A CASE STUDY ON PERCLUSION AND DISCOVERY OF SKIN MELANOMA RISK USING CLUSTERING TECHNIQUES

P. THANGARAJU[1], B.DEEPA[2]

[1]Asst. Professor , Department of Computer Applications, Bishop Heber College (Autonomous), Trichy-17.
[2]M.Phil, Scholar, Department of Computer Applications, Bishop Heber College (Autonomous), Trichy-17.

Abstract:

This paper presents a survey study on perclusion and discovery of skin melanoma risk using clustering techniques. Skin melanoma (cancer) is the uncontrolled growth of abnormal skin cells. Skin cancer can result in disfigurement and even death. Like other cancer, skin cancer also depends on some risk factors. So the discovery of skin melanoma is a multi layered problem. According to these risk factors group of people's data is gathered from different diagnostic centre which contains both cancer and non cancer patients information and gathered data is pre processed for duplicate and missing information. Then pre processing data is clustered using k- means clustering algorithm for separating relevant and non relevant data to skin cancer. Then significant frequent patterns are discovered using MAFIA algorithm. Finally implement a system using c#.net to predict skin melanoma risk level with suggestions which is easier, cost reducible and time savable.

Keywords: *Data mining, Skin cancer, Clustering, K-means, MAFIA, Significant Pattern.*

I. Introduction

Skin cancer is the uncontrolled growth of abnormal skin cells. It occurs when unrepaired DNA damage to skin cells triggers mutations, or genetic defects, that lead the skin cells to multiply rapidly and form malignant tumors. It is also the easiest to cure, if diagnosis and treated early. When allowed in progress, however skin cancer can result in disfigurement and even death.

Skin cancers are named for the type of cells where the cancer starts. The different types of skin cancers are,

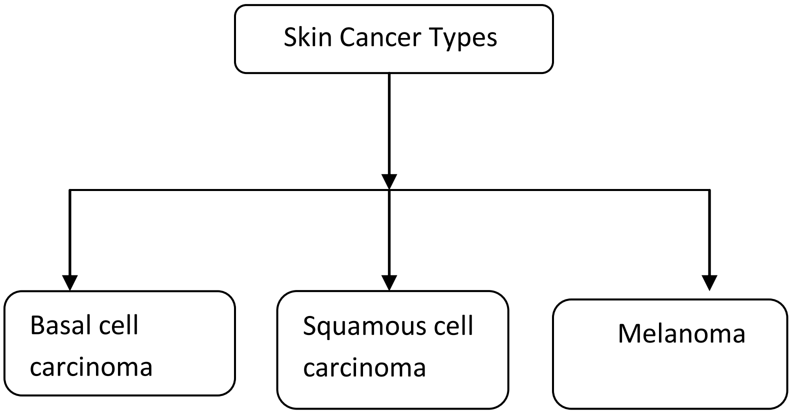

Figure 1: Types of skin cancer.

- Basal cell carcinoma:

These begin in parts of the skin that have been exposed to expressive harmful UV rays of the sun. The face is the most common place to find basal cell skin cancer. These are slow – growing and rarely spread to other parts of the body. More than 90 percent of the skin cancers found in Americans are basal cell carcinoma.

- Squamous cell carcinoma:

This cancer originates in squamous cells. This is the commonest type of skin cancer seen in dark skinned individuals. It is usually found in places that are not exposed to the sun. It usually occurs on parts of the skin that have been in the sun such as the head, face, ears, and neck. It sometimes spreads to other parts of the body [7].

- Melanoma:

This is the most risky type of skin cancer. It commonly extends to other major organs like liver, lung, brain, and bones. It is however much less common than the other types. Melanoma begins in Melanocytes (pigment cells) [7].

Skin cancer treatment based on the type and stage of the disease, the size and place of the tumor and general health and medical history of the patients. In most cases, the main aim of the skin cancer treatment is to demolish or eliminate the cancer completely. If skin cancer is found and treated early, it can be cured.

II Review of Literature

Muhammed Akmal Sapon , Khadijah Ismail and Suehazlyn Zainudin [1] presented a study of supervised learning algorithms of Artificial Neural Network on diabetes prediction. By using Regression Analysis, the performance of each algorithm is discussed. To validate the prediction accuracy, the prediction accuracy algorithm is calculated. The best performance is produced by Bayesian Regulation algorithm in the prediction of diabetes.

Zakaria Nouir, et . al [2] applied a new fuzzy clustering algorithm to a prediction tool of a third generation (3G) cellular radio network. Outcomes explain that the differences between simulation and the measurements can be diminished and the generalization capacity is improved to the proposed clustering algorithm. This algorithm performs better than the K- means algorithm. This algorithm is used to improve the generalization capabilities that use measurements to reduce the bias between reality and simulation and to prevent over- learning. This algorithm is used to predict the stochastic transfer function between s measurements and simulations. Hence improve the quality and precision of the simulations.

Doug Burdick , et. al [3] performed study for mining maximal frequent itemsets from a large transactional database by using MAFIA algorithm. When extracting long itemsets MAFIA performs best. MAFIA is highly optimized for mining long itemsets and on dense data consistently outperforms Depth project by ten to thirty and GenMax by two to ten.

Manaswini Pradhan and Dr. Ranjit Kumar Sahu [4] presented study for classifying diabetic patients into two classes by using one of the powerful method is Artificial Neural Network (ANN) based classification model. For feature selection, Genetic Algorithm (GA) is used for achieving better results. To find out the number of neurons in single hidden layered model, GA is used. Additionally the model is skilled with Genetic Algorithm (GA) and Back Propagation (BP) Algorithm and also accuracies of the classification are compared.

Carlos Ordonez, [5] in his work he shows that SQL implementation of the K- means algorithm works efficiently on the relational DBMS. It explains how to cluster large data sets defining and indexing tables to store and retrieve intermediate and final results, optimizing etc.,. The advantage in the proposed K-Means implementation can cluster large data sets and exhibits linear scalability. The final implementation is a naïve translation of K- means computation into SQL server as a framework to optimize the performance.

Jiang su and Harry Zhang, [6] proposed a fast decision tree learning algorithm that is based on a conditional independence assumption. The advantage of the new algorithm has a time complexity of O(m. n), where m is the size of the training data and n is the number of attributes. It avoids the vast model space searching, here the decision tree algorithms are based on the heuristic search. The advantage of the new proposed algorithm is to reduce the time complexity and faster than C4.5. The disadvantage is it doesn’t handle the missing values and it is inferior to C4.5 algorithm in handle missing values.

III Background

Data mining is the practice of examining large pre-existing databases in order to generate new information. Data Mining involves the following performance such as extract, transform, and load transaction data onto the data warehouse system, Store and manage the data in the multidimensional database system, Provide data access to business analysts and information technology professionals, Analyze the data by application software, and Present the data in a useful format.

The different steps in Data Mining are Selection, Preprocessing, Transformation, Data Mining and Pattern evaluation.

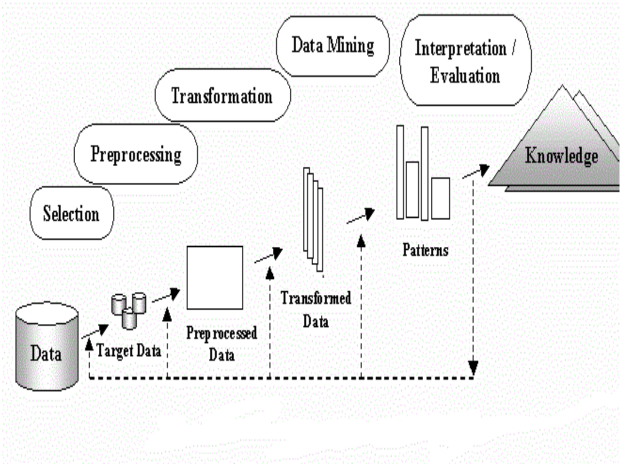

Figure 2: Steps of Data Mining Process

Data mining has some techniques to examine the data. They are classification, clustering, correlations, association rule etc and has been used intensively and widely by many organizations [11].

Data Pre-Processing is a main task of data mining. It is used to provide a appropriate analysis and also creating appropriate data for clustering by eliminating redundant data and supplies missing data according to the previous record. The main advantage of data pre-processing reduces memory [9].

Clustering is an unsupervised learning technique. It is a process of grouping the similar data into group. Clustering partitioned the dataset into similar and dissimilar dataset to skin cancer. MAFIA algorithm is used to find out frequent patterns easily and effectively than other algorithm. The datasets are responsible to skin cancer by using significant frequent pattern [9].

Because of unconsciousness about skin cancer and danger factors of skin cancer, everyday skin cancer patients are increasing hastily. The main aim of this survey is to build up a system which can be used by a person checking his/ her skin cancer danger level and provides suggestions according to his/her skin cancer danger level [9].

IV. Methodology

3.1 Data Pre-Processing:

Data pre-processing is a vital which is the monotonous task of data mining. The main aim of data pre-processing is formulating the suitable analysis and appropriate for clustering of collected data. It eliminates the redundant values and stores the missing values. It performs cleaning, normalization, transformation, feature extraction and selection, etc.

3.2 Data Clustering:

Data Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It is a major task of exploratory data mining and a common technique for statistical data analysis used in many fields, including machine learning pattern recognition, image analysis, information retrieval, and bioinformatics.

Data object is assigned to an unknown class that have unique feature and reduces the memory. It is the fundamental operation in data mining.

A clustering algorithm partitions a data set into several groups based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity.

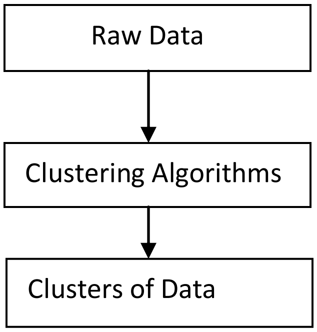

Figure 3: Stages of Clustering [10].

K- means is an unsupervised learning and iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached. Within a cluster, a centroid represents a cluster, which is a mean point within cluster [10].

Its main goal is to subset n observations into K clusters in which each observation belongs to the cluster with the nearest mean [10]. The result of this subset of the data space into Voronoi cells.

The numerical attributes only works efficiently in K-means algorithm. The most popular clustering tool which is used in industrial and scientific applications is K-means algorithm [10].

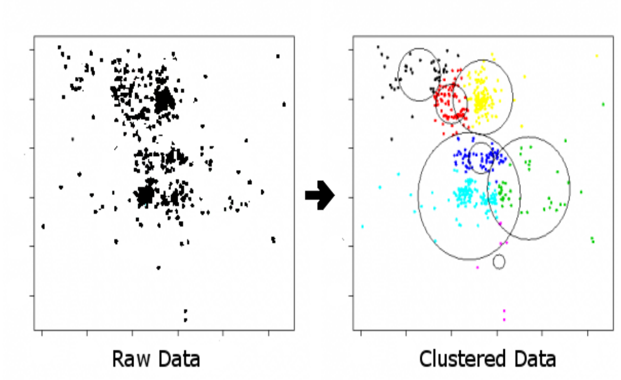

Figure 4: K-means algorithm example.

Data are separated into voronoi cells by K-means that assumes equal-sized cluster.

The fundamental algorithm is very easy [10]:

1. Select K points as initial centroid.

2. Repeat.

3. Form K clusters by assigning each point to its closest centriod.

4. Reassign the centroid of each cluster until centroid does not change.

3.3 Discover Frequent Pattern:

The most important, useful and significant topics of data mining is Discover frequent pattern. There are several algorithms are used to mine the frequent patterns from database such as classification, clustering, association and correlations etc such as Apriori, AprioriTid, Decision Tree and FP-Tree [9].

Maximal Frequent Itemset Algorithm (MAFIA) is used to extract the frequent pattern. The MAFIA algorithm is efficiently works than other algorithm such as Apriori, AprioriTid, Decision Tree and FP-Tree of mining the frequent pattern from clustered dataset [9].

3.4 Significant Pattern Find-out:

After determining the frequent pattern using MAFIA algorithm, the weightage significant patterns are extracted by using the Equation (1) [9]

$$S_{W(i)}=\sum(W_i*F_i) \qquad (1)$$

Here W_i is the attribute's weightage and F_i represents the number of frequency.

Then Significant Frequent Pattern is chosen by using the following Equation (2) [9]

$$SFP= S_W(n) \geq \varphi \text{ for all values of n} \qquad (2)$$

Here SFP denotes significant frequent pattern and φ represents significant weightage.

V Conclusion

This paper shows about how to prevent and detect the Skin cancer by using clustering techniques. k- means clustering algorithm for separating relevant and non relevant data to skin cancer. The major frequent patterns are discovered using MAFIA algorithm. In future implementation system using c#.net to predict skin melanoma risk level with suggestions which is easier, cost reducible and time savable than the existing work.

Reference:

[1] Muhammed Akmal Sapon , Khadijah Ismail and Suehazlyn Zainudin , Prediction of Diabetes using Artificial Neural Network, 2011 International Conference on Circuits, System and Simulation IPCSIT, Vol.7, pp. 299-303, Singapore, 2011.

[2] Zakaria Nouir, Berna Sayrac , Benoit Fourestie , Walid Tabbara and Francoise Brouaye, "Generalization Capabilities Enhancement of a Learning System by Fuzzy Space Clustering" , Journal of Communications, Vol. 2, No. 6, pp. 30-37, November 2007.

[3] Doug Burdick, Manuel Calimlim and Johannes Gehrke, "MAFIA: A Performance Study of Mining

Maximal Frequent Itemsets" Proceedings of the 17th International Conference on Data Engineering, pp. 443-452, April 02-06, 2001.

[4] Manaswini Pradhan, Dr. Ranjit Kumar Sahu, " Predict the onset of diabetes disease using Artificial Neural Network(ANN)", International Journal of Computer Science & Emerging Technologies (E-ISSN: 2044-6004), pp.303-311 Volume 2, Issue 2, April 2011.

[5] Carlos Ordonez, " Programming the K –means Clustering Algorithm in SQL", Proc. ACM Int'1 Conf. Knowledge Discovery and Data Mining, pp. 823-828, 2004.

[6] Jiang Su and Harry Zhang, " A Fast Decision Tree Learning Algorithm ", American Association for Artificial Intelligence, 2006.

[7] www.news-medical.net/health/skin-cancer-classification.aspx

[8]Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers,second Edition, (2006).

[9] Kawsar Ahmed, Tasnuba Jesmin and Md. Zamilur Rahman, " Early Prevention and Detection of Skin Cancer Risk using Data Mining" International Journal of Computer Applications (0975 – 8887) Volume 62– No.4, January 2013.

[10] Amandeep Kaur Mann and Navneet Kaur, " Survey Paper on Clustering Techniques", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013.

[11] Kawsar Ahmed, Abdullah-Al-Emran, Tasnuba Jesmin, Roushney Fatima Mukti,, Md Zamilur Rahman and Farzana Ahmed, "Early Detection of Lung Cancer Risk Using Data Mining", Asian Pacific Journal of Cancer Prevention, Vol 14, 2013.