International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 6 | ISSN : 2456-3307

A Hybrid Approach for the Fertility Rate Analysis In Human Beings Using Classification Algorithms

K. Malathi[1], Mrs. K. Sivaranjani[2]

[1]M.Phil. Scholar, Department of Computer Science, Bishop Heber College (Autonomous), Tiruchirappalli, Tamil Nadu, India

[2]Assistant Professor, Department of Information Technology, Bishop Heber College (Autonomous), Tiruchirappalli, Tamil Nadu, India

ABSTRACT

A decline in human sperm quality and quantity has been reported in numerous Western countries. This observation was also accompanied by an increase in urogenital malformations. The need for epidemiological studies dealing with unbiased populations in order to understand the causes of these observations is obvious. In this work three classification techniques of Data Mining are combine using Human disease datasets from University of California, Irvine (UCI) Machine Learning Repository. Accuracy and time complexity for execution by each classifier is observed. These algorithms were Naïve Bayes, SVM (Support Vector Machines) and hybrid classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers. This thesis discussed various techniques which are able to classify with future human semen analysis data will increase or decrease better than level of significance. Also, it investigated various global events and their issues classify on Human disease. It supports numerically and graphically. Classification is used to classify each item in a set of data into one of predefined set of classes or groups.

Keywords : Naïve Bayes, SVM, Hybrid Classifier

I. INTRODUCTION

The Healthcare industry is among the most information intensive industries. Medical information, knowledge and data keep growing on a daily basis. It has been predictable that a sharp care hospital may generate five terabytes of data a year. The aptitude to use these data to extract useful information for quality healthcare is crucial. Medical informatics plays a very important role in the use of clinical data. In such discoveries pattern recognition is important for the diagnosis of new diseases and the study of different patterns found when classification of data takes place. Computer assisted information retrieval may help support quality decision making and to avoid human error. Although human decision-making is often optimal, it is poor when there are huge amounts of data to be classified. Also efficiency and accuracy of decisions will decrease when humans are put into stress and immense work. Imagine a doctor who has to examine 5 patient records; he or she will go through them with ease. But if the number of records increases from 5 to 50 with a time constraint, it is almost certain that the accuracy with which the doctor delivers the results will not be as high as the ones obtained when he had only five records to be analyzed.

This lead to the use of data mining in medical informatics, the database that is found in the hospitals, namely, the hospital information systems (HIS) containing massive amounts of information which includes patients information, data from laboratories which keeps on growing year after year. With the help of data mining methods, useful patterns of information can be found within the data, which will be utilized for further research and evaluation of reports. The other question that arises is how to classify or group this massive amount of data. Automatic classification is done based on similarities present in the data. The automatic classification technique is only proven truihtful if the conclusion that is drawn by the automatic classifier is acceptable to the clinician or the end user.

In both case studies presented in this thesis we know the categories or outcome with respect to the different cases, thus we will concentrate mainly on supervised learning methods in data mining. Suppose information regarding classification or outcomes of the cases were not present, the result would be the use of unsupervised learning methods. Although none of the data makes any sense to the complier or the machine learning algorithms, text data are rather easier for classification and categorization than other types of data. Also with text data, results are more accurate and are obtained more quickly than with other types of data.

Lastly, some of the data mining algorithms make use of rules, which are required for categorization. Rules are obtained based on patterns present in the training data set, which are extracted by the various data mining algorithms. This rule-based stage can be performed on a desktop. Once these rules are obtained they can be stored on a PDA. Inputs regarding the patient can be fed to the PDA and classification of the input can take place based on the rules stored in the device in real time.

II. RELATED WORK

In [1] Azam Asilian Bidgoli, Hossein Ebrahimpour Komleh, Seyed jalaleddin Mousavirad et al presents Infertility problem is an important issue in recent decades. Semen analysis is one of the principle tasks to evaluate male partner fertility potential. The artificial neural network (ANN) is a powerful data mining tool that can be used for this goal. A genetic algorithm to optimize the structure of artificial neural network to classify the semen samples is present in this concept. The appropriate structure of artificial neural network is obtained based on this method. The method outperforms SVM, decision tree and naive Bayesian. Thus, this paper attempts to resolve it by using the bootstrap method. The performance of the proposed algorithm is significantly better than the previous works. The accuracy and area under curve (AUC) of method 84.83% and 0.84 respectively that are higher than in our experiments on a real fertility diagnosis dataset that is a good improvement compared with other classification methods.

In [5] Macmillan Simfukwe, Douglas Kunda, andChristopher Chembe et al presents One of the most worldwide health care concerns in the final decades has been the very low in the fertility rates. The problem is said to be more severe among the male population. Research has shown that environmental factors and life style habits have an impact on the quality of semen. Orthodox analysis of seminal quality utilizes a laboratory approach, regarding exclusive tests, which are also infrequently unable to process by the patient. In this paper they intend Naïve Bayes and Artificial Neural Network classifiers for the categorization of seminal quality, based on surroundings factors and life style habits. Comparisons between the two classifier models illustrate that their accuracy rate is the identical and stands at 80%, on the training set.

In [6] Maria Luisa Davila Garcia, Daniel A. Paredes Soto, Lyudmila S. Mihaylova et al presents in this paper used the framework is based on Speeded Up Robust Features (SURF), combines with Bag of Features (BOF) models and histogram of Oriented Gradients (HOG) is used to extract features and support vector machines is applied for their classification. The approach proposed is able to analyze images from unstained and motile sperm cells. A reliable approach to analyze the large amount of data generated from the SURF feature extraction. The analysis comprises the selection of the most representative features that best describe each of the four classes defined in this work for the classification of new instances. The extracted from the sperm cell images into normal, abnormal, non cell categories.Using this framework the author achieve classification results with an accuracy of 90% with the SURF approach compared to 78 % with the HOG approach.

In [7] Mendoza-Palechor, Fabio E, Ariza-Colpas, Paola P, Sepulveda-Ojeda, Jorge A et al presents semen analysis is standard for routine diagnosis of infertile couples studies through the sperm count, and it´s strongly related to male infertility, and they also express the importance of sperm concentration in male infertility, since it is a relevant factor in diagnosis of this disease. To accomplish this, it was necessary to make a segmentation process of the patients with or without fertility problems through the simple k-means method, and then the data mining techniques trees decision, support vector machine.Bayesian networks and k-nearest neighbors were applied and obtained that three methods had TpRate 100%, FpRate 0%, Precision 100%, Recall 100%, the fourth method, using Bayesian networks had TpRate 98.4%, FpRate 1.5%. Precision 88 % Recall 88%. These results clearly show these methods can achieve high percentages in all metrics, and confirms that the proposed method can be efficient and accurate to detect fertility rates in patients, improving the results accomplished by researchers in previous studies.

In [8] Muratori, M., Marchiani, S., Tamburrino,L., Forti, G., Luconi, M., Baldi, E presents a result of advancements in ART, understanding the potential implications of genetic disorders for infertile couples is critical. Even though ART method allows infertile male to father their own child without knowing the cause of their infertility, it also carries the potential risk of transmission of genetic or epigenetic aberrations to the offspring). Whether advanced ART techniques are associated with increased birth defects is still debated, and search for alternative options should go on. There is no evidence to advise one particular treatment option over another. The choice should be based on hospital facilities, convenience for the patient, medical staff, costs and drop-out levels.

In [10] Pagani, R., Cocuzza, M., Agarwa,l A presents Andrology or the male counterpart of gynaecology has gradually emerged as a speciality in its own right . The histories of progress in the fields of genetics and andrology are rich and include many breakthroughs. Many different causes underlie male factor infertility and its treatment is difficult at best. The clinician first needs a thorough understanding of male reproductive anatomy and physiology. Most hormonal imbalances can be readily identified and successfully treated no surgically. Additionally, a large number of patients require surgery to improve sperm production or to improve sperm delivery. However, the treatment of men with unexplained idiopathic infertility remains a challenge. There are three kinds of infertility treatment: Medical treatment, Surgical treatment and Assisted Reproductive Technology (ART).

In [12] EM Van Raemdonck, Ata-ur-Remand, M. Luisa Davila-Garcia, Lyudmila Mihaylova et al presents an algorithm for analyzing the morphology of motile sperm. The analysis of human sperm as part of infertility investigations or assisted conception treatments is a labor intensive process reliant upon

the skill of the observer and as such prone to human error. Therefore, there is a force to expand automated systems that can sufficiently assess the concentration, motility and morphology of live sperm. A computer assisted sperm analysis algorithm is proposed for accurately and efficiently analyzing the morphology of sperm. Techniques for eliminating the background, segmentation of the cells and template matching techniques are successfully used to first classify the cells into motile and immotile cells and secondly the motile cells are classified into normal and abnormal cells. The performance of the proposed algorithm is evaluated and accurate classification of cells with the accuracy of 86%.

III. METHODOLOGY

A. Classification

DataMining, or information Discovery in Databases (KDD) because it is as well known, is that the nontrivial extraction of implied, previously unknown, and almost certainly helpful info from knowledge. This encompasses a number of various technical approaches, like clustering, data account, learning classification rules, finding dependency networks, analyzing changes, and detection anomalies.

B. SVM Classifier

Support Vector Machine (SVM) could be a supervised machine learning algorithm which might be used for each classification and regression challenges. However, it's largely employed in classification issues.During this algorithm, here plot every data item as a degree in n-dimensional area(where n is range of options you have) with the worth of every feature being the worth of a specific coordinate. Then, the perform classification by finding the hyper-plane that differentiate the two categories very well.

Classifying knowledge could be a common task in machine learning. Suppose some given knowledge points every belong to 1 of 2 categories, and therefore the goal is to make your mind up that category a new information are in. within the case of support vector machines, a knowledge purpose is viewed as a p-dimensional vector p numbers), and that we wish to understand whether or not we will separate such points with a (p-1)-dimensional hyperplane. this is often known as a linear classifier. There square measure several hyperplanes that may classify the information. One cheap selection because the best hyperplane is that the one that represents the biggest separation, or margin, between the 2 classes. thus we decide the hyperplane in order that the gap from it to the nearest information on either side is maximized. If such a hyperplane exists, it's referred to as the most-margin hyperplane and therefore the linear classifier it defines is thought as a maximum margin classifier; or equivalently, the perceptron of best stability.

initialize yi = YI for i ∈ I
REPEAT
compute SVM solution w, b for data set with imputed labels
compute outputs fi = hw, xii + b for all xi in positive bags
set yi = sgn(fi) for every i ∈ I, YI = 1
FOR (every positive bag BI)
IF (P i∈I (1 + yi)/2 == 0)
compute i ∗ = arg maxi∈I fi
set yi∗ = 1
END
END
WHILE (imputed labels have changed)
OUTPUT (w, b)

C. Naive Bayes Classification

Naive Bayes is a easy probabilistic classifier based on applying Bayes' theorem (or Bayes's rule) with strong independence (naive) suppositions.

Bayes's rule:

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

Where,

P(H) is the prior probability of H.

P(E) is the prior probability of E.

P(H|E) is the posterior probability of H given E.

P(E|H) is the posterior probability of E given H.

Since the denominator P(E) is the probability of the proof without any knowledge of the event A, and since the theory A can be true or false.

D. HYBRID algorithm

These hybrid are said to the severance or the classifier restraint for the text classification. Best unraveling hybrid plays necessary role within the distinguishing the support vector. As before acting the development of the best hybrid. Here insert all the coaching knowledge points. There are several hyper planes generated with the assistance of support vector machine values. There are an infinite number of hybrid (the dotted lines) may well be generated, however there's just one hybrid (the solid line) that may optimally separate the info points from completely different classes. These values are known with the assistance of size; value allotted for the group action in GB or MB, the frequency of the dataset is taken. As these values iteratively notice for all records beneath the sphere. thence the dataset size remains same no more elimination is performed during this module. Finally, were distinguishing the minimum privacy protective value. That the minimum answer mentioned herein is somewhat pseudo minimum as a result of an upper bound of joint privacy leakage is simply an approximation of its actual price.

candidateSV = { closest pair from opposite classes }
while there are violating points do
Find a violator
candidateSV = candidateSV S violator
if any αp < 0 due to addition of c to S then
candidateSV = candidateSV \ p
repeat till all such points are pruned
end if
end while

IV. EXPERIMENTAL RESULT AND DISCUSSION

A. Dataset Used

The experiment is conducted over three stages of the proposed work namely existing feature selection and classification algorithms and the proposed feature selection and classification algorithm. In this study, a dataset for breast cancer is derived and used for the experimentation from the UCI machine learning repository database. The dataset comprised of ten attributes with 287 instances. This dataset was given as an input to the most popular data mining tool WEKA 3.6 for analyzing the correct accuracy prediction of various classification algorithms. Table 2.describes the attributes with its possible range.

Table 2. Human Activity Data Attributes

S. No	Attribute Name	Description
1	Class	no-recurrence-events, recurrence-events
2	age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
3	Child	0-1
4	accident	1-2
5	smoking	1-2, 2-1
6	Season	1,2,3,4 (Winter, Spring, Monsoon, Autumn)
7	Alcohol	0,1
8	edrel	small glands located on top of each kidney

Table II shows that accuracy of various individuals and ensemble model with three data partitions. Figure 4.1 also shows that Error rate and classification accuracy of different individuals and ensemble models

Here applied different classification techniques to Human activity dataset and the error results obtained is tabulated in table given below.

Table 2. Combined of Classification Algorithms

S. No	Error rate	SVM	Naïve Bayes	Hybrid
1	Kappa statistic	0	0	0
2	Mean absolute error	0.1762	0.2781	0.1573
3	Root mean squared error	0.4198	0.33389	0.3241
4	Relative absolute error	62.8956	99.3891	99.58
5	Root relative squared error	98.7585	99.7782	99.86
6	Time (s)	0.11	0.02	0.02
7	Accuracy	83.2776	83.2776	84.2457

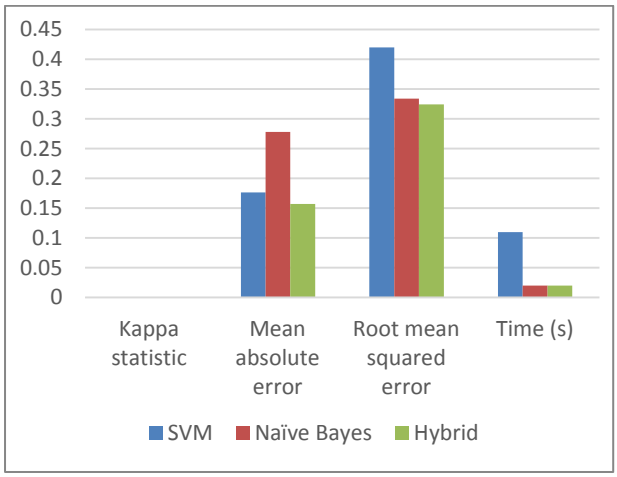

Figure 4.1. Classification using SVM, Naïve Bayes and Hybrid

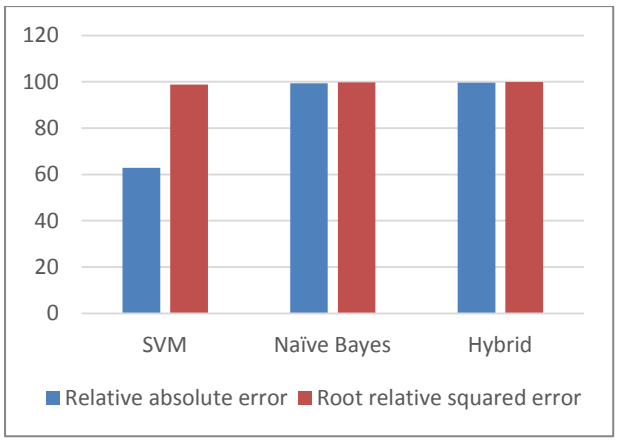

Figure 4.2. Classification using SVM, Naive Bayes and Hybrid

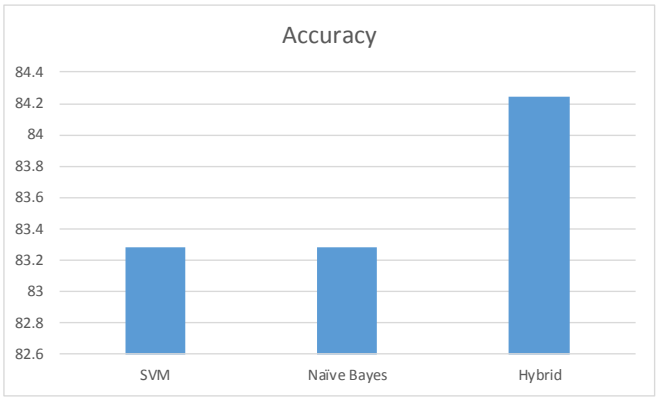

Figure 4.3 .Classification Accuracy using SVM, Naïve Bayes and Hybrid

Naive Bayes and Hybrid

The fast classification time by Hybrid algorithm is due to the absence of calculation in its classification process. The classification model is created outside the application, using Weka data mining tool and Java. And the model is converted into data before being incorporated into the application. Classification by way of following the classify is faster than the ones that need calculation as in the case of Naïve Bayes and Hybrid.

V. CONCLUSION

A classification method is used in searching the alternative design. There are three classifiers used in this experiment namely Naïve Bayes, SVM, and Hybrid. This experiment shows that Hybrid is the fastest and SVM is the slowest. The fast classification time of Hybrid because there is no calculation in its classification. Human Activity classification based on a very challenging and crucial task in the field of health care. Data mining techniques will be the gem stone in healthcare sector. It is able to achieve a classification accuracy, which matches or outperforms most of the proposed techniques described in the literature. Various data mining techniques have proven to be very helpful in decision making. The research undertook an experience on application of three data mining algorithm to classify the human semen analysis and to compare the based method of classification. Results depend on the evaluation methodology focused on a data classification tasks. New experiments should be carried out to explain why the naive Bayes behave so well on one against one classification tasks in contrast to its behavior on one against all tasks. Here are also interested to understand more precisely SVM behavior as it exhibited an uncommon performance pattern shaped as a wave when the size of the feature space increases. Finally, to recommend a Hybrid classifier with suitable parameter settings, a way to characterize classification tasks should be investigated, eventually via the use of a meta-learning strategy.

The proposed framework model can be used to analyze the existing work, identify gaps and provide scope for further works. The researchers may use the model to identify the existing area of research in the field of data mining in other dataset and use of other classification algorithms. Additionally, the potential of imaging in the submicron regime enables the characterization of ultra structures as new diagnostic details within samples in routine pathology that are not visible using conventional medical imaging techniques or light microscopy in semen diagnostics. As a future work Dimensionality reduction can be applied to the data set so that it will reduce number of test and time required to diagnose the disease. In future, Applying classification in image techniques produces higher accuracy results or ratio.

VI. REFERENCES

[1]. Azam Asilian Bidgoli1, Hossein Ebrahimpour Komleh1, Seyed jalaleddin Mousavirad "Seminal Quality Prediction using Optimized Artificial Neural Network with Genetic Algorithm" International journal of andrology 2015.

[2]. D. R. Grow, S. Oehninger, H. J. Seltman, J. P. Toner, R. J. Swanson, T. F. Kruger, and S. J. Muasher, "Sperm morphology as diagnosed by strict criteria: probing the impact of teratozoospermia on fertilization rate and pregnancy outcome in a large in vitro fertilization population," Fertility and Sterility, vol. 62, no. 3, pp. 559–567, 1994.

[3]. A. S. Hamberger L., K. Lundin and B. Soderlund, "Indications for intracytoplasmic sperm injection," Human Reproduction, vol. 13, no. 6, pp. 128–133, 1998.

[4]. K. Lundin, B. Sderlund, and L. Hamberger, "The relationship between sperm morphology and rates of fertilization, pregnancy and spontaneous abortion in an invitro fertilization/intracytoplasmic sperm injection programme," Human Reproduction, vol. 12, no. 12, pp. 2676–2681, 1997.

[5]. Macmillan SimfukweP 1 P, Douglas KundaP 1 P and Christopher Chembe "Comparing Naive Bayes Method and Artificial Neural Network for Semen Quality Categorization" Information Technology, Information Systems and Electrical Engineering (ICITISEE) 2017.

[6]. Maria Luisa Davila Garcia, Daniel A. Paredes Soto, Lyudmila S. Mihaylova. "A Bag of Features Based Approach for Classification of Motile Sperm Cells" IEEE International Conference on Internet of Things (iThings).

[7]. Mendoza-Palechor, Fabio E.; 2 Ariza-Colpas, Paola P.; 3 Sepulveda-Ojeda, Jorge A. De-la-Hoz-Manotas, Alexis, 5 Piñeres Melo, Marlon "Fertility Analysis Method Based on Supervised and Unsupervised Data Mining Techniques" IEEE International Conference on Data Science and Advanced Analytics (DSAA) 2015.

[8]. Muratori, M., Marchiani, S., Tamburrino,L., Forti, G., Luconi, M., Baldi, E. Markers of human sperm functions in the ICSI era. Frontiers in Biosciences. 16:1344-1363.

[9]. S. Oehninger, A. A. Acosta, M. Morshedi, L. Veeck, R J. Swanson, K. Simmons, and Z. Rosenwaks, "Corrective measures and pregnancy outcome in vitro fertilization in patients with severe sperm morphology abnormalities," Fertility and Sterility, vol. 50, no. 2, pp. 283–287, 1988.

[10]. Pagani, R., Cocuzza, M., Agarwa,l A.. Medical and surgical treatment of male infertility. Archives of Medical Sciences. 1A: S70-S83, 2009

[11]. F. X. Su Hai and L. Yang, "Robust cell detection of histopathological brain tumor images using sparse reconstruction and adaptive dictionary selection," Medical Imaging, IEEE Transactions, vol. 35, no. 6, pp. 1575–1586, 2016.

[12]. EM Van Raemdonck, Ata-ur-Rehman, M. Luisa Davila-Garcia, Lyudmila Mihaylova, Robert F. Harrison , Allan Pacey "An Algorithm for Morphological Classification of Motile Human Sperm Lore"IEEE CONFERENCE ON Sensor Data Fusion: Trends, Solutions, Applications (SDF) 2015.