



Odkrywanie reguł asocjacji z medycznych baz danych – podejście klasyczne i ewolucyjne

Halina Kwaśnicka, Kajetan Świtalski

Wrocław University of Technology, Institute of Applied Informatics,
Wyb. Wyspiańskiego 27, 50-370 Wrocław
halina.kwasnicka@pwr.wroc.pl, <http://www.ci.pwr.wroc.pl/~kwasnick>

Streszczenie. W pracy omówiono problematykę generowania reguł asocjacji z medycznych baz danych z wykorzystaniem autorskiej metody generowania reguł asocjacji wykorzystującej algorytm genetyczny EGAR, w porównaniu z klasycznym algorytmem FPTree. Dla celów badawczych opracowano program komputerowy, który jest stosunkowo uniwersalnym narzędziem do tego zadania. W pracy porównano efekty obu metod wykorzystując rzeczywiste zbiory danych medycznych z wrocławskiej kliniki.

1 Wprowadzenie

W posiadaniu różnych instytucji znajdują się obszerne zbiory danych, zgromadzone w postaci baz danych bądź hurtowni. Dane te ukrywają często obszerną i bardzo ważną wiedzę, której jednak nie można odczytać z nich bezpośrednio z użyciem standardowych środków. Zrodziła się więc potrzeba zastosowania wydajniejszych rozwiązań, dających lepsze rezultaty. Stąd też pojawienie się nowej dyscypliny naukowej nazywanej dzisiaj drążeniem danych (ang. Data Mining), a będącej częścią szerszego procesu – odkrywania wiedzy z baz danych (ang. Knowledge Discovering from Databases) [2, 10].

Obecnie ze specjalizowanych algorytmów drążenia danych i całego procesu pozyskiwania wiedzy z baz danych korzysta się w takich dziedzinach nauki jak: bankowość, marketing, telekomunikacja, energetyka, farmaceutyka, medycyna i wiele innych. We wszystkich dziedzinach można w zasadzie stosować te same metody drążenia danych, choć należy zwracać uwagę na specyfikę analizowanych danych – w niektórych zastosowaniach posiadane dane zawierają atrybuty zarówno o wartościach ciągłych, jak i symbolicznych, co należy uwzględnić w doborze metod. Niektóre metody nie nadają się do danych niepełnych, zaszumionych, itp., dlatego rodzaj danych jest równie istotny jak i cel drążenia danych. Zawsze w procesie drążenia danych powinni brać udział specjaliści z danej dziedziny, aby wydobyta wiedza mogła być oceniona i zweryfikowana. Czasami eksperci są w stanie postawić interesujące hipotezy, które metody drążenia danych potrafią zweryfikować.

Podstawowy cel tej pracy to zbadanie możliwości generowania reguł asocjacyjnych z medycznych baz danych za pomocą algorytmów genetycznych oraz porównanie użyteczności tej metody z podejściem klasycznym. Zaproponowano genetyczną, autorską metodę EGAR (Extended Genetic Association Rules). Słowo *Extended* oznacza, że autorzy oparli się na metodzie GAR [11]. Opracowany program komputerowy został tak zaprojektowany, aby praca z nim była łatwa i aby wyniki były prezentowane w czytelnej formie. Autorzy pracowali wcześniej z obiema wykorzystanymi tu bazami danych, realizując zadanie klasyfikacji [7, 8].

2 Generowanie reguł asocjacji jako zadanie drążenia danych

Za [15] możemy powiedzieć, że: Pozyskiwanie wiedzy z danych jest nietrywialnym procesem wyszukiwania rzeczywistych, nowych, potencjalnie użytecznych i zrozumiałych dla człowieka wzorców w zbiorach danych. Na proces pozyskiwania wiedzy z danych składa się sekwencja zadań, które powinny zapewnić pozyskanie użytecznej i zrozumiałej dla człowieka wiedzy:

1. Zrozumienie dziedziny problemu, zdefiniowanie celu pozyskiwania wiedzy
2. Pozyskanie potrzebnych danych
3. Wstępne przetworzenie danych – konsolidacja i oczyszczenie danych
4. Selekcja odpowiednich danych (redukcja danych) i ich wzbogacenie

5. Zakodowanie danych
6. Drażenie danych (wybór odpowiedniego zadania procesu drażenia, wybór odpowiedniej reprezentacji wiedzy i algorytmu eksploracji danych, uruchomienie algorytmu – przeszukiwanie danych i generowanie znalezionych wzorców)
7. Interpretacja, prezentacja i wyjaśnianie odkrytej wiedzy
8. Konsolidacja i praktyczne wykorzystywanie odkrytej wiedzy

Zadania drażenia danych można podzielić ze względu na postać otrzymywanej wiedzy lub na cel jej wykorzystania [15]. Najpopularniejsze zadania to: *klasyfikacja*, *klasteryzacja* danych, odkrywanie *zależności przyczynowych* oraz odkrywanie *reguł asocjacji (reguł związków)*. Każde z tych zadań może być realizowane stosując różne algorytmy [5, 13]. Niniejsza praca dotyczy odkrywania reguł asocjacji.

Odkrywanie reguł asocjacji jest najogólniejszym mechanizmem w dziedzinie drażenia danych i może być stosowane również w klasyfikacji przy pewnych założeniach. Typowa postać reguł asocjacji przypomina logiczną implikację:

Ciało → **Głowa** [*wsparcie, pewność*],

gdzie: **Ciało**, **Głowa** – zbiory atrybutów wraz z odpowiadającymi im wartościami, *wsparcie* (ang. support) – *wsparcie reguły*, *pewność* (ang. confidence) – *pewność (ufność) reguły*.

Wsparcie reguły jest to częstość współwystępowania wartości pewnych atrybutów w bazie danych; podawana jest procentowo. Jeśli *wsparcie* pewnej reguły wynosi 50%, oznacza to, że w połowie wszystkich danych z bazy, atrybuty występują z wartościami określonymi w ciele reguły. *Pewność* reguły jest to częstość współwystępowania wartości atrybutów z głowy reguły w tych rekordach bazy danych, w których występują wartości atrybutów z ciała reguły. Jeśli *pewność* reguły wynosi 50%, oznacza to, że w połowie rekordów, w których występują wartości atrybutów z ciała reguły, występują również wartości atrybutów z głowy reguły.

3 Metody generowania reguł asocjacji

Można wyodrębnić dwie zasadnicze grupy algorytmów odkrywania reguł asocjacji: algorytmy klasyczne oraz grupa algorytmów obliczeń miękkich (ang. Soft Computing). Algorytmy klasyczne są stosunkowo dobrze poznane, opisane i wykorzystywane w wielu systemach, natomiast algorytmy typu miękkiego stale się rozwijają i nie ma obecnie jednoznacznej metody pozwalającej na osiągnięcie najlepszych wyników. Najbardziej obiecującym podejściem typu *soft computing* wydają się być algorytmy genetyczne i to podejście jest wykorzystane w tej pracy.

3.1. Podejście klasyczne

Najprostszy algorytm pozyskiwania reguł to algorytm o nazwie *Apriori* [5]. Jest to algorytm generowania wzorców częstych, na podstawie których można w następnej kolejności wygenerować reguły asocjacji. Jest to iteracyjne podejście, w którym wzorce k -elementowe służą do pozyskania wzorców $(k+1)$ -elementowych. Wykorzystuje on podstawową własność wzorców częstych: każdy niepusty podzbiór zbioru częstego jest również zbiorem częstym, gdzie poprzez *zbiór częsty* (nazywany też *wzorcem częstym*) rozumiemy zbiór par <atrybut, wartość> współwystępujących w bazie z określoną częstością. Oznacza to, że jeżeli podzbiór pewnego zbioru nie jest częsty, to ten zbiór również nie jest częsty, zatem bezcelowe jest przeglądanie zbiorów będących nadzbiorem zbiorów nieczęstych, w poszukiwaniu wzorców częstych. Pozwala to zawęzić przestrzeń przeszukiwań poprzez generowanie „kandydatów” na wzorce częste $(k+1)$ -elementowe spośród nadzbiorów zbiorów częstych k -elementowych. Przy większej ilości danych wydajność *Apriori* jest niezadowolająca. Zaproponowano wiele jego modyfikacji mających na celu zwiększenie jego efektywności, jak np.: haszowanie, redukcję skanowania, itp. [5].

Algorytm drzewa wzorców częstych, zwany w skrócie *FPTree* (ang. *Frequent Pattern Tree*) jest algorytmem wydajnego generowania wzorców częstych [10]. Jego istotą jest struktura nazwana drzewem wzorców częstych (*FPTree*), która pozwala na kompresję danych zawartych w bazie.

Zadaniem algorytmu jest nie tylko stworzenie drzewa, ale też wydobycie tej informacji w postaci zrozumiałej dla użytkownika. Proces ten nazywany jest *drażeniem drzewa*, jest to algorytm rekurencyjny [10].

FPTree jest jednym z najwydajniejszych algorytmów klasycznych. Charakteryzuje się kompletnością (znajdowane są wszystkie wzorce o określonej częstości), jednokrotnym przeglądaniem danych, a co za tym

idzie – dużą wydajnością, brakiem konieczności generowania kandydatów, dużą kondensacją reprezentacji danych (dzięki strukturze FPTree), zachowaniem małego częstości wzorców. Istnieje wiele wersji algorytmu FPTree, różniących się między sobą głównie zastosowaną optymalizacją działania.

```
function draż (drzewo, wsparcie)
begin
if ( ma_jedną_gałęź(drzewo) ) generuj_wszystkie_wzorce_z(drzewo)
//generowane wzorce to wszystkie kombinacje elementów jedynej gałęzi
else if drzewo jest puste=return {}
for_each(elementy e drzewa) do
make wzorce_warunkowe for element e
//znajduje ścieżki zawierające e
filter wzorce_warunkowe, wsparcie
//pozostawia jedynie elementy wystarczającym wsparciu
make drzewo_warunkowe z wzorce_warunkowe
//podczas budowy drzewa warunkowego wzorce warunkowe są
// traktowane tak, jak rekordy podczas budowy drzewa głównego
nowe_wzorce = draż (drzewo_warunkowe, wsparcie)
for_each (nowe_wzorce) add element e
result += nowe_wzorce
end_for_each
return result
end
```

Rysunek 1. Pseudokod zaimplementowanego algorytmu FPTree

W pracy zastosowano prosty algorytm służący generowaniu reguł asocjacji na podstawie otrzymanych uprzednio wzorców częstych. Algorytm ten polega na wygenerowaniu dla każdego wzorca częstego **W** wszystkich kombinacji reguł typu: $A \Rightarrow B$, gdzie **A** to dowolny podzbiór zbioru częstego, a **B** to różnica zbiorów $W \setminus A$. Wsparcie tak utworzonej reguły równe jest częstości zbioru częstego, z którego została utworzona, natomiast pewność jest ilorzędem wsparcia zbioru częstego **W** i wsparcia zbioru **A**.

FPTree można stosować jedynie do atrybutów dyskretnych, w systemie opracowanym na potrzeby niniejszej pracy do dyskretyzacji atrybutów o dziedzinach ciągłych wykorzystano *algorytm dyskretyzacji według równej częstości*. Jest to prosty algorytm należący do grupy „bez nadzoru”, polega na dzieleniu zakresu wartości oryginalnego atrybutu na ustaloną z góry liczbę przedziałów. Granice tych przedziałów są dobierane tak, aby możliwie w każdym z nich znalazła się ta sama liczba przykładów uczących.

3.2. Algorytm genetyczny jako narzędzie drążenia danych (soft computing)

Algorytm genetyczny jest wzorowaną na naturalnej ewolucji metodą optymalizacyjną, nie gwarantującą znalezienia optymalnego rozwiązania [4, 6]. Jest jednak silną, uznaną metodą przeszukiwania, dającą często zadowalające wyniki.

Idea algorytmów genetycznych polega na zakodowaniu potencjalnego rozwiązania, utworzeniu populacji początkowych rozwiązań (nawet losowo), a następnie ich ‘ewoluowaniu’ stosując mechanizmy zaczerpnięte z biologii: lepsze osobniki są statystycznie częściej reprodukowane (mają więcej ‘dzieci’), pokolenie potomne (‘dzieci’) różni się od swoich rodziców wskutek losowego działania mechanizmów różnicujących – mutacji i krzyżowania. Po pewnym czasie ewolucji, przy dobrze dobranych parametrach ewolucji i ocenie osobników, otrzymujemy satysfakcjonujące (czasami optymalne) rozwiązania. Dzięki swojej sile, algorytmy genetyczne znalazły szerokie zastosowanie w rozwiązywaniu problemów szeregowania zadań, modelowania finansowego, optymalizacji funkcji, harmonogramowania, itp. Znajdują też szerokie zastosowanie w analizie danych medycznych, w tym, w drążeniu danych zgromadzonych w różnych klinikach świata [6, 7, 14].

Zaproponowany w niniejszej pracy algorytm genetyczny wykorzystuje doświadczenia autorów systemu GAR (Genetic Association Rules), który daje interesujące wyniki dla danych zawierających atrybuty z ciągłymi wartościami [11]. GAR nie generuje bezpośrednio reguł asocjacji, ale generuje wzorce częste, z których dopiero można wygenerować reguły asocjacji. Wiele baz danych, jak np. medyczne bazy danych, zawiera atrybuty zarówno o wartościach ciągłych jak i dyskretnych (symbolicznych), stąd nazwa zaproponowanej modyfikacji – EGAR (*Extended GAR*). Modyfikacji uległ zarówno sposób reprezentacji osobnika, jak i operatory genetyczne – mutacja i krzyżowanie. Tak przystosowany algorytm jest bardziej uniwersalny, może mieć szersze zastosowanie.

EGAR to podejście typu *Mitchigan* [4, 6] a więc genotyp osobnika zawiera informację o cząstkowym rozwiązaniu problemu – pojedynczym wzorcu częstym. Drażenie większej liczby wzorców polega na wielokrotnym przeprowadzaniu procesu ewolucji, przy czym każdorazowo z ostatniego pokolenia populacji wybierany jest najlepiej przystosowany osobnik, a reprezentowany przez niego wzorec częsty dopisany zostaje do globalnego zbioru wzorców częstych. Liczba wzorców pozyskanych z bazy zależy jest więc wyłącznie od liczby iteracji algorytmu.

Genotyp osobnika składa się z dwóch chromosomów, w którym zapisane są informacje o atrybutach występujących w zbiorze częstym oraz przedziałach, do których należą ich wartości. Jeden chromosom zawiera atrybuty ciągłe, gen w tym chromosomie jest trójką $\langle a, l, u \rangle$, gdzie a – atrybut, l – minimalna wartość (*lower value*), u – maksymalna wartość (*upper value*). Drugi chromosom tego samego osobnika zawiera atrybuty o dyskretnych wartościach, gen jest tu dwójką $\langle a, v \rangle$, gdzie a – atrybut, v – wartość atrybutu. Długość chromosomów może być różna u poszczególnych osobników, zależy od aktualnej liczby atrybutów (dyskretnych i ciągłych) w reprezentowanym przez osobnika wzorcu częstym.

Doświadczenie wskazuje, że w zadaniach drażenia danych lepiej sprawdzają się twarde metody selekcji, to znaczy takie, które w większym stopniu preferują najlepszych osobników. W prezentowanym podejściu zastosowano metodę próbkowania deterministycznego z elitaryzmem [4, 6]. Elitaryzm polega na preferowaniu najlepszych osobników, w tym wypadku, najlepszy osobnik z każdego pokolenia przechodzi bez zmian do następnego pokolenia (tzw. przeżywanie najlepszego). Jest to szczególnie istotne, biorąc pod uwagę fakt, że właśnie najlepiej oceniony osobnik dołączony zostaje do globalnego zbioru wzorców częstych.

Bardzo ważnym elementem jest funkcja oceny osobnika, tzn., na ile rozwiązanie zakodowane przez danego osobnika dobrze spełnia zadanie. W zadaniu generowania reguł asocjacji ocena ta jest wielokryterialna. Najbardziej naturalnym sposobem wydaje się być liczba pokrytych rekordów w bazie przez danego osobnika (kryterium *trafności*), ale wtedy preferowane byłyby osobniki trywialne, jedno-atrybutowe [8]. Dlatego też funkcja oceniająca jest uzależniona od innych miar, np. długości chromosomu (czyli liczby atrybutów występujących we wzorcu). Aby powiększyć miarę trafności generowanych wzorców ewolucja ma tendencję do poszerzania szerokości przedziałów wartości dla poszczególnych atrybutów aż do objęcia całej dziedziny atrybutu jednym przedziałem. Jest to oczywiście niepożądana cecha ewolucji, dlatego też w ocenie osobnika należy karać go za zbyt dużą *średnią amplitudę* przedziałów, gdzie średnia amplituda to średnia długość przedziałów zawartych w genach osobnika. Ta część oceny dotyczy chromosomu zawierającego atrybuty ciągłe. Aby nie generować wzorców częstych pokrywających te same rekordy w bazie, do funkcji oceny osobników włączono karę za pokrywanie już pokrytych rekordów. W tym celu, w każdej iteracji algorytmu, kiedy zwycięski wzorec (osobnik) dopisywany jest do zbioru wzorców częstych, pokrywane przez niego rekordy są oznaczane. Podsumowując, funkcja oceny (przystosowanie) i -tego osobnika wyraża się wzorem:

$$f_i = cov_i - a \cdot mark_i - b \cdot ampl_i + c \cdot nAtr_i$$

gdzie: cov_i – liczba rekordów pokrytych przez i -ty wzorec, $mark_i$ – liczba takich rekordów pokrytych przez i -ty wzorec rekordów, które były pokryte przez wcześniejsze wzorce, $ampl_i$ – średnia amplituda wzorca, $nAtr_i$ – liczba atrybutów w kodowanym wzorcu, a, b, c – wagi kar i nagród ustalone przez użytkownika.

Funkcja przystosowania zależy jest od obiektywnych miar jakości osobnika oraz od parametrów warunkujących wpływ tych miar, dlatego też EGAR wymaga dostrojenia wag przez użytkownika.

Zdefiniowano specjalizowane operatory dla poszczególnych chromosomów. Krzyżowanie osobników rodzicielskich daje dwóch potomków, z których lepszy (o większym przystosowaniu) jest wybierany do następnego pokolenia. Chromosomy zawierające atrybuty ciągłe są krzyżowane jak w GAR: chromosom pierwszego z potomków powstaje poprzez przepisanie wszystkich atrybutów z pierwszego kojarzonego osobnika, przy czym zakresy przedziałów wartości poszczególnych atrybutów sumowane są z odpowiednimi zakresami z chromosomu drugiego osobnika biorącego udział w krzyżowaniu, o ile w tym drugim atrybuty takie występują. Jeśli w chromosomie drugiego z krzyżowanych osobników nie występuje dany atrybut, to wartości krańców przedziału nie ulegają zmianie, to znaczy są takie same jak w pierwszym z kojarzonych osobników. Drugi chromosom potomny powstaje zaś poprzez przepisanie genów z drugiego chromosomu pierwszego osobnika biorącego udział w krzyżowaniu, a następnie dodanie wybranych w sposób losowy genów z drugiego chromosomu drugiego kojarzonego osobnika. W przypadku, gdy wylosowany do dodania gen reprezentuje atrybut istniejący już w tworzonego chromosomie, jego wartość zostaje zastąpiona nową z prawdopodobieństwem 50%. Analogicznie powstaje chromosom drugiego potomka, przy czym kojarzone osobniki zamieniane są miejscami.

Mutacja chromosomu z atrybutami ciągłymi polega na zmianie jednego lub większej liczby genów, poprzez modyfikację zawartych w nich krańców przedziałów. Wyróżnia się cztery możliwości mutacji genu: przesunięcie całego przedziału w lewo lub prawo, oraz zmniejszenie lub zwiększenie przedziału. Prawdopodobieństwo mutacji ustala się w granicach 1%, przy czym najczęściej przeprowadzana jest mutacja pojedynczego genu. Chromosom z atrybutami dyskretnymi podlega dwóm rodzajom mutacji: pierwsza polega na zastąpieniu dys-

kretniej wartości atrybutu inną, wskazaną na podstawie wartości tego atrybutu w losowo wybranym rekordzie bazy danych. Ten sposób mutacji uwzględnia rozkład wartości danego atrybutu w bazie danych, ponieważ większe jest prawdopodobieństwo wylosowania wartości atrybutu częściej występującego w bazie. Jest to istotne, bo zwiększa szanse na odpowiednie wsparcie osobnika podlegającego mutacji. Drugi rodzaj mutacji polega na zastąpieniu atrybutu reprezentowanego przez dany gen innym, losowo wybranym, dyskretnym atrybutem bazy oraz przypisaniu mu losowo wybranej wartości, przy czym w tym przypadku nowa wartość atrybutu jest losowana wprost z jego dziedziny. W ten sposób, zapewnia się, że każda wartość należąca do dziedziny atrybutu ma tę samą szansę na wylosowanie, dzięki czemu możliwe jest również odkrycie wzorców o mniejszym wsparciu.

4 Eksperymenty na medycznych bazach danych

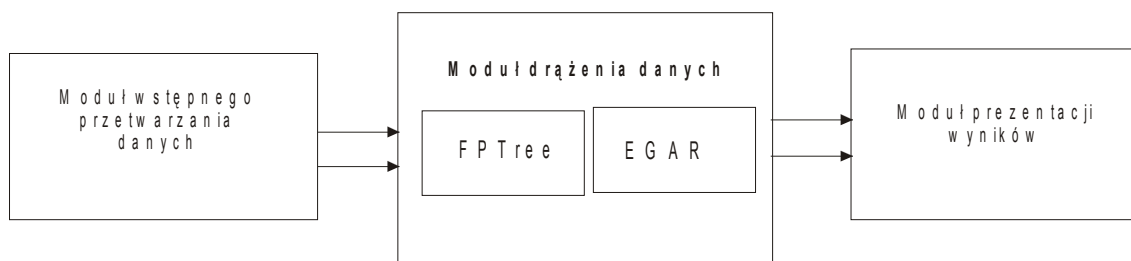
Do celów eksperymentalnych zaprojektowano i zaimplementowano obie omówione wcześniej metody generowania reguł asocjacyjnych na bazie wzorców częstych: FPTree i EGAR. Wykonany program komputerowy jest na tyle uniwersalny, że umożliwi badania na różnych bazach danych zawierających atrybuty zarówno dyskretne, jak i ciągłe.

4.1 System Antlia

System Antlia (nazwa wynika z zamiłowania astronomią współautora badań, Antlia jest nazwą gromady) jest wygodnym narzędziem do generowania reguł asocjacyjnych z danych zawierających zarówno atrybuty o wartościach ciągłych, jak i dyskretnych. Pod względem logicznym program zbudowany jest z trzech zasadniczych części: *modułu wstępnego przetwarzania danych*, *modułu drążenia danych* oraz *modułu prezentacji wyników*. Każda z tych części reprezentowana jest poprzez jedną z ikon w górnej części okna systemu. Użytkownik ma w każdej chwili możliwość przełączenia pomiędzy tymi modułami. *Antlia* umożliwia przeprowadzenie następujących operacji:

- wczytanie, wstępne odfiltrowanie oraz sortowanie danych,
- dyskretyzację danych metodą stałej częstości,
- definiowanie przez użytkownika liczby punktów podziału dyskretyzowanych dziedzin,
- zapis wstępnie przetworzonych danych w nowym pliku,
- drążenie reguł asocjacji metodą *FPTree* oraz autorską metodą *EGAR*,
- ustalanie wszystkich parametrów wybranej metody drążenia,
- śledzenie na wykresie wpływu wybranych parametrów na średnie przystosowanie populacji w algorytmie *EGAR*,
- przeglądanie, sortowanie oraz filtrowanie otrzymanych reguł asocjacji,
- zapisywanie reguł asocjacji do pliku, ich wydruk oraz podgląd wydruku.

Na rys. 2. przedstawiono ogólny schemat blokowy prezentowanego systemu. Program Antlia został napisany w środowisku Microsoft Visual Studio .NET 2002, w języku C++. W celu zapewnienia odpowiedniej wydajności, wszystkie algorytmy wykorzystywane w systemie zostały zoptymalizowane przy użyciu profilera, będącego częścią oprogramowania Compuware DevPartner Studio. Kod programu został obdarzony odpowiednim komentarzem w języku angielskim, według standardów DoxyGene.



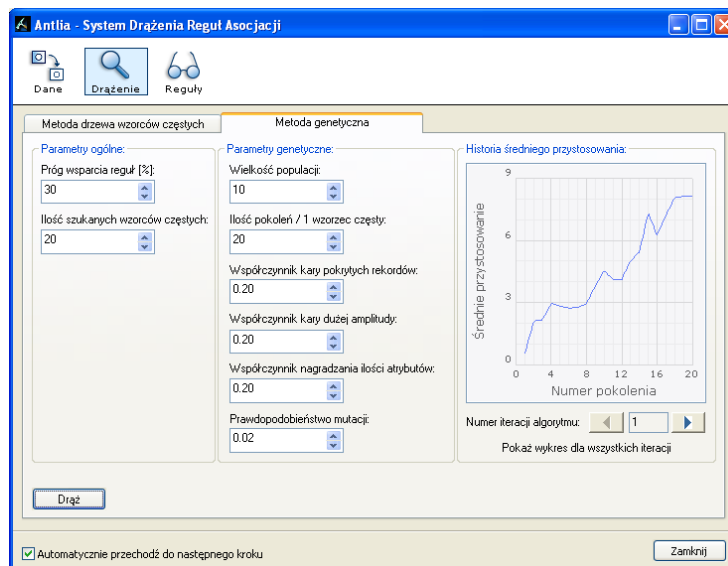
Rysunek 2. Podział systemu Antlia na moduły logiczne

Moduł wstępnego przetwarzania danych odpowiedzialny jest za odczytanie danych ze wskazanego pliku oraz ich wstępne przygotowanie do drążenia wybraną metodą. Program obsługuje własny format plików, przy-

gotowanie danych w tym formacie jest bardzo proste, sprowadza się do wyeksportowania danych w postaci pliku tekstowego, który musi spełniać dwa warunki: (1) w pierwszej linii pliku powinny znajdować się nazwy atrybutów oddzielone znakami tabulacji (nazwy mogą zawierać inne znaki białe np. spacje), (2) w każdej następnej linii pliku powinny znaleźć się wartości atrybutów oddzielone znakami tabulacji. Wczytane dane widoczne są w oknie programu w postaci tabeli. Dane te można sortować rosnąco lub malejąco według dowolnego atrybutu, możliwe jest również odfiltrowanie części rekordów. Ta część systemu umożliwia dyskretyzację atrybutów ciągłych, szczególnie ważną w przypadku drażenia reguł asocjacji metodą *FPTree*. Dyskretyzacja przeprowadzana jest *według stałej częstości*, przy czym liczba punktów podziału przeciwdziedziny atrybutu może być zdefiniowana przez użytkownika. Dodatkową możliwością jest wybór zakresu dyskretyzacji, daje to możliwość dyskretyzacji wszystkich atrybutów numerycznych, lub jedynie atrybutów o dziedzinach rzeczywistych.

Moduł drażenia danych daje użytkownikowi możliwość wyboru metody drażenia danych, dobrania jej parametrów, a następnie przeprowadzenia procesu drażenia reguł asocjacji. Moduł ten składa się z dwóch części, z których każda odpowiedzialna jest za osobną metodę drażenia danych. Użytkownik wybiera odpowiednią metodę drażenia danych poprzez wybór odpowiedniej zakładki. Dla metody *drzewa wzorców częstych*, głównymi parametrami ustalonymi przez użytkownika są: minimalne wsparcie oraz minimalna pewność szukanych reguł. Dodatkową możliwością oferowaną przez system jest wybór atrybutów, które, ze względu na cel drażenia danych, są interesujące w części konkluzji reguły. Opcja ta daje również możliwość przeprowadzenia procesu klasteryzacji danych, przy czym wartość atrybutu wybranego do części konkluzji (głowy reguły) określa przynależność do klasy, obiektu reprezentowanego przez rekord bazy danych.

Wybór drugiej zakładki w module drażenia danych powoduje przejście do *metody EGAR*. Dostępne tu pola umożliwiają ustalenie wszystkich parametrów algorytmu EGAR i przeprowadzenie drażenia danych w poszukiwaniu reguł asocjacji. Parametry konfigurujące algorytmu EGAR podzielone są na parametry ogólne, jak: minimalne wsparcie i liczba szukanych wzorców częstych oraz parametry specyficzne dla algorytmu genetycznego, czyli: wielkość populacji, liczba pokoleń przypadająca na wzorec częsty, współczynnik kary dla pokrytych rekordów, współczynnik kary dla dużej amplitudy, współczynnik nagradzania liczby pokrytych rekordów oraz prawdopodobieństwo mutacji. Użytkownik ma możliwość śledzenia wpływu parametrów na działanie algorytmu, poprzez wykres średniego przystosowania populacji na przestrzeni pokoleń (rys. 3.).



Rysunek 3. Ekran ustalania parametrów metody EGAR

Ostatnią część systemu Antlia, to moduł *prezentacji wyników*, służy do przeglądania i weryfikacji otrzymanych reguł asocjacji. Funkcje dostępne w tej części systemu umożliwiają użytkownikowi również sortowanie i filtrowanie reguł według osiąganego przez nie wsparcia i pewności. Możliwe jest zapisanie odfiltrowanych reguł asocjacji w wybranym przez użytkownika pliku.

4.2. Charakterystyka baz danych

Do eksperymentów wykorzystano medyczne bazy danych *sutek.xls* (dotyczy chorych kobiet na raka sutka) i *szyjka.xls* (dotyczy raka szyjki macicy). Obie bazy były przedmiotem wcześniejszych badań dla zadania klasyfikacji, z wykorzystaniem metod: analiza statystyczna [12], ewolucyjne generowanie reguł klasyfikujących [7] oraz wizualizacja danych [8]. W tych badaniach współpracujący lekarze wskazali cel badań i analizowali wyniki.

Wykorzystane bazy danych, jak większość rzeczywistych medycznych baz danych, nie są najłatwiejsze dla automatycznego pozyskiwania wiedzy, są stosunkowo małe, zawierają brakujące dane, zawierają atrybuty różnych typów: symboliczne, numeryczne – dyskretne i ciągłe. Większość danych opisuje tzw. typowe przypadki, co sprzyja generowaniu wiedzy typowej, znanej lekarzom. Baza *sutek.xls* zawiera 100 rekordów, dane dotyczą 21 atrybutów o różnych dziedzinach wartości. Baza *szyjka.xls* zawiera 530 rekordów, 12 atrybutów o różnych dziedzinach wartości. Należy podkreślić, że brak danych dla takich atrybutów, jak *wiek zgonu* jest oczywisty dla pacjentek, które jeszcze żyją, ale brak samej informacji o ponad 30 pacjentkach czy *żyją* jest większym problemem.

4.3 Analiza wyników

Na początku przeprowadzono drażnienie danych klasyczną metodą FPTree, która pozwala na wygenerowanie wszystkich reguł asocjacji o zadanych parametrach wsparcia i pewności.

Eksperymenty z metodą FPTree: Pierwsze eksperymenty odbyły się z bazą *sutek.xls*, bez dyskretyzacji i filtrowania danych, wsparcie i pewność ustalono na 70%. Otrzymano 340 reguł asocjacji o średnim wsparciu 73% i pewności 93%, ale – zgodnie z oczekiwaniami autorów, większość reguł była pewna, ale trywialna, np. jeśli nie nastąpił nawrót choroby, to pacjentka żyje lub, jeśli w wywiadzie krewnych nie stwierdzono raka sutka, to nie stwierdzono też również innego raka (92% pewność). W regułach, z powodu braku dyskretyzacji, nie występowały atrybuty o wartościach rzeczywistych, szukaliśmy reguł o wysokim wsparciu (70%), można oczekiwać więc, że tak wysokie wsparcie w medycznych bazach danych mają typowe przypadki, traktowane rutynowo przez lekarzy. Zatem, następne eksperymenty wykonano przy obniżonym wsparciu do 50%, ale – aby uzyskana wiedza nie była przypadkowa, podniesiono próg pewności do 95%.

Analizowano reguły po odfiltrowaniu tych ze wsparciem powyżej 70% (algorytm jest deterministyczny, te reguły mieliśmy w poprzednim eksperymencie), było ich 355, ich średnie wsparcie 56% a średnia pewność – aż 99%. Otrzymane reguły dotyczą w większości zależności okresu przeżycia pacjentek oraz czasu nawrotu choroby od stopnia złośliwości histopatologicznej. Dla laika jest to bardzo interesująca wiedza, jednakże dla lekarza nie stanowi ona wiedzy odkrywczej. Dlatego w kolejnym eksperymencie wskazano atrybuty, które mają znaleźć się w części konkluzji generowanych reguł asocjacji. Zgodnie z zainteresowaniem lekarzy w poprzednich badaniach, do konkluzji włączono atrybut *dlugość okresu przeżycia* oraz *czas do wystąpienia nawrotu*. Przy minimalnych wsparciu i pewności równych 70% uzyskano 56 reguł o średnim wsparciu 74% i pewności 96%. Jakkolwiek uzyskane reguły są w większości trywialne, zdarzają się też bardziej interesujące. Przykładowo, jeśli u bliskich krewnych nie było raka sutka oraz pacjentka w przeszłości nie chorowała na raka, to w większości przypadków (86%) ma czas przeżycia większy od 5 lat (maksymalny wyróżniony okres w danych).

Szczegółowa analiza wyników wskazuje, że w regułach występuje tylko maksymalny okres przeżycia, ponieważ występuje on najczęściej, a reguły o innej wartości nie są w stanie uzyskać wystarczającego wsparcia. Można w tej sytuacji odfiltrować z danych wstępnie te rekordy, które mają maksymalną wartość interesujących nas atrybutów lub znacznie obniżyć wymagany próg wsparcia. Pierwsze rozwiązanie wiąże się ze zbytnim zubożeniem zbioru danych – zbiór stanie się mało liczny, w drugim – będzie bardzo dużo częstych wzorców i czas drażnienia znacznie się wydłuży, uzyskana liczba reguł będzie zbyt duża, aby je móc przeanalizować. Zastosowano 'ręczną' dyskretyzację danych numerycznych, dla *dlugość przeżycia* i *czas nawrotu* wyodrębniono dwa przedziały, jednowartościowy o największej wartości i drugi, obejmujący pozostałe wartości. Uzyskane reguły osiągnęły 100 % pewność i wsparcie blisko 16%. Poza trywialnymi regułami uzyskano też interesujące, np. jeśli nastąpi wznova z odległymi przerzutami to okres przeżycia jest mniejszy niż 5 lat.

Aby sprawdzić wpływ dyskretyzacji na generowanie reguł asocjacji metodą FPTree dokonano dyskretyzacji wszystkich danych ciągłych w bazie. Obniżono parametr wsparcia do 30%, bo po dyskretyzacji wartości ciągłych, uzyskanie wysokiego wsparcia nie jest możliwe (automatyczne podzielenie przedziału wartości na n zbiorów powoduje, że maksymalne wsparcie może wynosić $100/n$ procent). Uzyskane reguły zawierały atrybuty rzeczywiste, ich średnie wsparcie to 39% a pewność 88%. Reguła, którą należałoby dać lekarzom do interpretacji to: jeśli stopień zaawansowania klinicznego wg UICC jest 2b oraz wielkość guza należy do przedziału [3,5 cm – 7,5 cm] to w wywiadzie u bliskich krewnych nie stwierdzono raka. Dyskretyzacja

wszystkich danych numerycznych powoduje generowanie dużej liczby reguł, trudno je analizować, ciężko znaleźć wśród nich interesujące.

Wykorzystanie bazy szyjka.xls powodowało te same problemy, co poprzednia baza – generowanie dużej liczby dość znanych i trywialnych reguł. Do ciekawszych reguł można zaliczyć regułę ukazującą zależność pomiędzy miejscem zamieszkania a rodzajem raka: jeśli pacjentka mieszka w mieście, to rodzaj histopatologiczny nowotworu jest *Ca plano* (pewność 91,18%). Atrybuty numeryczne zawierają mało powtarzające się wartości, dlatego też te atrybuty nie występują w odkrywanych regułach, mimo obniżania wartości wsparcia do 25%. Postępując podobnie, jak dla bazy sutek.xls, dokonując dyskretyzacji danych numerycznych i ograniczając listę atrybutów mogących wystąpić w części konkluzji, otrzymano kilka ciekawych reguł, np. jeśli w drugim leczeniu zastosowano izotop promieniotwórczy – rad, a rodzaj histopatologiczny był *Ca plano*, to pacjentka nie przeżyła.

Eksperymenty z metodą EGAR: Algorytm ten wymaga ustalania znacznie większej liczby parametrów niż FPTree, co wymaga większego zaangażowania ze strony użytkownika. Jednocześnie, metoda ta zwalnia użytkownika z konieczności dyskretyzacji atrybutów, jako że sama może dobierać granice przedziałów wartości atrybutów ciągłych (numerycznych). Jak zostało wspomniane w opisie programu, pozwala on na podgląd średniego przystosowania w populacji. Informacja ta jest ważna dla użytkownika, bowiem może on śledzić, czy przy dobranych parametrach ewolucja przebiega odpowiednio, tzn., czy nie dochodzi do przedwczesnej zbieżności, albo czy nasz algorytm nie przypomina błędzenia losowego zamiast ukierunkowanych zmian. Autorzy starali się tak dobrać domyślne parametry metody EGAR, aby w większości przypadków możliwe było uzyskanie interesujących reguł. Jak pamiętamy, w tym algorytmie liczba uzyskiwanych wzorców częstych jest parametrem metody, w każdej iteracji znajdowany jest jeden wzorec częsty. Z punktu widzenia algorytmów genetycznych, istotnymi badaniami są eksperymenty mające na celu pokazanie wrażliwości metody na parametry operatorów genetycznych, wielkość populacji itp. Z uwagi na ograniczone miejsce, nie będziemy przytaczać wszystkich wykonanych eksperymentów. Skupimy się na podsumowaniu tych badań. Warto zwrócić uwagę, że baza sutek.xls zawiera sporo atrybutów numerycznych, co powinno pozwolić na weryfikację zdolności metody do samodzielnego szukania odpowiednich przedziałów wartości tych atrybutów w regułach. Interesowano się głównie regułami o wysokiej pewności i niezbyt wysokim wsparciu, mając nadzieję, że właśnie w tej grupie znajdują się interesujące, tzn. odkrywcze reguły.

Bolączką tej metody jest skłonność do generowania reguł zbyt ogólnych, nie reprezentujących żadnej istotnej wiedzy. Wynika to z faktu, że EGAR, starając się spełnić wymóg dużego wsparcia dla generowanych reguł, jest skłonny ustalać przedziały wartości dla atrybutów numerycznych możliwie szerokie, nawet jeden przedział dla całej dziedziny. Aby temu przeciwdziałać, należy odpowiednio dobrać parametr *współczynnik kary dużej amplitudy*. Zbyt mała jego wartość spowoduje generowanie szerokich przedziałów wartości dla atrybutów numerycznych, zbyt duża wartość może zablokować proces ewolucji. Należy dbać, aby następował wzrost średniego przystosowania populacji w kolejnych iteracjach algorytmu. Zbytne ujednoczenie się osobników w populacji może być spowodowane za małym prawdopodobieństwem mutacji. Pamiętajmy również o odpowiednim ustaleniu parametru *współczynnik kary pokrytych rekordów*, który nie pozwala na to, aby kolejne generowane reguły pokrywały te same wzorce w bazie danych – reguły te reprezentowałyby tę samą wiedzę. Z istoty algorytmu wynika, iż w pierwszych pokoleniach ewolucji głównie wsparcie i pewność generowanych reguł odgrywa rolę w ewolucji, później stopniowo do głosu dochodzą parametry związane z pokrywaniem wzorców już pokrytych oraz z amplitudą przedziałów dla atrybutów numerycznych. Pozwala to na odkrywanie reguł o stosunkowo dużym wsparciu, które nie były generowane metodą FPTree, np., że pacjentki z zawartością białka *nm23* nie większą niż 6 przeżywały nie mniej niż 3 lata (92% pewność) lub u pacjentek, których bliscy krewni nie chorowali na raka, czas wznowy był nie krótszy niż 3 lata. Eksperymenty wykazały, że odpowiedni dobór współczynników kar za pokrywanie już pokrytych reguł oraz za amplitudę szerokości przedziałów wartości atrybutów numerycznych pozwala na wyeliminowanie nieefektywnych przebiegów ewolucji, tzn. takich, w których średnie przystosowanie populacji pozostaje na niskim poziomie. Jeśli parametr *nagroda ilości atrybutów* jest zbyt niski, to generowane reguły są stosunkowo krótkie. Z jednej strony jest to pozytywna cecha, bo generowane są reguły dość ogólne, ale z drugiej – takie reguły zwykle nie zawierają interesującej wiedzy. Preferując nieco dłuższe reguły, pozwalając na stosunkowo wysokie kary za dużą amplitudę, możemy uzyskać efektywne ewolucje oraz interesujące reguły, zawierające atrybuty numeryczne.

Przeprowadzono badanie efektywności metody EGAR na bazie sutek, w której atrybuty numeryczne zostały zdyskretyzowane wcześniej (EGAR nie wymaga dyskretyzacji). W wyniku otrzymano tylko 14 reguł, o stosunkowo wysokiej pewności, ale niebezpieczeństwem było utykanie algorytmu w lokalnym optimum – już po kilku pokoleniach średnie przystosowanie populacji przestało wzrastać.

Na obu bazach danych eksperymenty wykazują, że w wzrost prawdopodobieństwa mutacji pomaga zlikwidować przedwczesną zbieżność populacji. Jest to obserwacja zgodna z przesłankami teoretycznymi, to właśnie mutacja wnosi nowe wartości genów do populacji. Na uwagę zasługuje obserwacja (zgodna z teoretyczną ana-

lizą metody), że drażnienie reguł z baz dyskretnych jest procesem bardziej losowym z uwagi na fakt, że losowość wprowadzanych podczas mutacji genów odgrywa dużą rolę i przeszkadza to w ukierunkowaniu ewolucji.

5. Podsumowanie

Przeprowadzone eksperymenty oraz dokładna analiza metod pozwala stwierdzić, że metoda FPTree jest stosunkowo efektywną metodą dla pozyskiwania reguł asocjacji z baz danych o atrybutach dyskretnych. Możliwe jest stosunkowo szybko uzyskanie **wszystkich** wzorców częstych, a więc i zawartych w bazie reguł asocjacji, o zadanych wartościach minimalnego poziomu wsparcia i pewności. Stosowanie EGAR w takim przypadku nie gwarantuje znalezienia wszystkich reguł, co przemawia na jego niekorzyść. Zaprojektowany program pozwala na wskazanie atrybutów interesujących w części konkluzji, co pozwala na ukierunkowanie procesu generacji reguł. Jednakże dla drażnienia wiedzy z danych numerycznych, zwłaszcza ciągłych, metoda EGAR lepiej się sprawdza. Elastyczność automatycznego doboru przedziałów wartości poszczególnych atrybutów wykazała przewagę nad sztucznym podziałem ich dziedzin na ustaloną z góry liczbę przedziałów, umożliwia to uzyskiwanie reguł asocjacji o wyższym wsparciu. Należy też podkreślić, że EGAR jest metodą wrażliwą na parametry, jej stosowanie wymaga od użytkownika pewnego wysiłku i wyczucia. Możliwość podglądania wykresów przebiegu ewolucji ułatwia odpowiedni dobór parametrów. Biorąc pod uwagę, że mamy do czynienia z medycznymi bazami danych, należy zwrócić uwagę na fakt, że większość rekordów w bazie to bardziej typowe przypadki, dlatego też chcąc znaleźć odkrywcze reguły należy tak dobrać samą metodę pozyskiwania wiedzy, jak i jej parametry, aby nie promować zbyt wysokiego wsparcia reguł, ponieważ doprowadzi to do reguł reprezentujących wiedzę dość dobrze znaną i stosowaną przez lekarzy. Trzeba uciec się do generowania reguł niezbyt ogólnych, o wysokiej pewności i niezbyt wysokim wsparciu.

Stworzone narzędzie badawcze może służyć do pozyskiwania reguł z baz danych, dając możliwości wyboru samej metody, faktu i sposobu dyskretyzacji danych, filtrowania i sortowania danych wykorzystywanych do drażnienia, filtrowania i sortowania pozyskanych reguł, ich wygodnej prezentacji, również w postaci nadającej się do druku. System, poprzez wskazanie atrybutów do konkluzji generowanych reguł może być wykorzystany do generowania reguł klasyfikacji. Planowane jest rozbudowanie systemu o moduł drażnienia danych poprzez wizualizację danych, wykorzystując wyniki uzyskane w [8].

References

1. Aggarwal R., Prasad V.: *A tree projection algorithm for generation of frequent itemsets*. Journal of Parallel and Distributed Computing, 2001.
2. Francisci D., Brisson L., Collard M.: *A Scalar Evolutionary Approach to Rule Extraction*. Laboratoire Informatique Sinaux et Systemes, 2003.
3. Freitas A.: *A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery*. Pontificia Universidade Catolica do Parana, 2002.
4. Goldberg D. E., *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, 1989.
5. Han J., Kamber M.: *Data Mining: Concepts and Techniques*. Morgan Kauf., 2000.
6. Kwaśnicka Halina: *Obliczenia ewolucyjne w medycynie*. W: Kompendium informatyki medycznej. Red. Radosław Zajdel [i in.]. [Bielsko-Biała]: Alfa Medica Press, corp. 2003 s. 365-402, 2003.
7. Kwaśnicka H., Markowska-Kaczmar U., Matkowski R., Dryl J., Mikołajczyk P., Tomasiak J.: *Rule Discovery from Medical Data Using Genetic Algorithm*. Fourth International ICSC Symposium on Engineering of Intelligent Systems, Portugal, 2004.
8. Kwaśnicka H., Markowska-Kaczmar U., Matkowski R., Dryl J., Mikołajczyk P., Tomasiak J.: *Discovering Dependencies in Medical Data by Visualisation*. International ICSC Symposium on Engineering of Intelligent Systems, Portugal, 2004.
9. Lavington S.H., Freitas A.: *Mining Very Large Databases with Parallel Processing*. Kluwer, 1998.
10. Mao R., Yin Y., Pei P.: *Data Mining and Knowledge Discovery*. Kluwer Academic Publishers, 2004.
11. Mata J., Alvarez J. L., Riquelme J. C.: *An Evolutionary Algorithm to Discover Numeric Association Rules*. Universidad de Huelva, 2001.
12. Matkowski R.: *Wartość prognostyczna ekspresji receptora estrogenowego i produktu genu nm23 w komórkach raka przewodowego. Metody drażnienia danych w zastosowaniach medycznych i ich korelacja z wybranymi parametrami klinicznymi*. Akademia Medyczna we Wrocławiu, 2002 .
13. Olvia Parr Rud: *Data Mining Cookbook*. Wiley Computer Publishing, John Wiley & Sons, Inc, 2001
14. Peña-Reyes C. A., Sipper M. *Evolutionary computation in medicine: an overview*. Artificial Intelligence in Medicine 2000; **19** (1):1-23.
15. Piatetsky-Shapiro G., Frawley W.: *Knowledge Discovery from Databases*. Cambridge MA, 1991.