Artif Intell Rev
DOI 10.1007/s10462-016-9483-9

Applications of association rule mining in health informatics: a survey

Wasif Altaf[1] · **Muhammad Shahbaz**[2] · **Aziz Guergachi**[3]

Abstract Association rule mining is an effective data mining technique which has been used widely in health informatics research right from its introduction. Since health informatics has received a lot of attention from researchers in last decade, and it has developed various sub-domains, so it is interesting as well as essential to review state of the art health informatics research. As knowledge discovery researchers and practitioners have applied an array of data mining techniques for knowledge extraction from health data, so the application of association rule mining techniques to health informatics domain has been focused and studied in detail in this survey. Through critical analysis of applications of association rule mining literature for health informatics from 2005 to 2014, it has been explored that, instead of the more efficient alternative approaches, the Apriori algorithm is still a widely used frequent itemset generation technique for application of association rule mining for health informatics. Moreover, other limitations related to applications of association rule mining for health informatics have also been identified and recommendations have been made to mitigate those limitations. Furthermore, the algorithms and tools utilized for application of association rule mining have also been identified, conclusions have been drawn from the literature surveyed, and future research directions have been presented.

Keywords Association rule mining · Health informatics · Knowledge discovery · Intelligent systems and applications

✉ Wasif Altaf
s8waalta@stud.uni-saarland.de

[1] Department of Computer Science, Universität des Saarlandes, Saarbruecken, Germany

[2] Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, Pakistan

[3] Ted Rogers School of Management, Ryerson University, Toronto, ON, Canada

Published online: 09 May 2016

1 Introduction

The current era of human history, due to digital revolution, computerization and internet technologies, is known as information age. Although, the information age, has in fact been based on vast explosion of mostly unstructured data. Various approaches for processing this data, and putting meaning and effectively using it, have been explored by knowledge discovery (KD) and data mining (DM) researchers over the years. As Russell and Norvig (2009) stated that, continued developments in statistics, machine learning and pattern recognition led to advent of DM technology. Defined by Han et al. (2011), DM is known as the process of discovering interesting patterns and knowledge from large amounts of data.

In data science research community, it is a well-known fact that, DM is an interdisciplinary field which has vast practical applications across domains such as Environmental Science, Security and Cryptography, Manufacturing, Transportation, Robotics and Intelligent Agents, Business Intelligence (BI), Healthcare Industry, Market Analysis, Energy Industry, Software Engineering (SE), Search Engines, Natural Language Processing (NLP), Distributed Systems (DS), Speech Processing and Image Processing to name a few. It may superficially be impossible to count down and holistically enlist the application domains of DM technology. Due to DM's vast applicability and usability, it has been researched well and matured to meet the needs of the hour.

Need is the mother of invention—and one of the main reasons behind DM's need is the tremendous amount of heterogeneous and complex data generated by real-world systems. Since, there may exist, useful and relevant (interesting) patterns in data, as well as useless or irrelevant (uninteresting) patterns in data, so only interesting, useful and meaningful patterns in data may be focused, and the uninteresting patterns maybe discarded.

Typically, DM may be implemented for classification, clustering, and association rule mining (ARM) purposes. Classification is a supervised machine learning technique aimed at predicting class labels of test dataset instances based on class labels and knowledge extracted from training dataset. Opposed to classification, unsupervised machine learning techniques organize similar instances in groups (clusters), where clusters have no assigned labels. Clustering approaches intend to maximize inter-cluster dissimilarity and intra-cluster similarity. Whereas, ARM techniques are used to model dependencies between data items contained in the datasets. In last decade, all these three basic types of DM techniques have been well researched and successfully applied in practical systems related to various aforementioned and other domains. However, in this survey we have focused on application of ARM in Health Informatics (HI) domain only.

Even though, HI domain is in its very nature different from manufacturing domain, but it is interesting that the ARM techniques used for various manufacturing processes, quality and maintenance improvement tasks as used by Köksal et al. (2011), Kamsu-Foguem et al. (2013), Ruiz et al. (2014) can be applied to HI domain too. The cause lies in the similarity of both domains' prime aims, which are, maintenance and improvement. Similarly, ARM has also been applied in transportation domain by Mirabadi and Sharifian (2010), Maquee et al. (2012) for maintenance and accident data analysis purposes aiming at exploration of the causes which lead to problems. The same concept can be used in HI domain for exploring causes (symptoms) which lead to certain problems (diseases or conditions).

Moreover, (Iqbal et al. 2016) surveyed state of the art intelligent agents' applications in health care domain, in similar direction of sequential pattern mining using ARM (Fournier-Viger et al. 2012) has been applied in intelligent agents domain by Nkambou et al. (2011), Faghihi et al. (2012) for guiding learners in problem-solving situations, and better understanding of learners' mistakes by exploring the causes of the learner's mistakes. Such type of sequential pattern mining using ARM can be applied in HI for deeper understanding of the causes behind subjects' health related issues and for better guidance of the subjects for solving their health issues. In addition to these, (Srinivasan and Ramakrishnan 2011) applied evolutionary multi objective optimization for association and classification rules mining to medical, biology, physics, chemistry, finance, space research, electricity datasets etc. While (Badrinath et al. 2016) utilized hybrid algorithms using ARM and AdaBoost for mining various diseases' datasets. To sum up, ARM and hybrid techniques applied in various other domains can also be applied in HI domain to achieve various goals.

As health information systems (HISs) generate large amounts of complex and heterogeneous data, so this makes the case for application of ARM for exploratory analysis of the data for generation of novel, interesting and hidden patterns. In addition to the health data being generated by HISs, health data may also be available from other sources such as World Wide Web (WWW), surveys and medical equipment (such as electrocardiogram machines). ARM can be applied directly to structured data (such as a database relation), or unstructured data (such as text or images) which are first converted into structured data and then ARM is applied.

In this survey, ARM basics and interestingness measures for rules have been discussed briefly in Sect. 2. Following that, HI domain has been introduced in Sect. 3. The survey of literature focusing the applications of ARM for HI has been presented in detail in Sect. 4. Moreover, the limitations of ARM tools and techniques for HI have been identified in Sect. 5 along with respective recommendations for mitigation of those limitations. Conclusions from the survey have been drawn and future directions for research have been presented in Sect. 6. Finally, references have been provided at the end for the cited research work.

2 Association rule mining

Recurring relationships in data may be explored through frequent pattern mining, and the concept can be described in simplified way through the traditional market basket example, where it is a common consumer behavior to buy {***bread***, ***butter***} or {***milk***, ***bread***}. Discovery of implicit patterns as such is sought through application of frequent pattern mining techniques. However, there may exist associations in data such that presence or absence of some specific item may relate to (or result in) presence or absence of some other item. For example, a patient who suffers ***diabetes*** is somehow likely to suffer ***hypertension*** as well. Exploration of such associations is aimed through association rule mining (ARM), where implications such as ***diabetes*** ⇒ ***hypertension*** are discovered. More formally, association rules (ARs) are of the form such as given in Eqs. 1 and 2.

$$diabieties \Rightarrow hypertension\,[support = 3\,\%, confidence = 75\,\%]\,. \tag{1}$$

$$smoking \Rightarrow \neg exercise\,[support = 5\,\%, confidence = 85\,\%] \tag{2}$$

Support and Confidence are used to measure interestingness of ARs, where f AR given in Eq. 1, support of 3% means that ***diabetes*** and ***hypertension*** occurred together in 3 % of all the transactions contained in the database. While confidence of 75 % means that, the patients who suffers ***diabetes***, 75 % of the times, suffer ***hypertension*** as well. This can also be interpreted from expectedness point of view that, of all the patients who suffer diabetes, it can be said with 75 % confidence that they will suffer hypertension as well. Hever, for the association rule given in Eq. 2, support of 5 % means that, of all the transactions contained in the database there are 5 % transactions in which the occurrence of smoking and absence of exercise has been encountered together. While, 85 % confidence value means that, of all the patients who smoke, 85 % of them do not take exercise. From expectedness point of view, it can be said that, for a patient who smokes, with 85 % confidence it can be said that he or she does not take exercise. For further details related to ARM approaches, (Li et al. 2008; Fürnkranz and Kliegr 2015) may be seen.

2.1 Interestingness measures of association rules

By definition, ARs which satisfy minimum support and minimum confidence thresholds are known as strong ARs. Although it has been shown by Han et al. (2011) that the interestingness of ARs is a subjective matter, and strong association rules may not be interesting in some situations, or can be misleading ARs.

Since the number of ARs extracted through mining process may be very large, so practically, the interestingness analysis of association rules through a manual subjective process is impossible. Hence objective interestingness measures are used for modeling subjective interestingness of ARs. Han et al. (2011) presented a comparative analysis of lift, χ^2, all_confidence, max_confidence, Kulczynski, and cosine measures for evaluation of interestingness of association rules. Where, it was concluded by the researchers that, for interestingness analysis, the Kulczynski measure may be used in conjunction with imbalance ratio measure. However, according to (Ohsaki et al. 2007) the accuracy, relative risk, uncovered negative, peculiarity, and chi-square measure for one quadrant are the interestingness measures used by medical experts for modelling their subjective interest in ARs related to medical domain.

Additionally, according to (Bouker et al. 2012, 2013, 2014) AR selection from the extracted ARs can be performed effectively without favoring one measure or another, or without caring about heterogeneity of the measures or e threshold values for the utilized measures. We now present the formulae for commonly used lift interestingness measure in Eq. 2.

$$lift\,(A \Rightarrow B) = \frac{P\,(A \cup B)}{P\,(A)\,P\,(B)} \tag{3}$$

3 Health informatics

In this section, we introduce Health Informatics (HI) which is the defining science of health information technology (IT). Coira (2003) defined health informatics in multiple ways by focusing information and communication technologies (ICT) involved in the health IT and expressed HI as “the logic of healthcare”. However, in recent context, as expressed by Savel and Foldy (2012), HI defines the science behind health IT. Although practically, HI is a young emerging field it still has developed very quickly, often overlapping, subdomains such as clinical informatics, nursing informatics, dental informatics, consumer health informatics, laboratory informatics, bioinformatics, pediatric informatics, anesthesia informatics,ental health informatics, infectious disease informatics, cancer informatics, behavioral healthcare informatics, trauma informatics, patient informatics, public health informatics etc. This expanse underscores the role and opportunities of health informatics in healthcare industry.

The practical commonly known examples of HI can be hospital information systems, digital medical equipment (such as X-ray imaging systems, 3D digital CT (X-ray computed tomography) scanners, medical resonance imaging (MRI) scanners, digital blood pressure monitors, digital glucose meters and jet nebulizers). Advanced applications of HI can be intelligent diagnosis systems, secure internal and external document sharing, intelligent cohort discovery tools, intelligent patient monitoring systems, and intelligent clinical decision support systems to name a few.

Since, through survey and critical analysis, research trends, research problems, limitations and future opportunities can be identified. So, it is of prime importance to survey current state of the art for HI systems, tools and techniques as it is a rapidly growing and briskly developing field, which has direct impact on living beings specifically humans. It is also noteworthy that, HI has, due to its growth, been subdivided into various domains aforementioned, due to which it is practically impossible to conduct a holistically comprehensive survey of HI in general. Due to this limitation, the literature surveyed in this paper has been focused to ARM for HI only. Moreover, the literature which we have considered important or significant has been focused in the survey.

4 Applications of association rule mining in health informatics

In this section, we present the detailed survey of applications of ARM in HI domain. The research articles reviewed have been discussed in ascending chronological order, where the year-wise frequencies of the research articles reviewed for each year have been presented in Fig. 1.

We have included research articles regarding applications of ARM in HI domain for over an extended period of time i.e. from 2005 to 2014. Only interesting and important literature has been focused and included in the study, as it is impractical as well as impossible to include every relevant article in the study. Moreover, for each of the articles included, we have focused our discussion on presentation of summarized article theme, introduction to dataset(s), outline of the proposed algorithm or approach, tools and technologies, experimental settings, results, conclusion and future work. The length of the discussion for each article varies based on our own perceived uniqueness/interestingness of that particular article.

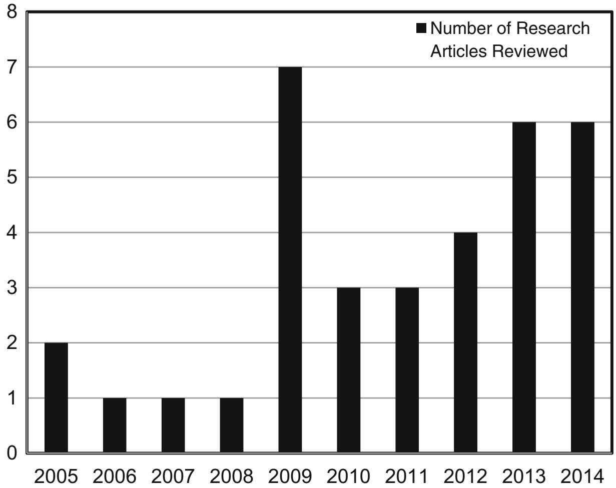


Fig. 1 Year-wise frequencies of research articles reviewed from 2005–2014 having more focus on recent years' significant publications

In ***analysis between lifestyle, family medical history and medical abnormalities using data mining method—association rule analysis***, Ogasawara et al. (2005) utilized ARM to study the relationships between lifestyles, family medical histories and medical abnormalities. Six lifestyle variables namely overweight, drinking, smoking, meals, physical exercise and sleeping time were considered for analysis. While medical history attributes namely hypertension, diabetes, cardiovascular disease, cerebrovascular disease, and liver disease were taken into account. Also, six medical abnormalities namely high blood pressure, hyperchoresterolemia, hypertrigriceridemia, high blood sugar, hyperuricemia, and liver dysfunction were used. Medical examination data of 7 years was collected for 5350 male employees of a company. The age group 40–49 was aimed for study, because in Japan, the adult diseases or lifestyle diseases are primary causes of deaths, and they develop and progress from 40 years and onwards.

Initially, the data for 5350 employees was collected in Microsoft Access database tables. Then, the integrated data was imported into Microsoft Excel sheet for analysis. The integrated data was association rule mined and logistics regression model (LRM) was also applied to compare the performance of both techniques. It was found by researchers that, association rule mining technique proved to be more useful in finding effective combinations of risk factors in terms of lifestyle diseases, than logistic regression model. A total of 4371 rules were extracted by using the proposed ARM technique. The authors concluded that, based on the study it can be argued that, AR analysis can play significantly important role in explaining combinations of risk factors related to lifestyle diseases than conventional modeling using LRM.

In ***mining interesting association rules in medical images***, Pan et al. (2005) have presented ARM algorithm for mining interesting association rules from medical images. They digitized 618 CT Scan images for utilization as dataset. They used a three-step approach to mine CT scan images. In step one, the dataset was preprocessed and through progressive water immersion method, regions of interest (ROI) were found. Regions of interest (Objects) were then converted into tables by combining them with location, size and other descriptors. Following this, in step two, frequent itemsets were generated and association rules (AR) were found. Authors defined image based support, object based support and confidence measures with respect to ARM of images. Sufficient frequent itemsets based on two supports in medical images (SFIMI) algorithm was proposed for frequent itemsets generation. Generating ARs from the sufficient itemsets (GAR) algorithm was proposed for generating association rules based on minimum confidence value. Lastly, in step three, post-processing was performed and uninteresting ARs were pruned out.

Results achieved through experimentation were analyzed based on domain knowledge, and it was found by researchers that, after pruning uninteresting rules, the extracted ARs were interesting and meaningful. Although, the results achieved by researchers were very basic in nature, but it was argued that the general framework proposed might be utilized in other domains too.

In ***constraining and summarizing association rules in medical data***, Ordonez et al. (2006) studied and researched the problem of the generation of large number of ARs from medical datasets. As, ARM algorithms may result in generation of extremely large number of ARs, especially when the minimum support and minimum confidence values are low, so a reduction or constraining technique may necessarily be required for effective analysis of the generated ARs. A dataset of 655 patients' heart disease problems, state and other details consisting of 113 attributes was used by the researchers. After preprocessing, out of the total 113 attributes, only 25 medical attributes were used for processing purposes.

ARs were constrained by usage of association rule size threshold, restriction of itemsets for appearance in premise or consequence, restriction of itemset combinations appearance in premise or consequence and lift measure usage. While association rules results were summarized by finding cover rules. The association rule size threshold meant that only the rules of size k or less would be kept for further processing, where k specifies the number of items in the AR. Restriction of itemsets for appearance in premise or consequence means that certain itemsets would be allowed appearance in premise or consequence only. Restriction of itemset combinations appearance in premise or consequence means that certain itemset combinations shall be allowed to appear in premise or consequence. While the lift measure was used by the researchers to restrict ARs to the rules with higher premise consequence dependence. Searching and replacement of cover rules for summarization means that the rules with same consequent are searched, and the premises for them are analyzed for hierarchical relationships. The subset premises are eliminated to keep the superset premises for the same consequent, hence resulting in summarized and lower number of ARs.

For experimentation purposes, the researchers used minimum support value of 1 %, minimum confidence value of 70 %, maximum rule size of 4, minimum lift value of 1.20, and minimum lift value of 2.0 for cover rules. Overall, the ARs with minimum confidence value of 0.90 were considered significant. Through experimentation it was found that the constraining and summarizing approach proved to be effective in significantly reducing the number of ARs generated and running time. Using high confidence and high lift values interesting ARs as well as rule covers were extracted for existence or absence of heart diseases. The authors concluded that their proposed framework could be applied in other domains, as well as for other medical datasets.

In ***evaluation of rule interestingness measures in medical knowledge discovery in databases***, Ohsaki et al. (2007) performed detailed analysis of practicability of rule interestingness measures in medical knowledge discovery in databases. As there exist numerous rule interestingness measures in literature and are used by researchers and practitioners, so practical usefulness analysis from a domain expert's point of view was required, so as to evaluate that specific interestingness measures made more sense to domain expert while reviewing an association rule. The researchers used two clinical datasets for experimentation purposes, one of hepatitis and the other of meningitis. The performance of rule interestingness measures was analyzed based on the interest shown by domain expert in the rule, and the results obtained through specific interestingness measure.

Through analysis, it was discovered that, out of total 40 interestingness measures analyzed, the accuracy, relative risk, uncovered negative, peculiarity, and chi-square measure for one quadrant are the interestingness measures useful from medical domain expert's point of view. Moreover, it was also discovered that medical knowledge discovery from databases could be advanced by utilization of other interestingness measures by medical domain experts. Researchers also aimed at evaluating the usefulness of user-interfaces (UIs) used for medical knowledge discovery using the interestingness measures.

In ***a text mining technique using association rule extraction***, Mahgoub et al. (2008) applied ARM to unstructured textual data for extraction of useful patterns related to Bird Flu disease. The extraction of association rules from text (EART) technique proposed by the researchers consisted of three main phases namely text preprocessing phase, ARM phase and visualization phase. We first discuss the dataset used, and then the application of EART technique to dataset.

The dataset utilized for testing and evaluation of technique contained 100 webpages collected from July 2006 to October 2006 from news sources such as Reuters, BBC, Yahoo, etc. The news stories were regarding geographical spread of bird flu virus and its effects on humans and birds. The corpus contained 30,000 single words and was of 440 KB in size.

The EART approach was based on three phases. Firstly, in text preprocessing phase, the system first converts dataset into XML format, then filtration is performed on stop words, then stemming is performed, followed by normalization step in which punctuation marks and special characters are replaced by space characters. Then as last text preprocessing step, indexing is done by using term document matrix through application of tf-idf weighting scheme. Secondly, in association rule mining phase, the system applies GARW algorithm (Generating Association Rules based on Weighting Scheme) where minimum support of 0.20, minimum confidence of 0.80 and weighting of 0.70 was used. GARW is a variant of Apriori algorithm, the only difference is that GARW uses weighting scheme too. Lastly, in visualization phase, the ARs extracted were presented in textual, tabular and graphical forms.

The EART system was developed in C# language, and experimentation was performed on Microsoft Windows XP Professional, on a 2.2 GHz Pentium 4 PC with 512 MB of RAM. Through experimentation, ARs were extracted and it was found by authors that the ARs generated didn't need domain knowledge for comprehension as they were easily understandable. The authors then compared the performance of EART with simple Apriori based system (which does not use weighting scheme), where the EART system was found to outperform Apriori when compared in terms of execution time taken by both systems on various minimum support values.

In ***mining healthcare data with temporal association rules: improvements and assessment for a practical use***, Concaro et al. (2009a) presented a novel approach for temporal association rule mining (TARM) of clinical and administrative healthcare data. Since clinical and administrative data complement each other in certain ways, so extraction of wealthy knowledge from the heterogeneous data was aimed. In particular, Regional Healthcare Agency (ASL) of Pavia, a part of Italian National Healthcare System, collected data of about 1300 diabetes patients through their General Practitioners (GPs) where patients pay visits from time to time. The data was collected from January 2007 to October 2008, and contained around 5000 inspections, the results from medical tests, and the information about current medical care, consisting of total 11 attributes for clinical data. However, the administrative data contained information about reimbursements.

Authors integrated clinical and administrative data, since administrative data was episodic in nature, so clinical data was preprocessed, state detections (high/low blood pressure), trend identifications (decreasing/normal/increasing glycaemia) and drug abstractions were formed so as to align the events according to the (episodic) nature of the administrative data. The temporal granularity was set to one day for healthcare episodes, which resulted in 110,000 healthcare episodes. Analysis was carried out through TARM over temporal sequences. The rule extraction was performed through a variant of Apriori algorithm, using support-confidence framework. Association rule templates were used to provide antecedent and consequent selection for the extracted rules, so that clinicians could manipulate rules. The ARM was performed at multiple levels, according to the levels defined by ATC coding system. The minimum support and minimum confidence were set to 1 percent and 30 percent respectively, while minimum improvement was used as interestingness measure to prune-out the uninteresting or meaningless rules as part of post-processing step.

The multilevel ARs extracted were examined by a clinician, and were marked for their meaningfulness or meaninglessness. It was found, through analysis that, in first case, the ARs extracted represent true clinical knowledge based relationships, so those ARs verified the current medical knowledge. Whereas in second case, where the ARs extracted didn't not make sense based on the current medical knowledge, those ARs could be the starting point of required further analysis.

In ***temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment***, Concaro et al. (2009b) applied TARM for analysis of costs related to diabetes mellitus. Using a process similar to (Concaro et al. 2009a, 2011) researchers integrated clinical and administrative data obtained from ASL of Pavia Italy, and then applied a similar approach to mine interesting rules using Apriori variant algorithm based on support-confidence framework. Authors redefined support and confidence measures to suit their ARM task. Through experimentation, interesting ARs were extracted for treatment expenses related to three age groups, i.e. 45–65, 65–75 and over 75 years age groups. It was found that, the age group 65–75 was the most expensive age group; moreover glycated hemoglobin was positively related to the most expensive profiles in all the age groups.

Additionally, authors also studied the total expected treatment costs' relevance to the value of glycated hemoglobin. It was found that, higher the glycated hemoglobin value, the higher were the total expected costs for all three age groups. It was also found that, the total expected costs for 65–75 year age group and over 75 years age group differed minimally for case with high or very high glycated hemoglobin values. For more background details of this work Concaro et al. (2009a, 2011) may as well be studied.

In ***mining administrative and clinical diabetes data with temporal association rules***, Concaro et al. (2009c)—a work partially related to (Concaro et al. 2009a, b), authors have presented application of TARM technique for analysis of care delivery flow of diabetes mellitus by extraction of temporal associations between diagnostics and therapeutic treatments. Authors used the dataset utilized by Concaro et al. (2009a). An Apriori variant approach was used for mining ARs, where support and confidence measures were redefined for temporal ARM.

In experimental settings, the support value of 1 percent and confidence value of 30 % was used. Authors showed through experimentation and the obtained results that, interesting ARs extracted could be used for evaluation of care delivery flow for specific diseases. And the methodology could be deployed to analyze or refine the clinical or administrative practices related to diabetes mellitus which result in unwanted or unsatisfactory results.

In ***supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques***, Ribeiro et al. (2009) applied ARM techniques for improving content-based image retrieval (CBIR) and computer-aided diagnosis (CAD). Two techniques were proposed by researchers, first, feature selection through association rules (FAR) was proposed for CBIR, and second, Image Diagnosis Enhancement through Association Rules (IDEA) was proposed for CAD. FAR was proposed to improve precision of CBIR system through continuous feature selection and weighting feature vectors using statistical association rules. While the researchers proposed the IDEA system for automatic suggestion of diagnosis for new images.

The FAR system used StARMiner algorithm, whereas IDEA system used Apriori algorithm based ARM. MRI dataset containing 704 images was used for evaluating FAR; the system was trained by using 176 images, and the testing was performed by using 528 MRI images. Through experimentation, it was concluded that FAR resulted in higher precision for CBIR than traditional techniques.

The IDEA system was evaluated using regions of interest (ROIs) dataset which contained 446 images, and a comparative analysis of IDEA system with Naïve Bayes, C4.5 and 1-Nearest Neighbor classifier showed that IDEA resulted in highest accuracy and sensitivity. It was also held by researchers that radiologists found IDEA system to be very helpful in diagnosis. Hence, the IDEA system could be deployed in real world. The researchers concluded that ARM techniques could be effectively employed in real world CBIR and CAD systems to increase radiologists' confidence in decision making.

In ***analysis of health care data using different data mining techniques***, Gosain and Kumar (2009) developed Antiretroviral Therapy (ART) system aimed at better management of human immunodeficiency virus (HIV) positive patients. They applied ARM using Microsoft Association Rules algorithm and DT using Microsoft Decision Trees algorithm on 672 HIV patients' data selected from HIV Database (HIVDB). Their methodology was based on CRISP-DM. They performed data cleaning, handled missing values, tackled noisy data issue, integrated data from various relations, and selected important or more relevant data attributes in data preparation step. For application of ARM and DT algorithms, data manipulation, data evaluation and data visualization, authors utilized Microsoft Business Intelligence and Microsoft SQL Server 2008. The attributes focused by them were limited to 9 only. In this regard, DT results obtained were very basic, and it was concluded, based on data analysis that, men were highly affected by HIV than women. Although, the conclusion that mostly illiterate or less educated people were facing HIV does not seem scientifically improvised, as the figures presented by authors illustrate that, only "Professional, Postgraduate & above" people were less affected by HIV, and patients with rest of the six levels of education had mixed levels of high HIV occurrences in their respective categories. Surprisingly, based on the data analysis, it is clear that people who do not belong to "Professional, Postgraduate & above" category, were more likely to be victimized by HIV, even graduates.

In ***association rules based data mining on test data of physical health standard***, Yu (2009) applied Microsoft Association Rules algorithm on public health standard (PHS) data of students to predict the main attribute(s) which lead to higher physical performance of students. Author utilized Microsoft SQL Server 2005 Integration Services to integrate Microsoft Excel based dataset (XLS file) in Microsoft SQL Server 2005 relation for application of Microsoft Association Rules algorithm. The dataset contained five main attributes, out of which "total score" attribute was to be predicted based on the remaining four attributes named as vital capacity, grip strength, standing long jump and step test. Through application of Microsoft Association Rules algorithm it was found that "standing long jump" was the attribute which has largest impact on the "total score" of physical health standard of a student.

In ***home health tele-monitoring system based on data mining***, Xianhai and Cunxi (2009) proposed a basic system for home healthcare monitoring using ARM. The system is based on client-server architecture, and has been implemented through socket programming using Java. The client terminal resides in patients' home, which uses various sensors for capturing patient's blood pressure, pulse and temperature and displays these measures on client's terminal as well it communicates the captured data to server. The server stores the data in data warehouse for further processing and data mining tasks. The server resides at hospital, where alarms maybe generated based on patient's condition, and the doctors can advise cures (medication and treatment) to patients through server, which then communicates these cures to client's terminal. Moreover, the ARs have been presented by researchers to illustrate that ARs can be applied on doctor's advice to patients regarding management of hypertension (as an example), though, the details regarding application of ARM and the results obtained have not been provided or discussed by authors. Also, the system architecture is very rudimentary, and the implementation details are also obscure. Due to which the contribution of research work may be considered insignificant and unclear.

In ***mining and post-processing of association rules in the atherosclerosis risk domain***, Berka and Rauch (2010) applied meta-learning techniques to association rules in the atherosclerosis risk domain. In other words, they applied ARM to the ARs extracted from atherosclerosis data to generate rules about rules for more effective comprehension of the

ARs generated in the first step. Authors used STULONG dataset[1] for ARM and subsequently meta-learning. STULONG contains data regarding 1400 male patients aimed at studying the risk factors involved in atherosclerosis. The authors created four relations namely ENTRY, CONTROL, LETTER and DEATH for transformation of STULONG dataset into digital form.

The researchers used WEKA's implementation of Apriori algorithm for application of ARM to the digitized STULONG dataset, and then the extracted ARs were post-processed by application of ARM to the rules extracted in prior step. For both ARM steps, the minimum support used was 10 %, while the minimum confidence was set to 70 %. Through experimentation, qualitative and quantitative meta-rules were extracted, and it was shown by authors that, in situations where large numbers of ARs are extracted and the analysis of those ARs becomes a problematic and tedious task, the meta-rules extraction can prove to be useful. Meta-rules were shown to reduce the number of the "original" ARs and ease in giving descriptive meanings to the original ARs. As future work, authors stated that they would further research the application of meta-learning to ARs with complex structure.

In ***hybrid medical image classification using association rule mining with decision tree algorithm***, Rajendran and Madheswaran (2010) proposed hybrid association rule classifier (HARC) based on ARM and decision tree (DT) algorithm. Authors used HARC for medical image mining on CT scan brain image data collected from suspected brain tumor patients. The problem statement was to correctly classify CT scan images as Normal, Benign or Malignant.

The authors used four step approach, which consisted of preprocessing, feature extraction, ARM and hybrid classification. In preprocessing step, the CT scan images' dataset was preprocessed using image processing techniques for normalization, noise reduction and completion. Then, in feature extraction step, the texture features were obtained from preprocessed image objects and were processed and stored in transactional database. While, in ARM step, the transactional database containing texture features' data was mined using Frequent Pattern Tree (FP-Tree) based ARM. Moreover, in hybrid classification step, DT approach was applied to transactional database instances for each of the rules generated in ARM step.

The HARC approach was evaluated by using training and testing phases. In both the training and the testing phases, preprocessing and feature extraction was performed. While features extracted were saved to transactional database in training phase only. Rest of the approach was similar to training and testing phase. The results obtained in testing phase, showed that, HARC outperformed ARM and DT based CT scan classifications, and achieved accuracy of 95 %, while sensitivity and specificity of 97 and 96 % were achieved, respectively.

In ***using associative classifiers for predictive analysis in health care data mining***, Soni and Vyas (2010) presented a theoretical review of associative classifiers for HI. Associative classifiers use association rule mining for classification purposes, hence the mechanism is similar to that of DTs on a higher level. Authors also briefly reviewed positive and negative associative classifiers, temporal associative classifiers, associative classifiers using fuzzy association rules and weighted associative classifiers for HI. Although, the work presented by researchers is a very basic review of literature on associative classifiers for HI, and no experimentation or proposed work has been presented.

In ***mining health care administrative data with temporal association rules on hybrid events***, a work relevant to Concaro et al. (2009a, 2011) proposed a technique for mining healthcare administrative data through TARM approach. Two datasets were used by

[1] EuroMISE Project STULONG, http://euromise.vse.cz/stulong-en/index.php. Accessed 5th December 2014.

researchers for experimental evaluation of the proposed approach, synthetic dataset and dataset of diabetic patients collected from Regional Healthcare Agency (ASL) of Pavia, Italy. Temporal granularity of one day was chosen for creating, managing and processing the episodes of care. Seven temporal operators namely BEFORE, MEETS, OVERLAPS, FINISHED-BY, EQUALS, STARTS and PRECEDES were used for positioning and associations of episodes of care. Episodes were essentially point-of-time based episodes or interval based episodes. The rule extraction was based on Apriori variant methodology, as the nature of data was different from the normal transactional data so the support and confidence measures were redefined to comply with the algorithmic requirements. Through post-processing, the uninteresting and meaningless rules were pruned out by using minimum improvement measure.

The proposed approach was compared (by authors) with interval based and point based approaches presented in literature, through experimental evaluation on synthetic dataset. It was found that the proposed approach outperformed both of the other techniques. Moreover, the real experimental evaluation of the approach was performed on Pavia ASL dataset, where diabetes was the focus of the study. The analysis was performed from two dimensions, first, to generate diabetes specific temporal ARs, second, to analyze the compliance of provision of healthcare services to medical recommendations. The temporal association rules mined (452 ARs) were analyzed by clinical expert and marked as clinically interesting, clinically irrelevant, or clinically unspecific. While, through temporal association rule analysis, it was also found that the compliance to medical recommendations was low for diabetes, and hence was questionable. As future work, authors proposed to improve upon the post-processing step by inclusion of ontologies, and by automatic validation of the ARs from online scientific literature.

In ***intelligent and effective heart disease prediction system using weighted associative classifiers***, Soni et al. (2011)—an extension of work done by Soni and Vyas (2010), authors experimented with weighted associative classifiers for heart disease prediction. They experimented with Cleveland Heart Disease Dataset[2] available online from UCI machine learning datasets, and used 303 records and 13 attributes for training and testing purposes, where attribute weights were assigned by domain expert. An intelligent heart disease prediction system (IHDPS) using weighted associative classifier was developed, where the system's GUI was based on Java, and Microsoft Access was used as a transactional database.

The IHDPS system's accuracy measure value was reported and compared with relevant systems, for three datasets namely, heart disease dataset, hepatitis dataset and cancer dataset. It was reported that IHDPS achieved the highest average accuracy as compared to other systems. Authors concluded that weighted associative classification is an effective approach for prediction of cardiac diseases, and the proposed system's performance was based on training data, since the 40 % instances were "No Heart Disease" so the system performed very well in correct classification of "No Heart Disease" cases when tested.

In ***a hierarchical model for association rule mining of sequential events: an approach to automated medical symptom prediction***, McCormick et al. (2011) presented hierarchical association rule mining (HARM) technique to predict sequential events. Authors applied the HARM technique for automated symptom prediction. The basic idea of the technique was to mine ARs as such ***symptom* 1 *and symptom* 2** $\Rightarrow$ ***symptom* 3**. where based on the history of a patient with symptom 1 and symptom 2, it can be predicted that the next symptom to be faced by patient might be symptom 3.

[2] Cleveland Heart Disease Dataset, https://archive.ics.uci.edu/ml/datasets/Heart+Disease. Accessed 5th December 2014.

The algorithm used for ARM didn't use constant support value; on the other hand, adjusted confidence measure was introduced for capturing those ARs for which the candidate itemsets generated could have low support and high confidence measures. This adjustment was aimed at capturing rare or infrequent itemsets. Authors used Gibbs sampling algorithm for online update of dynamic data.

HARM was applied to around 2300 patients' data, all older than 40 years, with a total of 42,000 encounters. The predictive performance of HARM was the highest with partially observed patients, than with new patients. This means that the symptom history on patient does have an effect in predicting next probable symptom. Through HARM, it was also found that myocardial infraction has, in addition to high cholesterol and hypertension issues, a relation with ethnicity too. It was also found that, age factor also plays a role in myocardial infraction, and people elder than 50 were more likely to be affected by myocardial infraction. As well it was found that a certain treatment (drugs) resulted in more chances of myocardial infraction, due to which specific drugs were ceased off the market.

In ***exploration of the association rules mining technique for the signal detection of adverse drug events in spontaneous reporting systems***, Wang et al. (2012) applied ARM for analysis of adverse drug events (ADEs). The datasets were generated from spontaneous reporting systems (SRS), for Shanghai, from January 2009 to December 2009. The datasets used for mining contained 24,297 reports after pre-processing and cleaning steps. The basic aim of the study was to extract association rules of the form ***drug* ⇒ *ADE***, where the premise drug was constrained to be of size of a single drug only, and the consequence would be the adverse drug event which could be caused by usage of that drug in some specific conditions.

Authors utilized SAS 9.1.3 for ARM and experimentation purposes. The datasets were first anonymized and personal and identification information was removed. Minimum support value was set to 3 percent, while minimum lift was selected to be 1.2 according to largest Youden's index. ARs containing multiple drugs as premise were pruned out, premise containing only single drug were used for analysis.

Researchers compared ARM results with those of various statistical approaches, and found that ARM performed better in identification of ADEs, but generated high false positives too. Still, the rules generated by ARM were backed by pharmacological literature,nd it was concluded by researchers that ARM could be used as a decision support system (DSS), and not as a decision making mechanism when dealing with ADEs.

In ***mining association rules from large datasets towards disease prediction***, Srinivas et al. (2012) presented positive and negative ARM for disease prediction. Authors have proposed defining association strength (DAST) algorithm, for mining frequent as well as rare itemsets. DAST was tested using synthetic dataset, while the experimentation was performed by utilizing a dataset collected directly from medical practitioners. The dataset used for experimentation contained around 1000 records and six attributes. Since, the experimentation dataset was very basic, so the results obtained were not novel and insignificant. Authors concluded that they would use DAST on real time data in future.

In ***knowledge discovery from mining association rules for heart disease prediction***, Jabbar et al. (2012) proposed heart attack prediction using boolean matrix (HAPBM) algorithm for cardiovascular disease (CVD) prediction using ARM. HAPBM is a variant of Apriori algorithm, the main difference being the conversion of discretized dataset into boolean matrix, and then frequent itemset generation from boolean matrix. For evaluation purposes, dataset was collected from various corporate hospitals of Andhra Pradesh, India.

Through experimentation, frequent itemsets were generated where maximum frequent itemsets size was 6. Authors presented CVD patterns for Andhra Pradesh, where results produced were empirical and novel. And through comparative analysis, it was also shown that HAPBM's execution time was comparatively better than that of Apriori algorithm.

In ***comparative study of association rule mining and MiSTIC in extracting spatio-temporal disease occurrences patterns***, Raheja and Rajan (2012) comparatively analyzed association rule mining and MiSTIC (Mining Spatio-Temporally Invariant Core Regions) approaches for extracting spatio-temporal disease occurrence patterns in Florida, California and New York. Salmonellosis disease was studied and the dataset used by authors contained numbers of people affected and rates of affected people out of 100,000 of the population, of the three cities. The dataset contained 50, 10 and 17 years of data for Florida, California and New York, respectively. Authors first preprocessed the dataset and then applied ARM by using Apriori algorithm, with minimum support and minimum confidence values of 1 and 40 percent respectively. It was found by researchers that the rules generated by this method were not very meaningful. Hence economy, demographics and environmental data was also collected and made part of the dataset and then ARM using Apriori was applied on the consolidated dataset. Through this method, even though a large number of rules were extracted, but according to authors, their meaningfulness was still very limited and dependent on domain knowledge or domain expertise.

Through application of MiSTIC approach, which essentially relied on finding cores with continues neighborhood and cores with a defined radius, was found to be simplistic and hence effective in mapping and analysis of spatio-temporal dimensions of the Salmonellosis etiology. It was found by researchers through analyses that Salmonellosis' rates were highly related to urbanization, the fact which was not revealed through ARM. Hence it was concluded by authors that spatio-temporal ARM algorithms need further exploration so that they can be used for more effective mining of data which contains spatial and temporal dimensions.

In ***discovering medical knowledge using association rule mining in young adults with acute myocardial infraction***, Lee et al. (2013) utilized ARM for discovering medical knowledge from acute myocardial infraction (AMI) patients. The authors chose 1247 young adults' data from Korean AMI registry (KAMIR) which contained 14,885 enrolled patients with 141 risk factors. They selected patients of age 45 years or younger, because most studies take into account the elder patients only, or elder and younger patients collectively. Moreover, 12 risk factors related to blood factors were selected out of the total 141 risk factors, to find associations between blood factors and disease history in young AMI patients. The approach used for ARM by researcher was a variant of Apriori algorithm, and standard support-confidence framework was used for frequent itemsets and ARs generation. However, the pruning step was performed by using lift, leverage, and conviction interestingness measures.

Through experimental evaluation, results comparative to various other state-of-the-art research-works were achieved. It was found that smoking, diabetes and hypertension were the significant risk factors involved in AMI in young adults. The results achieved were compared to relevant studies and it was shown that the ARM approach was useful for extraction of useful and meaningful associations in large dataset. And interesting associations found between blood factors and disease histories were a significant contribution of the study. Although, it can be argued that only 12 risk factors out of total 141 were taken into account, so the remaining 129 risk factors could also prove to be useful in mining implicit, interesting and novel ARs.

In ***mining medical data to identify frequent diseases using Apriori algorithm***, Ilayaraja and Meyyappan (2013) utilized ARM based on Apriori algorithm for identification of frequent diseases in particular geographical area for a particular time period. They used hospital information system (HIS) for collection of data that contained 1246 patient's medical health records. They implemented the proposed technique using WEKA. 1216 patient records containing 29 attributes for year 2012 were analyzed. Through application of Apriori algorithm frequent itemsets for diseases were generated. Authors concluded in their work that they also catered the geographical dimensions of frequently occurring diseases; on the other hand, no such thing was factually made part of the research article, due to which their work may be regarded as insignificant.

In ***extraction of positive and negative association rules from text: a temporal approach***, Mahmood et al. (2013) extracted positive and negative temporal association rules (TARs) from textual blogs. Authors selected medical blogs from different online sources and mined temporal associations along following three dimensions, i.e. seasonal/periodic associations, event-based associations, established associations. Seasonal associations are the associations between disease and symptoms which occur at a certain period of time in a year, over the years. Event-based associations depend on the occurrence of some specific event such as presidential elections, disaster, hurricane, famine or earthquake etc. Established associations are the ones which are equally likely to occur throughout all the seasons. Authors held that division of ARs into the above stated three temporal dimensions can greatly help medical practitioners in dealing with specific disease cases.

Authors presented a three step approach to mine TARs from medical blogs. In first step, dataset was partitioned into subsets for certain time periods. In second step, each subset was mined for association rules extraction. FP-Tree algorithm was used for generation of frequent and infrequent itemsets from data subsets. Then, positive and negative association rules (PNARs) were extracted from frequent and infrequent itemsets. In third step, positive and negative association rules generated in step two were analyzed from temporal interestingness point of view. Support, confidence and lift were used as interestingness measures, and ARs extracted were categorized into three categories namely seasonal/periodic association rules, event-based association rules and established association rules.

Results found from the application of three step algorithm were presented for each of the aforementioned three categories separately. The support and confidence thresholds for each category were also presented by authors. On the other hand, authors spared on the details of how they performed preprocessing of the medical textual blogs into a dataset viable for application of ARM algorithm presented.

In ***association rules of data mining application for respiratory illness by air pollution database***, Payus et al. (2013) mined air pollution database for extracting reasons behind respiratory illness, in Kuala Lumpur Malaysia. The dataset used was obtained from Malaysian Ministry of Health and the Department of Environment, Malaysia, and contained 7 attributes and 1000 instances.

The ARM process consisted of five step approach. In first step, data selection, seven attributes were selected for further preprocessing and analysis. In second step, preprocessing was performed, and the numerical data was discretized using equal frequency binning method. In third step, Apriori algorithm was applied on preprocessed data using WEKA 3.7 with minimum support and minimum confidence value of 0.1. In fourth step, evaluation was performed on the results obtained from step three. Totally 42 rules were generated, 17 rules out of these belonged to “normal” hospitalized patients, 24 belonged to “moderate” hospitalized patients and 1 rule belonged to “high” hospitalized patient. In step five, the knowledge extraction from the results and evaluation was done, and it was found that carbon monoxide (CO), temperature and particulate matter (PM10) were strongly correlated with the number of patients suffering from respiratory diseases.

It is however noteworthy that, the dataset utilized by authors did not contain temporal dimension, even though the dataset provided by Malaysian government agencies contained timestamps of the data instances. But the timestamps were discarded by authors as part of the data selection step.

In ***data mining applications in medical image mining: an analysis of breast cancer using weighted rule mining and classifiers***, Kavipriya and Gomathy (2013) proposed a technique for medical image mining for analysis of breast cancer using weighted association rule mining (WARM). They presented a four step approach, in first step, i.e. data preprocessing was performed to handle missing values and noisy data. In second step, approximation process has been performed to discretize the data based on expert's opinion. In third step, pruning-classification association rule (PCAR) has been used to prune out infrequent itemsets, and the number of candidate itemsets is greatly reduced. In fourth step, ARM has been applied to predict the disease severity.

Authors presented that, they tested their system using breast cancer diagnosis dataset provided by UCI machine learning repository, but they have not provided any results of the experimentation. Only abstract structure of the system was proposed. So, the work done was basic and theoretical only, and hence cannot be further evaluated for its novelty, or practical ability.

In ***using semantic-based association rule mining for improving clinical text retrieval***, Babashzadeh et al. (2013) have utilized semantics based ARM for improving clinical text retrieval. Their research was to take up TREC Medical Record challenge 2011 so they utilized TREC 2011 Medical Track Dataset[3] for experimental evaluation of the proposed approach. TREC Medical Record challenge 2011 dataset contained query set, electronic medical records of patients and relevant judgments. The electronic medical records contained 17,267 visits and 101,711 reports. MetaMap was used for concept location for the dataset, and the concepts identified were used for indexing. While query context modeling was done as a two-step process, where in first step, ARM was used to extract the rules which satisfied query concepts modeled by UMLS Methathesaurus. In second step, the extracted rules were weighted and ranked according to their semantic similarity to the query concepts.

Through experimentation, it was shown by the researchers that the semantics based ARM effectively enhanced the retrieval performance as compared to baseline systems in which terms based indexing, concepts based indexing, terms-concepts based indexing, query expansion using Rocchio algorithm and naïve methods were used for query context modeling.

In ***an association rule mining-based framework for understanding lifestyle risk behaviors***, Park et al. (2014) have studied inter-correlation between lifestyle risk factors for South Korea. They used 4th Korean national health and nutrition examination survey (FKNHANES) dataset of 14,833 young adults of ages more than 20 years. The FKNHANES data was collected from 2007 to 2009 and contained data for 5908 men and 8925 women. The lifestyle risk factors considered by researchers were frequent snacking, breakfast skipping, inadequate sleep, obesity, physical activity, heavy drinking and current smoking. Moreover, the gender, age, education, marital status and level of income were also used. ARM was used to extract interesting patterns from dataset, and then logistic regressing was used for predicting multiple lifestyle groups. SAS® Version 9.3 and SAS® Enterprise Miner® Version 4.3 were used by authors for ARM and construction of model figures, respectively. The minimum support value was set to be 2 percent, while the minimum confidence values of 50 and 60 % were used for women and men, respectively.

It was found that, the males who drink heavily, skip breakfast, are physically inactive, and obese were more likely to be current smokers as well. While for women, the physical inac-

[3] TREC 2011 Medical Track Dataset, http://trec.nist.gov/data/medical2011.html. Accessed 5th December 2014.

tivity, obesity and current smoking factors were associated with insufficient sleep. Authors concluded that the results of the study could be used to generate more effective interventional and awareness programs. For example, the non-smoking campaigns may focus on awareness of strenuous physical exercises and proper breakfasting to achieve better outcomes.

In ***analyzing lifestyle and environmental factors on semen fertility using association rule mining,*** Anwar and Ahmed (2014) utilized ARM for analysis of effects of lifestyle and environmental parameters on the semen fertility. Authors utilized the dataset from UCI machine learning repository, where the dataset contained data for 100 individuals of ages between 18 and 36 years. The dataset used by researchers contained attributes such as smoking habit, surgical treatment, number of hours sitting per day, season in which analysis was performed, age at the time of analysis, childish diseases, accident or serious trauma, high fever in the last year and the output diagnosis (normal or altered). First of all, the researchers preprocessed the data, and performed data cleaning and data normalization. Then, ARs were mined using Apriori algorithm based approach by using XLMiner tool.

The authors performed experimentation and analyze with multiple support and confidence values, and reported to have found interesting association rules, but they concluded that none of the parameters considered in the study have any effect on semen fertility abnormality for the 18–36 years age group. Authors also concluded that more variables may be included in the dataset to find useful results.

In ***association rule mining based clinical observations***, Rashid et al. (2014) have presented a basic study for analysis of clinical observations by using ARM. Authors have proposed a basic framework for analysis of healthcare enterprise data. Their framework is based on online transaction processing (OLTP) and clinical state correlation prediction (CSCP). CSCP gets data from OLTP and processes it for analysis of comorbidity using Apriori algorithm. As researchers used synthetic dataset for experimentation on prototype system, so practicability of the CSCP system can be argued. Moreover, as the CSCP system is based on Apriori algorithm so its performance or applicability for larger datasets can be criticized as well.

In ***significant patterns for oral cancer detection: association rule on clinical examination and history data***, Sharma and Om (2014) applied ARM on patients' clinical examination and history data for oral cancer detection. The dataset used was collected from various hospitals and registries for over 5 years between 2004 and 2009. Authors first performed data cleaning and integration as part of data preparation process. The consolidated dataset contained 1025 records having 33 attributes for each record. The attributes taken into account were tobacco smoking, tobacco chewing, ulcer, loosening of teeth, ulcer, burning sensation, hypertension, diabetes, size of mass etc. For experimentation purposes, researchers used Apriori algorithm provided by WEKA 3.7.9 tool. For experimentation settings, the minimum support value was set to 10 % and minimum confidence value was set to 90 %.

It was concluded based on experimentation that the diagnosis of oral cancer will be supported by results obtained. And the results obtained shall help addressing the real issue of latency of diagnosis due to which mortality rates are very high. Moreover, the authors also intend to apply the proposed approach on other datasets for extraction of valuable knowledge.

In ***mining surprising patterns and their explanations in clinical data***, Kuo et al. (2014) have proposed an approach for mining surprising and unexpected association rules from clinical data. Their approach has been aimed at provision of solution to the common problem of ARM process producing huge number of rules; often large in number hence their analysis could be very time and resource consuming, or even impossible. The approach proposed by authors reduces the number of association rules generated by considering the rules which are relevant to the domain knowledge (DK) of the domain expert using the system proposed and implemented. Hence, the system requires domain expert for the creation of knowledgebase (KB), and only the association rules relevant to the system's KB are extracted. Such rules have been considered as unexpected or surprising association rules.

Researchers used Australia and New Zealand Dialysis and Transplant (ANZDATA) dataset—which has been accumulated and funded by Australian and New Zealand governments since 1980. As of 2010, the number of attributes contained in dataset is 96, while for total 19,220 patients, the number of records contained in the dataset is 217,803. However, after the preprocessing process compromising of data selection, transformation and merging, the number of attributes used for mining was reduced to 39.

The interactive process for mining unexpected yet interesting association rules based on domain expert's DK used minimum support and minimum confidence values of 60 %. The process resulted in 6 unexpected association rules, which were then analyzed by domain expert. Upon detailed analysis by domain expert, 2 ARs were found to be surprising, while the remaining 4 ARs were considered as meaningless. Surprising and meaningless association rules extracted by researchers have been exemplified in Eqs. 3 and 4 respectively. Researchers also analyzed the reasons behind the surprising and meaningless association rules, and concluded that knowledge gaps may be the reason behind meaninglessness or surprisingness of the ARs, as the unexpected rules mined are related to facts contained in KB only.

$$\begin{aligned} lungb &= No \rightarrow txwlcat \\ &= NotonTransplantList[support = 65\,\%, confidence \\ &= 74\,\%, Unexpectedness = 7\%] \end{aligned} \tag{4}$$

$$\begin{aligned} lungb &= No \rightarrow dryweight \\ &= 53.2 - 101.5\,[support = 63\,\%, confidence \\ &\quad = 71\,\%, Unexpectedness = 9\,\%] \end{aligned} \tag{5}$$

The authors concluded that unexpectedness may be considered at all stages of the KDD process so that surprising knowledge may be extracted more effectively. Moreover, as the unexpectedness of association rules mined is dependent on the domain expert's knowledge and expertise, so the approaches proposed are dynamic in nature as it considers the unexpectedness relative to the DK. However, the DK may be different for different domain experts, and hence varying results might be achieved for different domain experts.

In ***negative and positive association rules mining from text using frequent and infrequent itemsets***, Mahmood et al. (2014) proposed an approach for mining positive and negative association rules from medical blogs using Apriori based frequent and infrequent itemset mining. Authors used a two-step approach for mining ARs, where in first step, textual blogs were preprocessed, inverse document frequency (IDF) based pruning was performed, and then frequent and infrequent itemsets were identified by using the proposed Apriori variant approach namely Apriori FISinFIS. In second step, interesting association rules were extracted from frequent and infrequent itemsets mined in preceding step.

The algorithm proposed was implemented using Java, and multiple support and confidence values were used as experimental settings. Through experimentation on 1926 medical blogs related to cancer, it was shown that the proposed approach proved to be effective and efficient in mining positive as well as negative association rules from unstructured text. Authors reported that, with decreasing support values, the number of frequent itemsets generated by Apriori FISinFIS also decreases, opposed to that, the number of generated infrequent

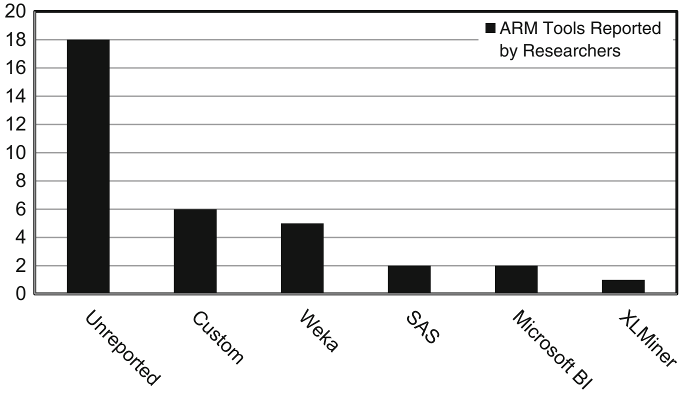


Fig. 2 Tools used for application of association rule mining for health informatics. Of all the researches reviewed in this survey, a significant number of researches did not report the implementation used, while many utilized customized implementations for ARM

itemsets increases. Moreover, it was concluded that, generation of positive association rules from infrequent itemsets can be very useful in mining interesting association rules from text.

4.1 Association rule mining tools used for health informatics

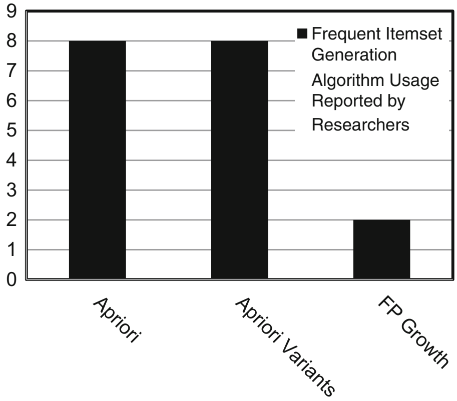
In Fig. 2, the tools used for application of ARM for health informatics are shown along with the respective number of researches using those tools. From all the literature reviewed, it has been observed that, most researchers did not report the tools used for ARM. Among the reported work WEKA has been found to be the most utilized tool for application of ARM for health informatics, which has been used more frequently than SAS, Microsoft Business Intelligence and XLMiner tools. However, customized system implementations using Java, C#, C or any other language tools, categorized as Custom Implementation, have also been used by researchers due to unavailability of more advanced ARM algorithms and techniques' implementations.

Since most of the researchers did not report the ARM tools used, it is still unclear from the survey that which specific tool is the most popular one. However, it is interesting to note that apart from wide availability of open source ARM software, many researchers still used custom implementations. The requirement of using custom implementations is rooted in unavailability of advanced ARM techniques' implementations e.g. AR post-processing techniques. On the other hand, commercial implementations such as SAS, Microsoft Business Intelligence suite and XLMiner are closed source and hence less flexible for customizations and analysis of the actual implementations—still requiring custom implementations of advanced ARM techniques.

4.2 Frequent itemset generation algorithms used for application of ARM for health informatics

Various algorithms used for frequent itemset generation for application of ARM for health informatics have been presented from the literature reviewed in Sect. 4 and are shown in Fig. 3. Although it has been proved by Han et al. (2000, 2011) that FP-Growth approach is more efficient than Apriori algorithm, it has still been found that Apriori algorithm and its variants have been used widely than FP-Growth approach. In order to increase performance,

Fig. 3 Frequent itemset generation algorithms used for application of ARM for health informatics. All the researches reviewed in this paper, Apriori and its variants have been utilized by most of the researchers, only few researchers have utilized FP Growth algorithm—which is highly efficient alternative for frequent and infrequent itemset mining from large datasets when compared to Apriori algorithm's performance



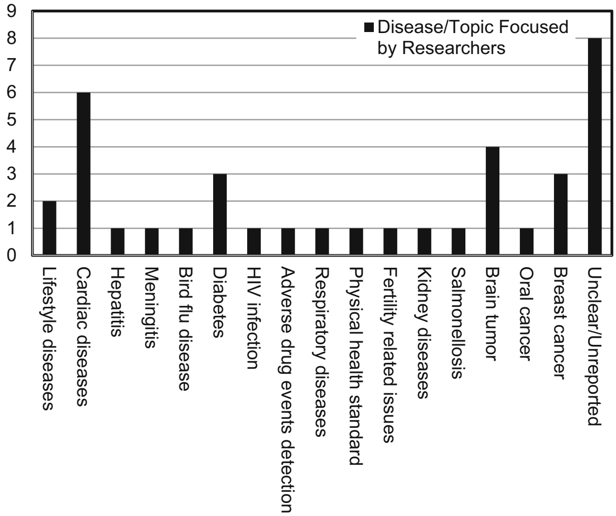


Fig. 4 Classification of research topics for application of ARM for health informatics. ARM has been extensively applied for knowledge generation related to cardiac diseases and their underlying causes, and also to various forms of cancers, both are major causes of deaths in humans

efficiency and scale, researchers and practitioners should utilize FP-Growth, parallel FP-Growth (Li et al. 2008) or more efficient alternatives instead of Apriori or its customizations.

Moreover, it has been observed from the literature surveyed in Sect. 4 that, none of the researches utilized distributed, cloud or intercloud computing for enhancement of scale and performance of frequent itemset or ARM. In real world applications such as HISs where higher volumes of data are required to be mined, simpler personal computing based approaches are likely to perform inefficiently, hence requiring utilization of cloud or intercloud computing resources or parallel processing based approaches.

4.3 Classification of research topics for application of ARM for health informatics

The literature reviewed has been classified with respect to the area of application of ARM for specific diseases or health related problems/issues and has been presented in Fig. 4. It can be clearly seen that ARM has been applied to an array of diseases or health related issues, specifically the cardiac diseases (Ordonez et al. 2006; Berka and Rauch 2010; Soni and Vyas 2010; McCormick et al. 2011; Jabbar et al. 2012; Lee et al. 2013) have been researched relatively more frequently. It can be observed that diseases such as lifestyle diseases (Ogasawara et al. 2005; Park et al. 2014) and diabetes (Concaro et al. 2009b, c, 2011), which frequently lead to cardiac problems, have also specifically been focused because of their implicit associations with cardiac problems. The results achieved related to cardiac related diseases are frequently novel, interesting and have real-world implications. It has been observed, through knowledge sharing and unsystematic interviewing that the results could be used to improve subjects' health as most of the subjects were unaware of the underlying causes of their cardiac issues. In this regard, knowledge sharing related to lifestyle diseases through public campaigning is necessary as well as mandatory.

In addition to cardiac related diseases and issues, various forms of tumors have been researched too. ARM has been applied for brain tumors' analysis (Pan et al. 2005; Ribeiro et al. 2009; Rajendran and Madheswaran 2010; Mahmood et al. 2014), breast cancers' analysis (Ribeiro et al. 2009; Kavipriya and Gomathy 2013), and oral cancers' analysis (Sharma and Om 2014), making it another application domain where it has been well experimented. It has been observed that tumors related knowledge generation was comparatively less successful than that of cardiac diseases related knowledge generation.

Additionally ARM, being an effective exploratory data analysis and knowledge generation technique, has been applied to Bird Flu (Mahgoub et al. 2008). While it has been used to explore epidemics such as dengue fever (Buczak et al. 2012; Thangam and Vanniappan 2015) and virus outbreaks such as Ebola virus (Go et al. 2014), which cause public crises, ARM application to these problems, according to our knowledge, do not have acceptable response-times. Since in case of such health crises, timely response is a key factor towards provision of healthcare, so ARM tools and technologies should be made more flexible to provide timely response.

Moreover, opposed to application of ARM to structured data, it has also been applied to unstructured data such as image and text data by various researchers. However, a few researchers have not clearly reported the specific application domain(s) and hence have been classified as Unclear/Unreported. Additionally, it can also be observed that the diseases' list to which ARM has been applied is very limited, this leaves opportunities for researchers as well as practitioners for mining interesting, unexpected and surprising association rules for other pervasive diseases including the diseases which sometimes cause public health crises.

4.4 Nature of datasets used for application of ARM for health informatics

Figure 5 displays the nature of datasets used by researchers for application of ARM for HI. The datasets sizes used in the researches were normally not huge for those researches which have reported the number of instances in the datasets used for experimentation purposes. It may however be observed that, the successful application of Apriori algorithm instead of the more efficient itemset mining algorithms may lie in the fact that limited number of instances (not more than three hundred thousand) were contained in the datasets. On the other hand, in data-driven age, when HISs produce giga-bytes of data daily, more and more data may be made available for experimentation, and efficient frequent/infrequent itemset and ARM algorithms utilizing big data may be used to mine these datasets. In this direction, recently (Chen et al. 2015) mined ARs in big data using gene expression data, in bio-informatics domain. While (Moens et al. 2013; Li 2014; Lin et al. 2015; Zitouni et al. 2015) utilized frequent itemset mining in big data in various application domains, Moradi and Keyvanpour (2015) applied

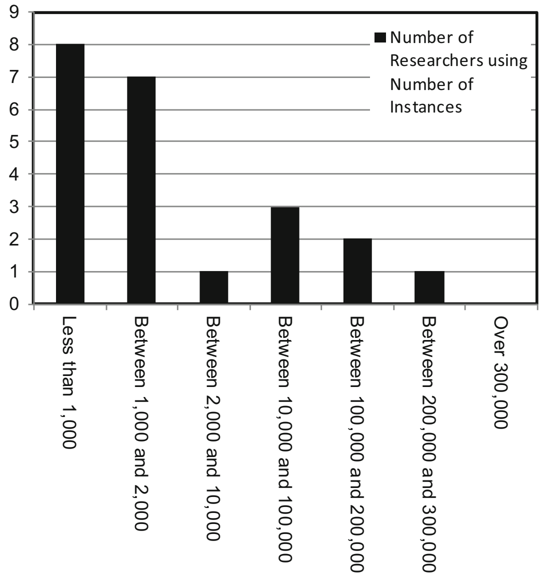


Fig. 5 Nature of datasets used for application of ARM for health informatics. Most of the researchers have applied ARM to small-sized medical and health related datasets (containing less than 2000 instances)

ARM to XML documents successfully, and (Vukićević et al. 2014) experimented with cloud based meta-learning for predictive modeling of biomedical data, therefore such techniques are required to be adopted for application of ARM for HI domain as well.

4.5 Preprocessing and post-processing techniques used for application of ARM for health informatics

In Fig. 6, the preprocessing and post-processing techniques used by researchers for application of ARM to structured and unstructured data for HI have been presented. It can be observed that preprocessing is an essential ARM task performed by most of the researchers with different levels of refinement. However, relatively fewer researchers have performed post-processing. In HI datasets the personal identification information of the patients have been removed and hence this preprocessing task is normally not performed by researchers, which suggests that the utilized datasets may not inherently contain any personal or identification information.

Moreover, the application of the presented post-processing techniques mainly suggests two distinct purposes. First, to reduce the number of ARs generated by pruning or restriction of size as the number of ARs generated are normally too large to analyze them effectively. While, the other one is, to mine important, surprising or unexpected ARs from the ARs generated through ranking, weighting, restriction of itemset(s) or meta-learning approaches as the irrelevant, unimportant or meaningless ARs generated may normally be huge in number. However, both of these issues have been addressed by Bouker et al. (2012, 2013, 2014) in a more flexible way, in terms that, the users neither have to worry about the heterogeneity and abundance of the interestingness measures, nor about the threshold values for the used interestingness measures. Surprisingly, in all the literature reviewed, none of the researchers have utilized the proposed post-processing methodology.

Additionally, the post-processing techniques used by the researchers as seen in the literature reviewed, suggest their unavailability in ARM tools discussed in Sect. 4.1, hence

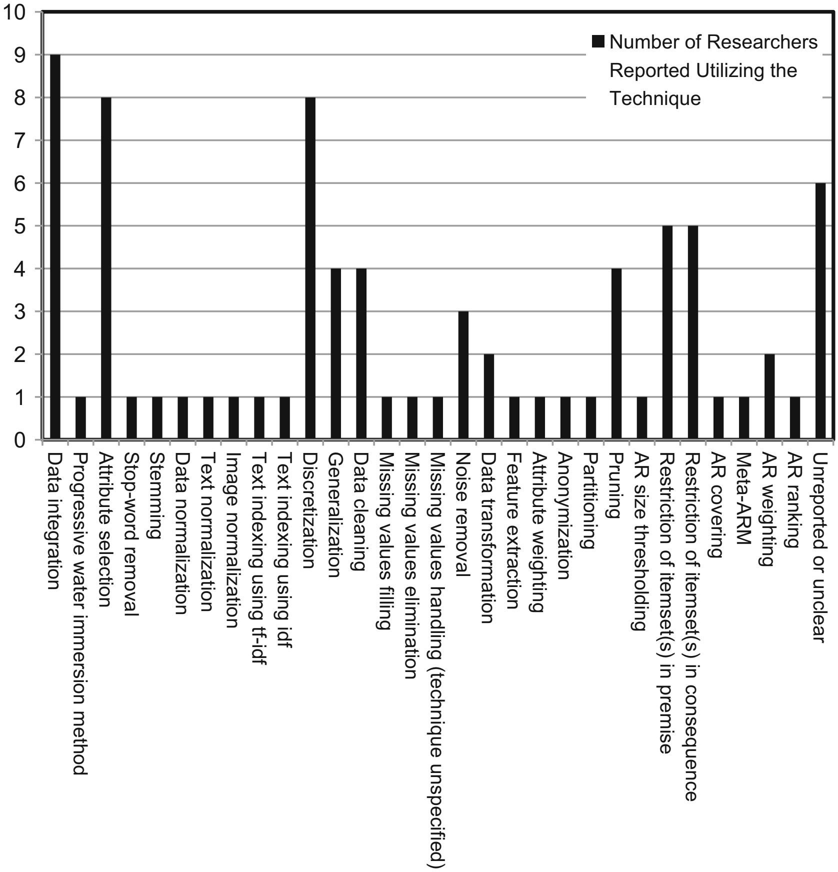


Fig. 6 Preprocessing and post-processing techniques used for application of ARM for health informatics. Literature reviewed suggests that preprocessing techniques have been extensively utilized by researchers, while limited post-processing techniques have been utilized, where the focus was restriction of itemsets in consequence and premise, pruning meta-ARM, weighting and ranking

requiring custom implementation. In order to improve the results of ARM process, support reusability and reduce implementation time and effort, the state of the art ARM tools may be extended to provide the required implementations.

5 Limitations of association rule mining applications in health informatics and relevant recommendations

In this section, the limitations of ARM applications in health informatics domain have been identified and respective recommendations have been made for mitigation of those limitations.

5.1 Domain expert requirement

The association rules generated for health informatics (as well as the ones generated in general) require subjective interestingness analysis from domain experts. The subjective interestingness analysis of ARs can be a tedious and expensive process given the large number of ARs generated through the ARM process. Hence, tools and techniques are required to be developed to automate this step, so that the need of domain expert can be minimized. One possible solution to this problem can be the utilization of Web for knowledgebase (KB) development, so that the domain knowledge (DK) gained through that KB can be utilized for automated evaluation of ARs' interestingness and meaningfulness. In addition to Web, other resources such as databases and digital libraries may aid in this KB development and DK gaining process too.

5.2 ARM systems for HI may only be DSS not automated decision making systems

As ARM is an exploratory data analysis technique which is used mainly for mining implicit and novel relationships in data. Healthcare domain is a sensitive field as it has direct effects on living beings, so it is of utmost importance that the applications of ARM for health informatics are error-free. Due to this high accuracy requirement, ARM may be used only for DSS and not for decision-making. This problem has been addressed by Bouker et al. (2012, 2013, 2014) and the researchers suggested that, the proper utilization of the abundant interestingness measures and threshold free approaches leads to higher accuracies in ARM. Moreover, Oliveira et al. (2014) explored various Artificial Intelligence (AI) based approaches for clinical guidelines' development and implementation enhancing decision support but still lacking trustworthy decision making. However, these approaches have still to be applied and evaluated in real-world HISs.

5.3 ARM performance should be improved

It is interesting as well as surprising to note that HI research community has been prone to use Apriori approach for frequent itemset generation (for ARM) as opposed to the more efficient FP-Growth approach. Since, Apriori approach has been known to be slower and less efficient when compared to FP-Growth, especially when applied to databases of large sizes, or with smaller support values, so it should be ceased to be used and FP-Growth, parallel FP-Growth (Li et al. 2008) or other efficient approaches should be used for frequent itemset generation for HI.

5.4 More data is required to be generated and made available

ARM is a technique applied to data and the results obtained are based on data quantity, quality and specificity. Through detailed review, it has been observed that, data may be limited due to which the support may be low for the surprising ARs, or the level of believe on the ARs may be weak. To overcome this limitation, more data is required to be generated, though it can be expensive and tedious task, but it can significantly improve the results by simply increasing the support count of surprising ARs. Data specificity and relatedness to the problem domain is another very important factor for application of ARM for HI. If data is specific to the problem domain, the chances of surprising or unexpected ARs may be limited. However, if data is less specific to the problem domain, then there may be more chances of surprising or unexpected ARs, with a disadvantage of possibly large number of meaningless ARs. It has been observed through literature that, there are cases where some or many data attributes have been discarded by researchers without any analysis based on prior knowledge or believes. So the unexpectedness or surprisingness of the ARs may be limited in those cases, hence the level of specificity of the data related to the problem domain may also be considered for more effective results.

5.5 ARM for HI should utilize Big Data

Since health information systems produce large amounts of complex, heterogeneous variable and variety of data, which is not being AR mined at a large scale. Specifically, recent advances in intercloud technology (Glott et al. 2011) and large-scale eHealth applications (Radu et al. 2015) make case for big data analytics and exploratory data analysis. Due to extensive processing, storage and security features provided by cloud, deep ARM may be explored to produce better results. Moreover, ARM for HI should also utilize big data using distributed machine learning such as Apache Spark$^{\mathrm{TM}}$ (Zaharia et al. 2010), for significant performance improvements.

6 Conclusion and future work

Exploratory data analysis for HI domain is an emerging field, and ARM can play a vital role in shaping its future. This however needs not only the ARM tools and techniques to be improved, but also the health information systems to be more flexible, informed, beneficial and open (still managing the data privacy and security concerns). As it has been found that, slower frequent itemset mining approaches such as Apriori and its variants are still being used frequently by the researchers because the datasets used are normally very limited in sizes. These approaches should be deprecated, and the more efficient alternatives available for the task should be utilized, both by researchers as well as practitioners, not only for considerable increase in performance but for mining larger datasets effectively. It is surprising to note that gigabytes of data produced daily by HISs is not being AR mined, this poses lots of opportunities for AR mining, analyses and consequently health improvements.

It has also been found that, association rule mining techniques have considerably advanced over the years, hence more advanced ARM tools for PNARM, STARM, Automated AR subjective interestingness analysis (AR semantic analysis), comprehensive interestingness analysis, and AR post-processing, should be developed and made available for research and development community, eliminating the requirement of custom development. AR post-processing techniques such as AR weighting, AR ranking, itemset restrictions in premise and/or consequence, meta-ARM may all be made available in the state-of-the-art tools. This will also enhance the ARM application effectiveness and provide a uniform platform for bench marking. All the aforementioned types of tools related advancements have array of dimensions, which creates commercialization opportunity for software development community too.

It has been found that, ARM applications for HI are still decision support systems, but not the much needed decision making systems. One main reason is the lack of accuracy due to which HISs cannot rely on ARM for decision making tasks. We believe that, in order to improve accuracy, HISs should utilize big data, cloud or intercloud computing for capturing, storing, transferring, sharing, analyzing and managing the patients data (without compromising the privacy of the data), so that the scale of ARM applications for analyses of health data can be extended to larger and complex databases as opposed to the traditional spreadsheets or the frequently used but limited sized relational data. Moreover, for improving accuracy and surprisingness, threshold independent and interestingness measures exhaustive approaches may also be explored further.

Additionally, it has been analyzed that ARM has been applied to relatively small number of commonly known but high mortality rate diseases. It is the time now to start exploring it to mine interesting, surprising and unexpected ARs for other diseases as well, because the availability of the resources with time is also improving. It is anticipated that in near future we would see that researchers will explore scalability issues in ARM using big data, cloud and/or intercloud computing tools and techniques for HI data analysis in general and for accuracy enhancement in particular. Specifically, the application of parallel FP-Growth based ARM for HI or health crises management in the cloud or intercloud may be researched.

References

Anwar MA, Ahmed N (2014) Analyzing lifestyle and environmental factors on semen fertility using association rule mining. Inf Knowl Manag 4(2):15–21

Babashzadeh A, Daoud M, Huang J (2013) Using semantic-based association rule mining for improving clinical text retrieval. Health Information Science, Springer, Berlin, pp 186–197

Badrinath N, Gopinath G, Ravichandran KS, Soundhar RG (2016) Estimation of automatic detection of erythemato-squamous diseases through adaboost and its hybrid classifiers. Artificial Intelligence Review, Springer Science+Business Media Dordrecht 45:471–488. doi:10.1007/s10462-015-9436-8

Berka P, Rauch J (2010) Mining and post-processing of association rules in the atherosclerosis risk domain. Information Technology in Bio-and Medical Informatics, Springer, Berlin, pp 110–117

Bouker S, Saidi R, Ben Yahia S, Mephu Nguifo E (2014) Mining undominated association rules through interestingness measures. Int J Artif Intell Tools 23(04):1460011

Bouker S, Saidi R, Ben-Yahia S, Mephu-Nguifo E (2012) Ranking and selecting association rules based on dominance relationship. In: 2012 IEEE 24th international conference on tools with artificial intelligence, pp 658–665

Bouker S, Saidi R, Ben-Yahia S, Mephu-Nguifo E (2013) Towards a semantic and statistical selection of association rules. arXiv preprint arXiv:1305.5824

Buczak AL, Koshute PT, Babin SM, Feighner BH, Lewis SH (2012) A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. BMC Med Inf Decis Making 12:124. doi:10.1186/1472-6947-12-124

Chen Y, Li F, Fan J (2015) Mining association rules in big data with NGEP. Cluster Comput 18(2):577–585. doi:10.1007/s10586-014-0419-3

Coira E (2003) Guide to health informatics. CRC Press, Boca Raton

Concaro S, Sacchi L, Cerra C, Fratino P, Bellazzi P (2011) Mining health care administrative data with temporal association rules on hybrid events. Methods Inf Med 50(2):166

Concaro S, Sacchi L, Cerra C, Bellazzi R (2009a) Mining administrative and clinical diabetes data with temporal association rules. MIE August 2009, pp 574–578

Concaro S, Sacchi L, Cerra C, Fratino P, Bellazzi R (2009b) Mining healthcare data with temporal association rules: Improvements and assessment for a practical use. Artificial Intelligence in Medicine. Springer, Berlin, pp 16–25

Concaro S, Sacchi L, Cerra C, Stefanelli M, Fratino P, Bellazzi R (2009c) Temporal data mining for the assessment of the costs related to diabetes mellitus pharmacological treatment. In: 2009 AMIA annual symposium, San Francisco, pp 119–123

Faghihi U, Fournier-Viger P, Nkambou R (2012) A computational model for causal learning in cognitive agents. Knowl Based Syst 30:48–56

Fournier-Viger P, Faghihi U, Nkambou R, Mephu Nguifo E (2012) CMRules: mining sequential rules common to several sequences. Knowl Based Syst 25(1):63–76

Fürnkranz J, Kliegr T (2015) A brief overview of rule learning. In: Rule technologies: foundations, tools, and applications. Springer International Publishing, New York, pp 54–69

Glott R, Husmann E, Sadeghi AR, Schunter M (2011) Trustworthy clouds underpinning the future internet. Springer, Berlin, pp 209–221. doi:10.1007/978-3-642-20898-0_15

Go E, Lee S, Yoon T (2014) Analysis of Ebolavirus with decision tree and Apriori algorithm. Int J Mach Learn Comput 4(6):543–546. doi:10.7763/IJMLC.2014.V4.470

Gosain A, Kumar A (2009) Analysis of health care data using different data mining techniques. In: International conference on intelligent agent & multi-agent systems, pp 1–6

Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. ACM SIGMOD Record 29(2):1–12

Han J, Kamber M, Pei J (2011) Data mining: concepts and techniques, 3rd edn. Morgan Kaufmann Publishers, Los Altos

Ilayaraja M, Meyyappan T (2013) Mining medical data to identify frequent diseases using Apriori algorithm. In: 2013 International conference on pattern recognition, informatics and mobile engineering (PRIME), pp 194–199

Iqbal S, Altaf W, Aslam M, Mahmood W, Khan MUG (2016) Application of intelligent agents in health-care: review. Artificial Intelligence Review. Springer Science+Business Media, Dordrecht, pp 1–30. doi:10.1007/s10462-016-9457-y

Jabbar MA, Chandra P, Deekshatulu BL (2012) Knowledge discovery from mining association rules for heart disease prediction. J Theor Appl Inf Technol 41(2):45–53

Kamsu-Foguem B, Rigal F, Mauget F (2013) Mining association rules for the quality improvement of the production process. Expert Syst Appl 40(4):1034–1045

Kavipriya A, Gomathy B (2013) Data mining applications in medical image mining: an analysis of breast cancer using weighted rule mining and classifiers. IOSR J Comput Eng 8(4):18–23

Köksal G, Batmaz I, Testik MC (2011) A review of data mining applications for quality improvement in manufacturing industry. Expert Syst Appl 38(10):13448–13467

Kuo YT, Lonie A, Pearce AR, Sonenberg L (2014) Mining surprising patterns and their explanations in clinical data. Appl Artif Intel 28(2):111–138

Lee DG, Ryu KS, Bashir M, Bae JW, Ryu KH (2013) Discovering medical knowledge using association rule mining in young adults with acute myocardial infarction. J Med Syst 37(2):1–10

Li X (2014) An algorithm for mining frequent itemsets from library big data. J Softw 9(9):2361–2365. doi:10.4304/jsw.9.9.2361-2365

Lin YC, Wu CW, Tseng VS (2015) Mining high utility itemsets in big data. Adv Knowl Discov Data Mining 9078:649–661. doi:10.1007/978-3-319-18032-8_51

Li H, Wang Y, Zhang D, Zhang M, Chang EY (2008) Pfp: parallel fp-growth for query recommendation. In: 2008 ACM conference on recommender systems. ACM, New York, pp 107–114. doi:10.1145/1454008.1454027

Mahgoub H, Rösner D, Ismail N, Torkey F (2008) A text mining technique using association rules extraction. International J Comput Intel 4(1):21–28

Mahmood S, Shahbaz M, Rehman ZU (2013) Extraction of positive and negative association rules from text: a temporal approach. Pak J Sci 65(3):407–413

Mahmood S, Shahbaz M, Guergachi A (2014) Negative and positive association rules mining from text using frequent and infrequent itemsets. Sci World J. doi:10.1155/2014/973750

Maquee A, Shojaie AA, Mosaddar D (2012) Clustering and association rules in analyzing the efficiency of maintenance system of an urban bus network. Int J Syst Assur Eng Manag 3(3):175–183

McCormick T, Rudin C, Madigan D (2011) A hierarchical model for association rule mining of sequential events: an approach to automated medical symptom prediction. SSRN eLibrary. doi:10.2139/ssrn.1736062

Mirabadi A, Sharifian S (2010) Application of association rules in iranian railways (rai) accident data analysis. Safety Sci 48(10):1427–1435

Moens S, Aksehirli E, Goethals B (2013) Frequent itemset mining for big data. In: IEEE international conference on big data, pp 111–118. doi:10.1109/BigData.2013.6691742

Moradi M, Keyvanpour MR (2015) An analytical review of XML association rules mining. Artificial Intelligence Review, Springer Science+Business Media, Dordrecht 43(2):277–300. doi:10.1007/s10462-012-9376-5

Nkambou R, Fournier-Viger P, Mephu Nguifo E (2011) Learning task models in ill-defined domain using an hybrid knowledge discovery framework. Knowl Based Syst 24(1):176–185

Ogasawara M, Sugimori H, Iida Y, Yoshida K (2005) Analysis between lifestyle, family medical history and medical abnormalities using data mining method—association rule analysis. Knowledge-Based Intelligent Information and Engineering Systems. Springer, Berlin, pp 161–171

Ohsaki M, Abe H, Tsumoto S, Yokoi H, Yamaguchi T (2007) Evaluation of rule interestingness measures in medical knowledge discovery in databases. Artif Intell Med 41(3):177–196

Oliveira T, Novais P, Neves J (2014) Development and implementation of clinical guidelines: an artificial intelligence perspective. Artificial Intelligence Review, Springer Science+Business Media, Dordrecht 42(4):999–1027. doi:10.1007/s10462-012-9376-5

Ordonez C, Ezquerra N, Santana CA (2006) Constraining and summarizing association rules in medical data. Knowl Inf Syst 9(3):1–2

Pan H, Li J, Wei Z (2005) Mining interesting association rules in medical images. Advanced data mining and applications. Springer, Berlin

Park SH, Jang SY, Kim H, Lee SW (2014) An association rule mining-based framework for understanding lifestyle risk behaviors. PloS one 9(2):e88859

Payus C, Sulaiman N, Shahani M, Bakar AA (2013) Association rules of data mining application for respiratory illness by air pollution database. Int J Basic Appl Sci 13(3):11–16

Radu A, Costan A, Iancu B, Dadarlat V, Peculea A (2015) Intercloud platform for connecting and managing heterogeneous services with applications for e-health. In: 2015 Conference on grid, cloud & high performance computing in science (ROLCG), Cluj-Napoca. doi:10.1109/ROLCG.2015.7367229

Raheja V, Rajan KS (2012) Comparative study of association rule mining and MiSTIC in extracting spatio-temporal disease occurrences patterns. In: 2012 IEEE 12th international conference on data mining workshops (ICDMW), pp 813–820

Rajendran P, Madheswaran M (2010) An improved image mining technique for brain tumour classification using efficient classifier. arXiv preprint arXiv:1001.1988

Rashid MA, Hoque MT, Sattar A (2014) Association rules mining based clinical observations. arXiv preprint arXiv:1401.2571

Ribeiro MX, Bugatti PH, Traina C Jr, Marques P, Rosa NA, Traina AM (2009) Supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques. Data Knowl Eng 68(12):1370–1382

Ruiz PP, Kamsu-Foguem B, Grabot B (2014) Generating knowledge in maintenance from Experience Feedback. Knowl Based Syst. doi:10.1016/j.knosys.2014.02.002

Russell S, Norvig P (2009) Artificial intelligence: a modern approach, 3rd edn. Prentice Hall, Englewood Cliffs

Savel TG, Foldy S (2012) The role of public health informatics in enhancing public health surveillance. MMWR Surveill Summ 61:20–24

Sharma N, Om H (2014) Significant patterns for oral cancer detection: association rule on clinical examination and history data. Netw Model Anal Health Inf Bioinf 3(1):1–13

Soni J, Ansari U, Sharma D, Soni S (2011) Intelligent and effective heart disease prediction system using weighted associative classifiers. Int J Comput Sci Eng 3(6):2385–2392

Soni S, Vyas OP (2010) Using associative classifiers for predictive analysis in health care data mining. Int J Comput Appl 4(5):33–37

Srinivasan S, Ramakrishnan S (2011) Evolutionary multi objective optimization for rule mining: a review. Artificial Intelligence Review, Springer Science+Business Media B.V 36(3):205–248. doi:10.1007/s10462-011-9212-3

Srinivas K, Rao GR, Govardhan A (2012) Mining association rules from large datasets towards disease prediction. Int Proc Comput Sci Inf Technol 27:22–26

Thangam M, Vanniappan B (2015) Mining association rules in dengue gene sequence with latent periodicity. Comput Biol J. doi:10.1155/2015/839692

Vukićević M, Radovanović S, Milovanović M, Minović M (2014) Cloud based metalearning system for predictive modeling of biomedical data. Sci World J. doi:10.1155/2014/859279

Wang C, Guo XJ, Xu JF, Wu C, Sun YL, Ye XF, Qian W, Ma XQ, Du WM, He J (2012) Exploration of the association rules mining technique for the signal detection of adverse drug events in spontaneous reporting systems. PloS ONE 7(7):e40561

Xianhai J, Cunxi X (2009) Home health telemonitoring system based on data mining. Int Forum Inf Technol Appl 2:431–434

Yu L (2009) Association rules based data mining on test data of physical health standard. Int Joint Conf Comput Sci Optimiz 2:322–324

Zaharia M, Chowdhury M, Franklin MJ, Shenker S, Stoica I (2010) Spark: cluster computing with working sets. In: HotCloud'10 2nd USENIX conference on Hot topics in cloud computing, p 10

Zitouni M, Akbarinia R, Yahia SB, Masseglia F (2015) A prime number based approach for closed frequent itemset mining in big data. In: 26th International conference on database and expert systems applications (DEXA'2015), vol 9261, pp 509–516. doi:10.1007/978-3-22849-5_35