A Novel Approach for optimization of Feature Selection

Duha A. Al-Darras, Suhail M. Odeh, Henry J. Chaya
Department of Computer Information Systems
Bethlehem University
Bethlehem, Palestine
Duha.a.aldarras@gmail.com, sodeh@bethlehem.edu, hchaya@bethlehem.edu

Abstract: *the accuracy of many classification problems is crucial. The number of features for collected data is increasing, and the need to find the best features to be used to increase the accuracy of classification is a necessity. There are several methods of feature selection, but none of them give the absolute best solution and most of them fall in the trap of local optima. This paper presents a new method that searches for the absolute best solution, or a solution which will give a higher classification accuracy rate by using a novel approach that divides the features into two groups: first group and second group of features. After that the method finds the best combination from the two groups to give the maximum accuracy rate. The purpose from this method is to select and find the best feature/s as individual or in groups.*

Keywords: *Feature selection, Machine learning, 1-KNN, Optimization, Genetic Algorithm.*

1. Introduction

Data classification, which is a process of separating data into distinct categories according to a set of known attributes, is used in many applications such as medicine. For example, there are different types of skin diseases: solar keratosis "precancerous", basal cell "cancer' and psoriasis. In a medical diagnosis of skin condition it is crucial to classify the class of the skin disease correctly, since treatment is based on the result of the diagnosis. Normally, the classification algorithms are based on one of the following methods: Decision Trees, Support Vector machines "SVM", Neural Network "NN", Naïve Bayes, Logistic Regression, K- Nearest Neighbours "KNN".

The complexity of a classifier grows exponentially with the number of features. With a large number of features, the performance of the classification methods degenerates. Therefore, feature selection methods try to find the minimum set of attributes to maximize the classification performance [4]. Ranking Methods, Search Based Feature Selection, Greedy Feature Flip Algorithm "G-Flip" are examples of these methods[3][10]. Genetic Algorithm "GA" has been also used as a feature selection method [2] [8] [7]. GA, which is a subset of Evolutionary Algorithms "EA", uses techniques inspired by biological evolution, such as reproduction, mutation, recombination and selection, to generate solutions for optimization problems.

In a genetic algorithm, a population of candidate solutions to an optimization problem is improved toward better solutions. Each candidate solution has a set of properties, which can be mutated and modified. The evolution usually starts from a population of randomly generated individuals. This is an iterative process, with the population in each iteration called a generation. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population. The algorithm in this research will not start with a random subset of features. First, it selects two sub sets of features according to a set of conditions. Then it finds the best combination/s of features from these two subsets.

Even though GA is an optimizer algorithm, in many problems it may have a tendency to converge towards local optima rather than global optimum. This means that the solution it gives is not the best; it is just better compared with another solution. In this paper an optimizer algorithm to give the best solution has been developed. GA will be used as a reference for comparing.

2. Related Research

Many researches that have been done in the features selection field come up with algorithms like G-Flip, GA, sequential search methods: sequential forward selection SFS, sequential backward selection SBS, plus L minus r (L-r)[7][9]. These three algorithms start with a random subset of features then try to modify this subset to reach a better one. Most of them fall in the trap of local maximum. In [10] they used G-Flip algorithm to maximize the margin based evaluation function. Their results showed that G-Flip converges to a local maximum after less than 20 times. Some other algorithms like Importance Score (IS), which is based on a greedy-like search, uses a set of rules to evaluate each feature (give the feature an importance score). In [2] the results suggested that Importance Score method has a high efficiency when dealing with little noise and small number of interacting features. It also suggested that IS has a tendency to get trapped on local peaks caused by noise or interdependencies among features.

What is aimed to be done in this research is to avoid starting with a random set of feature and to get rid of being trapped on local maximum.

3. Technical Approach

3.1. Methodology

- *Data sampling*: The data sampling method used in this research was dividing the data set into two parts: 20% of the data used for testing and the remaining 80% used for training the 1-KNN classifier. This sampling method gave the best result for training and testing the classifier.
- *Classification:* The classification method used in this research was first nearest neighbours (1-KNN) [8]. KNN was used since it is a very simple classifier that works well on basic recognition problems. The result of 1-KNN was examined by 3-KNN and 5-KNN and they almost gave the same conclusion as 1-KNN.(Fig.1)

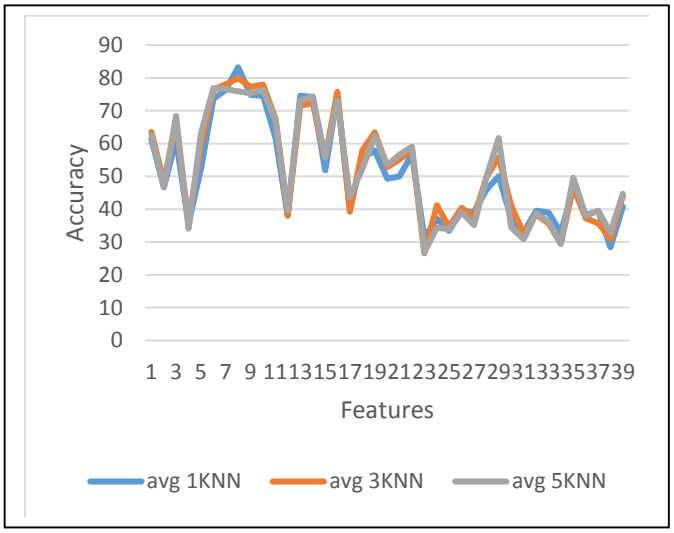

Figure 1. Comparison between average accuracies of 1KNN, 3KNN, 5KNN

- *First and second classes of features:* The selected features to be the best **N** named first class, denoted by **G** and the next best **N** named second class, denoted by $\hat{\mathbf{G}}$. There is a gap between the first and second class of features. In another words there is a distance between the minimum accuracy value of the first class features and maximum accuracy value of the second-class features. Till this point of research **N** should be between 5 and 8 (5<= **N** <= 8), which is a reasonable number.
- *Combinations:* In mathematics, a combination is a method of selecting items from a collection, such that the order of selection does not matter. In this research there was a need to find all combinations from the first and second class of features with all possible sizes. In mathematics, this is called a power set of **S**, which is all possible subsets of **S**. If **S** is a finite set with |**S**| = n elements, then the number of subsets of S is:

$$|\mathcal{P}(\boldsymbol{S})| = \mathbf{2^n} \quad (1)$$

In this research, the word combination will refer to a single set from the power set. If S is the set {x, y, z}, then the power set of S is:

$$\mathcal{P}(\boldsymbol{S}) = \{\{\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\} \quad (2)$$

3.2. Technical Steps

First the average accuracy rate was calculated for each feature by itself. Then the first and second class of features were selected according to the Average of all accuracies rate. The size of features in each class is N. After that the power set of the first class of features only was generated, in this work, N assumed to be equal 8, which results in 255 distinct combinations of different sizes varies between 1 to 8. For each combination 100 run of classification was done using the 80%-20% method for dividing the data and the 1-KNN classification method, after that the average of these runs was calculated. The same procedure was repeated with the second class of features. Then the power set of the union of the first and second classes of features was generated. This means that the size of combinations will be between 1 to 2N. From the results of these three groups of power sets, the power set of the union of the first and second classes of features gave the combinations with the highest accuracy averages. (Fig.2)

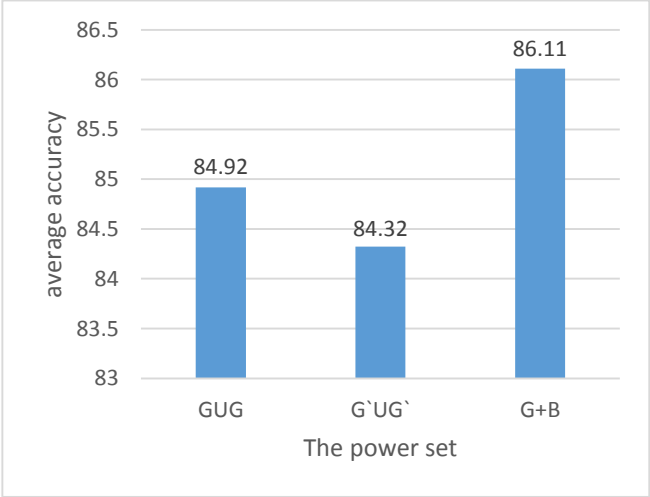

Figure 2. Comparison of different power sets accuracies

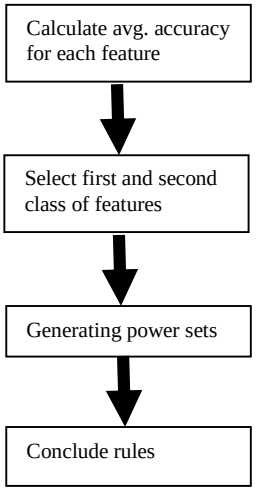

Figure 3: technical steps

The next step was to conclude rules and check their validity. These rules will be used to understand what happens when combining features together.
Symbols definitions:

- Gx and Gy: refers to two separate combinations from the first class only (no common features between them)
- Ĝ x and Ĝy: refers to two separate combinations from the second class only (no common features between them)
- g: refers to a single first class feature

Skin cancer dataset rules results:

1- Gx union Gy > Gx or Gy validation percent is 89.92%
2- Gx union Gy > Gx and Gy validation percent is 28.69%
3- *Ĝx union Ĝy > Ĝx or Ĝy validation percent is 98.02%*
4- Ĝx union Ĝy > Ĝx and Ĝy validation percent is 68.66%
5- Gx union Ĝy > Gx validation percent is 72.56 %
6- Gx union Ĝy > Ĝy validation percent is 70.10 %

The first and second rules results implied that if two first class features combinations were combined in one combination the resultant combination would give a higher classification accuracy than at least one of them (higher than Gx only or Gy only or both) in 89.92 times out of the overall situations when this rule happened (this is a strong rule). However, the resultant combination is higher than both Gx and Gy only in 28.69 times out of the overall situations when this rule happened (weak rule). Conclusion: the combination of two good attributes is not always good, in most of the times it is lower than one of them at least (lower than Gx only or Gy only or both)
Special cases from the previous rules:

- G_x *union* g_6 *union* $g_7 < G_x$
- G_x *union* g_6 *union* $g_8 < G_x$
- G_x *union* g_7 *union* $g_8 < G_x$
- G_x *union* g_6 *union* g_7 *union* $g_8 < G_x$

The Validation percent for all these special cases is 90.32%.This rule says that there are some features if combined with another combination, they would decrease the accuracy of that combination. This rule happened 90 times out of the overall situations when this rule is satisfiable (strong rule). Conclusion: there are some features if they were combined together they make a mess in the classification.
The third and fourth rules results implied that if two second class features combination were combined in one combination, the resultant combination would give a higher classification accuracy than at least one of them (higher than Ĝx only or Ĝy only or both) in 98.02times out of the overall situations when this rule happened (this is a strong rule). However, the resultant combination is higher than both Ĝx and Ĝy in 68.66 times out of the overall situations when this rule happened (also a strong rule). Conclusion : the combination of two second class feature is nearly always better than each one alone , in most of the times(in 68.66 times) it is higher than both of them (higher than Ĝx and Ĝy)
The fifth and sixth rules results implied that if a first and a second class combinations were combined in one combination, the resultant combination would give a higher classification accuracy than the first class features combination alone in 72.56 times of the overall times, and higher than the second class features combination alone in 70.10 times of the overall times. Conclusion: the combination of first and second class features is nearly better than each one alone.

Table1 includes the validation rules percent for the first and second data set. Rules 1, 3, 4, 5, and 6 have similar percentages for the two data sets. Rule 2 percentage for the second dataset was almost doubled. For skin cancer dataset its validation percent was 28.69%, for colon cancer dataset its validation percent was 52.53%. This happened because the average accuracies of features of colon cancer dataset were smaller than the average accuracies of skin cancer dataset. Also the gap between the first and second class of features for colon cancer dataset was about half of the skin cancer dataset gap (table2). This made rule 2 for the second dataset to be more similar to rule 4.

Table 1: rules validation percent for the two datasets used in the research

	Rules	Skin cancer	Colon cancer
1	**Gx union Gy > Gx or Gy**	89.92%	91.17%
2	**Gx union Gy > Gx and Gy**	28.69%	52.53%
3	**Ĝx union Ĝy > Ĝx or Ĝy**	98.02%	97.39%
4	**Ĝx union Ĝy > Ĝx and Ĝy**	68.66%	69.79%
5	**Gx union Ĝy > Gx**	72.56 %	61.02%
6	**Gx union Ĝy > Ĝy**	70.10 %	74.29%

Table 2: gap between first and second group of features

	min avg. accuracy in first group of features	max avg. accuracy in second group f features	gap
skin cancer	73.65	61.68	5.98
colon cancer	57.85	53.38	2.23

3.3. Tools and programming language

Programming language used in the implementation and testing the algorithms is JAVA. The Weka package is used to provide a collection of machine learning algorithms for data analysis and predictive modelling, as KNN. Microsoft Excel was used for analysing and calculating the statistical measurements of the results and data used in the work. Two PCs with different specification were used. The first one has dual-core CPU and 3G RAM, the second has i5-core CPU and 8G RAM.
From measurements results of the algorithm running time on the two PCs, it can be concluded that the size of RAM affects the running time of the algorithm. The algorithm generates all possible combinations of the first and second class features, which needs a RAM

with size greater than 3 in order to work efficiently. (table3)

Table 3: running time

PC	Time(minutes)
PC1	9
PC2	30

4. Experimental Results

Two data sets were used in this work. The first data set is extracted from images of three different skin lesions. These images were obtained by using the fluorescence technique from the Institute of Biophysics (University of Regensburg, Germany). These lesions can be classified into three groups: (1) Actinic Keratosis or malignant melanoma, a type of skin cancer known also as a solar keratosis, can be considered as the first step of the development of skin cancer). (2) Basal Cell Carcinoma is a cancer that begins in the deepest basal cell layer of the epidermis (the outer layer of the skin). (3) Psoriasis is a chronic skin condition that tends to run in families.

This data set contains 39 parameter plus the class label for 50 image of Actinic Keratosis, 50 image of Basal Cell and 65 image of Psoriasis [11]. The second data set is about colon cancer. This data set contains 83 parameter plus the class label for 62 instances. 22 of these instances belong to the positive class (does not have cancer) and the remaining 40 instances belong to the negative class (does have cancer).

Comparing Avg. accuracy: In this part the average accuracies of the combinations from the power set of the first class of features, were compared with the average accuracies of the combinations from the power set of the union of first and second-class of features. The maximum combination accuracy average for the power set of the first class was 84.9 while the minimum accuracy was 69.5; we can conclude that the midpoint of max and min accuracy is 77.2

Each combination accuracy from the power set of first class was compared with the midpoint of the first class. Then each combination accuracy from the power set of first and second class was compared with the midpoint of the first class. See table4 for results. The conclusion of the comparison is that the combinations from the power set of the union of first and second class gave better results than the combinations from the power set of the first class alone. This result shows that the combinations from the power set of first and second class together is containing the best combination of features.

Table 1: Midpoint comparison of skin cancer dataset

	greater than midpoint	smaller than midpoint
power set of first class	80.8%	19.2%
power set of the union of first and second class	98.2%	1.8 %

F-Measures: in statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0 [1]. The traditional F-measure or balanced F-score (F1 score) is the harmonic mean of precision and recall:

$$\boldsymbol{F_1 = 2 * \frac{\text{precision} * recall}{\text{precision} + recall}} \qquad (3)$$

Table 5: F1 average results for skin cancer dataset

	Class A	Class B	Class C
F1	0.538	0.533	0.999

Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. In simple terms, high precision means that an algorithm returned substantially more relevant results than irrelevant, while high recall means that an algorithm returned most of the relevant results.

Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items that were not labelled as belonging to the positive class but should have been).[6]

$$\boldsymbol{Precision = \frac{tp}{tp+fp}} \qquad (4)$$

$$\boldsymbol{Recall = \frac{tp}{tp+fn}} \qquad (5)$$

$$\boldsymbol{True\ negative\ rate = \frac{tn}{tn+fp}} \qquad (6)$$

$$\boldsymbol{Accuracy = \frac{tp+tn}{tp+tn+fp+fn}} \qquad (7)$$

tp.: true positive

fp.: false positive

tn.: true negative

fn.: false negative

It is possible to interpret precision and recall not as ratios but as probabilities:

Precision is the probability that a (randomly selected) retrieved document is relevant.

Recall is the probability that a (randomly selected) relevant document is retrieved in a search. [6]

The F1 test for skin cancer data set used in this work was above 0.5 and this is a good result, see table5. Which indicates that the classification results is reliable.

The size of first and second group is set to be between 5 as minimum and 8 as maximum. Let max be the maximum accuracy value of the first class of features, min be the minimum accuracy value of the first class of features, then:

$$\boldsymbol{gap = \frac{\max - min}{2}} \qquad (8)$$

The bound of G (or $\hat{G}$) is not specified until this point of work.

The main faced difficulty is the difference between the accuracy of one run and the average accuracy. The accuracy of one run is sometimes grater or smaller than the average accuracy of a specific combination by 5 to 20 point for the first data set used in the work. After running the algorithm for 100 times, for each run it gave a different combination to be the best. The average accuracies of the chosen combinations, in the 100 runs, were between 77.14 -85.92. Figure 4 shows the difference between one run accuracy and the average accuracy of three chosen combination in three runs.

Algorithm

1- Calculate Avg. accuracy for each feature by itself
2- Decide size of G (or $\hat{G}$)
3- Decide the bound of G (or $\hat{G}$)
4 - Do G and $\hat{G}$ combinations
5- Do one round of classification
6- Select the combination with highest accuracy

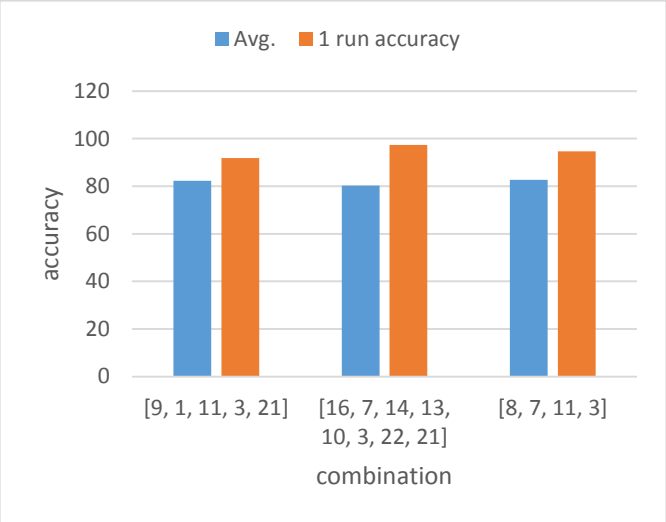

Figure 4: results of 3 runs of the algorithm

5. Conclusion

This paper introduces a new approach for optimization of feature selection. The results of work in this paper showed that if a feature is chosen to be the best by itself, it is not guaranteed that it is the absolute best among a combination of features. Sometimes a combination of good and bad features results in a higher accuracy than a single good feature, or a combination of good features only

From F1 measures, see table5, the validation of calculated classification accuracies for skin cancer dataset is good for the three classes of skin cancer.

At this point of work a first draft of the algorithm has been developed; however, it still needs to be improved to get better results and to overcome the faced problems. The algorithm will be improved by manipulating the following parameters: 1) Number of features to choose 2) Gap between first and second class of features. 3) Bound for the first and second class of features, in other words decide the maximum and minimum value for each class. When the algorithm is updated, it will be compared with GA in order to evaluate its results and performance.

5.1. Future Work

The work on this idea has not been finished yet. The future work will concentrate on generalizing the way of choosing features. The step of choosing first and second class of features will be replaced with choosing one class of features. These features will be chosen depending on a specific conditions.

References

[1] George Hripcsak, Adam S.Rothschild, "*agreement, the f-measure, and reliability in information retrieval*," Journal of the American Medical Informatics Association, vol. 12, pp. 296-298, 2005.

[2] Haleh Vafaie and Ibrahim F. Imam, *"Feature selection methods: genetic algorithms vs. greedy-like search,"* Journal of Communication and Computer.

[3] Isabelle Guyon, Masoud Nikravesh, Steve Gunn and Lotfi A. Zadeh, "*Feature extraction: foundations and applications,"* Springer-Verlag Berlin Heidelberg, vol. 207, 2006.

[4] Isabelle Guyon, Andr´e Elisseeff, "*An introduction to variable and feature selection,"* Journal of Machine Learning, 2003.

[5] Michael Buckland, Fredric Gey, "*The relationship between recall and precision,"* School of Library and Information Studies, University of California, Berkeley, Berkeley, CA 94720, 1994.

[6] Mineichi Kudo, Jack Sklansky, "*Comparison of algorithms that select features for pattern classifiers,"* Division of Systems and Information Engineering, Graduate School of Engineering, Hokkaido University, Kita 13, Nishi 8, Sapporo 060-8628, Japan. Department of Electrical Engineering, University of California, Irvine, CA 92697, USA. The Journal of the pattern recognition society, 1999.

[7] M. Pei1, E. D. Goodman1, W. F. Punch, "*Feature extraction using genetic algorithms,"* Case Center for Computer-Aided Engineering and Manufacturing, Department of Computer Science Genetic Algorithms Research and Applications Group (GARAGe), Michigan State University.

[8] P. Pudil, J. Novovicova, J. Kittler, *"Floating search methods in feature selection,"* Department of Electronic and Electrical Engineering, University of Surrey, Guildford, Surrey GU2 5XH. United Kingdom, 1993.

[9] Ran Gilad-Bachrachy, Amir Navotz,Naftali Tishbyy, "*Margin based feature selection - theory and algorithms,"* School of Computer Science and Engineering .Interdisciplinary Center for Neural Computation, Jerusalem.

[10] Suhail M. Odeh, "*Using an adaptive neuro-fuzzy inference system (AnFis) algorithm for automatic diagnosis of skin cancer,"* Journal of Communication and Computer, 2011.