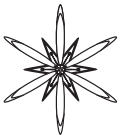
Modern Phytomorphology **6**: 155–160, 2014



УДК: 581.4: 004.94

МАШИННАЯ КЛАССИФИКАЦИЯ ВИДОВ РОДА *FICUS* L. НА ОСНОВЕ ФОРМ ЛИСТОВЫХ ПЛАСТИНОК

Александр З. Глухов и Иван И. Стрельников *

Аннотация. Актуальной задачей ботанической науки является разработка методов компьютерного определения видов растений. Идентификацию часто проводят на основе морфологии листовых пластинок. Перспективным является описание формы листьев через значения гармоник эллиптического Фурье разложения, но эффективность этого подхода требует дополнительной проверки. Необходима сравнительная оценка различных алгоритмов классификации. Работу проводили на основе 2812 изображений листьев 15 видов рода *Ficus* L. Для решения обозначенных задач установили оптимальный набор параметров проведения Фурье разложения. Определили, что наилучшие результаты классификации достигаются при использовании 18 гармоник Фурье. Количество опорных точек на контуре не влияло на результат выполнения моделей. Сравнили точность идентификации 30 классификационных алгоритмов. Самой высокой точностью классификации (98%) обладал алгоритм «случайного леса». Объединение нескольких классификационных алгоритмов путем «стогования» позволяет повысить эффективность распознавания форм листьев.

Ключевые слова: *Ficus,* форма листовой пластинки, эллиптический Фурье анализ, классификационные модели

*Донецкий ботанический сад НАН Украины, пр. Ильича, 110, г. Донецк, 83059, Украина; * ivanstrel87@gmail.com*

Введение

Определение вида растения является процедурой, широко востребованной в ботанической практике. Обычно, это рутинный процесс, требующий участия специалиста и не лишенный недостатков субъективного восприятия характеристик растения. Популярные сегодня молекулярные методы идентификации остаются малодоступными и дорогостоящими. В результате, сложился высокий интерес к разработке новых автоматических, мало затратных методов определения растений. Активное развитие прикладных отраслей компьютерного зрения, обработки изображений и распознавания образов, обеспечило перспективность идентификации растений на основе их изображений. Легкость получения цифровых фотографий и успехи в области искусственного интеллекта делают эти подходы доступными и эффективными. В приложениях распознавания образов форма объектов выступает наиболее информативным признаком. В рамках компьютерной классификации растений наибольший интерес представляет форма листовой пластинки, прежде всего из-за легкости отображения объекта в двухмерном пространстве и высокой стабильности признака в пределах вида.

В практической реализации возникают две базовые задачи: подбор оптимального метода описания формы и выбор алгоритма классификации. Для описания формы чаще всего используются классические морфометрические параметры на основе линейных размеров, периметра и площади (Wang 2005; Lee & Chen 2006; Singh *et. al.* 2010). В последние годы наблюдается растущий интерес к методам интерпретации формы через анализ конформации точек, формирующих контур объекта. Наиболее разработанным является эллиптическое Фурье разложение форм (Claude 2008). В работе J. Neto (2006), показана потенциально высокая эффективность классификации видов растений на основе дескрипторов Фурье. При этом нерешенной остается задача подбора оптимальной

комбинации параметров Фурье разложения. Отсутствуют данные и о сравнительной эффективности разных классификационных алгоритмов.

Основываясь на вышесказанном, были поставлены следующие задачи: определить оптимальное для классификации количество гармоник и опорных точек при Фурье разложении форм; дать сравнительную оценку 30 распространенных классификационных алгоритмов; проверить возможность применения процедуры стогования предсказательных моделей для повышения эффективности идентификации растений.

Материалы и методы исследований

Классификацию проводили на 2812 изображениях листовых пластинок принадлежащих к 15 видам рода *Ficus* L.: *F. benjamina* L., *F. binnendijkii* Miq., *F. craterostoma* Warb. ex Mildbr., *F. cyathistipula* Warb, *F. elastica* Roxb. ex Hornem., *F. macrophylla* Roxb., *F. microcarpa* L. f., *F. natalensis* Hochst. subsp. *leprieurii* (Miq.) C.C. Berg, *F. pumila* L., *F. religiosa* L., *F. rubiginosa* Desf. ex Vent., *F. sycomorus* L., *F. thonningii* Blume, *F. vallis-choudae* Delile, *F. watkinsiana* F.M. Bailey. Предварительную обработку сканированных изображений проводили в среде пакета FIJI (Schindelin *et al.* 2012). На этом этапе получали бинарные изображения листовых пластинок. Дальнейшую обработку осуществляли с использованием языка программирования R (R Core Team 2012). Общая схема получения гармоник эллиптического Фурье разложения соответствовала рекомендациям (Claude 2008).

Так как задача классификации видов растений подразумевает использование листьев с существенно различающимся морфологическим строением, была выбрана схема на базе псевдоопорных точек. В данном варианте, эффективность обучения классификационной модели может зависеть от двух параметров: количества опорных точек вдоль контура и количества определяемых гармоник Фурье разложения. Для нахождения оптимального соотношения этих параметров подготавливали исходный набор контуров листовых пластинок в шести вариантах, с расстановкой по 30, 40, 60, 100, 180 и 300 опорных точек на каждом контуре. После, для каждого варианта находили по 12, 18, 24, 30 гармоник эллиптического Фурье разложения. В результате получили 24 комбинации параметров. Значения гармоник из каждой комбинации использовали в качестве исходных данных для построения классификационных моделей. На этом этапе тестировали выполнение четырех распространенных алгоритмов: «опорная векторная машина с радиальным базисом», «случайный лес деревьев решений», «искусственная нейронная сеть» и «простой Бейес». Для классификации использовали схему обучения без контроля. Эффективность выполнения модели оценивали по результатам 5 случайных повторений 10-кратной перекрестной проверки. Оптимальную комбинацию определяли по результатам множественного дисперсионного анализа, значения гармоник этой комбинации использовали в дальнейшем.

На следующем этапе оценивали эффективность классификации 30 алгоритмов. Подбор оптимальных макропараметров осуществляли с применением библиотеки функций caret (Kuhn 2008). Перечень алгоритмов представлен в Табл. 1.

Для построения классификационных моделей использовали метод обучения с контролем. Для этого исходный набор данных делили на тренировочную и тестовую выборки по 75% и 25% от начального размера, соответственно. Подбор макропараметров и первичную оценку эффективности классификации проводили на основе 10-кратной перекрестной проверки. Вариант алгоритма с оптимальными параметрами обучали на всей тренировочной выборке.

Процедуру стогования моделей проводили на основе (Wolpert 1992).

Табл. 1. Перечень алгоритмов. * Сокращение названия алгоритма принятое в библиотеке caret языка программирования R.

Table 1. List of the algorithms. * Algorithm names abbreviations as accepted in the caret library of the R programming language.

№	Метод	Алгоритм*	Источник
1	Бэгинг (гибкий дискриминантный анализ)	bagFDA	Friedman 1991
2	Бэгинг (регрессионное дерево решений)	treebag	Hothorn *et. al.* 2004
3	Бейес метод	nb	Kuhn 2008
4	Деревья с поддержкой	gbm	Friedman 2002
5	Эластичная сеть	glmnet	Simon 2011
6	Гибкий дискриминантный анализ	fda	Hastie 2009
7	Общая линейная модель	glmStepAIC	Venables & Ripley 2002
8	Гетероскедастический дискриминантный анализ	had	Kumar & Andreou 1998
9	К ближайших соседей	knn	Venables & Ripley 2002
10	Обучаемая векторная дискретизация	lvq	Venables & Ripley 2002
11	Линейный дискриминантный анализ	sddaLDA	Kuhn 2008
12	Смешанный дискриминантный анализ	mda	Hastie 2009
13	Ближайших сходящихся центров	pam	Tibshirani *et. al.* 1999
14	Нейронная сеть	avNNet	Kuhn 2008
15	Нейронная сеть	nnet	Venables & Ripley 2002
16	Частичных наименьших квадратов	pls	Martens & Næs 1989
17	Дискриминантный анализ со штрафами	Pda	Hastie 2009
18	Дискриминантный анализ со штрафами	Pda2	Hastie 2009
19	Квадратичный дискриминантный анализ	sddaQDA	Kuhn 2008
20	Сеть функций с радиальным базисом	rbf	Kuhn 2008
21	Случайный лес	ORFridge	Menze *et. al.* 2011
22	Случайный лес	rf	Breiman 2001
23	Рекурсивное деление	ctree	Strobl *et. al.* 2009
24	Рекурсивное деление	rpart	Breiman *et. al.* 1984
25	Регулярный дискриминантный анализ	rda	Press *et. al.* 1992
26	Модель правил	C5.0Rules	Quinlan 1993
27	Самоорганизующиеся карты	bdk	Kuhn 2008
28	Рассеянный линейный дискриминантный анализ	sparseLDA	Phatak *et. al.* 2010
29	Машина опорных векторов	svmPoly	Kuhn 2008
30	Машина опорных векторов	svmRadial	Kuhn 2008

Отобрали 15 алгоритмов, проявивших наилучшую эффективность по результатам перекрестной проверки. Провели предсказание видов по значениям гармоник тренировочного множества всеми отобранными моделями. Результаты представили в виде вероятностей отнесения каждого из контуров листовой пластинки к каждому из 15 видов. Объединили предсказания всех моделей в одну таблицу, в результате каждому абрису соответствовал набор из 225 предсказанных значений. Полученное множество использовали при повторном обучении всех алгоритмов по

Рис. 1. Влияние количества гармоник на точность идентификации контуров листовых пластинок: **n** – количество наблюдений.

Fig. 1. The influence of the harmonics number on the accuracy of the leaf plate's identification: **n** – number of observations.

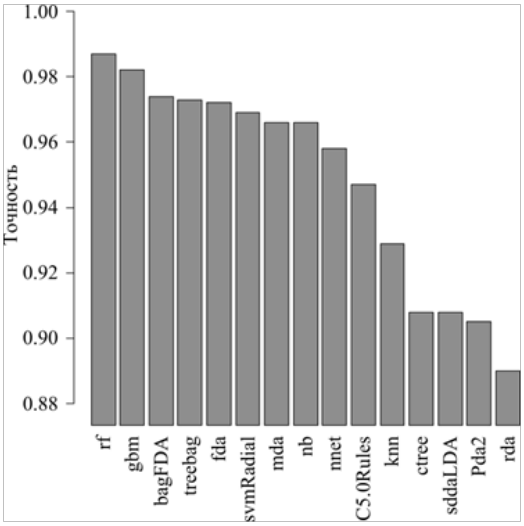


Рис. 2. Точность идентификации 15 лучших алгоритмов.

Fig. 2. The identification accuracy of the top 15 algorithms.

вышеописанной схеме. Для сравнения классификационной эффективности моделей индивидуальных алгоритмов и результатов стогования моделей сравнивали показатели перекрестной проверки и результаты классификации на тестовой выборке.

Результаты и их обсуждение

Оценили влияние количества опорных точек на контуре и количества находимых гармоник на эффективность идентификации вида. Согласно результатам дисперсионного анализа количество точек не влияет значимо на конечный результат обучения моделей. Поэтому для дальнейшего анализа использовали среднее значение – 100 точек. Количество гармоник Фурье разложения достоверно влияло на качество идентификации видов. Результаты теста по сумме выполнения всех классификационных моделей представлены на Рис. 1.

Из Рис. 1 видно, что наилучший результат идентификации достигается при использовании 18 гармоник. Взаимодействие двух параметров не обнаружено. Предварительная оценка значимости гармоник Фурье разложения показала, что для всех видов первые десять гармоник описывают не меньше 99% вариации форм. Это свидетельствует о значимости мелких деталей очертаний листа для точного определения вида. Можно предположить, что гармоники, следующие после восемнадцатой, не содержат специфической информации, то есть описывают шумность данных и не должны включаться в анализ.

Оценка эффективности идентификации индивидуальных алгоритмов показала, что 17 из них имеют точность выше 90%. Ошибки классификации этих моделей распределены равномерно. Следовательно, в выборке отсутствуют принципиально неразличимые виды. На рисунке 2 представлены показатели точности 15 лучших моделей.

Наилучший результат классификации продемонстрировал алгоритм rf – «случайный лес деревьев решений». Средний показатель точности этой модели по результатам 10-кратной перекрестной проверки составил 0,987 со стандартным отклонением в 0,01. Алгоритмы rbf,

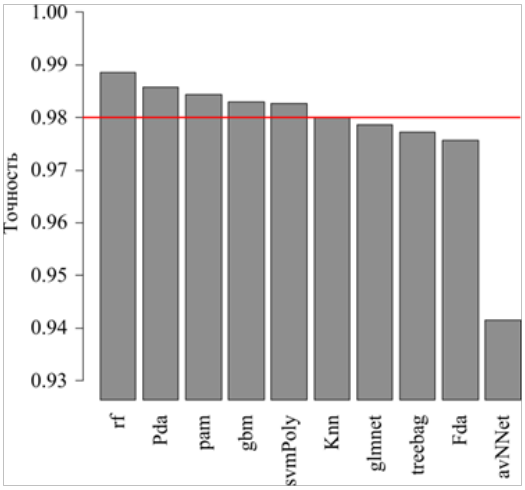


Рис. 3. Точность идентификации 10 лучших моделей после процедуры стогования. Красной линией обозначена максимальная точности индивидуальных моделей.

Fig. 3. The identification accuracy of the top 10 models after the stacking procedure. Red line assigns maximal accuracies for individual models.

ORFridge, glmStepAIC и sparseLDA показали точность меньше 0.5. Из предсказаний 15 лучших моделей сформировали новый обучающее множество и провели на нем повторное обучение моделей.

Согласно результатам перекрестной проверки предложенная схема позволила существенно улучшить качество классификационных моделей. Алгоритмы fda, gbm, glmnet, pam, Pda, rf, svmPoly и treebag продемонстрировали 100% среднюю точность идентификации с нулевым стандартным отклонением. Остальные модели кроме pls, rpart, ORFridge и rda имели точность выше 90%. Для объективной оценки улучшения классификаций проверили индивидуальные и множественные модели на тестовом множестве. Точность отдельных алгоритмов снизилась несущественно. Это свидетельствует об отсутствии эффекта переобучения. Соотношение между разными алгоритмами по точности осталось таким же, как и в случае с перекрестной оценкой. Самый высокий результат в 98% точных идентификаций был у rf. Сравнение результатов 10 лучших множественных моделей с максимальным показателем индивидуальных алгоритмов представлено на рисунке 3.

Результаты пяти алгоритмов после «процедуры стогования» превзошли максимальную точность индивидуальных моделей. Эффективность rf во множественном варианте составила 0,9886. Таким образом, стогование улучшило точность определения на 0,886. В масштабах выборки (701 контур) этот показатель может быть воспринят как несущественный. В абсолютных значениях индивидуальный rf допустил 14 ошибок, а rf после стогования – 8. Следовательно, количество неправильных классификаций уменьшилось на 43%.

Заключение

Гармоники эллиптического Фурье разложения форм листовых пластинок являются информативным параметром для проведения машинной идентификации видов растений. Оптимальные результаты классификации достигаются при использовании 18 гармоник. Высокая эффективность моделей подтверждает устойчивость признаков формы листьев в пределах вида. Наиболее перспективным алгоритмом для автоматического определения растений является случайный лес деревьев решений. Впервые применили метод стогования моделей в сфере компьютерной идентификации растений. Данный вариант объединения предсказаний способен значительно улучшить качество определения видов.

Цитируемые источники

Breiman L. 2001. Random forests. *Mach. Learn* **45 (1)**: 5–32.

Breiman L., Friedman J., Olshen R., Stone C. 1984. Classification and regression trees. Wadsworth.

Claude J. 2008. Morphometrics with R. Springer, New York.

Friedman J. 1991. Multivariate adaptive regression splines. *Ann. Stat.* **19 (1)**: 1–141.

Friedman J.H. 2002. Stochastic gradient boosting. *Comp. Stat. Data A.* **38 (4)**: 367–378.

Hastie N. 2009. Elements of statisical learning – data mining, inference and prediction (2nd edition). Springer, New York.

Hothorn T., Lausen B., Benner A., Radespiel-Troger M. 2004. Bagging survival trees. *Stat. Med.* **23 (1)**: 77–91.

Kuhn M. 2008. Building predictive models in R using the caret package. *J. Stat. Soft.* **28 (5)**: 1–26.

Kumar N., Andreou A. 1998. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Commun.* **25**: 283–297.

Lee C.-L., Chen S.-Y. 2006. Classification of leaf. *Int. J. Imaging Sys. Tech.* **16 (1)**: 15–23.

Martens H., Næs T. 1989. Multivariate calibration. Wiley, Chichester.

Menze B.H., Kelm B.M., Splitthoff D.N., Koethe U., Hamprecht F.A. 2011. On oblique random forests. *ECML PKDD'11 Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases. Vol.* **2**: 453-469.

Neto J.C. 2006. Plant species identification using Elliptic Fourier leaf shape analysis. *Comp. Electr. Agricult.* **50 (2)**: 121–134.

Phatak A., Kiiveri H., Clemmensen L.H., Wilson W.J. 2010. Constructing dependency networks using sparse linear regression. *Bioinformatics* **26 (12)**: 1576–1577.

Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. 1992. Numerical recipes in C. Cambridge University Press, Cambridge.

Quinlan R. 1993. Programs for machine learning. Morgan Kaufmann Publishers.

R Core Team 2012. R: A language and environment for statistical computing. R Foundation for statistical computing. Vienna, Austria. http://www.R-project.org/.

Schindelin J., Arganda-Carreras I., Frise E., Kaynig V., Longair M., Pietzsch T., Preibisch S., Rueden C., Saalfeld S., Schmid B., Tinevez J.Y., White D.J., Hartenstein V., Eliceiri K., Tomancak P., Cardona A. 2012. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**: 676–682.

Simon N. 2011. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Soft.* **39 (5)**: 1–13

Singh K., Gupta I., Gupta S. 2010. SVM-BDT PNN and Fourier moment technique for classification of leaf shape. *Int. J. Signal Process., Image Process. Pattern Recogn.* **3 (4)**: 67–78.

Strobl C., Malley J., Tutz G. 2009. An introduction to recursive partitioning. *Psy. Meth.* **14 (4)**: 323–348.

Tibshirani R., Hastie T., Narasimhan B., Chu G. 1999 Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. U.S.A.* **99 (10)**: 6567-6572. doi: 10.1073/pnas.082099299

Venables W.N., Ripley B.D. 2002. Modern applied statistics with S. 4th edition. Springer.

Wang X.-F. 2005. Recognition of leaf images based on shape features using a hypersphere classifier. *Adv. Intel. Computing.* **364**: 87–96.

Wolpert D. 1992. Stacked generalization. *Neural Networks* **5 (2)**: 241–259.

COMPUTER AIDED IDENTIFICATION OF THE *FICUS* L. SPECIES BY THE LAMINA SHAPE

Alexander Z. Gluhov & Ivan I. Strelnikov *

Abstract. The development of computer aided plant species determination is the urgent task of the botanical science. Identification is often bases on the morphology of the lamina. It is promising to describe the leaf shapes through the harmonic values of elliptic Fourier decomposition, but the effectiveness of this approach requires further verification. Another task is a comparative evaluation of different classification algorithms. The work was conducted on the 2812 leaves images of the 15 *Ficus* L. species. To solve the described tasks the optimal set of the Fourier decomposition parameters was determined. The best results are achievable by using the classification with 18 Fourier harmonics. Number of reference points on the outline does not affect the result of the models. We compared an identification accuracy of the 30 classification algorithms. Random forest algorithm had the highest classification accuracy – 98%. Combining different prediction algorithms by stacking improves the efficiency of the leaf shapes recognition.

Key words: *Ficus,* lamina shape, elliptic Fourier analysis, classification models

*Donetsk botanical garden NAS of Ukraine, Illich's Avenue 110, 83059 Donetsk, Ukraine; * ivanstrel87@gmail.com*