ЕЖЕГОДНИК-2012, Тр. ИГГ УрО РАН, вып. 160, 2013, с. 353–355

ПРИМЕНЕНИЕ ЛИНЕЙНОГО ДИСКРИМИНАНТНОГО АНАЛИЗА ДЛЯ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ПРОИСХОЖДЕНИЯ ИЗУМРУДА ПО ДАННЫМ РЕНТГЕНОФЛЮОРЕСЦЕНТНОГО АНАЛИЗА

© 2013 г. А. В. Поротников, М. П. Попов, Н. П. Горбунова

ВВЕДЕНИЕ

Определение происхождения драгоценных камней является одной из центральных задач геммологии. Среди способов решения этой задачи выделяют два направления: методы, основанные на изучении включений, и методы, основанные на изучении характеристик объекта в целом, в том числе элементный и разные виды спектрального анализа. Среди преимуществ второго направления можно выделить экспрессность при подходящем выборе аналитической процедуры и отсутствие непреодолимых препятствий для алгоритмизации. Рентгенофлюоресцентный анализ (РФА) монолитного образца имеет такие преимущества, как отсутствие пробоподготовки и разрушения образца и возможность проведения измерений в составе изделия. Однако минусом метода в данном случае является невозможность использования какого-либо стандарта, кроме внутреннего. Для алгоритмизации определения происхождения образца необходим поиск сигнатур (fingerprinting), то есть признаков, вычисляемых математически или алгоритмически и позволяющих принять решение о его происхождении без участия человека. Данная проблема относится к взаимопересекающимся областям знания, известным как искусственный интеллект, машинное обучение, математическая статистика и глубокий анализ данных. Далее будет использоваться терминология машинного обучения как наиболее адекватная проблеме.

Имеется единственная уникальная работа по использованию методов машинного обучения для классификации изумрудов [1]. В ней использовались данные элементного анализа 450 образцов, полученные электронным микрозондом, которые затем подавались на вход нейронной сети с одним скрытым уровнем. В соответствии с принятыми в сфере машинного обучения методами вначале нейронная сеть обучалась на 2/3 образцов, затем качество ее обучения проверялось на остальных образцах (кросс-валидация). Выполнялась классификация не по странам происхождения, а в соответствии с одной из классификаций месторождений по 5 категориям, цитируя [2]. Было получено успешное разделение изумрудов по типам месторождений с долей ошибок 3%.

Цель работы – исследование возможностей построения программно-аппаратного комплекса для определения происхождения изумруда персоналом без геолого-минералогической квалификации.

Объекты исследования. Исследовались 17 образцов берилла, из них 10 из России (месторождения Мариинское, Свердловское, Каменское, Черемша), 3 из Колумбии (2 – Чивор, 1 – Музо), 2 из Афганистана, 1 из Бразилии (Баия) и 1 из Китая. В соответствии с двумя из известных классификаций [3] данные образцы относятся к следующим типам (табл. 1).

МЕТОДЫ

Измерения проводили в лаборатории физических и химических методов исследования Института геологии и геохимии УрО РАН на рентгенофлуоресцентном энергодисперсионном спектрометре EDX-900HS фирмы SHIMADZU (Япония). Измерения выполняли в двух диапазонах: от Na до Sc, и от Ti до U; для каждого из них были выбраны оптимальные условия измерения спектров. В связи с произвольными размерами и формой образцов, затрудняющими использование стандартов, определяли не абсолютные концентрации, а значения относительных интенсивностей линий Al, Mg, Cr, V, Fe, Ca и K к интенсивности линии Si (табл. 2).

Для дальнейшей обработки был выбран линейный дискриминантный анализ (LDA). LDA является одним из самых простых методов машинного обучения и относится к методам обучения по прецедентам. Алгоритм LDA конструирует линейные комбинации исходных параметров объектов (дискриминантные функции) так, что значения этих функций максимально удалены друг от друга для объектов разных классов:

Таблица 1. Типы исследованных образцов в соответствии с различными классификациями

Происхождение	Классификация	
	Dereppe [1]	Schwarz, Giuliani [2]
Россия	1	1a
Колумбия	4	2b
Афганистан	3	2a
Бразилия (Баия)	1	1a
Китай	?	?

Таблица 2. Нормированные к Si интенсивности линий в спектрах РФА бериллов различного происхождения

№	Происхождение	Al/Si	K/Si	Ca/Si	Ti/Si	V/Si	Cr/Si	Mn/Si	Fe/Si	Ni/Si
С3–1	Колумбия	0.059	0.041	0.209	0	0.173	0.445	0.118	0.159	0
С3–2	Колумбия	0.059	0.5735	2.1176	0.1471	0.4853	0.5735	0.5147	0.6176	0.3088
С3–4	Афганистан	0.092	0.2372	0.6942	0.0076	0.0606	0.1014	0.0106	0.9394	0.0335
С3–5	Бразилия	0.067	2.1290	4.9802	0.1667	0.0397	0.0615	0.0159	8.5813	0.0794
С3–17	Мариинское	0.139	0.0162	0.0076	0.0000	0.0000	0.1205	0.0334	0.4075	0.0073
С2–1	Свердловское	0.050	0.0263	0.1447	0.5219	0.0044	0.1199	0.0117	2.7178	0.0497
С3–3	Колумбия	0.070	0.014	0.053	0	0.074	0.209	0.010	0.301	0.080
С3–4	Афганистан	0.123	0.434	4.311	0.023	0.050	0.090	0.059	2.592	0
С3–10	Каменское	0.076	0.517	0.144	0.068	0.013	0.235	0.070	4.746	0.005
С3–17	Мариинское	0.137	0.013	0.002	0	0.001	0.082	0.015	0.359	0.007
С3–19	Черемша	0.071	1.386	0.107	0	0.010	0.158	0.116	5.378	0.117
С3–25	Китай	0.075	0.030	1.726	0	0.174	0.010	0.021	1.114	0
С3–23	Мариинское	0.087	0.009	0.062	0	0	0.030	0.006	0.288	0.009
С2–2	Мариинское	0.059	0.255	0.387	0.137	0.0664	0.089	0.015	2.033	0.240
С2–3	Мариинское	0.080	0.116	0.117	0.072	0.0099	0.030	0.067	1.070	0.074
С2–6	Мариинское	0.066	0.005	0.064	0.302	0.0054	0.015	0.018	0.761	0.112
С2–8	Мариинское	0.091	0	0.065	0.251	0	0.026	0.038	0.380	0.038

$$f_{km} = \sum u_i X_{ikm},$$

где f_{km} – значение дискриминантной функции для m-го объекта k-го класса, u_{ik} – коэффициенты k-той дискриминантной функции, X_{ikm} – i-ый параметр m-го объекта k-го класса.

Далее объект относится к тому классу, дискриминантная функция которого имеет большее значение. Кроме этого, для наглядного представления результатов алгоритм выбирает подпространство (обычно 2-мерное, т. е. плоскость) в исходном n-мерном пространстве исходных данных, где n– число измеренных у объекта параметров, таким образом, что проекция исходных данных на это подпространство обладает теми же свойствами – минимизирует внутриклассовый и максимизирует межклассовый разброс проекций векторов.

Математические идеи, на которых основан расчет, включают в себя моделирование условных плотностей распределения параметров $P(X|y = k)$ для всех классов k и предсказание класса на основе правила Байеса.

$P(y|X) = P(X|y)\cdot P(y)/P(X) = P(X|y)\cdot P(Y)/(\sum P(X|\acute{y})\cdot p(\acute{y}))$.

В методе LDA $P(X|y)$ моделируется как распределение Гаусса и предполагается равенство ковариационных матриц всех классов.

Для обработки создана и использовалась оригинальная программа на основе реализации этого алгоритма в библиотеке scikit-learn [4].

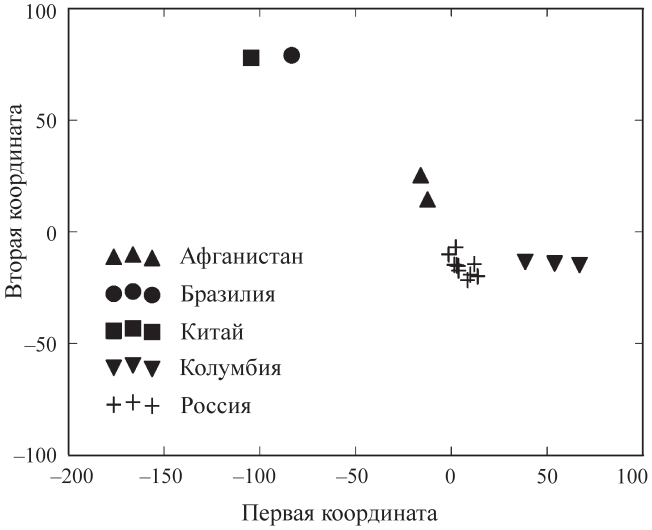

Рис. 1. Проекция методом LDA данных РФА на оптимальную плоскость демонстрирует успешное разделение всех классов. График в условных координатах на этой плоскости.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Результаты исследования состава образцов берилла приведены в табл. 2.

При дальнейшей обработке каждая страна происхождения считалась отдельным классом. Это может выглядеть спорным решением в свете известных классификаций (см. табл. 1), поскольку все из них, за исключением малоизвестной и неупомянутой классификации Bruno Sabot относят месторождения Бразилии (исключая Санта-Терезинья) и России к одному типу. С другой стороны, противоречивость этих классификаций отмечается разными авторами [3].

В результате обработки данных все образцы успешно разделены на классы, неверные отнесения отсутствуют. На проекции (рис. 1) также можно видеть линейную сепарабельность (возможность разделения с помощью прямых линий) всех классов. При этом российские образцы и бразильский оказались далеко друг от друга, что подтверждает бесполезность и противоречивость классификаций из табл. 1. По причине малого числа образцов этот

классификатор нельзя считать достаточно надежным, более того, по той же самой причине нельзя провести принятую в сфере машинного обучения кросс-валидацию, и результаты могут считаться лишь предварительными.

ВЫВОДЫ

Рентгенофлюоресцентные измерения монолитных образцов и последующее применение линейного дискриминантного анализа показывает свою перспективность для автоматического определения происхождения изумрудов. Однако малое число измерений пока не позволяет оценить надежность этого метода в соответствии с общепринятыми в машинном обучении критериями (с помощью кросс-валидации). В итоге наиболее важным является дальнейшее накопление экспериментальных и расчетных данных.

Работа выполнена в рамках междисциплинарного проекта УрО РАН № 12-М-235-2063 и при поддержке гранта РФФИ № 11-05-00035 в Центре коллективного пользования "Геоаналитик" УрО РАН.

СПИСОК ЛИТЕРАТУРЫ

1. *Dereppe J.M., Moreaux C., Chauvaux B. et al.* Classification of emeralds by artificial neural networks // Journal of Gemmology. 2000. № 27. P. 93–105.
2. *Schwarz D., Giuliani G.* Emerald deposits-a review // The Australian Gemmologist. 2001. № 21. P. 17–23
3. *Groat L.A. et al.* Emerald deposits and occurrences: A review // Ore Geology Reviews. 2008. № 34. P. 87–112.
4. *Pedregosa M. et al.* Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. № 12. P. 2825–2830.