

UNIVERSITATEA DIN CRAIOVA
Facultatea de Automatică, Calculatoare și Electronică

Lavinu Aurelian Bădulescu

- REZUMATUL TEZEI DE DOCTORAT -

**OPTIMIZAREA UNOR ALGORITMI
CU ARBORI DE DECIZIE ÎN „DATA MINING”**

Conducător științific: Acad. Prof. Univ. Dr. Ing. Mircea Petrescu

1. CUPRINS

Capitolul 1. INTRODUCERE

- 1.1. Extragerea de cunoștințe din bazele de date și „Data Mining”**
- 1.2. Problema clasificării**
- 1.3. Arborii de decizie**
- 1.4. Avantaje și dezavantaje ale arborilor de decizie**
- 1.5. Organizarea tezei**
- 1.6. Contribuții personale**

Capitolul 2. PREZENTAREA LITERATURII EXISTENTE ASUPRA ARBORILOR DE DECIZIE

- 2.1. Algoritmii AID, THAID, CHAID și FIRM**
- 2.2. Algoritmii CART**
- 2.3. ID3, C4.5 și alți algoritmi înrudiți**
- 2.4. Arbori de decizie leneși**
- 2.5. Algoritmii cu arbori de decizie din pachetul SAS**
- 2.6. Algoritmi cu arbori de decizie multidimensionali**
- 2.7. Algoritmi cu arbori de decizie multidimensionali oblici**
- 2.8. Algoritmi cu arbori de decizie omnidimensionali sau micști**
- 2.9. Algoritmi cu arbori de decizie uituci**
- 2.10. Algoritmi cu arbori de decizie paraleli**
- 2.11. Algoritmii cu arbori de decizie din pachetul WEKA**
- 2.12. Arbori de decizie cu algoritmi genetici**

- 2.13. Algoritmi cu arbori de decizie pentru fluxuri de date**
- 2.14. Algoritmi pentru arbori de decizie ortogonali**
- 2.15. Arbori de decizie hibridi**
- 2.16. Alți algoritmi cu arbori de decizie**
- 2.17. Arbori de decizie adaptivi online**
- 2.18. Algoritmi cu arbori de decizie distribuiți**

Capitolul 3. ALGORITMI DE INDUCERE ȘI REDUCERE A ARBORILOR DE DECIZIE ȘI CRITERII DE SELECȚIE A ATRIBUTULUI FOLOSITE ÎN EXPERIMENTE

3.1. Algoritmi de inducere și reducere a arborilor de decizie

3.2. Criterii pentru divizarea nodurilor folosite în experimente

- 3.2.1. Criterii bazate pe impuritate folosite în experimente
 - 3.2.1.1. Criteriul câștigului în informație folosit în experimente
 - 3.2.1.2. Criteriul Gini index folosit în experimente
 - 3.2.1.3. Criteriul Gini index modificat folosit în experimente
 - 3.2.1.4. Criteriul G2
 - 3.2.1.5. Criteriul DKM
 - 3.2.1.6. Criteriul câștigului în informație pătratic folosit în experimente
 - 3.2.1.7. Criteriul câștigului în informație ponderat folosit în experimente
 - 3.2.1.8. Criteriul câștigului în informație pătratic ponderat folosit în experimente
 - 3.2.1.9. Criteriul Relief folosit în experimente
 - 3.2.1.10. Criteriul importanței folosit în experimente
 - 3.2.1.11. Criteriul χ^2 folosit în experimente
 - 3.2.1.12. Criteriul câștigului specificității folosit în experimente
 - 3.2.1.13. Criteriul câștigului specificității ponderat folosit în experimente
 - 3.2.1.14. Criteriul lungimii minime a descrierii folosit în experimente
 - 3.2.1.14.1. Criteriul complexității stohastice folosit în experimente
- 3.2.2. Criterii bazate pe normalizarea impurității folosite în experimente
 - 3.2.2.1. Criteriul raportului câștigului în informație folosit în experimente
 - 3.2.2.2. Criteriul Gini index simetric folosit în experimente
 - 3.2.2.3. Criteriul distanței folosit în experimente
 - 3.2.2.3.1. Criteriul raportului câștigului în informație simetric folosit în experimente
 - 3.2.2.4. Criteriul raportului câștigului în informație pătratic folosit în experimente
 - 3.2.2.5. Criteriul raportului câștigului în informație pătratic simetric folosit în experimente
 - 3.2.2.6. Criteriul χ^2 normalizat folosit în experimente
 - 3.2.2.7. Criterii normate ale câștigului specificității folosite în experimente
- 3.2.3. Criterii binare folosite în experimente
 - 3.2.3.1. Criteriul înjumătățirii
 - 3.2.3.2. Criteriul mediei absolute a ponderii dovezii folosit în experimente
 - 3.2.3.3. Criteriul ortogonalității
 - 3.2.3.4. Criteriul Kolmogorov-Smirnov
 - 3.2.3.5. Criteriul AUC
- 3.2.4. Criteriile K2 și Bayes-Dirichlet folosite în experimente

3.3. Metode de reducere a arborilor de decizie folosite în experimente

- 3.3.1. Reducerea complexității costului
- 3.3.2. Reducerea cu scăderea erorii
- 3.3.3. Reducerea cu eroare minimă

- 3.3.4. Reducerea pesimistă folosită în experimente
- 3.3.5. Reducerea pe baza erorii sau reducerea cu nivel de încredere folosită în experimente
- 3.3.6. Reducerea optimă
- 3.3.7. Reducerea pe baza principiului lungimii minime a descrierii

3.4. Reguli de decizie folosite în experimente

3.5. Tratarea în experimente a atributelor cu valori lipsă

Capitolul 4. MATERIALELE ȘI METODA DE LUCRU FOLOSITE ÎN EXPERIMENTE

4.1. Bazele de date din folosite în experimente

- 4.1.1. Informații generale asupra bazei de date *Abalone* folosită în experimente
- 4.1.2. Informații generale asupra bazei de date *Cylinder Bands* folosită în experimente
- 4.1.3. Informații generale asupra bazei de date *Image Segmentation* folosită în experimente
- 4.1.4. Informații generale asupra bazei de date *Iris* folosită în experimente
- 4.1.5. Informații generale asupra bazei de date *Monk's Problem* folosită în experimente
- 4.1.6. Informații generale asupra bazei de date *Adult* folosită în experimente
- 4.1.7. Informații generale asupra bazei de date *Census Income* folosită în experimente
- 4.1.8. Informații generale asupra bazei de date *Forest Covertype* folosită în experimente

4.2. Metoda de lucru folosită în experimente. Sistemul de programe

- 4.2.1. Criterii de divizare utilizate în experimente
- 4.2.2. Programul utilizat în experimente pentru inducerea arborelui de decizie
- 4.2.3. Programul utilizat în experimente pentru reducerea arborelui de decizie
- 4.2.4. Programul utilizat în experimente pentru testarea preciziei de clasificare a arborelui de decizie
 - 4.2.4.1. Programul utilizat în experimente pentru calcularea matricei de confuzie
- 4.2.5. Programul utilizat în experimente pentru generarea regulilor de decizie asociate arborelui de decizie

Capitolul 5. REZULTATE EXPERIMENTALE ȘI DISCUȚII

5.1. Experimente pe baza de date *Abalone*

- 5.1.1. Experimente la inducerea arborilor de decizie cu diverse măsuri
 - 5.1.1.1. Experimente la extragerea regulilor de decizie din arborii de decizie nereduși
 - 5.1.1.2. Experimente la execuția arborilor de decizie nereduși
 - 5.1.1.2.1. Experimente la determinarea acurateții clasificării pe datele de antrenare
 - 5.1.1.2.2. Experimente la determinarea acurateții clasificării pe datele de test
- 5.1.2. Experimente la reducerea arborilor de decizie cu metoda de reducere cu nivel de încredere
 - 5.1.2.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși cu nivel de încredere
 - 5.1.2.2. Experimente la execuția pe datele de test a arborilor de decizie reduși cu nivel de încredere
- 5.1.3. Experimente la reducerea arborilor de decizie cu metoda de reducere pesimistă
 - 5.1.3.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși pesimist
 - 5.1.3.2. Experimente la execuția pe datele de test a arborilor de decizie reduși pesimist
- 5.1.4. Numărul regulilor de decizie și rata erorii de clasificare pentru cele trei tipuri de arbori de decizie
- 5.1.5. Experimente înrudite pe baza de date *Abalone*

5.2. Experimente pe baza de date *Cylinder bands*

- 5.2.1. Experimente la inducerea arborilor de decizie cu diverse măsuri
 - 5.2.1.1. Experimente la extragerea regulilor de decizie din arborii de decizie nereduși
 - 5.2.1.2. Experimente la execuția pe datele de test a arborilor de decizie nereduși
- 5.2.2. Experimente la reducerea arborilor de decizie cu metoda de reducere cu nivel de încredere
 - 5.2.2.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși cu nivel de încredere
 - 5.2.2.2. Experimente la execuția pe datele de test a arborilor de decizie reduși cu nivel de încredere
- 5.2.3. Experimente la reducerea arborilor de decizie cu metoda de reducere pesimistă
 - 5.2.3.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși pesimist
 - 5.2.3.2. Experimente la execuția pe datele de test a arborilor de decizie reduși pesimist

- 5.2.4. Numărul regulilor de decizie și rata erorii de clasificare pentru cele trei tipuri de arbori de decizie
- 5.2.5. Experimente înrudite pe baza de date *Cylinder Bands*

5.3. Experimente pe baza de date *Statlog. Image Segmentation / Satimage*

- 5.3.1. Experimente la inducerea arborilor de decizie cu diverse măsuri
 - 5.3.1.1. Experimente la extragerea regulilor de decizie din arborii de decizie nereduși
 - 5.3.1.2. Experimente la execuția pe datele de test a arborilor de decizie nereduși
- 5.3.2. Experimente la reducerea arborilor de decizie cu metoda de reducere cu nivel de încredere
 - 5.3.2.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși cu nivel de încredere
 - 5.3.2.2. Experimente la execuția pe datele de test a arborilor de decizie reduși cu nivel de încredere
- 5.3.3. Experimente la reducerea arborilor de decizie cu metoda de reducere pesimistă
 - 5.3.3.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși pesimist
 - 5.3.3.2. Experimente la execuția pe datele de test a arborilor de decizie reduși pesimist
- 5.3.4. Calcularea matricilor de confuzie pe baza rezultatelor obținute din experimente
- 5.3.5. Numărul regulilor de decizie și rata erorii de clasificare pentru cele trei tipuri de arbori de decizie
- 5.3.6. Experimente înrudite pe baza de date *Image Segmentation / Satimage* din Statlog

5.4. Experimente pe baza de date *Iris*

- 5.4.1. Experimente la inducerea arborilor de decizie cu diverse măsuri
 - 5.4.1.2. Experimente la extragerea regulilor de decizie din arborii de decizie nereduși
- 5.4.2. Experimente la reducerea arborilor de decizie cu metoda de reducere cu nivel de încredere
 - 5.4.2.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși cu nivel de încredere
- 5.4.3. Experimente la reducerea arborilor de decizie cu metoda de reducere pesimistă
 - 5.4.3.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși pesimist
- 5.4.4. Numărul regulilor de decizie pentru cele trei tipuri de arbori de decizie obținute din experimente

5.5. Experimente pe baza de date *Monk's problem*

- 5.5.1. Experimente la inducerea arborilor de decizie cu diverse măsuri
 - 5.5.1.1. Experimente la extragerea regulilor de decizie din arborii de decizie nereduși
 - 5.5.1.2. Experimente la execuția pe datele de test a arborilor de decizie nereduși
- 5.5.2. Experimente la reducerea arborilor de decizie cu metoda de reducere cu nivel de încredere
 - 5.5.2.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși cu nivel de încredere
 - 5.5.2.2. Experimente la execuția pe datele de test a arborilor de decizie reduși cu nivel de încredere
- 5.5.3. Experimente la reducerea arborilor de decizie cu metoda de reducere pesimistă
 - 5.5.3.1. Extragerea regulilor de decizie din arborii de decizie reduși cu metoda de reducere pesimistă
 - 5.5.3.2. Experimente la execuția pe datele de test a arborilor de decizie reduși pesimist
- 5.5.4. Numărul regulilor de decizie și rata erorii de clasificare pentru cele trei tipuri de arbori de decizie
- 5.5.5. Experimente înrudite pe baza de date *Monk 1*

5.6. Experimente pe baza de date *Adult*

- 5.6.1. Experimente la inducerea arborilor de decizie cu diverse măsuri
 - 5.6.1.1. Experimente la extragerea regulilor de decizie din arborii de decizie nereduși
 - 5.6.1.2. Experimente la execuția pe datele de test a arborilor de decizie nereduși
- 5.6.2. Experimente la reducerea arborilor de decizie cu metoda de reducere cu nivel de încredere
 - 5.6.2.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși cu nivel de încredere
 - 5.6.2.2. Experimente la execuția pe datele de test a arborilor de decizie reduși cu nivel de încredere
- 5.6.3. Experimente la reducerea arborilor de decizie cu metoda de reducere pesimistă
 - 5.6.3.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși pesimist
 - 5.6.3.2. Experimente la execuția pe datele de test a arborilor de decizie reduși pesimist
- 5.6.4. Numărul regulilor de decizie și rata erorii de clasificare pentru cele trei tipuri de arbori de decizie
- 5.6.5. Experimente înrudite pe baza de date *Adult*

5.7. Experimente pe baza de date *Census Income*

- 5.7.1. Experimente la inducerea arborilor de decizie cu diverse măsuri

- 5.7.1.1. Experimente la extragerea regulilor de decizie din arborii de decizie nereduși
- 5.7.1.2. Experimente la execuția pe datele de test a arborilor de decizie nereduși
- 5.7.2. Experimente la reducerea arborilor de decizie cu metoda de reducere cu nivel de încredere
 - 5.7.2.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși cu nivel de încredere
 - 5.7.2.2. Experimente la execuția pe datele de test a arborilor de decizie reduși cu nivel de încredere
- 5.7.3. Experimente la reducerea arborilor de decizie cu metoda de reducere pesimistă
 - 5.7.3.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși pesimist
 - 5.7.3.2. Experimente la execuția pe datele de test a arborilor de decizie reduși pesimist
- 5.7.4. Numărul regulilor de decizie și rata erorii de clasificare pentru cele trei tipuri de arbori de decizie
- 5.7.5. Experimente înrudite pe baza de date *Census Income*

5.8. Experimente pe baza de date *Forest Covertype*

- 5.8.1. Experimente la inducerea arborilor de decizie cu diverse măsuri
 - 5.8.1.1. Experimente la extragerea regulilor de decizie din arborii de decizie nereduși
 - 5.8.1.2. Experimente la execuția pe datele de test a arborilor de decizie nereduși
- 5.8.2. Experimente la reducerea arborilor de decizie cu metoda de reducere cu nivel de încredere
 - 5.8.2.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși cu nivel de încredere
 - 5.8.2.2. Experimente la execuția pe datele de test a arborilor de decizie reduși cu nivel de încredere
- 5.8.3. Experimente la reducerea arborilor de decizie cu metoda de reducere pesimistă
 - 5.8.3.1. Experimente la extragerea regulilor de decizie din arborii de decizie reduși pesimist
 - 5.8.3.2. Experimente la execuția pe datele de test a arborilor de decizie reduși pesimist
- 5.8.4. Numărul regulilor de decizie și rata erorii de clasificare pentru cele trei tipuri de arbori de decizie
- 5.8.5. Experimente înrudite pe baza de date *Forest Covertype*
- 5.8.6. Experimente pentru îmbunătățirea acurateții clasificării pe datele de test
 - 5.8.6.1. Experimente la inversarea setului de date de antrenare cu setul de date de test

Capitolul 6. CONCLUZII ALE EXPERIMENTELOR ȘI DIRECȚII PENTRU EXPERIMENTE VIITOARE

6.1. Concluzii ale experimentelor privind performanțele diferitelor măsuri utilizate la inducerea arborilor de decizie

6.2. Concluzii ale experimentelor privind performanțele metodelor de reducere

6.3. Direcții pentru experimente viitoare

Referințe

2. CUVINTE CHEIE: *data mining, problema clasificării, arbori de decizie, setul de date de antrenare, setul de date de test, clasă, etichete ale clasei, criteriul de selecție a atributului, extragerea cunoștințelor, reguli de decizie, arbore de decizie neredus, metode de reducere, arbore de decizie redus, rata erorii de clasificare, precizia clasificării.*

3. SINTEZE ALE PĂRȚILOR PRINCIPALE ALE TEZEI DE DOCTORAT

Obiectivele tezei de doctorat au vizat:

- optimizarea algoritmilor cu arbori de decizie (AD) în direcția găsirii unor criterii optime de divizare ale setului de date de la nivelul unui nod al AD și a unor metode optime de reducere a AD;
- dezvoltarea unui cadru algoritmic unificat pentru abordarea algoritmilor cu AD, a criteriilor de divizare și a metodelor de reducere;

- dezvoltarea unui cadru experimental în care, pe baza unui sistem propus de programe pentru inducerea, reducerea și executarea AD, a generării matricei de confuzie și a regulilor de decizie asociate AD, să fie realizată cercetarea.

Teza de doctorat este structurată pe șase capitole precedate de un cuprins al tezei și urmate de bibliografia utilizată la scrierea acesteia. Cuprinsul este structurat astfel încât, în problematica domeniului, include părțile: Introducere (Capitolul 1), Stadiul cunoașterii (Capitolul 2), Contribuții proprii (Capitolul 3, Capitolul 4 și Capitolul 5), Concluzii finale (Capitolul 6) și Bibliografie.

În **Capitolul 1, Introducere**, se realizează introducerea în problematica temei tratate. Astfel, se prezintă raportul dintre extragerea de cunoștințe din bazele de date și data mining (DM), problema clasificării, conceptele fundamentale privind AD, precum și avantajele și dezavantajele algoritmilor cu AD. În finalul capitolului se prezintă modul cum este organizată teza și care sunt principalele contribuții originale ale autorului ei. Una dintre principalele contribuții personale la nivel teoretic, este construirea unui cadru unificat în care au fost tratate unitar câteva zeci de criterii de divizare a setului de date de la nivelul unui nod al AD și, în același cadru matematic, prezentarea și a câtorva metode de reducere a AD.

Considerând că principalele abordări de optimizare a algoritmilor cu AD trebuie să meargă în primul rând pe două direcții: ale alegerii unui criteriu de divizare optim și a celei mai bune metode de reducere, contribuția personală majoră a tezei la nivel practic, este reprezentată de descoperirea, pe baza unui număr foarte mare de experimente personale, a unui criteriu de divizare care manifestă sistematic cele mai bune performanțe, comparativ cu câteva zeci de alte criterii luate în considerație în experimente. În același timp, s-a demonstrat superioritatea metodei de reducere a AD cu nivel de încredere. Astfel au fost concepuți noi algoritmi de AD, optimizați, reprezentând o contribuție personală a tezei.

O altă contribuție majoră a tezei de doctorat a fost reprezentată de dezvoltarea unui cadru experimental în care, pe baza unui sistem de programe pentru inducerea, reducerea și executarea AD, a generării matricei de confuzie și a regulilor de decizie asociate AD, au fost realizate numeroase experimente folosind baze de date cu caracteristici foarte diferite. Rezultatele experimentale au fost comparate cu numeroase rezultate ale altor cercetători care au experimentat pe aceleași baze de date, dar folosind alte tipuri de algoritmi de DM, constatându-se că rezultatele experimentelor personale sunt la fel de bune ca cele mai bune valori din literatură.

O altă contribuție personală importantă este reprezentată de studiul din Capitolul 2, rezultatul parcurgerii exhaustive și la zi a unui volum imens de lucrări în domeniul AD, coroborat cu un efort susținut de sintetizare și sistematizare a materialului lecturat. Acest capitol prezintă o sinteză personală a principalilor algoritmi de AD, realizată pe baza unei bibliografii exclusiv din literatura străină.

Capitolul 2, Prezentarea literaturii existente asupra arborilor de decizie, realizează o sinteză personală a principalilor algoritmi de AD, obținută în urma parcurgerii și sistematizării unei importante bibliografii din literatura străină. Astfel, sunt prezentați numeroși algoritmi de AD: algoritmi AID, THAID, CHAID și FIRM, algoritmul CART, ID3, C4.5 și alți algoritmi înrudiți, AD leneși, algoritmi cu AD din pachetul SAS, algoritmi cu AD multidimensionali, algoritmi cu AD multidimensionali oblici, algoritmi cu AD omnidimensionali, algoritmi cu AD uituci, algoritmi cu AD paraleli, algoritmi cu AD din pachetul WEKA, AD cu algoritmi genetici, algoritmi cu AD pentru fluxuri de date, algoritmi pentru AD ortogonali, AD hibridi, AD adaptivi online, algoritmi cu AD distribuți precum și alți algoritmi cu AD.

Capitolul 3, Algoritmi de inducere și reducere a arborilor de decizie și criterii de selecție a atributului folosite în experimente, reprezintă o contribuție originală, propunând un cadru algoritmic unificat pentru abordarea algoritmilor de AD și furnizează o descriere aprofundată a

zeci de măsuri de selecție a atributului pentru divizarea setului de date corespunzător unui nod și a câtorva metode de reducere a AD (reducerea complexității costului, reducerea cu scăderea erorii, reducerea cu eroare minimă, reducerea pesimistă, reducerea pe baza erorii/reducerea cu nivel de încredere, reducerea optimă și reducerea pe baza principiului lungimii minime a descrierii). Capitolul 3 prezintă algoritmi propuși pentru inducerea și reducerea AD, criterii bazate pe impuritate, criterii bazate pe normalizarea impurității, criterii binare, K2 și Bayes-Dirichlet utilizate ulterior în experimente. Totodată se prezintă modalitatea de transformare a AD în setul regulilor de decizie și modul de tratare a atributelor cu valori lipsă.

Capitolul 4, Materialele și metoda de lucru folosite în experimente, prezintă cele opt baze de date folosite în experimente, ale căror dimensiuni ajung până la aproape 600.000 de cazuri, 55 de atribute (continui și discrete), clase cu 7 etichete, numeroase cazuri cu valori lipsă și înregistrări duplicate sau contradictorii. Tot în acest capitol se prezintă sistemul de programe propus și utilizat în experimente: programul pentru inducerea AD, cel pentru reducerea AD, cel pentru testarea preciziei de clasificare a AD, cel pentru calcularea matricei de confuzie și cel pentru generarea regulilor de decizie asociate AD.

Capitolul 5, Rezultate experimentale și discuții, prezintă experimentele realizate pentru fiecare din cele 8 baze de date pe baza unui set identic de prelucrări. Performanțele fiecărui pas al prelucrărilor au fost prezentate, în tabele și diagrame însoțite de comentarii, la fiecare pas al experimentelor.

1. Primul pas al experimentelor a fost reprezentat de inducerea AD pe baza setului de date de antrenare. La acest pas au fost induși 28 de AD, corespunzător celor 28 de măsuri de selecție a atributului. Au fost luate în considerare valorile unor parametri ai procesului de construcție al AD: numărul de noduri al AD indus, numărul de atribute necesar inducerii AD, numărul de nivele ale AD, timpul de creștere a AD, mărimea fișierului conținând AD indus. S-au comparat și discutat aceste valori, s-au calculat indicatori statistici, s-au prezentat tabele și diagrame însoțite de comentarii, extrăgându-se concluzii asupra comportamentului diferit al celor 28 de AD induși cu cele 28 de măsuri.

2. La al doilea pas al experimentelor, acești 28 de AD în formă neredusă au fost prelucrați pentru a se extrage din ei cele 28 de seturi de reguli de decizie, câte un set de reguli de decizie pentru fiecare AD obținut la pasul întâi. Au fost luate în considerare valorile unor parametri ai procesului de extragere a regulilor de decizie din AD: numărul regulilor de decizie, timpul de construire a fișierului conținând regulile de decizie, timpul de citire a fișierului conținând AD, mărimea fișierului conținând setul de reguli de decizie. S-au comparat și discutat aceste valori, s-au calculat indicatori statistici, s-au prezentat tabele și diagrame însoțite de comentarii, extrăgându-se concluzii asupra comportamentului diferit al celor 28 de AD induși cu cele 28 de măsuri.

3. Al treilea pas al experimentelor, realizat pe fiecare bază de date, a presupus execuția celor 28 de AD induși la pasul întâi. Aceste experimente au verificat modelul de clasificare reprezentat de AD pe baza unui set de date de test, set de date necunoscut la pasul întâi, când a fost generat AD. Este cel mai important pas al experimentelor, deoarece acum se obține valoarea ratei erorii de clasificare pe datele de test sau precizia modelului. Pentru fiecare bază de date se obțin 28 de valori ale preciziei clasificării pe datele de test a AD în formă neredusă, corespunzător fiecăreia din cele 28 de măsuri cu care a fost indus AD. Prin compararea acestor valori se pot trage concluzii asupra performanțelor uneia sau alteia dintre măsurile de selecție a atributului.

4. Al patrulea pas al experimentelor a vizat reducerea celor 28 de AD, obținuți la pasul întâi, cu metoda de reducere cu nivel de încredere. Au fost luate în considerare valorile unor parametri ai procesului de reducere a AD: nivelul de încredere, numărul de atribute necesar construirii AD redus, numărul de noduri și numărul de nivele ale AD redus, timpul de reducere a fișierului conținând AD neredus, mărimea fișierului conținând AD redus. S-au comparat și discutat aceste

valori, s-au calculat indicatori statistici, s-au prezentat tabele și diagrame însoțite de comentarii, extrăgându-se concluzii asupra comportamentului diferit al celor 28 de AD reduși cu nivel de încredere.

5. Al cincilea pas al experimentelor a constat în extragerea regulilor de decizie din cei 28 de AD reduși cu metoda de reducere cu nivel de încredere la pasul patru. Prelucrările au semănat cu cele de la pasul 2.

6. Pasul al șaselea a presupus execuția pe datele de test a celor 28 de AD reduși cu metoda de reducere cu nivel de încredere la pasul patru. Au fost reținute valorile ratei erorii de clasificare pe datele de test, valori care ulterior au fost comparate cu valorile obținute la execuția pe datele de test a AD nereduși (vezi pasul 3) și cu valorile obținute la execuția pe datele de test a AD reduși cu metoda de reducere pesimistă (vezi pasul 7).

7. Al șaptelea pas al experimentelor a vizat reducerea celor 28 de AD, obținuți la pasul 1, cu metoda de reducere pesimistă. Prelucrările au semănat cu cele de la pasul 4.

8. Al optulea pas al experimentelor a constat în extragerea regulilor de decizie din cei 28 de AD reduși cu metoda de reducere pesimistă la pasul șapte. Prelucrările au semănat cu cele de la pasul 2.

9. Pasul al nouălea a presupus execuția pe datele de test a celor 28 de AD reduși cu metoda de reducere pesimistă la pasul șapte. Prelucrările au semănat cu cele de la pașii 3 și 6.

10. La pasul al zecelea s-a realizat o discuție asupra valorilor numărului de reguli de decizie și a valorilor ratei erorii de clasificare pe datele de test pentru cele trei tipuri de AD: neredus, redus cu metoda de reducere cu nivel de încredere și redus cu metoda de reducere pesimistă, evidențiind care măsură de selecție a atributului și care metodă de reducere a avut cel mai bun comportament în experimente.

11. În final, la pasul unsprezece, au fost trecute în revistă rezultatele unor lucrări reprezentative din literatura de specialitate ce prezintă teste de performanță pe baza de date respectivă și s-au comparat rezultatele obținute în acele experimente cu rezultatele obținute de cele 28 de măsuri și cele două metode de reducere. Din motive de spațiu, la o singură bază de date (*Image*) am calculat matricele de confuzie.

Toate experimentele pe cele 8 baze de date au fost realizate pe un calculator PC AMD Duron 995 MHz CPU cu 512 MB RAM, sub sistemul de operare Windows XP Professional Version 2002 Service Pack 2, iar codul a fost scris în limbajul de programare C folosind mediul de programare integrat Microsoft Visual C++ 6.0 Enterprise Edition.

În **Capitolul 6**, se prezintă **Concluzii ale experimentelor și direcții pentru experimente viitoare**. Astfel, folosind trei tehnici se compară performanțele la clasificare ale celor 28 de AD induși și reduși cu două metode de reducere pe cele opt baze de date.

Prima dată comparația se realizează pe baza valorilor mediei aritmetice și a abaterii standard a valorilor ratei erorii de clasificare pe datele de test pentru toate tipurile de AD, toate bazele de date și toate măsurile de selecție a atributului. Se constată că cel mai bine clasifică AD indus cu măsura *rciq* (raportul câștigului în informație pătratic), cu o medie a ratei erorii de clasificare pe datele de test de 18,59% (având și cea mai mică valoare pentru abaterea standard: 14,11), urmată de AD induși cu măsurile *rel* (relief), *k2*, *bd* (Bayes-Dirichlet), *lmdfa* (lungimea minimă a descrierii în frecvențe absolute) și *ersh* (complexitatea stohastică) cu valori mici pentru abaterea standard, iar cea mai slabă performanță medie o realizează AD indus cu măsura *ciqp* (câștigul în informație pătratic ponderat), cu o valoare medie de 31,12%.

AD indus cu măsura *rciq* are cea mai bună valoare a mediei ratei erorii de clasificare și cea mai mică abatere standard, deci cea mai mică împrăștiere a valorilor ratei erorii de clasificare pe datele de test în jurul mediei acestora. Acest rezultat ne indică faptul că performanța AD indus cu

măsura *rciq* nu este afectată prea mult de caracteristicile bazei de date. Adică, una din caracteristicile fundamentale ale măsurii *rciq* este independența de domeniu, caracteristică foarte importantă astăzi, când bazele de date sunt formate din date, cu atribute aparținând la domenii diferite, colectate împreună. Diversitatea domeniilor din alcătuirea bazelor de date este unul dintre motivele care au dus la creșterea necesității utilizării unor instrumente de extragere automată a cunoștințelor din bazele de date și a algoritmilor de învățare inductivă.

Rezultatele experimentelor ne arată că cea mai mică medie a numărului de reguli de decizie îl prezintă AD indus cu măsura *mapd* (media absolută a ponderii dovezii) cu o medie a numărului de reguli de decizie de 1.147,54, iar cea mai slabă performanță medie o realizează AD indus cu măsura *rcs* (raportul câștigului specificității) cu o performanță medie de 1.641,25. Să observăm că AD indus cu măsura *rciq*, care are cea mai bună comportare din punctul de vedere al preciziei clasificării pe datele de test, se plasează pe locul patru în ierarhia celor mai mici medii pentru numărul de reguli de decizie.

A doua modalitate de comparație a valorilor performanțelor de clasificare a celor 28 de AD a fost reprezentată de tehnica *win/tie/loss*. Și în acest caz, performanțele AD indus cu măsura *rciq* sunt totdeauna cele mai bune. AD indus cu măsura *rciq* prezintă cel mai mare număr de cazuri când are o valoare a performanței mai bună, decât toate celelalte măsuri și în același timp, cel mai mic număr de cazuri când are o valoare a performanței mai slabă decât a celorlalte măsuri. Următoarele cinci locuri sunt ocupate de AD induși cu alte cinci măsuri *csk*, *bd*, *k2*, *lmdfa* și *rel*, care însă nu mai prezintă proprietatea manifestată de măsura *rciq*, de a avea atât un număr mare de cazuri când depășesc valorile performanțelor AD induși cu toate celelalte măsuri, cât și un număr mic de cazuri când sunt depășite de valorile performanțelor celorlalți AD.

A treia modalitate de comparație a valorilor performanțelor de clasificare a celor 28 de AD a fost reprezentată de media geometrică a raportului ratei erorii de clasificare. Performanțele AD indus cu măsura *rciq* întrec, în toate cazurile considerate în experimente, performanțele AD induși cu celelalte 27 măsuri. Îi urmează, în ordine, valorile performanțelor realizate de AD induși cu măsurile *rel*, *gim* (gini index modificat), *bd*, *k2*, *lmdfa* și *csk*. Astfel constatăm că, indiferent de criteriul de comparație considerat, optimizarea algoritmilor cu AD prin utilizarea măsurii *rciq* induce AD cei mai performanți. Și optimizarea pe baza măsurilor *rel*, *k2*, *bd*, *lmdfa* și *csk* produce AD cu performanțe bune.

Din punctul de vedere al găsirii metodei de reducere optime, analiza valorilor obținute din experimentele prezentate în teză arată că AD redus cu metoda de reducere cu nivel de încredere obține sistematic cele mai bune valori ale preciziei clasificării pe datele de test, simultan cu generarea unui număr mic de reguli de decizie. Pe locul doi se situează performanțele AD redus pesimist, iar pe ultimul loc AD neredus. Să observăm că această diferențiere, care plasează totdeauna AD redus cu nivel de încredere pe primul loc și AD redus pesimist pe locul doi, este mult mai clară în privința numărului de reguli de decizie, decât în privința acurateții clasificării pe datele de test. În același timp, să mai observăm că reducerea îmbunătățește performanțele AD.

Experimentele pe care dorim să le efectuăm în continuare vor viza baze mult mai mari de date, cu foarte multe atribute și foarte multe cazuri, pe care dorim să verificăm concluziile pe care le-am obținut pentru bazele de date utilizate până în prezent. În același timp, vom include în setul de măsuri pe baza cărora se realizează divizarea unui nod și alte criterii de selecție a atributului și vom utiliza și alte metode de reducere a AD, în afara celor două utilizate în experimentele noastre. În același timp va trebui să ținem cont și de informațiile furnizate de matricea de confuzie.