مجلة العلوم البحتة والتطبيقية
Journal of Pure & Applied Sciences
www.Suj.sebhau.edu.ly ISSN 2521-9200
Received 12/05/2019 Revised 16/08/2019 Published online 17/11/2019

نحو اختيار امثل للخصائص لتحسين دقة التصنيف

عمر عبدالغني شيبة
قسم علوم الحاسوب– كلية تقنية المعلومات–جامعة سبها، ليبيا
للمراسلة: oma.shiba@sebhau.edu.ly

الملخص غالباً ما يستخدم تمثيل البيانات في مجالات تنقيب البيانات وتعلم الآلة العديد من الخصائص او الميزات ، التي ربما البعض منها فقط يرتبط بالهدف المرسوم. تم اقتراح العديد من الخوارزميات لاختيار المجموعة ذات العلاقة فقط من بين هذه الخصائص ، ولكن هذه الخوارزميات ليست جميعها مناسبة لحل مشكلة اختيار ميزة او خاصية معينة. الغرض من هذه الورقة هو تقليص عدد الخصائص عن طريق إزالة الخصائص الغير ذات صلة. إن اختيار مجموعة من الخصائص ذات الصلة يؤدي إلى تحسين أداء دقة التصنيف في مجال تنقيب البيانات، ولتحقيق هذا الغرض ، تم استخدام تقنية التشريح المستخدمة في لغات البرمجة لتقليل عدد الخصائص في بعض مجموعات البيانات المختارة. تشير النتائج التجريبية إلى أن التقنية المقترحة ادت الى تحسين دقة التصنيف بشكل ملحوظ مقارنةً بالطرق الأخرى المتمثلة في RELIEF مع خوارزمية التعلم الأساسية (C4.5) و (K-NN مع RELIEF) K-Nearest Neighbour ، و RELIEF مع Induction of Decision Tree (ID3) ، والتي تستخدم في الغالب في مهمة اختيار الخصائص في مجال تنقيب البيانات وتعلم الآلة.

الكلمات المفتاحية: اختيار الخصائص، تنقيب البيانات، التصنيف، تقنية التشريح، الخصائص الغير ذات صلة، دقة التصنيف.

Toward Optimal Feature Selection to Improve Classification Accuracy

Omar A. A. Shiba
Computer Science Department, Faculty Of Information Technology/Sebha University, Libya
Corresponding author: oma.shiba@sebhau.edu.ly

Abstract The representation of data in the fields of data mining and machine learning often uses many features, only a few of which may be related to the target concept. Many feature subset selection (FSS) algorithms have been proposed, but not all of them are appropriate for a given feature selection problem. The purpose of this paper is to eliminate the number of features by removing irrelevant once. Choosing a subset of the features may improve the performance of classification accuracy. To achieve this purpose, slicing technique used to reduce the number of features in some selected datasets. The experimental results indicate that the performance of proposed feature selection scheme is very good compared with other approaches which are RELIEF with Base Learning Algorithm (C4.5), RELIEF with K-Nearest Neighbour (K-NN) and RELIEF with Induction of Decision Tree Algorithm (ID3), which are mostly used in the feature selection task.

Keywords: Feature selection, data mining, classification, accuracy, slicing techniques, irrelevant features.

1. المقدمة

زيادة حجم البيانات من حيث عدد الحالات وعدد الخصائص أصبح تحدياً كبيراً لخوارزميات اختيار الخصائص ذات الصلة فقط [1]. للتعامل مع هذا النوع من المشاكل ، عادةً ما يتم استخدام طريقتين رئيسيين للتقليص وهما: استخراج الخصائص واختيار الخصائص [2]. تنشئ طريقة استخراج الخصائص مجموعة خصائص جديدة بينما تقوم طريقة اختيار الخصائص بالبحث عن مجموعة جيدة من الخصائص في المجموعة الأصلية عن طريق إزالة البيانات الغير ملائمة والمتكررة. الخصائص الغير ملائمة لا تحتوي على معلومات مفيدة حول مشكلة التصنيف في حين ان الخصائص المتكررة تحتوي على معلومات موجودة اصلا في خصائص أكثر اهمية وإفادة [3]. يعد اختيار الخصائص مفيداً في التعرف على الأنماط الإحصائية وتعلم الآلة وتنقيب البيانات والإحصائيات وخوارزميات التصنيف.

هناك العديد من التقنيات التي تم اقتراحها لإجراء معالجة مسبقة للخصائص مثل استخراج الخصائص [4] ، واختيار الخصائص [5] ، وإعطاء اوزان للخصائص [6].بناء الخصائص وتحويل الخصائص [7]. حيث يتحسن أداء معظم عمليات التصنيف عمليا عند إزالة الخصائص الغير ذات صلة أو غير الملائمة للحالة [8] [9]. بناءً على هذه الحقيقة ، ونتائج دقة التصنيف سابقا والتي حصل عليها باحثون آخرون تعتبر ليست جيدة بما يكفي، فكان لابد من استكشاف طريقة مثلى لتحسين دقة التصنيف. نتيجة لذلك تم اقتراح خوارزمية جديدة تقلل من عدد الخصائص التي من شأنها تحسين دقة التصنيف. حيث تقترح هذه الورقة خوارزمية اختيار الخصائص لإزالة الخصائص الغير ملائمة وإنتاج مجموعة فرعية صغيرة من الخصائص لتحسين دقة التصنيف. في حين ان بقية الورقة تم تنسيقها على النحو التالي. يستعرض

القسم الثاني معارف أساسية ودراسات سابقة متعلقة بالخوارزمية المقترحة. بينما يصف القسم الثالث الخوارزمية المقترحة. في حين تم استعراض ومناقشة النتائج التجريبية في القسم الرابع، وأخيراً قدم القسم الخامس بعض الاستنتاجات.

2. الأعمال ذات الصلة والدراسات السابقة

في هذا القسم ، معارف أساسية ودراسات سابقة متعلقة بالخوارزمية المقترحة وكذلك بمشكلة اختيار الخصائص سيتم التعريج عليها باختصار. فيما يتعلق بمشكلة اختيار الخاصية ، يوجد طريقتان عامتان: الفلتر filter و التجميع wrapper.حيث تتميز الطريقة الاولى بتطبيق منهجية مختلفة في طوري التدريب والاختبار. بينما الطريقة الثانية تطبق نفس الخوارزمية في العملية بأكملها. في الطريقة الاولى و خوارزمية FOCUS [10] يتم فحص المجموعة الجزئية واختيار تلك الخصائص التي ينتج عنها تقليل عدد الخصائص من بين جميع حالات التدريب. الخوارزمية الأخرى ذات الصلة هي Relief [6] وانواعها [7] تعتبر بعضاً من خوارزميات الفلترfilter المعروفة على نطاق واسع ، وهي خوارزمية عشوائية تقوم بتعيين وزن لكل خاصية استناداً إلى أقرب حالتين.

في استراتيجية التجميع wrapper ، اكثر خوارزميات البحث التتابعي شيوعا لاختيار الخصائص هي الاختيار التتابعي للأمام (FSS) والاختيار التتابعي للخلف (BSS). يبدأ FSS بصفر من الخصائص ويقوم بتقييم جميع المجموعات الفرعية للخصائص ،ويتم اضافة الخصائص ذات الأداء الأفضل تباعا. بينما يبدأ BSS بجميع الخصائص ويزيل بشكل متكرر الخاصية التي لا تؤدي إزالتها إلى زيادة النتائج [8]; [11] ;[12].

تتضمن استراتيجية التجميع خوارزمية التعلم كجزء من مهمة التقييم الخاصة بها. عادة ما توفر هذه الاستراتيجية دقة أفضل ولكنها أغلى من الناحية الحسابية مقارنة باستراتيجية الفلتر Filter [13].

اختيار الخاصية

الهدف من اختيار الخاصية هو تقليل عدد الخصائص المستخدمة لتوصيف مجموعة البيانات وذلك لتحسين أداء الخوارزمية في مهمة معينة.

تعريف 1: هي عملية اختيار أفضل مجموعة فرعية من الخصائص التي تصف الفرضية (في اسواء الحالات تكون هي نفس المجموعة الأصلية).

$$f \subset F \qquad (1)$$

حيث F في "معادلة 1" هي مجموعة "n" من الخصائص "الأصلية" و f` هي مجموعة "m".من الخصائص المستخرجة بواسطة خوارزمية اختيار الخصائص.

الخصائص الغير ذات صلة

لا تكون الخاصية ذات صلة إذا كانت لا تسهم بأي شيء في الفرضية المستهدفة ، أي أنها لا تقدم مساهمة ذات معنى في مهمة التصنيف خوارزميات الجار الاقرب لها حساسية بشكل خاص في تضمين خصائص غير ذات صلة في مجموعة البيانات. تم تحسين أداء معظم عمليات التصنيف عند إزالة الخصائص الغير ذات صلة أو غير الملائمة للحالة [8].

3. الخوارزمية المقترحة والشروط ذات الصلة

يستعرض هذا القسم بعض التعريفات الأساسية المتعلقة بخوارزمية التشريح المقترحة ؛

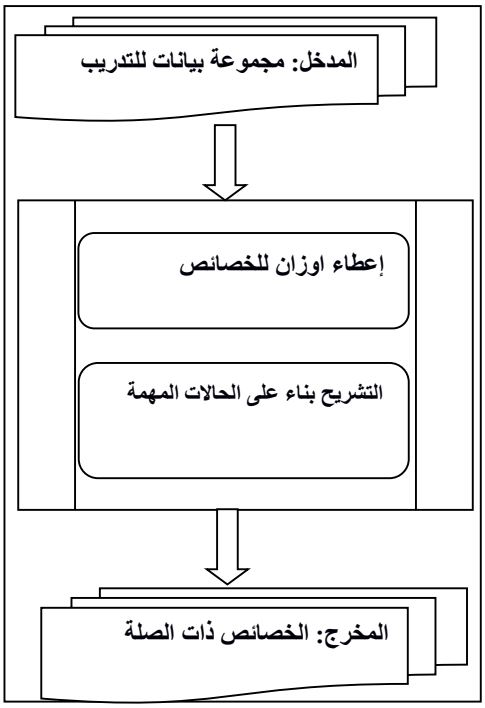
التعريف 2: تشريح الحالة: هي عملية للحصول تلقائياً على الأجزاء الفرعية (الخصائص) لحالة ذات معنى شامل.

التعريف 3: معيار تشريح الحالة: يشير إلى شروط حساب عملية تشريح الحالة، بناء على ماذا ولأي حالة مطلوب التشريح.

التعريف 4: الحالة المشرحة تحتوي على جميع الخصائص التي يمكن أن تكون لها علاقات مباشرة مع الخصائص المهمة في الحالة الجديدة.

الفكرة الأساسية للخوارزمية المقترحة

من الناحية النظرية ، الخوارزمية المقترحة هي شكل من أشكال خوارزميات الجار الأقرب. حيث تتمثل الخطوة الأولى في تعيين أوزان للحالات الجديدة وأيضاً لحالات التدريب في ملف البيانات. الخطوة الثانية هي تشريح الحالات بناء على خصائص مختارة. يعمل تشريح الحالات على إزالة تلك الخصائص التي لا صلة لها بالحالة المطروحة وأيضاً للحالات في ملف الحالات الخاصة بالتدريب. تشريح الحالة كما تم ذكره انفا يعني أننا مهتمون بالحصول على جزء من خصائص الحالة المقترحة للحل أو المسؤولة عن أجزاء معينة من حل الحالة المطروحة. وذلك عن طريق تشريح الحالة بناء على الخصائص المهمة، حيث يمكننا الحصول على حالة جديدة مع عدد قليل من الخصائص أو مع خصائص مهمة فقط. الشكل (1) يبين خطوات عملية التشريح.



شكل(1) يبين خطوات عملية التشريح

كما هو مبين في الشكل 1 ، فإن الوصف الموجز لعمل الخوارزمية المقترحة لاختيار الخصائص هو كما يلي:

• يتم إدراج الحالة في <مرحلة وزن الخصائص> لتخصيص الأوزان لكل خاصية في الحالة باستخدام طريقة إحصائية تسمى "الاحتمال الشرطي" والذي سيخصص أوزان ارتباط عالية لمعظم الخصائص ذات الصلة من بين الخصائص الأخرى.

• بعد تخصيص الأوزان ، ستحدث تقنية التشريح فعلها وفقًا لأوزان الخصائص وذلك عن طريق إزالة الخصائص الغير مهمة فقط.

• ناتج الخطوة الثانية هو مجموعة فرعية من الخصائص، والتي تحتاج إليها خوارزمية التصنيف.

تتكون الخوارزمية المقترحة من قاعدة بيانات، وطريقة حسابية واحدة لتحديد أهمية كل خاصية، وتقنية تشريح لاختيار لتحديد الخصائص المهمة لاتخاذ القرار.

تمثيل قاعدة البيانات

في المهام التقليدية لتعلم الآلة، يتم تمثيل البيانات كجدول من الحالات. حيث يتم وصف كل حالة من خلال عدد محدد من الخصائص ، جنباً إلى جنب مع تسمية تشير إلى فئتها او تصنيفها. الخصائص عادة واحدة من نوعين. الاسمية (القيم فيها هي عناصر في مجموعة غير مرتبة) أو رقمية (القيم فيها عبارة عن أرقام حقيقية).

تتطلب الخوارزمية المقترحة مجموعة من الحالات السابقة كمدخلات ؛ يتم تمثيل هذه المجموعة من الحالات كملف بيانات علائقية. كل حالة هي سجل في ملف البيانات، حيث يتكون هذا الجدول من جزءين. يستخدم الجزء الأول كمؤشرات لقيمة الجزء الثاني والذي يمثل متغير الهدف. الجدول (1) يظهر هيكلية ملف البيانات.

جدول (1) يوضح هيكلية ملف البيانات

	خاصية 1	خاصية 2	.....	خاصية N	فئة / تصنيف
حالة 1	.....	.....	.....	.....	

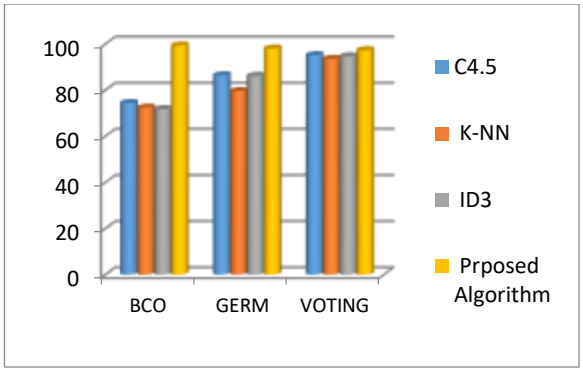
[.....predicators.....] [....Goal...]

4. النتائج التجريبية

في هذا القسم ،سيتم مقارنة الخوارزمية المقترحة كخوارزمية لاختيار الخصائص مقابل ثلاث خوارزميات لاختيار الخصائص والمتمثلة في في (RELIEF + C4.5) و (K-NN + RELIEF) ،(ID3 + RELIEF) ، والتي تستخدم في الغالب في مهمة اختيار الخصائص ، حيث تم إجراء التجارب على ثلاث مجموعات بيانات تم الحصول عليها من مستودع تعلم الالة [14]والموضحة في الجدول 2. عند الانتهاء من جميع التجارب على مجموعات البيانات الثلاثة المحددة من مجالات مختلفة ، فإنه يمكن ملاحظة أن جميع التقنيات تعطي دقة تصنيف جيدة. كذلك يمكن ملاحظة أن الخوارزمية المقترحة أفضل من خوارزميات اختيار الخصائص الأخرى، لأن اختيار الخاصية في الخوارزمية المقترحة يستند إلى أوزان الخصائص وتسمية الفئة ، حيث يعتمد اختيار الخصائص في الخوارزميات الأخرى على استخراج القواعد ، والتي تنتج أحياناً عددا من الخصائص التي لا صلة لها بالحالة قيد الدراسة، وفي بعض الأحيان تصبح ضعيفة للغاية وغير مدعومة من قبل أي حالة. يوضح الشكل 2 الفرق في دقة تصنيف الخوارزمية المقترحة مقابل الخوارزميات التي تم اختيارها في هذه الدراسة باستخدام RELIEF.

جدول(2) يوضح دقة التصنيف للخوارزمية المقترحة مقابل مجموعة من الخوارزميات الاخرى

Algorithm Datasets	C4.5+ RELIFE	K-NN+ RELIFE	ID3+ RELIFE	Proposed Algorithm
BCO	74.37	72.375	71.75	99.30
GERM	86.42	79.57	86.143	98.00
VOTING	95.125	93.5	94.625	97.30



شكل (2) يوضح الاختلاف في دقة التصنيف للخوارزميات المختارة

5. الخلاصة

وصفت هذه الورقة باختصار بعض الدراسات و الأعمال ذات الصلة بغرض اختيار مجموعة جزئية من الخصائص في حل مشكلات التصنيف. كما قدمت الورقة وناقشت التقنية المقترحة لنفس الغرض.

لقد تم اختبار خوارزمية اختيار الخصائص المقترحة على ثلاث مجموعات من بيانات حقيقية مع ثلاث خوارزميات مختلفة شائعة الاستخدام لاختيار مجموعة جزئية للخصائص. حيث أظهرت التجارب أن التقنية المقترحة تعمل على تحسين دقة التصنيف وقد أعطت نسبة عالية جداً بالنسبة لدقة التصنيف وذلك لأنها تستند إلى تشريح الخصائص وفقا للخصائص ذات الصلة فقط. ومع ذلك، لا يُنظر إلى هذه التقنية على أنها تقنية بديلة ، ولكنها تقنية مكملة للطرق التي تهدف لاختيار مجموعة جزئية من الخصائص.

المراجع

[1]- T. Hlaing, "Feature Selection and Fuzzy Decision Tree for Network Intrusion Detection," International Journal of Informatics and Communication Technology (IJ-ICT), vol/issue: 1(2), 2012, pp. 109-118.

[2]- P. A. Estévez, et al., "Normalized Mutual Information Feature Selection," IEEE Transactions on Neural Networks, vol/issue: 20(2), 2009, pp. 189-201.

[3]- Smita Chormunge, Sudarson Jena, "Efficient Feature Subset Selection Algorithm for High Dimensional Data" International Journal of Electrical and Computer Engineering (IJECE) Vol. 6, No. 4, 2016, pp. 1880-1888.

[4]- Liu, H., & Motoda, H. "Feature Extraction, Construction, and Selection, A Data Mining Perspective", Kluwer Academic Publisher. 1998.

[5]- Kohavi, Ron & John, George.H.. "Wrappers for Features Subset Selection. In Artificial Intelligence", 1(2), 1997, pp. 273-324.

[6]- Aha, D. W. "Feature Weighting for Lazy Learning Algorithms. In Feature Extraction, Construction, and Selection, A Data Mining Perspective, Huan, Liu and Hiroshi, Motoda (eds.) Kluwer Academic Publisher, 1998, pp. 13-32.

[7]- Zupan, Blaz and Bohanec, Marko and Demsar, Janez and Bratko, Ivan, "Feature Transformation by Function Decomposition", In IEEE Intelligent Systems, 13(2), 1998, pp. 38-43

[8]- Ming Dong, & Ravi Kothari. "Feature Subset Selection Using a New Definition of Classifiability Computer Science Department, Wayne State University, 2003.

[9]- Omar A. Shiba, Md. Nasir Sulaiman, Ali Mamat & Fatimah Ahmad, "An Efficient and Effective Case Classification Method Based On Slicing". International Journal of The Computer, the Internet and Management, 14(2), 2006, pp15-23.

[10]- J. R. Anderson and M. Matessa. Explorations of an incremental, bayesian algorithm for categorization. Machine Learning, 9, 1992, pp. 275–308.

[11]- Aha, D W. "Feature Selection for Case-Based Classification". AAAI Technical Report WS-94-01, 1994.

[12]- D. W. Aha and R. L. Banker, "A Comparative Evaluation of Sequential Feature Selection Algorithms", Proceeding of the Fifth International Workshop on Artificial Intelligence and Statistics, pp. 1-7, 1995

[13]- Raman, B., & Thomas R.Ioerger, "Instance Based Filter for Feature Selection. Journal of Machine Learning Research, 1, 2002, pp.1-23.

[14]- Murphy, P. M.(1996). UCI Repositories of Machine Learning and Domain Theories [online]. University of California, Irvine Available: http://www.isc.uci.edu/~mlearn/MLRepository.html, [2018, Apr. 12 - Date of brows].