RỪNG NGẪU NHIÊN CẢI TIẾN CHO PHÂN LOẠI DỮ LIỆU GIEN

Huỳnh Phụng Toàn[1], Nguyễn Vũ Lâm[2], Nguyễn Minh Trung[1] và Đỗ Thanh Nghị[3]

ABSTRACT

Our investigation aims to propose random trees to classify gene data which have very small amount of samples in very high dimensions and noise. The random forest algorithm proposed by Breiman is usually suited for classifying very-high-dimensional datasets. However, the classical majority rule of a decision tree degrades the classification accuracy of random forests. We have proposed to improve the classification performance of random forests by using in each leaf of the tree a local class labeling rule instead of the majority rule. The numerical test results on gene datasets from datam.i2r.a-star.edu.sg/datasets/krbd/ showed that that our proposal gives good classification results compared with classical random forests and support vector machine (SVM) in terms of Precision, Recall, F1 and Accuracy.

Keywords: Genes expression classification, Decision trees, Random forests, k nearest neighbors

Title: Improved random forests for classifying gene data

TÓM TẮT

Trong bài viết này, chúng tôi đề xuất giải thuật rừng ngẫu nhiên cải tiến cho phân lớp dữ liệu gien thường có rất ít các phần tử dữ liệu nhưng số chiều rất lớn và có nhiễu. Trong thực tế, giải thuật rừng ngẫu nhiên của Breiman thường được sử dụng cho phân lớp kiểu dữ liệu như dữ liệu gien. Tuy nhiên, do sử dụng luật bình chọn số đông ở nút lá của cây quyết định làm dự báo của rừng ngẫu nhiên bị giảm. Để cải thiện kết quả dự báo của rừng ngẫu nhiên, chúng tôi đề xuất thay thế luật bình chọn số đông bởi luật gán nhãn cục bộ. Kết quả thử nghiệm trên các tập dữ liệu gien từ site datam.i2r.a-star.edu.sg/datasets/krbd/ cho thấy rằng giải thuật rừng ngẫu nhiên cải tiến do chúng tôi đề xuất cho kết quả phân loại tốt khi so sánh với rừng ngẫu nhiên của cây quyết định C4.5 và máy học véctơ hỗ trợ dựa trên các tiêu chí Precision, Recall, F1, Accuracy.

Từ khóa: Phân loại dữ liệu gien, giải thuật học cây quyết định, rừng ngẫu nhiên, k láng giềng

1 GIỚI THIỆU

Phân lớp dữ liệu có số chiều lớn có nhiễu như dữ liệu gien (mỗi chiều cung cấp rất ít thông tin cho tách lớp) được biết là một trong 10 vấn đề khó của cộng đồng khai mỏ dữ liệu (Yang and Wu, 2006). Mô hình học phân lớp thường cho kết quả tốt trong khi huấn luyện lại cho kết quả rất thấp khi dự báo. Vấn đề khó khăn thường gặp chính là số chiều quá lớn lên đến hàng nghìn chiều thậm chí đến cả triệu và dữ liệu thường tách rời nhau trong không gian có số chiều lớn việc tìm mô hình phân lớp tốt có khả năng làm việc với dữ liệu có số chiều lớn là khó khăn do có quá

[1] Bộ môn Tin Học Ứng Dụng, khoa Khoa Học Tự Nhiên, Trường Đại học Cần Thơ
[2] Trung tâm Tin Học-Công Nghệ Phần Mềm, Trường Cao Đẳng Cộng Đồng Kiên Giang
[3] Bộ môn Khoa Học Máy Tính, khoa CNTT&TT, Trường Đại học Cần Thơ

nhiều khả năng lựa chọn mô hình. Việc tìm một mô hình phân lớp hiệu quả (phân lớp dữ liệu tốt trong tập thử) trong không gian giả thiết lớn là vấn đề khó. Đã có hai lớp giải thuật tiêu biểu là máy học véc tơ hỗ trợ của Vapnik (SVM (Vapnik, 1995)) và rừng ngẫu nhiên của (Breiman, 2001) được biết đến như là những giải thuật phân lớp hiệu quả các tập dữ liệu có số chiều lớn như dữ liệu gien.

Tiếp cận rừng ngẫu nhiên cho độ chính xác cao khi so sánh với các thuật toán học có giám sát hiện nay, bao gồm cả AdaBoost (Freund and Schapire, 1995), ArcX4 (Breiman, 1998) và SVM (Vapnik, 1995). Khi xử lý dữ liệu có số chiều lớn và số phần tử ít như dữ liệu gien thì rừng ngẫu nhiên và SVM là hai giải thuật học nhanh, chịu đựng nhiễu tốt và không bị tình trạng học vẹt, điều này ngược lại với AdaBoost, ArcX4 rất dễ bị học vẹt và ảnh hưởng lớn với nhiễu (Grove and Schuurmans, 1998). Tuy nhiên, luật quyết định ở nút lá của các cây trong rừng ngẫu nhiên dựa vào luật bình chọn số đông, điều này dẫn đến độ chính xác của giải thuật rừng ngẫu nhiên bị giảm khi phân lớp dữ liệu. Để khắc phục nhược điểm trên, chúng tôi đề xuất thay thế luật bình chọn số đông ở nút lá bằng luật gán nhãn cục bộ dựa trên giải thuật k láng giềng (Fix and Hodges, 1952). Giải thuật rừng ngẫu nhiên cải tiến do chúng tôi đề xuất thường cho kết quả phân lớp chính xác hơn so với giải thuật gốc. Kết quả thử nghiệm trên các tập dữ liệu gien (Jinyan and Huiqing, 2002) cho thấy rằng giải thuật rừng ngẫu nhiên cải tiến do chúng tôi đề xuất cho kết quả phân loại tốt khi so sánh với rừng ngẫu nhiên của cây quyết định C4.5 và máy học véctơ hỗ trợ dựa trên các tiêu chí Precision, Recall, F1, Accuracy.

Trong phần 2, chúng tôi sẽ trình bày tóm tắt giải thuật rừng ngẫu nhiên và giải thuật cải tiến cho phân lớp. Kết quả thực nghiệm sẽ được trình bày trong phần 3 trước phần kết luận và hướng phát triển trong phần 4.

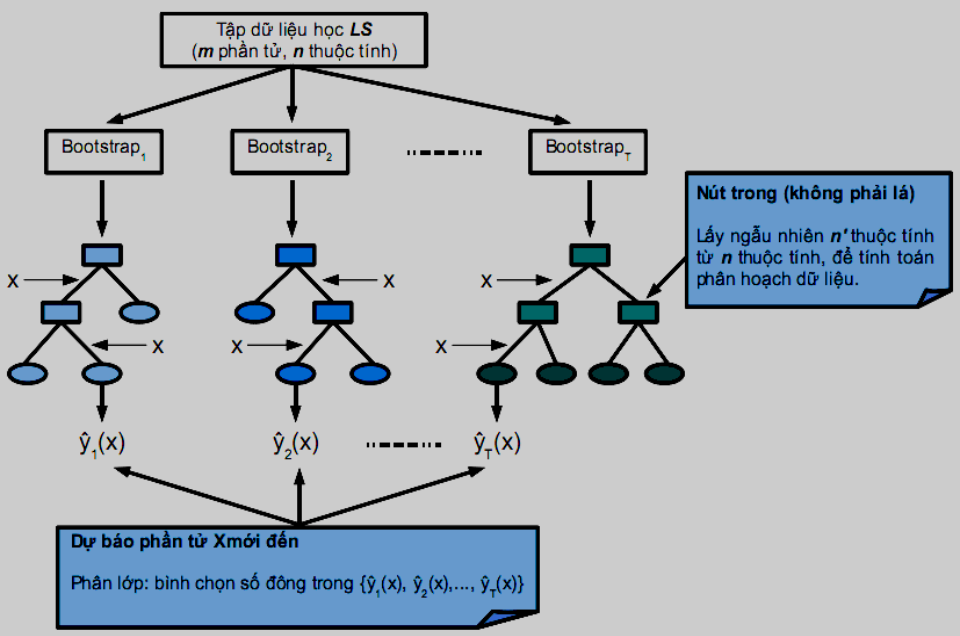
2 RỪNG NGẪU NHIÊN CẢI TIẾN CHO PHÂN LỚP DỮ LIỆU

Từ những năm 1990, cộng đồng máy học đã nghiên cứu cách để kết hợp nhiều mô hình phân loại thành tập hợp các mô hình phân loại để cho tính chính xác cao hơn so với chỉ một mô hình phân loại. Mục đích của các mô hình tập hợp là làm giảm variance và/hoặc bias của các giải thuật học. Bias là khái niệm về lỗi của mô hình học (không liên quan đến dữ liệu học) và variance là lỗi do tính biến thiên của mô hình so với tính ngẫu nhiên của các mẫu dữ liệu học. (Buntine, 1992) đã giới thiệu các kỹ thuật Bayes để giảm variance của các phương pháp học. Phương pháp xếp chồng (Wolpert, 1992) hướng tới việc cực tiểu hóa bias của các giải thuật học. Trong khi (Freund and Schapire, 1995) đưa ra Boosting, (Breiman, 1998) đề nghị ArcX4 để cùng giảm bias và variance, còn Bagging (Breiman, 1996) thì giảm variance của giải thuật học nhưng không làm tăng bias quá nhiều. Tiếp cận rừng ngẫu nhiên (Breiman, 2001) là một trong những phương pháp tập hợp mô hình thành công nhất. Giải thuật rừng ngẫu nhiên xây dựng cây không cắt nhánh nhằm giữ cho bias thấp và dùng tính ngẫu nhiên để điều khiển tính tương quan thấp giữa các cây trong rừng.

Rừng ngẫu nhiên (được mô tả trong hình 1) tạo ra một tập hợp các cây quyết định không cắt nhánh, mỗi cây được xây dựng trên tập mẫu bootstrap (lấy mẫu ngẫu

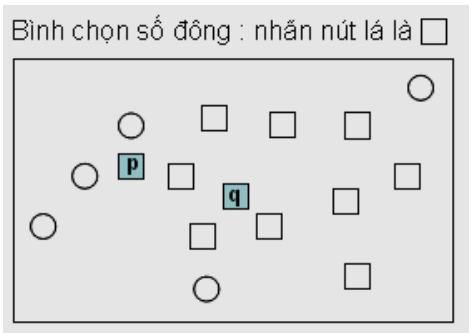
nhiên có hoàn lại), tại mỗi nút phân hoạch tốt nhất được thực hiện từ việc chọn ngẫu nhiên một tập con các thuộc tính.

Lỗi tổng quát của rừng phụ thuộc vào độ chính xác của từng cây thành viên trong rừng và sự phụ thuộc lẫn nhau giữa các cây thành viên. Giải thuật rừng ngẫu nhiên cho độ chính xác cao khi so sánh với các thuật toán học có giám sát hiện nay, chịu đựng nhiễu tốt.



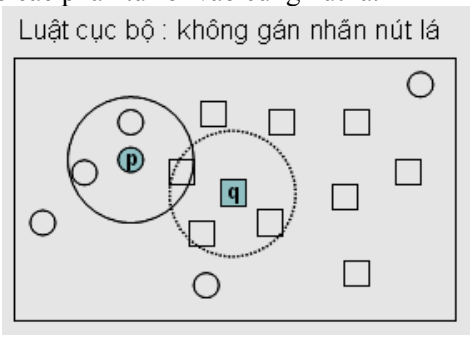
Hình 1: Giải thuật rừng ngẫu nhiên

Tuy nhiên, nếu chúng ta trở lại luật gán nhãn ở nút lá của các cây quyết định trong rừng ngẫu nhiên, hai giải thuật cây quyết định phổ biến là CART (Breiman *et al.*, 1984) và C4.5 (Quinlan, 1983) thường dùng chiến lược bình chọn số đông. Thời điểm xây dựng cây quyết định, nếu nút lá có chứa các phần tử dữ liệu của các lớp không thuần nhất, việc gán nhãn cho nút lá được tính cho nhãn của lớp có số lượng phần tử lớn nhất chứa trong nút lá. Xét ví dụ như hình 2, nút lá có chứa 14 phần tử trong đó lớp hình vuông có 9 phần tử và lớp hình tròn có 5 phần tử. Nút lá sẽ được gán nhãn là hình vuông do số phần tử lớp hình vuông nhiều hơn hình tròn. Chiến lược gán nhãn này làm cho luật quyết định không được chính xác. Khi phân lớp, phần tử nào rơi vào nút lá đều được gán nhãn của nút lá. Vì vậy, phần tử p, q được gán nhãn là vuông. Hiệu quả phân lớp không cao (phần tử p có thể sai).



Hình 2: Luật bình chọn số đông cho gán nhãn ở nút lá của cây quyết định, nút lá có nhãn là vuông, nên điểm p và q đều được phân lớp vuông

Để nâng cao hiệu quả phân lớp của cây quyết định trong giải thuật rừng ngẫu nhiên, chúng tôi đề xuất thay thế luật gán nhãn trên cơ sở bình chọn số đông bởi luật gán nhãn cục bộ sử dụng giải thuật k láng giềng (Fix and Hodges, 1952). Thay vì việc gán nhãn ở nút lá được thực hiện khi xây dựng cây, chúng tôi trì hoãn việc gán nhãn này lại. Nghĩa là nút lá vẫn chưa được gán nhãn. Chúng tôi chỉ thực hiện việc gán nhãn trong khi dự báo phần tử mới đến. Xét tại nút lá như hình 3 vẫn chưa được gán nhãn. Với luật quyết định cục bộ dựa trên 3 láng giềng. Khi phần tử dữ liệu mới đến chẳng hạn như p và q, rơi vào nút lá; chúng tôi thực hiện tìm 3 phần tử trong nút lá gần nhất với dữ liệu mới đến, sau đó mới thực hiện việc gán nhãn cho phần tử cần dự báo được dựa trên nhãn của các láng giềng. Khi phân lớp, phần tử p rơi vào nút lá, chúng ta tìm 3 láng giềng của p, gán nhãn cho p dựa trên bình chọn số đông từ 3 láng giềng, nhãn của p được gán là tròn. Tương tự, phần tử q được gán nhãn là vuông từ bình chọn số đông từ 3 láng giềng của nó. Luật quyết định này giúp cho việc phân lớp của cây đạt chính xác cao hơn vì trong chiến lược này, mặc dù các phần tử dự báo rơi vào cùng nút lá nhưng nhãn của nó có thể khác nhau trong khi chiến lược bình chọn số đông thường sử dụng trong cây quyết định lại gán cùng nhãn cho các phần tử rơi vào cùng nút lá.



Hình 3: Luật cục bộ sử dụng 3 láng giềng, nút lá chưa gán nhãn, điểm p, q được gán nhãn lần lượt là tròn, vuông dựa trên bình chọn số đông của 3 láng giềng

Việc cải tiến đơn giản chỉ thực hiện ở luật gán nhãn của cây quyết định và các chi tiết còn lại của giải thuật rừng ngẫu nhiên vẫn được giữ lại. Rừng ngẫu nhiên vì thế được cải thiện độ chính xác khi phân lớp dữ liệu gien có số chiều rất lớn.

3 KẾT QUẢ THỰC NGHIỆM

Để có thể đánh giá hiệu quả của giải thuật rừng ngẫu nhiên cải tiến, chúng tôi cài đặt giải thuật rừng ngẫu nhiên cây quyết định C4.5 (RF-C4.5) và giải thuật cải tiến (iRF-C4.5) bằng ngôn ngữ lập trình C/C++ có kế thừa từ mã nguồn của C4.5 được cung cấp bởi (Quinlan, 1993). Dữ liệu gien chúng tôi chạy thử nghiệm, có số chiều rất lớn, được lấy tại (Jinyan and Huiqing, 2002). Bên cạnh đó, chúng tôi quan sát kết quả của giải thuật cải tiến iRF-C4.5 trong thực nghiệm bằng cách so sánh với rừng ngẫu nhiên của cây quyết định C4.5, RF-C4.5 và máy học LibSVM (Chang and Lin, 2001). Tất cả các kết quả đều được thực hiện trên một máy tính cá nhân chạy hệ điều hành Linux.

Bảng 1: Mô tả các tập dữ liệu gien

ID	Tập dữ liệu	Số phần tử	Số chiều	Lớp	Nghi thức
1	ALL-AML-Leukemia	72	7129	ALL, AML	trn-tst
2	MLL-Leukemia	72	12582	MLL, rest	trn-tst
3	Breast Cancer	97	24481	relapse, non-relapse	trn-tst
4	Prostate Cancer	136	12600	cancer, normal	trn-tst
5	Lung Cancer	181	12533	cancer, normal	trn-tst
6	Diffuse Large B-Cell Lymphoma	47	4026	germinal, activated	loo
7	Subtypes of Acute Lymphoblastic (Hyperdip)	327	12558	Hyperdip, rest	trn-tst
8	Subtypes of Acute Lymphoblastic (TEL-AML1)	327	12558	TEL-AML1, rest	trn-tst
9	Subtypes of Acute Lymphoblastic (T-ALL)	327	12558	TEL-ALL, rest	trn-tst
10	Subtypes of Acute Lymphoblastic (Others)	327	12558	Others, diagnostic groups	trn-tst

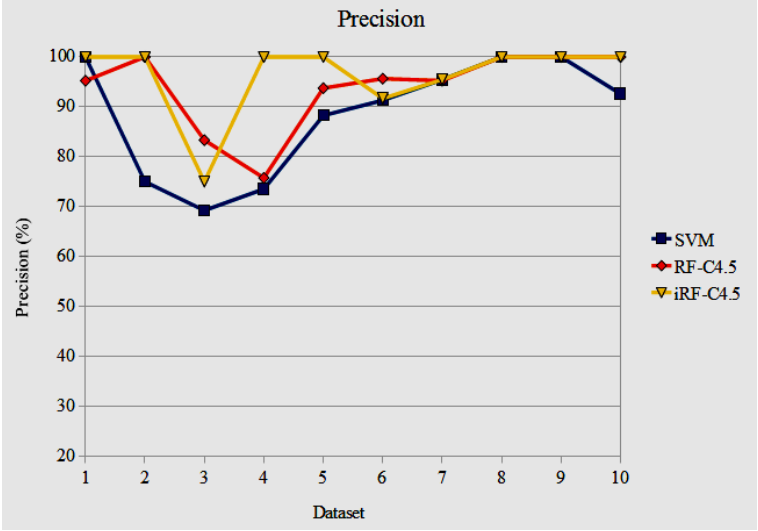
Chúng tôi tiến hành thực nghiệm trên 10 tập dữ liệu gien có số chiều rất lớn từ kho dữ liệu sinh-y học. Mô tả các tập dữ liệu được tìm thấy trong bảng 1. Chúng tôi chú ý đến các phương pháp kiểm tra được liệt kê trong cột cuối của bảng 1. Với những tập dữ liệu có sẵn tập học và tập thử, chúng tôi dùng tập học để thử điều chỉnh các tham số ở đầu vào của các giải thuật nhằm thu được độ chính xác tốt khi học. Sau đó, dùng mô hình thu được để phân lớp tập thử. Nếu tập học và tập thử không có sẵn, các giao thức kiểm tra chéo (cross-validation protocol) để đánh giá. Do các tập dữ liệu có ít hơn 300 phần tử, chúng tôi dùng giao thức kiểm tra chéo leave-one-out (loo). Tức là dùng một phần tử trong tập dữ liệu để thử, các phần tử khác dùng để học. Lặp lại đến khi tất cả các phần tử đều được dùng để thử một lần.

Để thấy rõ hơn tính hiệu quả của iRF-C4.5 so với RF-C4.5 và LibSVM, chúng tôi tiến hành phân tích hiệu quả của các thuật toán phân lớp dựa trên các tiêu chí như Precision, Recall, F1-measure và Accuracy (van Rijsbergen, 1979). Precision của một lớp là số phần tử dữ liệu được phân lớp đúng về lớp này chia cho tổng số phần tử dữ liệu được phân về lớp này. Recall của một lớp là số phần tử dữ liệu được phân lớp đúng về lớp này chia cho tổng số phần tử dữ liệu của lớp. F1-measure là tổng hợp của Precision và Recall, và được định nghĩa là hàm trung bình điều hòa giữa hai giá trị Precision và Recall ($F1 = \frac{2 \times \mathrm{Pr}ecision \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call}$). Độ chính xác Accuracy là số điểm dữ liệu được phân lớp đúng của tất cả các lớp chia cho tổng số điểm dữ liệu. Khi xây dựng mô hình, các giải thuật rừng ngẫu nhiên xây dựng 200 cây quyết định và luật cục bộ là 2 láng giềng (lý do chọn 200 cây để đảm bảo được độ chính xác cao). Riêng máy học SVM chỉ cần sử dụng hàm nhân tuyến tính là phân lớp tốt nhất các tập dữ liệu gien. Chúng tôi thu được kết quả như trình bày trong bảng 2.

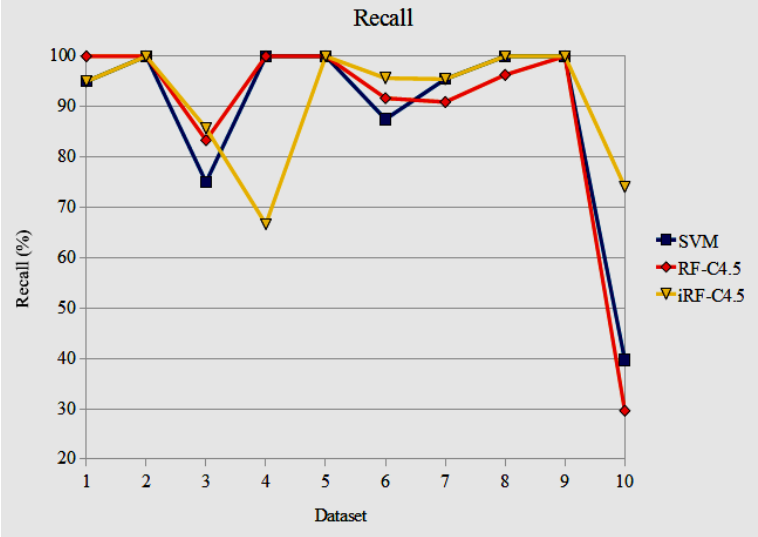
Bảng 2: Kết quả phân lớp của LibSVM, RF-C4.5 và iRF-C4.5

ID	Precision			Recall			F1-measure		
	Lib-SVM	RF-C4.5	iRF-C4.5	Lib-SVM	RF-C4.5	iRF-C4.5	Lib-SVM	RF-C4.5	iRF-C4.5
1	**100**	95.24	**100**	95	**100**	95	97.44	**97.56**	97.44
2	75	**100**	**100**	**100**	**100**	**100**	85.71	**100**	**100**
3	69.23	**83.33**	75	75	83.33	**85.71**	72	**83.33**	80
4	73.53	75.76	**100**	**100**	**100**	66.67	84.75	**86.21**	75
5	88.26	93.75	**100**	**100**	**100**	**100**	93.75	96.77	**100**
6	91.3	**95.65**	91.67	87.5	91.67	**95.65**	89.36	**93.62**	**93.62**
7	**95.45**	95.24	**95.45**	**95.45**	90.91	**95.45**	**95.45**	93.02	**95.45**
8	**100**	**100**	**100**	**100**	96.3	**100**	**100**	98.11	**100**
9	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
10	92.59	**100**	**100**	39.68	29.63	**74.07**	55.56	45.71	**76.92**

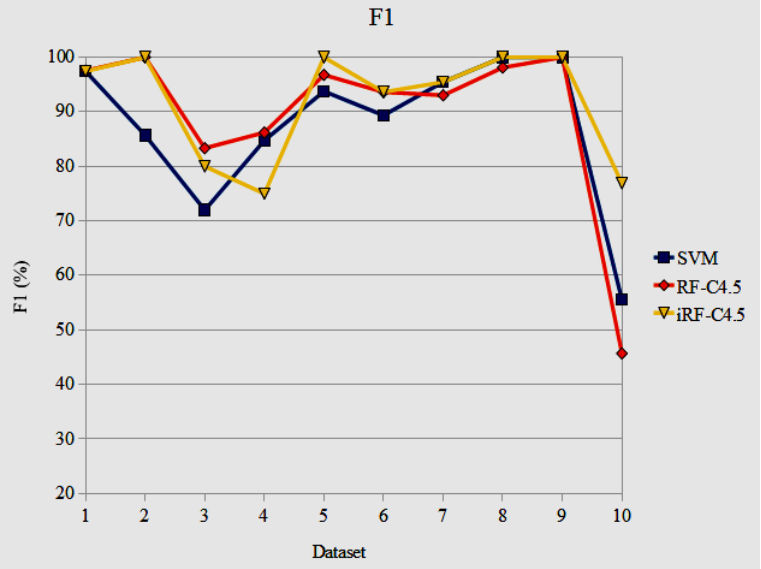
Nhìn vào bảng kết quả phân lớp để so sánh hiệu quả của giải thuật LibSVM, RF-C4.5 và iRF-C4.5, chúng ta có thể thấy rằng với tiêu chí Precision, giải thuật iRF-C4.5 cho kết quả tốt nhất 8/10 tập dữ liệu. Khi so sánh dựa vào tiêu chí Recall, iRF-C4.5 cũng cao hơn chiếm 8/10 tập dữ liệu.



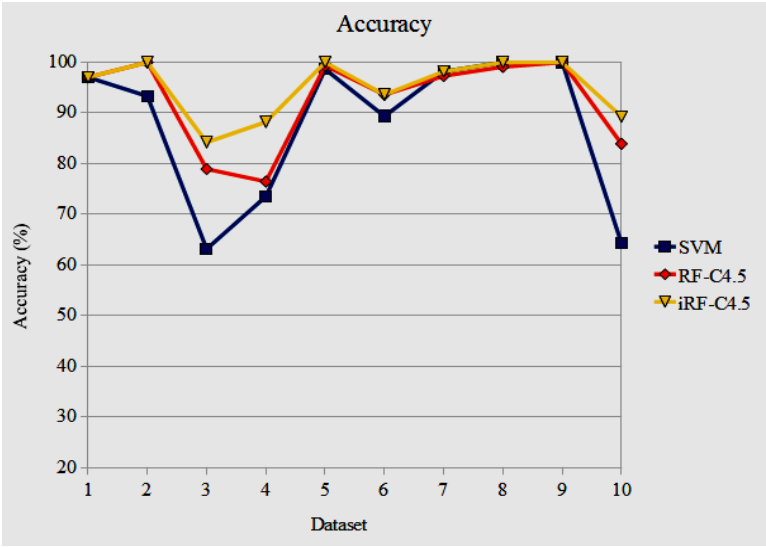
Đồ thị 1: So sánh tiêu chí Precision của 3 giải thuật trên 10 tập dữ liệu



Đồ thị 2: So sánh tiêu chí Recall của 3 giải thuật trên 10 tập dữ liệu



Đồ thị 3: So sánh tiêu chí F1 của 3 giải thuật trên 10 tập dữ liệu



Đồ thị 4: So sánh tiêu chí Accuracy của 3 giải thuật trên 10 tập dữ liệu

Xét trên tiêu chí F1 (trung bình điều hòa giữa hai giá trị Precision và Recall), iRF-C4.5 cho kết quả tốt nhất chiếm 7/10 tập dữ liệu so với LibSVM và RF-C4.5.

Nhìn vào đồ thị 1 trình bày trực quan so sánh kết quả Precision của LibSVM, RF-C4.5 và iRF-C4.5. Tương tự, các đồ thị 2, 3, 4 lần lượt trình bày so sánh kết quả Recall, F1 và Accuracy.

4 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày giải thuật rừng ngẫu nhiên cải tiến cho phép phân lớp hiệu quả các tập dữ liệu gien có rất ít số phần tử nhưng có số chiều rất lớn và mỗi chiều

cung cấp rất ít thông tin cho việc phân lớp. Ý tưởng xuất phát từ giải thuật rừng ngẫu nhiên do Breiman đề xuất, chúng tôi đề xuất thay thế luật bình chọn số đông cho việc gán nhãn ở nút lá bằng luật quyết định cục bộ dựa vào giải thuật k láng giềng. Kết quả thực nghiệm trên các tập dữ liệu gien cho thấy rằng giải thuật cải tiến iRF-C4.5 cho kết quả tốt trên tiêu chí về Precision, Recall, F1 và độ chính xác Accuracy khi so sánh với giải thuật gốc rừng ngẫu nhiên RF-C4.5 và cả giải thuật LibSVM. Kết quả thực nghiệm cho phép chúng tôi tin tưởng giải thuật rừng ngẫu nhiên cải tiến phân lớp hiệu quả các tập dữ liệu gien có số chiều lớn.

Trong tương lai, chúng tôi tiếp tục nghiên cứu các luật quyết định cục bộ dựa trên các giải thuật hiệu quả hơn k láng giềng. Ngoài nghiên cứu cải thiện chất lượng mô hình phân lớp, chúng tôi hứa hẹn sẽ tập trung cho cải tiến tốc độ học và phân lớp của giải thuật.

TÀI LIỆU THAM KHẢO

A.J. Grove and D. Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), 1998, pp. 692–699.

C.C. Chang and C.J. Lin. Libsvm – a library for support vector machines. 2001. http://www.csie.ntu.edu.tw/cjlin/libsvm.

C.V. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

D. Wolpert. Stacked generalization. *Neural Networks 5*, 1992, pp. 241–259.

Fix, E. and Hodges, J.: Discriminatoiry Analysis: Small Sample Performance. Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, USA, 1952.

J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

L. Breiman, J.H. Friedman, R.A. Olshen and C. Stone. *Classification and Regression Trees*. Wadsworth International, 1984.

L. Breiman. Arcing classifiers. *The annals of statistics*, 26(3): 801–849, 1998.

L. Breiman. Bagging predictors. *Machine Learning 24(2)*:123–140, 1996.

L. Breiman. Random forests. *Machine Learning 45(1)*:5–32, 2001.

L. Jinyan and L. Huiqing. Kent ridge bio-medical dataset repository. 2002, http://datam.i2r.a-star.edu.sg/datasets/krbd/.

Q. Yang and X. Wu. 10 Challenging Problems in Data Mining Research. *Journal of Information Technology and Decision Making 5(4)*:597-604, 2006.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

W. Buntine. Learning classification trees. *Statistics and Computing 2*, 1992, pp. 63–73.

Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Computational Learning Theory*, 1995, pp. 23–37.