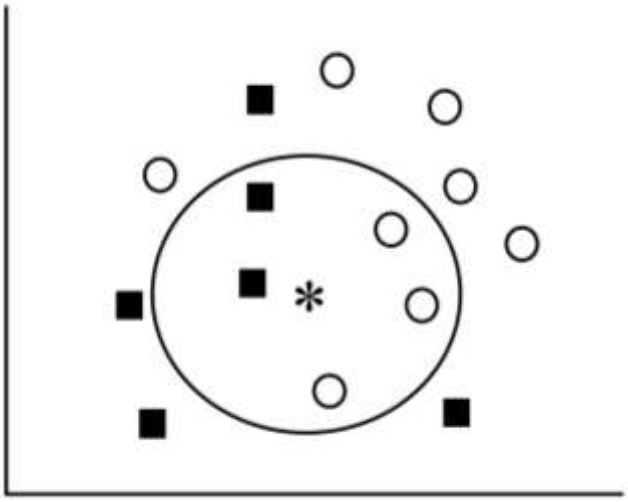
ب- اختيار أفضل التصنيفات من المجموعة المختارة في (أ) وقد تم الإشارة اليها بالرمز m مع اختيار القسم العائدة اليه بهدف تقليل قيمة اخطاء التصنيفات المختارة في (أ).

ج- فصل العقدة الى عقدتين فرعيتين تبعاً للمعيار المستخدم او القيم التصنيفية الأفضل التي تم اختيارها في (ب).

3. استخلاص المخرجات من جميع الأشجار من خلال إيجاد المجموعة $\{T_b\}_1^B$ وأخيرا فأنه عند نقطة معينة X فأن التصنيف ممكن حسب المعادلة التالية: [6]

$$f_{RF} = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (1)$$

خوارزمية الجوار الاقرب KNN k حالها حال اسلوب الغابة العشوائية من احد اساليب تعلم الالة التي تهتم بالانحدار لتصنيف البيانات من خلال تحديد احتمال انتماء نقطة البيانات الى اقرب المجموعات من خلال البعد والقرب في التشابه في الخصائص والميزات. تستخدم في الغالب المسافة الاقليدية بين مجموعات البيانات لتحديد الجوارات الاقرب والاستعلام من خلالها لإجراء افضل وادق نتائج التصنيفات. فعلى فرض وجود متجهين لخاصيتين تصنيفيتين x_i و y_i فالمسافة الاقليدية بينهما يحدد مدى التشابه والتقارب المطلوب للحكم والتصنيف كما يوضح ذلك الشكل (2) ادناه والمعادلة (2). نظرًا لأن KNN لا يتطلب مرحلة تدريب من خلال تحديد نموذج مسبق ولذلك فإن التنفيذ الرئيسي هو "البحث" عن أقرب جيران k التي بتنوعها ستؤدي الى نتائج تصنيفية مختلفة.



شكل (2): مفهوم الجوارات الاقرب K بالبساطة دون نموذج مسبق من حيث التشابه والاختلاف.

$$dist(A,B) = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{m}} \quad (2)$$

عندما m هو بعد وحجم المتجهين المترابطين اما A وB فهما المجموعتان التصنيفيتان التان تنتمي اليها المشاهدات. وبعد تحديد المسافات والتشابه والاختلاف فان التنبؤ بالصنف الذي تنتمي له المشاهدة من الممكن ان يتم عبر المعادلة (3) ادناه.

$$\hat{y} = \arg\min_{y=1,2,\ldots,k} \sum_{i=1}^{k} \hat{P}(j \mid x) C(y \mid j) \quad (3)$$

عندما $\hat{y}$ تمثل المتغير التنبؤي التصنيفي وان k تمثل عدد الاصناف اما $\hat{P}(j \mid x)$ هي الاحتمال اللاحق للفئة j للمشاهدة x وان $C(y \mid j)$ فهي الكلفة التصنيفية للمشاهدات y ضمن الفئة الصحيحة j.

النتائج والمناقشات

سيتم الاعتماد على استخدام الاداة (fitrensemple) في برنامج (MATLAB) بعد تفعيل الخاصية 'Reproducible' لبناء أنموذج الانحدار التجميعي(Regression Ensemple Model) للغابة العشوائية RF باستخدام عدة متغيرات تفسيرية ومتغير واحد معتمد. ان بيانات هذا البحث تتضمن بيانات سلاسل زمنية احادية المتغير(درجات الحرارة الصغرى وكميات التبخر) وسيتم

اعتماد مبدأ الارتباط الذاتي في السلاسل الزمنية لإنشاء متغيرات تفسيرية من كل متغير من متغيرات الدراسة وذلك من خلال استخدام التخلفات الزمنية للمتغير الاصلي كمتغيرات تفسيرية حيث سيكون لكل متغير من متغيرات الدراسة أربعة متغيرات تفسيرية (أربعة تخلفات زمنية) فيما سيكون نفس المتغير الاصلي هو المعتمد اذ سنعتمد على مبدأ التجميع والتوفيق بين النماذج فان اشجار الغابة العشوائية باستخدام (10) تجزءات للبيانات كحد اقصى والتي سيستفاد من توفيقها باستخلاص افضل النتائج. وسيتم استخدام (500) شجرة ثم توفيقها للحصول على افضل التصنيفات. كذلك الحال بالنسبة لأسلوب KNN فسيتم الاعتماد على استخدام الاداة (fitcknn) في برنامج (MATLAB) بعد تحديد عدد الجوارات اذ تم اخذ كل الاحتمالات الممكنة وكذلك تحديد المسافة الاقليدية حكماً للتشابه والاختلاف بين المجاميع التصنيفية والمشاهدات.

بعد الانتهاء من بناء نماذج RF وKNN فالخطوة التالية هي التنبؤ التصنيفي بعد تدريب البيانات واختبارها بعد تقسيم البيانات الى فترتين للتدريب والاختبار. البيانات هي في الاصل سلاسل زمنية للسطوع الشمسي SLR تم تصنيفها اعتمادا على حدود العتبة للمعدلات الشهرية والمعدل السنوي فتم الحصول بذلك على مجموعتين من البيانات احداهما مصنفة حسب المعدلات السنوية والاخرى حسب المعدلات الشهرية للسطوع الشمسي الكلي لمحافظة نينوى وكل صنف منهما مقسم الى فترة للتدريب وفترة في نهاية السلسلة للاختبار بواقع 362 مشاهدة للتدريب للعام 2018 و65 مشاهدة للاختبار. تم الحصول على البيانات من احد مراكز الارصاد الجوية الزراعية في محافظة الموصل. دقة التصنيفات دونت حسب الاسلوب المستخدم ومجاميع البيانات كما في الجدول (1) ادناه.

جدول (1): دقة التصنيفات باستخدام اسلوبي RF و KNN لتصنيف السطوع الشمسي

	التصنيف بالعتبة السنوية		التصنيف بالعتبات الشهرية	
الاسلوب	**التدريب**	**الاختبار**	**التدريب**	**الاختبار**
KNN	100.00%	90.77%	100.00%	56.92%
RF	99.17%	98.46%	99.72%	64.62%

اذ من الواضح تباين الدقة وتفاوتها من اسلوب لآخر ومن فترة لأخرى ولكن دقة التصنيف هي العنوان الابرز لكلا الاسلوبين بشكل عام. يشذ عن الدقة التصنيفية نتائج فترة الاختبار للبيانات المصنفة بالعتبات الشهرية اذ من الواضح تدني الدقة مقارنة بالمجاميع الاخرى من البيانات. ولهذا لا يمكن القول بالتفوق المطلق لاحد الاسلوبين ولكن الركازة والرصانة هي من نصيب الغابة العشوائية لأنها تعمل مثل اساليب التعلم العميق لكثرة عدد اشجار القرار المستخدمة.

الاستنتاجات

من خلال ما تقدم من نتائج واستعراضات فانه من الممكن استنتاج امكانية الحصول على دقة متناهية في التصنيف بخطأ تصنيفي يكون منعدما احيانا كما تم ملاحظته في النتائج. ولذلك فمن الممكن اقتراح كلا اسلوبي KNN و RF للحصول على افضل التصنيفات وبأقل الاخطاء كأساليب حديثة تعتمد مبدأ التنوع والتكرار بأعداد مهولة تسهم في الحصول على فرص اكبر لتعدي دقة التصنيفات.

المصادر

[1] Kornelsen, K. and P. Coulibaly, "Comparison of interpolation, statistical, and data-driven methods for imputation of missing values in a distributed soil moisture dataset". Journal of Hydrologic Engineering, 2014. **19**(1): p. 26-43.

[2] He, H., et al., "Ensemble learning for wind profile prediction with missing values". Neural Computing and Applications, 2013. **22**(2): p. 287-294.

[3] Cadenas, E. and W. Rivera, "Wind speed forecasting in the South Coast of Oaxaca, México". Renewable Energy, 2007. **32**(12): p. 211.2128-6.

[4] Mahmood, F.H. and G.S. Al-Hassany, "Study global solar radiation based on sunshine hours in Iraq". Iraqi Journal of Science, 2014. **55**(4A): p. 1663-1674.

[5] Chaichan, M.T., et al., "The effect of dust components and contaminants on the performance of photovoltaic for the four regions in Iraq: a practical study". Renewable Energy and Environmental Sustainability, 2020. **5**: p. 3.

[6] Noureen, S., et al., "A comparative forecasting analysis of arima model vs random forest algorithm for a case study of small-scale industrial load". International Research Journal of Engineering and Technology, 2019. **6**(09): p. 1812-1821.

Journal of AL-Rafidain University College for Sciences (2023); Issue 54; 154 – 158

AL- Rafidain University College

PISSN: (1681-6870); EISSN: (2790-2293)

Journal of AL-Rafidain University College for Sciences

Available online at: https://www.jrucs.iq

JRUCS

Journal of AL-Rafidain University College for Sciences

Using K-Nearest Neighbor and Random Forest Approaches for Classifying Solar Radiation

Osamah B. Shukur	Muthanna S. Sulaiman
drosamahannon@uomosul.edu.iq	muthanna.sulaiman@uomosul.edu.iq
Department of Statistics and Informatics - College of Computer Science and Mathematics - University of Mosul, Nineva, Iraq	

Article Information

Article History:
Received: December, 15, 2022
Accepted: March, 3, 2023
Available Online: December, 31, 2023

Keywords:
Solar radiation (SLR), K-nearest Neighbor (KNN), Random Forest (RF), Classification.

Correspondence:
Osamah B. Shukur
drosamahannon@uomosul.edu.iq

Abstract

Studying climatic status and meteorological effects is important to identify climatic and environmental elements and their impacts in various fields of human life as well as other organisms. In this study, solar radiation (SLR) variables will be studied and classified based on their autoregressive variables by identifying the mathematical relationship among these variables using K-nearest Neighbor (KNN) and Random Forest (RF) techniques. Iraqi datasets taken from an agricultural meteorological station located in Mosul, Iraq, were used as a real case study. In these types of data, there are many obstacles, including nonlinearity and uncertainty, that will be the reasons for inaccurate classifications. The results of the comparisons explain that the RF approach and KNN in SLR classification have varied classification performance, while both of them produce highly accurate classification results. In conclusion, SLR can be accurately classified using RF and KNN techniques.

https://doi.org/10.55562/jrucs.v54i1.586