Feature selection using Forward Weighted Genetic algorithm

Mrs.S.Vydehi,
Professor & Head, Department of Computer Science
Dr.SNS Rajalakshmi College of Arts and Science, Coimbatore, Tamilnadu, India

Dr.M.Punithavalli
Associate Professor, Department of Computer Science
Bharathiar University, Coimbatore, Tamilnadu, India

Abstract- Feature Selection plays an imperative task in mining. It acts as a pre-processing technique which helps in extracting subset of features that are appropriate for mining. It also reduces the task of mining. In this paper the feature selection process is implemented using modified genetic algorithm. The Kullback-Leibler & Shennon Entropy is used as a Pre-processing technique in this paper, then, from that outcome the feature selection process is progressed. The proposed methodology helps in extracting accurate features from the health- care Dataset.

Keywords – Feature Selection, KLSE, Genetic Algorithm, fitness, Weighted fitness.

I. INTRODUCTION

Normally, the original dataset outnumbers in size, from these only a minimal data will be fit for the analysis. Considering the whole dataset for the research process would be inappropriate. The process of selecting the appropriate features from the original dataset is called Feature Selection. This paper briefly explains about the process of extracting the appropriate features using Genetic algorithm. On beforehand the Kullback-Leibler & Shennon Entropy (KLSE) is used where the Kullback finds the similarity measure and Shannon entropy measures the divergence of the data. The outcome from KLSE is piped into Forward Weighted Genetic algorithm (FWG) which helps in acquiring accurate features.

The paper is organized as follows. Proposed FWG is explained in section II. The Outcome of the research Experiment are presented in section III. Conclusion and Future remarks are given in section IV.

II. PROPOSED ALGORITHM

In this paper the work is framed into two process, the former is a pre-processing technique using KLSE and the latter is the feature selection process using genetic algorithm. The framework is shown in figure 1.

KLSE –

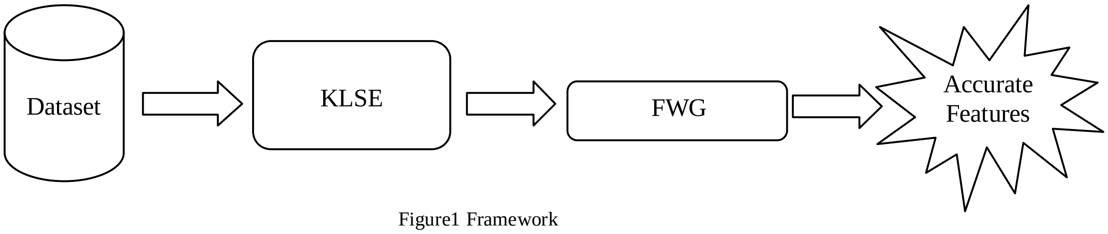

Figure1 Framework

The Kullback–Leibler is also referred as information divergence or relative entropy. The KL is a non-symmetric measure of the difference between two probability distributions *such as P* and *Q*. Specifically, the Kullback–Leibler divergence of *Q* from *P is*, denoted as $D_{KL}(P||Q)$. The KL distance) is termed as a natural distance that function from a "original or true" probability distribution (p), to a "target" probability distribution (q).

For discrete (not necessarily finite) probability distributions, p= $\{p_{1,}...p_n\}$ and q= $\{q_1, ..., q_n\}$, the KL-distance is defined to be

$$KL(p, q) = \Sigma_i\, p_i.\ \log_2({}^{p}{}_{i} \,/\, {}_{qi})$$

For continuous probability densities, the sum is replaced by an integral.

$$KL(p, p) = 0$$

$$KL(p, q) \geq 0$$

On the other hand the Shannon entropy is defined as a measure of the dissimilarity or divergence associated with a random variable. The Shannon entropy measures the uncertainty or divergence in a random variable. It is also referred as a “measure of surprise”. The Shannon entropy of (A) is given by

$$H(\boldsymbol{A}) = -\sum_{i=1}^{n} p_i \log_2 p_i.$$

The entropy quantifies the expected value of the information contained in a dataset and brings out the measure of dissimilarity between the dataset.

B. Forward Weighted Genetic Algorithm –

The Genetic algorithm is represented as a Search technique which helps in deriving approximate solutions to optimization and search problems. Conventionally, it is referred as Global search heuristics or Evolutionary Algorithm. The algorithm incorporates the operators like mutation, selection, and crossover (also called recombination). In Genetic space the dataset are represented in the form of 0’s and 1’s and Fitness function is used for feature selection. The steps in FWG is as follows.

Step 1 : Initial Population is generated from KLSE

Step 2 : In each generation, the fitness f = a_1+ P / N of every individual data in the population is evaluated,.

Step 3 : Weightage is calculated for each fitness.

Step 4 : Then the selected features are modified to form a **new population**.

Step 5 : Based on the weights the features are selected

Step6 : The algorithm terminates when a satisfactory fitness level has been reached for the population

Normally, in the traditional genetic algorithm the fitness are calculated for each data, wherein the fitness 1 > fitness 2 then fitness 1 is discarded, herein, the fitness 1 is not discarded instead weightage is calculated for each fitness and features are selected based on the weightage. The advantage of this method is, it minimizes the data loss when compared to the conventional outcomes. The experimental result of the proposed work is explained in the next section.

III. Experiment and Result

The experiment is carried out on a medical database (heart disease). The dataset constitutes 76 attributes in which 14 attributes are used. The results are measured with the parameter like accuracy and specificity are used to measure the performance. The accuracy of the cluster is given by

Accuracy = (TP + TN) / (TP+ TN+ FP+FN)

The specificity is defined as follows

Specificity= (TN) / (TN+FP)

- **True Positive** (TP)→ **Positive instances** correctly classified as **Positive.**
- **False Negative (FN)→ Positive instances** incorrectly classified as **Negative.**
- **True Negative (TN)→ Negative instances** correctly classified as **Negative.**
- **False Positive (FP)→ Negative instances** incorrectly classified as **Positive.**

Table 1. Methods & Results

	KL	KLSE	FWG + KL
Precision	0.8423	0.8921	**0.9234**
Sensitivity	0.8813	0.9125	**0.9362**
Accuracy	0.8567	0.9032	**0.9241**
Specificity	0.1187	0.0875	**0.0638**

The Table 1 shows the results of the Existing method Kullback–Leibler , Kullback–Leibler Shannon entropy without feature selection and Forward Weighted Genetic Algorithm with feature selection. The outcome of the proposed method excels in all the parameters compared to the existing method. The precision, sensitivity and accuracy of FWG+KL is desirably increased whereas the specificity measure is decreased accordingly compared to the existing methods. The outcome of the work proves that the proposed FWG+KL brings out the accurate measure in the heart disease dataset. The graphical representation of the proposed work is shown in figure 2.

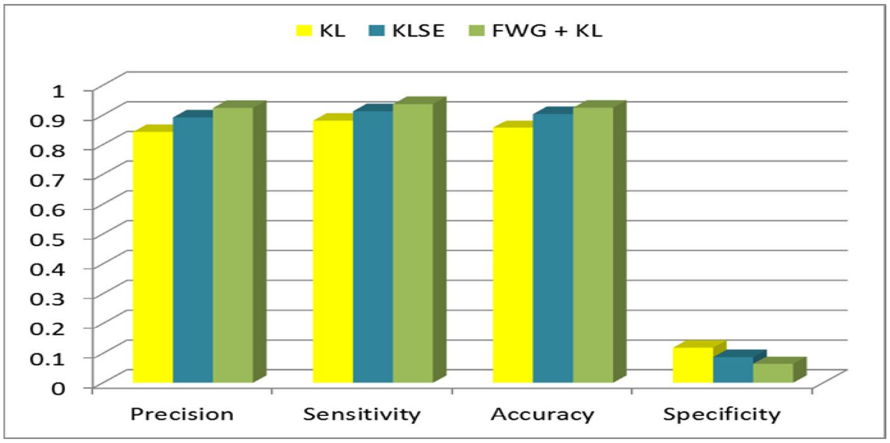

Figure 2. Result Statistics

IV.CONCLUSION

The conventional Genetic Algorithm normally results in Local optima problem, where the fitness are discarded when higher fitness is derived. But, in the proposed method the weightage of each fitness is calculated, based on the weightage the features are selected. The proposed FWG+KL brings out the accurate features also it minimizes the data loss. Further, the work can be progressed to clustering process, also, the feature selection process can be optimized.

REFERENCES

[1] S. Abiteboul, P.C. Kanellakis, and G. Grahne, “On the Representation and Querying of Sets of Possible Worlds,” *Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD), 1987*

[2] Banerjee A. and Jegelka S. and Sra S. “Approximation Algorithms for Bregman Co-clustering and Tensor Clustering”, 2008.

[3] Bin Jiang, Jian Pei, Senior Member, IEEE, Yufei Tao, Member, IEEE, and Xuemin Lin, Senior Member, IEEE, “Clustering Uncertain Data Based on Probability Distribution Similarity” *IEEE Transactions On Knowledge And Data Engineering*, Vol. 25, No. 4, APRIL 2011.

[4] Arindam Banerjee Srujana Merugu, Inderjit S. Dhillon Joydeep Ghosh, “Clustering with Bregman Divergences” *Journal of Machine Learning Research 6 (2005) 1705–1749 Submitted 10/03; Revised 2/05; Published 10/05.*

[5] Ye Yuan , Guoren Wang , Lei Chen , Haixun Wang , “Efficient Keyword Search on Uncertain Graph Data” *IEEE Transactions on Knowledge and Data Engineering,* Issue No.12 - Dec. (2013 vol.25) pp: 2767-2779, http://doi.ieeecomputersociety.org/10.1109/TKDE.2012.222

[6] Chunhui Zhu ; Tsinghua Univ., Beijing, China ; Fang Wen ; Jian Sun,”A rank-order distance based clustering algorithm for face tagging”, *IEEE International Conference*, June 2011.

[7] H.-P. Kriegel and M. Pfeifle, Density-Based Clustering of Uncertain Data,Proc. ACM SIGKDD Intl Conf. Knowledge Discovery in Data Mining (KDD) ,2005.

[8] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip, (2006) Efficient Clustering of Uncertain Data,,Proc. *Sixth Intl Conf. Data Mining (ICDM).*

[9] B.W. Silverman (1986) Density Estimation for Statistics and Data Analysis, Chapman and Hall.

[10] P.B. Volk, F. Rosenthal, M. Hahmann, D. Habich, and W. Lehner, (2009) Clustering Uncertain Data with Possible Worlds,, Proc. IEEE Intl Conf. Data Eng. (ICDE).

[11] H. Peng, F. Long, and C. Ding. “Feature selection based on mutual information: Criteria of maxdependency, maxrelevance, and min- redundancy.”*IEEE Transactions on Pattern Analysis and Machine Intelligence*,27(8):1226–1238,2005.

[12] Y. Sun, M. Robinson, R. Adams, R. teBoekhorst, A.Rust, and N.Davey. Using feature selection filtering methods for binding site Predictions. volume 1, pages566–571, July2006.

[13] Jason VanHulse,Amri NapoL itano, T.M,Khoshgoftaar “Feature Selection with High-Dimensional Imbalanced data” IEEE Conf. Data mining Tech.2009