IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 2, Ver. III (Mar.-Apr. 2017), PP 19-26
www.iosrjournals.org

Post Analysis Based AKD Using Domain Ontology for Disease Prediction

S.Antoinette Aroul Jeyanthi[1], Dr.S.Pannirselvam[2]
[1]*Department of Computer Science, Pope John Paul II College of Education, Pondicherry, India*
[2]*Department of Computer Science, Erode Arts & Science College, Erode, TamilNadu, India*

Abstract: *In present era, Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. Many of the ARM (Association rule Mining) approaches are well investigated in the literature, but it generates large number of association rules. If the dataset size is larger, then huge rules may occur, often it is a critical situation where decision making is difficult or unattainable because knowledge is not directly present in frequent patterns. This paper presents a hybrid model where post-analysis based domain ontology concept has been adopted to discover actionable knowledge from frequent patterns. For experimental study, we apply this approach on a clinical dataset of 1000 patients, contained symptoms having different diseases. Proposed approach follows three phase procedure in order to achieve actionable knowledge, in the first phase filter the uninteresting patterns, second phase rank the filtered patterns finally post analysis based approach has to be applied to discover actionable knowledge. The new approach is efficient and outperforms as compared to a previous AIRM algorithm in order to match knowledge discovery process.*

Keywords: *Actionable knowledge, Domain Ontology, Prior knowledge, Semantic distance,User belief*

I. Introduction

In health industry, Data Mining provides several benefits such as detection of the fraud in health insurance, availability of medical solution to the patients at lower cost, detection of causes of diseases and identification of medical treatment methods. It also helps the healthcare researchers for making efficient healthcare policies, constructing drug recommendation systems, developing health profiles of individuals etc. [1]. The data generated by the health organizations is very vast and complex due to which it is difficult to analyze the data in order to make important decision regarding patient health. This data contains details regarding hospitals, patients, medical claims, treatment cost etc. So, there is a need to generate a powerful tool for analyzing and extracting important information from this complex data. The outcome of Data Mining technologies are to provide benefits to healthcare organization for grouping the patients having similar type of diseases or health issues so that healthcare organization provides them effective treatments. It can also useful for predicting the length of stay of patients in hospital, for medical diagnosis and making plan for effective information system management.

Association is one of the most vital approach of data mining that is used to find out the frequent patterns, interesting relationships among a set of data items in the data repository. Association also has great impact in the healthcare field to detect the relationships among diseases, health state and symptoms. Ji et al., used association in order to discover infrequent casual relationships in Electronic health databases [2]. Healthcare organization widely used Association approach for discovering relationships between various diseases and drugs.

Traditional Association rule mining (ARM) approaches Apriori, FP-growth generates a huge number of frequent patterns which are not able to produce direct knowledge or inference. It is a serious issue when the clinical data are mined, where different combinations of symptoms may belongs to common disease because patient's symptoms may vary patient to patient but disease may be the same. Traditional ARM approaches fail due to their crispy nature, for example, if a patient has a high fever, low blood pressure, severe headache and cold, then it may be prone to malaria. This information cannot be discovered just by applying traditional ARM approaches which is concentrated on statistical data. Hence, a modified approach is needed to overcome this situation and efficient decision making process. This paper proposes a novel model for post-analysis based actionable knowledge discovery using domain ontology for disease prediction.

II. Related Work

Association Rules, ARs, reveal the relationships among items in a transaction of a database. However, the large number of generated ARs makes it difficult for decision makers to process, and interprets their utility. To tackle this limitation, several studies have been proposed to process the generated rules using objective as subjective methods and ontology.

2.1 Objective and Subjective methods for AR's Post-processing

In [3], Shaharanee et al. proposed the application of objective analysis to assess the generated rules. The approach consists in combining data mining and statistical measurement techniques (such as redundancy analysis, sampling and multivariate statistical analysis) to discard the insignificant rules. Alcala-Fdez et al. proposed a method based on rule covers to prune ARs [4]. The method defines subsets of rules describing the same transaction row. Then the rule set is reduced to its rule cover.

In [5], Franke introduced the notion of subsumed rules which consist in a set of rules having the same conclusion part and several additional conditions in the condition part. A similar approach was introduced in [6] where the authors proposed to combine two different algorithms of data mining applied on the same data set. By comparing condition and conclusion of both kind extracted rules, these rules are then categorized into robust, consistent, and noteworthy rules.

Other methods have proposed a rule-like formalism to model the user's expectations such as in [7]. Discovered rules are pruned/ filtered by comparing them to the user's expectations. In order to prune non-pertinent ARs, Concaro et al. have proposed in [8] a novel measure, called the Minimum Improvement measure. This measure describes the difference between the confidences of two rules. In fact, the rule can be pruned when the measure value is less than a fixed threshold.

In [9], the authors have proposed an iterative rule validation system based on several operators, including rule grouping, filtering, browsing, and redundant rule elimination. An original method was proposed in [10] to prune and organize rules with the same consequent. First, the algorithm transforms the database in an ARs base, and then meta-rules are extracted. These latter express the relations between two ARs and allow pruning/grouping of the discovered rules.

2.2. Ontologies in AR's Post-processing

Another set of existing methods applying ontologies in ARs post processing task have been proposed. In [11], the authors have focused on ontology-based ARs post processing to improve the integration of user's knowledge. This method is based on the use of a rule schema reflecting the user's expectation and an ontology involving concept constraints. The ARs evaluation is carried over the defined rule schemas in order to prune and filter rules.

For medical ARs filtering, authors in [12] proposed a hybrid pruning method involving the use of both objective and subjective analysis, with the latter involving the use of an ontology. The authors in [13] proposed a method that was applied using general medical-domain ontology constructed using the Unified Medical Language System with the goal of pruning already known rules.

In [14], the authors used ARs to point out dependence relationships between Gene Ontology terms using an annotation dataset and background knowledge. In [15], authors have proposed to group AR based on whether the rule items share relationships within a domain ontology. This method uses vector space modeling of rule elements and an ontology based semantic similarity measure. The latter is based on measuring the depth of the least common ancestor node of two concepts in the ontology.

In the perspective of computing ARs interestingness using domain ontology, the approach in [16] consists on calculating the conceptual distance by computing the number of edges of the shortest path between two concepts. The shorter the path is (from one concept to the other), the more similar the concepts are.

In [17], the authors proposed to combine concept similarity metrics, formulated using the domain ontology with traditional interestingness measures (support and confidence). The domain specific semantic similarity between two items i_1 and i_2 is defined as :

$$Sim(i_1, i_2) = \frac{\left(Dist(LCA(i_1, i_2), Root)\right)}{\left(Dist(i_1, i_2) + Dist(LCA(i_1, i_2), Root)\right)}$$

Where: LCA(i_1, i_2) is the lowest common ancestor of the concepts i_1 and i_2, Dist (LCA(i_1, i_2), Root) is the length of path from LCA(i_1, i_2) to the root and Dist(i_1, i_2) is a distance measure between i_1 and i_2 .

III. Proposed Framework

Post analysis-based AKD (PA-AKD) is a two-step pattern extraction. First, generally interesting patterns are mined from data sets by technical interestingness associated with the algorithms used. The mined general patterns are then pruned, distilled and summarized into operable business rules in terms of domain-specific business interestingness and involving domain knowledge. In the proposed model the domain knowledge is represented using ontology. The proposed model follows three phase procedure to achieve required Actionable Knowledge for Disease Prediction.

1. Filtering the uninteresting rules based on expert's beliefs.
2. Ranking the filtered rules using semantic conceptual distance based on domain ontology.
3. Integrating the above two phases and further analyse the ranked rules using past history of the patient ailment and predict the probable disease.

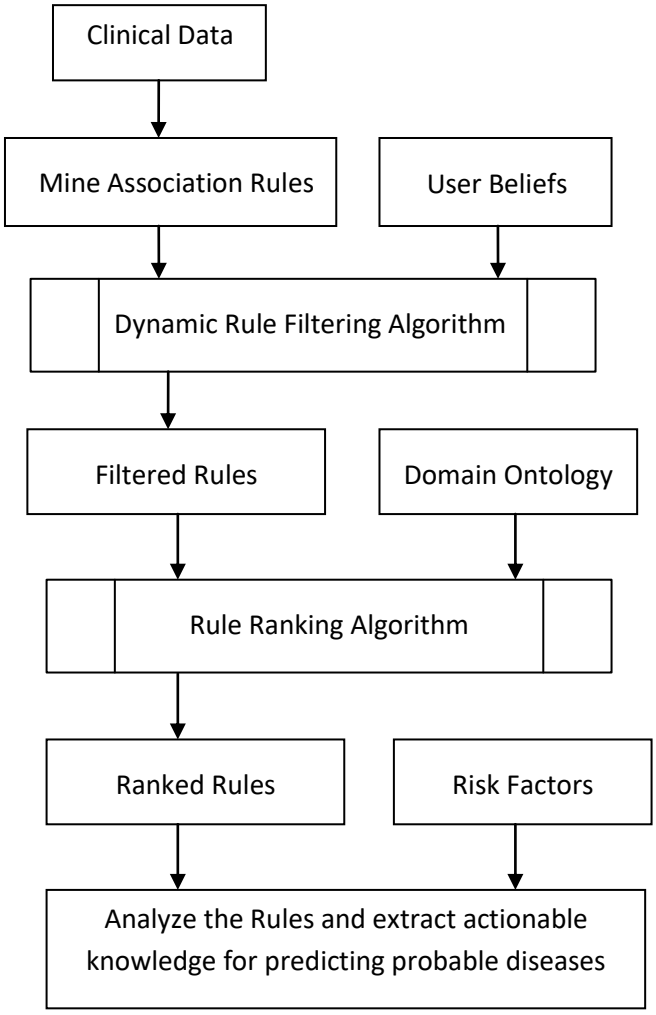

Figure 1. Process Flow of the Proposed PA-AKD model

3.1 Dynamic Rule filtering Technique using User Beliefs.

In this phase, in order to mine the dataset and to discover frequent patterns or frequently occurring symptoms, association rule mining has to be performed on the dataset. Under this process ARM algorithm generate frequent patterns as per predefined minimum threshold value and stored in frequent pattern base. The huge number of generating frequent patterns may be insufficient to draw conclusions; over here, there will be a need for an additional dynamic rule filtering technique to filter the uninteresting patterns based on the user's belief and reduce the size of the discovered patterns. The implementation and results of a dynamic rule filtering technique have been discussed in [18].

3.2. Ranking the Rules using Semantic Conceptual Distance based on Domain Ontology.

The semantic relationship exists between the attributes are captured and represented the domain knowledge using a Domain Ontology. The symptoms are categorized into three classes: Physical pain, Signs, Discomforts. The weight was set to one for all the symptoms that are directly associated to diseases. For the indirect symptoms the weight was increased by one at each level.

In this phase, the discovered rules are analyzed and detect the interrelation between various diseases and symptoms which are not directly associated in the dataset. Then the algorithm computes their semantic conceptual distance. The larger the distance, the more the rule is unexpected and therefore, interesting. The discovered rules are ranked based on the unexpectedness of the rule. The implementation of ranking the rules using semantic conceptual distance based on domain ontology has been discussed in [19].

3.3. PA-AKD using domain ontology for Disease Prediction

The Post analysis based AKD for Disease Prediction model adopt a simple and effective approach. This model integrates the first two phases and obtains the ranked rules based on subjective interestingness. The risk factors of the diseases are stored in the database. The user is asked to supply additional information about the patient like smoker, alcoholics and about other ailments like BP, obesity, diabetes, etc. These information are compared with the risk factors of the corresponding disease and based on the result, the decision is taken. The model is implemented using the following algorithm.

Step 1: Begin
Step 2: Load (DS), Min_Sup_Threshold, RS,RF
//Phase –I. Dynamic Rule Filtering
Step 3: Extract Filtered Rule Set (FRS) using User Beliefs
// Phase II. Ranking Algorithm
Step 4: Find the Semantic Conceptual Distance of discovered rules using Ontology
Step 5: Rank them based on unexpectedness.
// *Phase III. Analyse discovered rules and Finding Actionable Patterns*
Step 6: Read additional information and Compare it with Risk Factors based on that extract required knowledge.
Step 7: List the probable diseases for the symptoms.
The model analyzes the ranked rules based on the information given by the user and presents a list of probable diseases for the given symptoms.

IV. Experimental Evaluation

4.1 Experimentation

For experimental study a training data of 20 patients are considered. First phase of this approach says to perform the ARM algorithm on the dataset to find frequent pattern or frequent symptoms with minimum support 20%.This phase generates a huge number of frequent patterns which makes the knowledge discovery process too tedious. Here decision making is difficult or sometimes impossible because knowledge is not directly present in the frequent pattern base (containing huge number of frequent patterns) can't give knowledge directly in a practical sense. To resolve this problem, a dynamic rule filtering technique using user beliefs is used to filter uninteresting rules and therefore reduce the size.

Second phase of this model calculates the semantic conceptual distance of the filtered rules using domain ontology. If the distance is more, then the rules are unexpected. So it is considered as interesting. The model detects the indirect relationship between the symptoms. All the unexpected rules are on the top of the list. More importance is given to those rules because serious diseases are ignored due to less frequency. Common diseases are easily predicted using statistical significance. The table-4.1 shows top ten ranked rules.

TABLE-4.1: Top ten ranked rules

R1 : abdominal pain,loss of appetite,vomiting R2 : abdominal pain,diarrhea,bloated belly	Appendicitis Irritable bowel syndrome
R3 : vomitting,nausea,burning sensation in chest R4 : vomiting,nausea,burning sensation in chest R5 : abdominal pain,vomiting,headache	Indigestion Cornary Artery Gastrentertis
R6 : abdominal pain, nausea,vomiting,blood in urine R7 : abdominal pain, nausea,vomiting	Kidney stones Ulcers
R8 : fever,headache,muscle pain	Flu
R9 : increased urine output,thirst,hunger,fatigue	Diabetes
R10: abdominal pain, nausea,vomiting	pelvic inflammatory disease

Third phase of the model analyses the ranked rules with the help of additional information about the patient and his/her ailments. The task of disease prediction is not simple because, some or all symptoms of a frequent pattern may be responsible for the same disease (all or partial symptoms may belong to common disease). Apart from that, the analysis says that the exact matching procedure is adopted; it may drop many of interesting patterns. That may be responsible for incorrect prediction.

The concept of human cooperated mining is adopted. The real requirements for discovering actionable knowledge in constraint-based context determine that real data mining is more likely to be human involved rather than automated. Human involvement is embodied through cooperation between humans (including users and business analysts, mainly domain experts) and data mining system.

Case 1:

Consider the rules R3 and R4, both has same symptoms resulted in two different diseases. Ranks of both the rules are also same. At this juncture, the medical practitioner has to decide whether to choose R3 or R4.

R3 :vomitting,nausea,burningsensation in chest -->Indigestion

R4 : vomiting,nausea,burning sensation in chest -->Cornary Artery

In the given example, if the patient has BP then he may have a high possibility of Cornary Artery than indigestion.

Case 2:

Consider the rules R6 and R7; R6 has higher rank than R7.

R6 : abdominal pain, nausea,vomiting,blood in urine→Kidney stones

R7 : abdominal pain, nausea,vomiting →Ulcers

If the patient age is below 40, then there is no chance for Kidney stones even if the patient has a symptom of blood in urine. So the chances of ulcers are higher.

4.1 Results and Analysis

The result analysis has been carried out to analyze the possibilities of disease infection. Consider the patient reporting with the symptoms of abdominal pain and vomiting. The experimental study performing in section-4.1, lists the possibilities of the following diseases infection. Table-4.2 lists the probable diseases.

TABLE 4.2 – Probable diseases

Disease name	Rank			
	Male	Female	Age<40	Age>50
Appendicitis	6	6	6	6
IBS	5	5	5	5
Gastrentertis	4	4	4	4
Kidney stones	3	3	0	3
Ulcers	2	2	2	2
PID	0	1	0	0

By considering the age and sex of a patient, the analysis draws following knowledge which is helpful in the decision making process of predicting diseases for the medical practitioner. Figure 4.1 and figure 4.2 depicts the result analysis of the Table 4.2 over the training data.

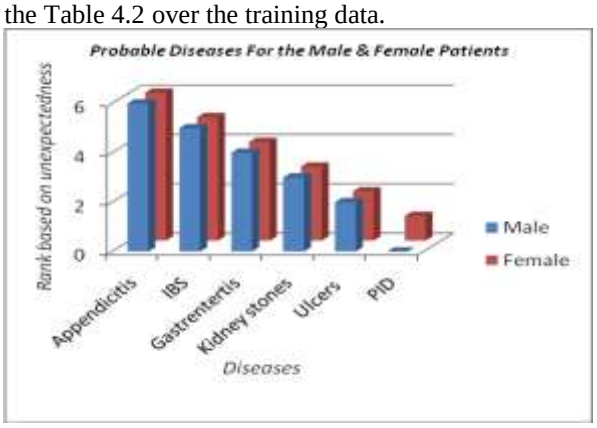

Figure 4.1 Probable diseases for male & Female Patients

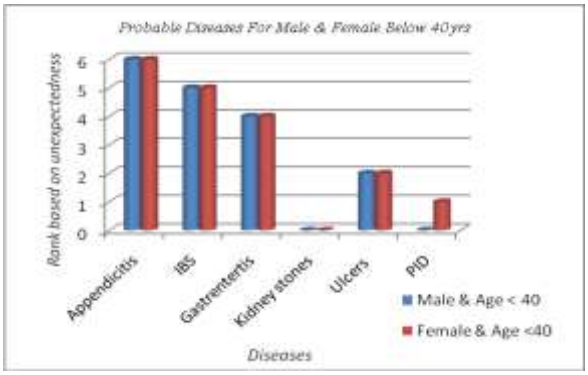

Figure 4.2 Probable diseases for patients below 40 years

V. PERFORMANCE EVALUATION

Evaluation of the learned knowledge usually involves measuring the performance using a test data set. The following measures are considered for evaluating the performance of the proposed model.

- Precision
- Recall
- False Discovery rate
- Accuracy

A total of 248 records with different medical attributes (factors) according to the type of disease were obtained from clinical database. This dataset contains data of two different diseases Coronary artery and Ulcers. The records were split into two datasets: training dataset (174 records) and testing dataset (74 records). To avoid bias, the records for each set were selected randomly. The trained model was evaluated against the test datasets for accuracy.

Confusion matrix is a tool used to measure the effectiveness of the model, which sorts all cases from the model into categories, by determining whether the predicted value matched the actual value. The rows in the matrix represent the predicted values for the model, whereas the columns represent the actual values. A classification matrix is an important tool for assessing the results of prediction because it makes it easy to understand and account for the effects of wrong predictions. By viewing the amount and percentages in each cell of this matrix, we can quickly see how often the model predicted accurately.

In our examples, for the first disease Coronary Artery, the test dataset contained 8 patients with Coronary Artery disease and 66 patients without coronary artery. Table 4.3 shows the results of the confusion matrix for the two models, the Naive Bayes and the proposed model.

Table 4.3 Confusion Matrix for the two models for Coronary Artery

Counts for Naive Bayes on Coronary Artery		
Predicted	FALSE	TRUE
FALSE	60	5
TRUE	6	3

Counts for Proposed Model on Coronary Artery		
Predicted	FALSE	TRUE
FALSE	63	2
TRUE	3	6

In our examples, for the second disease Ulcer, the test dataset contained 26 patients with Ulcer and 48 patients without ulcer. Table 4.4 shows the results of the confusion matrix for the two models, the Naive Bayes and the proposed model.

Table 4.4 Confusion Matrix for the two models for Ulcer.

Counts for Naive Bayes on Ulcer		
Predicted	FALSE	TRUE
FALSE	44	5
TRUE	4	21

Counts for Proposed Model on Ulcer		
Predicted	FALSE	TRUE
FALSE	47	2
TRUE	1	24

Table 4.5 shows the values of various measures such as precision, recall, false discovery rate and accuracy of the two models for both the diseases.

Table 4.5 Measures of the two models for both diseases.

	Coronary Artery		Ulcer	
Measures	Naive Bayes	Proposed	Naive Bayes	Proposed
Precision	0.9091	0.9545	0.8333	0.8750
Recall	0.9231	0.9692	0.8696	0.9130
FDR	0.0909	**0.0455**	0.1667	**0.1250**
Accuracy	0.8514	**0.9324**	0.8108	**0.8649**

The figure 4.3 and figure 4.4 shows the comparison chart of proposed model with Naive Bayes for the two diseases.

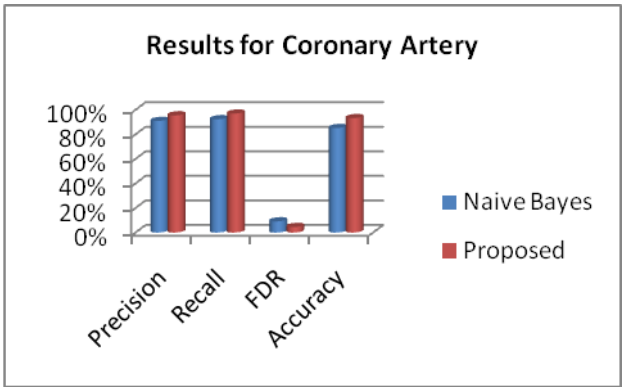

Figure 4.3 Performance Chart for Coronary Artery

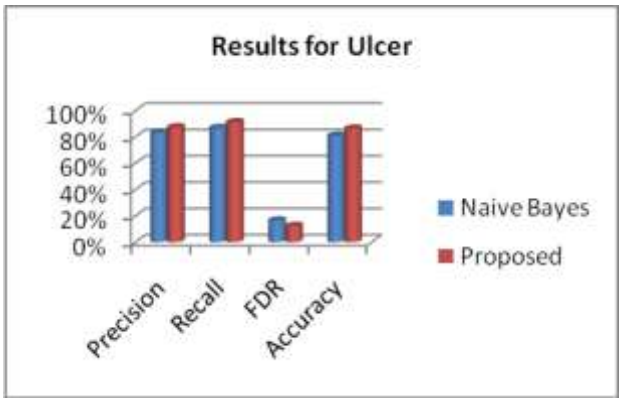

Figure 4.4 Performance Chart for Ulcer

The proposed model appears to be most effective as it has the highest percentage of correct predictions (95.45%) for patients with coronary artery and (87.50%) for ulcer, where the Naive Bayes has(90.91%) and (83.33%) for ulcer. The proposed model is also has less false discovery rate (0.04 and 0.12) for coronary artery and ulcer respectively, when compared to Naïve Bayes model. This is because the model incorporates patient's medical profile and the risk factors of the disease in the process of predicting the disease.

VI. Conclusion

The goal of this model is to give actionable knowledge which helps the practitioner to consider indirect relations between the symptoms and the other factors like age, sex, etc., which ought to be taken into account during the decision making process. Ignoring certain deceiving symptoms can prove to be fatal, as the disease might reach its untreatable stage. This approach makes the user give equal attention to all the symptoms. Thus it helps to prevent discomforts and deaths which result due to the mere carelessness of the decision maker. This hybrid model may play an essential role in the field medical data mining to obtain beforehand recognition and alert about the particular disease to prevent it.

References

[1]. H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, (2005).
[2]. J. Yanqing, H. Ying, J. Tran, P. Dews, A. Mansour and R. Michael Massanari, "Mining Infrequent Causal Associations in Electronic Health Databases", 11th IEEE International Conference on Data Mining Workshops, (2011).
[3]. M. Shaharanee, I. Nizal, H. Fedja et D. Tharam, «Interestingness measures for association rules based on statistical validity,» Knowledge-Based Systems, vol. 24, n° 13, pp. 386-392, 2011.
[4]. P. Manda, F. McCarthy, B. Nanduri et M. Bridges, «Information Theoretic Interestingness Measures for Cross-Ontology Data Mining in the Mouse Anatomy Ontology and the Gene Ontology,» Computational Engineering, Finance, and Science (cs.CE), pp. 116, 2015.
[5]. D. Franke, «System and method for efficiently generating association rules,» U.S. Patent 8,401,986,, n° %18, 2013.

[6]. S. Anthony, R. Karthik, R. Kulathur «Finding Persistent Strong Rules» Data Mining: Concept, Methodologis vol. 28, pp. 85-103, 2012.
[7]. B. Liu, W. Hsu«Visually Aided Exploration of Interesting Association Rules,» Knowledge Discovery, vol. 1574, pp. 26-28, 1999.
[8]. S. Concaro, L. Sacchi, C. Cerra, P. Fratino et R. Bellazzi, «Mining healthcare data with temporal association rules: Improvements and assessment for a practical use.,» Artificial Intelligence in Medicine, pp. 16-25, 2009.
[9]. G. Adomavicius et A. Tuzhilin., «Expert-Driven Validation of Rule» Data Mining and Knowledge Discovery, pp. 33-58, 2001.
[10]. A. Berrado et G. Runger., «Using metarules to organize discovered AR» Data Mining and Knowledge Discovery, pp. 409-431, 2007.
[11]. R. Shrikant et R. Agrawal, «Mining Generalized Association Rules,» Proc.21st Int conf Very Large Database, pp. 407-419, 1995.
[12]. G. Mansingh «The Role of Ontologies in Developing Knowledge Technologies,» KM for Development, pp. 145-156, 2015.
[13]. C. Marinica et F. Guillet, «Knowledge-based interactive postmining of association rules using ontologies,» IEEE Transactions on knowledge and data engineering, vol. 22, p. 784–797, 2010.
[14]. G. Mansingh «Using ontologies to facilitate post-processing of association rules,» Information Sciences, vol. 181, pp. 419-434, 2011.
[15]. S.Myhre «Additional gene ontology structure for improved biological reasoning,» Bioinformatics, vol. 22, n116, pp. 2020-2027, 2006
[16]. A.Kumar, B.Smith et C.Borgelt, «Dependence Relationships between Gene Ontology Terms based on TIGR Gene Product Annotations,» 3rd International Workshop on Computational Terminology, pp. 31-38, 2004.
[17]. P.Razan, T.Groza, J.Hunter et A.Zankl, «Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain,» Journal of Biomedical Semantics, vol. 5, n18, pp. 1-13, 2014.
[18]. S.Antoinette Aroul Jeyanthi, S.Pannirselvam, "Dynamic Rule Filtering Technique using user beliefs" International Journal of Computer Society & Information Science,(ISSN:1947-5500), Vol 13, Special Issue,pp1-4, July 2015
[19]. S.Antoinette Aroul Jeyanthi, S.Pannirselvam, "A Domain ontology Method for Semantic Conceptual Distance based on Rule Ranking Algorithm", International Journal of Innovative Technology and Creative Engineering, (ISSN:2045-8711),Vol-5,No.3,March 2015.