Int J Pharm Bio Sci 2013 Apr; 4(2): (B) 1281 - 1286

Research Article **Bioinfomatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

AN ENHANCED CLUSTERING ALGORITHM IMPLEMENTED ON BIOLOGICAL DATA IN DATA MINING

DR.T.NALINI [*1] AND S.REVATHI[2]

[1]Professor, Department of computer science and engineering, Bharath University, Chennai, India
[2] PG Scholar, Department of computer science and engineering, Bharath University, Chennai, India

ABSTRACT

Bioinformatics is defined as the use of computer science and information theory to analyze biological systems. When people are injected vaccines, they may feel some side effects. Vaccine Adverse Event Reporting System (VAERS) is an awareness program launched in United States to monitor the people regarding their post vaccination period. VAERS is used to track the adverse events such as their disease, symptom, location of origin, age, gender, etc associated with these vaccines. KMeans clustering algorithm is used to cluster the similar data items which helps the active biologists of VAERS to gather relevant information about their patients. It's very helpful to create awareness among the public regarding those vaccines. After each monitoring, the conditions of these patients are updated in the VAERS database. There are few queries listed below, that provides the similarity that exists between the patients of VAERS.

KEYWORDS: Biological data, VAERS, Data mining, KMeans clustering

DR.T.NALINI
Professor, Department of computer science and engineering,
Bharath University, Chennai, India

Corresponding author

INTRODUCTION

Bioinformatics[13] is an energetic field where the algorithms and methods of computer science are employed to solve the problems of biologists. Bioinformatics[2] is a unified discipline formed by the combination of biology, computer science and information technology. Information technology is used to manage and to perform analysis on the biological data. The two important functions of bioinformatics are storage and analysis. Biological data[8] is read using some complex machines and they are processed using algorithms of artificial intelligence, soft computing, data mining, etc. These data are stored in databases. They can be retrieved and used whenever needed and stored back into the database. Bioinformatics[1] was preceded by an area called computational biology. Computational biologists[11] shared their ideas to form algorithms that solve problems regarding biological data. They develop ideas related to computer science, where the algorithms developed are reliable. For example, when a large set of biological data are analyzed, it involves human-computer interaction because the needs of the biologists are very broad and complex.Data mining or knowledge discovery is the computer-oriented process of digging and analyzing large volumes of data and finally extracting the meaning of the data. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining approaches seem ideally suited for Bioinformatics[7]. The extensive databases of biological information create both challenges and opportunities for development of novel Knowledge Discovery in Databases (KDD) methods. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience.Clustering[3] is the one of the useful technique in data mining. Clustering can also be said as an unsupervised learning technique. Clustering forms a structure from the collection of data. Objects present in a cluster are similar to the objects of the same cluster and dissimilar to the objects of the other clusters. Few applications of clustering are marketing, insurance, city-planning, earth-quake studies, etc.The Vaccine Adverse Event Reporting System (VAERS[6]) is a national vaccine safety surveillance program co-sponsored by the Centers for Disease Control and Prevention (CDC) and the Food and Drug Administration (FDA). VAERS provides biological data regarding the patients registered, their symptoms, disease affected, and details of being hospitalized. These data are just maintained without being categorized. If these data are well categorized, the registered patients of VAERS can be given special preference. Thus, the diseases, symptoms, age category of patients, the location from where they originate, their condition after admission are all clustered here.

MATERIALS AND METHODS

(i) Simple KMeans Clustering:

KMeans[4] is an iterative clustering algorithm in which items are moved among a set of clusters until the desired set is reached. This can be viewed as a type of squared error algorithm.

The cluster mean of $K_i=\{t_{i1}, t_{i2}, \ldots\ldots, t_{im}\}$is defined as,

$$m_i = \frac{1}{m}\sum_{j=1}^{m} tij \qquad (1)$$

where $t_{i1}, t_{i2}\ldots t_{im}$ is the set of input elements, K is the desired number of clusters and m is the cluster mean obtained.A clustering algorithm finds some useful groups of components based on certain similarity. Also, the clustering algorithm attempts to find the centroid of a group of data sets. Most algorithms evaluate the distance between a point and the cluster

centroid. The output of any clustering algorithm is the statistical expression of the cluster centroid with the total number of elements in each of the cluster. The *k*-Means algorithm[5] is a distance-based clustering algorithm that partitions the data into a predetermined number of clusters. In almost all cases, the simple KMeans clustering algorithm takes more time to form clusters. So it is not suitable to be employed for large datasets.

(ii) VAERS Data:

Many medical researchers make use of VAERS[9] to study the effects of vaccination. In the present study, VAERS data is obtained to cluster the similarity that exists between the biological data. The vaccine production can be made faster and appropriate by clustering the similar types of patients based on certain criteria. So, the biologists are being helped in the preparation of vaccines. The following are some queries that are answered in this article, which will surely help the biologists in taking care of the registered patients of VAERS.

Query 1: The type of disease attacked for patients registered with VAERS.
Query 2: The types of symptoms each registered patient of VAERS have experienced.
Query 3: The origin or location of the registered patients of VAERS.
Query 4: The total number of patients who got recovered after being registered with VAERS.
Query 5: The total number of patients who died after being registered with VAERS.
Query 6: The sources of fund allocated to the registered patients of VAERS.
Query 7: The major portion of gender that is being registered for VAERS.
Query 8: The total number of patients who are disable in VAERS.
Query 9: The age group of registered patients of VAERS.

These queries are answered below, which can be used by the active biologists to help them in the production of vaccines for the registered patients of VAERS.

RESULTS AND DISCUSSIONS

(i) Experimental setup

VAERS[12] maintains a database of the people who are being adversely affected by the injected vaccines. The VAERS database is updated each year and the most recent updated data is considered for this research work. The primary objectives of VAERS are to detect new, unusual, or rare vaccine adverse events (VAEs), monitor increases in known adverse events, identify potential patient risk factors for particular types of adverse events, identify vaccine lots with increased numbers or types of reported adverse events and assess the safety of newly licensed vaccines.Weka is a collection of open source ML algorithms used for pre-processing, classifiers, clustering, and association rule. Weka is created by researchers at the University of Waikato in New Zealand. It is a Java based tool used in the field of data mining. It uses flat text files to describe the data. It can work with a wide variety of data files including its own ".arff" format and C4.5 file formats.The queries[10] mentioned above are solved here, which share a great knowledge with the biologists to assist them in the production of vaccines that will save the life of the registered patients of VAERS.

Query 1: The type of disease attacked for patients registered with VAERS.

Among the 16,792 registered patients of VAERS, each individual patient is affected with a different disease. So, they form 16,792 clusters. Each patient is registered with an adverse disease which is brought into the concern of VAERS. These details are updated onto the VAERS database. This database is updated is each year once.

Query 2: The types of symptoms each registered patient of VAERS have experienced.

Each patient of VAERS is affected with a different disease. So, obviously their symptoms are also independent. They all form an individual cluster resulting in a total of 16,792 clusters. These details are updated onto the

VAERS database. This database is updated is each year once.

Query 3: The origin or location of the registered patients of VAERS.

Among 16,792 registered patients of VAERS, the origin or location of them can be categorized into 54 different clusters say GA, VA, OR, AZ, LA, MD... The results show that, the registered patients of VAERS originate from different parts of the world.

Query 4: The total number of patients who got recovered after being registered with VAERS.

Table 1
Number of patients recovered After registering in VAERS

Cluster id	No of instances	% of cluster	Recovered
0	6273	37	Y
1	6785	40	U
2	3733	22	N

As shown in table 2, among the total 16,792 registered patients of VAERS 6,273 patients have recovered from their disease, 3,733 patients have not recovered and 6,785 patients are in unknown condition. The percentage of these clusters can be stated as 37% have recovered, 22% have not recovered and 40% are in unknown status. Also they are initialized with cluster id's 0 for recovered, 1 for unknown condition and 2 for not recovered.

Query 5: The total number of patients who died after being registered with VAERS.

Table 2
Number of patients died after registering in VAERS

Cluster id	No of instances	% of cluster	Died
0	9408	56	N
1	7383	44	Y

As shown in table 1, among the total 16,792 registered patients of VAERS 9,408 patients are alive and 7,383 patients have expired. The percentage of these clusters can be stated as 56% are alive and 44% have expired. Also they are initialized with cluster id's 0 and 1.

Query 6: The sources of fund allocated to the registered patients of VAERS.

Table 3
Different sources of fund offered to patients in VAERS

Cluster id	No of instances	% of cluster	Fund type
0	5818	35	PUB
1	1386	8	UNK
2	1205	7	MIL
3	2799	17	PVT
4	5583	33	OTH

As shown in table 3, among the total 16,792 registered patients of VAERS, 5 different sources of funds are being offered to produce vaccine for the patients registered with VAERS. 5,818 patients re offered fund from public, 2,799 patients are offered private fund, 1,205 patients are allotted military fund, 1,386 patients are offered fund from unknown sources and finally 5,583 patients are offered other different types of funds. The percentage of these clusters can be stated as 35% of public fund, 17% of private fund, 7% of the military fund, 8% of unknown fund and 33% of other sources of funding. Also they are initialized with cluster id's 0 for public fund, 1 for unknown sources of fund, 2 for military fund, 3 for private fund and 4 for other sources of fund.

Query 7: The major portion of gender that is being registered for VAERS.

Table 4
Gender of patients who are in need of vaccines

Cluster id	No of instances	% of cluster	Gender
0	8252	49	F
1	5803	35	U
2	2736	16	M

As shown in table 4, among the total 16,792 registered patients of VAERS, the major gender which is affected with many types of diseases is Female. 8,252 registered patients of VAERS are female, 2,736 patients are male and 5,803 patients are of unknown gender. The percentage of these clusters can be stated as 49% are female, 16% are male and 35% are of unknown gender. Also they are initialized with cluster id's 0 for female, 1 for unknown gender, 2 for male.

Query 8: The total number of patients who are disable in VAERS.

Table 5
Number of patients who are disable in VAERS

Cluster id	No of instances	% of cluster	Disable
0	5890	35	Y
1	10901	65	N

As shown in table 5, among the total 16,792 registered patients of VAERS 5,890 patients are physically disable and 10,901 patients are not physically disable. The percentage of these clusters can be stated as 35% of patients are physically disabled and 65% of patients are normal. Also they are initialized with cluster id's 0 for physically disable and 1 for normal patients who are not disable.

Query 9: The age group of registered patients of VAERS.

Among 16,792 registered patients of VAERS, the two major age groups are below 10 and above 50. This age group of people are highly affected by adverse diseases. They need to be cared more and their vaccines are to be detected faster.

CONCLUSION

Thus the queries are solved thereby helping the active biologists of VAERS to group similarly patients. The patients with similar symptom,

disease, age, location, disability, gender, etc are all clustered. This data can be used by the active biologists to care their patients. KMeans clustering algorithm is used to cluster the similar data items which helps the active biologists of VAERS to gather relevant information about their patients. It's very helpful to create awareness among public regarding those vaccines. Thus after each monitoring, the conditions of these patients are updated in the VAERS database.

ACKNOWLEDGEMENT

The authors would thank the management of Bharath University for their support and the Department of Computer Science and Engineering for their encouragement towards the research work.

REFERENCES

1 Li, J.; Wong, L. and Yang, Q. (2005). Data Mining in Bioinformatics, IEEE Intelligent System, IEEE Computer Society.

2 Hirschman, Lynette; C. Park, Jong; T., Junichi, Wong, L. and H. Wu., Cathy (2002) Accomplishments and challenges in literature data mining for biology, BIOINFORMATICS REVIEW, Vol. 18 no. 12, 1553–1561.

3 Han and Kamber (2006). Data Mining concepts and techniques, Morgan Kaufmann Publishers.

4 Eduardo Raul Hruschka, Ricardo J. G. B. Campello, Alex A. Freitas, and Andre C. Ponce Leon F. de Carvalho ," A Survey of Evolutionary Algorithms for Clustering", *IEEE Trans. Syst., Man, Cybern.—Part C: Appl. And Review,* Vol. 39, No. 2,PP.133-155,(2009).

5 R. Xu and D. Wunsch, "Survey of Clustering Algorithms", "IEEE Transactions on Neural networks", vol. 16, no. 3, May (2005).

6 Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection, J Am Med Inform Assoc; 18:631e638, doi:10.1136/amiajnl-2010-000022, (2013).

7 Tobias Scheffer and Ulf Leser, Data Mining and Text Mining for Bioinformatics, Proceedings of the European Workshop on Data Mining and Text Mining for Bioinformatics, (2003).

8 N.M. Luscombe, D. Greenbaum, M. Gerstein, Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, USA, Yearbook of Medical Informatics, (2001).

9 Krallinger, M., F. Leitner, et al. "Analysis of biological processes and diseases using text mining approaches." Methods in Molecular Biology 593: 341-382, (2010).

10 Sophia Ananiadou and John McNaught (editors) ,Text Mining for Biology and Biomedicine University of Manchester and UK National Centre for Text Mining) Boston and London: Artech House, xi+286 pp; hardbound, ISBN 1-58053-984-X, (2006).

11 Altman R et al. Text mining for biology — the way forward: Opinions from leading scientists. Genome Biology, 9(Suppl. 2): S7, (2008).

12 http://vaers.hhs.gov/data/vaersdatafiles/2012VAERSDATA.csv

13 Milind sale, Bioinformatics: History and Use, N. B. Navale college of commerce, lonavala, Proceedings of National Conference on Computational Neuroscience, International journal of Pharma and Bio Sciences, (2011).