Feature Selection for Prediction of HIV/AIDS using Data Mining Technique by Applying the Concept of Theory of Evidence

A.M. Saravanan **R.Vijaya** **C. Jothi Venkateswaran**

Abstract

The devastating disease HIV is well known as being the cause for development of Acquired Immunodeficiency Syndrome (AIDS). In the last 2 decades, over 60 million people have been infected with HIV, most of the people are identified and infected by HIV in the developing countries. In this paper Dempster Shafer (DS) theory is focused to identify the hidden information from the data set using the concept theory of evidence. The dataset are collected from the different Non-Government Organizations (NGO) and Government organizations of India. To analyze the dataset in particularly nearby areas of Vellore District of Tamil Nadu , India, the DS theory is used and to select the important attributes available in the dataset in order minimize the size of the data set.

Key words

Demographic, Geographic, Diagnostic, Treatment, Pharmacy, Provider and Utilization

1. Introduction

During 2006, around four million adults and children are estimated to have been infected with HIV in the world. By the end of the 2006, an estimated 10 million people were living with HIV/AIDS. India is having localized epidemics of HIV infection, National Family and Health Survey III (NFHS) data generated through population based survey. One of the state in India, Tamil Nadu has shown a considerable change in HIV scenario over a period. This is evident from the HIV sentinel surveillance (HSS) , Behavior Surveillance Survey(BSS), AIDS Prevention And Control (APAC) and recently from NFHS-III and Integrated Behavioral and Biological Assessment surveys (IBBA). Government of Tamil Nadu has initiated aggressive programme against the infection. Initial activities are focused on awareness generation.

HIV surveillance data in Tamil Nadu is generated through ICTC, clinics for patients on ART, STD clinics, blood banks and recorded deaths in addition to HSS , BSS and IBBA. Annual Sentinel surveillance is meant essentially for trend analysis and to pick up hot spots for HIV infection. Since there is no other dependable information available, sentinel surveillance is used for estimations as well. With this data a suitable data mining techniques are used to predict the status of HIV. The Dempster Shafer Theory of Evidence is used here to filter the dataset so as to do the effective data mining.

2. Dempster Shafer Theory

The Dempster-Shafer theory, also known as the theory of belief functions, is a generalization of the Bayesian theory of subjective probability [1]. Whereas the Bayesian theory requires probabilities for each question of interest, belief functions allow us to base degrees of belief for one question on probabilities for a related question. These degrees of belief may or may not have the mathematical properties of probabilities; how much they differ from probabilities will depend on how closely the two questions are related.

The Dempster-Shafer theory is based on two ideas: the idea of obtaining degrees of belief for one question from subjective probabilities for a related question [2], and Dempster's rule for combining such degrees of belief when they are based on independent items of evidence.

There are three important functions in Dempster-Shafer theory: the basic probability assignment function (bpa or m), the Belief function (Bel)[3], and the Plausibility function (Pl). The basic probability assignment (bpa) is a primitive of evidence theory.

Generally speaking, the term “basic probability assignment” does not refer to probability in the classical sense. The bpa, represented by m, defines a mapping of the power set to the interval between 0 and 1, where the bpa of the null set is 0 and the summation of the bpa’s of all the subsets of the power set is 1. The value of the bpa for a given set A (represented as m(A)), expresses the proportion of all relevant and available evidence that supports the claim that a particular element of X (the universal set) belongs to the set A but to no particular subset of A . The value of m(A) pertains only to the set A and makes no additional claims about any subsets of A. Any further evidence on the subsets of A would be represented by another bpa, i.e. B Ì A, m(B) would the bpa for the subset B. Formally, this description of m can be represented with the following three equations:

Manuscript received May 5, 2011
Manuscript revised May 20, 2011

$$m: P(X) \rightarrow [0,1] \qquad (1)$$

$$m(\Phi) = 0 \qquad (2)$$

$$\sum_{A \in P(X)} m(A) = 1 \qquad (3)$$

where P(X) represents the power set of X, Φ is the null set, and A is a set in the power set (AЄAP (X)).

From the basic probability assignment, the upper and lower bounds of an interval can be defined. This interval contains the precise probability of a set of interest (in the classical sense) and is bounded by two non additive continuous measures called Belief and Plausibility. The lower bound Belief for a set A is defined as the sum of all the basic probability assignments of the proper subsets (B) of the set of interest (A) (B C A). The upper bound, Plausibility, is the sum of all the basic probability assignments of the sets (B) that intersect the set of interest (A) (B ∩ A = Φ). Formally, for all sets A that are elements of the power set (AЄ P (X)),

$$Bel(A) = \sum_{B|B C A} m(B) \qquad (4)$$

$$Pl(A) = \sum_{B B \cap A \neq \emptyset} m(B) \qquad (5)$$

The two measures, Belief and Plausibility are non additive. This can be interpreted as is not required for the sum of all the Belief measures to be 1 and similarly for the sum of the Plausibility measures[2]. It is possible to obtain the basic probability assignment from the Belief measure with the following inverse function:

$$m(A) = \sum_{B|B \subseteq A} (-1)^{|A-B|} Bel(B) \qquad (6)$$

where |A-B| is the difference of the cardinality of the two sets.
In addition to deriving these measures from the basic probability assignment (m), these two measures can be derived from each other. For example, Plausibility can be derived from Belief [4] in the following way:

$$Pl(A) = 1 - Bel(\bar{A}) \qquad (7)$$

where A is the classical complement of A. This definition of Plausibility in terms of Belief comes from the fact that all basic assignments must sum to 1.

$$Bel(\bar{A}) = \sum_{B|B \subseteq \bar{A}} m(B) = \sum_{B|B \cap A = \emptyset} m(B) \qquad (8)$$

$$\sum_{B|B \cap A = \emptyset} m(B) = 1 - \sum_{B|B \cap A = \emptyset} m(B) \qquad (9)$$

From the definitions of Belief and Plausibility, it follows that Pl(A)=1-Bel(A). As a consequence of Equations 6 and 7, given any one of these measures (m(A), Bel(A), Pl(A)) it is possible to derive the values of the other two measures. The precise probability of an event (in the classical sense) lies within the lower and upper bounds of Belief and Plausibility, respectively.

$$Bel(A) = P(A) = Pl(A) \qquad (10)$$

3. Problem Description

The data is collected from TANSACS and Primary Health centre in and around Vellore, TamilNadu, India. Such large amount of data must be normalized using the Normalization technique in the database management system. Depends upon the characteristics of the variable in the dataset must be analyzed to take suitable decision to control the spreading of HIV. The vertical fragmentation has been utilized to reduce the size of the data by applying the concept of clustering in the Data mining. The DS theory of evidence is used to filter the data so as to extract the hidden information efficiently in the data set.

4. Methodology

The dataset for this paper to extract the hidden information is gathered from multiple sources from TANSACS and Primary Health centers nearby Vellore. The objective is that maintaining a uniform and unique record for each patient in the study population. The dataset includes variables viz., Demographic, Geographic, Diagnostic, Treatment, Pharmacy, Provider and Utilization.

Data Preparation

The data preparation procedure includes for cleaning and reducing the size of the data and transforming data to an unique record per patient. Domain experts with their expert knowledge are utilized to find the features which are having most probability of values by assigning the probability of mass values to them.

The Demographic variables can be denoted as De ={ d1,d2,d3, …..} and each variable may grouped into D1= {d1,d2…}, D2 ={d3,d4,…}. Then De= {D1,D2,…Dn}, where each D's will a subgroup demographic variables.

The Geographic variables G= {G1,G2,…Gn} and each G's will have a subgroup of Geographic variables. Diagnostic variable Di= {Di1,Di2…Din} , the Treatment variables T={T1,T2,…} , the Pharmacy variables P={P1,P2, …}, the Provider variables Pr= {Pr1,Pr2,…} and the Utilization variables U={U1,U2,…} are denoted and grouped.

The Dempster Shafer theory concept is applied to find the important variables from each clustered group. For the calculation four variables are taken to find the most important group for extraction of hidden information.

Mass Function (m) Assignment

The Domain experts are used to assign the probabilities for the variables based on the expert knowledge. The expert gives the following mass values to the variables (A) as shown in the Table 1.

Variables	Mass
{D1}	0.02
{G1}	0.02
{Di1}	0.03
{T1}	0.03
{D1,G1}	0.05
{D1,Di1}	0.05
{D1,T1}	0.05
{G1,Di1}	0.05
{G1,T1}	0.05
{D1,G1,Di1}	0.10
{D1,G1,T1}	0.10
{D1,Di1,T1}	0.10
{G1,Di1,T1}	0.10
{D1,G1,Di1,T1}	0.25
Total Mass value	**1.0**

Table 1. Probability of Mass Assignment

Belief in A (Bel) Function

The belief in A is the power set i.e., the sum of the masses of variables which are subsets of A. The subsets of A will be a power set and it has 2n -1(n = 4), set of variables (excluding empty subset) ie., Power set of A = {D1,G1,Di1,T1, {D1,G1}, {D1,Di1},{D1,T1} , …{D1,G1,Di1,T1}}

Now Belief in A

Bel({D1})= m({D1})= 0.02
Bel({D1,G1})=m({D1})+m({G1})+m({D1,G1})=0.02+ 0.02+0.05=0.09
Bel({D1,Di1})=m({D1})+m({Di1})+m({D1,Di1})=0.02 .+0.03+0.05=0.1

The calculations shows the belief function for subgroup of variables.

Plausibility of A : $P_l(A)$

The Plausibility of A , Pl(A) is the sum of all the masses of the sets that intersect with set A.

Pl({D1})= m({D1})+ m({D1,G1}) + m({D1,Di1}) + m({D1,T1}) + m({D1,G1,Di1}) + m(D1,G1,T1) + m({D1,Di1,T1}) + m({D1,G1,Di1,T1})

Pl({D1})= 0.02+ 0.05+0.05+ 0.5+0.10+ 0.10+ 0.10+ 0.20= 0.67

Disbelief (Doubt) in A : Dis(A)

The disbelief in A is (not A)

i.e., - Dis(A)= 1- $P_l(A)$

The calculated values are shown in the Table 2.

Variables	Mass	Belief	Pl(A)	Dis(A)
{D1}	0.02	0.02	0.67	0.33
{G1}	0.02	0.02	0.67	0.33
{Di1}	0.03	0.03	0.68	0.32
{T1}	0.03	0.03	0.68	0.32
{D1,G1}	0.05	0.09	0.89	0.11
{D1,Di1}	0.05	0.10	0.90	0.10
{D1,T1}	0.05	0.10	0.90	0.10
{G1,Di1}	0.05	0.10	0.90	0.10
{G1,T1}	0.05	0.10	0.90	0.10
{D1,G1,Di1}	0.10	0.11	0.91	0.09
{D1,G1,T1}	0.10	0.27	0.97	0.03
{D1,Di1,T1}	0.10	0.32	0.97	0.03
{G1,Di1,T1}	0.10	0.33	0.98	0.02
{D1,G1,Di1,T1}	0.25	0.33	0.98	0.02
Total Mass value	1.0	1.0	1.0	0

Table 2. Belief , Plausibility calculations and their values

Belief Interval of A

The uncertainty associated with a given subset A is defined by the belief interval [Bel(A) , Pl(A)]

Conclusion

In this paper the Dempster Shafer theory of evidence is used to extract the hidden information by filtering the features from the large data set. The Normalization and Vertical Fragmentation concept is used with the DS theory to filter the data set. In this paper four variables are focused and one domain expert is used to find the belief and plausibility values which gives to select the variables for data mining to understand the behavior of HIV. Use of more than one expert and more than four variables for mass assignment of the variables can be done in the further research work.

References

[1] Dempster, A. P., A generalization of Bayesian inference, *J. Roy. Stat. Soc. Ser.B* 30(2), 205-247, 1968.

[2] Shafer, G., *A Mathematical Theory of Evidence*, Princeton University Press,Princeton, N.J., 1976.

[3] Shafer, G., Perspectives on the theory and practice of belief Functions, *Int. J.Approx. Reasoning* 4(5 / 6), 323-362, 1990.

[4] Shafer, G., Rejoinders to comments on “Perspectives on the theory and practice of belief functions,” *Int. J. Approx. Reasoning* 6(3), 445-480, 1992

[5] HIV Risk Behavior Surveillance Survey in Rural Tamil Nadu (Wave – IV; 2006) by Social and Rural Research Institute , New Delhi.

[6] The Tamil Naud HIV Sentinel Surveillance Report – 2006 by Tamil Nadu State AIDS Control Society (TANSACS), Government of Tamil Nadu, India

[7] The Tamil Naud HIV Sentinel Surveillance Report – 2005 by Tamil Nadu State AIDS Control Society (TANSACS), Government of Tamil Nadu, India

[8] The Tamil Naud HIV Sentinel Surveillance Report – 2006 by Tamil Nadu State AIDS Control Society (TANSACS), Government of Tamil Nadu, India

[9] Behavioural Surveillance Survey Round II in Rural Tamil Nadu by TANSACS, Government of India, India.

A.M. Saravanan is an Assistant Professor in the Department of Computer Science, at Muthurangam Government Arts College (Autonomous), Vellore. He holds a Master Degree in Computer Science from the Bharathiar University, Coimbatore and a Master Degree in Information Technology at Punjabi University, Patiala. He has more than 16 years of teaching and 5 years of research experience. He guided more than 10 M.Phil. research scholars.

R. Vijaya is an Assistant Professor in the Department of Computer Science, at Arignar Anna Government Arts College for Women, Walajapet. She holds a Master Degree in Computer Application from the University of Madras in the year 2000. She has more than 12 years of teaching and 3 years of research experience.

Dr. C. Jothi Venkateswaran, is an Associate Professor and Head of the Post Graduate Department of Computer Science at Presidency College, Chennai. He has been serving more than 24 years of teaching experience and more than 10 years of research experience in the field of Data mining and Database Management System. He has published many articles in the National and International Journals of Computer Science and presented papers in many Conferences.