Asian Journal of Information Technology 5 (4) : 413-417, 2006

Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Technique

[1]Kuttiannan Thangavel,[2]P. Palanichamy Jaganathan and [3]P.O. Easmi
[1]Department of Mathematics, Gandhigram Rural Institute -Deemed University, Gandhigram-624 302, Tamilnadu, India
[2]Department of Computer Science and Applications, Gandhigram Rural Institute-Deemed University, Gandhigram -624 302, Tamilnadu, India
[3]Radiation Oncologist, Christian Fellowship Community Health Centre, Ambilikkai-624 612, Tamilnadu. India

Abstract: Data mining is an umbrella term referring to the process of discovering patterns in data, typically with the aid of powerful algorithms to automate part of the search. These methods come from the disciplines such as statistics, machine learning (Artificial Intelligence), pattern recognition, neural networks and databases. In particular this paper reveals out how the problem of cervical cancer diagnosis is approached by a data mining analyst with a background in machine learning. Application areas for this problem include analysis of telecommunications systems, discovering frequent buying patterns, analysis of patient's medical records, etc. In the health field, data mining applications have been growing considerably as it can be used to directly derive patterns, which are relevant to forecast different risk groups among the patients. To the best of our knowledge data mining technique such as clustering has not been used to analyse cervical cancer patients. Hence, in this paper we made an attempt to identify patterns from the database of the cervical cancer patients using clustering.

Key words: Data mining, cervical cancer, patterns, knowledge discovery, clustering, k-means

INTRODUCTION

In a sense, cancer is easy to detect- just kill the patient, perform through autopsy and you can likely to discover any existing cancer. Ofcourse, this detection method defeats the purpose, but it serves to illustrates the fact that cancer detection is a spectrum from maximally invasive, expensive and accurate methods to non-invasive, relatively inexpensive and possibly accurate methods. This paper presents an approach to predict non-linear groups in cervical cancer of patient records. The main aim is to discover patient groups at high risks of cervical cancer disease. Cancer of the cervix is having a devastating impact on woman's health around the world, especially in developing countries, where it is the most common cancer and the leading cause of death from cancer in woman. It is estimated that 500000 new cases occur every year world wide, the majority (80%) being in the developing world[1]. Although cervical cancer is preventable disease, it still remains a major burden on public health resources in India.

Accurate data on the magnitude of the morbidity and mortality from cervical cancer in the developing countries are scanty and are usually hospital based. Cancer registration in most of the low resource countries is difficult because of inadequate sources of information within the health delivery systems and lack of reliable population data for estimating the accurate incidence rates. To facilitate the proper planning of services for the prevention, early diagnosis and treatment of cervical cancer, the prediction of patterns in cervical cancer patients using data mining was carried out[2].

Cancer has become one of the ten leading causes of death in India. It is estimated that there are nearly 2- 2.5 million cancer cases at any given point of time. Over 7 lakh new cases and 3 lakh deaths occur annually due to cancer. Data from population-based registries under National Cancer Registry Program indicate that the leading sites of cancer are oral cavity, lungs, oesophagus and stomach amongst men and cervix, breast and oral cavity among women. Cancers of oral cavity and lungs in males and cervix and breast in females account for over 50% of all cancer deaths in India[3].

The objective of this paper is to establish which factors influence cervical cancer diagnosis and the treatment in that region. There is considerable interest in the use of computational techniques to aid in the detection and diagnosis the cervical cancer. Most computer-aided studies focus on screening and other tests like smear test, VIA and calposcopic test, since it is

Corresponding Author: Kuttiannan Thangavel, Department of Mathematics, Gandhigram Rural Institute -Deemed University, Gandhigram-624 302, Tamilnadu, India

the primary tool for the detection of cervical cancer. The decision to biopsy is difficult by the fact that it is painful, costly and patients have to hospitalise. In this work the k-means algorithm is used to cluster the cervical cancer patient's demographic data.

CERVICAL CANCER DISEASE AND RISK GROUP DETECTION

The cervix is the lower part of the womb (uterus) and is often called the neck of the womb. The womb is a muscular, pear-shaped organ at the top of the vagina. The lining of the womb is shed each month and results in bleeding called a period. These periods stop temporarily during pregnancy and will normally continue until a woman has the 'change of life' or menopause. Close to the cervix is a collection of lymph nodes.

The risk factors for cervical cancer widely known that includes inadequate screening, Human Papilloma Virus (HIV), multiple sexual partners, young age at intercourse or early marriage, male sexual behaviour: consumption and Tobacco, living habits and homostatic factors. In addition, it was detected that coexistence of risk factor increases the disease rate. Risk factors can be classified in to four categories, based on the evidence of their association with the disease, usefulness of measuring them and their responsiveness to intervention[4]. Category I consists of the most important risk factors for which high correlation with cervical cancer rate has been proved (white discharge hip-pain, smelly vaginal discharge, early marriage, sex at early age, multiple sexual partners). Category II includes risk factors for which the correlation with cervical cancer is likely (malnutrition, male sexual behaviours, husband's food habit and living condition). Category III is formed of risk factors associated with increased cervical cancer rate that, if modified, may decrease the risk (psychological factors), Category IV consists of risk factors associated with the increased cervical cancer rate, which can't be influenced (age, family history).

Now a days cervical cancer prevention relies practically on two significantly different concepts.:

- General education of the women population about known risk factors, especially about life style factors.
- Risk factor screening in general practice by data collection performed in the different stages:

- Collecting analytic information and physical examination results, including risk factors like age, positive family history, husband information and all.
- Collecting results of laboratory tests including pop smear test and biopsy test.
- Collecting direct investigation reports like VIA etc.

The data collected in general practice screening can be used as a basis for detecting patients at risk for cervical cancer. In Many cases with significantly pathological values, the decision is not difficult. However, the problem of disease prevention is to decide in case with slightly abnormal values and in cases when combination of different risk factors occurs.

DATA MINING AND KNOWLEDGE DISCOVERY

Human analysts with no special tools can no longer make sense of enormous volumes of data that require processing in order to make informed business decisions. Data mining automates the process of finding relationships and patterns in raw data and delivers results that can be either utilized in an automated decision support system or assessed by a human analyst.

The main reason for necessity of automated computer systems for intelligent data analysis is the enormous volume of existing and newly appearing data that require processing. The amount of data accumulated each day by various businesses, scientific and governmental organizations around the world is daunting. Hospital Scientific and business organizations store each day about 1 TB (terabyte!) of new information. It becomes impossible for human analysts to cope with such overwhelming amounts of data. Two other problems that surface when human analysts process data are the inadequacy of the human brain when searching for complex multifactor dependencies in data and the lack of objectiveness in such an analysis. A human expert is always a hostage of the previous experience of investigating other systems. Sometimes this helps, sometimes this hurts, but it is almost impossible to get rid of this fact.

One benefit of using automated data mining systems is that this process has a much lower cost than hiring an army of highly trained (and paid) professional statisticians. While data mining does not eliminate human participation in solving the task completely, it significantly simplifies the job and allows an analyst who is not a professional in statistics and programming to manage the process of extracting knowledge from data. The process of storing knowledge discovery in data bases (KDD)[5] consists of sequence of steps including problem understanding, data understanding and preparation, data mining result interpretation and evaluation finally the use of discovered knowledge.

CLUSTERING REVISITED

The need to analyse data for decision making is growing exponentially, since data collection through electronic version grow rapidly. Thus the field of data mining has emerged at the intersection of statistics, data bases and machine learning for development of the techniques to obtain information and knowledge from vast amounts of micro data, which are of numerical and categorical in nature. The development of hardware and software and the rapid computerisation of business have made capturing the data easy and digitised information are stored in database, this makes the collection and storing the data to grow at a phenomenal rate. As a result, traditional adhoc mixtures of statistical techniques and data management tools are no longer adequate for analysing such data.

Raw data is rarely of direct use. Its true value is predicted on the ability to extract information useful for decision support or exploration and understanding the phenomena governing the data source. One or more analyst may be intimating familiar with the data and with the help statistical techniques provide summaries and generate reports. Hence the analysts are acting as a sophisticated query processor. However such manual query processing has its own limitations as the size of data grows and the number of dimension increases. Since the scale of data manipulation, exploration and inference go beyond human capacities. Computing technologies become inevitable. Partitioning a set of objects into homogenous clusters is fundamental operation in Data mining[4] and the operation is needed in a number of Data Mining tasks such as unsupervised classification and Data Summation. This operation is also used in segmentation of large heterogeneous Data sets into smaller homogenous subsets that can be easily managed, separately modelled and analysed. Clustering is a popular approach used to implement this operation. Clustering methods partition a set of objects in the same cluster are more similar to each other than objects in different clusters according to some defined criteria. In statistical clustering methods[6,7], we use similarity measure to partition objects, were as in conceptual clustering methods[8,9], we cluster the objects according to the concept of objects.

The Data mining community has recently put a lot of efforts on developing fast algorithms for clustering large Data sets. Some popular algorithms include CLARA program[10], CLARANS[11], DBSCAN[12], BIRCH[13], and K-modes algorithm[12], K-prototypes[14], and PCBClu[15]. These algorithms are often revisions of some existing clustering methods by using some carefully designed search methods (e.g. combination of sampling procedure and the clustering program PAM in CLARA program, randomised search in CLARANS), organising structures (e.g. CF-Tree in BIRCH and PC-tree in PCBClu). Indices (e.g., R*-Tree in DBSCAN) and statistical methods (frequency and dissimilarity measure in K-prototypes and K-modes). These algorithms have shown some significant performance still based on complex schemes and procedures. They cannot be used to solve massive categorical data clustering problems as simple as K-means clustering methods in numerical domain.

The K-means based methods[16] are efficient for processing the large data sets, thus very attractive for Data mining. The major handicap for them is that they are often limited to numeric data. The reason is these algorithms optimise a cost function defined on the Euclidean distance measure between the data points and means of cluster[17]. Minimising the cost function by calculating means limits they used numerical data.

THE K-MEANS ALGORITHM

The K-means algorithm[7,15] is build upon the following operations.

Step 1: Choose initial cluster Centers Z1 , Z2,....Zk randomly from the n points
w1, w2, wN, wI ∈ Rm

Step 2: Assign point Wi, I = 1, 2, .. N to Cluster Cj = 1, 2,.. K
if and only if || Wq − Zq , P = 1,2, K. and J ≠ P. Ties are resolved arbitrarily

Step 3: Compute the new cluster centers Z1*, Z2* , ...Zk* as follows:
$Zi^* = (I/n) \Sigma Wj$ I = 1, 2, ... K.
Wj ∈ CJ

Step 4: If Zi * = Zi, I = 1, 2, ...K then terminate. Otherwise Zi ← Zi * and go to step 2.

Except for the first operation, the other three are repeatedly performed in the algorithm until the algorithm converges. Note that in case the process does not terminate normally at Step 4, then it is executed for a maximum fixed number of iterations.

The optimality of this algorithm can be estimated by Inter and Intra clustering metric values which has been calculated by sum of the Euclidean distances. Mathematically, the clustering intra metric ì for K clusters Cl, C2. .,Ck

$$\mu (C1,C2...,Ck) = \sum_{i=1}^{k} \sum_{xjeci} || Zi - Zj ||$$

Where Ci are Clusters and Zj are cluster centers And inter Cluster metric v for K clusters Cl,C2. ..,Ck

$$V (C1,C2...,Ck) = \sum_{i=1}^{k} \sum_{j=k+1} || Zi - Zj ||$$

The task of the proposed clustering technique is to search for the appropriate cluster centers Zl, Z2,. ..,Zk such that the clustering intra-cluster metric ì is minimised and inter cluster metric v is maximised.

There exist a few variants of algorithm which differ in selection of the initial K-means, dissimilarity calculations and strategies to calculate cluster means[6,18] .The sophisticated variants of the K-means algorithm include the well known ISODATA algorithm[19] and the fuzzy K-means algorithms[20,21].

Most K-means type algorithms have been proved convergent[16,22,23]. The K-means algorithm has the following important properties.

- It is efficient in processing large data sets. The computational complexity of the algorithm is O (tkmn), where m is the number of attributes, n is the number of objects, k is the number of clusters and t is the number of iterations over the whole data set. Usually, k, m, t * n. In clustering large data sets, the K-means algorithm is much faster than the hierarchical clustering algorithms whose general computational complexity is O (n2)[24].
- It often terminates at a local optimum[16,23]. To find out the global optimum, techniques such as deterministic annealing[24] and genetic algorithm [15] can be incorporated with the K-means algorithm.
- It works only on numeric values because it minimises a cost function by calculating the means of clusters.
- The clusters have convex shape[6]. Therefore, it is difficult to use the K-means algorithm to discover clusters with non-convex shapes.

The main difficulty in using the K-means algorithm is to specify the number of clusters. Some variants like ISODATA include procedure to search for the best K at the cost of some performance. But the extensive studies dealing with comparative analysis of different clustering methods suggests that there is no general strategy, which works equally well in different problem domain. However it has been found that it is usually beneficial to run schemes that are simpler and execute them several times like K-means, rather than using schemes that are very complex but need to be run only once.

The K-Means algorithm is best suited for data mining because of its efficiency in processing large data sets. However working only on numeric values limits its use in data mining because data sets in data mining often have categorical values

CERVICAL CANCER DATA SET AND MINING

The sample included women, 25 - 75 years of age, who were registered themselves as patients in Christian Fellowship Community Health Centre in Dindigul district Tamilnadu for a period of ten weeks. The CFCH Centre is providing health care services to large proportions of the cancer patients and is located in low income and historically undeserved neighbourhoods. The data were collected from the cancer patients, who are coming for treatment in the outpatient department and are admitted in the hospital for radiation therapy. The patients interviewed were proven cases of cervical cancer. The data collected consists of nearly fifty attributes and fairly represent all potentially and typically available information about a patient such as patient's food habit, white discharge, foul smelling vaginal discharge, husbands food habit and other behaviours, number of pregnancies, abortions, marriage age and all. In this study only patient records with complete data were included. Some of the data that are categorical in nature are then converted in to numerical data. The data are than clustered with k-means algorithm using MATLAB.

RESULTS AND DISCUSSION

The problem we investigate in this study is how to support a physician's decision of whether a biopsy is warranted. Perhaps by analysing existing or easily measured data about a patient we can develop some means by which a physician caring for a patient can better decide when to biopsy. The result is shown in Fig. 1

CONCLUSION

This model can be used in emergency room, were physician is not able to handle large amounts of data. It has been observed that the prediction of cervical cancer patient groups from existing or easily measured demographic data using clustering yield better solutions for the problem. The results thus obtained from this study are shown to be consistent with traditional medical diagnosis techniques and are really useful in prediction of non-linear groups, which are essentially different risk groups. The careful study on different risk groups let to the discovery of high-risk groups, which enables to take decision at individual, national level and at world level. If

this study is extended to a large geographical area, the highly influencing factors in different ethnic groups can also be easily identified. Also it is possible to link several attributes, which are directly or indirectly involved in the cervical cancer diagnosis.

REFERENCES

1. Parkin, D.M., P. Pisani and J. Ferly, 1993. Estimates of the worldwide incidence of eight major cancers in 1985. International journal of cancer, 54: 594-606.
2. Zvavahera, M. Chirenje, Simbarashe Rusakniko, Leah Kirumbi, Edward w. Ngwalle, Pulani Makuta-tlebere, Sam Kaggwa, Winnnie Mpaanju Shumbuaho and Lucy Makoae,Situation analysis for cervical cancer diagnosis and treatment in East, Central and Southern African Countries.
3. Rao, Y.N, Sudir Gupta and S.P. Agarwal, 2003. National cancer control programme: Current status and strategies, 50 years of cancer control in india, NCD Section,Director General Of Heath, Nirman Bhavan, New Delhi-110011.
5. Fayyad, U.M., G. Piatetsky-Shapiro, P. Smyth, 1996, From Data Mining to Knowledge Discovery: An Overview In: Fayyad U.M., G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in knowledge discovery and data mining, AAAI press, pp: 1-34.
6. Anderberg, M.R., 1973. Cluster Analysis for Applications, Academic Press.
7. Jain, A.K. and R.C. Dubes, 1988. Algorithm for Clustering Data, Prentice Hall.
8. Fisher, D.H., 1987. Knowledge Acquisition via incremental Conceptual Clustering, Machine Learning, 2: 139-172.
9. Michalski, R.S. and R.E. Stepp, 1983. Automated construction of Classifications: Conceptual clustering verses numerical Taxonomy, IEEE Trans. On Pattern Analysis and machine Intelligence, 5: 396-410.
10. Kaufman, L. and P.J. Rousseeuw, 1990. Finding Groups in Data -An Introduction to Cluster Analysis, Wiley.
11. Ng, R.T. and J. Han, 1994. Efficient and Effective Clustering Method for spatial Data Mining, Proceeding of the 2Oth VLDP Conference, Sntiago, Chile, pp: 144-155.
12. Huang. Z., 1997. Clustering Large data sets with mixed numeric and Categorical values, In the Proceedings of the First Pacific Asia Conference on Knowledge Discovery and Data Mining, Singapore, World Scientific, pp: 21-34.
14. Huang. Z., 1997. A fast clustering Algorithm to cluster very large categorical data sets in data mining, Proceedings of the SIGMOD workshop on Research Issues on Data Mining and Knowledge Discovery, Department of Computer Science, The University of British Colombia, Canada, pp: 1-8.
15. Ananthanarayana, V.S., M. Narasimha Murthy and D.K. Subramanian, 2001. Efficient, Clustering of Large Data Sets, Pattern Recognition, 34: 2561-2562-2563.
16. MacQueen, J.B., 1967. Some method, fu, classification and analysis of multivariate observations, Proceeding, of the fifth Berkely Symposium on Mathematical Statistics and Probability, pp: 281-297.
17. Everitt, B., 1974. Cluster Analysis, Heinemann Educational Books Ltd.
18. Bobrowski, L. and l.C. Bezdeck, 1991. c-means clustering with the l1 and l4 Norms, IEEE Transactions on Systems, Man and Cybernetics, 21: 545-554.
19. Ball, G.H. and D.J. Hall, 1967. A clustering technique for Summarizing Multivariate Data, Behavioral Science, 12: 153-155.
20. Ruspini, E.R., 1969. A new approach to clustering, information Control, 19: 22-32.
21. Ruspini, E.R., 1973. New experimental results in clustering, information services, 6: 273-284.
22. Bezdek, L.C., 1980. A convergence Theorem for the fuzzy ISODATA clustering algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2: 153-155.
23. Selim, S.Z. and M.A. Ismail, 1984. K-means type algorithm, A generalised 'convergence, Theorem and Characterisation, of local optimality, IEEE Transactions on Pattern Analysis, Machine Intelligence, 6: 81-87.
24. Kirkpatrick, S., C.D. Gelati and M.P. Vecchi, 1983. Optimisation by Simulated Anneling, Sci., 220: 281-297.
24. Murtagh, F., 1992. Comments on parallel algorithm, for, hierarchical clustering and cluster validity, IEEE Transaction on Pattern Analysis and Machine Intelligence, 4: 1056-1057.