ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 7, September 2012

Comparative Study of Different Feature Selection Algorithm in Small dataset Among KNN, FUZZY & GENETIC Algorithm.

Lal Bahadur pandey, Siddharth Choubey

Abstract— **In the earth of curse of dimensionality feature selection acting a very important role in reducing the entire feature collection with the partial subset of features Falling the number of features pave way for a variety of advantaged as well as simplify the assignment. Feature selection means finding the appropriate set of features which will give the majority of it to the solution with minimum or null error rate. Selected features are to be tested with the help of classifiers, so that the separation of selected features can be proved to be most favorable when compared to other features subsets separately as well as a group. Genetic algorithms are now days play a very important role among any other method in selecting the features based on the Theory of Evolution and on the "Survival of the fitness". It is a heuristic approach. To cooperate with the GA approach we have the classifiers which will go hand in hand to bring out the final set of features along with their calculation accuracy. In this paper we have analyze, classifiers and compare them with their act and the unit of correctness.**

Keywords—**Feature Selection, Knn, Genetic Algorithm, Fuzzy Rough Set classification.**

I. INTRODUCTION

In recent years, knowledge discovery in databases (KDD) and data mining became more and more important as the number and size of reachable data sources raise at extraordinary rates, and all industry segments from financial sectors to telecommunications rely on analysis of data to compete.By now, KDD is defined as a complex, iterative, and interactive process which requires more then loading data into an intelligent algorithm and waiting for automatically achieved result to deploy them.

It is clear that data preparation is one of the most important and time consuming phases in KDD. Research tasks such as data selection, data cleaning, data construction, data integration, and data formatting often decide the success of data mining arrangements.

In context of machine learning system there are several reasons to decrease the original training set to smaller one. The first of them is to decrease the noise in original data set because some learning algorithms may be noise fragile. The second reason to shrink the training set is to decrease the amount of calculation. The third and comparatively new reason to use Feature selection appeared together with new prototype selection algorithm.

Feature selection is a focusing task in the data preparation phase of KDD. It is one of the useful means of data reduction. Next section describes few feature selection algorithms then after the comparison study of these algorithms. Finally, in section 4 we reached to conclusion.

II. FEATURE SELECTION ALGORITHMS

A. *KNN Method for Feature Selection:*

K -Nearest Neighbor algorithm is an instance based supervises learning that has been used in many applications of data mining, statistical pattern recognition, image processing etc… KNN algorithm accomplishes very good performance in their experiments on unusual datasets. It is one of the most popular algorithms for pattern recognition. KNN algorithm is proved to accomplish good results in the experiments on unusual datasets. KNN algorithm improves the classification performance. KNN algorithm takes the k-neighbors and calculates the distance to classify the test samples instead of considering all the test samples. In KNN method k the umber of samples from the training set is generated and considered as initial population. The fitness value for the chromosomes is calculated so that the best fitness (highest fitness value) individuals are selected and they are stored. The most natural choice of the fitness function is some measure of the classification performance of the KNN rule. The value that obtained by calculating the distance between the best individuals is stored as global maximum. The process continues with a set of steps for a required no. of limited iterations.

The steps that are involved are:

1. Create a population.
2. Apply crossover and mutation operators.
3. Find the local minima.
4. Compare the value obtained for local minima with the existing global maximum.
5. The higher value is taken as global maximum.
6. The steps from 1 – 5 are repeated until the required number of iterations or the required global maximum.

It is proved that the classification correctness of KNN method is above 93%.

Manuscript received Sep 15, 2012.

***Lal Bahadur Pandey**, ME Scholar, Deptt. Of Computer Science & Engineering, Shri Shankaracharya College of Engineering & Technology Bhilai,India,+91-9981757033.*

***Siddharth Choubey**, Deptt. Of Computer Science & Engineering, Shri Shankaracharya College of Engineering & Technology, Bhilai, India, +91-9993373778.*

B. Fuzzy Sets for Feature Selection:

A rough set is a formal estimate of a crisp set, in terms of a pair of sets which give lower and upper estimate of the original set. The lower and upper estimated sets are crisp sets. Fuzzification is a process in which the data is represents using a function called membership function. The quantitative value is transfer into fuzzy sets. The two membership values (yes or no) are produced for each attribute from the membership function. Fuzzy rough sets are an extension for rough sets. They are applied in situations where classes are described as fuzzy sets on the feature space. Categorization exactness can also be obtained as a measure of how the data is classified in fuzzy recognition problems. Fuzzy rough sets for feature selection play an important role in many applications where the data under consideration is not discrete. In a classical rough set theory it is possible to consider real valued data. The inductive learning of a fuzzy rule-based classification system first of all determines a set of fuzzy rules from the set of instances and patterns. Each of these patterns is described by a set of features which are called as variables or characteristics.

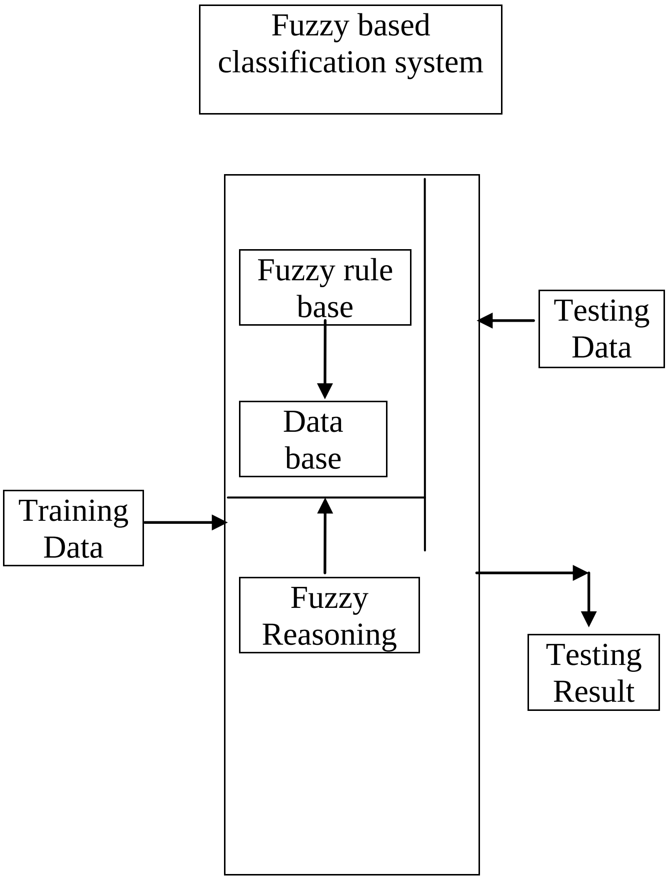

Figure 1: fuzzy based feature selection

C. Genetic Algorithm Approach:

Genetic Algorithm is an adaptive heuristic method of global-optimization searching and simulate the performance of the evolution process in environment. It maps the searching space into a inherent space. That is, every possible key is determined into a vector called a chromosome. One element of the vector represent a gene. All of the chromosomes make up of a population and are predictable according to the fitness function. A fitness value will be used to measure the "fitness" of a chromosome. Initial populations in the genetic process are accidentally created. GA then uses three operators to produce a next generation from the current generation: replica, intersect, and metamorphosis. GA eliminates the chromosomes of low fitness and keeps the ones of high fitness. This whole process is repetitive, and more chromosomes of high fitness move to the next production, until a good chromosome (individual) is found. The main objective of genetic feature selection stage is to reduce the dimensionality of the problem before the supervised inductive learning process. Among the many wrapper algorithms used, the Genetic Algorithm (GA), which solves optimization problems using the methods of evolution, specifically "survival of the fittest", has proved as a promising one. GA evaluates each individual's fitness as well the superiority of the solution. The fitter individuals are more eligible to enter into the next generation as a population. After a required number of generations the final set of optimal population with fittest chromosomes will emerge giving the solution.

Selection algorithms include a few vital components such as the initial population or the starting point in the feature space, search procedure, evaluation function or the fitness function and terminating condition. Initially all the features are taken into consideration. Later the subset of features can be found by evaluating all the possible solutions. These search procedures that are practical to implement are not definite to find the optimal subset of features. Genetic algorithms is one of the search actions which simulate natural evolution mechanisms of natural selection and natural inheritance are used to in order to find solution to a problem. The basic operations involved in genetic algorithm are maintaining a population of solutions, selecting better solutions for recombination with each other and use their offspring to replace poorer solutions. GA is a combinatorial search technique based on random and probabilistic measures. Features are selected using a fitness function and then combined via cross-over and transmutation operators to produce the next generation of subsets.

III. RESULT AND CONCLUSION

The graphical representation given below depicts the performance of all the combination with respect to the correctness of the model. It is clearly shown that the correctness in the case of k-nearest neighbor algorithms networks is the highest for small data set.

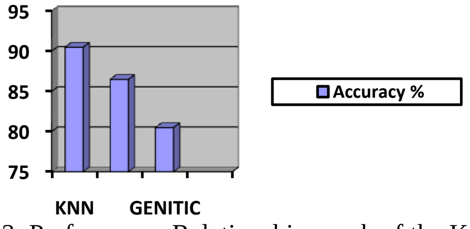

Figure 2: Performance Relationship graph of the Knn,

Fuzzy, Genetic algorithm and the classifiers.
KNN algorithm is an evolutionary algorithm that is capable of selecting the features for classification. This paper focus on the performance of the classifiers for the same data (small dataset) and the study shows that the KNN algorithms classifier prove to be the best when compare with other classifiers.

REFERENCES

1. Changjing Shang and Qiang Shen ,"Aiding Classification of Gene Expression Data with Feature Selection: A Comparative Study", International Journal of Computational Intelligence Research. ISSN 0973-1873 Vol.1, No.1 (2005), pp. 68–76 °c Research India Publications http://www.ijcir.info.

2. Yubin Kuang , "A Comparative Study on Feature Selection Methods and Their Applications in Causal Inference", Thesis for a diploma in Computer Science, 30 ECTS credits Department of Computer Science, Faculty of Science, Lund University(2009).

3. Ying Liu , "A Comparative Study on Feature Selection Methods for Drug Discovery", Georgia Institute of Technology, College of Computing, Atlanta, Georgia 30322, *J. Chem. Inf. Computer. Sci.* 2004, *44,* 1823-1828.

4. J.R. Méndez1, F. Fdez-Riverola1, F. Díaz2, E.L. Iglesias1, and J.M. Corchado "A Comparative Performance Study of Feature Selection Methods for the Anti-spam Filtering Domain", 1 Dept. Informática, University of Vigo, Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain {moncho.mendez, riverola, eva}@uvigo.es 2 Dept. Informática, University of Valladolid, Escuela Universitaria de Informática, Plaza Santa Eulalia, 9-11, 40005, Segovia, Spain fdiaz@infor.uva.es 3 Dept. Informática y Automática, University of Salamanca, Plaza de la Merced s/n, 37008, Salamanca, Spain corchado@usal.es.

5. Pengpeng Lin, Jun Zhang, Ingrid St. Omer, Huanjing Wang, and Jie Wang,"A comparative Study on Data Perturbation with Feature Selection",Proceedings of the international multiConference of engineers and Computer scientists(2011).

6. Jerzy Stefanowski," An Experimental Study of Methods Combining Multiple Classi¯ers - Diversi¯ed both by Feature Selection and Bootstrap Sampling", Institute of Computing Science, Pozna¶n University of Technology, ul. Piotrowo 3A, 60{965 Pozna¶n, Poland,Jerzy.Stefanowski@cs.put.poznan.pl.

7. *Tao Li, Chengliang Zhang and Mitsunori Ogihara," A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression", Computer Science Dept., University of Rochester, Rochester, NY 14627-0226, taoli@cs.rochester.edu,* Bioinformatics Advance Access published April 15, 2004.

8. Isabelle Guyon, Clopinet, California ,Constantin Aliferis, Vanderbilt University, Tennessee Andr¶e Elissee®, IBM ZÄurich, Switzerland" Causal feature selection"(March 2, 2007).

9. Lei Yu, Huan Liu ," Feature Selection for High-Dimensional Data:A Fast Correlation-Based Filter Solution", Department of Computer Science & Engineering, Arizona State University, Tempe, AZ 85287-5406, USA, leiyu@asu.edu, hliu@asu.edu.

10. Gongde Guo[1], Daniel Neagu[1] and Mark T.D. Cronin[2] ," Using *k*NN Model for Automatic Feature Selection" [1] Department of Computing, University of Bradford, Bradford, BD7 1DP, UK ,{G.Guo, D.Neagu}@Bradford.ac.uk , [2] School of Pharmacy and Chemistry, Liverpool John Moores University, L3 3AF, UK M.T.Cronin@Livjm.ac.uk .

11. Ruck D.W. S.K.Rogers and M.Kabrisky(1990) Feature Selection using a multilayer perceptron", Journal of Neural Network Computing. 40-48.

12. Sikora R., Piramuthu S., 2007. Framework for efficient feature selection in genetic algorithm based data mining. European Journal of Operational Research, 180(2), pp. 723-737.

13. Swiniarsk.R.W.i, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Lett. 24 (6) (2003) 833–849.

14. Verma, B. K., A neural network based technique to locate and classify micro calcifications in digital mammograms, IEEE World Congress on Computational Intelligence, WCCI'98, Anchorage, USA, 1998, pp. 2163-2168.

15. Weston J., Mukherjee S., Chapelle O., Ponntil M.,Poggio T., Vapnik V., 2001. Feature selection for SVMs.Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA.

16. Zhong.N, J. Dong, S. Ohsuga, Using rough sets with heuristics for feature selection, J. Intell. Inf. Syst. 16 (3) (2001) 199–214.

17. Zilberstein.S Using Anytime Algorithms in Intelligent Systems, AI Magazine, vol. 17, no. 3, pp. 73-83, 1996.

Lal Bahadur Pandey, B.E., M.E.(Pursuing) in Computer Technology & Application from Shri Shankaracharya College of Engineering & Technology, bhilai.India. , research areas are Data mining, parallel processing.

Mr. Siddhartha Choubey, received the M.Tech (C.T.A) from SSCET, CSVTU University, Bhilai, India in 2008. He is pursuing his Ph.D. in Image Processing from MATS University Raipur. He is working as Associate Professor in the department of Computer Science and Engg. in Shri Shankaracharya college of Engg. & Tech. Bhilai. His research interests include image processing and applications such as pattern recognition.