Indian Journal of Science and Technology, *Vol 10(20), DOI: 10.17485/ijst/2017/v10i20/111941, May 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Improved Rule Based Classifier Based on Decision Trees (IRBC-DT) for Gastric Cancer Data Classification

Thara Lakshmipathy[1]* and Gunasundari Ranganathan[2]

[1]Department of Computer Science, Karpagam University, KAHE, Coimbatore – 641 021, Tamil Nadu, India; kltharavijay@gmail.com
[2]Department of Information Technology, Karpagam University, Coimbatore – 641 021, Tamil Nadu, India; gunasoundar04@gmail.com

Abstract

Objectives: To design and develop an improved rule based classifier based on decision trees (IRBC-DT) for Gastric Cancer data classification with increased accuracy, hit rate and substantial reduction of elapsed time. **Methods/Analysis:** At the initial stage, IRBC-DT mingles a pair of techniques, namely the boosting and arbitrary sub-space, in order to build rules based on classification. As a result, the subsequent level divides the dataset into two parts in which the first set for training the data and the second one for pruning. Then a decision tree is built for analyzing the misclassified instances. **Findings:** Each feature is tested and assigned with precise weight for which the k-nearest neighbor classifier is applied, based on the weighted features. As a final point, the algorithm will get updated with the instances which contain the misclassified class labels. Once subsequent to analysis and updating the instances of misclassified class labels, the conflicting rules are checked and the same are removed. Attribute bagging is a set of classifier that operate on a sub-space of the original element space, and produces the class corresponds to the result of those unique classifiers. Random subspace scheme has a striking option of data classification that ensemble with considerably more number of features, such as cancer data. Also, boosting is modeled particularly for classification, which alters the weak classifiers into strong ones by means of an iterative process. Boosting mechanism makes use of selecting the apt classification in order to coalesce the complete classifier results. **Applications/Improvements:** IRBC-DT is implemented in MATLAB and can be applied in healthcare sector. From the results it is perceived that the method gains better performance than that of the existing algorithms for gastric cancer data classification.

Keywords: Accuracy, Decision Trees, Elapsed Time, Gastric Cancer, Hit Rate, IRBC-DT, Misclassified Instances

1. Introduction

In the field of recent computer science research arena, data mining always stands to be a thrust research area. To add feather to the cap of data mining, soft computing take part for the design and development of predictive data mining as well as descriptive data mining applications. It perks up significant extent for the healthcare industry for making patient – centric health systems and health professionals. Such systems and professionals is used to systematically utilize the data and analytics for identifying consoles and best practices that leads to improved patient care and also results in reduced costs. A survey on Data Mining Techniques that had been employed for Bio Medical Research, presents the significance of data mining algorithms in disease diagnosis process[1].

As per the statistics on death rate review, cancer is the deadliest disease which leads to death even in developed nations. Predominantly, gastric cancer takes up the fourth position of universal common cancer. At global level, gastric cancer is believed as one of the major causes of cancer deaths. This lays concrete on the motivation to assisting the issues distressing the happening of the infection owing to the commonness of the syndrome and the

*Author for correspondence

elevated death rate of gastric cancer. This research work aims in design and development of improved rule based classifier based on decision trees for gastric cancer data classification. The performance metrics such as accuracy, hit rate and elapsed run time are obtained for evaluation. The proposed classifier improves the accuracy, hit rate and also reduces the time taken for classification.

2. Background

There exist some methods that had successfully been recognized to choose, split and categorize subtypes of Gastric Cancer (GC) and to figure out a few symptomatic predicaments[1]. In early gastric cancer (EGC), swelling attack is narrow to the mucosa or submucosa paying little mind to the proximity of lymph hub metastasis or not. Gene expression analysis accepted a signature that separated EGC from usual tissue[2]. The process of breaking 124 tumors and adjacent mucosa tests, examined the sub-atomic elements of gastric cancer, which could be apparent to swiftly characterized premalignant and swelling subtypes, utilizing DNA microarray-based gene expression profiling[3]. The identifiable proof of atomic signatures that are standard for subtypes of gastric cancer and connected premalignant changes should allow examination of the means obligatory in the start and progress of gastric cancer. A conditional test on 1024 genes (52% up-controlled and 48% down-directed) that were variably communicated in 19 EGC tests when disparity and 9 typical tissues[4]. These up-control genes are incorporated in cell cycle, DNA handling, ribosome bio-genesis, and association of cytoskeleton, while the down-control genes are embroiled in particular elements of the gastric mucosa (assimilation, lipid digestion system, and G-protein-coupled receptor protein flagging pathway). Likewise, another research distinguished a 973-gene signature to divide EGC from typical tissue making use of the micro selection information from the synchronized tumor and neighboring cancer less tissues of 27 EGC patients[5]. The result showed that the up-directed genes in EGC tissues were associated with cell relocation and metastasis. After a Gene expression study, it was revealed that 60 genes were incessantly up or down-managed in sequence in typical adenoma, carcinoma and mucosa tests by seeming at the expression outline of these tissues from eight patient-coordinated sets. For that reason, atomic order comes into view to be remarkably encouraging for sub-atomic study of EGC[6].

The IM (intestinal metaplasia) and the ChG (unending gastritis) are integrated in middle of the road phase of GC, the preceding portrayed by a gene expression signature relate to mitochondria. At the same time, as the last depicted by sign of multiplication. Because of this similarity, ChG has the possibility to undergo enthusiastic test which predicts the signature that goes with the metabolic subtype of GC[7]. Naturally, the degree of difference communicated gene set amongst ChG and IM is to a great extent covered with the GC metabolic signature (P = 0.00085, hyper geometric test).

CUP (Cancer of unknown primary site) is a very much perceived medical issue, representing 3 to 5% of the entire dangerous epithelial cancer. Glass can be recognized in light of preserved tissue-particular gene expression[8]. It has been demonstrated that gene expression profiling can distinguish tissue of starting point with an exactness rate somewhere around 33% and 93%[9]. One similar study connected a 92-gene CUP in order to examine the tumor tests of the patients with CUP. Among 20 cases, 15 (75%) found to be effectively anticipated. That means that those anticipated CUPs assumed as genuine inactive essential locales which were recognized after the underlying finding of CUP. This measure has been effectively connected to numerous different cancers, for example, bosom, colorectal, and melanoma[10].

The genetic material supporting the signature techniques be able to be utilized for distinguishing particular medication for GC patients, i.e., focused on treatments. For a substantial planned test on n, equals to 289, a genetic material expression signature was created to anticipate the tissue of starting point in many patients with CUP. The middle survival period was 12.5 months for victims who got examine coordinated site-particular treatment contrasted and the utilization of empiric CUP regimens. The affected persons whose CUP destinations were anticipated to have additional responsive swelling sorts survived longer than those anticipated to have less responsive lump sorts[11].

A data set comprised samples about 819 (24 +ve and 795 -ve samples) and 31 features is sampled for the following techniques namely CN2 Rules, C4.5 and Naive Bayes to diagnose stomach cancer[12]. The outcome shows that HER-1 protein and sex are significant factors in the identification and categorization of GC. Investigational results shows that the standard sensitivity is greater than 50 and falls between 86 and 100 percentage. At the same time, it has the categorization correctness and specificity

is closer in the range between 65 and 70 percentages. A similar study used classification tree analysis to examine data from a population-based case–control study (1095 cases, 687 controls) accomplished in Connecticut, New Jersey, and Western Washington State[13]. The hierarchical clustering is applied on 14 accessible medical factors from 3 categories, such as the immune histo-chemistry data, medical background and the stage information of Caner. The results showed that that two clinical factors, HER-1 protein and sex, can clearly describe and distinguish these 3 groups[14].

3. Improved Rule Based Classifier based on Decision Trees (IRBC-DT)

As for as the supervised learning is concerned, information order is made out of a two-stage procedure of training and testing. A mining model/classifier is worked from the occurrences in training stage, where cases are related with class marks. At that point in testing stage, the classifier is utilized to foresee the class marks for the testing examples. The improved rule based classifier based on decision trees (IRBC-DT) is proposed in this research work. The IRBC-DT mingles the irregular subspace and boosting approaches in order to build a set of classification rules. Attribute bagging is a set of classifier each working in a subspace of the first element space, and yields the class in light of the yields of these individual classifiers. Irregular subspace plan is a striking decision for characterizing information sets with considerably more number of features, such as cancer data. Also, boosting is modeled particularly for characterization to change frail classifiers to solid ones which is an iterative process. Boosting mechanism makes use of selection for categorization in order to coalesce the yield of individual classifiers. In our projected IRBC - DT, we deem irregular subspace method for shun over fitting. The IRBC – DT has taken a series of k iterations to generate good classification rules that does not want any optimization technique(s).

Initially, an identical weight, $\frac{1}{N}$ will be given to each occasion, xi in unique preparing informational collection, D. Here N is the aggregate number of examples. The weights of preparing occurrences will be attuned based on how they are classified in all iterations. When an example x_i has been properly off the record then weight will get decreased. When misclassification happens, after that the weight value will get augmented. From this it can be observed instance's weight reflects based on the difficulty. During every iteration, a sub-data set D_j will be obtained from the new training information set D and preceding sub-data set D_{j-1} by highest weighted examples. A DT_j was created as of the sub-data set D_j with irregularly chosen features in every iteration. Then every rule will be obtained for every leaf node of DT_j. Every pathway in DT_j from the root to a leaf compares with a rule. The error rate of DT_j will be ascertained by the aggregate of weights of misclassified occurrences. When an example, x_i is misclassified, then err(x_i) is one. Or else, err (x_i) is zero, as example, x_i is properly off the record.

When the fault pace of DT_j is fewer than the threshold-value, then rules are extracted from DT_j. The threshold-value is 0.49. So, if a tree, DT_j properly off the record 50% of preparation examples as of sub-data set, D_j then the generated rules can be obtained from the tree. The weights of accurately ordered occurrences were refreshed in the wake of extricating rules. In the event that a case, x_i in j^{th} iteration will be properly off the record, it's weight will be multiplied by the fault $\left(\frac{error(DT_j)}{1-error(DT_j)}\right)$. Then the weights for all occasions (counting the misclassified examples) were standardized. Keeping in mind the end goal to standardize a weight in this work it is duplicated by the total of old weights, partitioned by the whole of new weights. Subsequently, the weights of misclassified cases are expanded and the weights of effectively ordered occurrences will be diminished. As a last point, a sub-information set D misclassified will be created from D_j with the misclassified instances.

3.1 Algorithm – 1: Working of IRBC-DT

The proposed IRBC-DT is used for mining the cancer data. A few littler specimens (or subsets) of information are taken from the informational collection. Every subset of information is utilized to build an arrangement of rules with the help of IRBC-DT, which gives a few arrangements of grouping rules as result. At that point the rules are inspected and used to combine to build the last arrangement of order rules to manage the tumor information. The IRBC-DT working is shown in Algorithm – 1. The proposed IRBC-DT is suitable for performing classification in cancer information, as new rules from new

information can be included with existing rules without irritating the current rules. Besides, rules can be executed in random manner.

In this proposed work the creative preparation information will be split into two parts: (a) Training set, and (b) Pruning set. More often than not, the two-thirds of training instances will be made use for training set and 33% of examples are utilized for pruning set. The order rules are created from cases in the developing set as it were. In order to analyze the misclassified instances, initially a decision tree will be built from the misclassified instances. After that each feature will be experienced in the tree, A j ∈ Dmisclassified is assigned with the weight. The features that are not tested will not be taken into consideration of kNN classifier. Next the k-nearest neighbor classifier is applied based on the weighted features.

***Algorithm - 1*: Working of IRBC-DT**

$Input: D = \{x_1, ..., x_i, ...x_N\}$, *training dataset*;
k is the number of iterations;
Output : $rule - set$
Method :
1 : $rule - set = NULL$;
2 : *for* $i = 1$ *to* N *do*
3 : $x_i = \frac{1}{N}$; // *Initializing weights*
4 : *end for*
5 : *for* $j = 1$ *to* k *do*
6 : *if* $j == 1$ *then*
7 : *create* D_j, *by sampling* D *with replacement*;
8 : *else*
9 : *create* D_j, *by* D_{j-1} *and* D *with* X;
10 : *end if*
11 : *build tree* $DT_j \leftarrow D_j$ *by randomly selected features*
12 : *compute error* (DT_j)
13 : *end if*
14 : *if error* $(DT_j) \geq$ *threshold then*
15 : *goto line* 6
16 : $rules \leftarrow DTj$; // *Rule Extraction*
17 : *end if*
18 : *for each* $x_i \in D_j$ *that was correctly classified do*
19 : *multiply the weight of* x_i *by* $\left(\frac{error(DT_j)}{1 - error(DT_j)}\right)$
20 : *end for*
21 : *normalize the weight of each* $x_i \in D_j$;
22 : $rule - set$;
23 : *end for*;
24 : *return* $rule - set$;
25 : *create* $sub - data\ set, D_{misclassified}$ *from* D_j;
26 : *analyze* $D_{misclassified}$ *employing* $A\lg orithm\ 2$.

3.2 Algorithm – 2: Mechanism of Analyzing the Misclassified Instances

At long last, the calculation will get refreshed with the class name of misclassified occasions. Once in the wake of investigating and refreshing the classes of misclassified occasions the opposing rules are checked and the same will be removed. The mechanism of analyzing the misclassified instances is given in Algorithm 2.

***Algorithm – 2:* Mechanism of analyzing the misclassified instances**

1. *Build a decision tree DT with* $D_{misclassified}$
2. *for each* $A_j \in D_{misclassified}$ *do*
3. *If* A_j *is tested in DT then*
4. *Assign weight to* A_j
5. *Else*
6. *Not to consider similarity measure*
7. *End if*
8. *End for*
9. *For each* $x_i \in D_{misclassified}$ *do*
10. *Identify* $x \in D$ *with the similarity weighted* $A = \{A_1, A_2, ..., A_n\}$;
11. Re*cognize the most frequent class*
12. *Assign* $x_i \leftarrow c_l$
13. *End for*

4. Dataset and Performance Evaluation

The dataset collected from leading cancer care hospital that includes the testimony of 470 patients, each of which has 29 features. All features are considered as pointers of gastric cancer for a patient, as per medicinal writing. Be that as it may, some of them have never been utilized as a part of information digging based methodologies for gastric disease finding. The features are orchestrated in four gatherings: individual attributes, conduct, systemic features and the stomach.

Accuracy, hit rate and elapsed run time are the performance metrics taken for comparison, based on which the proposed classifier IRBC-DT is compared with the other two existing algorithms namely Apriori algorithm[16] and Ontology based Apriori algorithm[15].

As far as the accuracy of performance evaluation is concerned, 4 parameters namely True Positive, True Negative, False Positive and False Negative are used to compute the accuracy percentage, as mentioned below:

Table 1. Performance evaluation on accuracy

Algorithm	True Positive	True Negative	False Positive	False Negative	Accuracy %
Apriori Algorithm	291	26	123	30	67.44 %
Ontology based Apriori Algorithm	330	15	54	71	73.40 %
IRBC-DT (Proposed Work)	391	42	16	21	92.12 %

True positive (TP) : Gastric cancer patients correctly identified as affected

True negative (TN) : Unaffected patients correctly identified as unaffected

False positive (FP) : Unaffected patients incorrectly identified as affected

False negative (FN) : Gastric cancer patients incorrectly identified as unaffected

TP+TN+FP+FN = 470

Table 1 illustrates the performance evaluation on accuracy and computes, Accuracy percentage = ((TP+TN) /470) x 100. Table 2 illustrates the performance evaluation on hit rate and elapsed time.

Table 2. Performance evaluation on hit rate and elapsed run time

Algorithms	Hit Rate	Elapsed Run time
Apriori Algorithm	64%	2090 seconds
Ontology based Apriori Algorithm	71%	125 seconds
IRBC-DT (Proposed Work)	88%	78 seconds

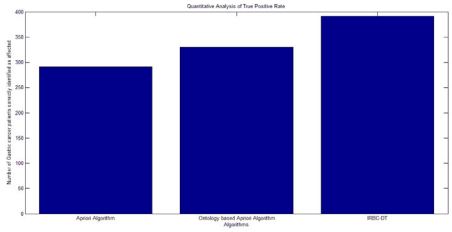

Figure 1(a). True positive: Gastric cancer patients correctly identified as affected.

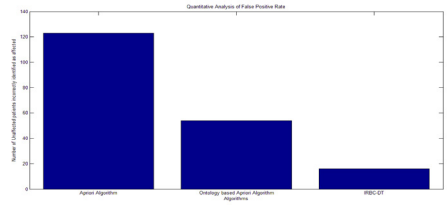

Figure 1(b). True negative: Unaffected patients correctly identified as unaffected.

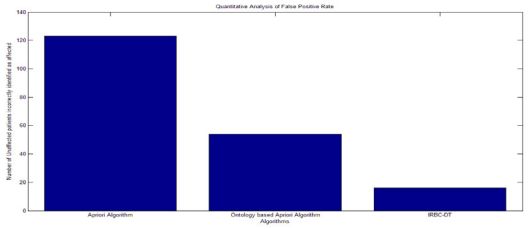

Figure 1(c). False positive: Unaffected patients incorrectly identified as affected.

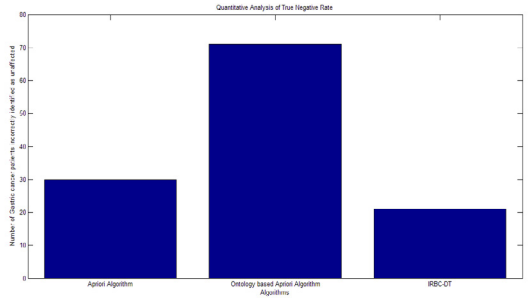

Figure 1(d). False negative: Gastric cancer patients incorrectly identified as unaffected.

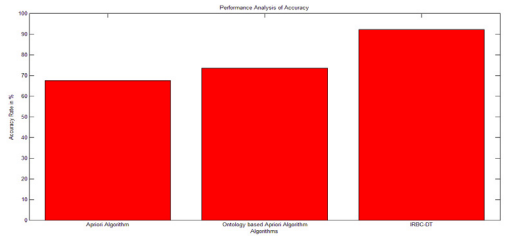

Figure 2. Accuracy of Apriori algorithm, Ontology based Apriori Algorithm and IRBC-DT.

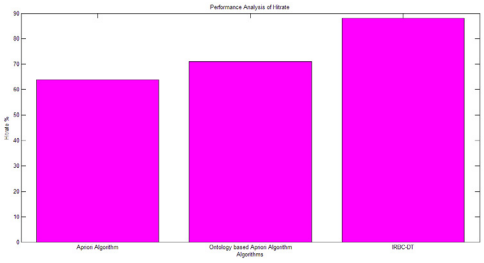

Figure 3. Hit rate of Apriori Algorithm, Ontology based Apriori Algorithm and IRBC-DT.

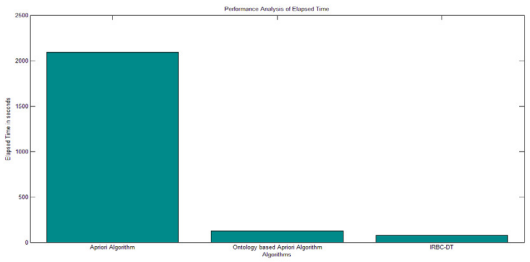

Figure 4. Elapsed time of Apriori Algorithm, Ontology based Apriori Algorithm & the proposed IRBC-DT.

5. Results and Discussion

Genuine positive, genuine negative, false positive, false negative are quantitatively analyzed and depicted in Figure 1. It is evident that the IRBC-DT model outperforms the other two algorithms in TP, TN and FP. Figure 2 portrays the accuracy rate of the algorithms and IRBC-DT outperforms other two algorithms and attain better classification accuracy of 92.12%. Figure 3 depicts performance analysis of hit rate of the algorithms and it is evident that the IRBC-DT outreaches than the two algorithms and obtains 88%. Figure 4 exposes the performance analysis in terms of elapsed time of execution of the algorithms and it is noteworthy that IRBC-DT consumes less time i.e. 78 seconds to classify 470 patient records.

6. Conclusion and Future Works

This article presents improved rule based classifier based on decision trees for gastric cancer data classification. Initially IRBC-DT merges the arbitrary subspace and boosting approaches for making an arrangement of grouping rules. Then, the dataset is divided into two parts namely training set and pruning set. In order to analyze the misclassified instances, a decision tree is created from the misclassified instances. Then every feature in the dataset is tested and assigned with the weight. The k-nearest neighbor is used on the weighted elements. Finally, the calculation gets refreshed with the class mark of misclassified cases. When breaking down and refreshing the classes of misclassified occasions the incongruous rules are ensured and the same are detached. The proposed IRBC-DT is implemented in MATLAB and the results portrays that IRBC-DT outperforms than that of ontology based algorithm and ontology based Apriori algorithm in terms of accuracy, hit rate and elapsed time. As far as FN is concerned, the IRBC-DT has little bit performance degradation and there is scope for future work to reduce the false negative.

7. Acknowledgement

We are extremely grateful to the Karpagam Academy of Higher Education, for the persistent guidance in research related activities. We would also like to extend our thanks to the management, for providing all sorts of support and necessary facilities to develop the research.

8. References

1. Thara L, Gunasundari R. Significance of Data mining techniques in disease diagnosis and Biomedical Research - A survey. The IIOAB Journal. 2016 Nov; 284–92.
2. Brettingham-Moore KH, Duong CP, Heriot AG, Thomas RJ, Phillips WA. Using gene expression profiling to predict response and prognosis in gastrointestinal cancers-the promise and the perils. Ann Surg Oncol. 2011 May; 1484–91. Crossref, PMid:21104326
3. Balasubramanian SP. Evaluation of the necessity for gastrectomy with lymph node dissection for patients with submucosal invasive gastric cancer. Br J Surg. 2001 Aug; 1133–4. PMid:11494983
4. Boussioutas A, Li H, Liu J, Waring P, Lade S, Holloway AJ, Taupin D, Gorringe K, Haviv I, Desmond PV, Bowtell DD. Distinctive patterns of gene expression in premalignant gastric mucosa and gastric cancer. Cancer Res. 2003 May; 2569–77. PMid:12750281
5. Vecchi M, Nuciforo P, Romagnoli S, Confalonieri S, Pellegrini C, Serio G, Quarto M, Capra M, Roviaro GC, Avesani CE, Corsi C, Coggi G, Di Fiore PP, Bosari S. Gene expression analysis of early and advanced gastric cancers. Oncogene. 2007 Jun; 4284–94. Crossref, PMid:17297478
6. Nam S, Lee J, Goh SH. Differential gene expression pattern in early gastric cancer by an integrative systematic approach. Int J Oncol. 2012 Nov; 1675–82. PMid:22961301, PMCid:PMC3982715
7. Kim H, Eun JW, Lee H, et al. Gene expression changes in patient-matched gastric normal mucosa, adenomas, and carcinomas. Exp Mol Pathol. 2010 Sep; 201–9. PMid:21185829
8. Lei Z, Tan IB, Das K. Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. Gastroenterology. 2013; 554–65. Crossref, PMid:23684942
9. Pavlidis N, Pentheroudakis G. Cancer of unknown primary site. Lancet. 2012; 1428–35. Crossref

10. Monzon FA, Koen TJ. Diagnosis of metastatic neoplasms: molecular approaches for identification of tissue of origin. Archieves of Pathology and Laboratory Medicine. 2010 Feb; 216–24.
11. Greco FA, Spigel DR, Yardley DA. Molecular profiling in unknown primary cancer: accuracy of tissue of origin prediction. Oncologist. 2010 Apr; 500–6. Crossref, PMid:20427384 PMCid:PMC3227979
12. Hainsworth JD, Rubin MS, Spigel DR. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon research institute. J Clin Oncol. 2013 Jan; 217–23. Crossref, PMid:23032625
13. Kirshners A, Parshutin S, Leja M. Research on application of data mining methods to diagnosing gastric cancer, advances in data mining. Proceedings of Industrial Conference on Data Mining, Lecture Notes in Computer Science. 2012; 7377:24–37. Crossref
14. Silvera SAN, Mayne ST, Marilie D, Gammon D. Diet and lifestyle factors and risk of subtypes of esophageal and gastric cancers: classification tree analysis. Ann Epidemiol. 2015 Jan; 50–7.
15. Wang X, Duren Z, Zhang C, et al. Clinical data analysis reveals three subtypes of gastric cancer. Proceedings of IEEE 6th international conference on systems biology, 2012. p. 315–20.
16. Mahmoodi SA, Mirzaie K, Mahmoudi SM. A new algorithm to extract hidden rules of gastric cancer data based on ontology. SpringerPlus. 2016 Mar; 5:312. Crossref, PMid:27066344 PMCid:PMC4786510
17. Rakesh A, Srikant R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB. 1994 Sep; 487–99. PMid:8054149