

BABES-BOLYAI UNIVERSITY
FACULTATEA DE MATEMATICĂ ȘI
INFORAMTICĂ

Bayesian Network

Authors:

Sergiu BREBAN Razvan SALAJAN

January 2, 2017

Contents

1	Data Mining	1
1.1	Informatia inseamna putere	1
1.2	Definitie	1
1.3	CISP	1
1.4	La ce se foloseste?	2
	Bibliography	4

List of Figures

1.1	CISP	2
-----	----------------	---

Chapter 1

Data Mining

1.1 Informatia inseamna putere

In ziua de azi, datorita serviciilor ieftine de stocare a datelor, stocarea de informatii devine o treaba usoara si deschide calea invatarii unor lucruri noi. Asadar obtinerea de noi cunostine se transforma in serviciul de extragere prin diferite metode de noi idei dintr-un ocean de informatii stocate care asteapta sa fie analizate. In zilele noastre majoritatea actiunilor cotidiene sunt inregistrate si stocate. Datorita numarului mare de oameni si de actiuni pe care un om le face se transmit date intr-un volum greu de imaginat. Avand disponibil un volum atat de imens de date partea mai grea devine analiza, invatarea si corelarea acestora. Desi lucrul cu date a fost folosit de mult timp de catre economisti, statisticieni sau meteorologi e de remarcat cresterea recenta al oportunitatilor de gasire al modelelor in date. Bazat pe [1] numarul datelor la nivel global stocat in bazele de date se dubleaza la fiecare 20 de luni. Avand in vedere volumul imens de date cu care suntem inundati si eficientizarea calculatoarelor de a suporta cautari are ca urmare cresterea interesului in data mining. Astfel, data mining, reprezinta o solutie la a gasi noi concepte, noi idei care pot ajuta la rezolvarea unor probleme din diferite ramuri, de exemplu, intr-un context comercial, obtinandu-se un avantaj comercial.

1.2 Definitie

Data mining reprezinta procesul de analiza si cercetare asupra unui volum imens de date stocat in baze de date sau alte tipuri de stocare de informatii. Procesul are ca scop descoperirea de noi cunostinte precum: modele, asocieri, comportamente, anomalii, structuri semnificative care sa ofere descoperiri noi, inovatoare. Procesul poate fi privit ca o cutie inchisa care produce rezultatul dorit. La fel cum si [2] precizeaza, procesul de data mining poate avea si rezultate nedorite daca o metoda este aplicata intr-un context nepotrivit sau modelele sunt construite pe presupuneri eronate.

1.3 CISP

"Cross-industry standard practice"(CISP) reprezinta un proces standard de abordare al solutiilor de tip data mining. Acest standard a fost creat cu scopul de a fi liber si de a fi utilizat de catre toti cei care folosesc data mining pentru a rezolva diferite probleme. A fost creat, cumva natural, datorita faptului ca fiecare incerca sa isi fac

propriul proces, astfel creandu-se haos. Asadar, CISP, a luat nastere ca fiind un standard care nu e specific unei industrii, aplicatii, sau unui tool. Bazat pe standardul CISP, un proiect de tip data mining consta din 6 etape dupa cum se poate vedea si din Figure 1.1. Procesul definit e unul iterativ care se poate adapta; fiecare compo-



FIGURE 1.1: CISP

nenta depinde de componenta precedenta iar in caz de iregularitati procesul se poate intoarce inapoi la pasul precedent.

1.4 La ce se foloseste?

Cerintele in care data mining e folosit s-au grupat in 6 clase (potrivit cu [2]).

- Descrierea: reprezinta modul prin care se cauta cai care precizeaza unele "tenduri" sau modele care se observa in date.
- Estimarea: se refera la a estima valoarea numerica a unui atribut tinta avand la dispozitie variabile predictorii care pot fi de tipul numerice sau categoriale(?). De exemplu se pot face estimari despre cat v-a cheltui o familie cu 2 copii pentru inceperea unui nou an scolar. De asemenea, un alt exemplu il reprezinta prezenta la vot pentru alegerile curente ale cetatenilor cu varsta cuprinsa intre [18-24] cu drept de vot bazat pe alegerile din anii trecuti.

- **Predictia:** este similara cu estimarea si clasificarea cu precizarea ca rezultatul predictiei o sa se intample in viitor. Exemple de predictii ar fi: pretul unei actiuni la bursa peste cateva luni sau cine va castiga campionatul mondial de fotbal. Bazat pe [2], in alegerile prezidentiale din America din 2012, echipa de campanie a lui Barack Obama s-a folosit de data mining. Acestia au folosit data mining pentru a identifica tipul sustinatorilor lui Barack Obama si s-au asigurat ca acestia merg la vot si au mai folosit pentru a face predictii despre voturile pentru Barack Obama in fiecare judet. Un exemplu de predictie pentru judetul Hamilton din statul Ohio considerat "indecis", a fost de 56.4% iar rezulultul final a fost de 56.6%.
- **Clasificarea:** este asemantoare cu estimarea cu diferenta ca variabila tinta, care se doreste sa fie clasificata, are ca valori categorii. Modelul de data mining examineaza datele care contini valori pentru attributele predictorii si atributul tinta.
- **Gruparea:** se refera la procesul de a grupa date, observatii, cazuri in clase asemantoare. O grupare reprezinta o colectie de obiecte care prezinta insusiri asemantoare intre ele. Gruparea reprezinta o sarcina diferita conceptual fata de clasificare, estimare sau predictia. In procesul de grupare nu exista un atribut tinta care sa fie clasificat. In schimb, gruparea, incearca sa imparta datele in grupuri sau clase relativ omogene, astfel incat similaritatiile elementelor dintr-un grup sunt maximizate iar disimilaritatille cu elemente straine grupului sunt minimizate.
- **Asocierea:** consta in a determina ce attribute "merg impreuna". Are ca scop gasirea de legaturi intre attribute. Este folosita cu succes in analiza pietei cumparatorilor din magazine pentru a gasi ce produse sunt cumparate impreuna si ce produce nu sunt. Asocierile sunt de forma "daca antecedentul atunci consecinta".

Bibliography

- [1] Mark A. Hall Ian H. Witten Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann; 3 edition (January 20, 2011), 2013.
- [2] Daniel T. Larose. *DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining*. John Wiley & Sons, 2014.