ISSN:0975-9646

Feature Selection Using Genetic Algorithm and Classification using Weka for Ovarian Cancer

Priyanka khare[1] Dr.Kavita Burse[2]

[1]*M.Tech. scholar, CSE, Oriental College of Technology Bhopal, India*
[2] *Director, Oriental College of Technology Bhopal, India*

Abstract— Data mining is the process of extracting use full information from the large datasets. In data mining classification is a technique used to predict group membership for data instance. The purpose of organization is to analyze the input data and to develop an accurate description or model for each class using the features present in the data. The aim of this paper to investigate the performance of different classification methods on clinical data. Before applying classification algorithm relevant feature are selected by applying genetic algorithm. Feature selection is the process of extracting relevant subset. After that weka tool is used for classifying the dataset.

Keywords data mining, weka, feature selection, genetic algorithm.

1. INTRODUCTION

The cost of medical and healthcare is rising very quickly than the willingness and the capacity to pay for it. Simultaneously, due to the accessibility of computers, progressively data is suitable accumulate. Such a huge amount of data cannot be processed by the expert in a short period to make diagnosis, prognosis and treatment schedules in less time. Therefore, data mining has become important to the medical healthcare world [1] [2]. A important step in the data mining is data pre-processing, as the feature of decisions is based on the value of data. Enhancing the medical database improve the feature of medical diagnosis. Data pre-processing steps are data cleaning, data integration, data transformation and data reduction (feature subset selection). A few attributes of datasets possibly redundant as their information may be contained in other attributes. More attributes can affect the computation time for the diagnosis accuracy. Some data in the dataset may not be useful for diagnosis and thus can be eliminated before learning. The goal of feature choosing is to find a least set of attributes so that the resulting probability distribution of the data classes is as close as likely to the original distribution obtained by all attributes [1]. Irrelevant, redundant, or noisy data can be removing through data reduction process. This reduction gives speed on data mining algorithm, and improving mining performance such as predictive accuracy and result comprehensibility [3]. In this work we present use of genetic algorithm for feature selection. The resulted reduces features gives as input to five classifiers.

2. RELATED WORK

The paper [4] presents a genetic programming based methodology to classify diabetes data. To facilitate the selection of features and for evaluating the effectiveness of diabetes features various methodologies have been used in this research.

Ovarian cancer diagnosis is a vital study because early detection and accuracy staging are the keys to increase the survival rate of the patient. In papers, [5] propose a novel hybrid intelligent system, that derives simple yet convincing fuzzy inference rules to diagnose ovarian cancer and determine its stage according to the level of seriousness.

3. PROPOSED FRAMEWORK

3.1 Feature Selection

Accessing useful data from subjectively large data collections or data streams is now of unique interest inside the data mining area. Researchers and practitioners understand that the feature selection is an important module to successful data mining. Feature choosing , as a process of selecting a relevant feature of original features according to definite condition, is an significant and regularly used as a decrease technique for data mining.

Feature selection has been an vigorous research area for decades in fields such as machine learning and data mining. A typical feature selection process consists of four basic steps namely, subset generation, subset estimation, stopping condition, and result confirmation. Subset generation is a search method that gives candidate feature subsets for evaluation based on a definite search strategy. Each candidate division is evaluated and compared with the previous best one according to an assured evaluation criterion. If the new one is better than the previous on then the previous one is removed. The process of subset generation and evaluation is frequent until a given stopping decisive factor is fulfilled, and then the particular most excellent subset typically needs to be validated by prior knowledge or different tests by means of synthetic and/or real world data sets. Feature selection can be establish in many areas of data mining such as classification, clustering, association rules, and regression. Feature collection algorithms intended with different evaluation criteria mostly fall into three categories [6]: the filter, wrapper, and hybrid models.

3.2 Genetic Algorithms

GA is a stochastic general search method, capable of effectively exploring large search spaces, which is usually required in case of attribute selection. Further, unlike many search algorithms, which perform a local, greedy search, GAs performs a global search. The Gas simulates the processes in natural systems for evolutions based on the principle of "survival of the fittest" given by Charles Darwin [7]. A genetic algorithm mainly composed of three operators: reproduction, crossover, and mutation. Reproduction selects good string (subset of input attributes); crossover combines good strings to try to generate better offspring's; mutation alters a string locally to attempt to create a better string. The string consists of binary bits: 1 to represent selection of attribute else 0 to drop that attribute. In each generation, the population is evaluated and tested for termination of the algorithm. If the termination criterion is not satisfied, the population is operated upon by the three GA operators and then re-evaluated. This process is repeated for specified number of generation.

3.3Weka tool

Weka (Waikato Environment for Knowledge Analysis) is a machine learning tool written in Java, developed at the University Of Waikato, New Zealand. Weka is also a bird found only on the islands of New Zealand. Weka is free software accessible under the GNU General Public License. The Weka workbench is a collection of state-of-the-art machine learning algorithms and data pre-processing tools. In weka algorithms can applied directly to a dataset or called from your own Java code. It contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also compatible for developing new machine learning schemes. In weka we can pre-process a dataset, feed it into a learning scheme, and analyse the resulting classifier and its performance, all without writing any program code at all. Getting to know data is integral part of the work, and many data visualization facilities and data pre-processing tools are provided [8].

Proposed Algorithm

```
Step 1: Start
Step 2: Read ovarian cancer dataset.
Step 3: Select best feature.
Step 4: Set the number of desired features.
Step 5: Set the fitness function.
Step 6: call the Genetic Algorithm
Step 6.1: Construction of the first generation
Step 6.2: Selection
   While stopping criterion not met do
Step 6.3: Crossover
Step 6.4: Mutation
Step 6.5: Selection
     End
Step 7: Weka classification
Step 7.1: Loading data in weka.
Step 7.2: Apply Interquartile Range
Step 7.3: Apply various algorithm
Step 7.4: Compare result.
Step 8: Compare accuracy
```

4. EXPERIMENTAL RESULTS

Ovarian cancer data set is used for data classification. Before applying classification algorithm relevant feature are selected by using feature selection methods. Feature selection is done by using genetic algorithm in Mat Lab and for classification weka tool is used.

The results are tabulated in Table 1and Table 2 .The overall ovarian cancer dataset features are reduced from a large dataset of size 15154×216 to 20×216. This reduced data set is used for classification, before the classification "Interquartile range "is applied so two more attribute named outlier and extreme values. After this we get the data set of size 22 x 216. Then the reduced dataset is loaded in weka to classify results the rules and the result is given in accuracy in percentage in table3.

TABLE 1. GENETIC ALGORITHM FOR OVARIAN CANCER

DATASET	NO. of Attributes	No. of instances	NO. of Classes
Ovarian Cancer (without GA)	15154	216	2(Benign, Cancer)
Ovarian Cancer (with GA)	20	216	2(Benign, Cancer)

TABLE 2. INTERQUARTILE RANGE

DATASET	NO. of Attributes	No. of instances	NO. of Classes
Ovarian Cancer (with GA)	22	216	2 (Benign, Cancer)

Table-3: DIFFERENT PERFORMANCE METRICES RUNNING IN WEKA

CLASSIFIER	CORRECTLY CLASSIFIED INSTANCES	TP RATE	FT RATE	PRECION	RECALL	F.MEASURE	ROCAREA	TIME
Bayesnet	133 (61.57%)	0.608	0	1	0.608	0.757	1	0.1sec
SMO	212 (98.14%)	1	1	0.981	1	0.991	0.5	0.7sec
Simple Logistic	216 (100%)	1	0	1	1	1	1	11.6sec
ONE-R	215 (99.53%)	0.995	0	1	0.995	0.998	0.998	0.2sec
ZERO- R	212 (98.14%)	1	1	0.981	1	0.991	0.198	0.2sec

In this study, we examine the performance of different classification methods that could generate accuracy and some error to diagnosis the data set. According to above Table 3, we can clearly see the best algorithm in WEKA is Bayesnet classifier with an accuracy of 61.57% because it takes 0.1 seconds for classifying the dataset. The total time required to build the model is also a crucial parameter in comparing the classification algorithm.

5. Conclusion And Future Work

The objective of this study is to evaluate and investigate FIVE selected classification algorithms based on WEKA. The best algorithm in WEKA is Bayesnet classifier with an accuracy of 61.57% because it takes 0.1 seconds for classifying the dataset. They are used in various healthcare units all over the world. The future work will be focused on using the other classification algorithms of data mining. It is a known fact that the performance of an algorithm is dependent on the domain and the type of the data set. Hence, the usage of other classification algorithms like machine learning will be explored in future.

References

[1] Adam Woznica, Phong Nguyen, Alexandros Kalousis, "Model mining for robust feature selection", KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM New York, NY, USA, PP 913-921, 2012.

[2] Asha Gowda Karegowda, M.A.Jayaram, A.S .Manjunath, "Feature Subset Selection using Cascaded GA & CFS: A Filter Approach in Supervised Learning", International Journal of Computer Applications (0975 – 8887), Vol. 23– No.2, June 2011.

[3] MIT Lincoln Laboratory: http://www.ll.mit.edu/IST/ idaval/.

[4] Muhammad Waqar Aslama, Zhechen Zhu, Asoke Kumar Nandi, “Feature generation using genetic programming with comparative partner selection for diabetes classification”, Expert Systems with Applications 40, Pages 5402–5412, Elsevier, 2013.

[5] Di Wanga, Chai Queka, Geok See Ng, “Ovarian cancer diagnosis using a hybrid intelligent system with simple yet convincing rules”, Applied Soft Computing 20, Pages 25–39, Elsevier, 2014.

[6] Ian H.Witten and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Second Edition, Morgan Kaufmann Publishers, Elsevier Inc. 2005.

[7] D. Goldberg,” Genetic Algorithms in Search, Optimization, and Machine learning”, Addison Wesley, 1989.

[8] Shilpa Dhanjibhai Serasiya and Neeraj Chaudhary,” Simulation of Various Classifications Results using WEKA” International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-1, Issue-3, August 2012