ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Symptom based Cancer diagnosis to assist patients using Naive Bayesian Classification

Vineetha V *, and Asha P.

Department of Computer Science and Engineering ,Sathyabama University, Chennai, India.

ABSTRACT

Mining with medical services is the exploration point now. Most researched topic is cancer, as affected rate of people hiked year by year. The survival rate is minimum due to late recognition of cancer as patients are not aware of the disease and its symptoms. The minimal survival rate is due to carelessness of people even after undergoing some unseen symptoms in them. This study focus on a guidance to the patients based on the symptoms they have undergone. Also provides an detailed assistance regarding the possibility of cancer, its prevention, treatment, side effects of the treatment, Best Hospitals and Doctors to prefer. To break down information, Naive Bayesian classification is used along with stemmer to provide a better and viable result. Accuracy is obtained by prioritizing the most relevant rules mined. Thus, this method provides a simple and a detailed guidance about cancer to the affected patients.

Keywords: Classifiation; Symptom; Sentence; Stemmer.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Mining useful data from a large amount of available raw medical data is a red-hot now. Individuals all over are stressed over their wellbeing conditions and to screen those illnesses, they experience some online applications, download virtual products, look in google, and so on. In this manner the principal check step is handled by the data accessible online or some devoted versatile applications. Just if this beginning confirmation step gives no better answer for mending the sickness or if the circumstance is not controlled, then they adjust for counseling a specialist.

The work additionally focuses on such an online data as a helpline for patients to screen their illness and to give a nitty gritty depiction about that malady, for example, cure, counteractive action, side effects, best hospitals and doctors to consult. Cancer is the horrible illness and it is the most expanding sicknesses as well. Be that as it may, the majority of the general population is not mindful of the data with respect to tumor such as, what are its manifestations? What all malignancies are there? What are its medicines and symptoms?. Mindfulness about the ailment is less as there are more sorts of growth, for example, lung, eye, kidney, liver, stomach, and so on. Despite the fact that malignancy is an appalling ailment, there is a cure when it is perceived at an early stage. The paper talks about the early stage growth help to help patients by giving them both the positive rules and negative rules, keeping in mind the end goal to give them proposals to counsel to further references i.e. a specialist.

Mining rules to concentrate positive rules are investigated by numerous methodologies were negative ones are not given much significance. Thus the study is about a medicinal aid that focuses on recovering both positive and negative rules in a superior way by compacting the general information accessible. Stemming strategies help in this pressure of information to half so as to diminish and spare memory space. To break down information, Naive Bayesian Classification method is talked about and its execution is assessed expressing as it is observed to be more viable, simple and less tedious than different routines. Hence, in this work segment of information gathered is painstakingly improved execution.

RELATED WORK

In 2015, Gyorgy et al. [1] proposed a calculation on Survival Association Rule (SAR) to discover survival results where Association Rule Mining (ARM) is needed. As ARM discovers co-occuring regular examples, yet neglects to consider other components like age, in light of the fact that relying upon the age figure, the hypertensive and hyperlipidemic variables do fluctuate. Hence SARM amplifies ARM by taking care of survival results, making conformity for confounders and also consolidating different elements to mine the powerful standards. Some of the summarization techniques such as RPGlobal, APRX-collection, Topk and Bottom p Summarization (BUS) techniques are analyzed and BUS is finalized as the best technique that apply summarization on the covering the patient records rather than considering only the expression rules. Also BUS controls redundancy and better in quality when compared to other techniques. Thus, the rules are more interpretable and more suitable to assess risk, but Prediction is not appropriate for individual patients.

Oana Frunza et al. [2] used a machine learning approach in 2012, to identify the disease-treatment relations. The work is categorized into two, one to identify and another to extract the relation. Initial step is to find the accurate model for predicting the input by classification method. For this purpose, few classification methodologies such as decision based technique, Adaptive learning model, probabilistic based technique, support vector machine and a classifier model is involved. The next step analyses Bag-Of-Word (BOW) representation, NLP with Biomedical and Medical concept to extract the disease-treatment relationship using very few information about the disease. The resultant is that, BOW representation when combined with any of the classification methodologies generated accurate results yet it lacks to provide diagnosis information about a particular disease.

In 2012, Idheba et al. [7] integrated two algorithms such as Positive Negative Association Rules (PNAR) and Interesting Multiple Level Minimum Support's (IMLMS) to a new approach called PNAR_IMLMS.The original IMLMS approach is slightly modified at prune step so as to remove meaningless rules, this generates interesting frequent and infrequent itemsets. Then correlation and Valid Association Rule based on Correlation Coefficient & Confidence (VARCC) measures are used to mine positive rules from

frequent itemsets and negative rules from both frequent and infrequent itemsets. Thus, valid positive and negative association rules are resulted avoiding uninteresting rules.

Shweta Kharya [16] in 2011, analzsed many classification techniques such as decision tree, Association Rule Mining (ARM) with Artificial Neural Network (ANN), Naive Bayes classifier for diagnosis and prognosis of breast cancer. Weka is used to experiment the different parameters used to form the tree and also to decide the splitting nodes. From this method, a significant associations are found by this tree, but evaluation is to be processed in a larger set for a higher degree of confidence. In ARM with ANN, a classifier is formed with ARM as a first step and a neural network consisting of 2 layers are formed based on the classification system to diagnose cancer. Naive Bayes classifier is used to predict the probability of the presence of cancer. From among these techniques, decision tree provides a better accurate result.

Ramasubbareddy et al. (2010) presented [9] an approach called Negative Association Rule (NAR) that uses Apriori to retrieve positive rules initially. Then from the rules retrieved, k negative itemsets are obtained. Later candidate generation and pruning is done to find the valid positive and negatively associated rules. Thus, this approach produces negatively associated rules from the positively associated rules reducing an extra scan to the database.

Yanguang et al.[12] introduces a new Interest_support_confidence approach which overcomes traditional Support_confidence that misleads association rules in 2009. The new mining method initially checked whether minimum interest has met and then correlation measure with the support measure is determined. This evaluation finds positive, negative and independent rules. After that the positive and negative rules are checked whether it satisfies minimum support and confidence. The only difficulty is, support and confidence for negative rules cannot be found directly as it includes absence of itemsets. Still the method generates a reduced set of positive association rules with more meaningful negative association rules.

PNAR algorithm was proposed by Honglei et al.[5] in 2008, that mines valid rules quicker through correlation coefficient measure and pruning strategies. At first, positive and negative rules are extracted from the frequent and infrequent itemsets. Using pruning strategy, interesting positive rules are mined that satisfies both minimum support and confidence measure along with a correlation coefficient in order to remove contradicting rules. Then interesting negative rules are mined as positive rules except that the minimum support and confidence is different. Thus, all valid association rules are found.

In 2008, Shi-ju et al. [10] described PNAR_MDB on P_S measure algorithm to mine association rules from multiple databases. From a large company, multiple database along with its weightage is retrieved. Thus support count is calculated with slight changes by including weight factor, but confidence remained same. Then itemsets are pruned with correlation measure, the survived rules are then passed on to undergo P-S measure for mining more interesting rules. The number of rules gets decreased with more interestingness that avoids knowledge conflicts within the database while mining association rules simultaneously.

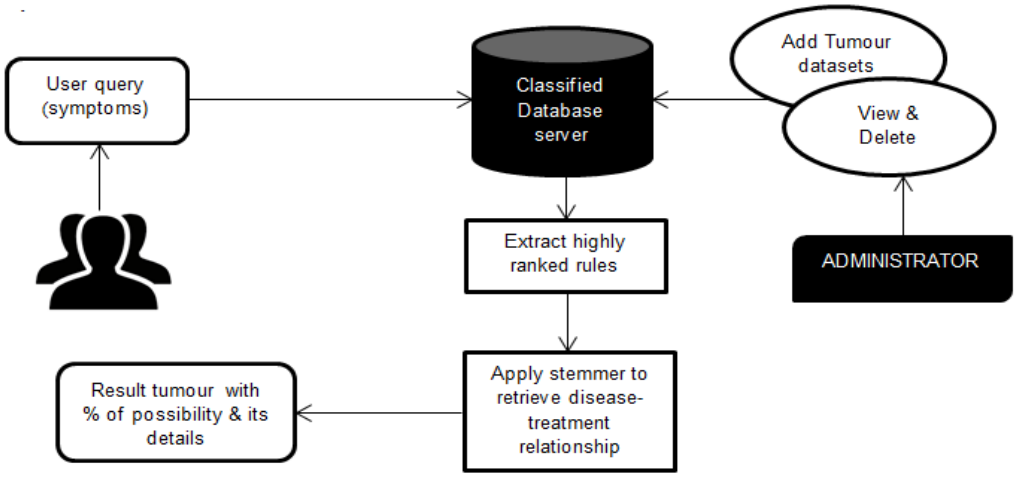
PROPOSED WORK

The Proposed System:

The general idea is about principle mining from existing data accessible about growth utilizing stemmer, bunching investigation and positioning in light of a weight assigned to each rule.

The proposed work in Fig 1, is all about assisting the user or a patient with all the information available about cancer. Initially the patient query is nothing but the unseen symptoms they experience in their day to day life and maybe they were so confused about its future outcome. This confusion can be clarified by inputting the undergone symptoms with comma separated. The query is then passed on to the server where all the cancer related data sets are classified and stored. The symptoms entered are matched with the dataset and then the relevant rules retrieved are ranked to prioritize them. Stemmer applies to the ranked rules to fetch the treatment regarding every cancer possibility. Then the resultant is the all possible cancer types along with its probability percentage to intimate the seriousness of the disease to the user.

ISSN: 0975-8585



Data set:

Cancer data set (D) is collected from UCI Data sets where all available types of cancer analyzed regarding the patient record and the symptoms are filtered making the dataset. Our dataset outlet consists of different symptoms with a weight provided and a comma separated combination of all possible symptoms called sentences (S) with two classes (C), 'C_1' and 'C_2'. C_1 is the 'yes' class that has all combinations of symptoms that insist that cancer has occurred where C_2 a 'no' class conveys a negative result to the user indicating the patients who been experiencing these symptoms, not being affected by cancer. Both these positive and negative sentences (PNS) are retrieved to the user, so as to provide some awareness among the people.

Table 1: Type of cancer and its symptoms as attributes.

NAME stomach cancer
ATTRIBUTE Indigestion-22
ATTRIBUTE Pain in abdomen-20
ATTRIBUTE Loss of appetite-18
ATTRIBUTE Weight loss-13
ATTRIBUTE Vomiting-10
ATTRIBUTE Sense of fullness-7
ATTRIBUTE Nausea-5
ATTRIBUTE Swelling-3
ATTRIBUTE Fatigue-2

Table 2: Combination of symptoms defining yes or no class that helps to predict cancer.

CLASS (YES, NO)
Indigestion,Loss of appetite,Weight loss,Vomiting,Sense of fullness,Nausea,Swelling,YES
Pain in abdomen,Loss of appetite,Weight loss,Vomiting,Nausea,Swelling,YES
Weight loss,Vomiting,Sense of fullness,Nausea,Fatigue,YES
Vomiting,Indigestion,Paininabdomen,Weight loss,Sense of fullness,Nausea,Swelling,YES
Pain in abdomen,Indigestion, Loss of appetite,Weight loss,Nausea,NO
Sense of fullness,Indigestion,Weight loss,Vomiting, Nausea,NO
Indigestion,Pain in abdomen,Loss of appetite,Weight loss,Vomiting,Sense of fullness,Nausea,NO
Nausea,Indigestion,Loss of appetite,Weight loss,Sense of fullness, NO

Table 3: Number of possible combinations of symptoms forming sentences.

	Bone	Eye	Kidney	Liver	Lung	Stomach
No of Symptoms	7	10	7	11	15	9
Combination of Symptoms	127	1023	127	2047	32767	511

ISSN: 0975-8585

Classifier:

To predict the user input to which classification it belongs to, (i.e) yes or no class. Classifier uses a classification algorithm in order to classify into categorical data. To commence with, there is a need for a training data set where the algorithm is applied to generate the classified rules. Training set D is provided applying the Naive Bayesian classification algorithm to produce the resultant classified rules. From the resultant rules, we can classify to which class the rules belongs. The classification is done based on the class label in the dataset. This is a sort of predicting the inputted query to analyze whether it belongs to a class C_1 or C_2.

Algorithm_ Bayesian_Classification()

1. Let D be the training set of sentences or tuples represented as S = { S_1, S_2,.. S_n). The 'n' measurements are made on the 'n' attributes such A_1, A_2,... A_n respectively.
2. Suppose that there are 'm' classes C_1, C_2,... C_m. Given a tuple S, the Naïve Bayesian Classifier will predict that S belongs to the class C_i iff **P** (C_i|S) > **P** (C_j|S). Thus we maximize P (C_i|S).

By Baye's Theorem,

$$\mathbf{P}(C_i|S) = \frac{\mathbf{P}(S|C_i)\mathbf{P}(C_i)}{\mathbf{P}(S)}$$

3. As **P** (S) is a constant for all classes, only {**P** (S|C_i) **P** (C_i)} need to be maximized.

$$\mathbf{P}(C_i) = \frac{\textbf{Presence of particular sample}}{\textbf{Total no. of samples}}$$

4. Evaluate **P** (S|C_i),

$$\mathbf{P}(S|C_i) = \prod_{k=1}^{n} \mathbf{P}(S_k|C_i)$$

Which means, **P** (S_k|C_i) = **P** (S_1|C_i) * **P** (S_2|C_i) *P (S_3|C_i)..... *P (S_n|C_i)

5. In order to predict the class label of S, **P** (S|C_i) **P** (C_i) is evaluated for every class C_i. The classifier predicts that the class label of tuple S is the class Ci, iff

P (S|C_i) **P** (C_i) > **P** (S|C_j) **P** (C_j).

Stemmer:

Stemming is a procedure of stripping so as to diminish the words to its root word or supplanting the prefix or postfix or even both. There are more number of stemmers, but here we utilize the affix stemming technique to diminish the word as it stripes both the postfix and prefix. Yet, while applying the stripping technique alone might bring about negligible words. Accordingly, we incorporate fasten substitution alongside stripping any place it is required. Stemming is applied to find the disease-treatment relationship after the related rules are being extracted. Stemmer is processed in order to compress the sentences retrieved from the dataset so as to incorporate an easy capturing of all related cancer information to the extracted relevant sentence or rules. Affix stemmer is a better and quicker technique, as it doesn't keep up a different lookup table along these lines sparing memory space.

Table 4: Applying the affix stripping stemmer to the input we get,

Input	Output- Stripping
Irregularities	Regular
Depression	Depress
Vomiting	Vomit
Muscle-cramps	Muscle-cramp
Swelling	Swell

ISSN: 0975-8585

Table 5: Applying the affix stripping and substitution stemmer to the input we get,

Input	Stripping	Output - substitution
Decreased	Decreas	Decrease
Troubling	Troubl	Trouble
Urination	Urinat	Urinate

Prioritizing by Ranking and Relevant Resultant:

Ranking is utilized to organize the relevant sentences separated from D. In the wake of separating the related principles from D, positioning is done to the extricated rules based upon the weightage given to each attribute in the group or a class. The weightage is based upon a few investigations of term recurrence where the term suggests the side effects that has been so visit with the patients who have experienced that kind of tumor. The yield to the user is top prioritized rules where they all the more much of the time happen inside of the patients. What's more, to alarm the patients with its malady earnestness, the sort of disease alongside the percentage(%) of plausibility is given therefore to the client. The resultant tumor sort can be seen further so as to help the client with its full depiction about the cure, counteractive action, side effects, best hospitals list and doctors to consult. Another part is the negative result, where the patients who have experienced the same kind of side effect and need not have prone to the risk of cancer.

Such part of the manifestations are likewise considered as exceptions and their event can be uncommon now and again moreover. These exceptions can likewise be considered as a quality to clients as they might likewise get to be one of such an anomaly.

EVALUATION RESULT

Evaluation Measures:

Accuracy, Basle measures such as Precision and Recall are the common measures to evaluate the functioning of the system. 'Accuracy' is to estimate the correctness of the proposed model, in retrieving the relevant list of cancer types of the users requested query of symptoms. Also accuracy is actually defined as a system functioning without error. Suppose that a system has retrieved a list of results based on the input, but how to check whether they are accurate or correct?. To satisfy this purpose, we use 'Precision' and 'Recall' measures. For measurement, we must analyze two terms such as relevant (RL) and retrieved (RT) result. Relevant is the set of result related to the user query where the other term is the set of result retrieved for the query. To commence with, Precision is the fraction of retrieved results that is relevant to the query. Thus, it is a quality measure and high precision is a term used for the algorithm which yields more relevant results than irrelevant instances.

$$\text{Precision} = \frac{RL \cap RT}{RT} \text{ (i.e) } \text{Precision} = \frac{\text{no.of relevant (correct) results}}{\text{Total no.of retrieved results}}$$

Next the Recall measure, which is the fraction of relevant result retrieved. Thus, this is a quantity measure that checks number of relevant instances retrieved. High Recall is used for a method that results in more number of relevant instances. F-measure is the harmonic mean of both Precision and Recall.

$$\text{Recall} = \frac{RL \cap RT}{RL} \text{ (i.e) } \text{Recall} = \frac{no.of\ relevant\ (correct)\ results}{Total\ no.of\ \ relevant\ results\ that\ should\ been\ returned}$$

Table 6: F-measure evaluation results

	Precision	Recall	F-Measure
Bone	96%	100%	98%
Eye	85%	94.44%	89.72%
Kidney	95%	95%	95%
Liver	88.89%	88.89%	88.89%
Lung	77.78%	87.50%	82.64%
Stomach	100%	88.89%	94.46%

ISSN: 0975-8585

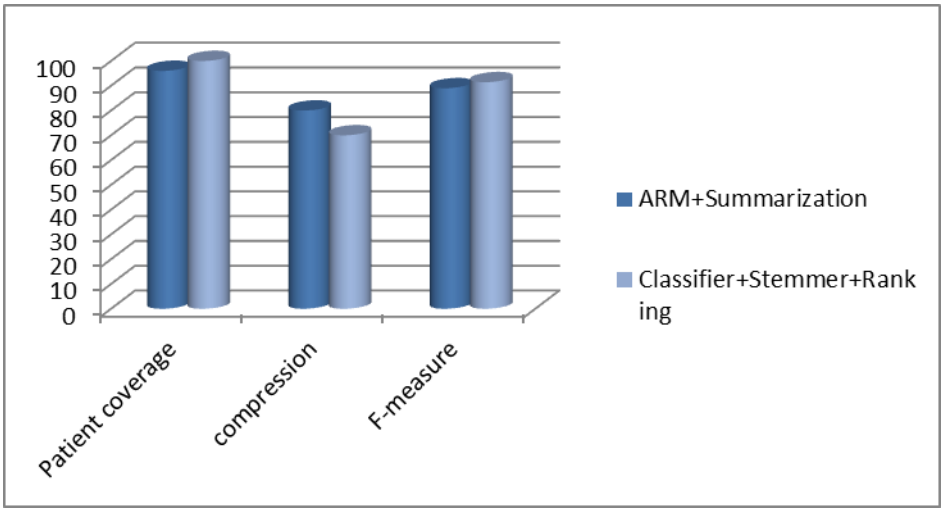


Fig. 2. Comparison of the Existing System and Proposed System.

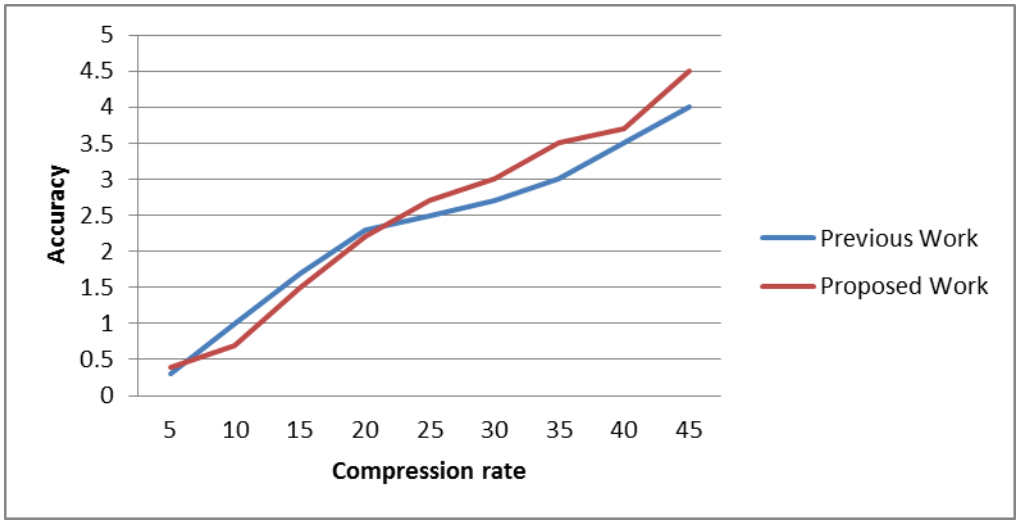


Fig. 3. Accuracy-compression rate.

The above graph Fig. 3. shows the accuracy prediction of cancer type with the risk level. The accuracy measure is actually dependent on the compression rate. In previous work, more summarization techniques are used with after with redundancies that may affect the quality of the result. In the proposed technique, stemming method is used for compressing the dataset which reduces more redundant rules than previous work [1]. Thus, from the evaluation result, the proposed method generates a correct and relevant result to the user query given. The relevant measure is more, as it result in better accurate system providing a guidance and awareness of the users who search for an assistance to cancer.

CONCLUSION AND FUTURE WORK

The result generated is in two parts, one is the possible cancer types that links to a detailed disease-treatment relationship, such as cure, counteractive action, side effects, best hospitals list and doctors to consult and another is the percentage of probability for each type of cancer retrieved. This output is very much useful for users to know their status of a disease by the probability count and a detailed guidance to be aware of it. Also the suffix stemmer utilized for both stripping and substitutions improves rate and rightness of stemming in this manner decreasing the memory space. Yet, it's restricted to some unpredictable structures and words. What's more, the Naive Bayesian Classification algorithm result in predicting the class of the query inputted that helps to provide the positive or the negative results. Bayesian Classification is utilized for better calculation time. Positioning the tenets gave preference of organized and applicable recovery of yield to the client keeping in mind the end goal to help them with exact and more pertinent result. Ranking with

weightage, provided an advantage of prioritized output retrieval to the user in order to assist them with accurate and more relevant result.

As a future work, more effective stemmer can be utilized as a part of request to handle the unpredictable and compound words. Persistent patient history of disease can be incorporated alongside the malignancy sorts in order to give a superior answer for the better treatment of the client.

REFERENCES

[1] Gyorgy J. Simon, Member, IEEE, Pedro J. Caraballo, Terry M. Therneau, Steven S. Cha, Regina Castro J, and Peter W. Li, "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus," in IEEE Transactions on Knowledge and Data Engineering, Vol. 27, No. 1, January 2015.

[2] Oana Frunza, Diana Inkpen, and Thomas Tran, "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts" in IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 6, June 2011.

[3] Craven M, "Learning to Extract Relations from Medline," Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.

[4] Gaizauskas R, Demetriou G, Artymiuk PJ and Willett P, "Protein Structures and Information Extraction from Biological Texts: The PASTA System," Bioinformatics, vol. 19, no. 1, pp. 135-143, 2003.

[5] Honglei Zhu and Zhigang Xu, "An Effective Algorithm for Mining Positive and Negative Association Rules," in International Conference on Computer Science and Software Engineering, 2008.

[6] Ginsberg J, Mohebbi Matthew H, Rajan SP, Lynnette B, Mark SS and Brilliant L, "Detecting Influenza Epidemics Using Search Engine Query Data," Nature, vol. 457, pp. 1012-1014, Feb 2009.

[7] Idheba Mohamad Ali Swesi O, Azuraliza Abu Bakar and Anis Suhailis Abdul Kadir, "Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets," in the 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012.

[8] Goadrich M, Oliphant L and Shavlik J, "Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction," Proc. 14th Int'l Conf. Inductive Logic Programming, 2004.

[9] Ramasubbareddy B, Dr. Govardhan A and Dr. Ramamohanreddy A, "Mining Positive and Negative Association Rules," in the 5th International Conference on Computer Science & Education Hefei, China, August 24–27, 2010.

[10] Shi-Ju Shang, Xiang-Jun Dong, Jie Li and Yuan-Yuan Zhao, "Mining Positive and Negative Association Rules in Multi-database Based on Minimum Interestingness," in International Conference on Intelligent Computation Technology and Automation, 2008.

[11] Rosario B and Hearst MA, "Semantic Relations in Bioscience Text," Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics, vol. 430, 2004.

[12] Yanguang Shen, Jie Liu and Zhiyong Yang, "Research on Positive and Negative Association Rules Based on "Interest-Support-Confidence" Framework," in IEEE 2009.

[13] Chandola V and Kumar V, "Summarization – Compressing data into an informative representation," Knowl. Inform. Syst., vol. 12, no. 3, pp. 355–378, 2006.

[14] Yin X and Han J, "CPAR: Classification based on predictive association rules," in Proc. SIAM Int. Conf. SDM, 2003.

[15] Shweta Kharya, "Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease," in International Journal of Computer Science, Engineering and Information Technology Vol.2, No.2, April 2012.