Vështrim i përgjithshëm mbi Data Mining, lidhja e saj me teknologjitë Machine Learning, DBMS, statistikat

Esmeralda HOXHA[1]
Llambrini SOTA

Abstrakt: *Të dhënat dhe informacionet luajnë një rol të rëndësishëm në aktivitetet e njeriut. Data Mining (DM) është procesi i nxjerrjes së njohurive, duke analizuar vëllime të mëdha të dhënash nga perspektiva të ndryshme dhe duke e përmbledhur atë në informacione të dobishme. Për shkak të rëndësisë që ka nxjerrja e informacionit nga magazina të mëdha të dhënash, DM është bërë një komponent thelbësor në fusha të ndryshme të jetës njerëzore. Avancimet në Statistika, Learning Machine, Inteligjencën artificiale etj. dhe zbatimet e DM në ditët e sotme kanë evoluar dhe këto zbatime kanë pasuruar fusha të ndryshme të jetës njerëzore, duke përfshirë biznesin, arsimin, mjekësinë, shkencën etj. Ky dokument diskuton për Data Mining; çfarë është DM; teknologjitë DM; qëllimet për përmirësime të ndryshme në fushën e DM nga e kaluara në të tashmen dhe shqyrton prirjet e ardhshme. Ky artikull hedh një vështrim të përgjithshëm dhe mbi DBMS dhe-statistikat.*

Fjalë kyç: ***Data Mining, KDD, Learning Meachine, DBMS, Statistikat, Black-box***

Abstract: *Information and Data play an important role in human activities. Data Mining (DM) is the process of extracting knoëledge by analyzing large volumes of data from different points of view and summarizes it into useful information. Due to the importance of extracting information from large data warehouses, DM has become an essential component in various fields of human life. The advances in Statistics, Machine Learning, Artificial Intelligence etc... and DM applications nowadays have evolved and these applications have enriched various fields of human life including business, education, medicine, science etc. This paper*

[1] Universiteti "Pavarësia", Vlorë, E-mail: esmeralda.hoxha@unipavaresia.edu.al

discusses about Data Mining, what is DM, DM technologies, goals, various improvements in the field of DM from the past to the present and examines future trend. This article takes an overview on the DBMS and statistics.

Keywords*: Data Mining, KDD, Learning Machine, DBMS, Statistics, Black -box*

1. Hyrje

Data Mining (Heikki, Mannila, 1996) është duke depërtuar në të gjitha fushat e jetës njerëzore, si p.sh biznes, industri, edukim, mjekësi dhe shkencë. Ardhja e teknologjisë së informacionit në fusha të ndryshme të jetës njerëzore ka çuar në magazinimin e vëllimeve të mëdha të të dhënave në formate të ndryshme, si në formë dokumentesh, imazhesh, regjistrime zanore, video, të dhënash shkencore, dhe formate të reja të dhënash. Të dhënat e mbledhura nga aplikime të ndryshme kërkojnë mekanizma të duhur të nxjerrjes së njohurive/ e informacionit nga magazinat e mëdha (ku ato janë grumbulluar) për marrjen e vendimeve më të mira. Zbulimi i njohurive në bazat e të dhënave KDD-Knowledge Discovery in Databases (Zbulime Njohurish në Bazën e të dhënave), i quajtur shpesh Data Mining, synon zbulimin e informacionit të dobishëm nga koleksionet e mëdha të të dhënave. Funksionalitete kryesore të DM janë duke aplikuar metoda të ndryshme dhe algoritme për të zbuluar dhe nxjerr modelet e ruajtura të të dhënave. Fusha e DM kanë përparuar dhe janë paraqitur edhe në fusha të reja të jetës njerëzore me integrime të ndryshme dhe avancimet në fushën e statistikave, Bazën e të dhënave (Data base), Machine Learning (Zbulime Njohurish në Bazën e të dhënave), inteligjencën artificiale dhe kapacitetet e llogaritjes (fuqitë llogaritëse) etj.

Fushat e ndryshme të aplikimit DM janë: Shkencat Humane (LS), CRM (Menaxhimi i marrëdhënieve të klientëve (Customer Relationship Management), Aplikimet Web, Prodhim, Financë/Bankare, Kompjuteri/Rrjeti/Siguria, Monitorim /Survejim, Mbështetja në Mësimdhënie, Modelet klimatike, Astronomi etj.

Në aspektin statistikor, DM do të shihej si një analizues automatik kompjuterik i informacionit nga një grup i madh dhe kompleks të dhënash. Statistika ka një ndikim të madh në industri, biznes dhe shkencë. Ajo, gjithashtu, ofron mundësi të mëdha kërkimore për zhvillimet e reja metodologjike. Statisticienët duhet të kenë interes të madh në DM. Statistika potencialisht mund të ketë një ndikim të madh në DM.

Shumë i rëndësishëm është dhe zhvillimi i sistemeve që përdorin të dyja, si të dhënat, ashtu edhe modelet për t'i qasur një kuptim më të pasur dhe të plotë kompleksitetit të botës reale, si dhe të mbështesë vendimet e marra prej botës reale. Ky model dhe orientimi i të dhënave kërkon zgjerime të konsiderueshme të teknologjive të bazës së të dhënave, siç janë integrimi i të dhënave, optimizimi i query-it dhe përpunimi, bashkëpunuesit analitikë.

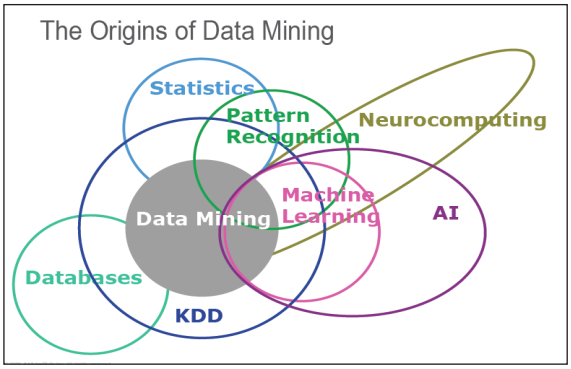
Çfarë është Data Mining?

Koncepti i DM është duke u bërë gjithnjë e më popullore si një mjet informativ i menaxhimit të biznesit, ku ai pritet të zbulojë strukturat e njohurive, që mund të udhëheqin vendime në kushtet e një sigurie të kufizuar. Kohët e fundit, është rritur interesi në zhvillimin e teknikave të reja analitike, projektuar posaçërisht për të adresuar çështjet relevante të DM të biznesit (p.sh.: Pemë e klasifikimit), por DM bazohet ende në parimet konceptuale e statistikore, përfshirë analizën tradicionale eksploruese të të dhënave, EDA. Analiza e të Dhënave Eksplorues (EDA) është një përqasje për analizimin e të dhënave, mbledhjen e karakteristikave të tyre kryesore, shpesh me metoda vizuale.

Megjithatë, një ndryshim i rëndësishëm ndërmjet DM dhe (EDA) është që DM është më shumë i orientuar drejt aplikimeve të natyrës së dukurive themelore (bazë). Me fjalë të tjera, Data Mining është relativisht më pak e interesuar për identifikimin e marrëdhënieve të veçanta, si ajo midis variabileve. Për shembull, zbulimin i natyrës së funksioneve themelore ose llojeve të veçanta të tyre, si dhe varësit interaktive midis llojeve te ndryshme të variabileve nuk janë qëllimi kryesor i DM. Në vend të kësaj, fokusi i DM është në gjetjen/prodhimin e një zgjidhje që mund të gjenerojë parashikime të dobishme. Prandaj, DM pranon, mes të tjerash, një "Black box" (Ky testim është një metodë e testimit te Software-it, që shqyrton funksionalitetin e një aplikimi, pa u marrë me strukturat apo punët e brendshme), për eksplorimin e të dhënave ose zbulimin e njohurive dhe përdor jo vetëm teknikat (EDA), por edhe teknika të tilla, si: rrjetet nervore, të cilat mund të gjenerojnë parashikimet e vlefshme, por nuk janë të aftë për të identifikuar natyrën specifike të ndërlidhjes mes variabileve, në të cilat janë bazuar parashikimet.

DM është konsideruar shpesh të jetë "një përzierje e statistikave, AI (inteligjencës artificiale), dhe Data base" (Pregibon, 1997, f. 8), e cila deri para pak kohësh nuk është njohur zakonisht si një fushë me interes për statisticienët.

FIGURA 1 ORIGJINA E DATA MINING



Data Mining është një proces analitik i projektuar për të shqyrtuar të dhënat (zakonisht sasi të mëdha të të dhënave - në mënyrë tipike të biznesit ose tregut - që njihen gjithashtu, si "të dhëna të mëdha") në kërkim të modeleve të qëndrueshme apo marrëdhënie sistematike ndërmjet variabileve dhe, pastaj, duke aplikuar modelet për të zbuluar nëngrupe të reja të dhënash. Qëllimi përfundimtar i DM është parashikimi – DM parashikues, është lloji më i zakonshëm i DM dhe ka kërkesat më të drejtpërdrejta të biznesit.

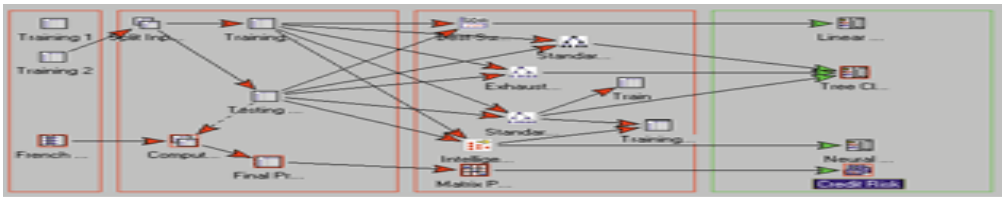
Procesi i DM përbëhet nga tri faza:

1. Eksplorim fillestar
2. Ndërtimi i modelit të identifikimit me miratim/verifikimin
3. Shpërndarja (d.m.th, zbatimi i modelit të të dhënave të reja në mënyrë që të gjenerojnë parashikimet).

Faza 1: Exploration. Kjo fazë, zakonisht, fillon me përgatitjen e të dhënave, që mund të përfshijë pastrim të të dhënave, transformime të dhënash, zgjedhjen e nëngrupeve të të dhënave - në rast se të dhënat përcaktohen me një numër të madh variabëlish ("fusha") - kryerjen e disa operacioneve paraprake. Kjo fazë është zgjedhja për të sjellë numrin e variabileve për një gamë të menaxhueshme (në varësi të metodave statistikore, të cilat janë duke u konsideruar).

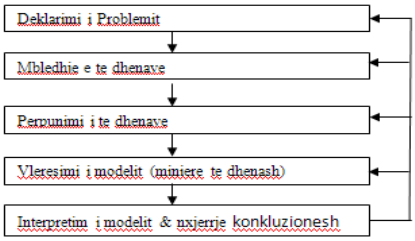
FIGURA 2. FAZAT NE PROCESIN E DATA MINING

Përvetësimi të dhënave	Përgatitja e të dhënave, pastrim, transformim	Analizimi i të dhënave, modelimi, klasifikimi, parashikimi	Raporte



Faza 2: Ndërtimi, Modeli dhe Miratimi. Kjo fazë ka parasysh modele të ndryshme dhe zgjedhjen e njërit prej më të mirëve, të bazuar në performancën e tyre parashikuese, (d.m.th., duke shpjeguar ndryshueshmërinë në fjalë dhe prodhimin e rezultateve të qëndrueshme të të gjithë mostrave). Kjo mund të tingëllojë si një operacion i thjeshtë, por në fakt, ai ndonjëherë përfshin një proces shumë të përpunuar. Ka një larmi të teknikave të zhvilluara për të arritur këtë qëllim - shumë prej të cilave janë bazuar në të ashtuquajturin "vlerësimin e modeleve konkurruese" d.m.th. duke aplikuar modele të ndryshme për të njëjtin grup të dhënash dhe pastaj, duke krahasuar performancën e tyre për të zgjedhur më të mirin. Këto teknika - të cilat janë konsideruar shpesh thelbi i DM parashikuese të dhënave - përfshijnë: Votimin, Rritjen, Stacking (Përgjithësime të stivosura), dhe Meta-Learning (Biggs,1985)

Faza 3: Shpërndarja. Në këtë fazë të fundit përdoret modeli i zgjedhur si më i mirë në fazën e mëparshme, duke e aplikuar atë për të dhëna të reja, në mënyrë që të gjenerohen parashikime lidhur me rezultatin e pritur.



Interpretimi i fazave të figurës së mësipërme është kryer më poshtë:

1. Deklarimi i problemit: Në këtë hap, një modelues zakonisht përcakton një sërë variabëlish, por pa njohur varësinë dhe formën e përgjithshme të kësaj varësie si hipoteza fillestare. Ky hap kërkon ekspertizë të kombinuar prej një fushë të aplikimit dhe një model datamining.
2. Mbledh të dhënat: Hapi i dytë është i përqendruar në atë, se si të dhënat janë të krijuara dhe grumbullimin apo mbledhjen.
3. Përpunimi i të dhënave (bazuar dhe në karakteristikat e tyre): Të dhënat zakonisht "të mblidhen" nga bazat e të dhënave ekzistuese, depo e të dhënave, dhe tregu i të dhënave. Përpunimi i të dhënave përfshin zakonisht, së paku, dy detyra të përbashkëta:
   a. Zbulimi i veçuar (dhe largimi) – Veçimi, është një karakteristikë e jashtëzakonshme e të dhënave, që nuk janë në përputhje me pjesën më të madhe të vëzhgimeve.
   b. Shkalla, shifrimi dhe përzgjedhje e karakteristikave.

4. Vlerësimi i modelit: Zgjedhja dhe zbatimi i teknikës së duhur të DM është detyra kryesore në këtë fazë. Ky proces nuk është i hapur, zakonisht, në praktikë, implementimi është i bazuar në modele të ndryshme, dhe zgjedhja e modelit më të mirë është një detyrë shtesë..
5. Interpretimi i modelit dhe nxjerrja e konkluzioneve: Në shumicën e rasteve, modelet DM duhet të ndihmojë në vendimmarrje. Prandaj, të tilla modele duhet të jenë të interpretueshme, në mënyrë që të jenë të dobishme, për shkak se njerëzit nuk mund të bazojnë vendimet e tyre në modele komplekse "black-box".

Zakonisht, modelet e thjeshta janë më të interpretueshme, por ata janë edhe më pak të sakta. Metodat moderne të DM pritet të japin rezultate shumë të sakta, duke përdorur modele me dimensione te larta. Problemi i interpretimit të këtyre modeleve, theksojmë, shumë i rëndësishëm, është konsideruar si një detyrë e veçantë, me teknika të veçanta për të vërtetuar rezultatet.

Koncepte Data Mining

Kompjuterët e sotme dhe mjetet përkatëse software, mbështetur në përpunimin e të dhënave, përcaktojnë me miliona mostra dhe qindra karakteristika. Grupet të mëdha të dhënash, duke përfshirë ato me lloje të dhënash të përziera, janë një mjedis tipik fillestar për zbatimin e të dhënave të teknikave DM. Kur një sasi e madhe të dhënash është ruajtur në një kompjuter, dikush nuk mund të nxitojë drejt e tek teknikat e DM, sepse, më parë duhet të zgjidhet problemi kryesor, i cilësisë së tyre. Gjithashtu, është e qartë se një analizë manuale e cilësisë nuk është e mundur në atë fazë. Prandaj, është e nevojshme për të përgatitur një analizë të cilësisë së të dhënave në fazat e hershme të procesit DM, zakonisht kjo është një detyrë që duhet të ndërmerret në fazën e përpunimit të të dhënave. Cilësia e të dhënave ka një efekt të thellë në imazhin e sistemit dhe përcakton modelin përkatës, që është përshkruar në mënyrë implicite, por, gjithashtu mund të kufizojë aftësinë e përdoruesit për të marrë vendime të informuara. Duke përdorur teknikat e DM, do të jetë e vështirë për të ndërmarrë ndryshime të mëdha cilësore në një organizatë, nëse të dhënat janë të një cilësie të dobët.

Ka një numër të treguesve të cilësisë së të dhënave:

1. Të dhënat duhet të jenë të sakta. Analisti duhet të kontrollojë se emri është shkruar drejt, kodi është në një gamë të caktuar, vlera është e plotësuar etj.
2. Të dhënat duhet të ruhen sipas tipit të të dhënave. Analisti duhet të sigurojë, se vlera numerike nuk është paraqitur në formën e karaktereve, se numrat e plotë nuk janë në formën e numrave reale etj.
3. Të dhënat duhet të kenë integritet. Updates nuk duhet humbur për shkak të konflikteve në mesin e përdoruesve të ndryshëm; backup-e të fuqishme dhe procedurat e shërimit duhet të zbatohet, në qoftë se ato nuk janë tashmë pjesë e Sistemit të Menaxhimit të Bazës së të dhënave (DBMS).
4. Të dhënat duhet të jenë të përputhshme (ose pajtueshme). Forma dhe përmbajtja duhet të jetë e njëjta pas integrimit të grupeve të mëdha të të dhënave nga burime të ndryshme.
5. Të dhënat nuk duhet të jenë me tepri. Në praktikë, të dhënat e tepërta duhet të minimizohen dhe të jenë të kontrollueshme. Rekordet e dublikuara duhet të eliminohen.
6. Të dhënat duhet të jenë më koherente. Komponenti kohë i të dhënave duhet të njihet në mënyrë të qartë nga të dhënat, apo të jetë i nënkuptuar nga mënyra e organizimit të tij.

7. Të dhënat duhet të jenë mirëkuptuara. Emërtimi i standardeve është i nevojshëm, por jo i vetmi kusht që të dhënat të kuptohen mirë. Përdoruesi duhet të dijë mirë, se të dhënat korrespondojnë me një domain (Një emër domain është një string identifikimi, që përcakton një fushë të autonomisë administrative, autoritet, ose kontrolli në internet) themelor.
8. Vendosja ose grumbullimi i të dhënave duhet të jetë i plotë. Humbja e të dhënave, siç ndodh në realitet, duhet të minimizohet. Humbja e të dhënave ul cilësinë e modeleve globale. Nga ana tjetër, disa teknika të DM janë të fuqishme mjaftueshëm për të mbështetur analizën e të dhënave me vlerat e zhdukura.

Çfarë realizon Data Mining për ne?

DM është përdorur për një shumëllojshmëri qëllimesh: si në sektorin private dhe atë publik, në industri të tilla, si: shërbimet bankare, sigurime, mjekësi dhe zakonisht një pakicë e përdorin DM për të zvogëluar shpenzimet, rritjen e kërkimeve dhe shitjeve. Për shembull, industritë e sigurimeve bankare përdorin aplikimet e DM për të zbuluar mashtrimin dhe për të ndihmuar në vlerësimin e rrezikut (p.sh., duke shënuar besueshmërinë). Nga përdorimi i të dhënave të konsumatorëve të mbledhura gjatë disa viteve, kompanitë mund të zhvillojnë modele që parashikojnë, nëse një klient gjykon, nëse është një kredi e mirë për të rriskuar, ose nëse një kërkesë mund të jetë mashtruese dhe duhet të hetohet më nga afër. Komuniteti mjekësor, mund të përdor DM për të ndihmuar të parashikojnë efektivitetin e një procedure apo ilaçi. Firmat farmaceutike përdorin DM për përbërësit kimik dhe materiale gjenetik, për të ndihmuar kërkime që udhëzojnë për trajtime të reja për sëmundjet. Shitësit me pakicë mund të përdorin informacionin e mbledhur nëpërmjet programeve të ngjashme (p.sh. kartat e blerësit të klubit, konkurse, etj.)

Data Mining:

- Identifikon perspektivat tuaja më të mira dhe më pas, i mban ose përdor ato si konsumatorë.
- Parashikon mundësit e shitjeve (cross-sell) dhe më pas bënë rekomandimet.
- Na ndihmon për të mësuar parametrat që ndikojnë në prirjet në shitje dhe kufijtë.
- Ndan në sektorë tregjet dhe personalizon komunikimet.

Kufizimet e DM

Edhe pse produktet DM mund të jenë mjete shumë të fuqishme, ato nuk janë aplikacione të vetëmjaftueshme. Për të qenë i suksesshëm, DM kërkon teknikë të kualifikuar dhe specialistë analitikë, që të mund të strukturojnë analizën dhe interpretojnë prodhimin që është nxjerrë (krijuar). Si rrjedhojë, kufizimet e DM janë kryesisht te të dhënat apo ndërvarësitë me to, sesa tek lidhja me teknologjinë.

Megjithëse DM-ja mund të ndihmojë në zbulimin e modeleve dhe marrëdhënieve, ajo nuk i tregon përdoruesit për vlerën ose rëndësinë e këtyre modeleve. Ky lloj vendimi ngelet të bëhet nga ana e përdoruesit. Në mënyrë të ngjashme, vlefshmëria e modeleve që janë zbuluar varet se si ato krahasohen me rrethanat e «botës reale». Për shembull, për të vlerësuar vlefshmërinë e DM-së, një aplikim i projektuar për të identifikuar të dyshuarit e mundshëm terrorist në një grumbull të madh individësh, përdoruesit mund të testojnë modelin e përdorur të të dhënave, që përfshin informacionin rreth terroristëve të njohur. Megjithatë, derisa të ri-afirmohet një profil i veçantë, nuk do të thotë se aplikimi do të identifikojë një të dyshuar, sjellja e të cilit devijon dukshëm nga modeli origjinal.

Një tjetër kufizim i DM është se, ndërsa ajo mund të identifikojë lidhjet ndërmjet sjelljeve dhe variabileve, nuk është e thënë të identifikojë një marrëdhënie rastësore. Për shembull, një aplikim mund të identifikojë modelin e një sjelljeje, si p.sh., prirja për të blerë bileta avioni vetëm pak para fluturimit, kur avioni është planifikuar të niset, kjo është e lidhur me karakteristika (arsye) të tilla, si: të ardhurat, niveli i edukimit, dhe përdorimi i internetit. Megjithatë, kjo nuk tregon se sjellja për blerjen e biletave është shkaktuar nga një ose më shumë prej këtyre variabileve. Në fakt, sjellja e individit mund të ndikohen nga disa arsye shtesë, si: përkushtimi (nevoja për të bërë udhëtime në një njoftim të shkurtër), statusi i familjes (një i afërm i sëmurë që ka nevojë për kujdes) etj.

Aplikimet e Data Mining

1. Kujdesi shëndetësor
   - Lufta kundër drogës: ka ndihmuar për të zbuluar trajtime më pak të shtrenjtë, por po aq efektive kundër drogës.
   - Mjekësi, Imazhe-diagnostifikim, monitorimi i tyre në kohë reale (p.sh., duke parashikuar gratë në rrezik të lartë...)
   - Identifikimin e klientëve të mundshëm për të blerë politika të reja, të përcaktojë modelet e sjelljes së klientëve të 'rrezikshëm'.
2. Biznesi dhe Financa
   - Bankat - për të zbuluar çfarë produktesh janë duke përdorur klientët, në mënyrë që ata të mund të ofrojnë përzierje të duhur të produkteve dhe shërbimeve për të përmbushur më mirë nevojat e konsumatorëve.
   - Kompanitë që përdorin kartë kredit, e-mail dhe materiale promovuese për njerëzit që kanë më shumë gjasa për t›u përgjigjur.
   - Huadhënësit, për të përcaktuar se cilët aplikantë kanë më shumë gjasa për t'u paracaktuar mbi një kredi.
3. Sport dhe bixhoz(lojëra fati)
   - Ekipet e Sportit – analizon të dhënat për të përcaktuar lojtarin më të favorshme në ndeshje dhe quan atë si lojtarin më të mirë.
   - Industrinë e Lojërave - analizon trendët e konsumatorëve kumarxhinj në kazino.
   - Tifozët e Sportit - parashikon se cilat ekipe do të zgjidhen për turneun, parashikon fituesit e lojës.
4. Arsim
   - Menaxhimi i regjistrimit - cilët prej studentëve do të marrin pjesë
   - Ruajtja / Diplomimi Analiza - cilët nxënës do të mbeten të regjistruar pas vitit të parë dhe / ose nëpërmjet diplomimit.
   - Parashikimi i Dhurimit- Donatorët dhe sa mund të dhurojnë ata.

2. Teknologjitë- Machine Learning, Dbms, Statistics. (Heikki, Mannila. 1996)

Që prej viteve 1980, teknika të bazuara në njohuri janë përdorur gjerësisht nga hulumtuesit e shkencës së informacionit. Këto teknika janë përpjekur të tërheqin kërkues, specialist të informacionit njohës të domain-ve, skema të klasifikimit të njohurive, strategji efektive kërkimi, query heuristikw (deduktiv, orientues, ndihmues), mundësitë e përpunimit të gjuhës

natyrale në rikuperimin e dokumentit, në dizajn sistemesh. Pavarësisht nga përdorueshmëria e tyre, sisteme të këtij lloji janë quajtur sisteme të performancës, pra, këto sisteme mund të performojnë vetëm për atë që janë programuar.

Machine Learning, teoria kompjuterike e të mësuarit, dhe të ngjashme si këto janë përdorur shpesh në kontekstin e DM, për të treguar zbatimin e algoritmeve të gjenerimit model-montim apo klasifikimin për DM të parashikuesve të të dhënave. Ndryshe nga analizat statistikore tradicionale të të dhënave, e cila zakonisht merret me vlerësimin e parametrave të popullsisë nga konkluzione statistikor, theksi në DM (dhe Machine Learning) është zakonisht në saktësinë e parashikimit, në klasifikimin e simbolit parashikues, pavarësisht, modeleve ose teknikat që janë përdorur për të gjeneruar parashikim, nëse është e interpretueshme apo e hapur për shpjegim të thjeshtë. Shembuj të mirë të llojeve të teknikave që aplikon DM, janë ato që përdoren në rrjete neutrale apo teknikat meta-learning, të tilla si: nxitja (simulimi) etj. Këto metoda, zakonisht përfshijnë montimin e shumë modeleve komplekse "gjenerike", që nuk janë të lidhura me ndonjë arsyetim apo teori të të kuptuarit të proceseve themelor.

Data Mining kombinon metoda dhe mjete nga të paktën tri fusha: Machine Learning, Statistika, dhe Bazat e të Dhënave. Ndonjëherë mund të dëgjojmë komentet e mëposhtme:

- DM është vetëm machine learning!
- DM është vetëm statistikë!
- Çfarë ka të bëjë DM me bazat e të dhënave?

Lidhjet e ngushta ndërmjet Machine Learning, statistikës dhe DM janë mjaft të dukshme. Të tre fusha synojnë gjetjen e rregullsisë interesante, modeleve, apo koncepteve nga të dhënat empirike. Metodat e Machine Learning (ML) formojnë thelbin e DM: pemët e vendimit ose rregulli i induksionit është një nga komponentët kryesorë të algoritmeve të DM.

Shumica e machine learning supozon se ka diçka për të mësuar. Në zbulimin e njohurive, nga ana tjetër, të dhënat janë gjëja primare, dhe ne nuk do të supozojmë se nuk do të ketë ndonjë strukturë të ndjeshme pas të dhënave.

- ML është aplikimi i algoritmeve kompjuterike që përmirësohen automatikisht nëpërmjet përvojës.
- ML mundëson analizë inteligjente të të dhënave, jashtëzakonisht të mëdha / magazinat e njohurive.
- ML mundëson kërkime të automatizuara për varësi shumë faktorësh kompleks në të dhëna. Makinat dhe sotware-et janë më të lira se 'njerëzit', procesi ML është i përsëritshëm, i përputhshëm dhe i fuqishëm në trajtimin e shumë detyrave në analizën e të dhënave.

Sistemet KDD zakonisht kanë synime mjaft modeste, në aspektin e kompleksitetit të njohurive të fituara. Në disa machine learning metodat e KDD mund të jenë të dobishëm edhe në mbledhjen e të dhënave të vogla. Për më tepër, burimi thelbësor i kompleksitetit në DM është zakonisht, jo numri i objekteve në bazën e të dhënave, por numri i atributeve: numri i modeleve të mundshme zakonisht rritet të paktën në mënyrë eksponenciale të numrit të atributeve.

Një term tjetër në modë është analiza eksploruese e të dhënave (EDA), e cila theksoi epërsinë e të dhënave si udhërrëfyese për procesin e analizës. KDD dhe EDA kanë qëllime dhe metoda shumë të ngjashme. Sipas perspektivës interesante statistikore mbi KDD, nga Plaku dhe Pregibon (J. Elder IV and D. Pregibon, p.83), fokusi i statistikave ka lëvizur gradualisht nga vlerësimi i modelit të zgjedhur. Në vend që të kërkoni për vlerat parametër që e bëjnë një model të të dhënave, edhe struktura model është pjesë e procesit të kërkimit. Përveç këtyre teknikave, komuniteti KDD ka shumë për të mësuar nga statistikat, p.sh., në trajtimin e pasigurisë.

Diferenca kryesore ndërmjet KDD dhe statistikës është ndoshta në përdorimin e gjerë të metodave të machine learning në KDD, në vëllimin e të dhënave, dhe në rolin e çështjeve kompjuterike komplekse në KDD. Për shembull, edhe metodat QTM & PKF kanë vështirësi në trajtimin e dhjetëra mijëra vlerave parametër; renditja e disa lloj preprocesorësh kombinator është e nevojshme për të bërë zgjedhjen e modelit për përpunimin e detyrës. Duket se kombinime të tilla të metodave mund të jenë të dobishme: teknika kombinatore janë përdorur për të krasitur hapësirën e kërkimit dhe metodat statistikore janë përdorur për të shqyrtuar pjesët e mbetura në detaje (Galil & Ukkonen,1995).

Duke përdorur Statistikat në Data Minierave

Data Minierave është një proces i centralizuar i të dhënave, kështu karakteristika e të dhënave përcakton se si duhet hartuar algoritmi. Ka gjithmonë probleme me të dhënat e botës reale, të cilat duhet të përballet data mining, ato mund të klasifikohen në pesë grupe:

- të dhëna ultra të mëdha
- të dhënat e zhurmshme (Noisy)
- të dhëna të paplota
- të dhëna të tepërta
- të dhëna dinamike

Teknikat e data driven, ose mbështeten në heuristics për të udhëzuar kërkimin e tyre nëpër hapësirën e madhe të marrëdhënieve të mundshme, ndërmjet kombinimeve të vlerave atribut, ose miratojnë një lloj metode reduktim - të dhënë për të bërë algoritëm më efikas.

Fokusi aktual i statistikave ka lëvizur gradualisht nga vlerësimi modelit te përzgjedhja e modelit. Pra, nuk fokusohemi vetëm tek kërkimi për vlerat e parametrit që e bëjnë një model t'i përshtat mirë të dhënat, por edhe te struktura model si pjesë e procesit të kërkimit. Ky trend i përshtat bukur qëllimet e Data Mining, kur njeriu nuk dëshiron të rregullojë paraprakisht strukturën e modelit. Në përparimet e fundit thuhet, se metodat Markov chain Monte Carlo (MCMC), bëjnë të mundur marrjen në konsideratë hapësira shumë më të mëdha të modelit sesa më parë.

Bazat e të dhënave dhe DBMS (Grup programesh që mundëson ruajtjen, modifikimin dhe nxjerrjen e informacionit nga një bazë të dhënash)

Në sistemet e menaxhimit të bazës së të dhënave, SQL ishin në dispozicion vetëm pas futjes së të dhënave krahasuese në vitet 1970. Nga administratorët e data bazave nuk mund të pritet që të monitorojnë ngarkesën e punës dhe të reagojnë me akordimet e duhura, prandaj automatizimi është thelbësor. Prandaj, këtu do të përdorim DBMS, i cili përdor machine learning për të modeluar dhe parashikuar ngarkesën e punës, si dhe hamendëson për të ardhmen.

DBMS-et shumë relacionale sigurojnë menu ose formojnë ndërfaqet bazë, të cilat nuk i kërkojnë përdoruesit të lëshojnë komanda të qarta si [SELECT]; [NGA]; [KU]. Me ndërfaqen është operuar, duke zgjedhur artikujt nga një menu ose plotësimit të artikujve në një formë, p.sh.,"query-nga-formë "(QBF). Duke e quajtur një formular në ekran dhe mbushjen në vlerat atribut të dëshiruar, një përdorues mund të përdori këtë formular për të formuluar pyetje të thjeshta në lidhje me tabela të veçanta. Të tilla menu drejtuar ndërfaqes kërkojnë trajnim të përdoruesve, por ata ende kanë nevojë të kenë një ide të qartë se çfarë ata duan. Kështu DBMS kanë hulumtuar dizajne të ndryshme për gjuhët natyrale query në DBMS.

Arkitektura e sistemit të bazës së të dhënave po pëson ndryshime revolucionare (Rationale and Architecture for a Compelling Application, 2005). Më e rëndësishmja, algoritmet dhe të dhënat janë duke u bashkuar me integrimin dhe gjuhët programimit me sistemin e bazës së të dhënave. Kjo i jep një sistem të zgjeruar objekt-relativ, ku operatorët relativ jo-procedurale manipulojnë grupe objektesh.

Duke u nisur nga kjo, çdo DBMS tani është një shërbim web. DBMS-të tani janë kontejnerët objekt. Radhët janë objektet e para që do shtohen. Sipas Mitchell (1982), këto radhë janë baza për përpunimin e transaksionit dhe rrjedhshmërisë të aplikim site-tit. Përtej kësaj, DBMS të ketë një kornizë për algoritme të DM dhe ML. Pemët e vendimit, rrjetat, platforma grumbullimi dhe analizat seri - kohë janë ndërtuar dhe shtuar në algoritmet e rinj. Algjebra e krahasuar (relative) është një mënyrë e përshtatshme për programimin e këtyre sistemeve. Sistemi i bazës së të dhënave pritet tani të jetë vetë-drejtues, vetë-shëruese, dhe gjithmonë lartë(up). Tabela e mëposhtme tregon teknikat dhe fushat aktuale te DM, për formate të ndryshme.

TABELA. 1 TEKNIKAT DHE FUSHAT AKTUALE TE DM

Tipat e Data Mining	Fushat e Aplikimit	Formati i të dhënave	Teknikat Data Mining/ Algorimat
Hipermedia data mining	Aplikacionet Internet dhe intranet	Të dhëna Hyper Text	Teknika të klasifikimit dhe grupimit
Data mining i kudogjendur	Aplikime të telefonave Mobile, PDA, Camera digitale	Të dhëna të kudogjendura	Teknikat tradicionale të data mining të dizenjuara prej Statistikave dhe Machine Learning
Data mining multimediale	Aplikacionet Audio/Video	Të dhëna multimediale	Rregulla të bazuara në Algoritme pemë vendimi
Data mining hapësinor	Rrjeti, Aplikimet GIS dhe sensorët në distancë	Të dhëna hapësinore	Teknikat hapësinore të grupimit, OALP hapësinore.
Data mining seri kohore	Aplikime financiare dhe biznesi	Të dhëna seri kohore	Algoritme me rregulla induksioni

DM Hyper text/Hypermedia: HyperText dhe Hypermedia e të dhënave është një përmbledhje e të dhënave nga Katalogët on-line, biblioteka digjitale, dhe bazat on-line të të dhënave informative, të cilat përfshijnë hyperlinks tekst, markup-s dhe format e tjera të të dhënave.

DM Multimedial: Të dhënat multimediale përfshijnë imazhe, video, audio dhe animimet.

DM Hapësinor: Të dhënat hapësinore përfshijnë të dhënat astronomike, të dhënat satelitore dhe të dhëna artizanale hapësinore.

DM Seritë-Kohore: Një seri kohore është një sekuencë e pikave të të dhënave, e matur në mënyrë tipike në raste të njëpasnjëshme, e shpërndarë në intervale kohore uniforme. Shembuj tipike përfshijnë çmimet e aksioneve, normat e këmbimit valutor, vëllimi i shitjeve të produktit, të matjeve bio-mjekësore, të dhënat mbi motin etj., të mbledhura gjatë kohës(rritëse monotone).

Roli i statistikës si një lojtar në atë që e quajmë "informacion relativ ose krahasues", në mënyrë të vazhdueshme do të zvogëlohet me kalimin e kohës (Elder & Pregibon, p.83).

Data Mining = Analizë Statistikore

DM mori një emër të keq fillimisht, sepse ajo u pa si "pastrim statistikor" ose një "ekspeditë peshkimi". DM u bë një praktikë e pranueshme, sepse përdoruesit e saj ushtruan rigorozisht statistikën në analizat e tyre.

"DM është procesi i përzgjedhjes, duke eksploruar dhe modeluar sasi të madhe të të dhënave për të zbuluar modele, të panjohura më parë, të të dhënave për avantazhin e biznesit." (Sas Instituti Inc).

"DM thjesht do të thotë të gjesh modele në të dhënat tuaja të biznesit, të cilat ju mund t'i përdorni për të bërë biznesin tuaj më të mirë"(SPSS Inc).

"DM është përdorimi i analizës statistikore dhe teknikave machine learning në një mënyrë gjysmë automatike, në koleksionet e mëdha të të dhënave". (Jorgensen & Gentleman, 1998).

Qëllimet Data Mining

DM është një mënyrë inovative për të fituar njohuri të reja dhe të vlefshme të biznesit, duke analizuar informacionin e mbajtur në data bazën e kompanisë tuaj. Këto njohuri mund t'ju mundësojnë për të identifikuar vendet më të mira të tregut, dhe kështu mbështesin dhe lehtësojnë marrjen e vendimeve të mirë informuara në biznese. Në thelb, DM është një mënyrë për të lëvruar informacionin që kompania juaj tashmë ka, në mënyrë që të planifikojnë një strategji të biznesit për të ardhmen.

DM futet dhe zbulon në thellësi të biznesit, duke përdorur analitikë të avancuar dhe teknikat e modelimit. Me DM, ju mund të bëni pyetje (query) shumë më të sofistikuara për të dhënat tuaja, sesa mund t'i bënit me metodat query-in konvencionale. Informacioni që DM ofron mund të çojë në një përmirësim të madh në cilësi dhe siguri të vendimmarrjes në biznes.

Vetëm DM i mundëson një banke krijimin e profileve të konsumatorëve, të cilët tashmë kanë një lloj profili llogarie bankare. Banka pastaj mund të përdorni DM për të gjetur konsumatorë të tjerë, që përputhen me këtë profil, në mënyrë që ajo të mund të synojë saktë një fushatë marketingu për ta.

DM mund të identifikojë modelet në të dhënat e kompanisë, për shembull, në regjistrat e blerjes në supermarket. Nëse, për shembull, konsumatorët blejnë një produkt, lindin pyetje cilin produkt do të blejë A, B apo C? Produkti C mund të ketë gjasa që të blihet më shumë? Përgjigjet e sakta të këtyre pyetjeve dhe pyetjeve të tjera si këto, janë ndihmë e paçmuar për strategjitë e marketingut.

DM mund të identifikojë karakteristikat e një grupi të njohur të klientëve, për shembull, ata që kanë një biografi të dëshmuar si rrezik të kreditit të varfër. Kompania pastaj mund të përdori këto karakteristika te klientët e rinj për të parashikuar, nëse edhe ata do të jenë rreziqet e kreditit të varfër.

- Parashikimi: DM mund të tregojnë se si atribute të caktuara brenda të dhënave do të sillen në të ardhmen. Në aplikime të tilla, logjika e biznesit është përdorur së bashku me të dhënat e minierave.
- Identifikimi: modelet e të dhënave mund të përdoren për të identifikuar ekzistencën e një artikulli, një ngjarje, ose një aktivitet. Për shembull, ndërhyrës, duke u përpjekur për të thyer një sistem, të mund të identifikohen nga programet e ekzekutuara, fotografitë që arrihen dhe koha e CPU për seancë.
- Klasifikimi: DM të ndarjes së të dhënave në mënyrë që, klasa të ndryshme apo kategori mund të identifikohen në bazë të kombinimit të parametrave.

- Optimizimi: Një qëllimi eventual DM mund të jetë për të optimizuar përdorimin e burimeve të kufizuara të tilla, si: të kohës, hapësirës, parave, ose materialeve dhe për të maksimalizuar variabile të prodhimit, të tilla, si: shitja apo fitimet sipas një grup të caktuar të kufizimeve.

Konkluzione

DM po del si një nga karakteristikat kryesore të sigurisë së shumë iniciativave. Shpesh përdoret si një mjet për zbulimin e mashtrimeve, vlerësimin e rrezikut dhe të produktit të pakicës. DM përfshin përdorimin e analizës së të dhënave në mjete të zbuluar, të panjohura më parë, në modelet e vlefshme dhe marrëdhëniet në grupe të dhënave të mëdha.

Në kontekstin e sigurisë së atdheut, DM është parë shpesh si një potencial për të identifikuar aktivitete terroriste, transfertat e parave dhe aktivitete të komunikimit, si dhe për të identifikuar dhe ndjekur vetë terroristët individuale, si nëpërmjet udhëtimit, ashtu edhe të dhënave të emigracionit.

Ndërsa të dhënat e DM përfaqësojnë një përparim të rëndësishëm në llojin e mjeteve analitike aktualisht në dispozicion, ka kufizime në aftësinë e saj. Një kufizim është, se edhe pse të dhënat e DM mund të ndihmojë të zbulojnë modelet dhe marrëdhëniet, ajo nuk i thotë përdoruesit vlerën apo rëndësinë e këtyre modeleve. Këto lloj vendimesh duhet të bëhen nga ana e përdoruesit. Një kufizim i dytë është se, ndërsa të dhënat e minierave mund të identifikojnë lidhjet mes sjelljeve dhe/ose variabileve, ajo nuk do të identifikojë një marrëdhënie shkakësore. Për të qenë i suksesshëm, DM ende kërkon specialistë të aftë teknikë dhe analitikë, të cilët mund të analizojë strukturën dhe interpretimin e prodhimit që është krijuar.

Të dhënat e minierave është duke u bërë gjithnjë e më e zakonshme në të dy sektorët private dhe publike. Industri të tilla, si: shërbimet bankare, sigurimeve, mjekësi dhe pakicë zakonisht përdorin të dhënat e minierave për të zvogëluar shpenzimet, rritjen e kërkimeve dhe rritjen e shitjeve. Dy përpjekjet që kanë tërhequr një nivel të lartë të interesit të kongresit përfshijnë, ndërgjegjësimin e informacionit në lidhje me terrorizmin (TIA) të projektit (Stop Now) dhe e ndihmuar nga kompjuterët pasagjer para kontrollit të Sistemit II (Capps II) e projektit Cancel Now (që zëvendësohet nga Safe Flight).

Një çështje është e cilësisë së të dhënave, që i referohet saktësisë dhe plotësisë së të dhënave të analizuara. Një çështje e dytë është ndërveprimit të softuerit të dhënave të minierave dhe bazat e të dhënave, duke u përdorur nga agjenci të ndryshme. Një çështje e tretë është zvarritje e misionit, apo përdorimi i të dhënave, për qëllime të tjera nga ato për të cilat, të dhënat janë mbledhur fillimisht. Një çështje e katërt është 'vetësia'. Pyetjet që mund të konsiderohen të përfshijnë shkallën në të cilën agjencitë qeveritare duhet të përdorin dhe përzierjen e të dhënave tregtare me të dhënat e qeverisë, nëse të dhënat e burimeve janë duke u përdorur për qëllime tjera, nga ato për të cilat fillimisht ishin dizenjuar dhe zbatimin e mundshëm të 'Aktit të Fshehtësisë' në këto nisma.

DM dhe zbulimi i njohurive janë aktualisht mjaft popullore. Ne përmbledhim disa nga drejtimet e lidhura të bazës së të dhënave kërkimore.

1. Zhvillimi i gjuhëve query me teknikat për të optimizuar pyetje model.
2. Përfaqësime që konsiderohen për klasa të ndryshme të modeleve.
3. Caching(Teknika efektive për të përmirësuar performancën e file-ve të sistemeve) strategji për përpunimin pyetje të lidhura mirë (fortë).
4. Kombinimet e DM dhe teknikave statistikore.

5. Duke përdorur fond njohurish (p.sh., metadata) (Biggs, 1985) në procesin e KDD.
6. Mjetet për zgjedhjen, grupim, dhe shfaqur e zbuluar njohuri.

Bibliografia

Biggs, J. B. (1985, Nëntor). The role of meta-learning in study process. British Journal of Educational Psychology, 55(3), 185-212.

Fayyad, U. M., Weir, N., & Djorgovski, S. (1993). Automated cataloging and analysis of sky survey image databases: the SKICAT system. CIKM '93 Proceedings of the second international conference on Information and knowledge management (pp. 527-536). New York: ACM.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37-54.

Galil, Z., & Ukkonen, E. (1995). Combinatorial Pattern Matching. 6th Annual Symposium, CPM 95,. Espoo: Springer-Verlag Berlin Heidelberg.

John F. Elder, I., & Pregibon, D. (1995). A Statistical Perspective on KDD. KDD-95 Proceeding (pp. 87-93). AAAI.

Krishnaswamy, S., Seng Wai, L., Rakotonirainy, A., Horovitz, O., & Mohamed Medhat, G. (2005). Towards situation-awareness and ubiquitous data mining for road safety: Rationale and architecture for a compelling application. Intelligent Vehicles and Road Infrastructure, (pp. 16-17). Melbourne.

Madras, N. (2002). Lectures on Monte Carlo Methods. Toronto: American Mathematical Society.

Mannila, H. (1996). Data mining: machine learning, statistics, and databases. Proceedings of 8th International Conference on Scientific and Statistical Data Base Management (pp. 2-9). Stockholm: IEEE.

Mitchell, T. (1982). Generalization as Search. Artificial Intelligence, 18(2), 127-134.

Piatetsky-Shapiro, G. (1999, Dhjetor). The Data-Mining Industry Coming of Age. Intelligent Systems and their Applications, IEEE, 14(6), 32-34.

Quinlan, J. R. (1993). Programs for machine learning. San Francisco: Morgan Kaufmann Publishers Inc.

Raedt, L. D., & Dehaspe, L. (1997). Clausal discovery. Machine Learning, 26(2), 99-146.

Raedt, L. D., & Džeroski, S. (1994, Tetor 31). First-order jk-clausal theories are PAC-learnable. Artificial Intelligence, 70(1), 375-392.