Investigation of Differential Evaluation Optimization Algorithm for Medical Data Classifications

Shymaa Mohammed Jameel,

Iraqi Commission for Computers& Information Informatics Institutes For Postgraduate Studies Baghdad, Iraq

Summary
Medical informatics is chiefly inductive and information intensive science where observation and analysis of comprehensive clinical data can lead to complex and powerful evidence based decision support systems Conversely, most medical institutions are still keeping a large amount of medical data in narrative form resulting in huge volume of potential information with limited or no utility and accessibility In this research, a meta-heuristics approached (Differential Evolution) as a proposed algorithm in order to enhance the quality of performance (accuracy) of medical applications that help medical staff to take the correct decision within a good enough computational time.

Key words:
Investigation, Medical Data

1. Introduction

Medical informatics is chiefly inductive and information intensive science where observation and analysis of comprehensive clinical data can lead to complex and powerful evidence based decision support systems. One of the primary goals of these automated systems is to make information more accessible, representative and meticulous in a quick span [4]. Furthermore, they have gained increased importance in the recent years as it can even outperform a human expert in some cases in diagnosing diseases as the process is highly subjective and fundamentally depends on the experiences of the assessor and his/her interpretation on the information[4].

Conversely, most medical institutions are still keeping a large amount of medical data in narrative form resulting in huge volume of potential information with limited or no utility and accessibility. An effort to exploit this data poses multiple challenges as it involves processing free text data with the presence of acronyms, synonyms, negation and dependence on temporality. Thus, in this work we try to explore different challenges that arise while trying to identify obese patients and the co-morbidities exhibited by them based on the narrative patient record.

In this research, a meta-heuristics approached (Differential Evolution) as a proposed algorithm in order to enhance the quality of performance (accuracy) of medical applications that help medical staff to take the correct decision within a good enough computational time. A several cases will be solves in this research taken from [14] such as for Breast Cancer (BC), Pima Indian Diabetes(PID), Hagerman Surgery Survival (HSS), Liver disorders(LD) , Wisconsin Breast Cancer (WBC), Statlog Heart (SH), Australian Credit Approval (ACA),Parkinsons (PA), SPECTF, German Credit Data (GCD) and Appendicitis(AP).

Up to date, human analysts need advance specail computational tools to process and comprehend such large and complex medical datasets.

2. Motivation

Recently,solving medical problems gets a higher priority for scienstist . The main question is How to solve medical problems in an automated way? Medical datamining becomes one of the most popular topics in datamining community.

The motivation of this research is to extract usefull knowledge from medical datasets and thus discoveringing decision-making insights for the diagnosis and treatement of diseases. The first challenge of these medical datasets is sortting the data and recognise it. In the typical setting a dataset of historic data, which describe some types of disease or a medical disorder, is assumed to be available. Medical datasets consist of records of patients describing physical and laboratory examinations related to that type of disease or medical disorder. So, the computational challenge is how to develop a diagnostic system, that could assist the diagnostic this type of disease based on the knowledge extracted from the historic medical datasets. Such datasets for Breast Cancer (BC), Pima Indian Diabetes(PID), Haberman Surgery Survival (HSS), Liver disorders(LD), Wisconsin Breast Cancer (WBC), Statlog Heart (SH), Australian Credit Approval (ACA), Parkinsons (PA), SPECTF, German Credit Data (GCD) and Appendicitis(AP).

Within a health care setting, it is often desirable from both clinical and operational perspective to capture the uncertainty and variability amongst a patient population, for example to predict individual patient outcomes, needs.

Manuscript received January 5, 2014
Manuscript revised January 20, 2014

Homogeneity brings the benefits of increased certainty in individual patient needs and resource utilization, thus providing an opportunity for both improved clinical diagnosis and more efficient planning and management of health care resources.

3. Literature review

A number of classification algorithms are considered and evaluated for their relative performances and practical usefulness on different types of health care datasets. The algorithms are evaluated using four criteria: accuracy, computational time, comprehensibility of the results and ease of use of the algorithm to relatively statistically naive medical users. The researchers have shown that there is not necessarily a single best classification tool, but instead the best performing algorithm will depend on the features of the dataset to be analyzed, with particular emphasis on health care data [13].

Moreove, current classification approaches may work well with some training datasets, while they may perform poorly with other datasets for no obvious reason. Pham and Triantaphyllou (2009) [1] [2] [3] argued that such approaches usually did not try to achieve a balance between fitting and generalization when they inferred models from datasets. Thus, the models they infer may suffer from over fitting and over generalization problems, and this causes their poor performance. Over fitting occurs when a model can accurately classify data points that are very closely related to the training data but performs poorly with data that are not closely related to the training data. Over generalization occurs when a model erroneously claims to be able to accurately classify vast amounts of data that are not closely related to the training data.

Usually, current classification approaches attempt to minimize the sum of false-negative and false-positive error rates without considering these two error rates in a weighted fashion. They also do not consider the case of having unclassifiable instances. To appreciate the magnitude of this situation, let us consider, for instance, the case of a diagnostic system for some serious diseases (say some kind of aggressive cancer). In a case like this a false- positive diagnosis would subject a patient to some emotional challenge and unnecessary medical tests and treatments. On the other hand, a false-negative diagnosis may cause loss of critical time, which in turn may turn out to be fatal to the patient. It is reasonable to argue here that these two cases of diagnostic errors should be associated with significantly different penalty costs (i.e., much higher for the false-negative case). Similar situations may occur when approving large lines of credit (as the current financial crisis is demonstrating), in oil exploration, issuing evacuation orders to avoid natural disasters (such as when a hurricane is approaching a vulnerable area), classification of targets as enemy or not, and so on. The type of unclassifiable cases is more subtle. Now the system does not make any diagnosis due to limited input information. However, in an extreme case a system may avoid any false-positive and false-negative types by reverting to unclassifiable outcomes for most diagnostic instances. That is, such a system would offer advice only when a new instance is of an obvious nature (i.e., either clearly positive or clearly negative) and avoid any challenging instance. This would result in high numbers of unclassifiable cases. Thus, this outcome should be related to a penalty value as well [2].

In this research, we proposed a new approache (algorithm) in order to enhanc the accuracy, computational time, comprehensibility of the results and ease of use of the algorithm to relatively statistically naive medical users. In order to achive this objective, a meta-huristics algorithms are able to solve these kind of NP-hard problems, The effectiveness of metaheuristic approaches in solving combinatorial optimisation problems particularly in various domians such as scheduling and timetabling problems have been proven in the literature [see 4,5,6,7,8,9,10,11,12]. Moreover, multiobjective algorithms such as Non-dominated Sorting Genetic Algorithm II (NSGA-II) are able to solve this kind of problems. Here in, we will applied such algorithms to sove medical problems such as Breast Cancer (BC), Pima Indian Diabetes(PID), Haberman Surgery Survival (HSS), Liver disorders(LD), Wisconsin Breast Cancer (WBC), Statlog Heart(SH), Australian Credit Approval (ACA), Parkinsons (PA), SPECTF, German Credit Data (GCD) and Appendicitis(AP) [14].

Some results of some previous studies on these data sets are shown in Table 1. As shown in Table 1, some algorithms have been proposed to solve these medical datasets. Such an approaches Early neural networks, fuzzy logic and Support Vector Machine.

There are two main research in this study is:

- Understanding the Medical data structure based on the problem statement.
- Investigating the Differential Evolution approach in order to enhance the accuracy for each case study.

Table1: Datasets used in the experiments.

Dataset	No. of points	No. of attributes	Previous studies	Accuracy (%)
Pima Indians Diabetes (PID)	768	8	Early Neural Networks	76.0
			IncNet	77.6
			Fuzzy approach	77.6
			Flexible Neural-Fuzzy Inference System	78.6
			Fuzzy Neural Networks	81.8
			Statlog Project	78.0
Haberman Surgery Survival (HSS)	306	3	SVMs using linear terms in the objective function	71.2
			Proximal SVMs	72.5
			Integer SVMs	62.7
			Logical functions	66.2
Wisconsin Breast Cancer (WBC)	286	10	C4.5	94.7
			Rule Induction approach	96.0
			Linear Discriminant Analysis approach	96.8
			SVMs	97.2
			Neuro-Fuzzy approach	95.1
			Fuzzy-GA approach	97.5
			Neuro-Rule approach	98.1
			Supervised Fuzzy Clustering	95.6
			Fuzzy Artificial Immune Recognition System	98.5
			Classification through ELECTRE and data mining	94.4
Liver-Disorder (LD)	345	6	RULES-4 algorithm	55.9
			C4.5	65.5
			Reduced SVMs	74.9
			SVMs	69.2
			Least Squares SVMs	94.3
			FAIRS	83.4
Statlog Heart (SH)	270	13	Different approaches in Statlog Project	76.7
			Attribute weighted artificial immune system	87.4
Australian Credit Approval (ACA)	690	14	C4.5	85.7
			Eight genetic programming approaches	83.0
			Different approaches in Statlog Project	86.9
			Extend Naive Bayes	76.7
			SVMs	85.5
Appendicitis (AP)	106	7	Predictive Value Maximization approach	89.6
			Fuzzy Rule-Based Classification System	84.0
			Nefclass	87.7
FourClass	862	2	Fuzzy Kernel Multiple Hyperspheres	99.8
			KNN-SVM	100.0
German Credit Data (GCD)	1000	24	Graph-based relational concept learner	71.5
			DTs	72.9
			SVMs	77.9
Ionosphere (INS)	351	34	Features selection in conjunction with ANNs and KNNs	90.6
			Integration between fuzzy class association rules and SVMs	89.2
Parkinsons (PA)	195	23	ANNs	81.3
			SVMs	91.4
SPECTF	267	45	CLIP 3	77.0
			Rough set-base multiple criteria linear programming	68.0

4. Objectives:

1. To study the current approaches used for medical data classifications. Highlighting the weakness of the current methods.
2. To develop a meta-heuristic algorithm to enhance the accuracy of medical data classifications.
3. To test the proposed approach using medical benchmark datasets (available online).
4. To evaluate the proposed approach using Weka software (an open-source Java application)

5. Methodology

The research has four main phases. It will begin with the literature surveys on medical data classifcations to get a

better understanding about this domain. Surveys on various metaheurictic approaches and its current use as one of the optimisation technique will also be carried out in order to model the proposed approaches and how they can be implemented as optimisation techniques in finding high accuracy value.

The second phase of this research will be the development of formal framework on metaheuristic approaches for medical applications. It involves the refinement of specific feature of metaheuristic approaches where the representation of the solution, the generation of the initial solution and how the improvement algorithms work.

The third phase of the research is the development of metaheuristic approaches and tested on a number of medical datasets [14]. The maximze the accuracy obtained to ehnace the performance of medical staff to take a decision.

Finally the results obtained from phase 3 will be compared with other available approaches in the literatures.This phase is important for the researcher to prove the hypothesis that the theory metaheuristics can be used competitively or better than other AI techniques in solving different medical problems.

6. Differential Evaluation Algorithm(DEA)

DEA is a stochastic, population-based optimisation algorithm introduced by Storn and Price in 1996. It developed to optimise real parameter, real valued functions General problem formulation is:

$X \neq \emptyset$, the minimisation problem is to find

$$x^* \in X \text{ such that } f(x^*) \leq f(x)\ \forall x \in X$$

where:

$$f(x^*) \neq -\infty$$

DEA is an Evolutionary Algorithm. This class also includes Genetic Algorithms, Evolutionary Strategies and Evolutionary Programming

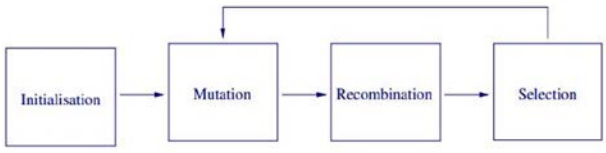

Figure 1: General Differential Evolution Algorithm Procedure

DE is a basic algorithm of the population that employed crossover, mutation and selection operators as in genetic algorithms. The main difference in obtaining better solutions is that genetic algorithms rely on the crossover operation, while in the DE algorithm it is based on the mutation operation. The algorithm uses mutation operation as a search mechanism and selection operation to direct the search towards the potential regions in the search space.

This population then is improved by applying mutation, crossover and selection operators. The main steps of the differential evolution algorithm are given in Figure 2.

Initialization
Evaluation
do while *(termination criterion are met)*
    Mutation
    Crossover
    Evaluation
    Selection
end do while

Figure 2 Differential Evolution algorithm

7. Conclusion

Finally the results obtained from phase 3 will be compared with other available approaches in the literatures.This phase is important for the researcher to prove the hypothesis that the theory metaheuristics can be used competitively or better than other AI techniques in solving different medical problems.

References:

[1] Storn, R. 1997. Differential Evolution, A Simple and Efficient Heuristic Strategy for Global Optimization over Continuous Spaces. Journal of Global Optimization, Vol. 11, Dordrecht, pp. 341-359.

[2] Pham HNA, Triantaphyllou E. Prediction of diabetes by employing a new data mining approach which balances fitting and generalization. In: Yin Lee R, editor. Studies in computation intelligence, vol. 131. Berlin, Germany: Springer; 2008. [p. 11–26, chapter 2].

[3] Pham HNA, Triantaphyllou E. An application of a new meta-heuristic for optimizing the classification accuracy when analyzing some medical datasets. Expert Systems with Applications 2009;36(5):9240–9.

[4] Pham HNA, Triantaphyllou E. The impact of overfitting and overgeneraliza- tion on the classification accuracy in data mining. In: Maimon O, Rokach L, editors. Soft computing for knowledge discovery and data mining. New York, NY, USA: Springer; 2007. [p. 391–431, part 4, chapter 5]. Abdullah S., Shaker K., Shaker, H. 2011. Investigating a Round Robin Strategy over Multi Algorithms in Optimising the Quality of University Course Timetables. International Journal of the Physical Sciences.

[5] Abdullah, S., Shaker, K., McCollum, B. and McMullan, P.: Dual Sequence Simulated Annealing with Round-Robin

Approach for University Course Timetabling. Evolutionary Computation in Combinatorial Optimization, (EVOCOP-2010), Lecture Notes in Computer Scieince, Springer, PP 1-10.

[6] Abdullah, S., Shaker, K., McCollum, B. and McMullan, P. 2010. Incorporating great deluge with Kempe chain neighbourhood structure for the enrolment-based course timetabling problem. Rough Set and Knowledge Technology (RSKT 2010). Lecture Notes in Artificial Intelligence, Springer, 70-77.

[7] Shaker, K. and Abdullah, S. 2010. Controlling Multi Algorithms Using Round Robin for University Course Timetabling Problem. Database Theory and Application, Bio-Science and Bio-Technology (DTA-2010), Lecture Notes in Computer Scieince, Springer, PP 47-55.

[8] Abdullah, S., Shaker, K., McCollum, B. and McMullan, P.: Construction of Course Timetables Based on Great Deluge and Tabu Search. 8th edition of the Meta-heuristic International Conference (MIC 2009) Hamburg, 2009.

[9] Shaker, K. Abdullah, S.: Incorporating Great Deluge Approach with Kempe Chain neighbourhood structure for Curriculum-Based Course Timetabling Problem. 2nd conference on Data Mining and Optimization (DMO '09), IEEE, PP 149 – 153, 2009.

[10] Salwani Abdullah, Hamza Turabieh, Barry McCollum and Edmund K Burke. An Investigation of a Genetic Algorithm and Sequential Local Search Approach for Curriculum-based Course Timetabling Problems. In the 4th Multidisciplinary International Scheduling Conference: Theory and Applications (MISTA2009), Dublin, August 2009, pp 727-731, 2009.

[11] B. McCollum, P.J. McMullan, A. J. Parkes, E.K. Burke, S. Abdullah, An Extended Great Deluge Approach to the Examination Timetabling Problem. In the 4th Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA 2009), Dublin, August 2009, pp 424-434, 2009.

[12] Salwani Abdullah and Hamza Turabieh. Generating University Course Timetable using Genetic Algorithms and Local Search. Accepted to be published in the International Conference on Convergence and Hybrid Information Technology, Busan, Korea. 11th -13th November 2008.

[13] Polat KA, Sahan S, Kodaz H, Gunes S. Breast cancer and liver disorders classification using artificial immune recognition system (airs) with performance evaluation by fuzzy resource allocation mechanism. Expert Systems with Applications 2007;32:172–83.

[14] http://archive.ics.uci.edu/ml/index.html

[15] S.I. Birbil, S.C. Fang. An electromagnetism-like mechanism for global optimization. Journal of Global Optimization 25, 263–282 (2003).