Empirical Analysis on Cancer Dataset with Machine Learning Algorithms

Chapter *in* Advances in Intelligent Systems and Computing · August 2019

DOI: 10.1007/978-981-13-0514-6_75

Panduranga vital Terlapu
Aditya Institute of Technology & Management
35 PUBLICATIONS **54** CITATIONS

M. Murali Krishna
SRI SIVANI COLLEGE OF ENGINEERING
7 PUBLICATIONS **36** CITATIONS

Analysis of Cancer Data Set with Statistical and Unsupervised Machine Learning Methods

T. Panduranga Vital, K. Dileep Kumar, H. V. Bhagya Sri and M. Murali Krishna

Abstract Research on cancer is very important where cancer is one of the leading diseases which causes more deaths worldwide. Data mining is very useful for analysis of medical data, especially in cancer. Statistical analysis results are very useful for assumptions, preventions and diagnosis of cancer. The main aim of this paper is analysing the cancer data set collected form zone 1 (Srikakulam, Vizianagaram, Visakhapatnam) of Andhra Pradesh for estimations or diagnosis and preventions of cancer disease. The analysis is very useful with its good results. In this, we use statistical and clustering methods like k-means and hierarchical and multidimensional scaling (MDS). As per the statistical reports, most of the women are affected by cancer than men in the zone 1 of AP. Most of the women cancer patients suffered from breast cancer, and most of the men cancer patients are affected by lung cancer. The analysis also gives interesting results about living styles and habits linked with cancer. Unsupervised machine learning algorithms also give the good results for predicting cancer. The hierarchical cluster study projections clearly describe the cause of occurring cancer in zone 1 districts of Andhra Pradesh that the main combination factors are smoke, drink, gutkha (chewing tobacco) and related job.

T. Panduranga Vital (✉) · K. Dileep Kumar · H. V. Bhagya Sri · M. Murali Krishna
Department of Computer Science & Engineering,
Sri Sivani College of Engineering, Srikakulam, Andhra Pradesh, India
e-mail: vital2927@gmail.com

K. Dileep Kumar
e-mail: kadamati.dileep@gmail.com

H. V. Bhagya Sri
e-mail: bhagya.hanumanthu@gmail.com

M. Murali Krishna
e-mail: maadugula@gmail.com

S. C. Satapathy et al. (eds.), *Smart Intelligent Computing and Applications*,
Smart Innovation, Systems and Technologies 104,
https://doi.org/10.1007/978-981-13-1921-1_27

1 Introduction

Data mining (DM) is an effective and efficient method for knowledge extraction from raw data. Raw data faces different difficulties that make traditional or conventional strategy improper for knowledge extraction [1, 2]. DM should have the capacity to handle different data types in all configuration formats. Medical or healthcare data mining is multifaceted field with commitment of medicine and data mining [3, 4]. The important medical characteristics that specified are: estimate the health costs, diagnosis and visualization; extract the hidden values from biomedicine information and find a relationship amongst diseases and drugs.

The development of information storage technology has prompted to produce a large amount of raw data that considers perspectives [5, 6]. These perspectives are mounting of modern storage equipment and algorithm advancement. Useful and effective knowledge can be gained by the normal raw data. Knowledge discovery is the insignificant extraction of verifiable, already obscure and helpful prospective information from the data [7]. Attractive components for extracted knowledge are sensible time multifaceted nature, understandability, precision and valuable result. The extracted knowledge gives the new information for further utilization discoveries. DM was initially considered as equivalent word of KDD [8, 9].

De Falco extracted knowledge by differential evolution in the form of if … then rules to predict the results of diseases. In this approach, results of DM were compared with a skilled oncologist's research work [10]. Survivability of breast cancer disease can be predicted by k-means. Results demonstrate that evaluation, stage of cancer, number of primaries and radiation are the most prognostic factors in [11].

Lung cancer is one of the frequently occurred and vulnerable cancers mostly affected by air pollution, cigarette smoking and cardiopulmonary syndrome identified by statistically significant and rich association techniques [12, 13]. Due to industrialization and urban development have increased human revelation to plentiful cyan genetic substances that was raised about their relationship to the aetiology of chronic diseases [14]. The medical scientists have long-utilized maps to track the spread of disease with using powerful novel tools [15] in geographic information system technology that assists to reveal far more than simply the 'where' and 'when' of an epidemic.

2 Methodologies

The data set has been analysed by DM techniques. The frequency of patients has been analysed for the collected data from zone 1 (Srikakulam, Vizainagaram and Visakhapatnam) districts of Andhra Pradesh, India. The questionnaire has been framed based on the present personal profile, living and food habits, travelling mode, previous history, internal and external factors, etc.

The data was collected with 1100 instances (550 cancer instances and 550 non-cancer) and 46 attributes and one class (place, age, cancer type, family history, drinking, smoking, tea, coffee, milk, job, morning-eat, lunch-eat, dinner-eat, travel, living status, fruits, vegetables, sleeping hours, tension, cooldrinks and ice cream, study, height, weight, BP, pains, hair loss, gutkha (chewing tobacco), marital status, blood group, treatment type, bathing, oils, fast food, other disease, morning walk, use mobile, drugs or using tablets, treatment mode, diagnosis mode, meditation, mosquito repellents, injuries, speak level, watching TV, think levels and class status of cancer (0 (no) for non-cancer and 1 (yes) for cancer).

K-Means (KM) Algorithm

The KM algorithm is dividing the set of observation points into k clusters [16]. Let R is the real number set and Rd said to be the d-dimensional object space. Given a finite set X that is determined by Eq. (1).

$$X = \{x_1, x_2 \ldots x_n\} \tag{1}$$

where ‘n’ indicates the object’s number.

The KM algorithm partitions the set into subset S, whose subsets are mentioned in Eq. (2)

$$S = \{S_1, S_2, \ldots\ldots\ldots, S_k\} \tag{2}$$

where k is a predefined number.

Each cluster C is represented by an object. It is mentioned in Eq. (3)

$$C = \{c_{1,} \ldots., c_k\} \tag{3}$$

where C is the centre set in the object area.

The Euclidean distance estimation measures using the distance amongst items and cluster centres. Equation (4) represents the objective function that it should be minimized.

$$f = \sum\nolimits_{i=1}^{k} \sum\nolimits_{i=1}^{n} (||x_i(j) - C_j||)^2 \tag{4}$$

where

$x_i(j)$ is a particular ith data point at jth cluster

C_j is centroid or cluster centre

n is the number of instances or data points

$||x_i(j) - C_j||$ is the Euclidean distance(ED) between $x_i(j)$ and C_j

The cluster centres are derived by Eq. (5)

$$C_j = \left(\frac{1}{n}\right) \sum_{x \in s}^{n} X_j \tag{5}$$

where C_j is the count of data objects.

Each cluster is characterized by its centre point called the centroid. Generally, the distance measures used in clusters do not represent the spatial distance values. In general, the solution is to find the global minimum which is the exhaustive choice of starting points. But the usage of many replicates with random initial point leads to a solution called global solution. A centroid is a point whose coordinates are obtained by calculating the mean of each coordinate of the points of samples classified into the clusters. The KM algorithm input is the number of clusters k and n objects and the output is set of k clusters which minimize the squared error function.

3 Results and Discussion

The statistical analysis provided the relationship of the collected data sets. As per analysis, there are more number of cancer instances in Visakhapatnam district compared to other two districts like Vizianagaram and Srikakulam. There are more number of female patients (64.3%) compared to male patients (35.7%).

As per the analysis of family history (hereditary), 19.4% of cancer patient's family members had cancer. Hence, cancer is related to both hereditary and metabolic.

Drinking alcohol and smoking habits cause of cancer. In this research, 26.2% of cancer patients have the drinking alcohol habit and 33.3% of cancer patients have smoking habit. Some other habits are like drinking tea and coffee, 83.5% of cancer patients drink tea and 23.4% of patients prefer coffee. Hence, drinking tea may be the reason for cancer occurrence.

The study results show more number of breast cancer (20.2%) patients followed by cervix (16.3%), stomach (7.3%) and blood (7.3%). Hence, there is a need to control these cancers in the zone 1 regions of AP. 44.2% of patients are house wives which are observed to be more when compared to working women. 56.5% of cancer patients live in urban region, whereas 43.5% of patients live in rural region. According to this, there are more number of cancer patients in urban region compared to rural region. Most of them are travelling by bus and auto, and most of the cancer patients like banana, non-vegetarian, rose and jasmine

Age, height and weight are some of the factors occurring cancer. In this analysis, most of the patients are in between the age of 35–60 years. Hence, cancer can be an ageing disease. In the height factor, the mean of the height is 153.97 ± 15.09 centimetres. In the weight factor, the mean of the weight is 54.89 ± 9.97 kilograms.

Most of the cancer patients have tension (54.8%). Most of the patients do not have education. Most of them are having low blood pressure compared to high BP.

Some of the cancer patients prefer cooldrinks rarely (61.3%). Some of the patients prefer to take ghutka (chewing tobacco). About 46.6% of cancer patients have pain in body. The study also shows that most of the patients have hair loss (73.9%).

Some interesting results that 91.7% of patients are married. Most of the patients are related to O+ blood group. There is more number of diabetic (21.2%) patients associated with cancer.

The most of the patients preferred allopathic (89.5) as primary treatment. Previously, most of the patients did not use drugs. The patients underwent treatment by chemotherapy. Most of the diagnosis was conducted by scanning and biopsy. Some other facts from the study that above 26% patients used to take palm oil, 58.1% do not take fast foods and 54.2% use steel as cooking vessel.

It shows that most of the patients do not go for morning walk (63.5%), does not have any habit of meditation (82.9%) and even playing games (64.9%). Above 84.5% of cancer patients are using mosquito repellents. About 71.4% of patients did not have major injuries in their lifetime. 66.7% of patients have medium speaking level, 53.6% of patients see TV medium and 54.8% of patients think medium.

The main cluster centroid uses k-means clustering algorithm by class and gender. The model is constructed with full training data and takes the 0.5 s. The two clustered instances are cluster 0 and cluster 1. The cluster 0 contains 42% instances, whereas cluster 1 contains 58% instances. The centroid values of full data attributes are 47.4802 (age), urban (living), yes (tension), 153.9742 (height), 54.8869 (weight), etc. The cluster 0 centroid's values are 55.3302 (age), rural (living), no (tension), 155.5991 (height), 55.5425 (weight), low (BP), yes (pains), more (hair loss) and so on. The cluster 1 centroid's values are 41.7808 (age), urban (living), yes (tension), 152.7945 (height), 54.411 (weight), normal (BP), no (pains), less (hair loss) and so on.

Figure 1 shows the k-means cluster analysis related to age attribute and cancer type attribute. In this, the blue colour elements represent the cluster 0 and the red colour elements represent the cluster 1. The square symbols represent the male elements, and the cross symbols represent the female elements. The clusters 0 and 1 are formed with patients' age between 35 and 70 years and patients of breast and lung cancers.

Figure 2 shows the scatter plot for cancer types with blood groups at regional level. Most of the cancer instances from Visakhapatnam and Vizianagaram are seen in A +ve, B +ve and O +ve blood groups. K-means cluster algorithm is used to analyse the scatter plot.

Figure 3 shows the multidimensional scaling (MDS) plot. As per the analysis, it has been observed that more members from Visakhapatnam are affected by cancer due to tensions.

Figure 4 shows the complete link hierarchical clustering for the cancer data set of zone 1 districts of Andhra Pradesh. In this, first compute the pairwise distances of two clusters named as C1, C2 and then choose the largest distance of all pairwise distances. In other words, if the cluster C1 contains m elements and the C2 cluster contains n elements, then compute the m*n pairwise distances and construct the hierarchical link for chosen largest distance for that element.

In this experiment, the complete link hierarchical clustering is grouping the ages of cancer instances with respect to the cancer type attribute. The distance between

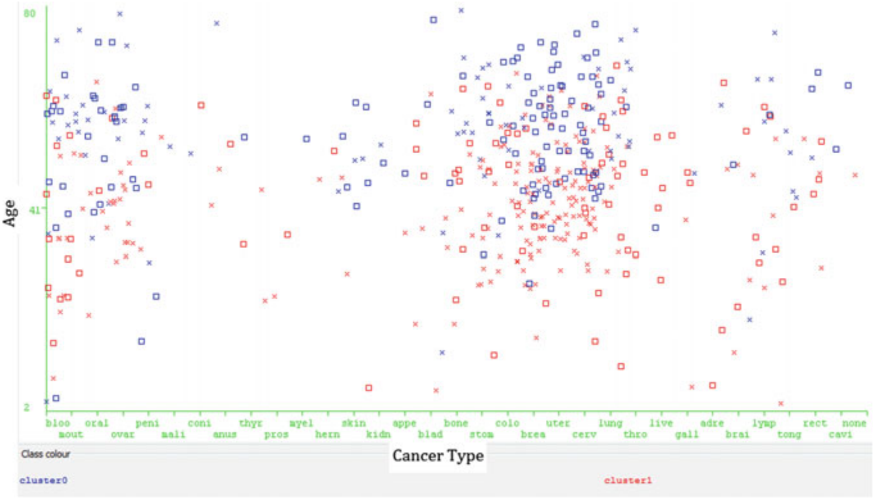

Fig. 1 K-means clustering related to cancer type and age

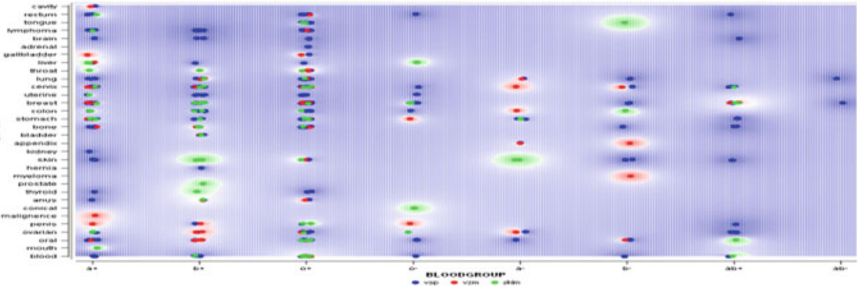

Fig. 2 Scatter plot for relationship to blood group, cancer type and place using k-means clustering

two clusters is determined as the maximum of pairwise distances. The maximum distance of two cluster elements (age 9 and age 9) is connected in hierarchical way at the distance 3.886854 (shown in green colour hierarchical cluster block). In the same way, all elements in green colour cluster elements are constructed in hierarchical manner. The maximum height of the green colour hierarchical cluster block is 5.646441. As well as, the maximum height of the pink colour cluster structure is 6.10689, and the blue colour cluster height is 6.4366. The blue and pink colour clusters are connected at the maximum height 6.591434 and treated as one cluster. It is connected with the green colour cluster at height 6.766052 in-depth value 5.

In the same way, complete link hierarchical clustering can be applied to the cancer type attribute (Fig. 4). The maximum distance of two cluster elements (blood and bone) is connected in hierarchical way at the distance 3.886854 in green

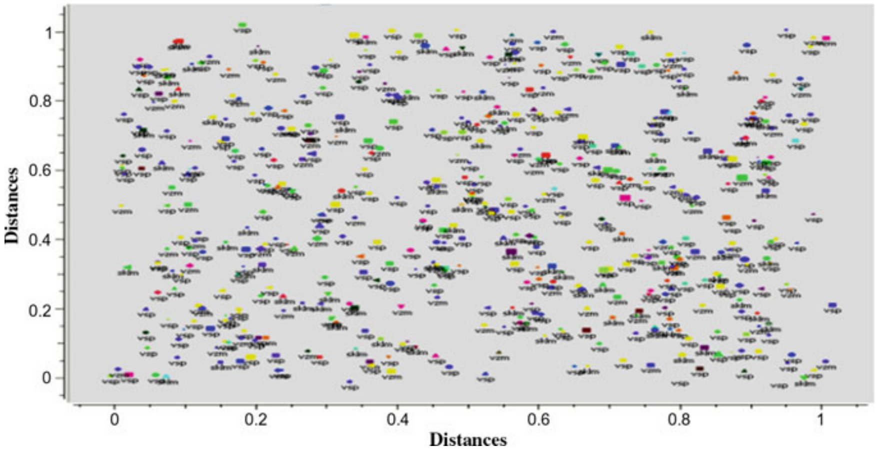

Fig. 3 MDS Plot

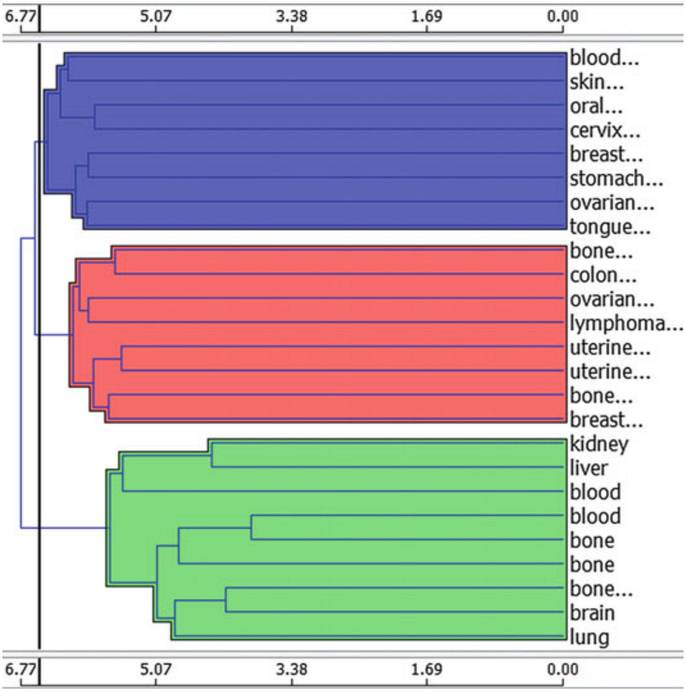

Fig. 4 Complete link hierarchical cluster with respect to cancer type attribute of cancer data set

colour cluster block. In the same way, all elements in green coloured cluster are constructed in hierarchical manner. The height of the green colour hierarchical cluster is 5.646441. As well as, the height of the pink colour cluster is 6.10689, and the blue colour cluster height is 6.4366. The blue and pink colour clusters are connected at the maximum height 6.591434 and treated as one cluster. It is connected by the green colour cluster at height 6.766052 in-depth 5.

Figure 5 shows the linear projection by using hierarchical clustering for cancer data set. The projections are measured with five parameters (projections with maximum five attributes) that are cancer type (cantype), smoke, drink, gutkha (chewing tobacco) and job. The most projections are interrelated with these attributes that this combination is the first rank with 92.98% predictable score of all

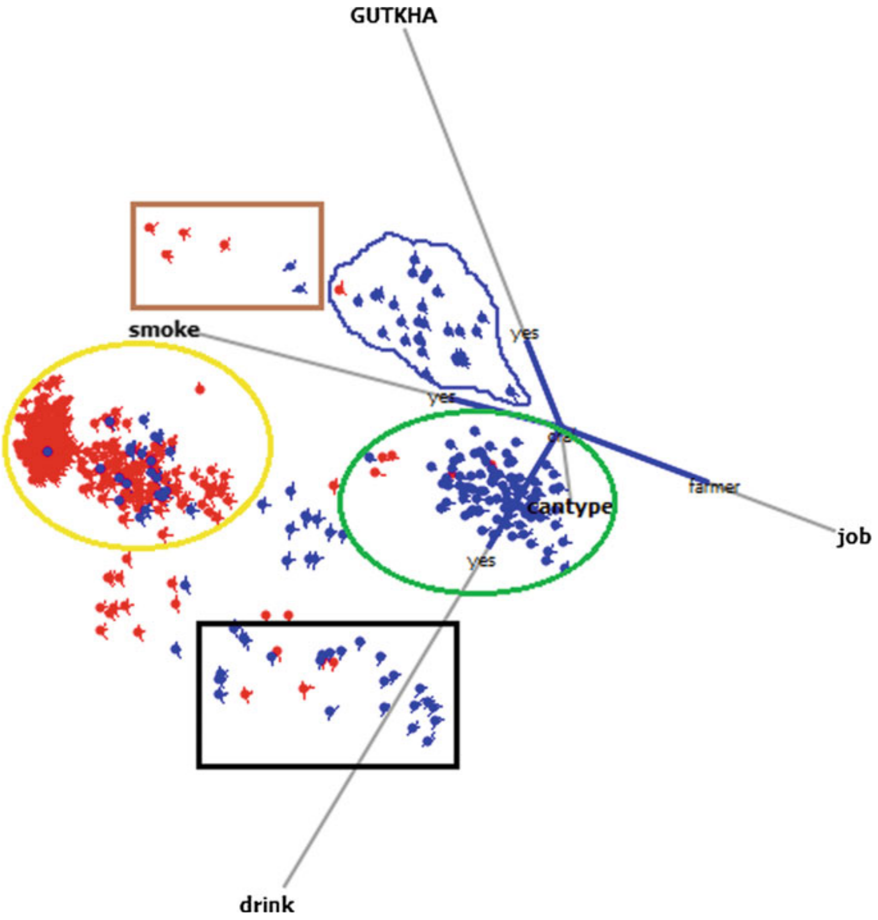

Fig. 5 Projections analysis of cancer data set

other tested possible placements of five visualized attributes in the projection list. For this, k-nearest neighbour method is used for projection evaluation and supervised principal component analysis for optimizing separation analysis. The study clearly describes the cause of occurring cancer in zone 1 districts of Andhra Pradesh that the main combination factors are smoke, drink, gutkha (chewing tobacco) and related job.

The figure shows the projections analysis of cancer data set. The red colour plotting points represent the female instances, and blue colour plotting points represent the male instances of cancer data set. The green colour circle block contains the most of male instances and complete instances in the block related to habit of smoking (yes), drinking (yes) and gutkha (yes). In this, many instances are associated with mouth, lung, colon and stomach cancers as per the parameter cancer type (cancer type attribute) computations, and it's data points indicates the lower level job status like farmer, auto driver related to job parameter (attribute).

The complete instances of yellow colour circle block correspond to the no habits like smoke (no), drink (no) and gutkha (no). Most of the data points of this yellow circle block are red colour that indicates the female instances of cancer data set. Most of the exemplars specify the breast and colon cancer on the cancer type (cantype) parameter and also describes the job parameter that corresponds to housewife, worker, and office assistant.

The black-bordered block data points show that drink parameter value is yes and remaining smoke and gutkha attributes values are no. The blue-bordered block data points show that drink parameter value is no and remaining smoke and gutkha parameters values are yes. The brown-bordered block data points show that drinking habit is yes and remaining smoke and gutkha attribute values are no.

4 Conclusion

The analysis with attributes age, sex, smoke, mosquito repellents, gutkha (chewing tobacco) thinking levels, tensions and food habits are major factors in cancer occurrences. In this, scatter and MDS plates are analysed using k-mean cluster algorithm. It gives the good result causing cancer. As well as, projections analysis is used by hierarchical clustering with Kruskal's algorithm and display attribute projections rankings of cancer data set. Further analysis will be conduct on prediction and curing of cancer.

Note: Authors have taken the consent from the concerned person/authority to use the materials, etc., in the paper. Authors will be solely responsible if any issues arise in future with regard to this.

References

1. Cruz, Joseph A., David S Wishart.: Applications of machine learning in cancer prediction and prognosis. Cancer inf. **2** (2006)
2. Hara., Ichimura, T.: Data mining by soft computing methods for the coronary heart disease database. In: Fourth International Workshop on Computational Intelligence and Application, IEEE SMC Hiroshima Chapter, Hiroshima University, Japan, 10–11 December (2008)
3. Rajkumar, Reena, G.S.: Diagnosis of heart disease using datamining algorithm. Glob. J. Comput. Sci. Technol. **10**(10) (2010)
4. Lenert, L., Lin, A., Olshen, R., Sugar, C.: Clustering in the Service of the Public's Health http://www-stat.stanford.edu/olshen/manuscripts/helsinki.PDF
5. Srinivas,K., Rani, B.K., Govrdhan, A.: Applications of data mining techniques in healthcare and prediction of heart attacks. Int. J. Comput. Sci. Eng. (IJCSE). **02**(02), 250–255 (2010)
6. Yan, H.: Development of a decision support system for heart disease diagnosis using multilayer perceptron. In: Proceedings of the 2003 International Symposium, vol. 5, pp. 709–712 (2003)
7. Sitar-Taut, V.A.: Using machine learning algorithms in cardiovascular disease risk evaluation. J. Appl. Comput. Sci. Math. (2009)
8. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Mag. **17**(3), 37 (1996)
9. Balasubramanian, T., Umarani, R.: An analysis on the impact of fluoride in human health (dental) using clustering data mining technique. In: Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, 21–23 March 2012
10. De Falco, I.: Differential Evolution for automatic rule extraction from medical databases. Appl. Soft Comput. **13**(2), 1265–1283 (2013)
11. Belciug, S., Gorunescu, F., Salem, A., Gorunescu, M.: Clustering-based approach for detecting breast cancer recurrence. In: 10th International Conference on Intelligent Systems Design and Applications (2010)
12. Vital, T., Panduranga, et al.: Data collection, statistical analysis and clustering studies of cancer dataset from viziayanagaram District, AP, India. ICT and critical infrastructure. In: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II. Springer, Cham, (2014)
13. Douglas, P.K., Harris, S., Yuille, A., Cohen, M.S.: Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief versus disbelief. Neuroimage **56**(2), 544–553 (2011)
14. Rekha Saxena, B.N.: Nagpal, M.K. Das, Aruna Srivastava, Sanjeev Kumar Gupta, Anil Kumar, A.T. Jeyaseelan, and Vijay Kumar Baraik. A spatial statistical approach to analyze malaria situation at micro level for priority control in Ranchi district, Jharkhand. Indian J. Med. Res. **136**(5), 776–782 (2012)
15. Andreeva, P.: Data modelling and specific rule generation via data mining techniques. In: International Conference on Computer Systems and Technologies – CompSysTech (2006)
16. Hamerly G., Elkan C.: Learning the K in K-means. In: Proceedings of the 17th Annual Conference on Neural Information Processing Systems, British Columbia, Canada (2003)