ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Data mining techniques for finding serious Adverse Events

G.V.Sriramakrishnan[1] and Latha Parthiban[2]
[1]Department Computer Science and Engineering, St Peter's University, Tamilnadu, India.
[2]Department Computer Science and Engineering, Pondicherry University Community College, Puducherry, India.
***Corresponding author: E-Mail:greatsri8@gmail.com**

ABSTRACT

Very fast growth in medical data needs huge electronic medical record forms that has patient's health history. Governments take necessary steps to gather the patient's health history to carry out research and be prepared for any disease outbreaks at large to the citizens. Research has shown that the disease outbreaks are due to the lifestyle, the living conditions and the treatment undergone during the past. Medical literature states that many drugs whose complete safety profile is unknown have been approved. Some drugs have shown serious adverse events, and subsequently withdrawn. There may be some drugs which still show adverse effects on the patients. This paper analyses the various data mining techniques to find adverse events.

KEY WORDS: Pharmacovigilance, Data mining, Adverse Events.

1. INTRODUCTION

In the present medical scenario medical records are being digitized and maintained as Electronic Medical Records (EMR) (Olsson, 2010; Forster, 2011; Kumar, 2013). The abundance of these records can be used for research on finding useful patterns relating to adverse effects, drug efficiency and related information. These reports can also give a view of the treatment process like any surgeries undergone and report any adverse events after the process. Drug use leading to better outcome is undoubtly accepted by all but there are some side effects also in the present days. These side effects are sometimes equally dangerous and may be fatal. These effects are being identified by the present system and a study on the medication process is being carried out.

In India the main drawback is the lack of data. Only in the present scenario are the hospitals digitizing the data. There is a severe resistance to adopt the new technology. The medical data is very specific (Yeleswarapu, 2014; Sarker, 2015; Benton, 2011). To mine medical data all information should be converted into numeric values. The specificity of the medical data lies in the fact that the attributes' (symptoms') values usually come from certain ranges (Harpaz, 2014; Abbasi, 2014; Nikfarjam, 2011; Tuarob, 2014; Marie Dupuch, 2015). The paper is organized as follows: Section 2 discusses the background work; section 3 discusses on data mining process for Pharmacovigilance; section 4 discusses the algorithms for medical applications; section 5 discusses the proposed work and section 6 the conclusions.

Literature review: The aim of pharmacovigilance is early detection of adverse drug events so that proper treatment is given as early as possible. Many Spontaneous Reporting System (SRS) database are available to identify these adverse events. The structure of FDA_AERS database structure is shown in figure 1. ADRs are caused by unexpected interaction of drugs with patients downstream pathway perturbations (Liebler, 2005). Most common belief is its interaction with resulting in ADR (Marie Dupuch, 2015; Liebler, 2005; Blagg, 2006). Fliri (2005), demonstrated that drugs having analogous *in vitro* protein binding had related side-effects which was confirmed by Campillos (2008). Scheiber (2009), also proved this by linking pathways by toxic compounds vs. those affected by nontoxic compounds. Fukuzaki (2009), proposed a method to predict ADRs using sub-pathways that share correlated modifications of gene-expression profiles in the presence of the drug of interest. Table 1 shows the frequently used semantic distance measure.

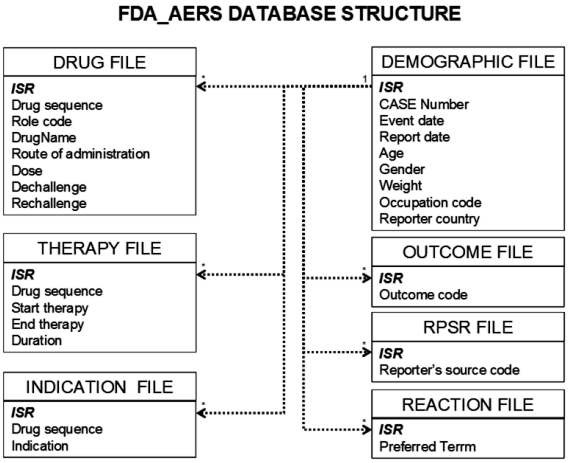

Figure.1.FDA_AERS structure

Most frequently used semantic similarity and distance algorithms. SP stands for shortest path, NCP stands for the nearest common parent, IC stands for information content. The + means that the technique mentioned in a given column is used in a given reference from the first column.

Table.1. Frequently used semantic distance measures

Measures	Resource	SP	NCP	Depth	Density	IC
Rada (1989)	MeSH	+	-	-	-	-
Sussna (1993)	WordNet	+	-	+	+	-
Zhong (2002)	Conceptual graphs	+	+	+	-	-
Wu and Palmer (1994)	WordNet	+	+	+	-	-
Jarmasz &Szpakowicz(2003)	Roget's Thesaurus	+	-	-	-	-
Resnik (1999)	WordNet	-	+	+	-	+
Leacock and Chodorow (1998)	WordNet	+	-	-	-	-
Jiang & Conrath (1997)	WordNet	+	+	+	+	+
Lin (1998)	WordNet	-	+	+	-	+
Hirst and st. Onge (1998)	WordNet	+	-	-	-	-
Steichen (2006)	Medical Ontology	-	+	+	+	+
Cho (2003)	WordNet	+	+	+	+	+
Yang (2005)	WordNet	-	-	+	-	-

SP:Shortest Path; NCP: nearest common parent; IC: information content

Data mining process for Pharmacovigilance: The data mining process for Pharmacovigilance using FDA Adverse Event Reporting System is shown in figure 2. FAERS database contains facts on adverse events reported from throughout the world. The mapping of data of this FDA_AERS is done using standard WHO drug dictionary. After mapping missing data are handled and duplicates are removed from which signal interpretation take place.

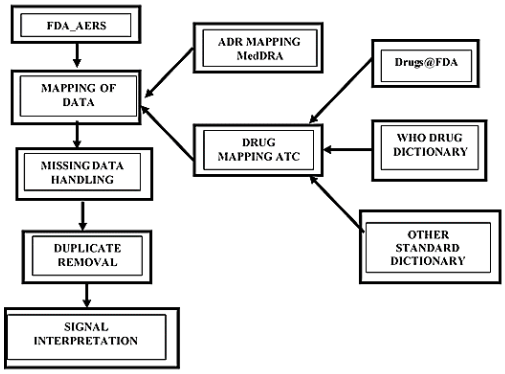

Figure.2. Data mining process

Data mining algorithms for medical applications: The main process involved in data mining is the usage of efficient algorithm. Two important standard algorithms ID3 and C4.5 are presented as follows.

The ID3 algorithm is presented as follows:

Given: $Sampl$ - the set of training examples ($Sampl \subseteq T$, T is decision table), d_k- the attribute which value is to be predicted by the tree, $S = \{s_1, \ldots, s_l\}$ the set of symptoms

Results: The decision tree $Tree$

BEGIN

1. If all examples are positive, Return the single-node tree $Root$, with label = + END
2. If all examples are negative, Return the single-node tree $Root$, with label = - END
3. If number of predicting attributes is empty, then Return the single node tree $Root$, with label = most common value of the target attribute in the examples END
4. Set $s_g := s_i$ the attribute with the highest $Gain(Sampl, s_i)$
5. Set s_g as a $Root$ of the decision tree
6. For each $v \in Values(s_g)$ DO 7-10
7. Add a new tree branch below Root, corresponding to the test $s_g = v_i$,
8. Let $examples(v_i)$, be the subset of examples that have the value v_i for s_g.
9. If $examples(v_i)$ is empty Then below this new branch add a leaf node with label = most common target value in the examples END
10. Below this new branch add the subtree ID3 ($examples(v_i)$, d_k, $S = \{s_1, \ldots, s_l\} \backslash s_g$)

END

Return $Root$

A single tree C4.5 generates classification rules from many decision trees. Many medical applications use C4.5 as missing values are handled by this algorithm

Given: $Sampl$ – the set of training examples ($Sampl \subseteq T$, T is decision table), d_k- the attribute which value is to be predicted by the tree, $S = \{s_1, \ldots, s_I\}$ the set of symptoms

Results: The decision tree $Tree$

BEGIN

1. For $1 \le k \le K$ compute decision frequencies $freq(d_k, Sampl)$
2. If all cases in $Sampl$ belongs to the same class d_k set the node is a leaf with associated class d_k.
3. Else set as a node the most frequent class d_k, and count the classification error a of the leaf is the weighted sum of the cases in $Sampl$, whose class is not d_k.
4. For $s_i, i \in (1, I)$ count $Gain(Sampl, s_i)$
5. Set $s_g := s_i$ the attribute with the highest $Gain(Sampl, s_i)$
6. If s_g is continuous find $Threshold$
7. For each $Sampl'$ in the splitting of $Sampl$ DO 8-9
8. If $Sampl' = \emptyset$ set as a child of s_i is a leaf
9. Else below this new branch add the subtree C4.5 for d_k and $S = \{s_1, \ldots, s_I\} \backslash s_g$
10. Compute errors of s_g

3. RESULTS OF DATA MINING ON FDA_AERS

Experimental results (Raschi, 2013) on FDA_AERS has shown that lot of duplicates and missing values are reported which has made it difficult to continue with analysis. Here analysis is done using drug-event pair based on drug event pair. Disproportionality is calculated in (Raschi, 2013) using cumulative strategy where noting of exact time point and complete assessment of adverse reports are evaluated cumulatively.

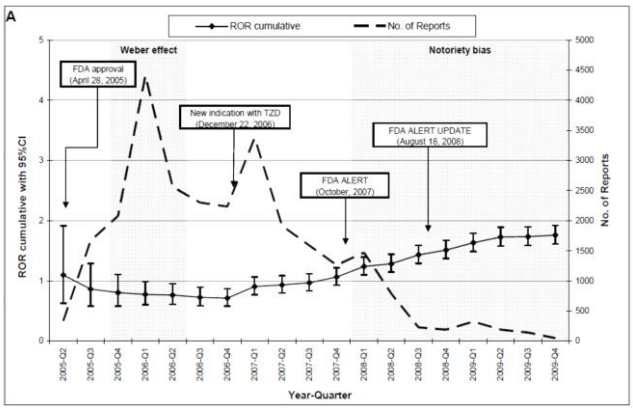

Figure.3. Time trends of ROR with 95%CI of Pancreatitis associated with exenatide use

4. CONCLUSION

With adverse effects on huge population available in FDA_AERS, researchers are looking into new ways to signal detection in pharmacovigilance. Proper analysis of FDA_AERS can help in saving patients precious life using stable drug safety assistance. Recently, plenty of work has been done to increase the limits of pharmacovigilance. Data mining algorithms helps in improving the accuracy of analysis but overall, results obtained from these algorithms should be measured with care and directed by proper clinical assessment. The review work done in this paper will help us to work on missed data and duplicate data removal after which analysis is planned to be done for different adverse effects.

REFERENCES

Abbasi A and Adjeroh D, IEEE Intelligent Systems, 2014, 60–80.

Benton A, Ungar L, Hill S, Hennessy S, Mao J and Chung A, Journal of Biomedical Informatics, 44, 2011, 989–996.

Blagg J, Annual Reports in Medicinal Chemistry, 41, 2006, 353-368.

Campillos M, Kuhn M, Gavin AC, Jensen LJ and Bork P, Science, 321(5886), 2008, 263-266.

Cho M, Choi J, Kim P, 4th International Conference in Advances in Web-Age Information Management, 2003, 381–388.

Fliri A F, Loging W T, Thadeio P F and Volkmann A, Nature Chemical Biology,1(7), 2005, 389-397.

Forster A J, Jennings A, Chow C, Leeder C, van Walraven C, Journal of the American Med Informatics Association, 19, 2011, 31– 38.

Fuzuzaki M, Seki M, Kashima H, and Sese J, IEEE International Conference on Bioinformatics and Biomedicine, 2009, 142-147.

Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, Jung K, LePendu P and N H Shah N H, Drug Safety, 37(10), 2014, 777-790.

Hirst G, St. D Onge D, Wordnet, An electronic lexical database, 1998, 305–332.

Jarmasz M and Szpakowicz S, Recent advances in natural language processing, 2003, 212– 219.

Jiang J and Conrath D, Research in computational linguistics, 1997, 19–33.

Kumar V, Journal of Pharmacovigilance, 1(1), 2013, 1–3.

Leacock C and Chodorow M, Wordnet: An electronic lexical database, 49(2), 1998, 265-283.

Liebler D C and Guengerich F P, Nature Reviews Drug Discovery, 4 (5), 2005, 410-420.

Lin D, International conference on machine learning, 1998, 296–304.

Marie Dupuch and Natalia Grabar, Journal of Biomedical Informatics, 54, 2015, 174–185.

Nikfarjam A, Gonzalez G, Proceedings of the American medical informatics association Annual symposium, 2011, 1019–1026.

Olsson S, Pal SN and Stergachis A and M Couper M, Drug Safety, 33(8), 2010, 689–703.

Rada R, Mili H, Bicknell E, M Blettner M, IEEE Transactions Systems, Man Cybernetics, 19(1), 1989, 17–30.

Raschi E, Piccinni C, Poluzzi E, Marceline G and De Ponti F, The association of pancreatitis with antidiabetic drug use: gaining insight through the FDA pharmacovigilance database. Acta Diabetologica, 50(4), 2013, 569 -577.

Resnik P, Journal of Artificial Intelligence Research, 11, 1999, 95–130.

Sarker A and Gonzalez G, Journal of Biomedical Informatics, 53, 2015, 196–207.

Scheiber J, Chen B, Milik M, Sukuru S C, Bender, Mikhailov D, Whitebread S, Hamon J, Azzaoui K, Urban L, Glick M, Davies J W, and Jenkins J L, Journal of Chemical Information and Modelling, 49(2), 2009, 308-317.

Steichen O, Daniel-Le Bozec C, Thieu M, Zapletal E, Jaulent M, Computers in Biology and Medicine, 36(7), 2006, 768–88.

Sussna M, Proceedings of the 2nd International conference on Information and Knowledge Management, 1993, 67–74.

Tuarob S, Tucker C S, Salathe M and Ram N, Journal of Biomedical Informatics, 49, 2014, 255–268.

Wu Z and Palmer M, Proceedings of associations for computational linguistics, 1994, 133–138.

Yang D and Powers D M W, Proceedings of the 28th Australasian computer science conference, 2005, 315–322.

Yeleswarapu S, Rao A, Joseph T, Saipradeep V G, Srinivasan R, BMC Medical Informatics and Decision Making, 14(13), 2014.

Zhong J, Zhu H, Li J and Yu Y, 10th International conference on conceptual structures, ICCS 2002. LNCS 2393. Springer Verlag, 2002, 92–106.