Indian Journal of Science and Technology, *Vol 9(4), DOI: 10.17485/ijst/2016/v9i4/87042, January 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Classification of Adverse Event Thyroid Cancer using Naïve Entropy and Association Function

K. V. Uma[1]* and S. Appavu alias Balamurugan[2]

[1]Department of Information Technology, Thiagarajar College of Engineering, Madurai – 625015, Tamil Nadu, India; kvuit@tce.edu
[2]Department of Information Technology, K.L.N. College of Information Technology, Sivagangai - 630612, Tamil Nadu, India; app_s@yahoo.com

Abstract

Background: Decision trees are a simple and powerful form of multiple variable analyses which allows predicting, explaining, describing, or classifying an outcome. The risk factors for Adverse event include demographic features of patients and concurrent illnesses, hypersensitivity to related drugs, drugs currently taken etc. **Methods:** The objective is to classify the Adverse event Thyroid cancer outcomes based on the risk factors. For that Decision tree based classifier model that uses Naïve entropy for calculating the information gain is proposed for classifying the adverse event .First the missing values in the dataset are handled using mean of nearby points. Along with that, Association function is used for determining the relative degree between the given attribute and class C. **Findings:** The proposed Classifier Model generates the If then Rules for adverse event outcome of Thyroid cancer. It considers only three attributes such as age, drug name and indication for using the drug for generating the tree structure. Hence the depth of the tree is reduced. The rules generated are grouped into categories .Then they are arranged in the descending order based on the number of occurrences of the rules. The top rules specify the major occurrence of adverse event for the combination of the attribute valus.Accuracy of the proposed classifier model is compared with that of J48, KNN and Naïve Bayes algorithm. The proposed model has better accuracy than that of other classifiers. **Application:** The different rules generated are stored in the database. In order to prevent adverse event thyroid cancer, the physicians can make use of this database and avoid medications in these combinations.

Keywords: Adverse Event, Association Function, Data Mining, Decision Tree, Entropy

1. Introduction

World Health Organisation has defined Adverse Drug Reaction, as a response to a drug that is noxious and unintended .And it occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or for modification of physiological function"[1]. The term Adverse Drug Event includes harm caused by the drug (adverse drug reactions and overdoses) and harm from the use of the drug (including dose reductions and discontinuations of drug therapy)"[2]. About 45,610 adverse drug events were reported in children less than 18 years of age. Out of these records about, 64% (29,298) indicated a serious injury. Adverse event in children grew over time from 6,320 in 2008 to 11,401 in 2012, increasing at the same rate as for adult patients. The rate was examined for year of age in children. The number was greatest in the first year of life, then declined and leveled off until adolescence, when cases again raised rapidly[3]. Adverse drug events can be prevented in older adults before hospitalization. Most emergency hospitalizations for an adverse drug events in older adults resulted were because of some of the commonly used medications. And some of the events resulted from medications are designated

*Author for correspondence

as high-risk or inappropriate. Proper medication of antithrombotic and antidiabetic drugs will reduce the hospitalizations for adverse drug events in older adults[4]. Adverse Drug Events may results from medication errors but not at all the case. Every drug introduced in the market has to undergo clinical trials in order to evaluate the safety issues of its usage. But the trials are limited in the number, duration, characteristics of patients exposed, and the type of data collected. The detailed and complete safety profile related with new drug cannot be fully established through clinical trials. The American Food and Drug Administration have defined an adverse event as serious, when the patients have any of the outcomes: 1.Death (DE), 2.Life-Threatening (LT), 3.Hospitalization (HO) (initial or prolonged), 4.Disability (DS), 5.Congenital Anomaly (CA), 6. Requires intervention to prevent permanent impairment or damage (RI), 7. Others (OT). The Agency for healthcare research and quality of U.S. Department of Health & Human Services reported that 28 percent to 95 percent of ADEs can be prevented by reducing medication errors through computerized monitoring systems. Computerized medication order entry will prevent 84 percent of dose, frequency, and route errors. Hence the related knowledge about adverse event will be used to make accurate decisions for the diagnosis and treatment of the outcomes of the adverse events.

2. Previous Work

The methods used to discover knowledge in Databases seems to be useful to analyze large clinical databases. In the Knowledge discovery process, the preprocessing step which comprises of data cleaning, handling of missing values, data transformation are important since it conditions the quality of the results obtained by data mining procedures. It consumes 80% of the whole project time. There are many tools and techniques available to analyze and handle inconsistent data and missing values. The preprocessing step was divided into 3 main stages such as data cleaning, explanatory study of missing values, choice of the procedure that has been used for handling missing values. The data cleaning stage was mainly focused on a system of logical rules to correct mistakes and on cluster analysis used in order to discard the poorly filled files. Multivariate statistical procedures were used to analyze the missing-data. Two methods, imputation by the most common value (mode) and imputation using decision trees that were used to deal with missing values were compared. A large medical diabetes database which comprises of 23,601 patients including numerous missing values was taken and applied with the two methods .A system constructed with logical rules were allowed for correcting mistakes on essential parameters (for example, the type of diabetes). Cluster analysis was able to identify only 10% of poorly filled files .For Dataset with low number of missing values (< 10%) and categories (< 4), imputation using decision trees provided better results than imputation by mode[5]. Classification methods such as Decision Tree Induction (DTI) can be modeled using training dataset for extracting rules to predict the outcomes of new patients. But the incompleteness of the data and high dimensionality of stored data was a problem while applying data mining methods. Another technique called as Canonical Correlation Analysis (CCA) was used prior to Decision tree Induction as a dimension reduction technique. It was used to preserve the feature of the original data by omitting non-essential data. Data from 3949 breast cancer patients were analyzed using CCA technique. Raw data were preprocessed by using a set of logical rules. Missing values were replaced by using the Expectation Maximization algorithm[6]. The characteristics of clinical data were analyzed. Then the sigmoid function was used to preprocess the original data. Self-organizing neural network was selected to model. After obtaining the modeling results, it was analyzed and compared with clinical diagnosis. The sigmoid function maintained the same geometry of raw data. The output generated by the modeling was almost similar with the clinical diagnosis. Preprocessing of clinical data has improved the quality of network input data and obstacles for further clinical data mining modeling were removed[7]. Association rule called as Unexpected Temporal Association Rules (UTARs) were used to describe association between the drug and adverse event. Interestingness measure called as residual-leverage and a novel case-based exclusion technique for its calculation was used to handle the unexpectedness of the event. It was combined with an event-oriented data preparation technique to handle the infrequency. Then an algorithm called MUTARC was used to find pairwise UTARs. It was applied to generate Adverse Drug Reaction (ADR) signals from real-world healthcare administrative databases[8]. Another important factor was to focus on developing data mining techniques for detecting adverse drug reactions related with drug-drug

interactions. In general, the occurrence of adverse drug reaction is unexpected and infrequent in health domain. To determine the relation between the drug-drug interactions n-wise unexpected temporal association rule was generated[9]. Another interestingness measure called as exclusive causal leverage, was used in the construction of experience based fuzzy RPD model. This measure was used to quantify the degree of association of a Casual Association Rule. This interestingness measure was designed to mask the undesirable effects caused by high frequency events. Hence this measure was applied to detect the causal associations between each of the three drugs (i.e., enalapril, pravastatin, and rosuvastatin) and the ICD9 codes[10]. There are associations termed as Multi-item Adverse Drug Event (ADE) associations which relate multiple drugs to multiple adverse events. One of the standards in pharmacovigilance was bivariate association analysis, where each single drug-adverse effect combination was studied separately. Multi-Item AADE associations was difficult to identify and it was prominent in pharmacovigilance studies. Here, an association rule mining was applied to the FDA's spontaneous Adverse Event Reporting System (AERS)[11]. The Adverse Event Reporting System (AERS) was an FDA database, which provides good information on voluntary reports of Adverse Drug Events (ADEs). In order to improve the mining capacity of AERS data for drug safety signal detection and promote semantic interoperability between the AERS and other data sources, normalization of data was done. The drug information in the AERS was normalized to RxNorm, which is a standard terminology source for medication. It uses a natural language processing medication extraction tool called as MedEx. Greedy algorithm was used to obtain class information from the National Drug File-Reference Terminology (NDF-RT). Adverse events were aggregated by mapping with the Preferred Term (PT) and System Organ Class (SOC) codes of Medical Dictionary for Regulatory Activities (MedDRA)[12]. By considering all the above references, in the proposed work the adverse event Thyroid cancer is considered. It is an uncommon type of cancer. Thyroid cancer is a disease, that occur when abnormal cells begin to grow in thyroid gland. There are situations where Thyroid cancer may occur due to wrong medications also i.e., it may be an adverse event also. So it is necessary to extract useful knowledge about the adverse event Thyroid cancer and make use of it such that further patients may not get affected.

3. Method

The steps needed in the construction of the system are given below (Figure: 1). Initially data collection is done. If needed they are integrated into a single file. Since a single adverse event is analyzed, the records related to that reaction are separated out. The missing values are handled. Then the preprocessed data is randomly split into test data and training data. Decision trees are constructed by using the different entropies. Then the accuracy of both the trees is calculated and compared with other classifier models. The steps involved in construction of a new classifier model, are given below.

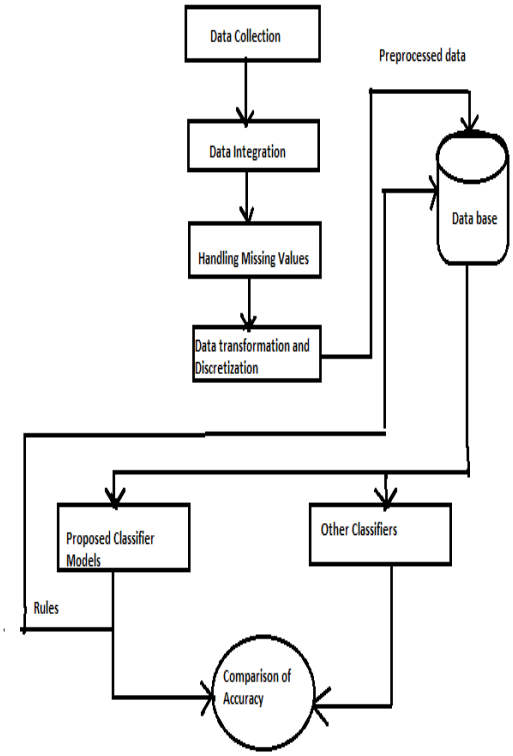

Figure 1. System Design.

Step 1: Data-preprocessing

a) Data-integration

The data relevant to adverse event are present in different files. They are integrated into a single file.

b) Handling Missing values

The main drawback in discarding the entire record having missing values may damage the reliability of the resulting classifier. Hence they are to be handled using proper techniques. Here the missing values are handled

using mean of nearby points (moving average). It replaces missing values with the mean of valid surrounding values. The span of nearby points is the number of valid values above and below the missing value used to compute the mean.

Step 2: The Naïve entropy is calculated as

$$Entropy(S) = H^{\wedge}N(Y) = \log n - \frac{1}{n}\sum_{k=1}^{k} h_k log h_k$$

The most commonly used entropy estimator is derived from empirical class probabilities. For each class k, we count the number of occurrences of that class as $h_k = \sum y \in YI(Y = k)$ where k is the no. of classes and n is the sum of all counts.

Step 3: Calculate the Information gain IG (V) using the formula:

$$IG = Entropy(S) - \sum_{v \in value(A)} \left(\frac{|S_v|}{|S|}\right) * Entropy(S_v)$$

Step 4: Calculate the Association Function (AF): Suppose A is an attribute of data set D, and C is the category attribute of D. the relation degree function between A and C can be expressed as:

$$AF(A) = \frac{1}{n} * \sum_{i=1}^{n} |X_{i1} - X_{i2}|$$

where X_{ij} indicates that attribute A of D takes the i-th value and category attribute C takes the sample number of the j-th value, n is the number of values attributes A takes.

Step 5: Calculate the normalization factor: Suppose that there are m attributes and each attribute relation degree function value is AF (1), AF (2) ….AF (m), respectively.

$$V(k) = \frac{AF(k)}{AF(1) + AF(2)...AF(m)}$$ for which $0 < k \leq m$.

Step 6: Calculate the new information gain:

New gain(S, A) = Gain(S, A)*V (k).

Now this new Gain can be used as a new criterion for attribute selection to construct decision tree.

Step 7: Construct the decision tree with the root node as the attribute which has the maximum information gain value.

Step 8: For the root attribute, if all class labels (values) belong to the same class, then it is the terminating condition.

Step 9: If there are different class label, then the tree is further branched with the next node as the attribute which has the next higher information gain value.

Step 10: The above step is repeated until the terminating condition holds.

4. Experiment and Results

The FDA Adverse Event Reporting System (FAERS)[13] is a database that contains information on adverse event and medication error reports submitted to FDA.

4.1 Data-Preprocessing

4.1.1 Data-Integration

The data is collected from FDA Adverse Event Reporting System. It contains the records related to adverse event. This dataset have five different files namely demographic file (patients demographic and administrative information), drug file (contains drug/biologic information for as many medications as were reported for the event), reaction file, outcome file (patient's outcome) and indications file (indications for the use of reported drugs). The patient records which have been reported the adverse event Thyroid cancer is retrieved from each file and these records are integrated into a single file. Then the reporter's personal information is removed in the integrated file. Only the patients information and the information about the medications such as drug name, indication for use of the drug, drug dosage etc. are considered.

4.1.2 Handling Missing Values

After integrating all the above mentioned tables, the Adverse Drug event dataset is found to have a large number of missing values. They are to be handled using proper techniques. Here the missing values are handled using mean of nearest point technique. The related patients' information is grouped together and the missing value of any of the attribute is filled with the mean value of that attribute in that group.

4.1.3 Data Transformation and Discretization

The data are to transformed or consolidated into forms appropriate for using the data mining techniques. Here, the attribute weight is given in different units such as lbs and kgs. Hence the weight values in lbs are converted to kgs for easy manipulation. Similarly age values are given in months, years and days. The values in months and days are converted into years. Discretization refers

to the process of converting or partitioning continuous attributes, features or variables to discretized or nominal attributes/features/variables /intervals. In the ADR dataset the variable wt (weight) and age has number of distinct values. It becomes impossible to visualize most of the branches in the decision tree and the number of rules generated from the algorithm will also increase considerably. All the other attributes have only few discrete values expect drug name which cannot be discretized. Hence the attributes age and weight are discretized. They are discretized using Class attribute interdependency maximization algorithm. The age is split into 4 intervals. They are 0 to 9 years, 10 to 33 years, 34 to 62 years and 62 to 86 years.

4.2 Training and Test Set

The dataset is randomly split into two parts called the training set and the test set. The training set is used to construct the classifier to discover potentially predictive relationships. The test is used to assess the strength and utility of the predictive relationship. In other words the classifier constructed from the training set is used to predict the classification for the instances in the test set. From the preprocessed dataset, about 80% data are considered as a training set for constructing the classifier model and 20% of the data are considered as a test data.

Table 1. Comparison of new classifier model with other classifiers

Measures	Proposed Classifier model (Uses Naïve Entropy)	J48	KNN	Naïve bayes
Accuracy	67.8%	43.40%	52.67%	48.75%
Precision	HO=0.486 DE=0.346 DS=0.10 LT=0.12 OT=0.813 RI=0.19	HO=0.441 DE=0.392 DS=0.18 OT=0.457 LT=0.271 RI=0	HO=0.335 DE=0.245 OT=0.347 LT=0.321 RI=0	H0=0.664 DE=0.391 DS=0.271 OT=0.587 LT=0.354 RI=0.195
Recall	HO=0.67 DE=0.565 DS=0.15 LT=0.07 OT=0.701 RI=0.123	HO=0.555 DE=0.296 DS=0.071 OT=0.448 LT=0.106 RI=0	HO=0.451 DE=0.345 DS=0=0.653 OT=0.478 LT=0.165 RI=0.123	HO=0.324 DE=0.744 DS=0.931 OT=0.525 LT=0.593 RI=0.949

4.3 Implementation of the Proposed Classifier Model

After the data is split into testing set and training set, the Naive entropy is calculated which gives the measure of impurities. Then the Information gain for all the attributes are found out. Then the association function value is calculated. From this, the new gain value is found out. The attribute drug name has higher gain value. It forms the root of the tree. Then the tree is grown by using the gain value of attributes in the decreasing order.

4.4 Rule Generation

Generating rules enables easier visualization of decision trees. A rule has two parts, an antecedent (if) and a consequent (then). The consequent part is the outcome that is found by the combination of antecedents. Some of the rules generated by Classifier model are given below.

1) IF DRUGNAME='HUMIRA' AND AGE ='62 TO 86' AND "INDI_PT='RHEUMATOID ARTHRITIS' THEN OUTC_COD='HO. (COUNT=12)

The drug Humira taken by the patients of age group 62 to 86, and indication for use of drug is 'RHEUMATOID ARTHRITIS lead to the outcome of Hospitalisation for 12 patients. The following rules are similar to this.

2) IF DRUGNAME='HUMIRA' AND AGE ='34 TO 62'AND INDI_PT='ANKYLOSING SPONDYLITIS' AND THEN OUTC_COD='HO'. (COUNT=16)
3) IF DRUGNAME='HUMIRA' AND INDI_PT='ANTIDEPRESSANT THERAPY' THEN OUTC_COD='OT'. (COUNT=6)
4) IF DRUGNAME='ZOMETA' AND AGE = '62 TO 86' AND INDI_PT='ATRIAL FIBRILLATION' THEN OUTC_COD='DE'. (COUNT=4)
5) IF DRUGNAME='ZOMETA' AND AGE='34 TO 62' AND INDI_PT='OSTEOPOROSIS' THEN OUTC_COD='OT'. (COUNT=24)
6) IF DRUGNAME='CELEXA' AND AGE='34 TO 62' AND INDI_PT='DEPRESSION' THEN OUTC_COD='OT'. (COUNT=2)
7) IF DRUGNAME='HYDROCHLOROTHIAZIDE' AND AGE ='34 TO 62' AND INDI_PT= 'LEFT VENTRICULAR DYSFUNCTION' THEN OUTC_COD=HO. (COUNT=15)
8) IF DRUGNAME='METHOTREXATE'AND AGE='34 TO 62' AND INDI_PT= 'RHEUMATOID ARTHRITIS' THEN OUTC_COD='OT'. (COUNT=14)

9) IF DRUGNAME='METHOTREXATE'AND AGE='62 TO 86' AND INDI_PT=' RHEUMATOID ARTHRITIS' THEN OUTC_COD='DE'.(COUNT=2)
10) IF DRUGNAME='VICTOZA'AND AGE = '34 TO 62' AND INDI_PT=TYPE 2 DIABETES MELLITUS' THEN OUTC_COD='OT'.(COUNT=24)
11) IF DRUGNAME='VICTOZA' AND AGE ='34 TO 62' AND INDI_PT='OVERWEIGHT' THEN OUTC_COD='OT'. (COUNT=3)
12) IF DRUGNAME='VICTOZA'AND AGE='34 TO 62'AND INDI_PT='TYPE 2 DIABETES MELLITUS' THEN OUTC_COD='OT'. (COUNT=12)
13) IF DRUGNAME='VICTOZA'AND AGE='62 TO 86' INDI_PT='TYPE 2 DIABETES MELLITUS' THEN OUTC_COD='OT'. (COUNT=23)
14) IF DRUGNAME='SYNTHROID' and AGE ='34 TO 62' AND INDI_PT='THYROIDECTOMY' THEN OUTC_COD='HO'.(COUNT=8)
15) IF DRUGNAME='SYNTHROID' and AGE ='34 TO 62' AND INDI_PT='THYROIDECTOMY' THEN OUTC_COD='OT'. (COUNT=15)

The different rules generated are stored in the database. In order to prevent adverse event thyroid cancer, the physicians can make use of this database and avoid medications in these combinations.

4.5 Calculate the Accuracy, Precision and Recall

A confusion matrix is a specific table layout that allows visualization of the performance of an algorithm. It is used to determine how many instances of class X were correctly classified as class X or misclassified as some other class. The rows correspond to correct classifications and the column corresponds to predicted classifications. By using the Confusion Matrix, the accuracy of the proposed classifier model is calculated. It is of about 68.8%. Table 1 shows the comparison of proposed classifier model with other models. It shows that the proposed model have better classification accuracy than that of other models.

5. Conclusion

Decision tree based classifier model proposed for classifying outcomes of the adverse event thyroid cancer have better classification accuracy than that of other classifier models .The missing values in the dataset are handled by mean of nearby points. Naïve entropy provides more appropriate measure of impurities. Along with the Information gain, Association function is used for determining the relative degree between the given attribute and class C. By storing these rules in the database and by projecting to the public, these types of medications can be avoided.

6. References

1. International drug monitoring: the role of national centres. World Health Organization. Technical Report Series. 1972; 498:1–25.
2. Nebeker JR, Barach P, Samore MH. Clarifying Adverse Drug Events: A Clinician's Guide to Terminology, Documentation, and Reporting. Ann Intern Medicine. 2004; 140(1):795–801.
3. Moore TJ, Cohen MR, Furberg CD, Mattison DR. Adverse drug events in children under age 18. Special Report on Children Quarter Watch. Institute for Safe Medication Practices. 2014.
4. Budnitz DS, Lovegrove MC, Shehab N, Richards CL. Emergency Hospitalizations for Adverse Drug Events in Older Americans. The New England Journal of Medicine. 2002-12; 365.
5. Duhamel A, Nuttens MC, Devos P, Picavet M, Beuscart R. A preprocessing method for improving data mining techniques: Application to a large medical diabetes database. Studies in Health Technology and Informatics. 2003; 95:269–74.
6. Razavi AR, Gill H, Ahlfeldt H, Shahsavar N. A Data Preprocessing Method to Increase Efficiency and Accuracy in Data Mining. Artificial Intelligence in Medicine, Lecture Notes in Computer Science. 2005; 3581(1):434–43.
7. Ang Q, Liu ZW, Wang W, Li K. Explored research on data preprocessing and mining technology for clinical data applications. 2010 The 2nd IEEE International Conference on Information Management and Engineering (ICIME). Chengdu, 2010. p. 327–30.
8. Jin H, Chen J, He H, Williams GJ, Kelman C, O'Keef CM. Mining Unexpected Temporal Associations: Applications in Detecting Adverse Drug Reactions. IEEE Transactions on Information Technology in Biomedicine. 2008; 12(4):488–500.
9. Shanmugapriya K, Shanmugapriya D, Summia Parveen H, Niranjani V. N-Unexpected Temporal Association Rule for Diagnosing Adverse Drug Reaction from Health Database. International Conference on Information and Intelligent Computing IPCSIT. 2011; 18(1):158–61.
10. Ji Y, Ying H, Tran J, Dews P, Ayman Mansour R. Massanari M. A Method for Mining Infrequent Causal Associations and Its Application in Finding Adverse Drug Reaction

Signal Pairs. IEEE transactions on Knowledge and Data Engineering. 2012; 25(4):721–33.

11. **Harpaz R**, **Chase** HS, **Friedman C.** Mining multi-item drug adverse effect associations in spontaneous reporting Systems. BMC Bioinformatics. 2010; 11(9):7.
12. **Wang L**, **Jiang G**, **Li D**, **Liu H**. Standardizing adverse drug event reporting data. *Journal of Biomedical Semantics*. 2014; **5(36):1–13.**
13. Dataset. Available from: http: // www.fda.gov / Drugs / GuidanceComplianceRegulatoryInformation / Surveillance / ucm082193. 2015 May 10.