

Fuzzy Association Rule Mining

¹Lekha, A., ²C.V. Srikrishna and ³Viji Vinod

¹Dr. M.G.R Educational Research Institute, Chennai, India-600095, Assistant Professor, Department of MCA, PESIT, Bangalore

²Department of MCA, PESIT, Bangalore-560085

³Department of MCA, Dr. MGR Educational and Research Institute, Chennai, India-600095, India

Article history

Received: 12-04-2014

Revised: 16-04-2014

Accepted: 25-07-2014

Corresponding Author:

Lekha. A.

Research Scholar, Dr M G R

Educational Research Institute,

Chennai, India-600095, Assistant

Professor, Department of MCA,

PESIT, Bangalore

Email: raoalekha@gmail.com

Abstract: The paper attempts to propose a fuzzy logic association algorithm to predict the risks involved in identifying diseases like breast cancer. Fuzzy logic algorithm is used to find association rules. The results of the study revealed that the prediction is better reliable than conventional methods.

Keywords: Association Rule Mining, Breast Cancer, Fuzzy Logic

Introduction

Fuzzy logic is an approach of data mining that involves computing the data based on the probable predictions and clustering. In the traditional approach it is done based on “true or false”. Algorithms that use fuzzy logic are increasingly being applied in several disciplines to help in mining databases. One of the potentially viable applications of Fuzzy logic algorithms found in clinical studies is the clustering of breast cancer data which enable oncologists to detect and evaluate breast cancer risks such as malignant tumors. Breast cancer is currently one of the major health problems as well as the leading cause of death amongst women worldwide (Dorf and Robert, 2010). It is known that breast cancer is the second largest cause of cancer deaths among women. It is also among the most curable cancer types if it is diagnosed early (Addeh and Ebrahimzadeh, 2012). Consequently early detection of cancer risks is one of the key ways of improving the prognosis of the disease.

Although there are a number radiological techniques such as mammography that can be used in the early detection of breast cancer risks, the enormous data generated by these techniques often make it difficult for radiologists to accurately evaluate breast cancer data (Chen and Chen, 2008). Clinical findings, tumor characteristics and molecular markers are integrated to identify different risk categories,

based on which treatment is planned for each individual case (Saleh *et al.*, 2011).

The major variables in breast cancer databases include the information related to the risk factors and mammographic findings. Although the primary causes of cancer are not yet known, there are a number of risk factors that have been identified and can therefore be fixed to particular class. Generally tumors can be malignant (cancerous) or benign (non-cancerous). In most cases, malignant tumors have rapid growth that often results in the destruction of normal tissues and their eventual spread to all parts of the body. On the other hand, benign tumors tend to be localized and grow slowly without any significant spread to the other parts of the body (Nguyen and Walker, 2003). Consequently the risk of breast cancer development is generally higher when malignant tumors are detected in an individual.

The major variables in breast cancer databases include the information related to the risk factors and mammographic findings. Although the primary causes of cancer are not yet known, there are a number of risk factors that have been identified and can therefore be fixed to particular class. Generally tumors can be malignant (cancerous) or benign (non-cancerous). In most cases, malignant tumors have rapid growth that often results in the destruction of normal tissues and their eventual spread to all parts of the body. On the other hand, benign tumors tend to be localized and grow slowly without any significant spread to the

other parts of the body (Michalski *et al.*, 1986). Consequently the risk of breast cancer development is generally higher when malignant tumors are detected in an individual.

Use of artificial intelligence techniques such as fuzzy clustering algorithms can significantly improve the diagnosis and evaluation of breast cancer risks through clustering of the particular data elements (Addeh and Ebrahimzadeh, 2012). Consequently the incorporation of fuzzy logic algorithms in data mining is a powerful tool that can be employed in the extraction, clustering, quantification and analysis of the data base information regarding the assessment and diagnosis of cancer risks.

When dealing with uncertainties in databases, fuzzy logic clustering algorithms can be used to cluster different elements of data into various membership levels depending on their closeness (Addeh and Ebrahimzadeh, 2012). For example, during the evaluation of breast cancer risks, mammogram data may possess some degree of fuzziness such as ill defined shapes, indistinct borders and different densities. In this regard, a fuzzy clustering algorithm can be one of the most effective ways of handling the fuzziness of data related to breast cancer. As an intelligent technique, Fuzzy logic data mining algorithms not only provide excellent analysis of the data but can also be used to develop accurate results that are easy to implement.

One of the greatest potential advantages of bringing the concept of fuzzy logic in data mining is that such algorithms can significantly be used in the modeling of inaccurate, non linear and complex data systems by implementing human knowledge and experience as a set of fuzzy rules that uses fuzzy variables for inference purposes (Nguyen and Walker, 2003). For example when using fuzzy algorithm for the prediction and clustering of breast cancer data, the human experience and knowledge related to breast cancer risks can be expressed as a set of inference rules of deduction that are then attached to the fuzzy logic system. Another important advantage of fuzzy algorithms systems for prediction and clustering of breast cancer data is that they usually have a significantly high inference speed (Gláucia *et al.*, 2012). There has been advances in new rule mining technique using fuzzy logic for mining medical data in order to serve the needs of Multidimensional Breast cancer Data applications. (Sankaradass and Arputharaj, 2011). This study proposes a fuzzy association algorithm that can be used in the data mining of breast cancer data and consequently in the evaluation and prediction of cancer risks in patients with suspected cancer cases.

Association Rule Mining

Mining association rule is one of the important research problems in data mining. There are many known algorithms for mining Boolean association rule such as Apriori, Apriori TID and Apriori Hybrid algorithms for mining association rule (Dorf and Robert, 2010). Although the current quantitative association rule mining algorithms can solve some of the problems introduced by quantitative attributes, they introduce some problems (Mukherjee, 2010). The first problem is caused by the sharp boundary between intervals. The algorithms either ignore or over-emphasize the elements near the boundary of the intervals in the mining process. The use of shape boundary interval is also not intuitive with respect to human perception (Castillo and Melin, 2008; Chen and Chen, 2007).

Fuzzy Healthy Association Rule Mining

If X and Y are sets of attributes of database and A, B are sets containing fuzzy sets which characterize X and Y respectively then a fuzzy association rule is of the form, "If X is A then Y is B". As in the binary association rule, "X is A" is called the antecedent of the rule while "Y is B" is called the consequent of the rule (Subramanyam and Goswami, 2006). The fuzzy association rule is easily understandable because of linguistic terms associated with the fuzzy set.

Many ARM algorithms are more concerned with efficient implementations than producing effective rules. In almost all ARM algorithms, thresholds (both confidence and support) are crisp values. This support specification may not suffice for queries and rule representations that require generating rules that have linguistic terms such as "low protein" Fuzzy approaches deal with quantitative attributes by mapping numeric values to real values.

Fuzzy Association Rule Algorithm:

- Candidate itemsets are generated with predefined minimum support and minimum confidence
- Implement the k-means clustering algorithm to find centroids or mid-points
- Using these centroids fuzzy sets and membership degree is calculated
- Fuzzy rules are generated using the fuzzy support and confidence found

Case Study

The breast cancer domain was obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. M. Zwitter and M. Soklic have

provided the data (http://www.hakank.org/weka/BC.arff (25063)). The Arff conversion of the data set was provided by Hakan Kjellerstrand (Fu *et al.*, 1998). This data set has 286 instances described by 9 attributes+one class attribute. The set includes 201 instances of one class and 85 instances of another class.

Results of the Experiment

The data set was processed by the Association Rule Mining algorithm with minimum support 0.25 and minimum confidence 0.9 to find the fuzzy rules.

The rules generated are given below in Table 1.

According to the above output, the percentage of the attribute node-cap with value 'no' along with the attribute inv-node with value '0-2' is 0.98. The attribute node-cap with the value 'no' alone occurs in 11 rules and the attribute inv-node with value '0-2' occurs in 11 rules along with the other attributes.

In the above output the attributes node-cap with value 'no' and inv-node with value '0-2' plays a major role in forming association rules. The output

changes with the change in the values of minimum support and minimum confidence.

It was processed by the Fuzzy Healthy ARM algorithm with minimum support 0.25 and minimum confidence 0.9 to find the fuzzy rules.

The rules generated are given below in Table 2.

Discussion

From the above table the analysis highlights the following information. The attribute age [40-99] plays a major role in generating the rules. The attributes tumorsize [15-59] and inv-node [20-59] take part in both sides of the rule generation. The attribute deg-malignant (Castillo and Melin, 2008) plays a major role in forming the remaining fuzzy rules.

Fuzzy Association Rule Mining only works on quantifiable data that are given in different ranges. To implement fuzzy logic in association rule mining clustering has to be applied first and later association rule mining. The analysis shows that the occurrences of cancer depend on the attributes of age and degree of malignant.

Table 1. Association data analysis for ARM algorithm

If	Then	Confidence measure
Node-cap-'no'	inv-nodes between 0-2	0.90
Inv-nodes between 0-2	node-cap-'no'	0.94
Node-cap-'no', breast-side-'left'	inv-nodes between 0-2	0.94
Breast side is 'left' and inv-node is '0-2'	node-cap 'no'	0.95
Node-cap is 'no' and class is 'no-recurrence-event'	inv-node '0-2'	0.93
Class is 'no-recurrence-events' and inv-node is '0-2'	node-cap 'no'	0.95
Menopause is 'premeno' and inv-node is '0-2'	node-cap 'no'	0.93
Irradiate is 'No' and class is 'no-recurrence-events'	node-cap 'no'	0.91
Node-cap is 'no' and irradiate is 'No'	inv-node '0-2'	0.94
Irradiate is 'No' and inv-node is '0-2'	node-cap 'no'	0.96
Irradiate is 'No', class is 'no-recurrence-events' and node-cap is 'no'	inv-node '0-2'	0.95
Ays irradiate is 'No', class is 'no-recurrence-events' and inv-node is '0-2'	node-cap 'no'	0.98

Table 2. Association data analysis for FHARM algorithm

If	Then	Confidence measure
Age between 40-59	Tumor size between 25-59	1.0
Age between 60-99	Tumor size between 25-59	1.0
Age between 40-59	Inv-node between 30-39	1.0
Age between 60-99	Inv-node between 30-39	1.0
Age between 40-59	Deg-malignant-2	1.0
Age between 60-99	Deg-malignant-2	1.0
Tumor size between 15-25	Inv-node between 30-39	1.0
Tumor size between 25-59	Inv-node between 30-39	1.0
Tumor size between 15-25	Deg-malignant-2	1.0
Tumor size between 25-59	Deg-malignant-2	1.0
Inv-node between 20-29	Deg-malignant-2	1.0
Inv-node between 30-59	Deg-malignant-2	1.0

Conclusion

Applying fuzzy logic on both association rule mining and Clustering techniques on this data set shows that attribute age [40-99] plays a major role in generating the rules and forming clusters. The analysis shows that for a few attributes the percentage of occurrence varies in different clusters. The attributes that take part in the determination of cancer are age [40-49][50-59], menopause with value being 'premeno', node-caps being 'no', degree of malignancy being (Castillo and Melin, 2008), breast quadrant being 'left-up' and irradiate being 'no'. FARM algorithm works for quantifiable data and not textual data.

Although the primary causes of cancer are not yet known, there are a number of risk factors that have been identified and can therefore be fixed to particular classes. Generally tumors can be malignant (cancerous) or benign (non-cancerous). In most cases, malignant tumors have rapid growth that often results in the destruction of normal tissues and their eventual spread to all parts of the body. The paper results in the findings that age, menopause and degree of malignancy are some of the reasons for breast cancer.

Acknowledgement

The researchers thank the principal and management of PESIT, for their continued support.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues involved.

References

- Addeh, J. and A. Ebrahimzadeh, 2012. Breast cancer recognition using a novel hybrid intelligent method. *J. Med. Signals Sens.*, 2: 95-102.
- Castillo, O. and P. Melin, 2008. *Type-2 Fuzzy Logic: Theory and Applications*. Berlin: Springer-Heidelberg, ISBN-10: 978-3-540-76283-6, pp: 29-43.
- Chen, Z. and G. Chen, 2007. An approach to classification based on fuzzy association rules. *Proceedings of the International Conference on Intelligent Systems and Knowledge Engineering, (ISKE' 07)*, pp: 15-16.

- Chen, Z. and G. Chen, 2008. Building an associative classifier based on fuzzy association rules. *Int. J. Comput. Sys.*, 1: 262-273. DOI: 10.1080/18756891
- Dorf, R.C. and H.B. Robert, 2010. *Modern Control Systems*. 12th Edn., Prentice Hall, ISBN-10: 0136024580, pp: 1104.
- Fu, A.W., M.H. Wong, S.C. Sze, W.C. Wong and W.L. Wong *et al.*, 1998. Finding fuzzy sets for the mining of fuzzy association rules for numerical attributes. *Int. Symp. Intelli., Data Eng. Learn.*, Hong Kong,
- Gláucia, R.M.A.S., C.R.M. Leite, A.M.G. Guerreiro and A.D.D. Neto. 2012. Fuzzy method for pre-diagnosis of breast cancer from the Fine Needle Aspirate analysis. *BioMed. Eng.* 11: 83-83. DOI: 10.1186/1475-925X-11-83
- Michalski, R.S., I. Mozetic, J. Hong and N. Lavrac, 1986. The multi-purpose incremental learning system aq 15 and its testing application to three medical domains. *Proceedings of the 5th National Conference on Artificial Intelligence*, Nov. 12, Philadelphia, PA: Morgan Kaufmann, pp: 1041-1045.
- Mukherjee, S., 2010. *The Emperor of All Maladies: A Biography of Cancer*. 1st Edn., New York, Simon and Schuster, ISBN-10: 13-978-1439170915, pp: 608.
- Nguyen, H.T. and C.L. Walker, 2003. *A First Course in Fuzzy and Neural Control*. 1st Edn., New York, Chapman and Hall, ISBN-10: 1420035525, pp: 312.
- Saleh, A.A.E., S.E. Barakat, A.A. Ebrahim Awad, 2011. A fuzzy decision support system for management of breast cancer. *Int. J. Adv. Comput. Sci. Appli.*, 2: 34-40.
- Subramanyam, R.B.V. and A. Goswami, 2006. Mining fuzzy quantitative association rules. *Expert Sys.*, 23: 212-225. DOI: 10.1111/j.1468-0394.2006.00402.x
- Sankaradass, V. and K. Arputharaj, 2011. A descriptive framework for the multidimensional medical data mining and representation. *J. Comput. Sci.*, 7: 519-525. DOI: 10.3844/jcssp.2011.519.525