

УДК [575.1 +581.14]:519.24

Кластерный и регрессионный анализ изменений количественных морфометрических признаков

©2007 Монтиле А.А. *, Шавнин С.А. **, Монтиле А.И. ***

Ботанический сад, Российская Академия Наук, Уральское Отделение, Екатеринбург, 620144, Россия

Аннотация. Обосновывается целесообразность использования изменений значений признаков при классификации объектов, предполагающей учет различной наследственности. Для преодоления трудностей классификации, связанных с невыпуклостью и наличием значительных пересечений областей локализации объектов в пространстве признаков, предлагается использовать функциональные (в том числе, динамические) зависимости между значениями признаков. Рассматриваются различные способы реализации такого рода процедур, базирующиеся на обобщении понятия политететического класса.

Ключевые слова: классификация, кластерный анализ, регрессионный анализ, обработка результатов морфометрических измерений

1. ВВЕДЕНИЕ

Практической основой материала, выносимого на обсуждение, послужило решение двух задач: отбор генетических форм сосны обыкновенной, обладающих наибольшей скоростью роста, и анализ влияния температуры среды на ростовые процессы сосны. При решении первой задачи доминировала необходимость построения классификации объектов наблюдения, базирующейся на измерениях приростов, которая отображала бы различные группы генотипов. При решении второй - необходимость выделения и определения функциональной зависимости динамики приростов от температурного градиента. В качестве объекта исследования, в первом случае, использовались сравнительные культуры сосны обыкновенной, выращенные из семян плюсовых деревьев. Для второй задачи - молодняк сосны, растущий на опытном участке вблизи факела сжигания попутного газа. При проведении исследований использовались пакеты STATISTICA, STATGRAPH, и специально разработанный пакет, ориентированный на решение задач политететической классификации, который включает в себя ряд специализированных, а также все стандартные методы кластеризации, описанные в [1], и обладает развитыми возможностями визуальной обработки.

2. ИСХОДНЫЕ ПРЕДПОЛОЖЕНИЯ

Подход к анализу динамики непрерывных количественных морфометрических признаков (важная компонента индивидуального морфогенеза) с целью определения характеристик, на основе которых можно идентифицировать различные группы генотипов, базировался на двух взаимосвязанных соображениях.

* org17@mail.ru

** sash@botgard.uran.ru

*** mai@usfeu.ru

С биологической точки зрения, наследуются (обуславливаются генотипом) не только качественные или количественные признаки, но и динамические характеристики непрерывно протекающих на различных уровнях организации организмов биологических процессов, включая, связанные с их адаптацией к изменяющимся условиям среды. В каждой конкретной ситуации, определяемой набором значений переменных факторов, различные генотипы могут предопределить сходную динамику изменений одного подмножества внешних признаков и различную другого. Само разбиение априорно не известно. Его выявление в качестве инварианта, который может быть сопоставлен определенной наследственности, представляется значительным результатом.

С точки зрения метода решения, в силу приведенных выше содержательных соображений, представляется необходимым построение политетической, в более широком по сравнению с [2] смысле, классификации объектов. Кроме того, классификацию целесообразно осуществлять не только на основе расстояний в пространстве признаков, но и с учетом сходства-различия функциональных зависимостей, связывающих между собой и со временем, значения признаков для отдельных объектов.

3. ПРАКТИЧЕСКИЕ ТРУДНОСТИ

"Вычленение" и последующее восстановление зависимостей признаков от конкретного фактора или их набора в соответствии с традиционной методикой предполагает предварительное выделение групп объектов, относительно которых с определенной долей уверенности можно утверждать о наличии и "одинаковости" искомых зависимостей. Для этого обычно и используется кластерный анализ в пространстве признаков. Зависимость результатов работы отдельных алгоритмов кластерного анализа от наличия (предположение компактности) и специфики группировок объектов в пространстве признаков может приводить к неверным результатам. Поэтому, в условиях априорной неопределенности обычно используют набор алгоритмов и выводы делаются на основании совокупности результатов работы всех алгоритмов. Далее определяется конкретный вид зависимостей между признаками.

Использование традиционной методики не увенчалось успехом, несмотря на многообразие использованных средств и дополнительные измерения. Выделяемые кластеры не соответствовали разбиению деревьев по любым содержательным основаниям. Объективная причина заключалась в том, что области локальных сгущений объектов в пространстве признаков были не выпуклые и перекрывались. При этом разброс объектов в пространстве признаков не имел статистического характера, а сущностно отображал широкий диапазон индивидуальной изменчивости фенотипа. Дополнительные затруднения представлял нелинейный и разный вид зависимостей между значениями признаков, явно определяемый при содержательном, неформальном выделении групп объектов. Однако, постоянно используемая визуализация пространства сверток признаков, вид и параметры, которых модифицировались, позволила обнаружить наличие различных по виду зависимостей отдельных параметров и их сверток от времени для объектов различных "смысловых" групп. При замене некоторых признаков разностями значений выполнялось требование компактности.

В качестве простейшей иллюстрации рассмотрим ситуацию, возникшую при анализе динамики морфометрических признаков, отображающей ростовые процессы сосны (*Pinus sylvestris* L.). Его целью являлось определение характеристик, с помощью которых можно идентифицировать различные по наследственности группы деревьев.

В качестве объекта исследования использовались сравнительные культуры сосны обыкновенной в возрасте 10 лет, созданные из семян плюсовых деревьев и растущие на одном участке в относительно одинаковых почвенных и микроклиматических

условиях. Было обследовано 3 семьи полусибсов от плюсовых деревьев с номером регистрации по госреестру 11, 1 и 52, далее семьи будут обозначаться 1-11, 2-1 и 3-52. Семья 1-11 представлена 29 деревьями, семья 2-1 – 24 деревьями, семья 3-52 – 28 деревьями. У каждого дерева выбирали 3 ветки с ориентациями: юг, юго-восток, восток в возрасте 6 лет и измеряли приросты веток по длине за 4 последних года.

Попытка разделить семейства с помощью кластерного анализа по приростам за 4 года не увенчалась успехом. Использовались методы ближайшего соседа, дальнего соседа, медиан, центроидов, попарного среднего, Уорда (Ward's) и К внутригрупповых средних с метриками Евклидовой, квадратом Евклидовой и Сити-блок (Манхэттенской). Методы медиан и попарного среднего в различных метриках устойчиво выделяют 2 кластера. Методы дальнего соседа и Уорда устойчиво выделяют 3 кластера. Ни один из перечисленных методов не дал желаемого результата. Выделяемые кластеры не соответствовали ни разбиению деревьев по семействам, ни разбиению их по ориентациям, ни разбиению по семействам с учетом ориентаций. В таблице 1 приведены результаты кластерного анализа для методов, выделяющих 3 кластера и различных метрик: для каждого выделенного кластера указано количество веток из каждого семейства, отнесенных к этому кластеру.

Таблица 1. Примеры распределения веток деревьев разных семейств по кластерам

Метод	Метрика	N кластера	Кол. из 1-11	Кол. из 2-1	Кол. из 3-52
Дальнего соседа STATGRAPH	Евклидова	1	28	22	19
		2	19	24	35
		3	20	17	11
	Сити-блок	1	46	41	36
		2	3	7	17
		3	18	15	12
Уорда (Ward's) STATGRAPH	Евклидова	1	32	19	20
		2	9	20	30
		3	26	24	15
	Сити-блок	1	36	20	23
		2	25	24	14
		3	6	19	28
К внутригрупповых средних STATISTICA	Евклидова	1	33	28	22
		2	24	20	14
		3	10	15	29

Детальный анализ данных по приростам веток позволил обнаружить взаимосвязь между приростами за два следующих друг за другом года, или влияние приростов за один год на приросты за последующий. Результаты обработки имеющихся данных позволяют утверждать, что такого рода зависимость имеет место как для каждого семейства в отдельности, так и для всех семейств, взятых вместе. Визуальное представление выборки с выделением представителей различных семейств, выявляет значительно пересекающиеся, но не совпадающие "полосы" объектов, относящихся к каждому из семейств. Поэтому естественным шагом является построение линейной регрессии, характеризующей такого рода зависимость для каждого отдельно взятого семейства. Данные о коэффициентах корреляции и коэффициентах, характеризующих углы наклона регрессионных прямых, для каждого семейства и для трех пар лет приведены в таблице 2.

Таким образом, можно сделать вывод, что в пределах одного возрастного периода семейства можно сопоставлять и различать, сравнивая коэффициенты, характеризующие угол наклона прямой регрессии. Но, подобный вывод, был получен

фактически на основе "удачной" визуализации подходящего сечения пространства признаков.

Таблица 2. Коэффициенты корреляции и углы наклона регрессионных прямых для семейств первой группы

Семья	Годы	2001-2002		2002-2003		2003-2004	
	Количество дер.(вет.)	Коэфф. корр.	Коэфф. накл.	Коэфф. корр.	Коэфф. накл.	Коэфф. корр.	Коэфф. накл.
Данные по веткам всех ориентаций							
1-11	29 (87)	0,634061	0,44799	0,753757	0,60611	0,767542	0,891995
2-1	24 (72)	0,881982	0,869691	0,845141	0,753524	0,669127	0,734226
3-52	28 (84)	0,810916	0,826786	0,764652	0,75395	0,791858	0,837371
Данные по веткам 1 ориентации							
1-11	19 (19)	0,601629	0,414227	0,680741	0,453464	0,800973	0,988473
2-1	23(23)	0,910845	0,9385	0,864787	0,710885	0,829153	1,0883
3-52	26(26)	0,799939	0,851902	0,781066	0,865472	0,808982	0,793672
Данные по веткам 2 ориентации							
1-11	27(27)	0,70594	0,54164	0,810325	0,62561	0,80892	0,96711
2-1	24(24)	0,861997	0,807957	0,865415	0,710191	0,543001	0,595786
3-52	27(27)	0,795133	0,714467	0,736536	0,632917	0,723572	0,760895
Данные по веткам 3 ориентации							
1-11	24(24)	0,601084	0,404139	0,805049	0,749979	0,732478	0,81167
2-1	18(18)	0,877424	0,844584	0,782712	0,818934	0,627101	0,624864
3-52	26(26)	0,866377	0,961228	0,782049	0,764722	0,816251	0,909175

Дополнительно следует отметить, что характеристикой, отвечающей конкретной наследственности, оказываются не сами приросты веток по длине (измеряемые признаки), а динамика изменений этих приростов (зависимость, одним из способов описания которой, является рассмотрение углов наклона описанной регрессионной прямой).

Возникающую при использовании традиционной методики проблему можно сформулировать следующим образом: восстанавливать регрессию имеет смысл только в пределах выделенных кластеров, но сама кластеризация возможна только с учетом различия/сходства у объектов регрессионной зависимости, связывающей между собой некоторое подмножество множества измеряемых признаков.

При решении задач, связанных с выделением групп объектов с одинаковой наследственностью на основе измерений морфометрических признаков, ситуация усложняется изменением регрессионных зависимостей, вследствие действия различных факторов, но "новый" вид регрессии, как правило по-прежнему совпадает для объектов одной группы. Проанализированные случаи несовпадения, позволили высказать предположение о нерелевантности, по крайней мере, частичной, фиксированного в пределах решаемой задачи определения подмножества признаков, на основании которых осуществлялась классификация. Трудности, связанные с определением и корректировкой используемого при классификации подмножества признаков, по мере поступления новых данных (повторные измерения всего множества показателей у тех же самых объектов, сопряженные с повторным решением одной и той же содержательной задачи), в значительной мере совпадают с такого же рода проблемой при выделении различных комплексов физиологических и морфометрических показателей, отображающих отличающиеся друг от друга индивидуально-типологические особенности процессов адаптации. Наличие нескольких устойчивых состояний - аттракторов, предполагает необходимость выделения их бассейнов в

фазовом пространстве, возможность которого в значительной мере зависит от выбранного подмножества переменных состояния. Для решения этой задачи нами была разработана формальная модель, в рамках которой построены необходимые алгоритмы [3], которые и использовались для определения и корректировки подмножества признаков при классификации.

4. ПУТИ ПРЕОДОЛЕНИЯ. МЕТОДИКА

4.1. Использование в качестве дополнительных признаков агрегированных показателей и изменений значений отдельных признаков за определенные интервалы времени

С точки зрения статистической обработки результатов измерений, исходным объектом является куб, одно измерение которого – объекты, второе – признаки, третье – время. При необходимости осуществляется пополнение набора признаков результатами сверток – агрегированных показателей, вычисление которых производится по задаваемой пользователем формуле. Переменные, входящие в формулу – это первичные признаки, либо ранее определенные агрегированные показатели. Условно говоря, обрабатывается ось признаков. В некотором смысле предлагаемый способ, является одним из вариантов общего подхода, предложенного М.М. Бонгардом [4].

Учитывая специфику решаемых задач – предположение о сходной динамике процессов у объектов с одинаковой наследственностью, нами практически всегда использовался конкретный вид свертки - разность измерений, произведенных в смежные моменты времени для одного и того же признака. Пополнения набора признаков и агрегированных показателей в данном случае происходит с использованием оси времени, очевидно, что разности – это простейший случай, более сложные варианты обычно вытекают из содержательных особенностей решаемых задач. При решении первой задачи ключевым оказалось использование изменений осевых приростов и приростов боковых веток, без учета ориентации. При решении второй - использование изменения свертки, отображающей прирост биомассы.

4.2. Политетическая классификация

Итоговое отнесение объектов к классу определяется по отношению их принадлежности к одним и тем же кластерам в различных кластерных конфигурациях, каждая из которых характеризуется конкретным подмножеством множества признаков (расширенного в соответствии с 4.1) и используемым методом кластеризации. При этом возможна дальнейшая детализация с точки зрения учета совместной комбинаторики способов определения функции расстояния/сходства и стратегий кластеризации, но на практике мы ограничились уровнем метода – сочетание конкретного способа определения расстояния и определенной стратегии. Строгая принадлежность обеспечивается транзитивностью отношения на множестве объектов выделяемого класса по фиксированному подмножеству признаков. Нестрогая - подразумевает уменьшение до задаваемого, исходя из содержательных соображений, порога количества признаков (классическая политетичность), по которым объекты относятся к одним и тем же кластерам построенных кластерных конфигураций.

При решении практических задач, было сочтено возможным, отменить требование обязательного отнесения каждого объекта, к какому либо классу, т.е. итоговое распределение объектов по найденным классам не обязано быть разбиением множества объектов. Как следствие, в результате проводимой классификации, классы, состоящие из одного объекта, не рассматривались, а соответствующие объекты считались неклассифицированными. С другой стороны к одному классу относились только объекты, попадающие в один и тот же кластер каждой кластерной конфигурации.

Совместный учет количества и состава подмножества признаков, на основании "совпадения" значений которых делается вывод о принадлежности объектов определенному кластеру кластерной конфигурации, а также количества и параметров кластерных конфигураций в которых несколько объектов принадлежат к одному и тому же кластеру, является предметом дальнейших исследований.

4.3. Выделение групп объектов на основе одинаковости функциональных зависимостей

Решение задачи кластеризации в такой постановке, подразумевает перебор всех подмножеств множества признаков для всех подмножеств множества объектов и попытки восстановления регрессии в каждом возможном случае. Очевидна потенциальная, но не практическая из-за трудоемкости, осуществимость такой процедуры. Поскольку приемлемый алгоритм решения такого рода задачи отсутствует, были использованы эвристические процедуры. Наиболее эффективным оказалось выделение подмножеств объектов и признаков на основе визуального анализа. В основном использовались два способа. Первый - отображение сечений пространства признаков, дополненного свертками, вид и состав признаков для которых определяется пользователем. Второй - двухмерная модификация данного отображения: одна из осей соответствует метрической функции расстояния (метрика выбирается пользователем), вторая – неметрической функции сходства (косинус между векторами), начало координат и ориентация осей задаются пользователем, либо определяется автоматически.

Если регрессия строится для 2 или 3 признаков, определение "одинаковости" функций проще всего производить визуально. Для большего количества признаков понятие "одинаковости" функциональных зависимостей может быть формализовано различными способами. Мы использовали аналогию со структурными методами распознавания. Одинаковыми считаются функции, задаваемые синтаксически эквивалентными формулами. Допустимый разброс числовых значений коэффициентов определяется суммарным отклонением и не может превосходить порог, определяемый пользователем.

Для рассмотренных ранее задач классификация объектов, в первом случае, была осуществлена на основе одинаковости линейной регрессии, связывающей прироста за два смежных года, во втором на основе одинаковости полиномиальной регрессии (разных порядков), связывающей прироста биомассы за год (свертка) с расстоянием от факела и годом.

СПИСОК ЛИТЕРАТУРЫ

1. Rencher A.C. *Methods of Multivariate Analysis*. – 2nd ed. New York: Wiley. 2002.
2. Сокал Р.Р. Кластер-анализ и классификация: предпосылки и основные направления. В сб.: *Классификация и кластер*. Ред. Райзин Д. В. М.: Мир. 1980. с. 7-19.
3. Монтиле А.И., Шавнин С.А., Монтиле А.А. Метод формирования комплексов физиологических и морфометрических показателей, отображающих особенности индивидуальных процессов адаптации. *Вестник Томского государственного университета*. 2006. **21**. с. 105-107.
4. Бонгард М.М. *Проблема узнавания*. М.: Наука. 1967.

Материал поступил в редакцию 03.04.2007, опубликован 24.04.2007