Classification of Renal Cancer using Principal Component Analysis (PCA) and K-Nearest Neighbour (KNN)

Nikita, Harsh Sadawarti, Balwinder Kaur
School of Engg. & Tech., CT University,
Ludhiana,

Abstract:-**The increasing growth of the renal cancer became the most crucial issue in the society. Renal cancer has to be predicted at its introductory stages to prevent the last or end stages. The early identification and treatment will keep the renal cancer from getting worse. There is no any awareness about its growth, its end effects and if it increased then how can it affects the health of a patient. Hence, there is requirement of advanced diagnostic system which assist to maintain the health of an individual. The main goal of this undertaken research is to merge the developed data reduction techniques and the supervised classification technique. The data reduction technique used in this research work for the detection of renal cancer is Principal Component Analysis (PCA) and similarly, the used supervised classification technique id K- Nearest Neighbour (KNN). The examination of the dataset as well as the observed result is done by considering several performance parameters such as classification accuracy, sensitivity, specificity and precision. The result deduced that the Principal Component Analysis (PCA) with K-Nearest Neighbour (KNN) has the better results with classification accuracy 93.33%.**

Keywords: Data mining, Classification, Principal component analysis, K-nearest neighbour, Renal cancer.

I. INTRODUCTION

The extraction of the required data or an information form a given dataset is done by the data mining techniques. The other terms that are used for the data mining are mining of knowledge form the database, data archaeology, extraction of knowledge and data analysis. The data mining is basically used to process the information stored in the database and grab the required knowledge from it. Various techniques of data mining assist to fetch the required knowledge from the large amount of data stored in the knowledge base. These techniques are classification, association and clustering [1].

The data mining algorithms also aids to implement the feature selection as well as classification. The dimension reduction and feature selection are both closely relate to each other. The main intent of the feature selection is to identify or choose only those features which are important in the dataset and rest of the features are neglected which are not relevant or duplicate features. The feature selection algorithms enhance the possibility of more accurate and fast operations of data mining algorithms as it reduces the dimensionality of the provided dataset [2].

Classification is a supervised learning in which the system learns from the given dataset and classify the new instances according to that particular provided dataset. In other words, in this learning, the system classify according to the knowledge that learnt by the system by the previously classified set of instances during the training period. The dataset provided to the system as a training data is already categorized into predefined groups and system will search those data to classify the new given instance and make patterns and according to that patterns the prediction will take place.

The data mining techniques are used in many domains. These algorithms are also used in the medical field to predict the trends of a patients' health [3]. In this study, principal component analysis with k-nearest neighbour is used for the classification of renal cancer. The organization of rest of the paper is as follow: in section 2, the literature review of renal cancer is illustrated and section 3 displays the overview about renal cancer. The principal component analysis algorithm and k-nearest neighbour are explained in the section 4 and section 5 respectively. The experimental results for the developed research work is presented in the section 6 and the research work in concluded in section 7.

Literature Review

This section explained the literature review of renal cancer in which several researchers used distinct methodology for the identification as well as diagnosis of renal or kidney cancer.

Golodetz, Voiculescu, and Cameron [4] detected the different requirements that used to make a tool a decision support system. This developed system can detect the renal cancer form a human body. However, in future, this proposed decision support system can also be used for other deadly cancers. The various techniques of image processing such as segamentation, registration, etc. are used to evaluate the growth of the cancer and the primary intent of this system is to assist the experts in the treatment of their patients.

Haas and Nathanson [5] determined the various genes which are inherited from the parents to the children and became a reson for the renal cancer. As the renal cancer is a herditry disease, then it is important to identify those mutated genes that can increse the possiblity of having renal cancer in future for a particular individual.hence, by recognize these genes with their proper and precise definition, the renal cancer can be detected easily and can also be avoid for more damage in future generations.

Tuncer and Alkan [6] constructed a decision support system which assists in the identification of kedney cancer by using the images of the kidney. This system compares the images of a health kidney with the abnormal kidney. The two steps i.e. segmentation and cancer detection are used in this process. The support vector machine (SVM) is used for the classification of various stages of the renal cancer.

Linguraru et al. [7] propsed a clinical tool which aids to classify the renal tumors. The image processing technique has been used on the computerised tomograpgy (CT) images for this classification. The proposed tool detects the different five types of the lesion of renal or kidney. The developed clinical tool permits the correct and accurate classification of canacer as well as cysts from the provided clinical dataset.

Abhilash and Chauhan [8] gave a prediction methodology that helps to monitor the respiration induced movements of urological organs. This technique developed by using a hybrid method i.e., adaptive neuro-fuzzy inference system. The image data is provided to the system. The prediction if the movement of the kidney is analyzed from the skin markers seen in the images of kidney in the dataset.

Tander, Ozmen, and Ozden [9] used the mutilayer preceptron neural network to predict the possibilty of reoccurance of the renal cancer. As the renal cancer can be reoccure even after the taking the cure of it. The developed network calculate the probability of renal cancer of a particular patient after 5 years od operation or transplant of the kidney.

Renal Cancer

The imperative organ of the human body is the kidney. The main responsibility of this organ is to remove the waste products from metabolic activities. The foreign elements in the body such as urea and creatinine are also eradicated by the kidney. Hence, the failure of the kidney can cause the major destruction to a human body [10]. The renal cancer is at rank 13 in all over the world among other cancers. Additionally, it is most frequent cancer in men with 9th rank. There are several number of patients affected by the renal cancer in 2012 in which 214,000 were men and 124,000 were women. The countries having high socio-economic development have approximately 70% of the cases of renal cancer. The renal cancer affects the men more as compared to women and also the children are least affected by this deadly disease. In 2012, 143,000 deaths had been examined that is due to renal cancer and because of these number of deaths, the renal cancer reach at the 16th rank among all other cancer which causes death [11].

The major risk factor of renal cancer is smoking. This is the main reason due to which renal cancer affects more men than ladies. Hence, there is 1.6% risk for development of this deadly disease [12]. This disease is seen in the nephrology general practice. The renal cancer is also known as renal cell carcinoma [13]. Renal cancer can be occurred due to genetic or hereditary. Hence, the probability of a patient having renal cancer increases if the patient has family background, which is suffering from this life-threatening disease. out of total number of cases of renal cases, 3% to 4% of cases are due having family history of renal cancer [5]. Additionally, if a patient get the treatment for the renal cancer then there are also chances to reoccur this disease again to that particular patient after few years. Hence, the patient have to do regular checkups and consultation with specialist to stop the reoccurrence of the renal cancer [9].

The researchers for the treatment of the renal or kidney cancer have used several techniques. The image processing techniques are used in which the magnetic resonance imaging (MRI) and computerised tomography (CT) images have been used for the identification of renal cancer by applying numerous techniques such as feature selection, segmentation, etc[4]. The fuzzy logic and the adaptive neuro-fuzzy inference system have been developed for the diagnosis as well as recognition of renal cancer at the introductory stages as if the renal cancer reach at its final stage then the treatment is more difficult [14].

Principal Component Analysis

The principal component analysis found the required pattern from the dataset and then compression of that particular dataset has been take place without any much lose of information. It reduces the dimensionality of the data by selecting the required features and neglect the irrelevant ones. The main applications of the principal component analysis are applications like patter recognition, face detection, compression of images or evaluating the required pattern from large amount of data. Hence, the principal component analysis is a crucial technique for feature selection as well as for data compression [15].

The overview of principal component analysis (PCA) is explained below:

Consider that {xt} where t=1,2,3,4,……,N. this is a n dimensional input data and the mean of this data is represented by μ. This means can be defined as:

$$\mu = \frac{1}{N}\sum_{t=1}^{N} x_t \quad \ldots.(1)$$

The covariance matrix of the assumed input data x_t is defined as following

$$C = \frac{1}{N}\sum_{t=1}^{N}(x_t - \mu)(x_t - \mu)^T \quad \ldots.(2)$$

Principal component analysis calculate the eigenvalue problem of C i.e., covariance matrix of x_t:

$$C_{vi} = \lambda_i v_i \quad \ldots.(3)$$

Where λ_i= eigen values and i=1,2,3,…,n

v_i= corresponding eigen values and i= 1,2,3…,n.

Now, calculate the m eigen values also known as principal directions corresponding to the m largest eigen values (m<n), to show the given data with reduced dimensional vectors.

Let

$$\phi = [v_1, v_2, \ldots., v_m]$$

And $$\Lambda = diag[\lambda_1, \lambda_2, \ldots, \lambda_m] \quad \ldots. (4)$$

Then

$$C\phi = \phi\Lambda \quad \ldots.(5)$$

The approximation precision of largest engine vector is represented by a parameter v. hence,

$$\frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \geq v \quad \ldots..(6)$$

From the equation (5) and (6), the number of eigen vactors can be identified. Hence, the reduced dimension feature vector of the new dataset of inputs x is calculated by:

$$x_f = \phi^t x \quad \ldots.(7)$$

K-Nearest Neighbor

The k-Nearest Neighbor is a supervised learning used in the pattern recognition for the classification as well as regression. In classification and regression, the K closest training instances are provided of inputs in the feature space. K-

nearest neighbour is an instance-based learning in which the system learn by examples or instances. The output of the K-Nearest Neighbor in case of classification is the class membership. The classification is being done according to the number of vote of neighbors. The class is known as single nearest, if the value of K is 1. All the neigbors are assigned by a weight of 1/d where d= distance of neighbour. The distance between two nearest neighbour is known as Euclidean distance and there is a straight line between the neighbours having shortest distance [16].

The drawback of the K-NN is that it is precise to the data, which is local configured. The feature extraction is the procedure of transferring the data set used as input to the collection of various features. Before applying the k-nearest neighbour algorithm in the feature space, the eradication is take place on the raw data. The various steps that are followed in k-nearest neighbor algorithm are shown in figure 1.

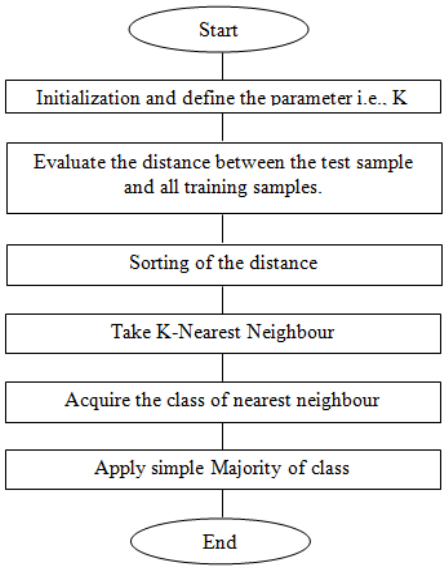

Figure 1: Steps of K-Nearest Neighbour algorithm.

RESULT

The methodology used for the proposed classification by using principal component analysis and k-nearest neighbour is described in figure 2.

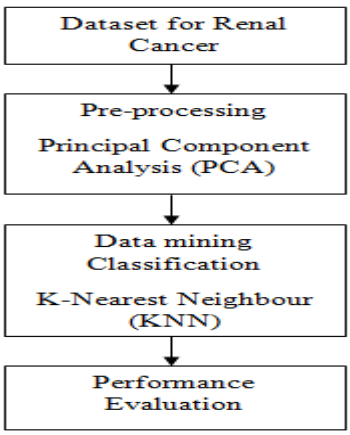

Figure 2: Methodology used for developed system

The various performance parameters have been considered to evaluate the performance of the developed system. These parameters are classification accuracy, precision, specificity and sensitivity.

The table 1 shows the values of these parameters that are evaluated

Performance Parameters	Value in percentage (%)
Classification Accuracy	93.33%
Sensitivity	94.20%
Specificity	92.59%
Precision	91.54%

Figure 5 represent the graphical view of calculated performance parameters of the proposed system.

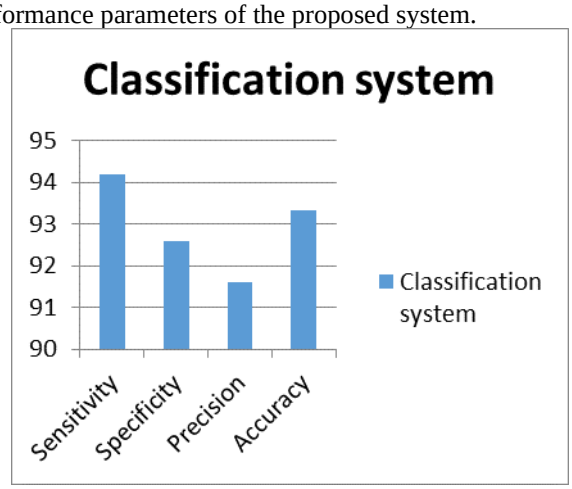

Figure 5: Calculated Parameters

CONCLUSION

The classification of the renal cancer has been done by using the principal component analysis (PCA) and K-Nearest Neighbour (KNN) in this research work. The experimental result calculated the performance of this system and according to it, the accuracy of classification for these techniques is 93.33%. From this research work , it is concluded that the classification of renal cancer can be done by these used techniques.

REFERENCES

[1] P. Sinha and P. Sinha, “Comparative Study of Chronic Kidney Disease Prediction using KNN and SVM,” International Journal of Engineering Research and, vol. V4, no. 12, 2015.

[2] A. Nway Oo, “Classification of Chronic Kidney Disease (CKD) Using Rule based Classifier and PCA,” International Journal of Management, Technology And Engineering, vol. 8, no. VII, Jul. 2018.

[3] A. K. S. S. B. P. Dr. D. P. Shukla, “A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level”, int. jour. eng. com. sci, vol. 2, no. 09, Sep. 2013.

[4] S. Golodetz, I. Voiculescu, and S. Cameron, “A Proposed Decision-Support System for (Renal) Cancer Imaging,” pp. 361–366, 2007.

[5] N. B. Haas and K. L. Nathanson, “Hereditary Kidney Cancer Syndromes,” Advances in Chronic Kidney Disease, vol. 21, no. 1, pp. 81–90, 2014.

[6] S. A. Tuncer and A. Alkan, “A decision support system for detection of the renal cell cancer in the kidney,” Measurement: Journal of the International Measurement Confederation, vol. 123, no. April, pp. 298–303, 2018.

[7] M. Linguraru, S. Wang, F. Shah, R. Gautam, J. Peterson, W. Linehan, and R. Summers, “Computer-aided renal cancer quantification and classification from contrast-enhanced CT via

histograms of curvature-related features," 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009.

[8] R. H. Abhilash and S. Chauhan, "Respiration-Induced Movement Correlation for Synchronous Noninvasive Renal Cancer Surgery," vol. 59, no. 7, pp. 1478–1486, 2012.

[9] B. Tander, A. Ozmen, and E. Ozden, "Neural network design for the recurrence prediction of post-operative non-metastatic kidney cancer patients," 2015 9th International Conference on Electrical and Electronics Engineering (ELECO), pp. 162–165, 2015.

[10] V. S. Anu.Batra, Usha.Batra, "A Review to Predictive Methodology to Diagbose Chronic Kidney Disease," 2016.

[11] A. Technopole, "Kidney Cancer : Diagnosis and Treatment," vol. 2, pp. 325–331, 2019.

[12] M. Atif, M. S. Alsalhi, S. Devanesan, V. Masilamani, K. Farhat, and D. Rabah, "SC," Photodiagnosis and Photodynamic Therapy, 2018.

[13] R. H. Weiss, "Metabolomics and Metabolic Reprogramming in Kidney Cancer," Seminars in Nephrology, vol. 38, no. 2, pp. 175–182, 2018.

[14] H. Ahmadi, M. Gholamzadeh, L. Shahmoradi, and M. Nilashi, "Computer Methods and Programs in Biomedicine Diseases diagnosis using fuzzy logic methods : A systematic and meta-analysis review," Computer Methods and Programs in Biomedicine, vol. 161, pp. 145–172, 2018.

[15] E. Oja, "Principal components, minor components, and linear neural networks," Neural Networks, vol. 5, no. 6, pp. 927–935, 1992.

[16] R. K. Leung, Y. Wang, R. C. Ma, A. O. Luk, V. Lam, M. Ng, W. Y. So, S. K. Tsui, and J. C. Chan, "Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case–control cohort analysis," BMC Nephrology, vol. 14, no. 1, 2013.