Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 9, September 2016, pg.67 – 71

Enhancing Clustering Mechanism by Customised Expectation– Maximization Algorithm: A Review

Jyoti
M.Tech Student
Department of Computer Science & Applications
Maharishi Dayanand University,
jyoti@nandal.in

Dr. Rajender Singh Chhillar
Professor & Former Head
Department of Computer Science & Applications
Maharishi Dayanand University,
Chhillar02@gmail.com

Abstract: Big data[1] is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, data curtain, search, sharing, storage, transfer, visualization, querying & information privacy. term often refers simply to use of predictive analytics or certain other advanced methods to extract value from data, & seldom to a particular size of data set. Accuracy in big data might lead to more confident decision making, & better decisions could result in greater operational efficiency, cost reduction & reduced risk. Data mining[7] is central step in a process called knowledge discovery in databases, namely step in which modeling techniques are include. Research areas like artificial intelligence, machine learning, & soft computing have contributed to its arsenal of methods. In our opinion fuzzy approaches could play an important role in data mining, because they given comprehensible results (although this goal is maybe because this is sometimes hard to achieve with other methods). In addition, approaches studied in data mining have mainly been oriented at highly structured & precise data. However, we expect that analysis of more complex heterogeneous information source like texts, images, rule bases etc. would become more important in near future. Therefore we give an outlook on information mining, we have see as an extension of data mining to treat difficult data in heterogeneous information sources, & argue that fuzzy systems are useful in meeting challenges of information mining. **Soft computing** is use of inexact solutions to computationally hard tasks such as solution of NP-complete problems, there is no known algorithm that could compute an exact solution in polynomial time. process of knowledge in databases, often also called **data mining**[9], is first important step in knowledge management technology. End users of these tools & systems are at all levels of management operative workers & managers. & these are their demands on processing & analysis of data & information that affect development of these tools.

Keywords—Data mining, web mining, web intelligence, knowledge discovery, fuzzy logic, K-mean

[1] Introduction

Analysis of data sets could find new correlations to "spot business trends, prevent diseases, combat crime & so on." Scientists, business executives, practitioners of medicine, advertising & governments alike common meet are difficulties with large data sets in areas including web search, finance & business informatics. Scientists encounter limitations in including meteorology, genomics, connectomics, difficult physics simulations, biology & environmental research. Data sets are growing rapidly in part because they are increasingly gathered by less & numerous information sensing mobile devices, aerial software logs, microphones, radio-frequency identification (RFID) readers & wireless sensor networks.[5][6] world's technological capacity to store more information has roughly doubled every 40 months since 1980s; as of 2012, every day 2.5 Exabyte's of data are created.[8] Relational database management systems & desktop statistics packages often have difficulty handling big data. work instead requires parallel software running on tens, even thousands of servers". What is considered "big data" varies depending on capabilities of users & their tools, & expanding capabilities make big data a moving target. "For some business facing hundreds of gigabytes of data for first time might trigger a need to reconsider data management options. For others, this might take tens or hundreds of TB before data size becomes a important consideration. **Data mining** (the analysis step of "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is computational process of discovering system in large data sets involving methods at intersection of artificial intelligence, statistics, & database systems. overall goal of data mining is process of extract information from a data set & transform this into an understandable structure for further use. Aside from raw analysis step, this involves database & data management aspects, data pre-processing, model & inference considerations, interestingness metrics, complexity, post-processing of discovered structures, visualization, & online updating. term is a misnomer, because goal is extraction of patterns & knowledge from large amount of data, not extraction of data itself. It also is a buzzword & is frequently applied to any form of big range data or information processing (collection, extraction, analysis, & statistics) as well as computer decision support system, including artificial intelligence, machine, & business intelligence. Important book "Data mining: special machine learning tools & techniques with Java originally to be named just special machine learning, & term "data mining" was only added for marketing reasons. Often more general terms data analysis or when referring to actual methods, artificial intelligence & machine learning. The actual data mining work is automatic or semi-automatic analysis of big amount of data to extract before unknown interesting patterns such as groups of data records unusual records (anomaly detection) & dependencies. This generally involves using database approach such as spatial indices. These patterns could then be seen as a kind of summary of input data, & might be used in further analysis or, for example, in machine learning & predictive analytics

[2] Motivation & Problem statement

Web Intelligence[2] based Google Analytics is known as a service offered by Google that generates detailed about a website's traffic & traffic sources & measures conversions & sales. It's most widely used website statistics service. Basic service is free of charge & a premium version is available for a fee. Google Analytics might track visitors from all referrers, including search engines direct visits & referring sites. It also tracks email marketing, & digital collateral such as links within PDF documents. Regular article **of Google Analytics** Integrated with AdWords, users might now review online campaigns by landing page quality & conversions (goals). Goals might include sales, viewing a specific page, or downloading data. **Google Analytics** approach is to show high-level, dashboard -type data/information for casual user, & more in-depth data/information further into report set. Google Analytics analysis might identify poorly performing pages with techniques/technology such as funnel visualization, where visitors came how long they stayed & their geographical position. It also provides more including custom visitor segmentation. Google Analytics e-commerce reporting might track sales activity & performance. e-commerce reports show a site's transactions, revenue, & many other commerce-related metrics. **Dashboards** give you a summary of many reports on a page. Start within a dashboard with your most important performance indicators then create detailed dashboards for other special topics like search

engine optimization. Dashboards use to drop widgets for fast, easy customization. Challenging problem in Web Intelligence[4] is how to deal with uncertainty of information on wired & wireless Web. Adapting existing soft computing solutions, when appropriate for WI applications, must incorporate a robust notion of learning that would scale to Web, adapt to individual user requirements, & personalize interfaces. Ongoing efforts exist to integrate logic, artificial neural networks, probabilistic & statistical reasoning, fuzzy sets, rough sets, granular computing, genetic algorithm, & other methodologies in soft computing paradigm, to construct a hybrid approach/system for Web intelligence[5] Internet-level communication, infrastructure, & security protocols. Web is regarded as a computer- network system. WI techniques/technology for this level include Web data/information prefacing systems built upon Web surfing patterns to resolve issue of Web latency. intelligence of Web prefetching comes from an adaptive learning process based on observation & characterization of user surfing behaviour.2. Interface-level multimedia presentation standards. Web is regarded as an interface for human–Internet interaction. WI techniques/technology for this level are used to develop intelligent Web interfaces in which capabilities of adaptive cross-language processing, personalized multimedia representation & multimodal data/information processing are required. 3. Knowledge-level information processing & management tools. Web is regarded as a distributed data/knowledge base. We need to develop semantic markup languages to represent semantic contents of Web available in machine-understandable formats for agent-based autonomic computing, such as searching, aggregation, classification, filtering, managing, mining, & discovery on Web.

[3] Survey of earlier work

The use of data mining techniques in manufacturing began in 1990s & this has gradually progressed by receiving attention from production community. These techniques are now used in many different areas in manufacturing engineering to extract knowledge for use in predictive maintenance, fault detection, design, production, quality assurance, scheduling, & decision support systems. Data could be analyzed to identify hidden patterns in parameters that control manufacturing processes or to determine & improve quality of products. A major advantage of data mining is that required data for analysis could be collected during normal operations of manufacturing process being studied & this is therefore generally not necessary to introduce dedicated processes for data collection. Since importance of data mining in manufacturing has clearly increased over last 20 years, this is now appropriate to critically review its history & application. Data mining techniques becomes basic element of modern business. Although idea is not new, new technologies & implemented standards make a contribution to their growing popularity. Regarding to mining model usage SQL Server 2005 stands breakthrough in this area. Thanks to DMX language either programmers or database administrators are able to create Data Mining Systems in simple way. Although economical & business publications are very fruitful of data mining approaches, described problem is presented rather weak in international publications. Nethertheless some industrial appliances of data mining technology were considered in (Duebel, C., 2003). Industrial usage of data mining techniques opens new possibilities in decision making not only for top level management, but also for advisory or control systems. Several prediction, classification or even anomaly detection algorithms implementation might become lucrative tool for industrial process appropriate stages optimization, that combines diagnosis & control functions. reviewed literature shows that there is a rapid growth in application of data mining in industry & manufacturing. However, there is still slow adoption of this technology in some industries for several reasons including both difficulties in determining type of data mining function to be performed in any particular knowledge area & question of choice most appropriate data mining technique regarding to many possibilities. **Wadena Wójcik & Konrad Gromaszek (*Lublin University of Technology, Poland) introduced* "Data Mining Industrial Applications".** Data mining is blend of concepts & algorithms from machine learning, statistics, artificial intelligence, & data management. With emergence of data mining,

researchers & practitioners began applying this technology on data from different areas such as banking, finance, retail, marketing, insurance, fraud detection, science, engineering, etc., to discover any hidden relationships or patterns.

Jiawei Han & Jing Gao University of Illinois at Urbana-Champaign wrote paper on "Research Challenges for Data Mining in Science & Engineering"

With rapid development of computer & information technology in last several decades, an enormous amount of data in science & engineering has been & would continuously be generated in massive scale, either being stored in gigantic storage devices or °owing into & out of system in form of data streams. Moreover, such data has been made widely available, e.g., via Internet. Such tremendous amount of data, in order of tera- to peta-bytes, has fundamentally changed science & engineering, transforming many disciplines from data-poor to increasingly data-rich, & calling for new, data-intensive methods to conduct research in science & engineering. In this paper, they discuss research challenges in science & engineering, from data mining perspective,

[4]Tools & technology used

HARDWARE

1. CPU 1Ghz or more
2. HARDDISK (5GB Free space)
3. DVD ROM
4. MONITOR
5. KEYBOARD/MOUSE
6. **SOFTWARE** WINDOWS 7/8
7. MATLAB
8. DOT NET FRAMEWORK

Expectation maximization algorithm[7]

The EM algorithm is used to find maximum likelihood parameters of a element of data model in cases where equations cannot be solved directly. Typically these approach involve latent variables in addition to unknown parameters & known data observations. That is, either there are missing values among data, or model could be formulated more simply by assuming existence of additional unobserved data points. For example, a mixture model could be described more simply by assuming that each observed data point has been a consistent unobserved data point, or latent variable, specifying mixture component that each data point belongs to.

Properties [8]

Speaking of an expectation (E) step is a bit of a misnomer. What is calculated in first step are fixed, data-dependent parameters of function *Q*. Once parameters of*Q* are known, it is fully determined & is maximized in second (M) step of an EM algorithm. Although an EM iteration does increase observed data (i.e. marginal) likelihood function there is no guarantee that sequence converges to a maximum likelihood estimator. For multimodal distributions, this means that an EM algorithm might converge to a local maximum of observed data likelihood function, depending on starting values.

Application

EM[10] is frequently used for data clustering in machine learning & computer vision. In natural language processing, two well-known instances of algorithm are Baum-Welch algorithm & inside-outside algorithm for unsupervised induction of probabilistic free grammars. EM algorithm (and its faster variant Ordered subset expectation maximization) is also large range used in medical image reconstruction, especially in positron emission tomography & single photon released energy computed tomography. See below for other faster variants of EM.

Variants

A number of methods have been proposed to accelerate sometimes slow convergence of EM algorithm, such as those using conjugate gradient & modified Newton Raphson techniques. Additionally EM[7] could be used with constrained estimation techniques. Expectation conditional maximization replaces each M step with a sequence of conditional maximization (CM) steps in which each parameter θ_i is maximized each, conditionally on

other parameters remaining fixed. This idea is further extended in generalized expected standard maximization algorithm, in which one only seeks an increase in objective function *F* for both E step & M step under alternative description.[14] GEM is further developed in a distributed environment & shows promising results.

[5]Conclusion

The Internet of Things[3] concept arises from need to manage, automate, & explore all devices, instruments, & sensors in world. In order to make wise decisions both for people & for things in IoT, data mining technologies are open to all people with IoT technologies for decision making support & system optimization. Data mining involves discovering novel, interesting, & potentially useful patterns from data & applying algorithms to extraction of hidden information Due to increasing amount of data available online, World Wide Web has becoming one of most valuable resources for information retrievals & knowledge discoveries. Web mining technologies are right solutions for knowledge discovery on Web. knowledge extracted from Web could be used to raise performances for Web information retrievals, question answering, & Web based data warehousing. overall goal of data mining process is to extract information from a data set & transform this into an understandable structure. In data mining K-means clustering algorithm is one of efficient unsupervised learning algorithms to solve well-known clustering problems. disadvantage in k-means algorithm is that, accuracy & efficiency is varied with choice of initial clustering centers on choosing this randomly.

[6]References

1. *Hellerstein, Joe (9 November 2008).* "Parallel Programming in Age of Big Data". *Gigaom Blog.*
2. J. Liu, N. Zhong, Y. Y. Yao, Z. W. Ras, wisdom web: new challenges for web intelligence (WI), J. Intell. Inform. Sys.,20(1): 5–9, 2003.
3. Congiusta, A. Pugliese, D. Talia, & P. Trunfio, Designing GridServices for distributed knowledge discovery, Web Intel. Agent Sys, 1(2): 91–104, 2003.
4. J. A. Hendler & E. A. Feigenbaum, Knowledge is power: semantic web vision, in N. Zhong, et al. (eds.), Web Intelligence: Research & Development, LNAI 2198, Springer, 2001, 18–29.
5. N. Zhong & J. Liu (eds.), Intelligent Technologies for Information Analysis, New York: Springer, 2004.
6. Bouckaert, Remco R.; Frank, Eibe; Hall, Mark A.; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2010). "WEKA Experiences with a Java open-source project".
7. Journal of Machine Learning Research **11**: 2533–2541. original title, "Practical machine learning", was changed ... term "data mining" was [added] primarily for marketing reasons.
8. Mena, Jesús (2011). Machine Learning Forensics for Law Enforcement, Security, & Intelligence. Boca Raton, FL: CRC Press (Taylor & Francis Group). ISBN 978-1-4398-6069-4.
9. Piatetsky-Shapiro, Gregory; Parker, Gary (2011). "Lesson: Data Mining, & Knowledge Discovery: An Introduction". Introduction to Data Mining. KD Nuggets. Retrieved 30 August 2012.
10. Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, & Algorithms. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.
11. "Microsoft Academic Search: Top conferences in data mining". Microsoft Academic Search.
12. "Google Scholar: Top publications - Data Mining & Analysis". Google Scholar.
13. Proceedings, International Conferences on Knowledge Discovery & Data Mining, ACM, New York.
14. SIGKDD Explorations, ACM, New York.

