ГОДИШНИК НА МИННО-ГЕОЛОЖКИЯ УНИВЕРСИТЕТ “СВ. ИВАН РИЛСКИ”, ТОМ 59, СВ. I, Геология и геофизика, 2016
ANNUAL OF THE UNIVERSITY OF MINING AND GEOLOGY “ST. IVAN RILSKI”, Vol. 59, Part I, Geology and geophysics, 2016

# ОСНОВНИ ДЕЙНОСТИ ПРИЛАГАНИ В ПРОЦЕСА ЗА РАЗПОЗНАВАНЕ НА ОБРАЗИ ПРИ РЕШАВАНЕ НА ПРОГНОЗНИ ГЕОЛОЖКИ ЗАДАЧИ

***Атанас Кисьов***

*Минно-геоложки университет "Св. Иван Рилски", 1700София, e-mail: at.kisyov@gmail.com*

**РЕЗЮМЕ.** Теорията за разпознаване на образи в геологията и по-конкретно за класифициране на обекти по съвкупност от различни признаци (геофизични, геоложки, геохимични и др.) интензивно се развива и намира все по-широка практическа реализация. Разпознаването на образи е машинно-ориентирана методология, позволяваща бърз и многократен анализ на разнообразна информация и осигуряваща получаване на количествени резултати. Приложението на методите използвани за разпознаване на образи в геоложкото прогнозиране изисква прилагането на последователност от някои основни дейности, свързани с подбора на територията и обекта на прогнозиране, събирането и класифицирането на признаци използвани за описание на прогнозните обекти, както и формиране на еталонни обекти, които да бъдат използвани в процеса на класификация

**MAIN ACTIVITIES APPLIED IN THE PROCESS OF PATTERN RECOGNITION USED FOR SOLVING PROGNOSTICATION GEOLOGICAL TASKS**

***Atanas Kisyov***

*University of Mining and Geology “St. Ivan Rilski”, 1700 Sofia, e-mail: at.kisyov@gmail.com*

**ABSTRACT.** The theory of pattern recognition in geology and in particular for classification of objects in a set of different characteristics (geophysical, geological, geochemical, etc.) is developing intensively and is having a wide-ranging practical implementation. The pattern recognition is a machine-oriented methodology providing a rapid and repetitive analysis of diverse information and ensuring acquirement of quantitative results. The application of pattern recognition in geological prognostication requires a sequence of some major activities related to the selection of the territory and the subject of estimation, to the collection and classification of the characteristics used for description of the studied sites and to the definition of reference sites that can be used in the classification process.

## Въведение

Решаването на всички прогнозни задачи винаги, в една или друга степен, е свързано с елемент на неопределеност. Например, традиционната концепция за металогенно прогнозиране се основава на представата, че при отчитане само на няколко основни геоложки фактора най-ефективно може да се реши задачата, независимо от обстоятелството, че практически всички геоложки и геофизични данни по своята природа имат статистически характер. Освен това, функционалните връзки между геоложките и геофизичните параметри и орудяванията са сложни и многовариантни. Ето защо все по-убедително се налага идеята за насочване на стратегията на прогнозирането към комплексното обработване и анализиране на цялата налична информация за изследваните площи и последващо приложение на методите за разпознаване на образите.

По-голямата част от методите за разпознаване на образи се основават на принципа на пряката аналогия. Това е традиционен подход и при геопрогнозните задачи. Като правило, при тях се сравняват изследваните по комплекс от геолого-геофизични признаци участъци с известни еталонни промишлени находища или неперспективни участъци.

## Методична схема на прогнозната задача

Приложението на методите за разпознаване на образи в геоложкото прогнозиране изисква осъществяването на следните основни дейности:

**Избор на територията и обекта на прогнозиране**

Металогенното прогнозиране се осъществява за територията на съответните рудоносни региони, структурно-металогенни зони или провинции, рудни райони, рудни полета и рудни находища. Обектът на прогнозиране притежава специфичен набор от търсещи признаци.

**Избор на подход при прогнозирането**

Възможни са два подхода към геоложките условия от позицията на прогнозирането със средствата за разпознаване на образи - площно и обектно прогнозиране. (Дуда и Харт, 1976; Duda and Hart, 2000).

*Площното прогнозиране* предвижда разделянето на изучаваната територия на равни по площ участъци (елементарни участъци), всеки, от които, с комплекс от признаци, се сравнява с еталонни участъци.

*Обектното прогнозиране* се състои в оценка перспективността на отделни геоложки обекти или геоложки

структури. За целта те се съпоставят с аналогични еталонни обекти (например, в границите, на които има находища на полезни изкопаеми или с детайлни работи е доказано тяхното отсъствие).

Изборът на конкретен подход зависи от целите на прогнозирането, от степента на изученост на територията и от естеството на първичните материали.

При голямата детайлност на геолого-геофизичните работи и необходимост да се оцени конкретната перспективност на отделни обекти (структури, интрузиви, разломи, рудни полета и др.) е целесъобразно приложението на обектния подход при прогнозирането.

Площният подход е подходящ главно при регионалното и полурегионалното прогнозиране върху площи, които включват различни структури.

***Определяне на елементарния участък на прогнозиране***

Ако бъде приет площният подход за прогнозиране от изключително важно значение е коректният избор на елементарни участъци. За такива обикновено се приемат клетки с еднакъв размер и еднаква форма, съвкупността от които обхваща цялата изследвана територия. Размерът на елементарните участъци зависи главно от мащаба на прогнозните проучвания.

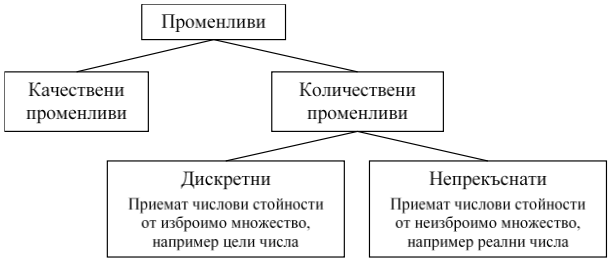
***Набиране на изходна съвкупност от признаци за описание на прогнозните обекти***

Един от най-сложните и важни етапи на количественото металогенно прогнозиране е набирането на изходния комплекс от признаци, които се използват за описание и съпоставяне на елементарните участъци. Това е неформализуем процес, при който са решаващи опитът и знанията на специалиста. Съществуват обаче някои съображения, които трябва да се отчитат. Съгласно с принципите на статистическата обработка и анализа, геоложките параметри могат да се разглеждат като признаци, ако е възможно еднозначно да бъдат описани за всички елементарни участъци (Silverman, 1986; Scott, 1997).

Данните за прогнозирането се вземат от първичните геоложки материали – геоложки, геофизични, геохимични, топографски и други карти, както и от продукти на тяхната обработка. Видът на тези данни, които в терминологията на автоматизираното прогнозиране се наричат признаци, е извънредно разнообразен.

Данните се вземат чрез измерване на определена характеристика върху група от обекти. Когато измерванията на тази характеристика могат да приемат различни числови или нечислови стойности, тази характеристика се нарича променлива. От друга страна, ако характеристиката е една и съща за всеки елемент на групата, то тя е константа. Константите не подлежат на статистически анализ.

Променливите могат да бъдат класифицирани по различни начини. Класификацията на променливите според техния тип е илюстрирана на фигура 1.




**Фиг. 1. Класификацията на променливите според техния тип**

Основното разграничение на променливите е на *количествени* (наричани още *числови* или *вариационни*) и *качествени* (наричани още *категорийни* или *атрибутивни*).

*Количествени променливи* са тези, чиито стойности се изразяват с числа в съответна мярка. Количествените променливи се поделят на *непрекъснати* (наричани още *индискретни* или *неизброими*) и *дискретни* (наричани още *прекъснати* или *изброими*).

Числова променлива, чийто допустими стойности са неизброими се нарича *непрекъсната променлива* (*continuous variable*). Случайната величина X е непрекъсната ако може да приема всички възможни стойности в даден краен или безкраен интервал. Вероятността за приемане на всяка допустима стойност x е нула (P(X=x)=0 за всяко x). За една непрекъсната случайна величина комулативната функция на разпределение е непрекъсната (без скок). Непрекъснатата променлива може да приема всяка една стойност върху скалата на измерване. Следователно, често казвайки, че дадена променлива е непрекъсната ще имаме предвид не толкова конкретния резултат от измерването, а по-скоро модела, който използваме при анализа и интерпретацията на тази променлива. В този случай ще предполагаме, че няма прекъсвания между отделните стойности на измерителната скала. На практика не може да се измери непрекъсната променлива с безкрайна точност и затова тя се апроксимира с изброима, дискретна променлива (Ту и Гонсалес, 1978).

Числова променлива, чийто допустими стойности са изброим брой се нарича *дискретна променлива* (*discrete variable*). Типичен пример са променливите, чийто стойности са подмножество на *целите числа* (*integer*). Случайната величина X е дискретна, ако с определена вероятност може да приема отделни, изолирани възможни стойности. Дискретната променлива може да приема стойности само от едно изброимо (крайно или безкрайно) множество.

*Качествени променливи* са тези, чиито стойности нямат числов израз, а се дават описателно, словесно. Такива са например петрографска разновидност и др.

Особен вид качествени променливи са тези, които имат само две възможни стойности, представляващи проста алтернатива. Такъв е например случаят, когато дадена единица или има (“да”), или няма („не”) дадено качество или свойство. Такива променливи се наричат *дихотомни* или *алтернативни (*Byholt and Hjort, 1996).

Следва да се отбележи, че методите за анализ на данните са тясно обвързани с начина на измерване. При анализа на данните на преден план излизат знанията ни за стойностите на свойствата на обектите. Тези стойности се получават чрез измерване на нужното свойство на обекта. Под измерване ще разбираме присвояването на определени символи на изучаваното свойство в съответствие с предварително зададено правило. Тези символи могат да бъдат букви, цифри или числа в зависимост от скалата на измерване, която се определя от типа на променливата – количествена или качествена.

Обсъждането на скалите на измерване е твърде важно, когато се предвижда използване на статистически методи за анализ. Скалата, по която се измерва дадена променлива е фактор от голямо значение за определянето на подходящи методи за анализ на данните във всяко едно изследване. Разделянето на скалите за измерване се извършва в съответствие със степента на точност на измерването, или с други думи, в съответствие с количеството информация, което те носят (Myashita et al.,1993; Devroye et al., 1996).

Измерванията могат да бъдат класифицирани в съответствие с *типа* или *йерархията* на мерната скала. Съществуват четири типа скали за измерване като всяка следваща е по-точна и по-информативна от предходната в йерархията. *Качествените* променливи се измерват в *номинална* и/или *рангова* скала, а *количествените* – в *интервална* и/или *скала на отношенията*. Първите две скали се наричат *неметрични*, а вторите две – *метрични* (респ. има неметрични и метрични данни).

За типовете скали може да бъде обобщено както следва:

- номиналната скала класифицира без да задава наредба;
- ранговата скала класифицира като задава и линейна наредба;
- интервалната скала класифицира като и задава линейна наредба, но допълнително задава относителна единица на измерването и относително начало на измерването;
- скалата на отношенията класифицира като задава и линейна наредба, освен това има зададена единица на измерването и абсолютно начало на това измерване.

***Описание на всички елементарни участъци по набора от изходни признаци***

Автоматизираният процес на обработка и анализ на информацията изисква нейното провеждане във форма удобна за математическа обработка. Това най-често е числовата форма на данните. За нейното постигане данните предварително се кодират. Кодовете трябва да удовлетворяват и изискванията за стандартизиране на данните – еднозначност, кратност, адитивност, прогресивност и др.

Съществено методично значение има начинът на представяне на признаците при описание на елементарните участъци. За количествените признаци (например разпределението на потенциалните геофизични полета) се фиксира стойността на параметъра за всеки от елементарните участъци. При качествените признаци обаче (например наличие на определена геоложка формация) фактологичното описание съвпада с алтернативна констатация, т.е. възможно е само едно от две взаимоизключващи се стойности (1 или 0, т.е. "да" или "не"). Този вид признаци, именувани най-често картографически, е целесъобразно да бъдат обработени по подходящ начин, за да се получи стойностно описание и да се създадат предпоставки за по-пълно извличане на полезната информация (Ruiz and Lopez-de-Teruel, 2001; Pao, 1990).

Признаците се описват количествено чрез кодиране на изходните променливи в определен брой интервали - *градации на признаците.* Броят и границите на градациите се подбират конкретно за всеки признак така, че възможно най-пълно и достоверно да описват статистическото му разпределение. При това се търси оптимален компромис между две противоречиви тенденции:

- по възможност да се локализират областите с екстремални стойности на променливата;
- да се постигнат условия, при които плътността на разпределение на обособените градации да не благоприятства за забавяне на малък брой стойности на променливата, информативността на които може да се определи достоверно надеждно(Ту и Гонсалес, 1978).

***Формиране на съвкупности от еталонни обекти***

Формираното признаково пространство (масивът от описанията на елементарните участъци) в общия случай се разбива на три области: еталони от перспективни (рудни) участъци, еталони от неперспективни (безрудни) участъци и участъци, на които трябва да се оцени перспективността.

Подходът към решаването на задачата за избор на еталони от двата класификационни класа - рудни (положителни) и безрудни (отрицателни), има голямо значение. Това е много важен методичен въпрос на обработката, от който зависи оптималното използване на априорната информация. Специално внимание заслужават два аспекта на този въпрос - пространственото разположение на еталоните и формирането на съвкупност от еталони за представяне на клас безрудни. Вероятностно-статистическите схеми, които могат да се използват за разпознаване, изискват съвпадение на многомерните разпределения за еталоните и разпознаваните обекти в пространството на диагностиращите признаци. Това изискване се изпълнява удовлетворително, ако еталонните и разпознаваните обекти са разположени достатъчно равномерно в изследваната територия.

Докато въпросът за избор на еталони от първи клас - рудни, се решава еднозначно, съвкупността от обекти от втори клас - безрудни, винаги се формира в значителна степен условно. Възможни са два вариантни подхода. При първия еталоните от втори клас се набират от добре изучени участъци, на които не е установена промишлена минерализация. При втория се анализира разпределението на признаците за цялата територия, в която няма промишлени орудявания. При първия вариант много трудно се изпълнява условието за равнозначно и равномерно отделяне на отрицателните еталони върху изследваната територия. Вторият вариант се основава на

предпоставката, че зад границите на известните промишлени минерализации орудяванията се срещат твърде рядко, което не винаги е валидно за цялата територия.

***Избор на метод за разпознаване на образи***

Разпознаването на образите е научна област, изучаваща функционирането на системи, които извличат общите характеристики на съвкупност от обекти, както и методите за създаване на такива системи. (Duda etal., 2000; Devroyeetal., 1996). За решението на проблема за разпознаване на образи (*pattern recognition*) се използват широк кръг от методи за извличане на знания (*knowledge discovery*) в големи бази данни (*data mining*).

Опитът показва, че могат да се обособят две постановки в процеса на разпознаването:

- Първата предвижда създаването на поне две обучаващи извадки на базата на еталонни обекти. Предполага се, че еталонните обекти, обединени във всяка от групите, са проявление на аналогични геоложки фактори и различията им имат случаен и незакономерен характер. Обединяването на разнородни обекти не позволява да се получи удовлетворително, както във формално, така и в геоложко отношение, прогнозно решение. Затова в райони, в които се наблюдава голямо разнообразие на геоложките условия се допуска по-голям брой обучаващи извадки. Присъствието на обучаващи групи от елементарни обекти в процедурата по разпознаването отделя клас задачи, означавани като „*разпознаване на образи с обучение*“.
- Втората постановка предвижда използването на така наречена *автоматична класификация* (*разпознаване без обучение*). Прилага се в слабо изучени територии, където е затруднено комплектоването на еталонни групи обекти.

При използването на алгоритми за разпознаване на образи се изхожда от свойството на признаковото пространство. Приема се, че постановката на задачата за разпознаване е детерминистична, когато пространството на признаците може да се раздели на непресичащи се обекти всеки, от които дефинира един от класовете. Ако съществува смесване на класовете в пространството на признаците, тяхното разделение е осъществимо на основата на статистическите постановки за решение на задачите за разпознаване на образи. Този подход към свойствата на признаковото пространство разделя алгоритмите за разпознаване на образи на две групи: детерминистични и статистически (Kohonen, 1986; Mulleretal., 2001).

Използват се и други класификации за разделяне на алгоритмите. Съществува алтернативна група алгоритми, при която е съществена информацията, описваща структурата на всеки обект. При тях от процедурата за разпознаване се изисква не само да отнесе обекта към определен клас, но и да се опишат тези негови страни, които изключват отнасянето му към друг клас. Такъв подход се нарича „синтактичен“ (структурен).

Приложението на статистическите решения е свързана с по-големи трудности, отколкото на детерминистичния подход. При статистическите алгоритми трябва да се отчита независимостта на признаците и закона на разпределението им, обемът на обучаващите извадки и т.н. Статистическиятанализ, обаче, преодолява пречещото действие на многото случайни фактори, които маскират характерните различия между класовете и затрудняват тяхното разделяне.

В сравнение с вероятностния, детерминистичният подход е по-слабо разработен, но въпреки това има широка област на приложение. Той е освободен от редица ограничения и изисквания, предявявани към изходните данни. Детерминистичните (евристични) алгоритми имат за цел намирането на най-кратките пътища към решаването на прогнозните задачи (Rumelhart and McClelland,1986).

Приема се, че ефективността на разпознаването на образи зависи от особеностите при използването на разрешаващата схема в дадена задача, а не от индивидуалните особености на алгоритмите, осъществяващи конкретната схема на разпознаването. При еднакви други условия, най-подходящи са тези алгоритми, които разделят класовете с помощта на най-прости решаващи правила. Поради това, при решението на дадена прогнозна задача изследователят е заинтересован да разполага с набор разнообразни алгоритми. Впоследствие се избира като работен този алгоритъм, който при конкретните условия е най-ефективен и дава най-просто и надеждно решение.

***Разпознаване и класификация***

Същинското разпознаване се състои в построяване на формален търсещо-оценъчен критерий въз основа на избрания алгоритъм за разпознаване на образи. Преди да послужи за прогнозиране, решаващото правило се подлага на съдържателен анализ. Най-често оценката на качествата на решаващото правило се извършва по резултатите от класифицирането на контролни обекти, които не са използвани в операцията по обучението. Специалистите анализират резултатите от контролните решения, отчитайки броят на грешките и отказите за класификация, в непосредствена връзка с геоложкия смисъл на разделението на обектите. При необходимост решаващото правило се коригира. За целта може да се преформулира признаковото пространство, да се промени състава на обучаващите извадки, или да се смени избраният алгоритъм. Тези процеси се повтарят, докато се получат удовлетворителни резултати. Въз основа на уточненото решаващо правило се извършва класификацията, т.е. автоматичното разделяне на изследваните територии или обекти на класове с геоложко съдържание (Orr, 1996; Haykin,1998;).

***Оценка на надеждността на решенията***

Използването на тестова група обекти с известна принадлежност е най-често прилаганият метод за оценка на надеждността на разпознаването. Такъв подход е оправдан при достатъчно големи представителни групи от обекти с известна геоложка принадлежност. В противен случай се използва методът на изключването от обучаващото множество на единични обекти и разглеждането им в последствие като контролни. Процедурата се повтаря за всички обекти включени в

обучението. При статистическите методи, качеството на решаващото правило се оценява по класификационната грешка, която има вероятностен характер.

## Заключение

Приложението на методите за разпознаване на образи, се налага като един съвременен подход към анализа и извличането на закономерности при обработката на масиви геоложки и геофизични данни и приложението им за прогнозиране на находища на полезни изкопаеми.

В тази връзка, чрез приложението на методамогат да се решат следните задачи:

• Разработването на методи за откриване и интерпретация на закономерности в масиви данни;

• Създаването на формален език за построяване на модели, описващи закономерностите в данните и приложението му за предсказване и търсене в бази данни;

• Създаването на методи за прогнозиране на рудни находища, при наложени моделни ограничения;

## Литература


Дуда, Р., П. Харт. *Распознавание образов и анализ сцен. М., Мир.* 1976. – 509с.

Ту, Дж., Р. Гонсалес. *Принципы распознавания образов. М., Мир.* 1978. – 412с.

Byholt, M., N.L. Hjort. Sometimes nonparametric beat parametric even when the model is right. - *Statistical Research Report, Dep. Of Mathematics, University of Oslo,* No. 18, 1996. -1-20.

Devroye, L., L. Gyorfi, G. Lugosi. A probabilisticTheory of Pattern Recognition*, Springer.* 1996. - 654p.

Duda, R., P. Hart, D. Stork. *Pattern Classification, 2nd ed., John Wiley & Sons.* 2000. - 680p.

Haykin, S. *Neural Networks: A Comprehensive Foundation, Prentice Hall.* 1998. - 842p.

Kohonen, T. Learning Vector Quantization, Helsinki University of Technology, - *Lab of Comp. and Information Science, Report TKK-F-A-601. 1986. - 220-227.*

Muller, K.R., S. Mika, G. Ratch, K. Tsuda, B. Scholkopf. *AnIntroduction to Kernel Based Learning Algorithms, IEEE Transaction on Neural Networks, vol.12, No.2, March,* 2001. - 181-198.

Myashita, Y., Z. Li, S. Sasaki. Chemical Pattern Recognition and multivariate analisys for QSAR studies. - *Trends in Analitycal Chemistry, vol. 12, no.2.* 1993. - 55-63.

Orr, M.J.L. *Introduction to radial basis networks.TR Centre for Cognitive Science, Univ. Of Edinburgh, Scotland, 1996. http://anc.ed.ac.uk/~mjo/.*

Pao, Y. H. *Adaptive Pattern Recognition and Neural Networks, Addison-Wesley Publishing, Reading, Massachusetts.* 1990. - 599p.

Ruiz, A., P.E. Lopez-de-Teruel. Nonlinear Kernel-Based Statistical Pattern Analisys*. - IEEE Transaction on Neural Networks, vol.12, No.1,* 2001 .- 16-32.

Rumelhart, D.E., J.E. McClelland. Parallel Distributed Processing. - *MIT Press, Cambridge, Massachusetts.* 1986. - 1024-1077.

Scott, D.W., *Density Estimation, Rice University, Houston, TX.* 1993 - 335-339. (http://rice.edu)

Silverman, B. W. *Density Estimation for Statistics and Data Analysis*, Chapman and Hall.1986. – 22p.



Статията е рецензирана от гл. ас. д-р Мая Григорова и е препоръчана за публикуване от кат. "Приложна геофизика".