Grzegorz Wojarnik

Wydział Nauk Ekonomicznych i Zarządzania
Uniwersytet Szczeciński

Koncepcja wykorzystania algorytmów genetycznych jako metody analizy danych medycznych gromadzonych w ramach Internetowego Konta Pacjenta

1. Wstęp

Internetowe Konto Pacjenta (IKP) to projekt, którego celem jest umożliwienie gromadzenia w jednym miejscu danych na temat stanu zdrowia pacjenta. Projekt ten zakłada rejestrację w centralnej bazie danych wszystkich danych związanych ze stanem zdrowia obywateli Polski. Będą tutaj rejestrowane dane na temat skierowań, recept, zaświadczeń, zaleceń, historia zdrowia i choroby, badania diagnostyczne poszczególnych użytkowników systemu.

IKP stanowi fragment projektu P1, czyli Elektronicznej Platformy Gromadzenia, Analizy i Udostępnienia Zasobów Cyfrowych o Zdarzeniach Medycznych, którego zadaniem jest budowa elektronicznej platformy usług publicznych w zakresie ochrony zdrowia umożliwiającej organom administracji publicznej i obywatelom gromadzenie, analizę i udostępnianie zasobów cyfrowych o zdarzeniach medycznych, w zakresie zgodnym z ustawą z dnia 28 kwietnia 2011 r. o systemie informacji w ochronie zdrowia.

Dane gromadzone przez system IKP mają uwzględniać m.in. różnorodne klasyfikacje medyczne, takie jak ICD-9 (Międzynarodowa Klasyfikacja Procedur Medycznych), ICD-10 (Międzynarodowa Klasyfikacja Chorób), ATC (klasyfikacja leków). Każda z tych klasyfikacji zawiera po kilka tysięcy pozycji, którymi mogą być opisane rekordy dotyczące zdrowia i choroby w danych pacjenta. Podczas procesu analizy danych taka olbrzymia liczba atrybutów będzie wymagała

wykorzystania metod sztucznej inteligencji, takich jak np. algorytmy genetyczne, których celem będzie poszukiwanie kombinacji czynników wpływających m.in. na pojawienie się określonej jednostki chorobowej czy stan zdrowia wybranej grupy osób lub też w ogóle występowanie danego czynnika chorobowego w zespole z innymi.

Algorytmy genetyczne są dziedziną sztucznej inteligencji, w której przyjęto jako paradygmat założenia zaczerpnięte z teorii ewolucji. Zauważono bowiem, że ewolucja opiera się na zmienności populacji osobników, która to zmienność zapewnia różnorodność zestawu cech osobników, co z kolei pozwala, mówiąc w największym uogólnieniu, na przetrwanie oraz przekazywanie swoich cech tym osobnikom, które są najlepiej przystosowane do otaczającego je środowiska. Właśnie cecha adaptacji decyduje o tym, w jakim stopniu poszczególne osobniki są zdolne do przeżycia w otaczającym je środowisku[1]. Kluczowym aspektem konstrukcji i w dalszej kolejności działania algorytmów genetycznych jest sposób kodowania danych, które są nośnikiem informacji genetycznych opisujących danego osobnika. Kodowanie to natomiast implikuje strukturę danych, która opisuje dany problem, a ściśle jedno rozwiązanie danego problemu. W algorytmach genetycznych bowiem chodzi o wyłonienie spośród generowanych rozwiązań tego, które najlepiej spośród wszystkich odpowiada na postawiony problem. W przyrodzie podstawową strukturą odpowiadającą za kodowanie informacji o danym osobniku jest kwas dezoksyrybonukleinowy (popularnie zwany DNA), w którym najmniejszymi nośnikami informacji są cztery zasady, czyli: adenina, guanina, cytozyna oraz tymina. Każda z sekwencji tych czterech zasad koduje poszczególne geny lub ich grupy. Jednak na potrzeby algorytmów genetycznych oczywiste jest, że nie ma obowiązku kodowania danych z wykorzystaniem notacji czwórkowej, ale najczęściej jest wykorzystywane kodowanie binarne (zero-jedynkowe) lub liczbami całkowitymi. Algorytmy genetyczne są heurystycznymi metodami stochastycznym, w związku z czym wyniki ich działania nie są dokładnie powtarzalne. Należy więc uwzględnić statystycznie wystarczającą liczbę eksperymentów w celu wyciągnięcia właściwych wniosków na temat skuteczności wartości parametrów dla poszczególnych operatorów genetycznych.

Równocześnie należy założyć, że z biegiem czasu, po wdrożeniu (które jest planowane na koniec 2014 r.), system IKP będzie obejmował swoim zasięgiem

[1] G. Wojarnik, *Wykorzystanie metod sztucznej inteligencji w zastosowaniach internetowych*, w: *Społeczeństwo informacyjne – problemy rozwoju*, red. A. Szewczyk, Difin, Warszawa 2007; G. Wojarnik, *Metody oceny jakości algorytmów genetycznych*, w: *Technologie informacyjne dla społeczeństwa*, red. W. Chmielarz, T. Parys, Wyższa Szkoła Ekonomiczno-Informatyczna w Warszawie, Warszawa 2009.

coraz większą liczbę osób, aby objąć ostatecznie 100% populacji ludności Polski. Dane zgromadzone na potrzeby IKP staną się nieocenionym źródłem danych analitycznych na temat stanu zdrowia mieszkańców kraju.

Wykorzystanie tych danych do gruntownych analiz, które będą mogły być wykonywane również za pośrednictwem zaproponowanego w pracy ewolucyjnego systemu analizy danych, będzie skutkowało:

- poprawą zdrowia obywateli,
- poprawą jakości świadczeń zdrowotnych,
- oszczędnością czasu lekarzy i pacjentów,
- podwyższeniem jakości procesów planowania i rozwoju w systemie ochrony zdrowia.

Z jednej strony wszystkie wyżej wymienione efekty będą przekładały się na konkretne kwoty, które są obecnie wydawane w warunkach dużej niepewności informacyjnej, niejako „po omacku". Z drugiej strony po wdrożeniu systemu efekt pozyskiwania reguł poszerzających wiedzę na temat zależności występujących pomiędzy zmiennymi medycznymi wprowadzanymi do systemu IKP pozwoli – dzięki wszechstronnej analizie danych – na skierowanie ograniczonych środków pieniężnych w te miejsca, które najbardziej tego wymagają.

Dlatego też warto przyjrzeć się możliwościom wykorzystania metod sztucznej inteligencji, a w szczególności algorytmów genetycznych, do analizy danych systemu IKP, który może stać się podstawą do odkrywania nieoczekiwanych reguł towarzyszących warunkom funkcjonowania opieki medycznej w Polsce. Podsumowując, należy stwierdzić, że celem artykułu jest przedstawienie koncepcji wykorzystania algorytmów genetycznych jako centralnej metody systemu analizy danych. System ten pozwoli swoim użytkownikom na odkrywanie nawet nieoczekiwanych powiązań i zależności, które niewątpliwie mogą być poznane dzięki spodziewanej olbrzymiej ilości danych gromadzonych w ramach systemu Indywidualnego Konta Pacjenta.

2. Algorytm genetyczny jako metoda eksploracji danych medycznych

Można założyć, że algorytmy genetyczne winny być wykorzystane wszędzie tam, gdzie jest znana ogólna reguła dająca szansę rozwiązania problemu szczegółowego, dlatego można stwierdzić, że – zgodnie ze stwierdzeniem R. Tadeusie-

wicza – należy je zaliczyć do metod indukcyjnych rozwiązywania problemów[2]. Z kolei za J. Łęskim można podać, że algorytmy lub metody ewolucyjne to metody naśladujące proces naturalnej ewolucji, polegającej na dokonujących się w populacjach organizmów żywych zmianach, których kierunkiem jest przystosowanie lub adaptacja w celu zwiększenia szans przeżycia i reprodukcji organizmów[3].

Najbardziej popularnym rodzajem metod ewolucyjnych są algorytmy genetyczne. Prekursorem takiego podejścia jest J.H. Holland, który opublikował w 1962 r. pracę *Outline for a logical theory of adaptive systems*. Przedstawił w niej podstawy systemów adaptacyjnych, mogących zmieniać się w reakcji ze środowiskiem, w którym funkcjonują[4]. Właśnie dzięki tej pracy badawczej można zauważyć, że działanie algorytmów genetycznych głównie opiera się na sterowanym przez funkcję oceny (ang. *fitness function*) procesie w dużej mierze wykorzystującym rachunek prawdopodobieństwa, w efekcie którego generowane są rozwiązania coraz bardziej zgodne z założonym celem.

Działanie algorytmu genetycznego można przedstawić przez wykonanie następujących kroków[5]:

1) inicjalizację populacji,
2) obliczenie wartości funkcji dopasowania każdego osobnika z populacji,
3) reprodukcję wybranych osobników w celu stworzenia nowej populacji,
4) przeprowadzenie operacji krzyżowania i mutacji na nowej populacji,
5) przejście do kroku 2, o ile nie zajdzie warunek kończący przetwarzanie.

W algorytmie genetycznym każdy przedstawiciel populacji jest niejako prezentacją jednego rozwiązania badanego problemu. Jakość takiego rozwiązania jest określana na bazie miary jego dopasowania do kryteriów branych pod uwagę w trakcie oceny, która jest funkcją oceny danego osobnika do badanego problemu, zatem odnosi się do wartości reprezentowanej przez chromosom (osobnika) podczas obliczania wartości przystosowania[6]. W jednym przebiegu algorytmu genetycznego jest generowana nowa populacja stanowiąca zbiór osobników najlepiej przystosowanych do stworzonych warunków, których eg-

---

[2] R. Tadeusiewicz, *Odkrywanie właściwości sieci neuronowych*, Polska Akademia Umiejętności, Kraków 2007.

[3] J. Łęski, *Systemy neuronowo-rozmyte*, Wydawnictwa Naukowo-Techniczne, Warszawa 2008, s. 16.

[4] K. De Jong, D.B. Fogel, H.P. Schwefel, *A history of evolutionary computation w Handbook of Evolutionary Computation*, Oxford University Press, Oxford 1997, s. A2.3:4.

[5] J. Kennedy, R.C. Eberhart, Y. Shi, *Swarm intelligence*, Morgan Kaufman Publishers, San Francisco 2001, s. 147.

[6] D.T. Larose, *Metody i modele eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa 2008, s. 256.

zemplifikacją jest funkcja oceny. Zapewnienie elementów zmienności następuje na etapie operacji genetycznych, które przez użycie operatorów genetycznych, takich jak krzyżowanie czy mutacja, wprowadzają zmiany w genotypie osobników, a więc są to zmiany, które dotykają poszczególnych właściwości osobnika lub osobników wytypowanych do zmian.

Zasada działania klasycznego algorytmu genetycznego jest najprostszym podejściem do algorytmów genetycznych. Można stwierdzić, że jego działanie może być sterowane zarówno wieloma parametrami opisującymi funkcjonowanie tego algorytmu, jak i parametrami poszczególnych operatorów genetycznych oraz warunków brzegowych generowanych rozwiązań. W zaprezentowanym schemacie działania klasycznego algorytmu genetycznego zostały zastosowane operacje, w których są wykorzystywane funkcje losowe (probabilistyczne). W algorytmie tym generowanie populacji początkowej losowo tworzy zbiór startowy osobników. Zostaną one poddane dalszym zmianom w celu osiągnięcia rozwiązania, które – zgodnie z założeniem – będzie możliwie bliskie rozwiązaniu optymalnemu. Operator krzyżowania umożliwia przypadkowe kojarzenie osobników w celu wygenerowania osobników potomnych, których geny zostaną wymieszane ze swoich genomów (czyli materiałów genetycznych). Natomiast celem wykorzystania operatora mutacji jest wprowadzenie losowych zmian o mniejszym zasięgu, ponieważ losowo wprowadza niewielkie zmiany w genomie osobnika.

Dlaczego – pomimo że GA (ang. *genetic algorithm*, algorytm genetyczny) nie gwarantuje wyłonienia rozwiązania optymalnego – generowane wyniki najczęściej okazują się przydatne w rozwiązaniu badanego problemu? Po pierwsze, dlatego że podejścia klasyczne i zindywidualizowane zazwyczaj nie będą możliwe do przeprowadzenia, a paradygmaty GA będą przydatne w wielu różnych sytuacjach. Inny powód jest taki, że paradygmaty GA są generalnie dosyć wytrzymałe. Wytrzymałość oznacza tutaj, że algorytm może być używany do rozwiązywania wielu problemów, a nawet wielu rodzajów problemów i wymaga on minimalnej liczby specjalnych dopasowań uwzględniających specyficzne jakości określonego problemu. W typowym przypadku algorytm ewolucyjny wymaga określenia długości wektorów rozwiązania problemu, szczegółów dotyczących ich kodowania oraz funkcji ewaluacji – reszta programu implementującego działanie algorytmu nie musi być zmieniana. Ponadto metodologie wytrzymałe są generalnie szybkie i łatwe do wdrożenia. To prowadzi do określenia zasady wystarczalności: jeżeli rozwiązanie jest wystarczająco dobre, wystarczająco szybkie oraz wystarczająco tanie, to jest właśnie wystarczające. W prawie wszystkich aplikacjach świata realnego są poszukiwane rozwiązania wystarczające i najczęściej są uznawane

za rozwiązania satysfakcjonujące. Oczywiście należy zauważyć, iż rozwiązanie wystarczająco dobre oznacza, że spełnia ono wymagania specyfikacji.

Jak pisze A.M. Kwiatkowska, celem działania algorytmu genetycznego może być również uzyskanie wiedzy na temat prawdziwych, nowych, mających znaczenie dla danego projektu nieodkrytych i nieznanych jeszcze praw oraz reguł, które rządzą badanymi zjawiskami[7]. Często jedynymi przesłankami badanych procesów, którymi dysponują badacze, są obserwacje oraz świadomość istnienia prawidłowości, które za tymi obserwacjami stoją. Dlatego właśnie algorytmy genetyczne dzięki swoim właściwościom pozwalają na dotarcie i odkrycie tych reguł. Algorytmy genetyczne, czy też szerzej metody ewolucyjne, znalazły szerokie zastosowanie w analizie danych medycznych. Na przykład Y. Yu[8] użył algorytmu genetycznego z rankingiem Pareto do optymalizacji planowania leczenia w terapii naświetlania. A. Petrovski i J. McCall[9] używają algorytmu ewolucyjnego SPEA (*Strength Pareto Evolutionary Algorithm*[10]) do optymalizacji leczenia chemoterapeutycznego – optymalizowane są dwa cele: zwalczanie nowotworu oraz czas przeżycia pacjenta.

Jednym z najbardziej obiecujących i szybko rosnących obszarów zastosowań algorytmów genetycznych jest analiza danych oraz prognozowania w biologii molekularnej. W dziedzinie tej algorytmy są używane m.in. do interpretacji danych rezonansu magnetycznego w badaniu struktury DNA[11], znajdowania poprawnego porządku w nieuporządkowanych sekwencjach DNA (Parsons, Forrest oraz Burks) oraz badania struktury protein[12].

Innym przykładem zastosowania algorytmów genetycznych jest diagnoza raka szyjki macicy. Rak szyjki macicy jest jednym z najczęstszych nowotworów przy częstości występowania na poziomie 6% wszystkich nowotworów u ko-

---

7 A.M. Kwiatkowska, *Systemy wspomagania decyzji w praktyce*, Wydawnictwo Naukowe PWN, Warszawa 2007, s. 101.

8 Y. Yu., *Multi-objective decision theory for computational optimization in radiation therapy*, „Medical Physics" 1997, vol. 24, s. 1445–1454.

9 A.Petrovski, J. McCall, *Multi-objective Optimisation of Cancer Chemotherapy Using Evolutionary Algorithms*, w: *First International Conference on Evolutionary Multi-Criterion Optimization*, red. E. Zitzler, K. Deb, L. Thiele, C.A. Coello Coello, D. Corne, Springer-Verlag, Lecture Notes in Computer Science 2001/1993, s. 531–545.

10 E. Zitzler, L.Thiele, *Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach*, „IEEE Transactions on Evolutionary Computation" 1999, vol. 3(4), s. 257–271.

11 C.B. Lucasius, G. Kateman, *Application of Genetic Algorithms in Chemometrics*, ICGA, San Mateo CA 1989, s. 170–176.

12 T.M. Mitchell, *An Introduction to Genetic Algorithms*, MIT Press, Massachusetts 1996, s. 46.

biet. Standardowy test na raka szyjki macicy to badanie wymazu Papanicolaou (lub „Pap"), które polega na oględzinach wymazu z komórek szyjki macicy pod mikroskopem[13]. Badanie wymazu „Pap" jest angażujące znaczne zasoby laboratoryjne oraz intensywne i długotrwałe, wymaga także dużej precyzji; wydaje się, że dobrze nadaje się do automatyzacji. Badania pokazują, że badanie to jest jednym z najbardziej klasycznych problemów automatycznej analizy obrazu. Należy przypuszczać, że taki automatyczny system powinien działać podobnie jak biegły cytolog, szukając na dostarczonym obrazie komórek zmian, które tradycyjnie znajdowane są przez człowieka. O algorytmie, którego zadaniem jest zautomatyzowane badanie obrazu wymazu z szyjki macicy, można powiedzieć, że należy do grupy algorytmów poszukujących zestawu pożądanych cech i właściwości.

N. Marvin i in.[14] używają dyfuzyjnego algorytmu genetycznego do wyprowadzenia modeli prognostycznych. Modele prognostyczne są stosowane w celu określenia, jakie prawdopodobieństwo przeżycia ma pacjent cierpiący na nietypowy rodzaj raka. Pod uwagę brane są trzy cele: maksymalizacja prawidłowej liczby prognoz przeżycia, maksymalizacja prawidłowej liczby prognoz śmierci oraz minimalizacja liczby użytych czynników.

V. Krmicek i M. Sebag stosują algorytm nazywany 4D-Miner do funkcjonalnego obrazowania mózgu, w tym przypadku maksymalizowane są 3 cele: długość, obszar oraz wyrównanie. Ze względu na naturę problemu rozwiązanie jest uważane za ważne nawet, jeżeli zdominowane jest ono odnośnie do wyrównania i obszaru, pod warunkiem, że umieszczone jest w innych rejonach mózgu[15].

H.A. Abbass używa ewolucyjnego podejścia sztucznej sieci neuronowej opartego na algorytmie PDE (*Pareto Differential Evolution* – algorytm ewolucji różniczkowej)[16] powiększonego przez poszukiwanie lokalne do przewidywania raka piersi. Minimalizowane są dwa cele – błąd i liczba ukrytych jednostek[17].

---

[13] J. Hallinan, *Feature Selection and Classification in the Diagnosis of Cervical Cancer*, The Practical Handbook Of Genetic Algorithms Applications, Chapman & Hall/CRC 2001.

[14] N. Marvin, M. Bower, J.E. Rowe, *An evolutionary approach to constructing prognostic models*, „Artificial Intelligence in Medicine" 1999, vol. 15(2), s. 155–165.

[15] V. Krmicek, M. Sebag, *Functional Brain Imaging with Multi-objective Multi-modal Evolutionary Optimization*, w: *Parallel Problem Solving from Nature – PPSN IX, 9th International Conference*, red. T.P. Runarsson, H.-G. Beyer, E. Burke, J.J. Merelo-Guerv´os, L.D. Whitley, X. Yao, Springer, Lecture Notes in Computer Science, vol. Reykjavik 2006/4193, s. 382–439.

[16] H.A. Abbass, R. Sarker, C. Newton, *PDE: A Pareto-frontier Differential Evolution Approach for Multi-objective Optimization Problems*, w: *Proceedings of the Congress on Evolutionary Computation (CEC'2001)*, IEEE Service Center, Piscataway, New Jersey 2001, s. 971–978.

[17] H.A. Abbass, M. Towsey, G. Finn, *C-net: a method for generating nondeterministic and dynamic multivariate decision trees*, „Knowledge and Information Systems" 2001, vol. 3,

Jak widać, metody sztucznej inteligencji są w szerokim zakresie wykorzystywane również w medycynie, dlatego warto podjąć wysiłek zmierzający do wykorzystania gromadzonych licznych danych na temat historii zdrowia i choroby poszczególnych osób do odkrywania często nieoczekiwanych zależności, które niewątpliwie są – czy też będą – ukryte w bazach danych zawierających dane na ten temat.

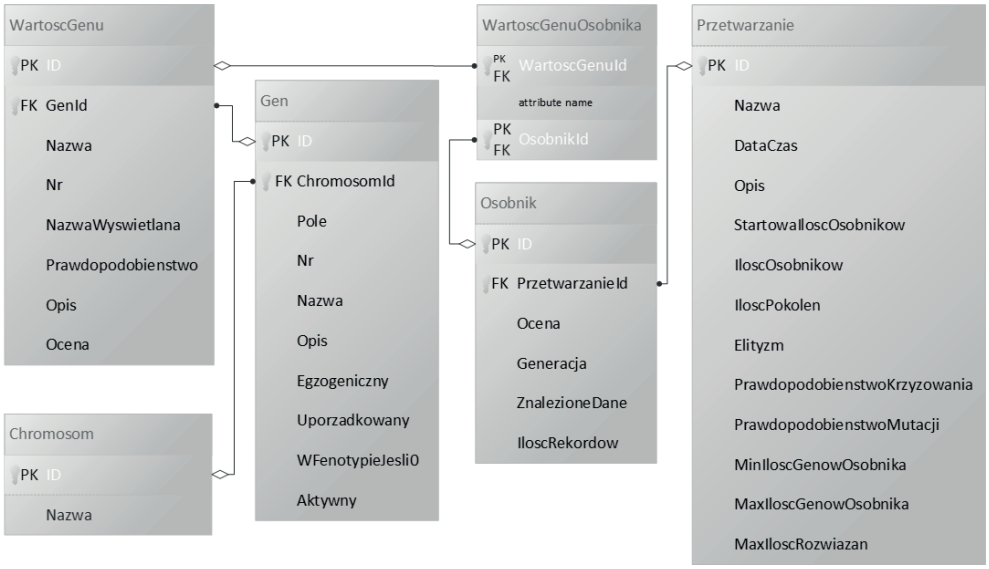
3. Koncepcja budowy ewolucyjnego systemu analizy danych opartego na algorytmie genetycznym

Wykorzystanie algorytmów genetycznych w medycynie jest na pewno związane z dynamicznym, złożonym i nieciągłym otoczeniem, które wymusza pozyskiwanie, przetwarzanie i wykorzystywanie coraz większych zbiorów danych i informacji w coraz krótszym czasie. Dlatego też prezentowana koncepcja może być określona jako element szerzej rozumianego systemu zarządzania wiedzą zgromadzoną na temat danych medycznych, ponieważ stanowi celowo skonstruowany system[18]. Celem tego systemu jest przenikanie procesów badawczych związanych z poznaniem reguł powiązanych ze stanem zdrowia i choroby osób, których dane będą gromadzone w ramach IKP. W związku z tym należy stwierdzić, że użyteczność zarządzania wiedzą dotyczącą danych medycznych będzie związana z wymiarem aplikacyjnym.

Proponowany model danych uwzględnia kodowanie algorytmu genetycznego za pomocą liczb całkowitych. Każdy gen genotypu posiada zestaw możliwych wartości (alleli), którym są przyporządkowane liczby całkowite o wartościach od 0 do *n*. Natomiast w skład genotypu wchodzą te liczby kodujące odpowiednią wartość danego genu, czyli allele. Należy dodać, że założono zmienną długość genotypu, co oznacza, że dane allele mogą znaleźć się w różnym locusie chromosomu. Algorytm genetyczny powinien operować właśnie na tych zakodowanych za pomocą liczb całkowitych wartościach każdego genu. Podejście takie czerpie z modelu danych gwiazdy w systemach typu OLAP, w którym znajduje się tabela faktów reprezentująca rejestrowane zdarzenie gospodarcze, natomiast tabele wymiarów otaczają tabelę faktów, zapewniając możliwość przetwarzania danych według konkretnego kontekstu, który jest przedstawiany przez dany wymiar.

s. 184–197.

[18] *Zarządzanie wiedzą w przedsiębiorstwie*, red. K. Perechuda, Wydawnictwo Naukowe PWN, Warszawa 2005, s. 62.



Rysunek 1. Ogólny model metadanych ewolucyjnego systemu analizy danych

Źródło: G. Wojarnik, *Ewolucyjny system analizy danych w warunkach adaptacyjnego środowiska zastosowań informatyki*, volumina.pl, Szczecin 2013.

Struktura danych opisana przez model zaprezentowany na powyższym rysunku pozwala przechowywać zarówno dane związane z rozwiązywanym problemem, jak i parametry pracy samego algorytmu. Dostępne geny opisywane są za pośrednictwem klasy **Gen**, która definiuje każdy z genów dostępnych dla badanego problemu. Klasa ta zawiera metainformacje pozwalające na przetwarzanie osobników w ramach działającego algorytmu genetycznego. Każdy gen może należeć do jednego chromosomu (klasa **Chromosom**), który powinien zawierać jeden lub więcej genów. Chromosomy pozwalają niejako grupować geny tak, aby przetwarzanie związane z wykorzystaniem operatorów genetycznych odbywało się na wybranych zestawach genów. Na potrzeby tej pracy przyjęto założenie, że każdy gen odpowiada odpowiedniemu polu (atrybutowi) z tabeli zawierającej już przetworzone dane z systemu źródłowego, które są zgodne z definicją genów zawartą w opisywanym modelu. W niniejszej pracy dane te noszą nazwę „danych weryfikacyjnych" – na ich podstawie jest dokonywane przetwarzanie danych w ramach algorytmu genetycznego, co zostanie przedstawione w dalszej części pracy. Atrybut **Pole** wskazuje więc, jakiego atrybutu danych weryfikacyjnych dotyczy dany gen. W chromosomie przetwarzaniu podlegają geny egzogeniczne, czyli takie, od których zależy wartość funkcji oceny. Geny, które nie są oznaczone jako egzogeniczne, są tymi, które stanowią kryteria oceny,

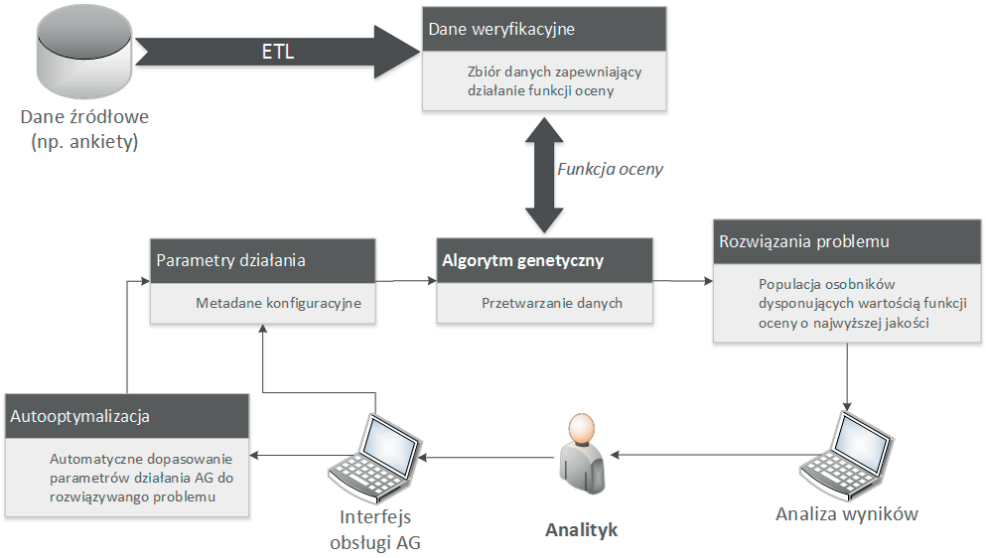
np. takim genem może być „wiek zgonu" lub „wiek zachorowania", o ile zostanie przyjęty jako czynnik zależny od kombinacji poszukiwanych czynników, które są właśnie opisywane przez zbiór zdefiniowanych genów.

Jako że celem działania proponowanego algorytmu jest wyłonienie cech oraz ich kombinacji, które w możliwie maksymalnym stopniu wpływają na badane zjawisko, analityk-ekspert wykorzystujący proponowane rozwiązanie może określić, czy dany gen ma wejść do puli wyłonionych cech, jeśli przyjmuje wartość 0 (zero), która jest podawana w ramach numeru (pole **Nr** klasy **WartoscGenow**). W związku z tym został zdefiniowany atrybut **WFenotypieJesli0**, który analityk może ustawić według własnej koncepcji. Z każdym genem w przypadku algorytmu genetycznego kodowanego za pomocą liczb całkowitych wiążą się dostępne wartości, które ten gen może przyjąć. W opisywanym modelu za wartości te odpowiada klasa **WartoscGenu**. Dzięki tej klasie można zdefiniować dostępne wartości, które będą stanowiły dziedzinę dla danego genu. Dziedzina ta będzie stanowiła podstawę generowania liczb całkowitych (atrybut **Nr**), które będą podlegały przetwarzaniu jako wartości genów podczas działania algorytmu genetycznego. Celem działania algorytmu genetycznego jest wyłonienie osobników, które możliwie najlepiej będą przystosowane do rozwiązania danego problemu. Egzemplifikacją tego jest ocena każdego wyłonionego osobnika z jego zestawem wartości poszczególnych genów. Atrybut **Ocena** informuje o wartości, jaką przyjmuje funkcja oceny dla danego osobnika stanowiącego rozwiązanie badanego problemu. Informacja na temat numeru kolejnego pokolenia, do którego taki osobnik został wygenerowany, jest dostępna dzięki atrybutowi **Generacja**.

W założonym modelu przyjęto takie rozwiązanie, że nie wszystkie geny i ich wartości, które mogą się pojawić u danego osobnika, są istotne, ale ważne są niektóre geny, niejako decydujące o wartości funkcji oceny, która jest wyrażona przez konkretną wartość będącą oceną wartościującą danego osobnika.

Celem opracowanej koncepcji jest prezentacja założeń systemu (którego strukturę przedstawiono na rysunku 2) wspierającego analizę danych umożliwiającą odkrywanie kombinacji czynników (rozumianych jako zmienne niezależne, egzogeniczne), które decydują o poziomie wartości zmiennych zależnych (endogenicznych, objaśnianych). System ten powinien pozwolić na analizę danych do wielu zastosowań w sposób możliwie przyjazny dla użytkownika i z zachowaniem minimalnej ingerencji użytkownika zmierzającej do przystosowania go do danego zastosowania. W związku z możliwością wystąpienia znacznej liczby możliwych kombinacji czynników już dla kilkudziesięciu cech niezależnych celowe jest wykorzystanie algorytmu genetycznego jako narzędzia

umożliwiającego poszukiwanie zestawu czynników o możliwie najlepszej jakości ze względu na przyjętą funkcję celu.



Rysunek 2. Struktura ewolucyjnego systemu analizy danych wykorzystującego dane weryfikacyjne jako podstawa funkcji oceny algorytmu genetycznego

Źródło: G. Wojarnik, *Ewolucyjny system analizy danych w warunkach adaptacyjnego środowiska zastosowań informatyki*, volumina.pl, Szczecin 2013.

Algorytm genetyczny stanowi serce systemu, ponieważ jest odpowiedzialny za przetwarzanie danych zmierzające do wygenerowania rozwiązań, które posłużą analitykowi podczas analizy danych. Ważne jest, aby system ten pozwalał na wszechstronne dopasowywanie parametrów do potrzeb analityka. Stąd analityk powinien dysponować interfejsem, za pomocą którego będzie mógł określać wartości parametrów działania algorytmu genetycznego. Jednocześnie należy wesprzeć użytkownika przez udostępnienie mechanizmu automatycznej optymalizacji parametrów algorytmu genetycznego, która pozwoli na dopasowanie parametrów jego działania tak, aby były one dobrane w sposób najbardziej przystający do badanego problemu.

W przedstawianej koncepcji podstawowym zadaniem algorytmu genetycznego jest generowanie rozwiązań, które będą dysponowały możliwie najwyższymi wartościami funkcji oceny. Oczywiste jest, że algorytm genetyczny nie jest w stanie przetworzyć całej dostępnej przestrzeni rozwiązań, ale mimo to daje satysfakcjonujące rozwiązania, z tym że liczba potrzebnych obliczeń zmierzających do

uzyskania rozwiązań jest nawet kilkadziesiąt tysięcy razy mniejsza niż w innym przypadku (tzn. w sytuacji, gdyby wykorzystać wobec problemu metodę *brute force*, która miałaby sprawdzić wszystkie możliwe kombinacje potencjalnych rozwiązań w celu wyłonienia najlepszych). Działanie funkcji oceny algorytmu genetycznego będzie opierało się na analizie danych weryfikacyjnych, które opisują badany problem. Dane są tworzone na bazie danych źródłowych, które pochodzą z systemów informacyjnych instytucji – organizacji, na której rzecz będzie działał opisywany system. Mogą to być np. dane ankietowe, które zostały zebrane w celu poznania zachowań lub cech badanego zjawiska. Często jest tak, że w ramach tych danych zbiera się bardzo wiele różnych cech, z których w kontekście badanego problemu tylko kilka (kilkanaście) w określonych kombinacjach wpływa na badany problem.

Dane weryfikacyjne uwzględniane podczas obliczania jakości generowanych przez algorytm genetyczny rozwiązań są umieszczone w pojedynczej tabeli o następującej strukturze:

$$\{\mathrm{ID}, \mathrm{GEN}_1, \mathrm{GEN}_2, \ldots, \mathrm{GEN}_n, Z\},$$

gdzie:
ID – identyfikator stanowiący klucz główny tabeli weryfikacyjnej,
GEN_k – wartość genu, który wskazuje na konkretną wartość spośród zdefiniowanych wartości danego genu,
Z – wartość funkcji oceny dla danego rekordu danych weryfikacyjnych.

W tym miejscu rozważań należy wskazać powiązanie pomiędzy atrybutem GEN_k a danymi na temat rozwiązywanego problemu, odwołując się do wcześniej przedstawionego modelu danych. Dane weryfikacyjne stanowią bazę działania funkcji oceny algorytmu genetycznego, ponieważ dla wygenerowanego zestawu cech będą wyszukane te dane weryfikacyjne, które będą zgodne z cechami osobnika wygenerowanego przez algorytm genetyczny. Wartość funkcji oceny dla wygenerowanego osobnika będzie obliczana właśnie na podstawie danych weryfikacyjnych odpowiadających cechom tego osobnika. Przydatny w zrozumieniu tego mechanizmu będzie poniższy przykład[19].

Dane weryfikacyjne składają się z n rekordów, trzech cech niezależnych (A, B, C) oraz jednej cechy zależnej (Z) – tabela poniżej.

[19] G. Wojarnik, *Ewolucyjny system analizy danych w warunkach adaptacyjnego środowiska zastosowań informatyki*, volumina.pl, Szczecin 2013.

Tabela 1. Ogólny schemat danych weryfikacyjnych dla funkcji oceny

Lp.	*A*	*B*	*C*	*Z*
1	A_1	B_1	C_1	Z_1
2	A_2	B_2	C_2	Z_2
...				
n	A_n	B_n	C_n	Z_n

Źródło: G. Wojarnik, *Ewolucyjny system analizy danych w warunkach adaptacyjnego środowiska zastosowań informatyki*, volumina.pl, Szczecin 2013.

Jeśli algorytm genetyczny wyłoni rozwiązanie $\{A_x, B_y\}$, to w zbiorze weryfikacyjnym zostaną znalezione wszystkie rekordy, których wartości *A* oraz *B* będą odpowiadały odpowiednim wartościom *A* i *B* w danych weryfikacyjnych. Dla takiego układu wartość funkcji oceny (*Z*) rozwiązania $\{A_x, B_y\}$ będzie obliczona według wzoru:

$$Z = (\sum_{k=1}^{m} Z_k)/m,$$

gdzie:

k – to indeks rekordu odszukanego w danych weryfikacyjnych,

Z_k – wartość funkcji oceny dla rekordu oznaczonego indeksem k danych weryfikacyjnych,

m – liczba rekordów znalezionych w zbiorze weryfikacyjnym, które odpowiadają wyłonionemu rozwiązaniu,

Z – wartość funkcji oceny dla rozwiązania $\{A_x, B_y\}$.

4. Zasady wykorzystania danych gromadzonych przez system IKP do analizy w ramach ewolucyjnego systemu analizy danych

Ogólne założenia systemu IKP znajdują się w dokumentacji konkursowej dostępnej na stronach Centrum Systemów Informacyjnych Ochrony Zdrowia (zakładka „konkurs"). Struktura danych proponowanego modelu danych bardzo dobrze nadaje się do uwzględnienia danych, które będą gromadzone w ramach systemu IKP; są to klasyfikacje medyczne, takie jak ICD-9 (Międzynarodowa

Klasyfikacja Procedur Medycznych), ICD-10 (Międzynarodowa Klasyfikacja Chorób) i ATC (klasyfikacja leków). Tworząc strukturę genomu, można podejść do tego na dwa sposoby:

1. Podejście pierwsze zakłada definicję dziedziny dla każdego genu jako wartości binarnych. Lista genów w tym podejściu stanowi odwzorowanie listy wartości poszczególnych klasyfikacji, gdzie oznaczenie TRUE określa pojawienie się danej klasyfikacji jako występującej w przypadku danego pacjenta. Wadą takiego rozwiązania jest stworzenie genomu o bardzo dużej liczbie genów wchodzących w jego skład – jest to suma liczby wszystkich kodów wyżej wymienionych klasyfikacji. Jednak przy dużej liczbie wpisów (kilkadziesiąt milionów osób w Polsce) można liczyć na odpowiednią liczbę danych, na podstawie których mogły być generowane satysfakcjonujące rozwiązania badanych problemów. Przykład fenotypu w tym podejściu to: Geny ICD-9: 6.31.2.1, 14.78.210, Geny ICD-10: M46.9, N20.2, Geny ATC: M 02 AA 13, R 05 CB 09.
2. Podejście drugie zakłada wprowadzenie do genomu genów wywodzących się nie ze wszystkich kodów klasyfikacji ICD-9, ICD-10 oraz ATC, ale wybranych podgrup celem ograniczenia liczby możliwych kombinacji potencjalnych fenotypów. Rozwiązanie takie pozwala na uzyskanie kombinacji cech bardziej ogólnych niż w przypadku pierwszego podejścia, ale za to umożliwia uzyskanie bardziej pewnych danych, dzięki przetwarzaniu mniejszego obszaru potencjalnych rozwiązań. Przykład fenotypu w podejściu drugim: Geny ICD-9: 6.31, 14.78, Geny ICD-10: M46, N20, Geny ATC: M 02, R 05.

Atrybut stanowiący wartość funkcji oceny mógłby być tworzony *ad hoc* w zależności od problemu, który byłby badany z wykorzystaniem danych zgromadzonych w IKP.

Fakt, że system IKP będzie działał na podstawie słowników i nie będą mogły pojawić się tam dane, które nie odpowiadają wartościom słownikowym, powoduje, że dostosowanie danych z systemu IKP do proponowanego podejścia analizy tych danych będzie mogło być wykonane niskim kosztem.

5. Podsumowanie

Bazy danych Internetowego Konta Pacjenta będą zawierały miliony rekordów, w których pojawią się setki, a nawet tysiące zmiennych opisujących stan zdrowia mieszkańców Polski. Jest mało prawdopodobne, aby wszystkie te zmienne były

zmiennymi niezależnymi, dlatego warto podejmować starania zmierzające do odkrywania zależności między nimi. Jednocześnie można przyjąć, że w przypadku analiz, których podstawą są dane zdrowotne, trudno się zdecydować na konkretne metody (czy to z powodu ich braku, czy odwrotnie – ich nadmiaru) służące ich eksploracji, które byłyby dopasowane do rozwiązywanych problemów na podstawie zgromadzonych danych. Ponadto dostępne metody analizy danych mogą sobie nie radzić z rozwiązaniem stawianego problemu z zakresu zarządzania systemem służby zdrowia lub analizy kosztów jego funkcjonowania zwłaszcza w sytuacji, gdy celem poszukiwań są związki między danymi, które są związkami nieoczekiwanymi lub niespodziewanymi.

Samo odkrywanie związków pomiędzy zmiennymi znajdującymi się bazie danych IKP będzie miało na celu poznanie reguł oraz praw, które rządzą zjawiskami zachodzącymi w polskim systemie opieki zdrowotnej. W dalszej kolejności znajomość tych reguł może pozwolić np. na: optymalizację kosztów, bardziej efektywne gospodarowanie dostępnymi zasobami, osiągnięcie większych zwrotów z poniesionych inwestycji czy oszczędności czasu lub zasobów potrzebnych do prowadzenia danej działalności, a w konsekwencji pozwoli na podwyższenie jakości usług medycznych świadczonych na rzecz mieszkańców kraju.

Równocześnie należy zwrócić uwagę na fakt, że w trakcie badania związków pomiędzy zmiennymi na temat choroby i zdrowia poszczególnych osób zarejestrowanych w bazie IKP najczęściej wybierane modele lub systemy są budowane w odniesieniu do konkretnego problemu lub metody, którą zamierza się ten problem rozwiązać. W związku z tym wysiłek potrzebny do zaimplementowania sposobu rozwiązania danego problemu w innym zastosowaniu jest wysiłkiem istotnym lub w ogóle polega na sporządzeniu po prostu nowego rozwiązania dostosowanego do tego innego zagadnienia, które ma zostać zbadane. Autor pracy wychodzi z założenia, że algorytmy genetyczne są metodami, które pozwalają na eksplorację danych zmierzającą do odkrywania nieoczekiwanych zależności pomiędzy zmiennymi w istniejących źródłach danych, w tym w tak bogatym w dane źródle, jakim będzie baza IKP. Dlatego też to algorytm genetyczny jest główną metodą w ramach opracowanej koncepcji ewolucyjnego systemu analizy danych, której zadaniem jest przetwarzanie danych w celu selekcji cech istotnych w kontekście badanego problemu. Można mieć nadzieję, że proponowane rozwiązanie będzie z powodzeniem wykorzystane do analizy danych systemu IKP.

Bibliografia

1. Abbass H.A., Sarker R., Newton C., *PDE: A Pareto-frontier Differential Evolution Approach for Multi-objective Optimization Problems*, w: *Proceedings of the Congress on Evolutionary Computation (CEC'2001)*, IEEE Service Center, Piscataway, New Jersey 2001.
2. Abbass H.A., Towsey M., Finn G., *C-net: a method for generating nondeterministic and dynamic multivariate decision trees*, „Knowledge and Information Systems" 2001, vol. 3.
3. De Jong K., Fogel D.B., Schwefel H.P., *A history of evolutionary computation w Handbook of Evolutionary Computation*, Oxford University Press, Oxford 1997.
4. Hallinan J., *Feature Selection and Classification in the Diagnosis of Cervical Cancer*, The Practical Handbook Of Genetic Algorithms Applications, Chapman & Hall/CRC 2001.
5. Kennedy J., Eberhart R.C., Shi Y., *Swarm intelligence*, Morgan Kaufman Publishers, San Francisco 2001.
6. Krmicek V., Sebag M., *Functional Brain Imaging with Multi-objective Multi-modal Evolutionary Optimization*, w: *Parallel Problem Solving from Nature – PPSN IX, 9th International Conference*, red. T.P. Runarsson, H.-G. Beyer, E. Burke, J.J. Merelo-Guerv´os, L.D. Whitley, X. Yao, Springer, Lecture Notes in Computer Science, vol. Reykjavik 2006/4193.
7. *Konkurs na „opracowanie koncepcji wykonania i wdrożenia prototypu Internetowego Konta Pacjenta (IKP)*, www.csioz.gov.pl/file.php?s=d2k/MTg=.
8. Kwiatkowska A.M., *Systemy wspomagania decyzji w praktyce*, Wydawnictwo Naukowe PWN, Warszawa 2007.
9. Larose D.T., *Metody i modele eksploracji danych*, Wydawnictwo Naukowe PWN, Warszawa 2008.
10. Lucasius C.B., Kateman G., *Application of Genetic Algorithms in Chemometrics*, ICGA, San Mateo CA 1989.
11. Łęski J., *Systemy neuronowo-rozmyte*, Wydawnictwa Naukowo-Techniczne, Warszawa 2008.
12. Marvin N., Bower M., Rowe J.E., *An evolutionary approach to constructing prognostic models*, „Artificial Intelligence in Medicine" 1999, vol. 15(2).
13. Mitchell T.M., *An Introduction to Genetic Algorithms*, MIT Press, Massachusetts 1996.
14. Petrovski A., McCall J., *Multi-objective Optimisation of Cancer Chemotherapy Using Evolutionary Algorithms*, w: *First International Conference on Evolutionary Multi-Criterion Optimization*, red. E. Zitzler, K. Deb, L. Thiele, C.A. Coello Coello, D. Corne, Springer-Verlag, Lecture Notes in Computer Science 2001/1993.

15. Tadeusiewicz R., *Odkrywanie właściwości sieci neuronowych*, Polska Akademia Umiejętności, Kraków 2007.
16. Wojarnik G., *Ewolucyjny system analizy danych w warunkach adaptacyjnego środowiska zastosowań informatyki*, volumina.pl, Szczecin 2013.
17. Wojarnik G., *Metody oceny jakości algorytmów genetycznych*, w: *Technologie informacyjne dla społeczeństwa*, red. W. Chmielarz, T. Parys, Wyższa Szkoła Ekonomiczno-Informatyczna w Warszawie, Warszawa 2009.
18. Wojarnik G., *Wykorzystanie metod sztucznej inteligencji w zastosowaniach internetowych*, w: *Społeczeństwo informacyjne – problemy rozwoju*, red. A. Szewczyk, Difin, Warszawa 2007.
19. Yu Y., *Multi-objective decision theory for computational optimization in radiation therapy*, „Medical Physics" 1997, vol. 24.
20. *Zarządzanie wiedzą w przedsiębiorstwie*, red. K. Perechuda, Wydawnictwo Naukowe PWN, Warszawa 2005.
21. Zitzler E., Thiele L., *Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach*, „IEEE Transactions on Evolutionary Computation" 1999, vol. 3(4).

* * *

The concept of using genetic algorithms as a method of analysing medical data collected under the Internet Patient's Account

Summary

The Internet Patient's Account (IKP) is a project whose aim is to enable the collection of data regarding a patient's medical condition in one place. This project involves the registration of all the data related to the health of Polish citizens in a central database. Therefore, it is worthwhile to make an attempt at an extensive use of data that will be collected by this system. The paper presents a concept of using the data analysis system involving a genetic algorithm, which will allow the exploration of the data collected under IKP.

Keywords: genetic algorithms, IT in healthcare, Internet Patient's Account, data analysis